ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

A Theoretical Model of Neocortical Association Areas as Large Autoassociators

Thesis submitted for the degree of
"Doctor Philosophiae"

CANDIDATE
Carlo Fulvi Mari

SUPERVISOR
Alessandro Treves

November 1998

SISSA - SCUOLA
NTERNAZIONALE
SUPERIORE
STUDI AVANZATI
TRIESTE
Via Beirut 2-4

TRIESTE

A Theoretical Model of Neocortical Association Areas as Large Autoassociators

Carlo Fulvi Mari [1]
SISSA - Neuroscience Programme
International School for Advanced Studies
Trieste - Italy

[1] email: C.Fulvi-Mari@Lboro.ac.uk
Present address: *Nonlinear and Complex Systems Group*, Dept. Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, U.K.

Contents

Introduction **2**

1 Neocortical modular structures **6**

1.1 Areas and columns 6

1.2 Minicolumns 13

1.3 Connections 15

1.3.1 Thalamo-cortical and cortico-thalamic projections 15

1.3.2 Cortico-cortical connections and parallelism 16

1.4 Organization and function of the columns 17

2 Modular autoassociators **19**

2.1 Autoassociative memory 19

2.2 Neocortical areas as large autoassociators 23

2.2.1 The cortical column as an autoassociator 23

2.3 Analysis of an 'isotropic' multimodular model 25

2.3.1 Modules and connectivity 25

2.3.2 Pattern structure 28

2.3.3 Metastable states 29

2.3.4 The phase diagram 29

2.4 May a modular autoassociator model neocortex? 31

3 A model for neocortical associative areas **35**

3.1 Two steps toward the suppression of memory glass 35

3.2 The mathematical model 38

3.3 Parameters of the model 43

3.4 The correlation of modular activity 44

3.5 Signal-to-Noise analysis 46

3.6 Results 47

3.7 Comments on the synaptic rule 51

Appendix 52

4 Bounds to the correlation between modules 53
4.1 Introduction and summary 53
4.2 Mathematical basis of the method 56
4.3 The randomly interacting system 59
4.4 Existing theorems 64
4.5 Conditions from self-averaging 64
4.6 A Fictitious Dynamics from a MRF 66
4.7 Analytical approach 68
4.8 Analogy with the Ising ferromagnet 70
4.9 Simulation results 71
4.10 Conclusions 74
Appendix A 76
Appendix B 76

5 Neuropsychology and the model: features, modular retrieval, and category-specific impairment 79
5.1 Features and category-specific impairments 79
5.2 A topographic computational model 81
5.3 The intriguing case of Alzheimer Disease 83
5.4 A non-topographic model for category-specific deficits 84
5.5 The modular neuronal model with selective semantic deficits 88

Acknowledgements 91

Bibliography 92

Introduction

In this work I have analyzed the plausibility of autoassociative neuronal networks as models for neocortical areas and their functions. The constructive elements of the model network are based on the neuroanatomical and physiological literature. I have tried to keep some fundamental characteristics of real neuronal systems in the modeling, such as the use of the neuron as the elementary unit of cerebral systems. I have concentrated my attention on the associative areas of neocortex, the regions in which autoassociative processes are plausibly more frequent and effective. In spite of the attempted neurophysiological realism, I have kept the model simple enough to allow for the analytical treatment of large ensembles of neurons and modules, and I have investigated the consequences of some assumptions on elementary cognitive processes postulated by the models, such as the retrieval of stored patterns.

In neuropsychology the use of the so called 'connectionist' models is very common. These models often succeed in reproducing cognitive deficits due to brain damage. They usually schematize the brain structures involved in the processes investigated with a set of *units* and *connections* between units. Each unit is thought to be responsible for a particular function, and the connections allow for the interaction and the co-operation of several functions toward the common cognitive goal. The network is constructed and trained so to perform some particular cognitive tasks; then the 'lesion' is introduced as a damage to a particular subset of units or connections, and the consequences of such operation on the 'cognitive' abilities of the network are analyzed. The aim is to find out a relation between the damaged subsystem and the cognitive deficit in the view of transposing such mapping to the real brain and the deficits observed in neurological patients. Quite often the units and the connections of connectionist models are given the dynamics and the distributed representation typical of neuronal networks. Nevertheless, their identification with neurophysiologically concrete neural systems is usually improbable.

I think that this gap between cognitive models and neuronal models should be filled, and the present work is an attempt to move toward this direction.

The first chapter of the thesis is a short overview of the anatomical and physiological data about mammalian neocortex, particular attention being paid to the evidences of modular anatomy and of a possibly modular organization subserving cognitive functions, in particular memory. The connectivity structure plays a crucial role in information processing.

In the second chapter, I introduce the concepts of autoassociative memory and of memory retrieval, also describing how some analytical models reproduce the properties of autoassociators. Some neocortical structures, due mainly to their connectivity, are plausibly working as autoassociators. Then a particular analytical model introduced earlier is described. Such a model is composed by a large set of interconnected modules of neurons, and it presents severe difficulties in performing retrieval of complex patterns. The authors of the original model discussed their results suggesting that the bad performance is a consequence of the implausibility of their model. In the last section of the second chapter I analyze the main contrasts of that model with biological systems, and hypotesize that the necessary modifications in order to make such model more realistic may also provide the network with a more realistic 'cognitive' performance.

