Statistical and dynamical properties of large cortical network models: insights into semantic memory and language

Emilio Kropff

July 2007

Contents

List of Figures

List of Tables

Acknowledgements

In the year 2002 I found myself in a desperate situation. In the worst part of the economic crisis in Argentina, professors in my university had temporally fled abroad and the perspectives of pursuing a scientific carrier, as I had always wished, were vanishing. Some years after my situation is pretty much the opposite: I am back into the track and exciting academic opportunities stand ahead. I want to thank Alessandro, responsible and mentor of this giant step.

Abstract

This thesis introduces several variants to the classical autoassociative memory model in order to capture different characteristics of large cortical networks, using semantic memory as a paradigmatic example in which to apply the results. Chapter 2 is devoted to the development of the sparse Potts model network as a simplification of a multi modular memory performing computations both at the local and the global level. If a network storing p global patterns has N local modules, each one active in S possible ways with a global sparseness a, and if each module is connected to c_M other modules, the storage capacity scales like $\alpha_c \equiv p_{max}/c_M \propto S^2/a$ with logarithmic corrections. Chapter 3 further introduces adaptation and correlations among patterns, as a result of which a *latching* dynamics appears, consistent in the spontaneous hopping between global attractor states after an initial cue-guided retrieval, somehow similar to a free association process. The complexity of the latching series depends on the equilibrium between self-excitation of the local networks and global inhibition represented by the parameter U. Finally, Chapter 4 develops a consistent way to store and retrieve correlated patterns, which works as long as any statistical dependence between units can be neglected. The *popularity* of units must be introduced into the learning rule, as a result of which a new property of associative memories appears: the robustness of a memory is inverse to the information it conveys. As in some accounts of semantic memory deficits, random damage results in selective impairments, associated to the entropy measure S_f of each memory, since the minimum connectivity required to sustain its retrieval is, in optimal conditions, $c_M \propto pS_f$, and still proportional to pS_f but possibly with a larger coefficient in the general case. Present in the entire thesis, but specially in this last Chapter, the conjecture stating that autoassociative memories are limited in the amount of information stored per synapse results consistent with the results.

Chapter 1

General Introduction

Though suffering at least a 1000-fold increase along the history of evolution, the mammalian cortex has conserved a remarkable degree of anatomical homogeneity, not present in more archaic regions of the brain. This observation suggests that the original success of cortical-based computations relays on a number of presumably simple and local principles, applicable to a geometrically increasing amount of tissue. The first of these principles to be assessed by neuroscientists was Hebbian plasticity, which remains until now the most general and well-studied characteristic of learning in the cortex [Hebb, 1949]. It is possible that other principles have as functional units local networks rather than single neurons, as suggested by a growing amount of evidence showing columns, minicolumns and hypercolumns as fundamental structures of information processing, a claim supported by the evolutionary perspective of cortex developing through the incorporation of new local networks rather than through their internal modification [Buxhoeveden and Casanova, 2002, Rakic, 1995]. While research on particular architectural details of different cortical areas has its maximum impact in the accurate description of low level computations, it is unveiling these general principles, instead, that might open us a door to understand cognitive aspects of the brain.

In their enlightening book, Braitenberg and Schuz [Braitenberg and Schuz, 1991] focus on several distinctive properties of the cortical anatomy to speculate about its possible computational function: the number of neurons and cortico-cortical connections is by far

larger than the number of input and output fibers; the excitatory synapses, which are the majority, are plastic and weak, connecting neurons of the same kind; information seems to be mixed following the principle of maximum convergence and divergence. All these elements put together suggest that the primary function of the cortex as a whole is that of forming long range associative memories.

Early in the history of neuroscience it became apparent that the cortex is divided into areas and subareas, many of them, especially in the somato-sensory and motor domains, associated to well defined functions. A network in which long-range synapses connect cortical neurons positioned far apart from each other is, thus, a multi-modular network, where modules have the double task of performing the corresponding local processing and participating in global computations. The discovery of mirror neurons in primates and the development of embodied theories of cognition have contributed with evidence supporting this dual interpretation of neural activity, even for high order processes such as semantic memory and language.

A way to simplify such a large and complex network performing at the same time local and global memory computations is the Potts model. In a Potts model each unit represents a local network activating one of S alternative local memory codes. The network connecting the different units through long range synapses is in charge of storing global memory states, thought of as particular combinations of local activation states. Chapter 2 introduces our studies of Potts networks [Kropff and Treves, 2005]. The initial step is to complete the founding work of Kanter [Kanter, 1988] by calculating the storage capacity of the simplest Potts network in the limit of $S \gg 1$. Next, a zero state is introduced in the Potts units, i.e. a state of activation of the local network that is not relevant for the global computations. The model must be defined in such a way that the performance is close to optimal, a point in which the previous literature is a bit confusing. The scaling of the storage capacity is then obtained by the traditional methods of replica analysis and highly diluted approximation, relating the result with the conjecture following which any optimized hebbian network can

store not more than a fraction of a bit of information per synapse.

Though it has been claimed with some fundament that memory formation is its primary function, it is largely accepted that cortex does much more than that. In particular, the human cortex is involved in high order processes such as problem solving or language. Chapter 3 analyzes an example of how the evolutionary improvement of a semantic memory network could have unexpectedly resulted in a new and more complex type of computation [Kropff and Treves, 2007a]. Indeed, introducing adaptation and correlation among patterns into a Potts model generates a new dynamical state - latching - characterized by the spontaneous chaining in time of related attractor states. A transition toward this self-sustained dynamics has been proposed [Treves, 2005] in relation to the proliferation of connections (and presumably of memory states) during the evolution of humans. The Chapter analyzes the complexity of the resulting 'symbolic' series in terms of the equilibrium between local self-excitation and global inhibition.

The networks presented in Chapters 2 and 3 can store correlated patterns only by virtue of a finite size effect. This explains the fact that the storage capacity in simulations [Treves, 2005] is much lower than the theoretical predictions for uncorrelated patterns [Kropff and Treves, 2005]. In fact, the literature lacks a purely hebbian learning rule that permits the effective storage and retrieval of general non trivially correlated memories. On the other side, several recent theories point out at correlation in the cortical representation of concepts as the key element to account for selective semantic memory deficits. Chapter 4 introduces a modification to previous models that solves the problem of storing and retrieving correlated patterns as long as different neurons can be regarded as statistically independent [Kropff and Treves, 2007b]. The resulting weights resemble those of the BCM learning rule [Bienenstock et al., 1982]. The estimation of the storage capacity in the highly diluted limit shows the side-effects of storing correlated memories: the robustness associated to the attractor states is not homogeneous, which results in selectivity in retrieval, just as observed in patients with semantic memory deficits.

The learning rule in Chapter 4 can be trivially extended to the Potts network of Chapter 2, presumably with little or no change in the overall behavior. In addition, it can host latching dynamics as presented in Chapter 3, completing the picture of statistical and dynamical properties of large scale networks in the cortex, of which those sustaining semantic memory are just a handy example to analyze.

Chapter 2

A Potts model of semantic memory

2.1 Introduction

Hebbian associative plasticity appears to be the major mechanism responsible for sculpting connections between pyramidal neurons in the cortex, for both short- and long-range systems of synapses. This and other lines of evidence [Braitenberg and Schuz, 1991] suggest that autoassociative memory retrieval is a general mechanism in the cortex, occurring not only at the level of local networks, but also in higher order processes involving many cortical areas. These areas are often regarded both from the anatomical and from the functional point of view as distinct but interacting modules, indicating that in order to model higher order processes we must first understand better how multimodular autoassociative memories may operate. In a class of models conceived along these lines, neurons in local modules, interconnected through short-range synapses, are capable of retrieving local activity patterns, which combined across the cortex and interacting through long-range synapses, compose global states of activity [O'Kane and Treves, 1992]. Since long-range synapses are also modified by associative plasticity, these states can be driven by attractor dynamics, and such networks are capable of retrieving previously learned global patterns.

This could serve as a simple model of semantic memory retrieval. The semantic memory system, as opposed to episodic memory, stores composite concepts, e.g. objects, and their

relationships. Although information about distinct features pertaining to a given object (e.g. its shape, smell, texture, function) may be processed in different areas of the cortex, a cue including only some of the features, e.g. the shape and color, may suffice to elicit retrieval of the entire memory representation of the object. Imaging studies show that, though distributed across the cortex, this activity is sparse and selective, and might involve regions associated to the concept being retrieved, even if not directly activated by the cue [Pulvermuller, 2002]. This process could well fit a description in terms of autoassociative multimodular memory retrieval. In this perspective, while a local module codes for diverse values of a given feature, a combination of features gives rise to a concept, which behaves as an attractor of the global network and is thus susceptible of retrieval. The two-level description that characterizes this view is the principal difference with other attempts to describe semantic memory in terms of featural representations [McRae et al., 1997].

In order to reduce the complexity of a full multimodular model [O'Kane and Treves, 1992, Fulvi Mari and Treves, 1998] one can consider a minimal model of semantic memory, which can be thought of as a global autoassociative memory in which the units, instead of representing, as usual, individual neurons, represent local cortical networks retrieving one of various (S) possible states of activity. The combined activity of these units generates a global state, which follows a retrieval dynamics. The first question arising from this proposal is how the global storage capacity of such a network is related to the different local and global parameters.

In the following section of this Chapter we present the model in mathematical terms. In the third section we compare, through a simple signal-to-noise analysis, different model variants proposed in the literature and extract the minimum requirements for a network of this kind to perform efficiently in terms of storage capacity. In the fourth section we analyze with more sophisticated techniques the simplest model endowed with a large capacity (the sparse Potts model) and, in particular, interesting cases such as the very sparse and the high-S limits. Following this we study modifications to the model that make it more realistic in

terms of connectivity. Finally, we relate the results from the previous sections to a simple information capacity analysis.

2.2 S-state fully connected networks

Autoassociative memories are networks of N units connected to one another by weighted synapses. These synapses are trained in such a way that the network presents, in the ideal case, a number p of preassigned attractor states, also called stored patterns, or memories, represented by the vectors $\vec{\xi}^{\mu}$, with $\mu = 1...p$. If the state of the network is forced into the vicinity of an attractor (e.g., by presenting a cue correlated with one of the stored patterns) the natural dynamics of the network converges toward the attractor, in state space, and the memory item is said to be retrieved. A substantial amount of the literature on attractor networks is devoted to study the relationship between the number and type of stored patterns and the quality of retrieval.

The state of a network at a given moment is given by the state of each of its units, σ_i for $i = 1...N$. The first quantitative analyses of autoassociative memories were of binary models [Amit, 1989], in which units could reach two possible states, $+1$ (active unit) and -1 (inactive unit), resembling Ising $\frac{1}{2}$ spins. In our case, in which units do not represent single neurons but rather local networks, we want active units to be able to reach one of S possible states, while inactive units remain in a 'zero' state. We thus choose the notation $\sigma_i = k$ for an active unit in state k and $\sigma_i = 0$ for an inactive unit. This particular choice has no effect on the results, since all quantities can be transformed to some other notation. On the other hand, the stored patterns $\vec{\xi}^{\mu}$ can be simply thought of as special states of the network. For this reason, it is natural to choose the same kind of representation for the activity of a unit i in pattern μ, ξ_i^{μ}.

Although in the first binary models of autoassociative memories patterns where constructed with a distribution of equally probable active and inactive units, the search of an accurate description of activity in the brain made it necessary to introduce sparse representa-

tions. This property of autoassociative memories is described by the sparseness a, defined as the average activity (the average fraction of active units) in the stored patterns. In our case, because we are assuming all S different activity states to be equally probable, we consider patterns defined by the following probability distribution:

$$\begin{aligned} P(\xi_i^\mu = 0) &= 1 - a \\ P(\xi_i^\mu = k) &= \tilde{a} \equiv \frac{a}{S} \end{aligned} \tag{2.1}$$

for any active state k. In this way the probability to find an active unit in a pattern is the sparseness a. For *sparse* codes, this quantity is closer to 0 than to 1.

Following the assumption of Hebbian learning and, as is usual for a simplified analysis, symmetry in the weights ($J_{ij} = J_{ji}$), a general form for the weights is

$$J_{ij}^{kl} = \frac{1}{E} \sum_{\mu=1}^{p} v_{\xi_i^\mu k} v_{\xi_j^\mu l}, \tag{2.2}$$

where E is some normalization constant and v_{mn} is an operator computing interactions between two states.

As one can notice, the long-range synapse weights in Eq. 2.2 have different values for different pre- and post- synaptic states k and l. In this way we do not intend to model the actual distribution of synapses going from one cortical area to another (since they connect neurons and not abstract states), but rather the general mechanism of communication between these areas. In a recent study [Mechelli et al., 2003], the authors have raised the issue of finding the most suitable description of global cortical networks in terms of single long-range synapses connecting distant local areas. Applying statistical tools (Dynamic Causal Modeling), they propose that MRI data can be described as produced by networks with category specific forward connections, roughly the kind of connections modelled by Eq. 2.2.

The state of generic unit i is determined by its local fields h_i^k, which sum the influences by other units in the network and are defined as

$$h_i^k = \sum_{j \neq i} \sum_l J_{ij}^{kl} u_{\sigma_j l} - U(1 - \delta_{k0}), \tag{2.3}$$

where we introduce the operators u_{mn}, analogous to v_{mn}, and a second (threshold) term, which has the function of regulating the activity level across the network [Buhmann et al., 1989, Tsodyks and Feigel'Man, 1988]. The unit i updates its state σ_i, with an asynchronous dynamics, in order to maximize the local field $h_i^{\sigma_i}$. In the general case, the probability to choose the state k is defined as

$$P(\sigma_i = k) = \frac{\exp(\beta h_i^k)}{\sum_{l=0}^{S} \exp(\beta h_i^l)}, \tag{2.4}$$

where β is a parameter analogous to an inverse temperature.

Finally, we can include all of these elements, as is usual for the study of attractor networks, into a Hamiltonian framework. The Hamiltonian representation of binary networks can be extended to S-state models as

$$H = -\frac{1}{2} \sum_{i,j \neq i} \sum_{k,l} J_{ij}^{kl} u_{\sigma_i k} u_{\sigma_j l} + U \sum_i \sum_{k \neq 0} u_{\sigma_i k}. \tag{2.5}$$

Note that for the case $S = 1$, Eq. 2.5 generalizes the Hamiltonians used in binary networks, given appropriate definitions of the weights J_{ij}^{kl} and of the operators u_{mn}.

We now specify a form for the u_{mn} and v_{mn} operators. In the simplest and most symmetric case these operators have two alternative values, depending on whether m and n are equal or different states

$$\begin{aligned} u_{mn} &= (\kappa_u \delta_{mn} + \lambda_u) \\ v_{mn} &= (\kappa_v \delta_{mn} + \lambda_v)(1 - \delta_{n0}), \end{aligned} \tag{2.6}$$

where we have introduced four parameters. Particular choices for these parameters define the different models in which we are interested, including several proposed in the literature. In the v operators, which define the value of the weights, we have included a factor which ensures $J_{ij}^{kl} = 0$ if either k or l are the zero state, to implement the idea that Hebbian learning occurs only with active states. As we will see below, this appears to be a crucial element in the model.

2.3 Signal-to-noise analysis

We now show that, within the group of models defined in the previous section, there is a family (which we call 'well behaved') that exploit multiple states and sparseness in an optimal way in terms of storage capacity or, as usual, of $\alpha \equiv p/N$. We begin by applying an adjusted version of the arguments developed in [Buhmann et al., 1989].

A signal-to-noise analysis is a simplified way to estimate the stability of stored patterns by studying what happens to a generic unit i during the perfect retrieval of a given pattern, assessing whether the state of this unit is likely to be stable or not. We can choose this retrieved pattern to be $\vec{\xi}^1$ without loss of generality. Eq. 2.3 can then be rewritten as

$$h_i^k = \frac{1}{E} v_{\xi_i^1 k} \sum_{j \neq i} \sum_l u_{\sigma_j l} v_{\xi_j^1 l} + \frac{1}{E} \sum_{\mu > 1} v_{\xi_i^\mu k} \sum_{j \neq i} \sum_l u_{\sigma_j l} v_{\xi_j^\mu l} - U(1 - \delta_{k0}), \tag{2.7}$$

where the terms in the RHS stand for signal (ς), noise (ρ) and threshold respectively. Generally speaking, if the field had only the signal part then the state would be stable, but the noise can destabilize it.

As usual in this kind of analysis, we consider the contribution of the noise term in Eq. 2.7 as if it were a normally distributed random variable, i.e. through its average and its standard deviation. In general both quantities scale like p, but in some special cases the average noise is zero and the standard deviation scales only like $\sqrt{p}$, which means that one can store more patterns, as the noise level is reduced. It is clear that the well behaved family of models which we are looking for must fit into this favorable situation. As we said, a necessary but not sufficient condition is the average of the noise to be zero. There are two ways of imposing this into the model. The first way is to make $\lambda_u = -\tilde{a}\kappa_u$, but in this case the standard deviation still scales like p. The second way is to use

$$\lambda_v = -\tilde{a}\kappa_v, \tag{2.8}$$

which makes the standard deviation scale like $\sqrt{p}$. Including this condition, the average signal and the standard deviation of the noise are

$$\varsigma = \frac{N\kappa_v^2}{E} \kappa_u \tilde{a}(1 - \tilde{a}) S(\delta_{\xi_i^1 k} - \tilde{a})(1 - \delta_{k0})$$

$$\rho \;=\; \frac{N\kappa_v^2}{E}\kappa_u\tilde{a}(1-\tilde{a})\sqrt{\alpha a\left\{1-\tilde{a}\left[1-\left(1-\frac{\lambda_u}{\tilde{a}\kappa_u}\right)^2\right]\left[\frac{1-a}{1-\tilde{a}}\right]\right\}}(1-\delta_{k0}), \qquad (2.9)$$

where terms of order $1/N$ have been discarded.

The storage capacity α_c can be estimated as the largest value of α for which $h_i^{\xi_i^1}$ is still likely to be the largest among all $S+1$ local fields. The situation is quite different depending on whether ξ_i^1 is in an active state or not, so one needs to analyze both cases. Note first that $h_i^0 = 0$, so if $\xi_i^1 = 0$ the rest of the local fields must be negative. For this to hold true at least within one standard deviation of the noise distribution we require $\varsigma - U \pm \rho < 0$, or in other words

$$a + \frac{U\,E}{N\kappa_v^2\kappa_u\tilde{a}(1-\tilde{a})} > \sqrt{\alpha a\left\{1-\tilde{a}\left[1-\left(1-\frac{\lambda_u}{\tilde{a}\kappa_u}\right)^2\right]\left[\frac{1-a}{1-\tilde{a}}\right]\right\}}, \qquad (2.10)$$

where we have adopted a positive κ_u.

In the case in which ξ_i^1 is not the zero state two conditions must be fulfilled, namely $h_i^{\xi_i^1} > h_i^0$ and $h_i^{\xi_i^1} > h_i^{k\neq\xi_i^1}$. These conditions can be condensed into

$$S(1-\tilde{a}) - \frac{U\,E}{N\kappa_v^2\kappa_u\tilde{a}(1-\tilde{a})} > \sqrt{\alpha a\left\{1-\tilde{a}\left[1-\left(1-\frac{\lambda_u}{\tilde{a}\kappa_u}\right)^2\right]\left[\frac{1-a}{1-\tilde{a}}\right]\right\}}. \qquad (2.11)$$

The most stringent of these 2 conditions determines α_c. Since in one case U has a negative sign and in the other a positive sign, the optimal point is reached by choosing a suitable threshold $U = \frac{N}{E}\kappa_v^2\kappa_u\tilde{a}(1-\tilde{a})\left[\frac{S}{2}-a\right]$ that makes both conditions equivalent. This choice determines a storage capacity of

$$\alpha_c \simeq \frac{S^2}{4a}\left\{1-\tilde{a}\left[1-\left(1-\frac{\lambda_u}{\tilde{a}\kappa_u}\right)^2\right]\left[\frac{1-a}{1-\tilde{a}}\right]\right\}^{-1}. \qquad (2.12)$$

Note that the expression between curly brackets is equal to or greater than $1-\tilde{a}$. As a consequence, the system remains optimal as long as this expression remains of order 1, which, considering always a to be closer to 0 than to 1, occurs when the expression $\left(1-\frac{\lambda_u}{\tilde{a}\kappa_u}\right)^2$ remains of an order not higher than 1. For this to be true we must impose

$$|\lambda_u| \lesssim \tilde{a}\kappa_u. \qquad (2.13)$$

We thus define the well behaved models as those which fulfil the conditions given by Eq. 2.8 and Eq. 2.13. This simple analysis indicates that the storage capacity of models in the well behaved family scales like S^2/a.

In the following subsections we examine different models proposed in literature, both within and outside the well behaved family.

2.3.1 Symmetric Potts model

The symmetric Potts model was the first S-state neural network to be proposed [Kanter, 1988]. Its units can reach S equivalent states but no zero state. Though simple, a model constructed with these elements is enough to show the S^2 behavior of the storage capacity, as we will see. It is defined by setting

$$\begin{aligned} a &= 1 \\ U &= 0, \end{aligned} \tag{2.14}$$

two conditions related to each other (if there is no zero state, the selectivity mechanism provided by the threshold is not necessary). Moreover $E = S^2 N$, which is just a normalization, and

$$\begin{aligned} \kappa_u &= \kappa_v = S \\ \lambda_u &= \lambda_v = -1. \end{aligned} \tag{2.15}$$

The conditions given by Eq. 2.8 and Eq. 2.13 are fulfilled, and the storage capacity in Eq. 2.12 is approximately

$$\alpha_c \approx \frac{S^2}{4}, \tag{2.16}$$

provided S is large enough. The symmetric Potts model is then a well behaved model of sparseness $a = 1$.

This model is studied analytically with replica tools in [Kanter, 1988], where the author finds an $S(S-1)$ behavior of the storage capacity for low values of S. Unfortunately, the

cited work lacks an analysis for high values of S, which is the interesting limit for modeling multi-modular networks. It is not too difficult, however, to clarify the behavior in this limit.

The replica storage capacity is defined as the highest value of α for which there is a solution to the equation

$$y = \frac{-1 + S\int Dz[\phi(z+y)]^{S-1}}{\sqrt{\frac{\alpha(S-1)}{S}} + \int zDz\{[\phi(z+y)]^{S-1} + (S-1)\phi(z-y)[\phi(z)]^{S-2}\}}, \tag{2.17}$$

where

$$\phi(z) \equiv \frac{1 + \mathrm{erf}(\frac{z}{\sqrt{2}})}{2}. \tag{2.18}$$

Throughout this work we use the gaussian differential $Dz \equiv \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz$, and the integration limits, if not specified, are -∞ and ∞.

We note that in Eq. 2.17 expressions of the form $[\phi(z)]^S$ can be approximated by displaced Heaviside functions for high values of S. Using this we obtain an approximated analytical expression for the storage capacity:

$$\alpha_c = \left[\frac{\phi(\sqrt{\frac{\pi}{2}})}{\sqrt{\frac{\pi}{2}} + \sqrt{2}\mathrm{erf}^{-1}(1 - \frac{\ln(2)}{S})}\right]^2 S^2. \tag{2.19}$$

The factor between brackets in this equation behaves like $\ln(S)^{-\frac{1}{2}}$ for high values of S, which means that the correction for high S to Kanter's low S approximation is a factor of order $\ln(S)^{-1}$.

We show in Fig. 2.1 the results of simulations of a symmetric Potts network ($N = 100$) contrasted with numerical solutions for Eq. 2.17, Kanter's low S approximation and our own high S approximation of Eq. 2.19. Each cross represents the results of a series of simulations with fixed values of S and N and varying p. For each value of p (varying in each case between a low limit of perfect retrieval and a high limit of no retrieval), the state of the network was set initially to exactly coincide with one of the stored patterns (chosen at random) and updated asynchronously until m reached a stable value, which was in all cases either close to 1 (retrieval) or close to 0 (no retrieval). We arbitrarily defined p_{max} in each

case as the value of p in which approximately 70% of the patterns were retrieved. In addition, the error $\triangle p_{max}$ was set as the difference between p_{max} and the value of p corresponding to a performance of 30%, and α_c was simply calculated as p_{max}/N. We did not plot the errors in Fig. 2.1 since they were smaller than the point size, but such error bars can be found in Fig. 2.2, showing simulations of a different network with similar criteria. The analytical predictions fit tightly the results of the simulations and numerical solutions, both for low and high S.

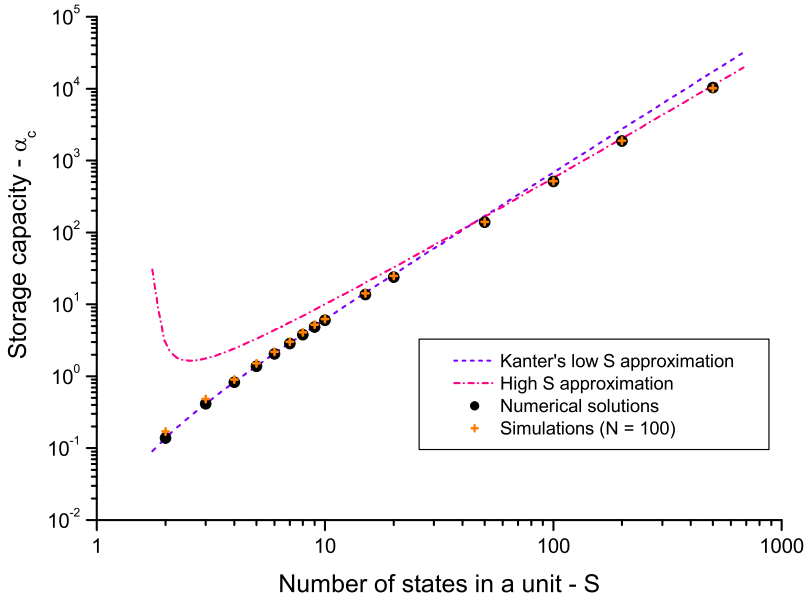

Figure 2.1: Storage capacity of a symmetric Potts network of $N = 100$ units for increasing S. Both axes are logarithmic. Black dots show numerical solutions for Eq. 2.17, which overlap almost perfectly with the simulations (plus signs). For low values of S ($S \lesssim 50$) Kanter's low S approximation fits well, while the high values of S are well fitted by Eq. 2.19.

2.3.2 Biased Potts model

This model is proposed and studied in [Bolle et al., 1993]. The authors extend the symmetric Potts model to an S-state network with arbitrary probability distribution for the states of the units in stored patterns. We adapt their formalism to the case of S equivalent states, a zero state and sparseness a. The parameters are then

$$\begin{aligned} U &= 0 \\ E &= N \\ u_{mn} &= ((S+1)\delta_{mn} - 1) \\ v_{mn} &= (\delta_{mn} - P_n) \end{aligned}, \tag{2.20}$$

where P_k is the probability of a unit in the stored patterns to be in state k. This model does not fit exactly our description because the v operators generate weights J_{ij}^{kl} that are not necessarily zero when k or l are zero. The signal to noise analysis for this situation shows a very poor storage capacity, scaling like a^2. If one adds a non-zero threshold ($U \sim a\,S$ in the optimal case) the storage capacity grows but remains of order 1. These two results show that allowing for non-zero weights to connect zero states is a drawback for the system. The poor performance can, however, be improved by multiplying the v operators by the corresponding $(1 - \delta_{n0})$ factors, and by adding a threshold. In this way, instead of Eq. 2.20 we introduce our definition, Eq. 2.6, for the v operators, with the values for κ's and λ's arising naturally from the model as

$$\begin{aligned} \kappa_u &= S+1 \\ \lambda_u &= -1 \\ \kappa_v &= 1 \\ \lambda_v &= -\tilde{a} \\ U &\sim aS. \end{aligned} \tag{2.21}$$

As in the symmetric Potts model, the condition given by Eq. 2.8 is fulfilled. However, the second condition (Eq. 2.13) can be approximated for high S by

$$a \gtrsim 1/(1 + 1/S) \sim 1, \tag{2.22}$$

which does not stand true for sparse coding. If, instead, $a \ll 1$, the critical value of α in Eq. 2.12 can be approximated as

$$\alpha_c \approx \frac{S^2}{4a}\left\{1 + \frac{1}{a\,S}\right\}^{-1}. \tag{2.23}$$

Hence the storage capacity of the biased Potts model can be preserved close to optimal by imposing an *ad hoc* relation between two parameters that are a priori independent, to assure $1 \ll a\,S$. In this particular situation the model is well behaved. In the opposite limit, when $a\,S \ll 1$, the storage capacity scales like S^3, which is inferior to the S^2/a behavior of the well behaved family.

2.3.3 Sparse Potts model

The simplest version of a well behaved model is perhaps the one introduced as a model for semantic memory [Treves, 2005], with the parameter values

$$\begin{aligned} E &= Na(1-\tilde{a}) \\ \kappa_u &= \kappa_v = 1 \\ \lambda_u &= 0 \\ \lambda_v &= -\tilde{a} \\ U &\sim 1/2. \end{aligned} \tag{2.24}$$

With these parameters, the sparse Potts model is clearly well behaved, and the storage capacity in Eq. 2.12 becomes

$$\alpha_c \simeq \frac{S^2}{4a}. \tag{2.25}$$

2.4 Replica analysis

Having introduced a simple model with optimal storage capacity, we can proceed to analyze the corrections to the signal-to-noise estimation by treating the problem in a more refined way with the classical replica method. The Hamiltonian in Eq. 2.5 can be rewritten for the sparse Potts model as

$$H = -\frac{1}{2}\sum_{i,j\neq i}\sum_{k,l} J_{ij}^{kl}\delta_{\sigma_i k}\delta_{\sigma_j l} + U\sum_i (1-\delta_{\sigma_i 0}), \tag{2.26}$$

with

$$J_{ij}^{kl} = \frac{1}{Na(1-\tilde{a})} \sum_{\mu=1}^{p} (\delta_{\xi_i^\mu k} - \tilde{a})(\delta_{\xi_j^\mu l} - \tilde{a})(1-\delta_{k0})(1-\delta_{l0}) \tag{2.27}$$

constructed using

$$v_{mn} = (\delta_{mn} - \tilde{a})(1-\delta_{n0}). \tag{2.28}$$

We consider the limit $p \to \infty$ and $N \to \infty$ with the ratio $\alpha \equiv \frac{p}{N}$ fixed. Patterns with index ν (μ) are condensed (not condensed). Following the replica analysis [Amit, 1989] the free energy can be calculated as

$$\begin{aligned} f = & \lim_{n\to 0} \frac{a(1-\tilde{a})}{2n} \sum_{\rho=1}^{n} \sum_{\nu} (m_\rho^\nu)^2 + \\ & + \frac{\alpha}{2n\beta} Tr\left(\ln[a(1-\tilde{a})(\mathbb{I} - \beta\tilde{a}\mathbf{q})]\right) + \frac{\alpha\beta\tilde{a}^2}{2n} \sum_{\rho,\lambda=1}^{n} q_{\rho\lambda} r_{\rho\lambda} + \frac{\tilde{a}}{n}(\frac{\alpha}{2} + U\,S) \sum_{\rho=1}^{n} q_{\rho\rho} - \\ & - \frac{1}{n\beta} \left\langle\left\langle \ln Tr_{\sigma_\rho} \exp\left\{ \beta \sum_{\rho=1}^{n} \sum_{\nu} m_\rho^\nu v_{\xi^\nu \sigma_\rho} + \frac{\alpha\beta^2}{2S(1-\tilde{a})} \sum_{\rho,\lambda=1}^{n} r_{\rho\lambda} \sum_k P_k v_{k\sigma_\rho} v_{k\sigma_\lambda} \right\} \right\rangle\right\rangle, \end{aligned} \tag{2.29}$$

where P_k is the probability of a neuron to be in state k in a stored pattern, as defined in Eq. 2.1. The order parameters m stand for the overlaps of the states with different patterns, and $q_{\rho\lambda}$ is analogous to the Edward-Anderson parameter [Edwards and Anderson, 1975], with the following definitions:

$$\begin{aligned} m_\rho^\nu &= \frac{1}{N\,a(1-\tilde{a})} \left\langle\left\langle \sum_{i=1}^{N} \left\langle v_{\xi_i^\nu \sigma_i^\rho} \right\rangle \right\rangle\right\rangle \\ q_{\rho\lambda} &= \frac{1}{N\,\tilde{a}\,a(1-\tilde{a})} \sum_{i=1}^{N} \left\langle\left\langle \sum_k P_k \left\langle v_{k\sigma_i^\rho} v_{k\sigma_i^\lambda} \right\rangle \right\rangle\right\rangle \\ r_{\rho\lambda} &= \frac{S(1-\tilde{a})}{\alpha} \sum_{\mu} \left\langle\left\langle m_\rho^\mu m_\lambda^\mu \right\rangle\right\rangle - \left(\frac{2S\,U}{\alpha} + 1\right) \frac{\delta_{\rho\lambda}}{\beta\tilde{a}}, \end{aligned} \tag{2.30}$$

in such a way that they are all of order 1. Consider, for example, that if $\sigma_i^\rho = \xi_i^\nu$ for all i then $m_\rho^\nu = 1$ on average, while $m_\rho^\nu = 0$ on average if both quantities are independent variables.

We now make two assumptions. First, we consider for simplicity that there is only one condensed pattern, making the index ν superflous. Second, we assume that there is replica

symmetry, and substitute

$$
\begin{aligned}
m_\rho^\nu &= m \\
q_{\rho\lambda} &= \begin{cases} q & if\ \rho \neq \lambda \\ \tilde{q} & if\ \rho = \lambda \end{cases} \\
r_{\rho\lambda} &= \begin{cases} r & if\ \rho \neq \lambda \\ \tilde{r} & if\ \rho = \lambda \end{cases}
\end{aligned}
\tag{2.31}
$$

Taking this into account, we arrive to the final expression for the free energy

$$
\begin{aligned}
f = a(1-\tilde{a})\frac{m^2}{2} + \frac{\alpha}{2\beta}\left[\ln\left(a(1-\tilde{a})\right) + \ln(1-\tilde{a}C) - \frac{\beta q\tilde{a}}{(1-\tilde{a}C)}\right] + \\
+\frac{\beta\alpha\tilde{a}^2}{2}(\tilde{q}\tilde{r} - qr) + \left[\frac{\alpha}{2} + S\,U\right]\tilde{q}\tilde{a} - \frac{1}{\beta}\left\langle\left\langle\int D\mathbf{z}\ln\left(1 + \sum_{\sigma\neq 0}\exp(\beta\mathcal{H}_\sigma^\xi)\right)\right\rangle\right\rangle,
\end{aligned}
\tag{2.32}
$$

where the finite-valued variable C has been introduced

$$
C \equiv \beta(\tilde{q} - q), \tag{2.33}
$$

in such a way that it is of order 1, and

$$
\mathcal{H}_\sigma^\xi \equiv m\, v_{\xi\sigma} - \frac{\alpha a}{S^2}\frac{\beta(r-\tilde{r})}{2}(1-\delta_{\sigma 0}) + \sum_k \sqrt{\frac{\alpha r\, P_k}{S(1-\tilde{a})}} z_k v_{k\sigma}. \tag{2.34}
$$

Both C and $\mathcal{H}_\sigma^\xi$ are variables that are typically found in a replica analysis. The latter, $\mathcal{H}_\sigma^\xi$, can be simply thought of as the mean field with which the network affects state σ in a given neuron if the condensed pattern in the same neuron is in state ξ (note that $\mathcal{H}_0^\xi = 0$). The parameter C, however, has no such an intuitive interpretation in this framework: it measures the difference between $\tilde{q}$, the mean square activity in a given replica, and q, the coactivation between two different realizations. It is interesting to point that this difference goes to 0 when $\beta \to \infty$, the zero temperature limit, so as to keep C of order 1. In the self consistent signal to noise analysis, a method that is based on the knowledge of the replica result and reaches to the same final equations with more intuitive derivations, C is related

to the derivative of the output of an average neuron with respect to variations in its mean field [Shiino and Fukai, 1992, Shiino and Fukai, 1993, Roudi and Treves, 2004].

We now derive the fixed-point equation for m as an example of how the limit $\beta \to \infty$ is taken. The equation for finite β is

$$m = \frac{1}{a(1-\tilde{a})} \left\langle \left\langle \int D\mathbf{z} \sum_{\sigma} v_{\xi\sigma} \left[\frac{1}{1 + \sum_{\rho \neq \sigma} \exp\left\{ \beta(\mathcal{H}_{\rho}^{\xi} - \mathcal{H}_{\sigma}^{\xi}) \right\}} \right] \right\rangle \right\rangle . \tag{2.35}$$

In the limit $\beta \to \infty$ the expression between brackets is 1 if $\mathcal{H}_{\sigma}^{\xi} > \mathcal{H}_{\rho}^{\xi}$ for every $\rho \neq \sigma$ and 0 otherwise. It can be thus expressed as a product of Heaviside functions. The equation for m at zero temperature is then

$$m = \frac{1}{a(1-\tilde{a})} \sum_{\sigma \neq 0} \left\langle \left\langle \int D\mathbf{z}\, v_{\xi\sigma} \prod_{\rho \neq \sigma} \Theta\left[\mathcal{H}_{\sigma}^{\xi} - \mathcal{H}_{\rho}^{\xi} \right] \right\rangle \right\rangle . \tag{2.36}$$

In the same way we derive the rest of the fixed point equations at zero temperature:

$$\begin{aligned} q \underset{\beta\to\infty}{\longrightarrow} \tilde{q} &= \tfrac{1}{a} \textstyle\sum_{\sigma\neq 0} \left\langle \left\langle \int D\mathbf{z} \prod_{\rho\neq\sigma} \Theta\left[\mathcal{H}_{\sigma}^{\xi} - \mathcal{H}_{\rho}^{\xi} \right] \right\rangle \right\rangle \\ C &= \tfrac{1}{\tilde{a}^2\sqrt{\alpha r}} \textstyle\sum_{\sigma\neq 0} \sum_{k} \left\langle \left\langle \int D\mathbf{z} \sqrt{\frac{P_k}{S(1-\tilde{a})}} v_{k\sigma} z_k \prod_{\rho\neq\sigma} \Theta\left[\mathcal{H}_{\sigma}^{\xi} - \mathcal{H}_{\rho}^{\xi} \right] \right\rangle \right\rangle \end{aligned} . \tag{2.37}$$

$$\tilde{r} \underset{\beta\to\infty}{\longrightarrow} r = \tfrac{q}{(1-\tilde{a}C)^2}$$

$$\beta(r - \tilde{r}) = 2U\tfrac{S^2}{a\alpha} - \tfrac{C}{1-\tilde{a}C}$$

The differences between r and $\tilde{r}$, and between q and $\tilde{q}$, are of order $\frac{1}{\beta}$. From the last equation it can be seen that the threshold U has the effect of changing the sign of $(r - \tilde{r})$ and allowing α to scale like $\frac{S^2}{a}$, with the variables C, r and $\tilde{r}$, as we have said, of order 1 with respect to a and S.

2.4.1 Reduced saddle-point equations

It is possible to calculate the averages in Eqs. 2.37 by reducing the problem to the following variables, which represent respectively signal and noise contributions

$$y \equiv m\sqrt{\frac{S^2}{\alpha a}\frac{(1-\tilde{a})}{r}} \equiv m\sqrt{\frac{(1-\tilde{a})}{\tilde{\alpha} r}} \tag{2.38}$$

$$x \equiv \frac{\tilde{\alpha}\beta(r-\tilde{r})}{2}\sqrt{\frac{(1-\tilde{a})}{\tilde{\alpha} r}}, \tag{2.39}$$

where we have introduced the normalized (order 1) storage capacity $\tilde{\alpha} \equiv \alpha a/S^2$, which clarifies that both variables x and y are also of order 1.

At the saddle point, using equations 2.37, we obtain

$$\begin{array}{rcl} y & = & \sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left(\frac{m}{\sqrt{q}+C\sqrt{r}}\right) \\ x & = & \sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left[U-\tilde{\alpha}C\sqrt{\frac{r}{q}}\right]\left[\frac{1}{\sqrt{q}+\tilde{a}C\sqrt{r}}\right] \end{array}, \tag{2.40}$$

which shows that the relevant quantities to describe the system are m, q, and $C\sqrt{r}$. Following this we compute the averages and get from Eq. 2.37 the corresponding equations in terms of y and x

$$\begin{aligned} q &= \frac{(1-a)}{\tilde{a}}\int Dw \int_{y\tilde{a}+x-i\sqrt{\tilde{a}}w}^{\infty} Dz\phi(z)^S + \\ &+ \int Dw \int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}w}^{\infty} Dz\phi(z+y)^S + (S-1)\int Dw \int_{y\tilde{a}+x-i\sqrt{\tilde{a}}w}^{\infty} Dz\phi(z-y)\phi(z)^{S-1} \\ m &= \frac{1}{1-\tilde{a}}\int Dw \int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}w}^{\infty} Dz\phi(z+y)^S - q\frac{\tilde{a}}{1-\tilde{a}} \\ C\sqrt{r} &= \frac{1}{\sqrt{\tilde{\alpha}(1-\tilde{a})}}\left\{\frac{(1-a)}{\tilde{a}}\int Dw \int_{y\tilde{a}+x-i\sqrt{\tilde{a}}w}^{\infty} Dz(z+i\sqrt{\tilde{a}}w)\phi(z)^S + \right. \\ &+ \int Dw \int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}w}^{\infty} Dz(z+i\sqrt{\tilde{a}}w)\phi(z+y)^S + \\ &\left. + (S-1)\int Dw \int_{y\tilde{a}+x-i\sqrt{\tilde{a}}w}^{\infty} Dz(z+i\sqrt{\tilde{a}}w)\phi(z-y)\phi(z)^{S-1}\right\}. \end{aligned} \tag{2.41}$$

Putting together Eqs. 2.40 and 2.41 one can construct the system of two equations that determine the storage capacity. We show an example of their solution in Fig. 2.2 for the parameters $U = 0.5$, $S = 5$ and varying sparseness, contrasting it with simulations of a

network of $N = 5000$ units. This figure shows quite a good agreement between simulations and numerical solutions for a region of the sparseness parameter a, whereas for $a < 0.3$ finite size effects appear, resulting in a lower storage capacity than predicted theoretically.

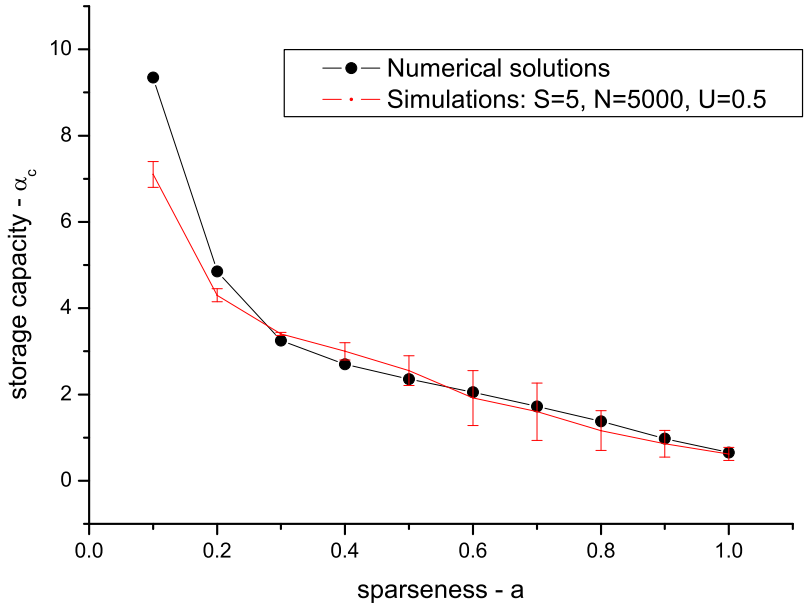

Figure 2.2: Dependence of the storage capacity of a sparse Potts network of $N = 5000$ units on the sparseness a. The black dots show numerical solutions of Eqs. 2.40 and 2.41, while the red line shows the result of simulations. For very sparse simulations (low values of a) finite size effects are observed, which make the storage capacity lower than predicted by the equations.

2.4.2 Limit case

Given that the equations presented in the previous subsection are quite complex, we now analyze the simpler and interesting limit case $\tilde{a} \ll 1$. Though it is not evident from the equations, the normalized storage capacity $\tilde{\alpha}_c$ goes to zero in a logarithmic way as $\tilde{a}$ goes to zero, which means that the storage capacity is not as high as the simple signal to noise analysis of section 3 might suggest. Our analysis of the replica equations for the symmetric Potts model (Eq. 2.19) showing logarithmic corrections is an example of this. We now analyze as another example the sparse Potts model in the case $U = 0.5$.

For the limit of $\tilde{a} \ll 1$ one can approximate Eqs. 2.41 by

$$m \approx \phi(y-x) \tag{2.42a}$$

$$q \approx \frac{(1-a)}{\tilde{a}}\phi(-x) + \phi(y-x) \tag{2.42b}$$

$$C\sqrt{r} \approx \frac{1}{\sqrt{2\pi\tilde{\alpha}}}\left\{\frac{(1-a)}{\tilde{a}}\exp\left(-\frac{x^2}{2}\right) + \exp\left(-\frac{(y-x)^2}{2}\right)\right\}, \tag{2.42c}$$

which is still quite a complex system. We can now make some self consistent assumptions. First we note that, considering x and y as variables that diverge logarithmically as $\tilde{a}$ goes to zero, Eqs. 2.42b and 2.42c indicate that $\sqrt{q} \gg C\sqrt{r}$. Second, for $U = 1/2$ it is possible to consider $x \approx y$, and thus, from Eq. 2.42a, $y \approx 1/\sqrt{2\tilde{\alpha}}$ and $x \approx \varepsilon/\sqrt{2\tilde{\alpha}}$, where ε is a correcting factor for x which is close to 1. With this in mind, and taking into account that $\tilde{\alpha}$ goes to zero with $\tilde{a}$, we can approximate Eq. 2.42b and Eq. 2.42c by keeping only the second term in the first case and only the first term in the second. The equations for y and x can be derived from Eqs. 2.42b and 2.40:

$$y = \sqrt{\frac{\phi(y-x)}{\tilde{\alpha}}} \tag{2.43}$$

$$x = \left[2U - \frac{1-a}{\tilde{a}}\sqrt{\frac{\tilde{\alpha}}{\pi}}\exp(-\frac{x^2}{2})\right]\frac{1}{\sqrt{2\tilde{\alpha}}}. \tag{2.44}$$

Replacing x by $\varepsilon/\sqrt{2\tilde{\alpha}}$ (and ε by 1 where irrelevant) we can approximate $\tilde{\alpha}$ as

$$\tilde{\alpha} = \frac{1}{4\ \ln\left(\frac{1}{(2U-\varepsilon)\tilde{a}}\right)}. \tag{2.45}$$

Next, we posit that $\tilde{a}^{-1}$ is the larger factor in the logarithm, while $(2U-\varepsilon)^{-1}$ gives a correction. A rough approximation for α_c is then

$$\alpha_c = \frac{S^2}{4\,a\ln\left(\frac{1}{\tilde{a}}\right)}, \tag{2.46}$$

which, inserted in 2.44, gives

$$(2U-\varepsilon) = (1-a)\left[4\pi\ln\left(\frac{1}{\tilde{a}}\right)\right]^{-\frac{1}{2}}. \tag{2.47}$$

This expression can be re-inserted into 2.45 in order to get a more refined solution of of Eq. 2.44 than the one given by Eq. 2.46 (which only takes into account the leading factor $\tilde{a}^{-1}$):

$$\alpha = \frac{S^2}{4\, a \ln\left(\frac{2}{\tilde{a}}\sqrt{\ln\left(\frac{1}{\tilde{a}}\right)}\right)}. \tag{2.48}$$

We show in Fig. 2.3 that the approximation given by Eq. 2.48 fits quite well the numerical solution of the sparse Potts model's storage capacity, particularly for very low values of $\tilde{a}$.

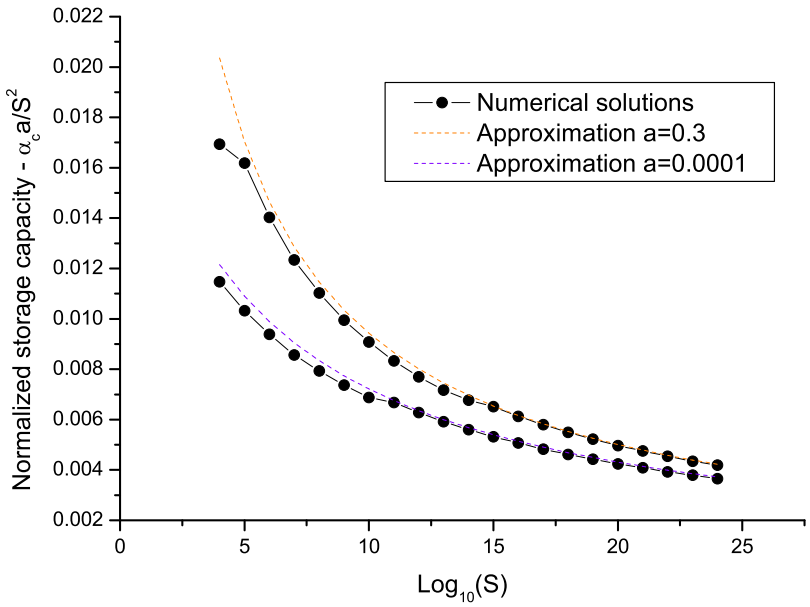

Figure 2.3: Corrections to the $\frac{S^2}{a}$ behavior of the storage capacity of a sparse Potts network for very low values of $\tilde{a}$ in the $U = 0.5$ case. The normalized storage capacity $\alpha_c a/S^2$ is represented, with black dots from numerical solving Eqs. 2.40 and 2.41 for two values of the sparseness: $a = 0.3$ and $a = 0.0001$; with color lines from the corresponding approximation given by Eq. 2.48. Note that to get a good fit we had to go to either very high values of S or very low values of a. Outside of this limit, the solution includes other corrections that we have neglected in the multiple steps that led to Eq. 2.48.

2.5 Diluted networks

In this section we present two modifications to our model which make the network biologically more plausible in terms of connectivity.

First, after considering, to a zeroth order approximation, the long range cortical network as a *fully connected* network, we now wish to describe it, to a better approximation, as a network in which the probability that two units are connected is c_M/N. Traditionally, analytic studies have focused on two soluble cases: the fully connected, which we have studied in the previous sections ($c_M = N$), and the highly diluted ($c_M \lesssim \log(N)$). A recent work has shown, however, that the intermediate case is also analytically treatable and that the storage capacity of an intermediate random network, regardless the symmetry in the weights, stands between the storage capacity of the limit cases [Roudi and Treves, 2004]. Supported by this result, we will focus on the (easier) solution for the highly diluted case, and consider any intermediate situation to be between the two limits.

The second modification reflects the notion that, although the function of long range connections is to transmit information about the state of a local network to another one, this transmission might not be perfectly efficient. We thus introduce an efficacy e, the probability that, in the reduced Potts model, a given state of the pre-synaptic unit is connected with a given state of the post-synaptic one.

Introducing these two modifications, the weights of the sparse Potts model become

$$J_{ij}^{kl} = \frac{C_{ij}^{kl}}{c_M e a(1-\tilde{a})} \sum_{\mu} v_{\xi_i^\mu k} v_{\xi_j^\mu l}, \tag{2.49}$$

where C_{ij}^{kl} is 1 if a connection going from state l in unit j to state k in unit i exists, and 0 otherwise (the average of C_{ij}^{kl} is ec_M/N).

The local field for the unit i and the state k can be analyzed into a signal, a noise and a threshold part, just as in Eq. 2.3,

$$h_i^k = \sum_{jl} J_{ij}^{kl} \delta_{\sigma_j l} - (1-\delta_{k0})U = (1-\delta_{k0}) \left\{ (\delta_{\xi_i^1 k} - \tilde{a}) m_i^k + N_k - U \right\}, \tag{2.50}$$

where

$$m_i^k \equiv \frac{1}{c_M e a (1-\tilde{a})} \sum_j C_{ij}^{k\sigma_j} (\delta_{\xi_j^1 \sigma_j} - \tilde{a})(1-\delta_{\sigma_j 0}). \tag{2.51}$$

Generally, when studying highly diluted networks, the noise term N_k can be treated directly as a normaly distributed random variable, because the states of different neurons are uncor-

related. In this case, N_k can not be considered as a random variable but rather as a weighted sum of normally distributed random variables η_l,

$$N_k \equiv \sum_{l=0}^{S}(\delta_{lk}-\tilde{a})\left\{\sum_{\mu>1}\frac{\delta_{\xi_i^\mu l}}{c_M\, e(1-\tilde{a})a}\sum_j C_{ij}^{k\sigma_j}(\delta_{\xi_j^\mu\sigma_j}-\tilde{a})(1-\delta_{\sigma_j 0})\right\} \equiv \sum_l(\delta_{lk}-\tilde{a})\eta_l. \quad (2.52)$$

The mean of η_l is zero for all l and its standard deviation is

$$\left\langle \eta_l^2 \right\rangle = \frac{\alpha N P_l q_i^k}{S c_M\, e(1-\tilde{a})}, \quad (2.53)$$

with

$$q_i^k \equiv \frac{1}{c_M\, e\, a}\sum_j C_{ij}^{k\sigma_j}(1-\delta_{\sigma_j 0}). \quad (2.54)$$

Note that m_i^k and q_i^k are analogous to m_ρ^ν and $q_{\rho\lambda}$ used in Section 4. If $c_M\, e$ is large enough these quantities tend to be independent of i and k:

$$\begin{aligned} m_i^k \to m &\equiv \frac{1}{Na(1-\tilde{a})}\sum_j(\delta_{\xi_j^1\sigma_j}-\tilde{a})(1-\delta_{\sigma_j 0}) \\ q_i^k \to q &\equiv \frac{1}{N\, a}\sum_j(1-\delta_{\sigma_j 0}). \end{aligned} \quad (2.55)$$

Following the analysis of highly diluted networks in [Derrida et al., 1987], the retrievable stable states of the network are given by the equations

$$\begin{aligned} m &= \frac{1}{a(1-\tilde{a})}\left\langle\left\langle \int D\mathbf{z}\sum_\sigma v_{\xi\sigma}\left[\frac{1}{1+\sum_{\rho\neq\sigma}\exp\left\{\beta(h_\rho^\xi - h_\sigma^\xi)\right\}}\right]\right\rangle\right\rangle \\ q &= \frac{1}{a}\sum_{\sigma\neq 0}\left\langle\left\langle \int D\mathbf{z}\left[\frac{1}{1+\sum_{\rho\neq\sigma}\exp\left\{\beta(h_\rho^\xi - h_\sigma^\xi)\right\}}\right]\right\rangle\right\rangle, \end{aligned} \quad (2.56)$$

where the local field, as in Eq. 2.50 is

$$h_\rho^\xi = m\, v_{\xi\rho} - U(1-\delta_{\rho 0}) + \sum_k\sqrt{\frac{\alpha N P_k q}{S c_M\, e(1-\tilde{a})}}z_k v_{k\rho}. \quad (2.57)$$

These equations are equivalent to those obtained with the replica method (which in the zero temperature limit are Eqs. 2.37 and Eq. 2.34 respectively) if one considers $C = 0$ (and, thus, $r = q$) and an effective value of α given by $\alpha_{eff} = p/(c_M\, e)$.

Comparing this result with that for the fully connected model one notes that, as $\tilde{a} \to 0$, the influence of C in the overall equations becomes negligible (this can be guessed already in Eq.2.40). Therefore if the coding is very sparse, the fully connected and the highly diluted networks become equivalent, and consequently also the intermediate networks. We show this in Fig. 2.4. As the parameter $\tilde{a}$ goes to zero, the storage capacity of the fully connected and the highly diluted limit models converge.

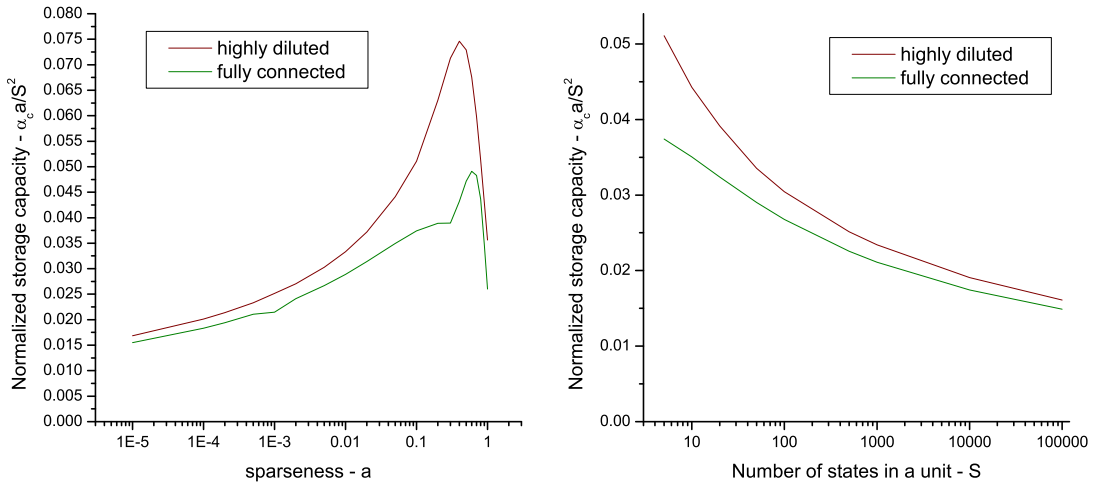

Figure 2.4: A comparison of the storage capacity of a fully connected and of a highly diluted sparse Potts networks. Numerical solutions to the corresponding equations with $U = 0.5$. Left, the dependence of the storage capacity, in the two cases, on the sparseness a, with $S = 5$. Right, the dependence on the number of states per unit S, with $a = 0.1$. In both cases we plot the normalized storage capacity, to focus only on the corrections to the S^2/a behavior. Note that as $\tilde{a} \to 0$ the storage capacity of the two types of network converges to the same result.

2.6 Information capacity

We have shown that the storage capacity of well behaved models scales roughly like S^2/a, while in the two particular examples that we analyzed in full with the replica method, Eqs. 2.19 and 2.46, there is a correction that makes it

$$\alpha_c \propto \frac{S^2}{a \ \ln(\frac{1}{\tilde{a}})} \tag{2.58}$$

for high values of S and low values of a. We now discuss why this is reasonable in the general case from the information storage point of view.

It is widely believed, though not proved, that autoassociative memory networks can store a maximum of information equivalent to a fraction of a bit per synapse. In our model the total number of synaptic variables is given by the different combinations of indexes of the weights J_{ij}^{kl},

$$number\ of\ synaptic\ variables = N\ c_M\ S^2 e. \tag{2.59}$$

On the other hand, the information in a retrieved pattern is N times the contribution of a single unit, which, using the distribution in Eq. 2.1, can be bounded by Shannon's entropy,

$$H = -\sum_{x \in distribution} P(x) \ln\left(P(x)\right) = -\left[(1-a)\ln(1-a) + a\ \ln(\tilde{a})\right]. \tag{2.60}$$

The upper bound on the retrievable information over p patterns is, then,

$$information \leq -p\ N\ \left[(1-a)\ln(1-a) + a\ \ln(\tilde{a})\right]. \tag{2.61}$$

The first term between square brackets in this expression is negligible with respect to the second term provided a is small enough and S is large enough. In this way we can approximate

$$\frac{information}{number\ of\ synaptic\ variables} \leq -\frac{pNa\ln(\tilde{a})}{Nc_M S^2 e} = -\frac{\alpha a \ln(\tilde{a})}{S^2} \leq -\frac{\alpha_c a \ln(\tilde{a})}{S^2} \tag{2.62}$$

.

This result, combined with Eq. 2.58, shows that the storage capacity of our model is consistent with the idea that the information per synaptic variable is at most a fraction of a bit.

2.7 Discussion

The capacity to store information in any device, and in particular the capacity to store concepts in the human brain, is limited. We have shown in a minimal model of semantic

memory, and in progressive steps, how one can expect the storage capacity to behave depending on the parameters of the system: a global parameter - the sparseness a - and a local parameter - the number of local retrieval states S, or, in other words, the storage capacity within a module. The S^2/a behaviour, with its corresponding logarithmic corrections, can be thought of as the combination of two separate results: the a^{-1} behaviour due to sparseness and the S^2 behaviour of the Potts model, which combine in a simple way. We have shown, however, that it is not trivial to define a model that combines these aspects correctly, and that the key is how the state operators are defined. From this study we have deduced the minimum requirements of any model of this kind in order to have a high capacity. Furthermore, through the argument of information capacity we present the well behaved family as representative of general Hebbian models with the same degree of complexity.

The featural representation approach has been so far successful in explaining several phenomena associated to semantic memory, like similarity priming, feature verification, categorization and conceptual combination [McRae, 2005, McRae et al., 1997]. The present work demonstrates that the advantage of the use of features in allowing the representation of a large number of concepts can be realized in a simple associative memory network. More quantitatively, our calculation specifies that in the Potts model the number of concepts that can be stored is neither linear [O'Kane and Treves, 1992] nor an arbitrary power [Schyns and Rodet, 1997] of the number S of values a feature can take, but quadratic in S.

In the case of non-unitary sparseness, one can associate the necessity of introducing a threshold (U) term, whatever its exact form in the local field or the Hamiltonian, with a criterion of selectivity, which is actually observed in the representation of concepts in the brain, as pointed out in the introduction. The threshold behaviour, which is a typical characteristic of neurons, appears to be also necessary at the level of local networks in order to maintain activity low in the less representative modules. The origin of such a threshold has not been discussed in this Chapter. However, a comment on this issue can be made regarding the internal dynamics of local networks. One can show that, as extensively described in the

literature [Amit, 1989], only when the state of a local autoassociative network is driven by external fields sufficiently close to an attractor (inside one of the S basins of attraction) the local system may end up retrieving a pattern on its own, a process that from the global network point of view corresponds to the activation of a unit. The local basin boundary acts in the full system as an effective threshold, roughly equivalent to the simple U term we introduced in the local field of our reduced system. Whether this threshold mechanism is enough, or some addition must be made, can be assessed by studying, in the future, the complete multimodular network without reducing it to Potts units.

Chapter 3

Correlated patterns: latching dynamics

3.1 Recursion and infinity

The unique capacity of humans for language has fascinated scientists for decades. Its emergence, dated by many experts between 10^5 and 4×10^4 years ago, does not seem to arise from the evolution of a distinct and specialized system in the human brain, though many adaptations may have accompanied its progress both inside the brain (for example the general increase in volume, in number of neurons, in connectivity) and outside (for example, the specialization of the human tongue, to facilitate more specific vocalizations). As suggested by a recent experiment [Barsalou, 2005], it is unlikely that the structure of our semantic system differ radically from that of other primates, who have been separated from us by a few million years of evolutionary history. What is it, then, that "suddenly" triggered in humans the capacity for language, and for other related cognitive abilities? This is a matter of great discussion for neurolinguistics and cognitive science, and rather unexplored territory from the point of view of neural computation.

A fruitful approach to this question requires, in our view, abandoning unequivocally the quest for brain devices specialized for language, and directing attention instead to general cortical *mechanisms* that might have evolved in humans, even independently of language, and that may have offered novel facilities to handle information, but within a cortical environment

that has retained a similar structure. A recent review [Hauser et al., 2002] has focused on the identification of the components which are at the same time indispensable for language and uniquely human. The authors reduce this set to a unique element: a computational mechanism for *recursion*, which provides for the generation of an infinite range of expressions, or sequences, of arbitrary length, out of a finite set of elements. A related but more general proposal [Amati and Shallice, 2007], accounts for a variety of cognitive abilities, including language, as enabled by *projectuality*, a uniquely human capacity for producing arbitrarily extended abstract plans that obey certain complex syntactic rules, expressible in terms of a sort of Feynman diagrams.

Thus, it seems that a transition from no recursion to recursion, or from finite to infinite recursion, is a good candidate to be identified as the "smoking gun" that has led to the explosive affirmation of language as a uniquely human faculty. A semantic memory network model has been introduced [Treves, 2005] as an hypothesis about the neural basis of this transition, a model which we have begun to describe quantitatively, from the memory capacity point of view [Kropff and Treves, 2005]. The *latching* dynamics characterizing this network model, which is its essential feature as a model of recursion, can be reduced to a complex and structured set of transitions. Our purpose in this Chapter is to offer a first description of this complexity and to investigate the parameters that control it.

3.2 Semantic memory

As opposed to episodic memory, which retains time-labelled information about personally experienced episodes and events, semantic memory is responsible for retaining facts of common knowledge or general validity, concepts and their relationships, making them available to higher cortical functions such as language. The problem of the organization of such a system has been central to cognitive neuropsychology since its birth. Fundamental studies like [Warrington and Shallice, 1984, Warrington and McCarthy, 1987] have begun to reveal the functional structure of semantic memory through the analysis of patients with differ-

ent brain lesions. Due to methodological reasons related to the paradigm of single-case studies on one side, and to the complexity of functional imaging on the other, there has been always a natural bias toward localization of semantic phenomena, and toward theories with a functionally fragmented view of the operation of the brain. A most radical one among these views is the Domain-Specific Theory [Caramazza and Shelton, 1998], based on the idea that rather than one system, evolution has created in the human brain different systems in charge of representing different concept categories. On the other extreme, recent proposals based on featural representations of concepts [McRae et al., 1997, Greer et al., 2001] tend to describe semantic memory as a single system, where phenomena such as category specific memory impairments arise from the inhomogeneous statistical distribution of features across concepts. This view opens a new perspective for mathematical descriptions and even quantitative predictions of semantic phenomena, as in [Sartori et al., 2005].

Featural representations imply that concepts are represented in the human brain mostly through the combined representation of their associated features. Unlike concepts, thought of as complex structures with an extended cortical representation, features are conceived as more localized, perhaps to a single cortical area (e.g. visual, or somato-sensory) and are *a priori* independent from one another. As proposed in [O'Kane and Treves, 1992], one can model feature retrieval as implemented in the activity of a local cortical network, which by means of its short-range connection system converges to one of its dynamical attractors, i.e. it retrieves one of many alternative activity patterns stored locally. Once the cortex is able to locally store and retrieve features, in different areas, it can associate them through Hebbian plasticity in its long-range synaptic system. Concepts are presented to the brain multi-modally, and thus multi-modal associations are *learned* through an integrated version of the Hebbian principle, reading: 'features that are active together wire together'. The association of features through long-range synapses leads to the formation of global attractor states of activity, which are the stable cortical representations of concepts, and which can then be associatively retrieved. The view that the semantic system operates

through attractor dynamics in global recurrent associative networks accounts for various phenomena described in the last few years such as, for example, the activation of motor areas following the presentation of different non-motor cues associated to an action concept [Pulvermuller, 2001].

3.3 Potts-networks

The Hebbian learning principle appears to inform synaptic plasticity in cortical synapses between pyramidal cells, both on short-range and on long-range connections, making appealing the proposal by [Braitenberg and Schuz, 1991], namely that to a first order approximation the cortex can be considered as a two-level, local and global, autoassociative memory. Furthermore, we have sketched above how featural representations can make use of this two-level architecture in order to articulate representations of multi-modal concepts in terms of the compositional representation of local features. The anatomical and cognitive perspectives can be fused into a reduced "Potts" network model of semantic memory [Treves, 2005].

In this model, local autoassociative-memory networks are not described explicitly, but rather they are assumed to make use of short-range Hebbian synapses to each retrieve one of S different and alternative local features, corresponding to S local attractor states. The activity of the local network i can then be described synthetically by an analog "Potts" unit, i.e. a unit that can be correlated to various degrees with any one of S local attractor states. The state variable of the unit, σ_i, is thus a vector in S dimensions, where each component of the vector measures how well the corresponding feature is being retrieved by the local network. The possibility of no significant retrieval – no convergence and hence no correlation with any local attractor state – can be added through an additional 'zero-state' dimension. Because the local state cannot be fully correlated, simultaneously, with all S features and with the zero state, one can use a simple normalization $\sum_{k=0}^{S} \sigma_i^k = 1$. Having introduced such Potts units as models of local network activity, in the following we will use the terms 'local network' and 'unit' as synonyms.

The global network, which stores the representation of concepts, is comprised of N (Potts) units connected to one another through long range synapses. This network is intended to store p global activity patterns, as global attractor states that represent concepts. When global pattern ξ^μ is being retrieved, the state of the local network i is the local attractor state $\sigma_{\mathbf{i}} \equiv \xi_i^\mu$, retrieving feature ξ_i^μ, a discrete value which ranges form 0 to S (the zero value standing for no contribution of this group of features to concept μ). As shown in [Kropff and Treves, 2005], such a compositional representation of concepts as sparse constellations of features (with a global sparsity parameter a measuring the average fraction of features active in describing a concept) leads to the desired global attractor states when long range connections have associated weights J_{ij}^{kl}

$$J_{ij}^{kl} = \frac{C_{ij}}{c_M a(1-\frac{a}{S})} \sum_{\mu=1}^{p} (\delta_{\xi_i^\mu k} - \frac{a}{S})(\delta_{\xi_j^\mu l} - \frac{a}{S})(1-\delta_{k0})(1-\delta_{l0}), \tag{3.1}$$

which can be interpreted as resulting from Hebbian learning. In this expression each element of the connection matrix C_{ij} is 1 if there is a connection between units i and j, and 0 otherwise (the diagonal of this matrix is filled with zeros), while c_M stands for the average number of connections arriving to a given Potts unit (i.e., local network) i. In this model, the maximum number of patterns, or concepts, which the network can store and retrieve scales roughly like $c_M S^2/a$. We refer to [Kropff and Treves, 2005] for an extensive analysis of the storage capacity of the Potts model.

3.4 Latching

Here we are interested in studying not the storage capacity but rather the dynamics of such a Potts model of a semantic network. Latching dynamics emerges as a consequence of incorporating two additional crucial elements in the Potts model: neuronal adaptation and correlation among attractors. Intuitively, latching may follow from the fact that all neurons active in the successful retrieval of some concept tend to *adapt*, leading to a drop in their activity and a consequent tendency of the corresponding Potts units to drift away from their

local attractor state. At the same time, though, the residual activity of several Potts units can act as a cue for the retrieval of patterns *correlated* to the current global attractor. As usual with autoassociative memory networks, however, the retrieval of a given pattern competes, through an effective inhibition mechanism, with the retrieval of other patterns. One can then imagine a scenario in which two conditions are fulfilled simultaneously: the global activity associated with a decaying pattern is weak enough to release in part the inhibition preventing convergence toward other attractors; but, as an effective cue, it is strong enough to trigger the retrieval of a new, sufficiently correlated pattern. In such a regime of operation, after the first, externally cued retrieval, the network may latch to a new attractor, and when it decays out of it yet to a new one, and so on, experiencing the concatenation in time of successive memory pattern retrievals (See Fig. 3.3). This concatenated spontaneous retrieval is an interesting model for the neural basis of a simple form of infinite recursion, the process postulated to be at the core of cognitive capacities including language.

Several interesting issues arise in trying to describe latching dynamics. The range of parameters enabling latching is one of them, which we will not address here, but rather leave for future communications. Here, we concentrate on a first description of the *complexity* of latching dynamics, and on which parameters control it. As we show, latching transitions are neither deterministic nor random, and they do not depend solely on the correlation between consecutive attractor states. Furthermore, there is strong asymmetry in the transition matrix. These properties can be controlled by a threshold parameter U.

3.5 Adaptation

In retrieval dynamics without adaptation, units are updated with the rule

$$\sigma_i^k = \frac{\exp(\beta(h_i^k + U\delta_{k0}))}{\sum_{l=0}^{S} \exp(\beta(h_i^l + U\delta_{l0}))} \tag{3.2}$$

under the influence of a tensorial local "current" signal which sums the weighted inputs from other units, with a fixed threshold U favouring the zero state

$$h_i^k = \sum_{j=1}^{N} \sum_{l=0}^{S} J_{ij}^{kl} \sigma_j^l. \tag{3.3}$$

To model firing rate adaptation, however, we introduce a modification in the individual Potts unit dynamics. The update rule

$$\sigma_i^k = \frac{\exp(\beta(r_i^k + U\delta_{k0}))}{\sum_{l=0}^{S} \exp(\beta(r_i^l + U\delta_{l0}))} \tag{3.4}$$

is now mediated, for $k \neq 0$, by the vectors r (the "fields" which integrate the h "currents") and θ (the dynamic thresholds specific to each state), which are integrated in time

$$r_i^k(t+1) = r_i^k(t) + b_1[h_i^k(t) - \theta_i^k(t) - r_i^k(t)] \tag{3.5}$$

$$\theta_i^k(t+1) = \theta_i^k(t) + b_2[\sigma_i^k(t) - \theta_i^k(t)]. \tag{3.6}$$

While θ averages σ in a typical time of b_2^{-1} steps, r averages $h - \theta$ in a typical time of b_1^{-1} steps. We also include a non zero local field for the zero state, driven by the integration of the total activity of unit i in all non zero directions, $(1 - \sigma_i^0)$:

$$r_i^0(t+1) = r_i^0(t) + b_3[1 - \sigma_i^0(t) - r_i^0(t)]. \tag{3.7}$$

Together with the fixed threshold U, this local field for the zero state regulates the unit activity in time, preventing local "overheating". A fixed threshold U of order 1 is crucial to ensure a large storage capacity (as shown in [Tsodyks and Feigel'Man, 1988]) and to enable unambiguous memory retrieval.

A final element we include is an effective self-coupling J_{ii}^{kk}, constant for every i and $k \neq 0$, which adds stability to the local network.

For the simulations in this Chapter we have set the parameters $b_1 = 0.1$, $b_2 = 0.005$, $b_3 = 1$ and $J_{ii}^{kk} = 1.8$.

3.6 Correlated distributions

Representations of concepts in the human brain are thought not to be randomly correlated, but rather to present a correlational structure that reflects the shared features between different concepts. In other words, an important part of the correlation between semantic representations may not be arbitrarily generated by the brain, but rather "imported" with the inputs that the semantic system receives from the outside (the correlations in the way we sense the 'real world'). If one assumes that the basic mechanism underlying semantic memory is autoassociative Hebbian learning, it remains unclear how the brain deals with the abrupt decay in storage capacity that these correlations would imply [1]. It is possible that rather than orthogonalizing the correlated input [Srivastava and Edwards, 2000, Srivastava and Edwards, 2004], the strategy of the cortex may be to retain the information about correlations, presumably to make use of it, perhaps as sketched above, to favor latching dynamics.

A standard mathematical procedure to introduce model correlations in a group of p patterns is through a hierarchical construct. Patterns are defined using one or more generations of *parents*, from which they descend, emulating a genetic tree. Since many patterns share the same parents, the generation process introduces correlations among descendant patterns, which are simpler for one-parent families and more complex in the case of multiple parents. We adopt a multi-parent scheme [Treves, 2005]. In addition, our parents are meant to represent semantic category generators, relating directly the correlation between patterns to categorization in a real semantic system, so as to preserve a possibility to link the correlational statistics of our model to observations in the cognitive neuroscience of semantic memory.

[1] We are studying possible solutions to this issue, which will be discussed elsewhere

3.6.1 Quantitative description of correlations

To characterize statistically the resulting set of patterns we introduce the two-pattern correlation distributions

$$\mathcal{C}_0 = \langle \mathcal{C}_0^{\mu\nu} \rangle_{\mu \neq \nu} = \langle \sum_{i=1}^{N} \delta_{\xi_i^\mu \xi_i^\nu} \delta_{\xi_i^\nu 0} \rangle_{\mu \neq \nu} \tag{3.8}$$

and

$$\mathcal{C}_1 = \langle \mathcal{C}_1^{\mu\nu} \rangle_{\mu \neq \nu} = \langle \sum_{i=1}^{N} \delta_{\xi_i^\mu \xi_i^\nu} (1 - \delta_{\xi_i^\nu 0}) \rangle_{\mu \neq \nu}, \tag{3.9}$$

where $\mathcal{C}_0$ takes into account only inactive units and $\mathcal{C}_1$ active units.

To estimate these distributions we now make some assumptions about the process of generation of patterns [Treves, 2005]. A set of Π parents, each active over a random assortment of f Potts units, is generated randomly. Each parent favors a particular direction in Potts space with different (exponentialy decaying) strength. An important quantity for the statistical description is the occupation number of a unit n_i, namely the number of parents active on it. All these parents struggle with varying strength in order to determine the final value of ξ_i^μ, the state of unit i in pattern μ, and this process is repeated for every μ. The occupation number can be thought of as deriving from a series of Bernoulli processes, resulting in a binomial distribution

$$P(n_i = k) = B(k; \Pi, \frac{f}{N}) \equiv \left[\frac{f}{N}\right]^k \left[1 - \frac{f}{N}\right]^{\Pi - k} \begin{pmatrix} \Pi \\ k \end{pmatrix} \tag{3.10}$$

where $B(k; N, p)$ is the binomial distribution, i.e. the probability of winning k times in N trials, with p the probability of winning in one trial. The binomial coefficient is, as usual,

$$\begin{pmatrix} \Pi \\ k \end{pmatrix} \equiv \frac{\Pi!}{k!(\Pi - k)!} \tag{3.11}$$

Next we define the sparsity-by-occupation-number, a_k, as the average activity within the subset of units with a given occupation number k, or, in other words, the fraction of active units divided by the fraction of units (active and inactive) within this subset. The sparsity-by-occupation-number can be modelled by noticing that [Treves, 2005] assumes a process of

filling the occupation levels from the highest to the lowest. The highest occupation levels have $a_k \sim 1$, and the sparsity-by-occupation-number rapidly decreases with lower occupation number. To put this description into a mathematical formulation, we can consider g to be a constant efficiency parameter in the filling of occupation levels. Then, if k_{max} is the highest occupied level, the model reads

$$a_k = a - g \sum_{l=k}^{k_{max}} P(n_i = l) \tag{3.12}$$

until k reaches k_{min}, defined as the value for which $a = g \sum_{l=k}^{k_{max}} P(n_i = l)$. If $k < k_{min}$ or $k > k_{max}$, $a_k = 0$. The constant k_{max} can be directly estimated from Eq. 3.10 as the highest value of k for which the rounded value of $NP(n_i = k) \geq 1$. In Fig. 3.1 we show actual measures and estimates using this model for the sparsity-by-occupation-number, the distribution of the total activity by occupation number and distribution of units by occupation number. The 3 graphs were constructed by fitting the single parameter g, which is the same in all cases, and seems to be stable when varying parameters such as Π or f.

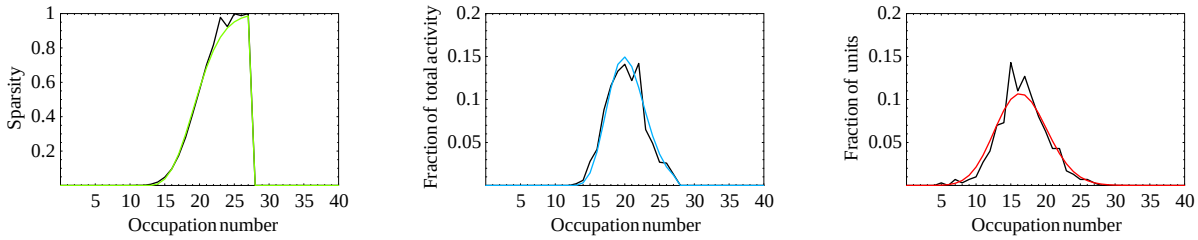

Figure 3.1: From left to right: sparsity-by-occupation-number, fraction of the total activity by occupation number, fraction of units by occupation number. The parameter values are $N = 300$, $p = 200$, $S = 2$, $a = 0.25$, $y = 0.25$, $f = 50$ and $\Pi = 100$. In black: simulations, in color: analytical estimates. g is set to 0.28.

If Π is large, units tend to have large occupation numbers and progressively low levels of occupation become empty. This is analytically convenient, since higher levels of occupation are easier to treat than lower levels, as the competition among many parents can be treated statistically. The distributions $\mathcal{C}_0^{\mu\nu}$ and $\mathcal{C}_1^{\mu\nu}$ can be thought of as generated by the subpopulations of independent occupation levels. Inside of each occupation level patterns can be

considered as randomly correlated, with a_k/S being the probability of a unit to be in a given state. In this way the mean values are

$$\begin{aligned} \mathcal{C}_0 &= \textstyle\sum_k NP(n_i = k)(1 - a_k)^2 \quad = N(1 - 2a) + \sum_k NP(n_i = k)a_k^2 \\ \mathcal{C}_1 &= \quad \textstyle\sum_k NP(n_i = k)a_k^2/S. \end{aligned} \tag{3.13}$$

It is interesting to have a complete picture of the distributions $\mathcal{C}_0^{\mu\nu}$ and $\mathcal{C}_1^{\mu\nu}$. They can be thought of as the combination of individual distributions

$$\begin{aligned} P(\mathcal{C}_0^k = n_k) &= B[n_k; NP(n_i = k), (a_k)^2] \\ P(\mathcal{C}_1^k = n_k) &= B[n_k; NP(n_i = k), (a_k)^2/S] \end{aligned} \tag{3.14}$$

each corresponding to the occupation level k, where in the case of $\mathcal{C}_0$ a base of $N(1 - 2a)$ must be added, corresponding to units that coincide in zero activity regardless of Bernoulli trials, as shown in Eq. 3.13. These contributions can not be considered as Gaussian, since $NP(n_i = k)$ is not necessarily large. Nevertheless, they can still be considered as independent distributions, and the sum of a large number of them can be considered as a distribution with mean given by the sum of the individual means (as shown in Eq. 3.13) and variance given by the sum of the individual variances

$$\begin{aligned} var(\mathcal{C}_0) &= \textstyle\sum_k NP(n_i = k)a_k^2(1 - a_k^2) \\ var(\mathcal{C}_1) &= \textstyle\sum_k NP(n_i = k)a_k^2/S(1 - a_k^2/S). \end{aligned} \tag{3.15}$$

We show actual and estimated distributions in Fig. 3.2. In both cases we used the means and variances given by Eqs. 3.13 and 3.15, but while for $\mathcal{C}_0^{\mu\nu}$ the estimate is a Gaussian distribution, for $\mathcal{C}_1^{\mu\nu}$, which is clearly non-Gaussian given the proximity of the values to zero, we used the corresponding Binomial distribution.

3.7 Transitions

We ran a large set of simulations using the dynamics explained in section 3.5. First of all, we created a set of $p = 50$ patterns using the algorithm described in section 3.6. This set of patterns was used during all the simulations. Each simulation started by giving an initial

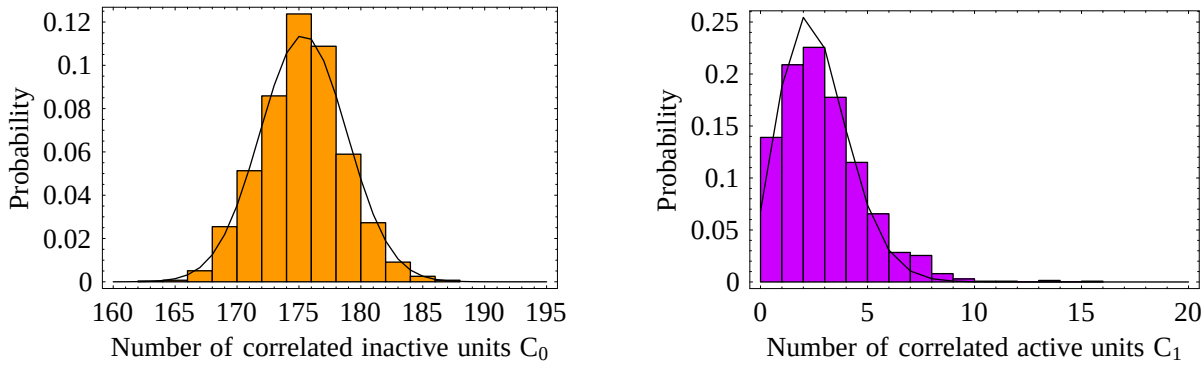

Figure 3.2: Distributions of $\mathcal{C}_0^{\mu\nu}$ (left) and $\mathcal{C}_1^{\mu\nu}$ (right) for$N = 300$, $p = 50$, $S = 10$, $a = 0.25$, $f = 50$ and $\Pi = 100$. In black: analytical estimates using Eqs. 3.10 to 3.15, and $g = 0.47$.

cue to the network (as an additional term in the local field) in order to induce the retrieval of one of the stored patterns. The network was then left free to evolve until, eventually, either the activity decreased to zero or else each unit was updated a maximum of 50000 times – keeping track of latching events. The simulation was run 50 times for each cued pattern, with different random seeds, and all 50 patterns were used as the cued pattern. In this way, we collected a dataset of latching events, with which we constructed the transition probability matrix M. We calculated M for 3 different values of the threshold $U = 0.5$, 0.4 and 0.3. In Fig. 3.3 we show examples of the latching behavior in the 3 cases.

The probability matrix is a square matrix with $p + 1 = 51$ rows and columns, the additional one corresponding to the "null" attractor, with each unit in the zero state. To estimate the transition probability between state μ and ν we counted the times a latching event between these two attractors appeared in the dataset. We added a transition to the "null" state whenever global activity decayed to zero, and assumed a probability of 1 for the transition from the null state to itself. Finally, given that M_{ij} represents the probability of having a latching transition from global attractors i and j, the sum of matrix elements over each row was normalized to 1.

A first interesting result is the distribution of correlations between attractors, parametrized by the numbers of units in the same state, $\mathcal{C}_0^{\mu\nu}$ and $\mathcal{C}_1^{\mu\nu}$. We computed these distributions using a) the whole set of patterns and b) the dataset of latching events. In the first

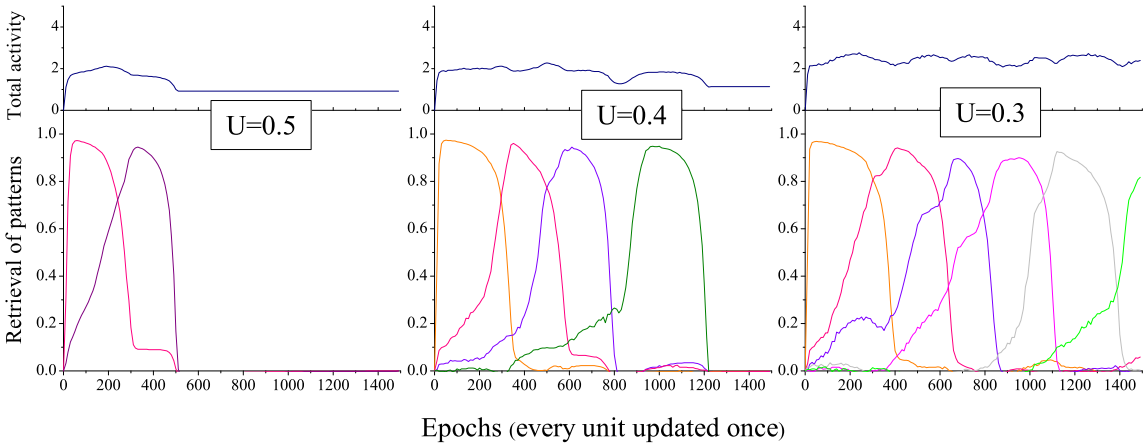

Figure 3.3: Examples of latching dynamics for the 3 values of U: 0.5, 0.4 and 0.3 (from left to right). Top plots: the evolution of the sum of all the activity in the network. Bottom: overlap of the state with the most relevant patterns. Each color corresponds to a different pattern.

case each pair of patterns enters the average once and only once. In the second case, only pairs of attractors visited one after another in a latching event are considered, with a weight proportional to their frequency of occurrence of a transition between them in the dataset. Fig. 3.4 shows the comparison between histograms. Notice that, while $\mathcal{C}_0^{\mu\nu}$ has a similar distribution in both cases, $\mathcal{C}_1^{\mu\nu}$ is shifted toward greater values in the dataset of latching events. This means that latching occurs preferentially between patterns that are correlated over active units. We show this in Fig. 3.4 (right) through the ratio of the probability obtained in b) over the probability obtained in a). The resulting function is clearly increasing with higher correlations.

The next interesting result is that the transition probability matrix is not symmetric, indicating that the correlation between two consecutive attractors, which is itself symmetric by definition, is not the only factor determining latching. To quantify this observation, we introduce a norm for matrices, by adding the absolute value of all of its elements, excluding the rows and columns related to the "null" attractor (which make the matrix asymmetric by construction) $\| M \| \equiv \sum_{\mu\nu} \mid M_{\mu\nu} \mid$. We then calculate $\| M - M^t \|$, which turns out to be

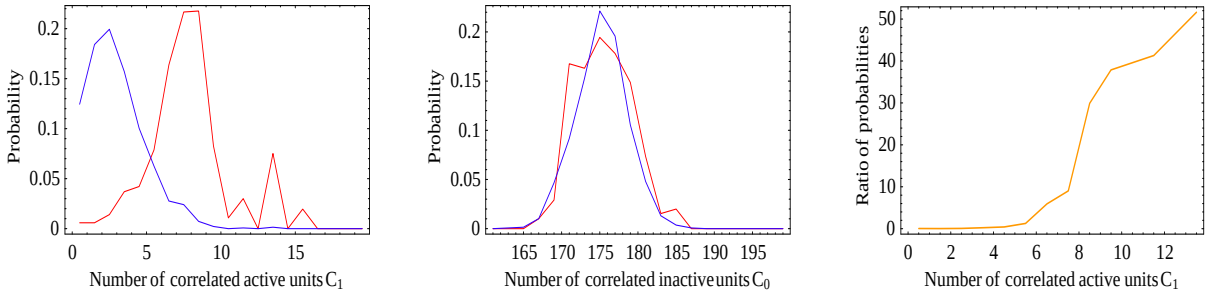

Figure 3.4: Distribution of $\mathcal{C}_1^{\mu\nu}$ (left) and $\mathcal{C}_0^{\mu\nu}$ (center) using the whole set of patterns (blue) and the dataset of latching events (red). Right: the ratio of the two probabilities shown in the left, showing a clear tendency for latching to occur between highly correlated attractors.

U	$\frac{\|M-M^t\|}{\|M\|}$
0.3	0.9
0.4	1.1
0.5	1.1

Table 3.1: Asymmetry of the transition probability matrix (excluding the "null" attractor) measured as the norm of the difference between M and M^t divided by the norm of M. As the threshold U diminishes, the matrix is more symmetric, due to randomness.

of the same order as $\| M \|$. We show this in Tab. 3.1. In addition, we observe that as the threshold U diminishes and randomness grows, the transition probability matrix gets more symmetric.

As M is a transition probability matrix, the eigenvalues of M can be shown to have a modulus lower than or equal to one. Because of the construction of the matrix, the eigenvalue corresponding to the zero pattern, which projects entirely into itself, is $\lambda_0 = 1$. In the general case, when applying the transition matrix n times to an initial pattern η, the result can be decomposed as

$$M^n \hat{\mathbf{x}}_\eta = A D^n A^{-1} \hat{\mathbf{x}}_\eta = A^{-1}_{0\eta} \hat{\mathbf{x}}_0 + \sum_{k=1}^{p} \lambda_k^n A^{-1}_{k\eta} \mathbf{v}_k \tag{3.16}$$

where D is the diagonal matrix of eigenvalues of M, A is the basis change matrix with the eigenvectors of M as columns, λ_k is the k'th eigenvalue of M (with $|\lambda_k| \leq 1$, a property of probability matrixes following the Perron-Frobenius theorem), $\mathbf{v}_k$ the corresponding eigen-

vector and $\hat{\mathbf{x}}_\eta$ is the unitary versor with elements $(\hat{\mathbf{x}}_\eta)_i = \delta_{i\eta}$. From this expression we can conclude that, for large values of n, activity will eventually decay to the "null" attractor, unless some non-null eigenvector of M has an eigenvalue of modulus 1 (at least 1 such eigenvector exists in any probability matrix). Whenever this is not the case, the decay time is given by the second largest eigenvalue of M. More specifically, for any eigenvalue λ_k, the number of transitions for its eigenspace to decay, for example, to 0.1 of its original amplitude is given by

$$n_{dec} = \log_{\lambda_k}(0.1). \tag{3.17}$$

In Tab. 3.2 we show n_{dec} for the second and the third largest eigenvalues, and for our 3 sample values of U. The highest number of transitions in this figure, corresponding to $U = 0.3$, almost corresponds to the length of our simulations (the convergence to an attractor and subsequent drift away from it take, with these parameters, between 300 and 500 updates of each unit, which multiplied by $n_{dec} \sim 50$ is of the same order as the 50000 updates we set as the maximum duration of the simulation). As a consequence, this eigenvalue might actually be underestimated, and in fact closer to 1. The emergence of unitary eigenvalues in the matrix, apart from the one corresponding to the null state, is of great interest, because it would indicate the transition from high-order (but finite) recursion to infinite recursion. More analysis is required to understand this transition, and it will be reported elsewhere. In particular, the threshold U seems to be more effective in controlling the complexity of latching transitions, rather than the order of recursion. The way the latter depends on other parameters, like c_M and S, has been sketched in [Treves, 2005].

One measure of the complexity of transitions is Shannon's information measure, computed over each row of M. We define

$$I_\mu = \frac{1}{\log_2(p+1)} \sum_{\nu=1}^{p+1} M_{\mu\nu} \log_2(\frac{1}{M_{\mu\nu}}). \tag{3.18}$$

Then $I_\mu \sim 0$ both if the attractor μ generates no latching (and thus decays to zero) or if it latches to another fixed attractor, deterministically. On the other hand, if the process of

U	λ_2	λ_3	n_{λ_2}	n_{λ_3}
0.3	0.96	0.57	56.4	4.1
0.4	0.62	0.47	4.8	3.0
0.5	0.4	0.36	2.5	2.3

Table 3.2: Second and third largest eigenvalues of M and the corresponding decay times n_{dec}, as defined in Eq. 3.17, calculated for the 3 values of U.

latching is completely random, $I_\mu = 1$. Fig. 3.5 shows an histogram with the distribution of I_μ for $U = 0.4$, and the mean of this distribution for our 3 values of U.

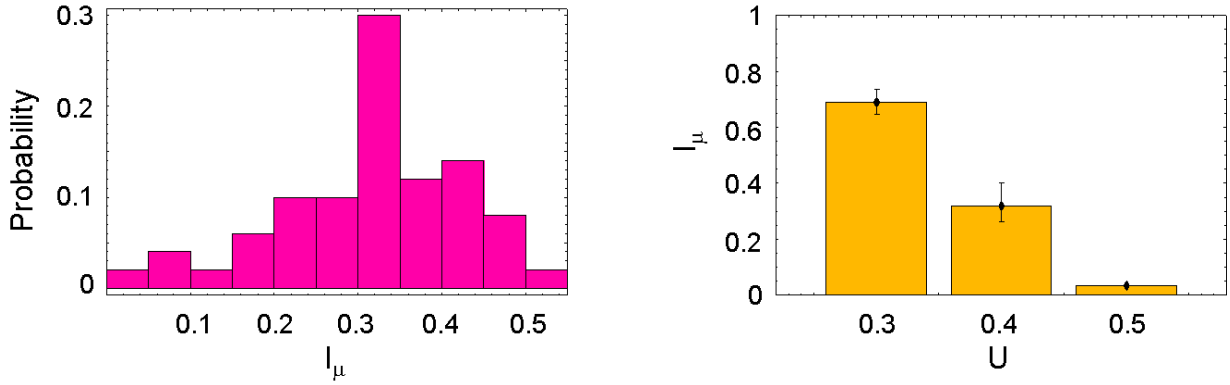

Figure 3.5: Left: distribution of I_μ for $U = 0.4$. Right: mean and quartiles of I_μ (containing the central half of the data) for the 3 sample values of U (right). The values chosen for the threshold span a large range between determinism ($I = 0$) and randomness ($I = 1$).

3.8 Discussion

During the last years, a tendency has been established in cognitive neuroscience toward analyzing semantic phenomena in terms of the distribution of correlations in the featural representation of concepts. This emerging perspective has opened a new domain for the quantitative modelling of higher order processes, that has so far been only partially explored. Here, following up on our previous reports [Treves, 2005, Kropff and Treves, 2005], we have attempted to sketch a mathematical framework to help better understand latch-

ing dynamics in the context of the reduced Potts model. The model itself is based on the idea that associative memory retrieval operates throughout the cortex at two levels [Braitenberg and Schuz, 1991], and as a generic functional mechanism rather than as a separate dedicated system [Fuster, 1999]. In this spirit, we have suggested a rough description of how attractor dynamics in the network model gives rise to a complex and structured set of transitions, that could be regarded as a model of infinite recursion. This complexity, grounded in the correlation between patterns, is controlled mainly by the threshold, that also sets the global activity in the network. An appropriate value of the threshold ensures the transient coexistence of decaying and newly emerging attractors at critical points in the retrieval process, when latching between attractors takes place.

Two additional aspects of latching dynamics, which are only weakly related to the control parameter studied here, still need to be studied in detail: differences between non-recurrent and recurrent networks on the one hand; and the cross-over from finite to infinite recursion on the other. These two issues are of a very dissimilar nature. While the latter, amounting to a percolation phase transition, can be described with the tools presented here, as sketched in section 3.7, the former requires a better comprehension of single retrieval dynamics. Both studies are in progress, and will be the object of future communications.

Though complexity and recursion are both aspects of latching dynamics, as described above, they are independent, as the following example can clarify. When correlations are very strong, and the control parameters set a low level of activity, the dynamics can show a tendency toward determinism, not in the sense of converging to the null attractor, but rather as a sustained cyclic activity involving small groups of patterns. Ideally, one could even find several eigenvalues of the transition probability matrix equal to 1 (associated with infinite recursion), and still a low complexity $< I_\mu >$ in the transitions. This kind of behavior does not seem to be interesting, though, in relation to the phenomena we want to model. The inverse pattern of behavior, corresponding to high complexity of transitions but low recursion, has not yet been observed by us. The interesting regime to study language, we

predict, is that of chaotic dynamics, where the divergence between neighbouring trajectories can be controlled by subtle cognitive factors.

Finally, though the control of complexity was presented here as involving the manipulation of a single parameter (the threshold U), which is actually enough to span the whole space of dynamical network behaviors, this control relies in fact on balancing U with other parameters, the most important of which is the self interaction term of the Potts units, J_{ii}^{kk}. If J_{ii}^{kk} increases, it tends to stabilize the current attractor, adding rigidity to the system. This balance between threshold and self interaction is of major importance in order to consider, in the future, the dynamics of the complete network, without the reduction to Potts units. The self interaction of Potts units is related to the capacity of local networks to maintain specific "delay" activity in the face of external input or, in other words, to the ratio of strengths between long- and short-range synaptic connections, in a full model including single neurons.

Chapter 4

Correlated patterns: consequences of their effective storage

4.1 Introduction

Autoassociative memory networks can store patterns of neural activity by modifying the synaptic weights that interconnect neurons [Hopfield, 1982, Amit, 1989], following the simple rule first stated by Donald O. Hebb: *neurons that fire together wire together* [Hebb, 1949]. Once a pattern of activity is stored, it becomes an attractor of the dynamics of the system. Evidence of attractor behavior has been reported in the rat hippocampus *in vivo* [Wills et al., 2005]. Such memory mechanisms have been proposed to be present throughout the cortex, where hebbian plasticity plays a major role.

The theoretical and computational literature studying variations of the original Hopfield model [Hopfield, 1982] is profuse. Advantages toward optimality or biological plausibility have been demonstrated by varying the learning rule, the neuron model, the architecture or connectivity scheme and the statistics of the input data. The resulting changes in the behavior of the network, however, are often quantitative rather than qualitative. Attractor networks are robust systems that depend only weakly on details. Any optimized attractor network, in fact, appears to be able to retrieve a total amount of information that is never

more than a fraction of a bit per synaptic variable. This limit, consistent with insight obtained with the Gardner approach [Gardner, 1988] but never fully proven, implies that the 'storage capacity' of any associative memory network is constrained by the number of independently modifiable synapses it is endowed with. A suboptimal organization can easily underutilize such capacity, but no clever arrangement can do better than that. Crossing the capacity limit induces a 'phase transition' into total amnesia, destroying the attractor dynamics that would lead to memory states.

Subtler memory deficits than an overall collapse have been reported in the neuropsychological literature, such as category specific effects in the semantic memory system. Patients with partial damage in the cortical networks sustaining semantic memory are found to lose preferentially some concepts rather than others (typically *animals* rather than *tools* or *living* rather than *non-living* things). Initially, research on these effects produced two major antagonistic accounts: the sensory-functional theory [Warrington and Shallice, 1984, Warrington and McCarthy, 1987] and the domain specific theory [Caramazza and Shelton, 1998]. Roughly, they hypothesize that different categories of concepts are localized within partially different (the former) or completely different (the latter) cortical networks. Damage to particular areas would then produce a deficit in the corresponding category of concepts. Attempts to validate some predictions of these theories have not been successful, and an alternative view has emerged in the last few years that, although formulated in various ways, basically hypothesizes that the crucial factor to understand category specific effects is the correlation among items of semantic information, presumed to be stored in one extended and only weakly heterogeneous network [Devlin et al., 2002, Tyler et al., 2000, Sartori and Lombardi, 2004, McRae et al., 1997]. According to this view, random damage to the network would produce selective impairments not because one category is more localized within the damaged area than the other, but rather because differences in the structure of correlations make some categories more vulnerable to damage than others. This explanation has been formulated in a qualitative rather than quantitative formulation. The object of the present study is

to fill this gap with a theory that produces systematic quantitative predictions applicable, in principle, to these and other memory networks storing correlated information. We focus on mathematical models that allow to assess the hypothesis in its 'pure' form, without discussing further other accounts of category specific deficits, found in the literature, which may of course offer complementary elements to an integrated explanation of empirical results.

Most models of attractor networks consider patterns that, for the sake of the analysis, are generated by a simple random process, uncorrelated with each other. Some exceptions appeared during the 80's, when interest grew around the storage of patterns derived from hierarchical trees [Parga and Virasoro, 1986, Gutfreund, 1988]. In particular, Virasoro [Virasoro, 1988] relates the behavior of networks of general architecture to *prosopagnosia*, an impairment in certain patients to identify individual stimuli (e.g., faces) but not to categorize them. Interestingly, his model indicates that prosopagnosia is not prevalent in networks endowed with Hebbian-plasticity. Other developments have described perceptron-like or other local rules to store generally correlated patterns [Gardner et al., 1989, Diederich and Opper, 1987, Srivastava and Edwards, 2004] or patterns with specifically spatial correlation [Monasson, 1992]. More recently, Tsodyks and collaborators [Blumenfeld et al., 2006] have studied a Hopfield memory in which a sequence of morphs between two uncorrelated patterns is stored. In their work, the use of a saliency function favouring unexpected over expected patterns, during learning, can result in the formation of a continuous one-dimensional attractor that spans the space between two original memories. Such fusion of basins of attraction is an interesting phenomenon that we leave for a later extension of this work. In this report, we assume that the elements stored in semantic memory are discrete by construction.

In summary, we aim to show here how a modified version of the standard 'Hebbian' plasticity rule enables an autoassociative network to store and retrieve correlated memories, and how a side effect of the need to use this modified learning rule is the emergence of substantial variability in the resistance of individual memories to damage, which, as we discuss, could explain the prevailing trends of category specific memory impairments observed

in patients.

4.1.1 The model

We consider a network with N neurons and $C < N$ afferent synaptic connections per neuron. The network stores p patterns, and the parameter $\alpha = p/C$ measures its memory load. As for classical analyses [Amit, 1989], we take the 'thermodynamic' limit ($p \to \infty$, $C \to \infty$, $N \to \infty$, α constant, C/N constant) in which the equilibrium properties of the network depend on α rather than separately on N, C and p.

The activity of neuron i is described by the variable σ_i, with $i = 1...N$. Each of the p patterns is a particular state of activation of the network. The activity of neuron i in pattern μ is described by ξ_i^μ, with $\mu = 1...p$. The perfect retrieval of pattern μ is thus characterized by $\sigma_i = \xi_i^\mu$ for all i. For the sake of simplicity, we will assume binary patterns, where $\xi_i^\mu = 0$ if the neuron is silent and $\xi_i^\mu = 1$ if the neuron fires. Consistently, the activity states of neurons will be limited by $0 \leq \sigma_i \leq 1$. Extensions of this work to e.g. threshold-linear units [Treves, 1990] or to Potts units [Kropff and Treves, 2005] are left for further analyses, though, as usual with attractor networks, there is no reason to expect large differences in the qualitative behavior of the system.

We assume that a fraction a of the neurons is activated in each pattern, $a = \Sigma_i \xi_i^\mu / N$ for $\mu = 1 \ldots p$. This *sparseness* parameter is critical in determining the storage capacity of any associative memory network [Treves and Rolls, 1991].

Each neuron receives C synaptic inputs. To describe the architecture of connections we use a random matrix with elements $c_{ij} = 1$ if a synaptic connection between post-synaptic neuron i and pre-synaptic neuron j exists and $c_{ij} = 0$ otherwise, with $c_{ii} = 0$ for all i, a requirement for most attractor network models to function. In addition, synapses have associated weights J_{ij}.

The influence of the network activity on a given neuron i is represented by the field

$$h_i = \sum_{j=1}^{N} c_{ij} J_{ij} \sigma_j \tag{4.1}$$

which enters a sigmoidal activation function when updating the activity of the neuron

$$\sigma_i = \{1 + \exp \beta (U - h_i)\}^{-1} \tag{4.2}$$

where β is an inverse temperature parameter and U is a threshold parameter, which must be kept of order 1 (given the appropriate scaling of the weigths that we will adopt) in order to have a storage capacity close to optimal [Buhmann et al., 1989, Tsodyks and Feigel'Man, 1988]. If $U \ll 1$ all the neurons tend to activate, somewhat similarly to what happens during an epileptic seizure. If, on the other extreme, $U \gg 1$, all neurons tend to be silent. In both extreme situations the effect of U on the network is much stronger than that of the attractors. When U is of order 1, on the contrary, the attractors dominate the dynamics of the network, keeping the total activity of the network near the sparseness a even for transient states, independently of small variations of U.

The learning rule that defines the weights J_{ij} in classical models reflects the Hebbian principle: every pattern in which both neurons i and j are active contributes positively to J_{ij}. In addition, in order to optimize storage, the rule may include some prior information about pattern statistics. In a one-shot learning paradigm, with uncorrelated patterns, the optimal rule uses the sparseness a as a 'learning threshold' [Tsodyks and Feigel'Man, 1988],

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^{p} \left(\xi_i^\mu - a\right) \left(\xi_j^\mu - a\right). \tag{4.3}$$

Note that this 'classical' rule includes implausible positive contributions when both pre- and post-synaptic neurons are silent, and neglects a baseline value for synaptic weights, necessary to keep them positive excitatory weights. Both are simplifications convenient for the mathematical analysis, which have been discussed elsewhere (e.g., in [Treves and Rolls, 1991]) and they will be assumed in the present model as well, though, as we will show, the first and more critical one will not be necessary once we introduce our modified rule.

The above rule has been effectively used to store patterns drawn at random from the distribution with probability

$$P\left(\xi_i^\mu\right) = a\delta\left(\xi_i^\mu - 1\right) + (1-a)\,\delta\left(\xi_i^\mu\right) \tag{4.4}$$

independently for each unit i and pattern μ. In such conditions, the storage capacity of the network is $\alpha_c \propto a^{-1}$. This result assumes the limit of low sparseness, $a \ll 1$, which is the interesting case to model brain function, limit that we will also take in the rest of this work.

Patterns that are correlated, unlike what is implied by the probability distribution in Eq. 4.4, cannot however be stored effectively in a network with weights given by Eq. 4.3. For example, patterns intended to model correlated semantic memory representations have been considered for a long time 'impossible to store' in an attractor network [McRae et al., 1997, Cree et al., 1999, Cree et al., 2006].

4.1.2 Network damage in the model

Semantic impairments can result from damage of very diverse nature, like Herpes Encephalitis, brain abscess, anoxia, stroke, head injury and dementia of Alzheimer type, this last characterized by a progressive and widespread damage. How can we represent damage in our model network in a general way?

The model literature on attractor networks shows that the stability of memories depends on the parameter $\alpha = p/C$ as explained above, where p can be considered in this case as fixed and equal to the number of concepts stored in the semantic memory of a patient. The sparseness a also plays an important role, since the critical value of α, or the storage capacity α_c, varies inversely to a. In addition, we will show in this work that the distribution of popularity a_i across neurons (the fraction of patterns in which each neuron i is active) is a crucial determinant of the storage capacity when memories are correlated. However, it is interesting to notice that both in the modelling literature and in this work, the total number of neurons in the network N is not a determinant factor for the stability of memories, as long as it is large enough to apply statistics.

In our model, random damage to a memory network might affect only C (if the damage is focalized on synapses) or N and C in the same proportion (if the damage is focalized on neurons), while the sparseness a and the distribution of popularity (see below) should, to a first approximation, remain unchanged due to randomness. Since N does not determine the stability of memories, here we simply model network damage as a decrease in the number of connections per neuron, C. Interestingly, forgetting in an intact network could be thought of as the modification of an increasing number of synaptic weights to values that are uncorrelated with the learned ones, and modeled in a similar way. The selective damage of an arbitrary group of synapses or neurons, instead, cannot be modelled simply as a decrease in C, and could lead to different and interesting results that are, however, outside the scope of this work.

4.2 Results

4.2.1 A rule for storing correlated distributions of patterns

We consider a distribution of patterns in which Eq. 4.4 no longer applies, although, to simplify the analysis, we still assume patterns to have a fixed mean activity, as quantified by the sparseness a (the more general case is treated in [Kropff, 2007], resulting in a more complicated analysis but no qualitative changes in the conclusions). We propose a learning rule similar to the one in Eq. 4.3 with the variant that now learning thresholds are specific to each neuron,

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^{p} \left(\xi_i^\mu - a_i^{post}\right) \left(\xi_j^\mu - a_j^{pre}\right). \tag{4.5}$$

Let us use a signal-to-noise analysis to identify appropriate values for such thresholds. The field in Eq. 4.1 can be split into a signal and a noise part by assuming, without loss of generality, that pattern 1 is being retrieved (σ_j similar to ξ_j^1 for all j):

$$h_i = \frac{1}{Ca} \left(\xi_i^1 - a_i^{post}\right) \sum_{j=1}^{N} c_{ij} \left(\xi_j^1 - a_j^{pre}\right) \sigma_j + \frac{1}{Ca} \sum_{\mu=2}^{p} \left(\xi_i^\mu - a_i^{post}\right) \sum_{j=1}^{N} c_{ij} \left(\xi_j^\mu - a_j^{pre}\right) \sigma_j \tag{4.6}$$

where the first term in the RHS is the signal and the second term is the noise. As usual, the signal is a single macroscopic term that drives activity toward the desired attractor state, while a sum of many microscopic contributions comprises the noise. To analyze the latter we assume that ξ_i^μ and ξ_j^μ are statistically independent variables, as long as $i \neq j$ (whereas we *do not* require ξ_i^μ and ξ_i^ν to be independent; on the contrary, the aim is to handle their correlation). If this condition of independence among units, which is central to our analysis, is fulfilled, the noise term can be viewed, to a first approximation, as generated by a gaussian distribution with mean

$$\ll noise \gg = \frac{p-1}{Ca} \sum_{j=1}^{N} c_{ij}\sigma_j \left(\ll \xi_i^\mu \gg_\mu - a_i^{post} \right) \left(\ll \xi_j^\mu \gg_\mu - a_j^{pre} \right). \tag{4.7}$$

If this mean is different from zero, the noise scales up with p, which is the first cause of the performance collapse mentioned above (the optimal one-shot learning rule for uncorrelated patterns has $a_k^{post} = a_k^{pre} = a$ for all k, which results in general in a mean noise different from 0). For $\ll noise \gg$ in Eq. 4.7 to vanish, at least to leading order in p, we must choose either $a_i^{post} = \ll \xi_i^\mu \gg_\mu$ or $a_j^{pre} = \ll \xi_j^\mu \gg_\mu$. We choose the latter

$$a_i^{pre} = a_i \equiv \frac{1}{p} \sum_{\mu=1}^{p} \xi_i^\mu \tag{4.8}$$

where we have introduced $0 \leq a_i \leq 1$, the *popularity* of neuron i, that measures how shared is the activity of this neuron among the patterns in memory. Once this particular choice has been made, one sees from Eq. 4.5 that the contribution of a_i^{post} to the field h_i vanishes, and its exact value is irrelevant. We then choose $a_i^{post} = 0$ for all i.

The next step is to analyze how the variance of the noise distribution scales up with p and C. We have

$$\ll (noise - \ll noise \gg)^2 \gg = \frac{1}{C^2 a^2} \sum_{\mu,\nu=2}^{p} \xi_i^\mu \xi_i^\nu \sum_{j,k=1}^{N} c_{ij} c_{ik} \sigma_j \sigma_k \left(\xi_j^\mu - a_j \right) \left(\xi_k^\nu - a_k \right) \tag{4.9}$$

which can be divided into four contributions that scale differently with p and C, depending

on whether or not j and k on one side and μ and ν on the other are equal:

$$\ll (noise- \ll noise \gg)^2 \gg = \frac{1}{C^2a^2}\sum_{\mu=2}^{p}\xi_i^\mu\sum_{j=1}^{N}c_{ij}\sigma_j^2\left(\xi_j^\mu-a_j\right)^2 + \tag{4.10}$$

$$+ \frac{1}{C^2a^2}\sum_{\mu\neq\nu=2}^{p}\xi_i^\mu\xi_i^\nu\sum_{j=1}^{N}c_{ij}\sigma_j^2\left(\xi_j^\mu-a_j\right)\left(\xi_j^\nu-a_j\right) +$$

$$+ \frac{1}{C^2a^2}\sum_{\mu=2}^{p}\xi_i^\mu\sum_{j\neq k=1}^{N}c_{ij}c_{ik}\sigma_j\sigma_k\left(\xi_j^\mu-a_j\right)\left(\xi_k^\mu-a_k\right) +$$

$$+ \frac{1}{C^2a^2}\sum_{\mu\neq\nu=2}^{p}\xi_i^\mu\xi_i^\nu\sum_{j\neq k=1}^{N}c_{ij}c_{ik}\sigma_j\sigma_k\left(\xi_j^\mu-a_j\right)\left(\xi_k^\nu-a_k\right).$$

The first term in the RHS scales like $(p-1)/C \simeq \alpha$, the second one like $(p-1)(p-2)/C$, the third one like $(p-1)$ and the fourth like $(p-1)(p-2)$. Remembering, however, our definition of popularity in Eq. 4.8, and the statistical independence between neurons, one can see that the leading contributions to the second to fourth term vanish. The remaining dependency of the variance on α is similar to the one found in classical models of autoassociative memory with independent or randomly correlated patterns, indicating that the new rule

$$J_{ij} = \frac{1}{Ca}\sum_{\mu=1}^{p}\xi_i^\mu\left(\xi_j^\mu-a_j\right) \tag{4.11}$$

is a generalization of the Hopfield model appropriate to the storage of correlated patterns.

Figure 4.1 shows simulations of networks of different size and connectivity, employing either the classical or our modified learning rule, to store either uncorrelated or correlated memories, as described in *Methods*. The hierarchical algorithm described in [Kropff and Treves, 2007a] allows us to construct datasets of different p and N values with approximately the same correlation statistics. The four curves result from the combination of the two different learning rules, the standard rule in Eq. 4.3 and the one in Eq. 4.11, with two types of pattern distribution, correlated or not. With the standard, one-shot learning rule, the number of uncorrelated patterns constructed using Eq. 4.4 that can be stored and correctly retrieved, p_{max}, grows linearly with the connectivity C. With non-trivial correlations among patterns, however, the storage capacity collapses: rather than scaling linearly with C, p_{max} even decreases toward 0 for very high values of C. This catastrophe is reversed when the popularity

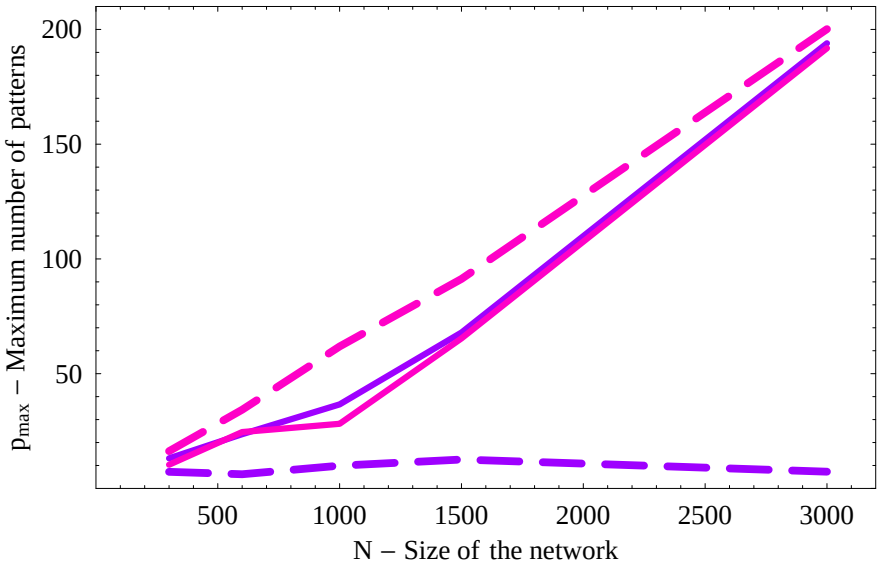

Figure 4.1: The critical value p_{max} measured as the value of p at which 70% of the patterns are retrieved succesfully. We show p_{max} as a function of N using the proportion $C = 0.17N$ for the four combinations of two learning rules and two types of dataset. Violet: one shot 'standard' learning rule of Eq. 4.5. Pink: modified rule of Eq. 4.11. Solid: trivial distribution of randomly correlated patterns obtained from Eq. 4.4. Dashed: non-trivially correlated patterns obtained using a hierarchical algorithm. In three cases the scaling of p_{max} with C is linear, as in the classical result. Only in the case of one-shot learning of correlated patterns there is a storage collapse.

a_i replaces the sparseness a as a learning threshold, bringing p_{max} back to its usual linear dependence on C. The linear dependence of course holds also when the more advanced rule is applied to the original dataset of uncorrelated (i.e., randomly correlated) patterns. Finally, it is important to note that the success in retrieving patterns stored with the rule of Eq. 4.11 does not depend on the algorithm that we used to construct the patterns, but rather shows the generality of the rule, as we do not include in it information about how patterns are constructed. We have tested the modified network with other sets of patterns (such as the random patterns in the same Figure or those described in *Methods*: patterns resulting from setting arbitrary popularity distributions across neurons as shown in Figure 4.3 or patterns taken from the semantic feature norms of McRae and colleagues [Kropff, 2007, McRae, 2005]) always reaching levels of retrieval that are consistent with the predictions of the theory.

Having defined the optimal model for the storage of correlated memories, we analyze in the following sections the storage properties and its consequences through mean field

equations. We note that the average of the popularity across neurons is $\sum_{j=0}^{N} a_j/N = a \ll 1$. In the interesting limit we will consider the popularity a_i generally near 0, and only exceptionally close to 1.

4.2.2 Retrieval with no interference: $\alpha \simeq 0$

If a pattern is being retrieved in a network with very low memory load ($\alpha \simeq 0$), the interference due to the storage of other patterns is negligible. The field in Eq. 4.1 is driven by a single term corresponding to the contribution of the pattern that is being tested for retrieval (which we call pattern 1), or, in other words, the signal term,

$$h_i \simeq \xi_i^1 \left[\frac{1}{Ca} \sum_{j=1}^{N} c_{ij} \left(\xi_j^1 - a_j \right) \sigma_j \right]. \tag{4.12}$$

This can be re-expressed by defining the variables

$$m_i^\mu \equiv \frac{1}{Ca} \sum_{j=1}^{N} c_{ij} \left(\xi_j^\mu - a_j \right) \sigma_j \tag{4.13}$$

and by noticing that, since N and C are large (in the thermodynamic limit both tend to infinity) and c_{ij} is a random connectivity matrix,

$$m_i^1 \to m \equiv \frac{1}{Na} \sum_{j=1}^{N} \left(\xi_j^1 - a_j \right) \sigma_j, \tag{4.14}$$

that is, the average of $(\xi_j^1 - a_j)\sigma_j$ across neurons. The variable m always refers to the pattern that is being tested for retrieval, and it measures its overlap with the state of the network.

Inserting Eq. 4.14 into Eq. 4.12 we obtain

$$h_i \simeq \xi_i^1 m. \tag{4.15}$$

This expression can be inserted into Eq. 4.2 to obtain the updated value of σ_j for all neurons $j = 1 \ldots N$. If the state of the network is stable, σ_j does not change with updating, so it can be reinserted into Eq. 4.14, yielding a single equation that describes the stable attractor states of the system

$$m = \frac{1}{Na} \sum_{j=1}^{N} \left(\xi_j^1 - a_j \right) \left[1 + \exp \beta \left(U - \xi_j^1 m \right) \right]^{-1}. \tag{4.16}$$

Splitting the sum into the aN terms in which $\xi_j^1 = 1$ and the $(1-a)N$ terms in which $\xi_j^1 = 0$, we can rewrite it as

$$m = \left(1 - a^1\right) \left\{ [1 + \exp \beta \left(U - m\right)]^{-1} - [1 + \exp \beta U]^{-1} \right\} \tag{4.17}$$

where the new parameter $0 \leq a^\mu \leq 1$ can be thought of either as the average popularity of the neurons active in pattern μ or as the average overlap between pattern μ and the other patterns:

$$a^\mu \equiv \frac{1}{Na} \sum_{j=1}^{N} \xi_j^\mu a_j = \frac{1}{p} \sum_{\nu=1}^{p} \left[\frac{1}{Na} \sum_{j=1}^{N} \xi_j^\mu \xi_j^\nu \right]. \tag{4.18}$$

Note that for the interesting limit of very sparse activity, in most cases $a^\mu \ll 1$. From the definition of m in Eq. 4.14 it can be noted that $m = 1 - a^1 \simeq 1$ for perfect retrieval (i.e., $\{\sigma_j\} \equiv \{\xi_j^1\}$) and $m = a - a^\sigma \simeq 0$ if the activity σ of the network has sparseness a but is unrelated to ξ^1, i.e., retrieval fails.

Eq. 4.17 always admits the solution $m = 0$, and it may have another stable solution depending on two combinations of parameters: βU and $\beta(1 - a^1)$. Whenever this non-zero solution exists, retrieval is possible. In Figure 4.2 we show, as a function of the two parameters, the highest value of m that solves Eq. 4.17. A first order phase transition is observed: given a fixed value of βU there is a critical value of $\beta(1-a^1)$ below which the only solution to Eq. 4.17 is $m = 0$, i.e., no retrieval. In the 'zero-temperature' ($\beta \to \infty$) limit, the condition for the existence of a non-zero solution in Eq. 4.17 reduces to $m = (1 - a^1) \geq U$, showing that at the critical point $a_c^1 = 1 - U$. Clearly, the choice $U = 0$ would permit the retrieval of patterns with arbitrary values of a^1 (which is, by definition, not larger than 1), but as shown in [Buhmann et al., 1989, Tsodyks and Feigel'Man, 1988] and in the following sections, a threshold value of order 1 is necessary to obtain an extensive storage capacity, close to optimal, when interference due to the storage of other patterns is not negligible.

An intuitive explanation of Figure 4.2 would be the following. The learning rule in Eq. 4.11 implies that the network is less *confident* of any neuron j with high popularity, since its positive contributions to outgoing weights are proportional to $1 - a_j$. This implies that the more popular is, on average, the ensemble of neurons underlying a given memory (as

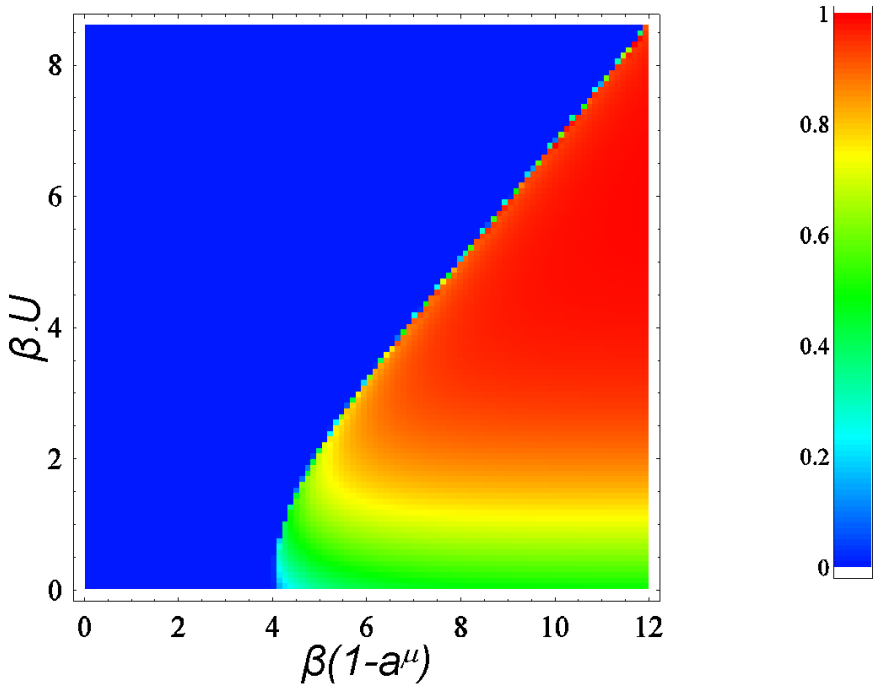

Figure 4.2: Numerical solutions of Eq. 4.17 varying the two relevant parameters: $\beta(1-a^1)$ on the x axis and βU on the y axis. A first order phase transition is observed in the value of m that solves Eq. 4.17. In the limit $\beta \to \infty$ the transition occurs along the identity line $1 - a^1 = U$.

expressed by its a^1 value), the less able it is to sustain, through neural activity, the corresponding attractor state. When the average activating signal is smaller than the threshold U, retrieval is no longer possible.

4.2.3 Retrieval with interference: diluted networks

To treat the case of extensive storage, p scaling up with C, we consider the so called *highly diluted* approximation, which is valid when either $C \ll N$ ('diluted', i.e. sparse connectivity proper, [Derrida et al., 1987]) or $a \ll 1$ (very sparse activity, [Treves and Rolls, 1991]). There are two independent motivations to study such a limit: on one side it approximates real cortical networks, with their sparse connectivity and sparse firing, on the other, calculations are much simpler than for fully connected networks, enabling deeper analysis and wider generalization. In addition, one obtains in this limit differential equations for the dynamical evolution of all relevant variables, valid also outside of equilibrium [Derrida et al., 1987].

Such an approach is outside the scope of this work, and it is left for future studies. It is worth mentioning that some experimental work on semantic memory [Sartori and Lombardi, 2004, Sartori et al., 2005] is based on a dynamical view of the networks involved in semantic processing, as it focuses on the type of input cues that can lead to successful retrieval.

The highly diluted approximation takes into account in the field h_i a signal term and a gaussian noise, while neglecting the effect of a second source of noise due to the propagation of neural activity around closed loops of synaptic connections. These effects scale in general like $\alpha a C/N$ [Roudi and Treves, 2004, Kropff, 2007], and are therefore negligible as $C/N \to 0$, $a \ll 1$ or, as in the previous section, $\alpha \simeq 0$.

In Eq. 4.11 we had already obtained an expression of the variance of the noise part of the field h_i when considering it to be purely gaussian. After computing the average over μ in the surviving first term, we obtain

$$\ll (noise- \ll noise \gg)^2 \gg = \alpha \; a_i \left[\frac{1}{Ca^2} \sum_{j=1}^{N} c_{ij} a_j \left(1 - a_j\right) \sigma_j^2 \right]. \tag{4.19}$$

The expression between square brackets depends on i only through the connectivity matrix c_{ij}. As in Eq. 4.14, we can take advantage of the fact that c_{ij} is random and C large, and replace the sum with an average over all neurons. We can conclude that $\ll (noise- \ll noise \gg)^2 \gg = \alpha a_i q$, where we define

$$q \equiv \frac{1}{Na^2} \sum_{j=1}^{N} a_j \left(1 - a_j\right) \sigma_j^2. \tag{4.20}$$

The local field then becomes

$$h_i = \xi_i^1 m + \sqrt{\alpha a_i q} z_i \tag{4.21}$$

where z_i may be assumed to be drawn from a normal distribution with mean 0 and variance 1, statistically independent with all other variables [1]. To describe attractors of the system,

[1] In the simplest signal-to-noise approach [Kropff and Treves, 2005] two 'worst-case' conditions must be met in order to have stable attractors: $h_i = m - \sqrt{variance} > U$ for values of i in which $\xi_i^1 = 1$ and $h_i = \sqrt{variance} < U$ for $\xi_i^1 = 0$. This shows that the optimal value of U is $m/2 \simeq (1 - a^\mu)/2$, which depends

as previously, we insert the field into Eq. 4.2 to obtain the stable value of σ_j, which can be re-inserted into the definition of m in Eq. 4.14,

$$m = \frac{1}{Na} \sum_{j=1}^{N} \left(\xi_j^1 - a_j\right) \left[1 + \exp \beta \left(U - \xi_j^1 m - \sqrt{\alpha a_j q} z_j\right)\right]^{-1}. \tag{4.22}$$

Making use of the independence of z_j with respect to a_j and ξ_j^1, we can take its average. The highly diluted version of Eq. 4.16 is then

$$m = \frac{1}{Na} \sum_{j=1}^{N} \left(\xi_j^1 - a_j\right) \int_{-\infty}^{\infty} Dz \left[1 + \exp \beta \left(U - \xi_j^1 m - \sqrt{\alpha a_j q} z\right)\right]^{-1} \tag{4.23}$$

where the gaussian differential is

$$Dz \equiv dz \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{4.24}$$

expressing the distribution of z_j.

In the following, for simplicity, we will take the limit of zero temperature, $\beta \to \infty$. The equation for m becomes

$$m = \frac{1}{Na} \sum_{j=1}^{N} \left(\xi_j^1 - a_j\right) \phi\left(\frac{\xi_j^1 m - U}{\sqrt{\alpha a_j q}}\right) \tag{4.25}$$

where

$$\phi\left(y\right) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{y}{\sqrt{2}}\right)\right) \tag{4.26}$$

is a sigmoidal function increasing monotonically from 0 to 1, with $\phi(0) = 1/2$. Since in Eq. 4.25 the terms are not linear in a_j, it is not straightforward to obtain the new version of Eq. 4.17. To do so we must first introduce the distribution of popularity across neurons, given by the probability

$$F\left(x\right) \equiv P\left(a_j = x\right), \tag{4.27}$$

and the distribution of popularity across neurons that are active in the pattern we are testing for retrieval,

$$f\left(x\right) \equiv P\left(a_j = x | \xi_j^1 = 1\right). \tag{4.28}$$

on global rather than local information. Interesting corrections in which the optimal value of U depends on a_i and is thus different for each neuron might come out of considering the non-diluted case, including an additional term in the local field h_i as mentioned above.

The purpose of introducing these distributions is to convert a discrete set of popularities $\{a_j\}$ into a continuous distribution, where the popularity is represented by the variable x. Since N is large, we can transform the sum in Eq. 4.25 into an integral over these distributions. As a result we obtain the equation

$$\begin{aligned} m &= \int_0^1 dx f(x) \left\{ (1-x)\phi\left(\frac{m-U}{\sqrt{\alpha x q}}\right) + x\phi\left(\frac{-U}{\sqrt{\alpha x q}}\right)\right\} - \\ &\quad -\frac{1}{a}\int_0^1 dx F(x) x \phi\left(\frac{-U}{\sqrt{\alpha x q}}\right), \end{aligned} \tag{4.29}$$

which extends Eq. 4.17 to the case of non negligible interference.

Since this equation depends not only on m but also on q, we need a second equation to close the system and univocally describe the stable states of the network. From the definition of q in Eq. 4.20 we can repeat the steps 4.22 to 4.25 and obtain, for stable states and in the limit of zero temperature,

$$q = \frac{1}{Na^2}\sum_{j=1}^{N} a_j\,(1-a_j)\left[\phi\left(\frac{\xi_j^1 m - U}{\sqrt{\alpha a_j q}}\right)\right]^2. \tag{4.30}$$

Introducing again the distributions of popularity – steps 4.25 to 4.29 – we can simplify this expression into

$$\begin{aligned} q &= \frac{1}{a}\int_0^1 dx f(x) x(1-x)\left\{\phi\left(\frac{m-U}{\sqrt{\alpha x q}}\right) - \phi\left(\frac{-U}{\sqrt{\alpha x q}}\right)\right\} + \\ &\quad +\frac{1}{a^2}\int_0^1 dx F(x) x(1-x)\phi\left(\frac{-U}{\sqrt{\alpha x q}}\right). \end{aligned} \tag{4.31}$$

Eqs. 4.29 and 4.31 describe the stable states of the network in this 'diluted' approximation. As in the noiseless case, a phase transition separates regions of parameter space where a solution with $m \sim 1 - a^1$ exists from regions where the only solution is $m = q = 0$. The latter can now be reached by increasing $\alpha = p/C$, i.e. the memory load. In other words, the phase transition to no retrieval determines the storage capacity of the system. If $f(x) = F(x) = \delta(x-a)$, which is the case for uncorrelated patterns, the classical equations for highly diluted binary networks [Buhmann et al., 1989, Tsodyks and Feigel'Man, 1988] are re-obtained, and the critical value of the memory load scales like

$$\alpha_c \propto \frac{1}{a\ln(1/a)} \tag{4.32}$$

for the relevant sparse limit $a \ll 1$.

How does this classical result generalize to the case of correlated representations?

4.2.4 The storage capacity

Already at first glance, the system of Eqs. 4.29 and 4.31, which determine the storage capacity of a network with correlated patterns, reveals a new property of associative memories. In both equations, the second term in the RHS depends on $F(x)$ and is thus common to the retrieval of any pattern. However, the RHS of both equations depends also on $f(x)$, the distribution of popularity among neurons active in the pattern that is being retrieved. In the general case, this distribution is different for every pattern, so that *the stability properties of the associated attractors will differ from pattern to pattern.*

To understand this idea it is convenient to think about the storage capacity as p/C_{min} (the minimum connectivity necessary to sustain retrieval) rather than as p_{max}/C (the maximum number of patterns that can be stored). In this view, each of p memory states stored in a network has an associated value of C_{min} that depends on its own statistical properties and on the statistical properties of the whole dataset. Any particular pattern can be retrieved only if the actual connectivity level C is higher than the value of C_{min} associated to it.

This view is of particular interest to analyze category specific deficits in semantic memory. We can think of p as being relatively fixed, corresponding, in the model, roughly to all the concepts acquired by a healthy subject during an entire life. A mild and non-selective damage of the network might decrease the parameter C, which would selectively affect the memories with a high value of C_{min}, while sparing the others.

An entropy characterization of the noise

To analyze Eqs. 4.29 and 4.31 we first consider that α and U are small enough to ensure that the retrieval is possible and that $\phi\left(\frac{m-U}{\sqrt{\alpha x q}}\right) \sim 1$ and $\phi\left(\frac{-U}{\sqrt{\alpha x q}}\right) \sim 0$. Following this, any pattern that we choose to test for retrieval has $m \simeq 1 - a^1$, as we had found for $\alpha \simeq 0$ and a

value of the noise variable q that is proportional to the average of $a_j(1-a_j)$ over the neurons that are active in the pattern (as can be seen from Eqs. 4.30 or 4.31), or in other words,

$$S_f \equiv \int_0^1 x(1-x)f(x). \tag{4.33}$$

Similarly to Shannon's entropy, S_f, and in consequence the noise variable q, approaches 0 if neurons in the distribution are all either very popular or unpopular in their firing, while it is maximum ($S_f = 1/4$) when $f(x) = \delta(x - 1/2)$, i.e. all neurons have popularity $a_i = 1/2$ [2]. Thus, a pattern will be better retrieved if a) it includes as unpopular neurons as possible (as shown previously, to ensure $m = 1 - a^1 > U$) and b) its neurons have a low 'entropy' value S_f, in order to minimize the noise $q \simeq S_f/a$.

An intuitive explanation of this comes from the analysis of the influence of neuron j as noise in the field h_i, proportional to $\sum_{\mu\neq 1} \xi_i^\mu(\xi_j^\mu - a_j)$ as shown in Eq. 4.6. If the popularity of neuron j is very low, terms of this noise where $\xi_j^\mu = 1$ are large contributions (proportional to $1 - a_j$), but very infrequent, while terms in which $\xi_j^\mu = 0$ are very frequent but only proportional to $a_j \ll 1$. The exact opposite pattern emerges if neuron j is very popular. As a result of this, in both cases the noise is very low. In the extreme of $a_j = 0$ or $a_j = 1$ the noise is exactly zero, since contributions of order 1 occur with probability 0 and inversely. In such a case the dynamics of the network is guided purely by the signal terms, that take h_i toward the correct value for retrieval. The case in which the noise is maximal is when the probability of neuron j to be active is $a_j = 1/2$ and each term of the contribution of neuron j to the noise in the field h_i is proportional to $1 - a_j = 1/2$ or $a_j = 1/2$. Finally, since the noise is also proportional to σ_j and pattern 1 is being retrieved, this effect is important only for the neurons j that are active in this pattern, explaining fully Eq. 4.33.

[2] Technically, this function applied to a single unit is Tsallis' entropy with parameter $q = 2$. Note, however, that Tsallis' entropy is not additive for independent events, while our S_f is clearly a normalized extensive quantity.

Popularity distributions $F(x)$ with low variance

As α increases, the assumption $\phi\left[(m-U)/\sqrt{\alpha x q}\right] \sim 1$ becomes eventually incorrect and for some critical value α_c a retrieval solution with $m \sim 1 - a$[1] no longer exists. A generally fair approximation when studying storage capacity is to assume that α_c scales inversely to the factor that accompanies α in the argument of ϕ, which in this case is xq. However, since x is a variable that spans the whole range from 0 to 1, the approximation is not useful in itself. In more general terms, α_c should scale inversely to $x_f q$, with $0 < x_f < 1$ some intermediate value with a strong dependence on $f(x)$. In this section we consider the case in which the variance of $F(x)$ is small enough to allow the approximation of x by its average a in the argument of ϕ, while in *Methods* we analyze some more general examples.

Our first order approximation, assuming α inverse to aq and $q \simeq S_f/a$, leads to

$$\alpha_c \propto \frac{1}{S_f}. \tag{4.34}$$

In line with what we had explained intuitively, the storage capacity, or C_{min}/p, is inverse to the entropy S_f of the pattern. In the classical case of randomly correlated patterns $S_f = a(1-a) \sim a$ (again, assuming cortical activity to be sparse, the interesting approximation is always $a \ll 1$), which leads to the Tsodyks and Feigel'man result in Eq. 4.32, without the logarithmic correction.

This correction appears only when $\phi\left(-U/\sqrt{\alpha a q}\right)$ starts to be significantly different from 0. The largest contribution is the one given by the second term in the RHS of Eq. 4.31, since it is not negligible when $\phi\left(-U/\sqrt{\alpha a q}\right)$ is of order a (considering $a \ll 1$), while the other neglected terms are only relevant when $\phi\left(-U/\sqrt{\alpha a q}\right)$ is of order 1. Again, we use the approximation of low variance, so the term we are interested in becomes

$$\mathcal{T}_2 = \frac{1}{a^2}\phi\left(\frac{-U}{\sqrt{\alpha a q}}\right)\int_0^1 dx F(x)x(1-x) \equiv \frac{1}{a^2}\phi\left(\frac{-U}{\sqrt{\alpha a q}}\right)S_F, \tag{4.35}$$

where, similarly to S_f, we define S_F as the entropy of the distribution $F(x)$. This term is near 0 for very small values of α, where q is dominated by the first term of Eq. 4.31,

which can still be considered as S_f/a, and it becomes significant only when both terms are of comparable magnitude. If this happens at values of α that are smaller than the one indicated by Eq. 4.34, the correction introduced by this term is relevant. To estimate this correction we impose the first and second terms of Eq. 4.31 to be about equal ($\mathcal{T}_2 \simeq S_f/a$) and consider $a \ll 1$, which leads to

$$\phi\left(-\frac{U}{\sqrt{\alpha_c S_f}}\right) \simeq \frac{aS_f}{S_F}. \tag{4.36}$$

Inverting the function ϕ we obtain α_c as

$$\alpha_c \simeq \frac{1}{2S_f}\left[\frac{U}{\mathrm{erf}^{-1}\left(1-\frac{2aS_f}{S_F}\right)}\right]^2. \tag{4.37}$$

The inverse error function can be approximated as

$$\mathrm{erf}^{-1}(1-y) \sim \sqrt{\ln\left(\sqrt{\frac{2}{\pi}}\frac{1}{y}\right)} \tag{4.38}$$

for small values of y. Since $F(x)$ has low variance, $S_f, S_F \sim a \ll 1$ and aS_f/S_F can be taken to be a small quantity. We then approximate

$$\alpha_c \simeq \frac{1}{2S_f}\left[\frac{U^2}{\ln\left(\frac{S_F}{\sqrt{2\pi}aS_f}\right)}\right] \propto \frac{1}{S_f \ln\left(\frac{S_F}{aS_f}\right)}. \tag{4.39}$$

If this scaling of α_c is lower than indicated by Eq. 4.34 (or, in other words, if $\ln(S_F/(aS_f)) > 1$) this correction is relevant. Finally, in the case of trivial correlations $f(x) = F(x) = \delta(x-a)$ and consequently $S_f = S_F \simeq a$. The full classical result of Eq. 4.32 is then reproduced by Eq. 4.39, indicating that the latter is a generalization of the former.

In *Methods* we find expressions similar to 4.39 for wider distributions of $F(x)$. As we show, the slower the decay of the tail of a smooth distribution $F(x)$ with increasing x, the poorer is performance in terms of storage capacity. If the decay of $F(x)$ is exponential or faster, the $1/S_f$ scaling of Eq. 4.39 holds with at most a larger logarithmic correction. If the decay is a power-law, instead, the scaling is much poorer: $\alpha_c \propto a/S_f$, with, as usual, $a \ll 1$.

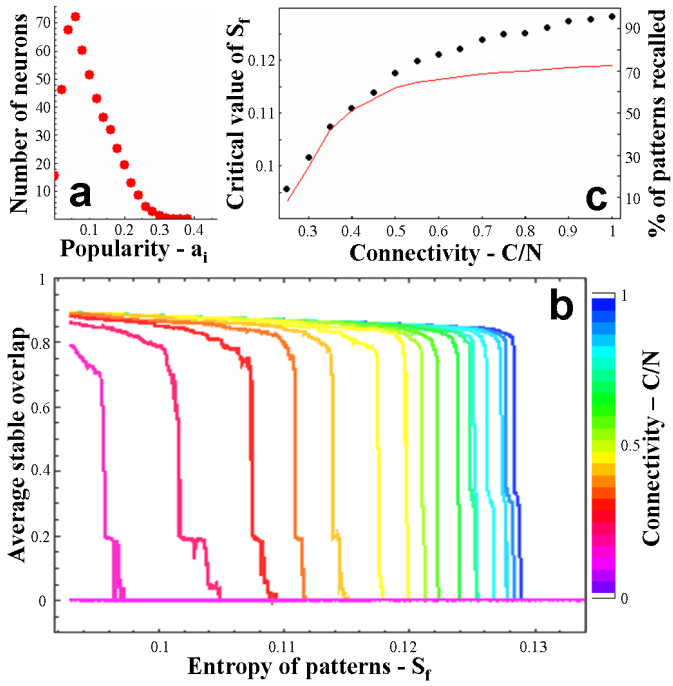

Figure 4.3: Simulations of the storage capacity of a network storing patterns with an arbitrary correlation distribution F(x). The parameters are $N = 500$, $p = 50$, $a = 0.1$, $U = 0.35$ and variable C. For all values of C each pattern is tested 10 times for stability, with different connectivity matrices c_{ij}. **a** Popularity distribution across the whole network, $F(x)$. Note that neurons with $a_i = 0$ do not really participate in network dynamics, making the effective values of C and N slightly lower. **b** Stable value of m for each pattern vs. its S_f value. The data has been smoothed by taking the median over a moving window. From blue toward violet: connectivity C/N starting with 1 and decreasing in steps of 0.05. For each color, the graph shows that some patterns are retrieved while others are not, corresponding to low and high values of S_f. The critical value of S_f at which the transition occurs moves to the left as the connectivity is reduced, which, as explained in the Introduction, is the strongest effect of random network damage. **c** Storage capacity computed from the step-like transitions in **b**. Black dots, left axis: critical value of S_f vs. connectivity, showing the maximum retrievable S_f supported by the C connections of the network. Red line, right axis: percent of patterns with a value of S_f lower than the critical one.

Informative memories are less robust

In Figure 4.3 we show results of simulations using a distribution of correlated patterns (see details in *Methods*), focusing on how the successful retrieval of a pattern depends on its S_f value, and how a decrease in C results in the selective lost of memories. This illustrates how the effective memory load of a network depends not only on the number of patterns that are being stored but also on how *informative* they are. An autoassociative memory could store virtually infinite patterns, for example, if they were constructed in such a way that all of the

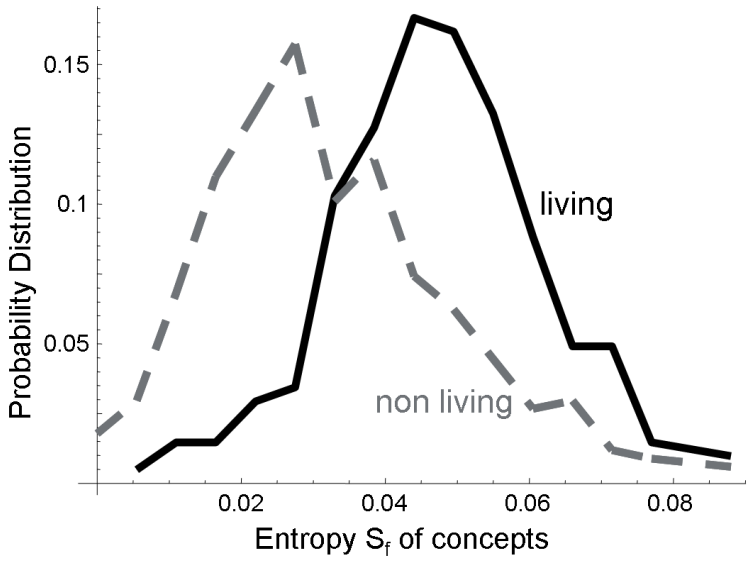

Figure 4.4: Distribution of S_f in concepts belonging to the 'living' and the 'non living' categories obtained from the feature norms of McRae and colleagues [McRae et al., 2005]. Living things have a distribution centered at higher values of S_f, which in terms of our analysis means that they are more informative but also more susceptible to damage, as observed in patient studies.

neurons contributed vanishing entropy, and hence were minimally informative: this would be the case if some neurons were active in nearly every pattern, while others in none, keeping the mean activity fixed to a value a. This result is in agreement with the notion that any associative memory network is ultimately constrained in the amount of information each of its synapses may store [Gardner, 1988].

The other interesting aspect of Eqs. 4.29 and 4.31 is that memory patterns are rather independent from one another in their retrievability. In the process of lowering C (which is, as explained in *Introduction*, the strongest effect of random network damage in our model) any pattern with a low value of S_f would be retrieved even when most of the other patterns have become irretrievable. Generally speaking, *informative memories are lost, while non-informative ones are kept.*

This model thus offers a quantitative explanation of category specific effects, along principles similar to those suggested, in a non mechanistic way, by several previous studies [Tyler et al., 2000, Sartori and Lombardi, 2004, McRae et al., 1997]. In our network, the

classical dichotomy would be verified if the semantic representations of *living* things had on average higher values of S_f than those of *nonliving* things, a plausible assumption that can be assessed using evidence in the relevant literature. As an example, we analyze the feature norms of McRae and colleagues, experimentally obtained representations of 541 concepts in terms of 2526 features [McRae et al., 2005] (see *Methods*). In Figure 4.4 we show that the distributions of S_f in the two categories overlap, but they are centered around different values of S_f, with living things on average more informative, hence more vulnerable to damage – a trend that is consistent with our analysis [3].

4.3 Discussion

Several experimental studies investigating semantic memory from the perspective of feature representation suggest that the representation of concepts in the human brain present non-trivial correlations [Vinson and Vigliocco, 2002, Garrard et al., 2001], presumably reflecting to some extent non-trivial statistical properties of objects in the real world or in the way we perceive them. It has not yet been proposed, however, how a plausible memory network could store reliably such representations; while attempts to model the storage of feature norms (experimentally obtained prototypes mimicking concept representations) with attractor networks have had success only using small sets of memories [McRae et al., 1997, Cree et al., 1999, Cree et al., 2006]. We propose here a way in which a purely Hebbian autoassociative memory could store and retrieve sets of correlated representations of any size, using a number of connections per neuron C that increases proportionally with p.

Interestingly enough, our learning rule is not quite appropriate for a one-shot learning

[3] One could feel tempted to store the patterns obtained from these norms in a network in order to simulate damage in a more direct way. However, the performance of the network is very poor due to the fact that the popularity distribution of the norms $F(x)$ has a power-law decay. This poor performance does not contradict the theory developed here, but rather validates it, as elaborated in [Kropff, 2007].

process, since it requires to calculate statistical properties of the dataset - the popularity of neurons - *before* learning the patterns. In the case of semantic memory, concepts are acquired through a long time experience and through the repeated exposure to diverse versions of the input, allowing, if necessary, for a continuous updating of popularity estimates. Episodic memory, on the other hand, requires one-shot learning, leaving no time for a learning rule like ours to deal with the correlation between memories. Associative networks may have evolved in other directions to enable the on-line storage of episodes and events. Evidence has recently been obtained [Leutgeb et al., 2007] supporting the suggestion that the dentate gyrus acts as an orthogonalizing device in the heart of the medial temporal lobe episodic memory system [Treves and Rolls, 1992]. The hippocampus could then function as an orthogonalized buffer, that helps neocortical networks acquire correlated memories through an off-line process. It has been proposed [Marr, 1971, Wilson and McNaughton, 1994, Hinton et al., 1995] that it is during sleep that the hippocampus transfers to cortical areas the statistical biases of the input, in a process of *consolidation*. While one-shot learning of a large dataset of orthogonal or randomly correlated patterns can be achieved through the 'standard' rule of Eq. 4.3, the learning or stabilization of correlated memories in their final cortical destination may be consolidated by a learning process that reflects what in our model we have defined as the popularity of different neurons. Such consolidation may well accompany the spontaneous retrieval of representations stored in the hippocampus [Squire and Zola-Morgan, 1991, McClelland et al., 1995].

Our results show that correlated representations can be stored at a cost: memories lose homogeneity, some remaining robust and others becoming weak in an inverse relation to the information they convey. These side effects should be observed in any associative memory system that is understood to store correlated patterns directly, and absent if information is first equalized through pattern orthogonalization.

Conversely, one may ask: are there benefits in representing correlated memories as they are, without recoding them into a more abstract, orthogonalized space? We have

shown in a previous study [Kropff and Treves, 2007a] that correlation plays a major role in driving a *latching* dynamics in a model of large cortical networks, in a process that could be a model of free association, and that might also underly the capacity for language [Treves, 2005]. Also, semantic priming has been shown to be guided by correlation [Vigliocco et al., 2004, Cree et al., 1999], selectively facilitating or inhibiting the retrieval of concepts, and potentially compensating for impaired episodic access [Ciaramelli et al., 2006]. On the other hand, embodied theories of cognition suggest that far from creating a neural structure of its own, the semantic system evolved on the same neural substrates that already had a primary function (visual, tactile or motor processing, etc.), for which correlation in the representation, even if useful, would be an inevitable outcome of their history.

Some predictions of our theory could perhaps be tested experimentally. The most immediate result to test is the relationship between the distribution of patterns and their relative robustness. The distribution of neural activity of different memory representations is however not available, for obvious technical reasons. Imaging techniques do not offer the required resolution, and collecting adequate statistics from single unit recordings in animals appears prohibitive. Nevertheless, other measurable quantities could yield an estimate of relevant statistical properties of the distribution: priming effects, for example, are related to the correlations between memory items. A second way to test the theory could be to assess the retrieval of a memory by a partial cue, similarly to what has been proposed in [Sartori and Lombardi, 2004], where the authors associate retrievability with a particular statistical measure: the *semantic relevance* of the cue. A third possibility could be to measure the speed of retrieval, which can be related to Eqs. 4.29 and 4.31 and, again, to the specific cue that the network receives to trigger recall. In this last case, however, retrieval activity in the semantic system should be isolated from other processes, such as categorization, which could take place automatically, affecting the overall timing. Probing different systems other than semantic memory might also be a possibility, since our conclusions are general to any associative network with correlated memories. If a set of stimuli with con-

trolled correlations were to be constructed (for example a set of pictures of caricature faces with exchangeable features), the memory of subjects trained with these stimuli could be tested for retrievability. The time-to-forget should then be related to the robustness, and inversely to the information content of each item, while with orthogonalized representations forgetting should be equalized.

4.4 Methods

4.4.1 Sets of patterns used in simulations

In the simulations shown in Fig. 4.1 a hierarchical algorithm was used to generate the patterns. The main idea is to produce, in the first place, a generation of random 'parent' patterns which are not part of the dataset but are used to influence with different strength a second generation, $\{\xi^\mu\}$ (more details and a full analysis of the statistics of the resulting patterns can be found in [Kropff and Treves, 2007a]). The reason to use this particular algorithm is that we needed a distribution of patterns with approximately the same correlation properties independently of p and N. Following our studies in [Kropff and Treves, 2007a], this is the case with the above algorithm, as long as p and N are not too small and asymptotic statistics applies.

For the simulations in Fig. 4.3 we needed higher levels of correlation than the ones that we could obtain with the algorithm described above, so as to illustrate the effects of large variability in the S_f values of the patterns. On the other hand, we did not require in this case patterns with more than one value of p and N. We then chose an algorithm that sets approximately an arbitrary popularity distribution over neurons. We chose

$$P(a_i) = \frac{1}{a} \exp\left(-\frac{a_i}{a}\right), \tag{4.40}$$

as the target distribution of popularity $F(x)$, with $\langle P(a_i) \rangle \simeq a$. Since the total number of

patterns is p, we defined the function

$$n_k = NP\,(k/p) \tag{4.41}$$

expressing, when rounded to the closest integer, how many neurons should be active in k patterns. For values of $n_k > 0.5$, we assigned a target popularity $a_i = k/p$ to $round(n_k)$ arbitrary neurons. To construct each pattern μ we initially set all neurons in the pattern to be inactive. Then we picked neuron i at random and set $\xi_i^\mu = 1$ with probability P_i, until aN neurons had been set to be active for each pattern. Finite size effects caused the actual distribution of popularity, shown in Fig. 4.3a, to be slightly different from the target one in Eq. 4.40, specially for low values of popularity. Since this region of the distribution is the less interesting one (see Section 4.4.3), we did not modify the patterns further.

The feature norms analyzed in Fig. 4.4 were downloaded from the *Psychonomic Society Archive of Norms, Stimuli, and Data* web site, www.psychonomic.org/archive, with the consent of the authors. The norms list $p = 541$ concepts relating several of $N = 2526$ features to each one of them. To each concept we associated a μ index and to each feature a i index. We set $\xi_i^\mu = 1$ if feature i was included in the description of pattern μ and $\xi_i^\mu = 0$ otherwise. Since not all patterns are associated with the same number of features, the sparseness is not constant across patterns. The average sparseness is $a \simeq 0.006$ equivalent to ~ 15 features per concept. For each concept, S_f is calculated as the average value of $a_i(1 - a_i)$ among the features that comprise it.

4.4.2 Testing the stability of memories

The stability of a memory item should be tested irrespective of how accurate a cue it needs in order to be retrieved. For this reason, we used the full original pattern as a cue, which is a good approximation of its attractor. The initial state, thus, is set to coincide with the tested pattern. In each update step, a neuron i is chosen at random and updated using the rule in Eq. 4.2, keeping track of m, whose initial value is close to 1 by construction. Initially, m varies rapidly, but it eventually converges to a stable value, either near 1 or near 0. A

proof of this is the step like transition in the stable values of m, shown in Figure 4.3b. The simulation stops when the variation of m is smaller than a threshold, which we set small enough to give three digits accuracy in m.

4.4.3 Storage capacity of more general distributions

As we have shown in *Results*, the important quantity to estimate in order to find the scaling of the storage capacity of a memory network with correlated patterns is the second term in the RHS of Eq. 4.31

$$\mathcal{T}_2 = \frac{1}{a^2}\int_0^1 dx F(x)x(1-x)\phi\left(\frac{-U}{\sqrt{\alpha x q}}\right). \tag{4.42}$$

The factor $\phi\left(-U/\sqrt{\alpha x q}\right)$ is 0 when $x = 0$ and reaches its maximum when $x = 1$. On the other side, since we consider the sparse limit $a \ll 1$ the distribution $F(x)$ is concentrated toward small values of x. For these two reasons, the interesting part of any smooth distribution function $F(x)$ is the decay of its tail with increasing x. We study in this section two interesting cases: exponential and power-law distributions. Keeping in mind that the exact behavior of $F(x)$ for small values of x is less relevant, these results can be generalized to any distribution function with such tails.

Exponential distribution

The exponential distribution

$$F(x) = \frac{\exp(-x/a)}{a} \tag{4.43}$$

is normalized to 1 and has mean equal to a – apart from a small correction of order $\exp(-1/a)$, which we neglect for simplicity. Its variance is about a^2, with a correction of the same order. Finally, $S_f \simeq a(1-2a)$. The critical second term in the RHS of Eq. 4.31 is

$$\mathcal{T}_2 = \frac{1}{a^2}\int_0^1 dx F(x)x(1-x)\int_{-\infty}^{\sqrt{y/x}} Dz = \frac{1}{a^2}\int_{\sqrt{y}}^{\infty} Dz \int_{y/z^2}^1 dx \frac{\exp(-x/a)}{a} x(1-x) \tag{4.44}$$

where we have inverted the integration order. Dz is the gaussian differential defined in Eq. 4.24 and $y = U^2 a/(\alpha S_f)$. The inner integral in the right-most side of the equation confirms

that the value of $F(x)$ for small x is less relevant than its decay for large x. The RHS is now integrable, resulting in

$$\mathcal{T}_2 = \frac{1}{a^2}\int_{\sqrt{y}}^{\infty} dz \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2} - \frac{y}{az^2}\right)\left[S_F + \frac{y}{z^2}\left(1 - \frac{y}{z^2} - 2a\right)\right]. \tag{4.45}$$

This expression can be integrated a second time, but its analytical expression is too complicated to include here. It is enough to mention that the largest contribution is proportional to $\exp\left(-\sqrt{2y/a}\right)$

$$\mathcal{T}_2 \simeq \frac{1}{2a^2}\exp\left(-\sqrt{\frac{2y}{a}}\right)\left(S_F + \sqrt{\frac{ay}{2}} - \frac{a}{2}\sqrt{\frac{ay}{2}} + \frac{ay}{2} - 2a\sqrt{\frac{ay}{2}}\right). \tag{4.46}$$

Assuming $2y/a \sim 1$ modulo some logarithmic correction (that we consider inside the exponential and neglect elsewhere) this results in

$$\mathcal{T}_2 \simeq \exp\left(-\sqrt{\frac{2y}{a}}\right)\frac{3}{4a^2}S_F. \tag{4.47}$$

Since only y depends on α_c it is easy to see from this equation that indeed $2y/a \sim 1$ modulo logarithmic corrections, making the previous assumption self-consistent. The storage capacity can be obtained by making the RHS of Eq. 4.47, as in the previous section, equal to S_f/a,

$$\alpha_c \simeq \frac{2U^2}{S_f\left[\ln\left(\frac{3S_F}{4aS_f}\right)\right]^2} \propto \frac{1}{S_f\left[\ln\left(\frac{S_F}{aS_f}\right)\right]^2}. \tag{4.48}$$

Note that the square on the logarithmic factor makes this storage capacity lower than the one found for $F(x)$ distributions of very low variance. Again, the correction is valid as long as the logarithm is large, in other words $\ln(S_F/aS_f) > 1$. If this condition is not met, the storage capacity scales like $1/S_f$.

Power law distribution

We define the power law distribution

$$F(x) = \begin{cases} 0 & \text{if } x < d \\ cx^{-\gamma} & \text{if } x > d \end{cases} \tag{4.49}$$

with $\gamma > 2$ and d a small cutoff value that prevents the integral of $F(x)$ from diverging. The conditions for normalization and mean are

$$1 = c\left(\frac{d^{1-\gamma}-1}{\gamma-1}\right) \tag{4.50}$$

$$a = c\left(\frac{d^{2-\gamma}-1}{\gamma-2}\right). \tag{4.51}$$

There is no simple analytical expression for c, d or S_F in terms of a and γ.

We want to compute

$$\mathcal{T}_2 = \frac{1}{a^2}\int_d^1 dx\ c\ x^{-\gamma}x(1-x)\phi\left(-\sqrt{\frac{y}{x}}\right) \tag{4.52}$$

where, again, $y = U^2 a/(\alpha S_f)$. $\mathcal{T}_2$ is integrable, resulting in

$$\begin{aligned}
\mathcal{T}_2 &= \frac{c}{a^2}\phi\left[-\sqrt{y}\right]\left(\frac{1}{\gamma-3}-\frac{1}{\gamma-2}\right)+\frac{c}{a^2}\phi\left[-\sqrt{\frac{y}{d}}\right]d^{2-\gamma}\left(\frac{d}{\gamma-3}-\frac{1}{\gamma-2}\right)- \\
&- \frac{c}{a^2(\gamma-3)}\left(\frac{1}{2\sqrt{\pi}}\left(\frac{y}{2}\right)^{3-\gamma}\left\{\Gamma\left[-\frac{5}{2}+\gamma,\frac{y}{2}\right]-\Gamma\left[-\frac{5}{2}+\gamma,\frac{y}{2d}\right]\right\}\right)+ \\
&+ \frac{c}{a^2(\gamma-2)}\left(\frac{1}{2\sqrt{\pi}}\left(\frac{y}{2}\right)^{3-\gamma}\left\{\Gamma\left[-\frac{3}{2}+\gamma,\frac{y}{2}\right]-\Gamma\left[-\frac{3}{2}+\gamma,\frac{y}{2d}\right]\right\}\right)
\end{aligned} \tag{4.53}$$

where $\Gamma[,]$ is the incomplete gamma function. The following series expansions are useful

$$\phi[-\sqrt{y}] = \frac{\exp(-\frac{y}{2})}{\sqrt{2\pi y}}\left\{1+\sum_{k=1}^{\infty}\left[\prod_{j=1}^{k}(2j-1)\right](-y)^{-k}\right\} \tag{4.54}$$

$$\frac{1}{2\sqrt{\pi}}\left(\frac{y}{2}\right)^{n-\gamma}\Gamma\left[-n+\frac{1}{2}+\gamma,\frac{y}{2}\right] = \frac{\exp(-\frac{y}{2})}{\sqrt{2\pi y}}\left\{1+\sum_{k=1}^{\infty}\left[\prod_{j=1}^{k}(2j-1+2(n-\gamma))\right](-y)^{-k}\right\}.$$

$\mathcal{T}_2$ is different from 0 only to order y^{-2} inside the curly brackets. At this order of approximation

$$\mathcal{T}_2 \simeq \frac{4c\exp(-y/2)}{a^2\sqrt{2\pi y^5}} \tag{4.55}$$

neglecting a similar term including the factor $\sqrt{d^5}\exp\left(-\frac{y}{2d}\right)$. As previously, the storage capacity can be estimated as

$$\alpha_c \propto \frac{a}{S_f \ln\left(\frac{a^{\gamma-2}}{S_f}\right)} \tag{4.56}$$

where we have used $c \propto a^{\gamma-1}$. If the logarithm is of order 1 or smaller the storage capacity scales simply like a/S_f.

Chapter 5

General Discussion

The present thesis work investigates different aspects of the proposal following which large cortical networks, and in particular those devoted to storing and retrieving semantic memory, function as autoassociative memories at two different levels, local and global. Such a proposal can be interpreted as a convergence point between two very different lines of thought. On one side, the speculations of Braitenberg and Schuz about possible computing mechanisms based on anatomical studies of the mammalian cortex. On the other, the feature representation framework, put in the center of the scene during the past decade by several groups of neuropsychologists inspired mostly on behavioral and imaging studies in humans.

Is this architecture better than that of a simple attractor network (equivalent to a Potts network with $S = 1$) in terms of storage capacity? In other words, has evolution found in this architecture a solution for storing more memories? Though one might be tempted to answer positively by the S^2 scaling of the storage capacity found in Chapter 2, a few considerations must be made before doing so. If the human cortex was thought to be a simple attractor network with approximately $C \sim 10^4$ connections per neuron and a sparseness of, for example, $a \sim 0.01$, the maximum number of storable and retrievable memories would be around $p_{max} \sim 0.1C/a \sim 10^5$. This number, of the order, for example, of the total number of words contained in the lexicon of a normal language, seems too tight as an upper limit. In order to calculate a similar estimation for a multi-modular network

it is necessary to introduce the detailed description rather than just the Potts approximation. Johansson and Lansner have optimized the architecture of such a network in a simplified system with non overlapping minicolumns clustering into modules and interconnected by binary synapses [Johansson and Lansner, 2007a, Johansson and Lansner, 2007b]. Inserting their optimized parameters into our model would result in a network with $S = 100$ and 4×10^3 out of the 10^4 connections per neuron dedicated to inter-modular communication. With such parameters, the estimation for a Potts network would be $p_{max} \sim 10^8$. However, this result depends strongly on their assumption of 5 synapses being enough to serve to communicate two given pre and post synaptic microcolumns, or, translated to the Potts network, two distant 'states'. Whether or not this number of connections, valid for their system, is enough in a non simplified two level autoassociative memory [O'Kane and Treves, 1992, Fulvi Mari and Treves, 1998], can only be assessed by studying the full model.

Chapter 3 shows that the complexity of the 'symbolic' series resulting from latching dynamics in a Potts memory network depends on the equilibrium between local self excitation of modules and global inhibition. It further proposes a non-linear dynamics approach to study the bifurcation generating the transition from a non latching to a latching system and a statistical study of the latching probability matrix to asses the possible transition toward infinite latching. We have made some partial progress in both directions [Russo et al., 2007]. In a mean field approach, we have obtained the dynamical equations that describe the behavior of the macroscopic variables across time in the case of a network storing two patterns. We have identified three regions in the space of parameters that correspond to three different latching mechanisms, a distinction that Chapter 3 does not make. A good marker for differentiating the regions is the value of m at which the transition is produced, named λ. In the same paper, large sets of simulations with many patterns in the fashion of those in Chapter 3 are analyzed, but this time considering the λ value of each transition. In a surprising consistency, three peaks appear in the resulting distribution of λ, corresponding to

the three regions observed with the simplified differential equations. The most frequent kind of transition is the one associated to the measure of correlation $\mathcal{C}_1^{\mu\nu}$ introduced in Chapter 3. Next in frequency of occurrence are transitions related to the units that are active in both patterns but in different states. A unit that is active in a given state and adapts seems to have some tendency to 'bounce' into another state of activity, especially when S is small. A last pathological kind of transition is actually a stable cycle during which several patterns are brought to a constant high level of retrieval that serves as a baseline to very deterministic oscillations, which seem to be a dynamical counterpart of the spurious states found in classical Hopfield networks, undesired attractor states resulting from combinations of several stored patterns. The differential equations predicting the dynamics of interaction between two memory states seem so far to be the best way to study, in the future, the dependence of each kind of latching transitions on the equilibrium between local self excitation and global inhibition. The results may then be extrapolated to the more general simulations, as in the cited paper, adding if necessary a noise term in the field representing the influence of the rest of the attractor states.

Chapter 4 introduces a new property of autoassociative networks. Each stored pattern has an associated storage capacity that is inverse to the information it carries. In addition, the slower is the decay of the distribution of popularity $F(x)$ toward high values of x, the worse is the overall performance of the network. For fast decaying popularity distributions the minimal connectivity per neuron necessary to sustain the retrieval of a memory is simply $C_{min} \propto pS_f$, while slower decays can produce much higher minimal connectivity levels.

The idea that feature correlations are the cause of selectivity in semantic impairments is not the only account of such deficits. The other main lines of thought are the sensory-functional theory [Warrington and Shallice, 1984, Warrington and McCarthy, 1987] and the domain specific theory [Caramazza and Shelton, 1998]. The work in Chapter 4 does not help to elucidate which of these accounts is better. It limits to build a quantitative theory of correlated memory storage, parallel to the the more qualitative formulations developed,

for example, in [Tyler et al., 2000, Tyler and Moss, 2001, McRae and Cree, 2002] or other quantitative but non mechanistic accounts in the same line [Sartori and Lombardi, 2004, Sartori et al., 2005].

In the simplest studies, patients with semantic impairments have been reported to have deficits in handling concepts related to either living or non living things. Capitani and colleagues [Capitani et al., 2003] have analyzed the whole literature concluding that 77% of the studies report impairments with living things against 23% reporting the opposite trend. Any account of these results has to explain first of all two facts: a) there is a double dissociation and b) there is a higher probability of impairments related to living things. So far, the present approach has explained the latter but not the first of these observations. However, it is possible that other not yet studied ways to damage the network (for example selective damage of connections or random damage of neurons) result in the opposite pattern of selective loss, given that, as shown in Chapter 4, the storage of correlated representations brakes the symmetry between categories. Future studies will be directed to ellucidate this possiblity.

Another unsolved point in this work is the actual storage of feature norms in a simulated network. Some partial results using the norms of McRae and colleagues [McRae et al., 2005] show that this might not be a trivial problem. The popularity distribution $F(x)$ in this group of patterns is a power law, the worst of the situations analyzed in Chapter 4. The solution of the equations in this Chapter applied to the feature norms predicts a $\sim 50\%$ performance in the retrieval of the stored memories. In simulations using a fully connected network (since N, the total number of features, is fixed, one must set $C = N$ to exploit the resources to the maximum), no retrieval is achieved regardless of the choice of parameters. The discrepancy between predictions and simulations is explained by the fact that the correction to the equations due to the full connectivity is not negligible and very strongly dependent on correlations. In [Kropff, 2007] I show several strategies to improve the performance of the network, based on the idea of pushing $F(x)$ toward low values of x. These are: the elimination

of the few most popular neurons, the expantion of the representation by incorporating new very unpopular neurons and the modulation of the strength of synapses by a function that depends on the popularity of the pre-synaptic neuron. All of these approaches can push the performance of the network near to 100%. In particular, the latter might be a possible cortical mechanism to counterbalance the effect of correlations, though it is difficult to imagine how to test in an experiment such a hypothesis. For reasonable popularity distributions $F(x)$ it is enough to modulate the weights with an exponentially decaying function of the popularity of the pre-synaptic neuron to get an effective $F(x)$ that decays exponentially or faster, reaching the optimality limit in terms of how the storage capacity scales with S_f. In fact, the performance of the network does not depend strongly on the particular choice of the modulatory function as long as it decays fast enough. Again, as throughout this thesis, there is a balance between storage capacity and retrievable information. The storage capacity increases at the expense of having a lower effective information in the representations, a quantity that should be defined in future works as the information that can actively participate in the retrieval of a memory, given that informative neurons, due to the modulation, are taken less into account than uninformative ones.

Bibliography

[Amati and Shallice, 2007] Amati, D. and Shallice, T. (2007). On the emergence of modern humans. *Cognition*, 103:358–85.

[Amit, 1989] Amit, D. J. (1989). *Modelling Brain Function: the World of Attractor Neural Networks*. Cambridge University Press.

[Barsalou, 2005] Barsalou, L. W. (2005). Continuity of the conceptual system across species. *Trends Cogn Sci.*, 9(7):309–11.

[Bienenstock et al., 1982] Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2:32–48.

[Blumenfeld et al., 2006] Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52(2):383–394.

[Bolle et al., 1993] Bolle, D., Cools, R., Dupont, P., and Huyghebaert, J. (1993). Mean-field theory for the Q-state Potts-glass neural network with biased patterns . *Journal of Physics A: Mathematical General*, 26:549–562.

[Braitenberg and Schuz, 1991] Braitenberg, V. and Schuz, A. (1991). *Anatomy of the Cortex: Statistics and Geometry*. Springer Verlag.

[Buhmann et al., 1989] Buhmann, J., Divko, R., and Schulten, K. (1989). Associative memory with high information content. *Phys Rev A*, 39:2689–2692.

[Buxhoeveden and Casanova, 2002] Buxhoeveden, D. P. and Casanova, M. F. (2002). The minicolumn hypothesis in neuroscience. *Brain*, 125:935–951.

[Capitani et al., 2003] Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). What are the facts of semantic category-specific deficits? a critical review of the clinical evidence. *Cognitive Neuropsychology*, 20:213–261.

[Caramazza and Shelton, 1998] Caramazza, A. and Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. Cogn. Neurosci.*, 10(1):1–34.

[Ciaramelli et al., 2006] Ciaramelli, C., Lauro-Grotto, R., and Treves, A. (2006). Dissociating episodic from semantic access mode by mutual information measures: evidence from aging and alzheimer's disease. *J Physiol Paris*, 100:142–53.

[Cree et al., 2006] Cree, G. S., McNorgan, C., and McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning : Implications for theories of semantic memory. *Journal of Experimental Psychology*, 32:643–658.

[Cree et al., 1999] Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

[Derrida et al., 1987] Derrida, B., Gardner, E. J., and Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *Europhysics Letters*, 4:167–173.

[Devlin et al., 2002] Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, M. J., and Tyler, L. K. (2002). Is there an anatomical basis for category-specificity? semantic memory studies in pet and fmri. *Neuropsychologia*, 40(1):54–75.

[Diederich and Opper, 1987] Diederich, S. and Opper, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58(9):949–952.

[Edwards and Anderson, 1975] Edwards, S. F. and Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F Metal Physics*, 5:965–974.

[Fulvi Mari and Treves, 1998] Fulvi Mari, C. and Treves, A. (1998). Modeling neocortical areas with a modular neural network. *Biosystems*, 48:47–55.

[Fuster, 1999] Fuster, J. M. (1999). *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate*. MIT Press.

[Gardner, 1988] Gardner, E. J. (1988). The space of interactions in neural network models. *J. Phys. A: Math. Gen.*, 21:257–270.

[Gardner et al., 1989] Gardner, E. J., Stroud, N., and Wallace, D. J. (1989). Training with noise and the storage of correlated patterns in a neural network model. *J. Phys. A: Math. Gen.*, 22:2019–2030.

[Garrard et al., 2001] Garrard, P., Ralph, M. A. L., Hodges, J. R., and Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18:125 – 174.

[Greer et al., 2001] Greer, M. J., van Casteren, M., McLellan, S. A., Moss, H. E., Rodd, J., Rogers, T., and Tyler, L. K. (2001). *The emergence of semantic categories from distributed featural representations*, pages 386–391. Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society. Erlbaum, Mahwah, NJ.

[Gutfreund, 1988] Gutfreund, H. (1988). Neural networks with hierarchically correlated patterns. *Phys. Rev. A*, 37(2):570–577.

[Hauser et al., 2002] Hauser, M., Chomsky, N., and Fitch, W. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298:1569–1579.

[Hebb, 1949] Hebb, D. (1949). *The organization of behavior*. Wiley: New York.

[Hinton et al., 1995] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

[Hopfield, 1982] Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational habilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.

[Johansson and Lansner, 2007a] Johansson, C. and Lansner, A. (2007a). Imposing biological constraints onto an abstract neocortical attractor network model. *Neural Computation*, 19:1871–1896.

[Johansson and Lansner, 2007b] Johansson, C. and Lansner, A. (2007b). Towards cortex sized artificial neural systems. *Neural Networks*, 20:48–61.

[Kanter, 1988] Kanter, I. (1988). Potts-glass models of neural networks. *Phys Rev A*, 37:2739–2742.

[Kropff, 2007] Kropff, E. (2007). Full solution for the storage of correlated memories in an autoassociative memory. Manuscript to appear in the proceedings of the international meeting "Closing the gap between neurophysiology and behaviour: A computational modelling approach", Birmingham, May 2007.

[Kropff and Treves, 2005] Kropff, E. and Treves, A. (2005). The storage capacity of potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(08):P08010.

[Kropff and Treves, 2007a] Kropff, E. and Treves, A. (2007a). The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185.

[Kropff and Treves, 2007b] Kropff, E. and Treves, A. (2007b). Uninformative memories will prevail: the storage of correlated representations and its consequences. submited.

[Leutgeb et al., 2007] Leutgeb, J. K., Leutgeb, S., Moser, M.-B., and Moser, E. I. (2007). Pattern separation in the dentate gyrus and ca3 of the hippocampus. *Science*, 315:961 – 966.

[Marr, 1971] Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 262:23–81.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychol. Rev.*, 1023:419–457.

[McRae, 2005] McRae, K. (2005). *Psychology of Learning and Motivation*, volume 45, chapter 2, pages 41–82. Elsevier.

[McRae and Cree, 2002] McRae, K. and Cree, G. S. (2002). *Factors underlying category-specific semantic deficits*, pages 211–249. Category-specificity in mind and brain. Hove, UK: Psychology Press.

[McRae et al., 2005] McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559.

[McRae et al., 1997] McRae, K., de Sa, V., and Seidemberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99–130.

[Mechelli et al., 2003] Mechelli, A., Price, C. J., Noppeney, U., and Friston, K. J. (2003). A Dynamic Causal Modeling Study on Category Effects: Bottom-Up or Top-Down Mediation? *J. Cogn. Neurosci.*, 15(7):925–934.

[Monasson, 1992] Monasson, R. (1992). Properties of neural networks storing spatially correlated patterns. *J. Phys. A: Math. Gen.*, 25:3701–3720.

[O'Kane and Treves, 1992] O'Kane, D. and Treves, A. (1992). Short and long range connections in autoassociative memory. *Journal of Physics A: Mathematical General*, A 25:5055–5069.

[O'Kane and Treves, 1992] O'Kane, D. and Treves, A. (1992). Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems*, 3:379–384.

[Parga and Virasoro, 1986] Parga, N. and Virasoro, M. A. (1986). The ultrametric organization of memories in a neural network. *J. Physique*, 47(11):1857–1864.

[Pulvermuller, 2001] Pulvermuller, F. (2001). Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, 5:517–524.

[Pulvermuller, 2002] Pulvermuller, F. (2002). A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Progress in Neurobiology*, 67:85–111.

[Rakic, 1995] Rakic, P. (1995). A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *TINS*, 18:383–388.

[Roudi and Treves, 2004] Roudi, Y. and Treves, A. (2004). An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010.

[Russo et al., 2007] Russo, E., Treves, A., and Kropff, E. (2007). Free association transitions in models of cortical latching dynamics. Manuscript in preparation.

[Sartori and Lombardi, 2004] Sartori, G. and Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16(3):439–452.

[Sartori et al., 2005] Sartori, G., Polezzi, D., Mameli, F., and Lombardi, L. (2005). Feature type effects in semantic memory: An event related potentials study. *Neurosci. Lett.*, 390(3):139–144.

[Schyns and Rodet, 1997] Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23:681–696.

[Shiino and Fukai, 1992] Shiino, M. and Fukai, T. (1992). Self-consistent signal-to-noise analysis and its application to analogue neural network with asymmetric connections. *J. Phys. A*, 25:L375.

[Shiino and Fukai, 1993] Shiino, M. and Fukai, T. (1993). Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Phys. Rev. E*, 48:867.

[Squire and Zola-Morgan, 1991] Squire, L. R. and Zola-Morgan, S. (1991). The Medial Temporal Lobe Memory System. *Science*, 253:1380–1386.

[Srivastava and Edwards, 2000] Srivastava, V. and Edwards, S. F. (2000). A model of how the brain discriminates and categorises. *Physica A*, 276:352–358.

[Srivastava and Edwards, 2004] Srivastava, V. and Edwards, S. F. (2004). A mathematical model of capacious and efficient memory that survives trauma. *Physica A*, 333:465 – 477.

[Treves, 1990] Treves, A. (1990). Graded-response neurons and information encodings in autoassociative memories. *Phys Rev A*, 42:2418 – 2430.

[Treves, 2005] Treves, A. (2005). Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology*, 6:276–291.

[Treves and Rolls, 1991] Treves, A. and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network*, 2:371–397.

[Treves and Rolls, 1992] Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, 2:189 – 199.

[Tsodyks and Feigel'Man, 1988] Tsodyks, M. V. and Feigel'Man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6:101–105.

[Tyler and Moss, 2001] Tyler, L. K. and Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, 5:244–252.

[Tyler et al., 2000] Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., and Levy, J. P. (2000). Conceptual structure and the structure of concepts: a distributed account of category-specific deficits. *Brain and Language*, 75(2):195–231.

[Vigliocco et al., 2004] Vigliocco, G., Vinson, D. P., Lewis, W., and Garret, M. F. (2004). Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4):422–88.

[Vinson and Vigliocco, 2002] Vinson, D. P. and Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics*, 15:317–351.

[Virasoro, 1988] Virasoro, M. A. (1988). The effect of synapses destruction on categorization by neural networks. *Europhys. Lett.*, 7(4):293–298.

[Warrington and Shallice, 1984] Warrington, E. and Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107(3):829–854.

[Warrington and McCarthy, 1987] Warrington, E. K. and McCarthy, R. A. (1987). Categories of knowledge. further fractionations and an attempted integration. *Brain*, 110(5):1273–1296.

[Wills et al., 2005] Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.

[Wilson and McNaughton, 1994] Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265:676 – 679.