In chapter 3, I describe a model with more realistic 'anatomy' and 'physiology', that has been studied by me and Alessandro Treves. We obtain a marked improvement of the retrieval abilities in comparison with the model described in the previous chapter, mainly due to the refinement of the extramodular connectivity structure and the introduction of correlation between the activity of connected modules. The results obtained seem to give new support to the modeling of neocortical areas by modular autoassociators. The relation between the co-variance of modular activity and the underlying long-range connectivity is thought to represent semantic knowledge.

Chapter 4 contains a mathematical demonstration of the validity of some assumptions that are of fundamental importance in the model considered in the previous chapter. Indeed, in that model the suppression of the noxious states that marred the performance of the model of chapter 2 is strongly dependent on the average correlation of the activities of adjacent modules. In this chapter the mathematical consistency of the induced set of marginal probabilities is proven and an upper bound to the average correlation is provided. This upper bound is fully compatible with the improvements obtained with the new modular network.

In chapter 5, I attempt to reproduce peculiar cognitive deficits, due e.g. to Alzheimer's disease, within our new model. In particular, I try to verify the hypotheses recently proposed in the neuropsychology literature about the basic phenomena that are responsible for category-specific impairments.

In conclusion, I think that our results sustain the notion of autoassociative multimodular neuronal networks as candidates for modeling neocortical association areas. The obstacle of glassy states seems to be overcome and a proposal is formulated for finding deep relations between the semantic system and the neuronal correlates.

Chapter 1

Neocortical modular structures

It is well known that the mammalian neocortex is far from being a homogeneous sheet of neural tissue. Beside the vertical organization in six *layers* (Braitenberg and Shutz, 1991), one can find cytoarchitectonical and physiological differences moving throughout the cortex parallel to its surface (Kaas, 1987; Fuster, 1997; Mountcastle, 1997).

1.1 Areas and columns

At the scale of a few millimeters, anatomical differences sharply define *areas* (Kaas, 1987); this is the scale usually considered by lesion studies, i.e. in neuropsychology. Indeed it has been known for long that a lesion onto the cortical tissue may provoke cognitive deficits whose pattern strongly depends on the location of the damage. Neuropsychological studies of patients with brain damages have allowed the construction of an approximate map of the cortex in which anatomically distinguishing characteristics of a region quite often correspond to peculiar cognitive functions that are presumably subserved by that region.

Neurophysiological studies investigate more finely the biological structures subserving cortical functions, and at the scale of one millimeter these studies reveal the columnar organization of neurons and neural projections (Mountcastle, 1997).

In the *somato-sensory* cortex, microelectrode penetrations normal to the pial surface reveal that the neurons, tested along the way, have very similar receptive fields on the skin. On the contrary, if the electrode moves parallel to the surface, the receptive fields of the encountered neuron are very similar for a penetration length of 300-500 μm, but then abrouptly change and another sequence of neurons is found, whose receptive fields are very similar to each other and evidently distinct from those of the preceding sequence (fig. 1.1).

V. B. Mountcastle

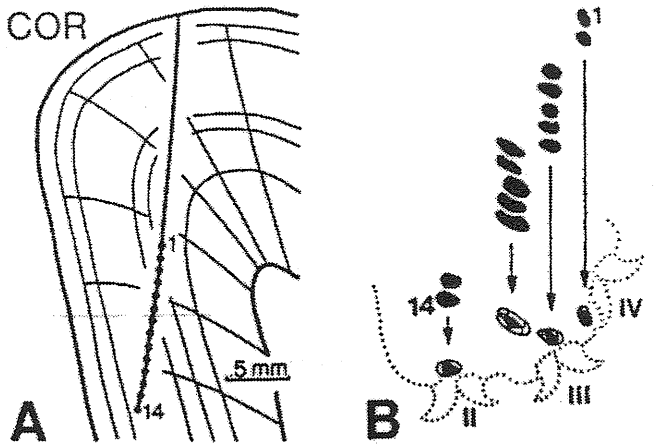

Fig. 6 Location of the peripheral receptive fields of 14 neurons observed in a microelectrode penetration made tangentially through the somatic sensory cortex of the cat. (A) The electrode passed nearly parallel to the pial surface. (B) There are three abrupt shifts in receptive field locations, as the electrode traversed a series of adjacent columns. (From Favorow, 1991, with permission from MacMillan.)

Figure 1.1:

A well defined columnar organization is also found in the *visual* cortex. For example, it is well known that any neuron in area V1 responds preferentially to visual stimuli shown to one of the eyes and in a well defined restricted visual field of that eye. Besides, when tested with short straight line stimuli, it is selectively sensitive to the lines at a particular angle of orientation. When the electrode penetrates vertically the visual area V1, it encounters neurons with the same ocular dominance, the same visual field, and with the same preferential line stimuli direction (exceptions may occur in layer IV). If instead the penetration is oblique to the surface (fig. 1.2), sharp changes in some of the perceptual preferences are detected. In particular, if the electrode moves in certain directions, an abrupt change of the directional selectivity happens at regular intervals such that the preferential angle of orientation varies monotonically until the 180 degrees round is completed. The complicate distribution of *iso-directional hypercolumns* crossing *ocular dominance* columns is still the object of intensive studies, both theoretical and experimental.

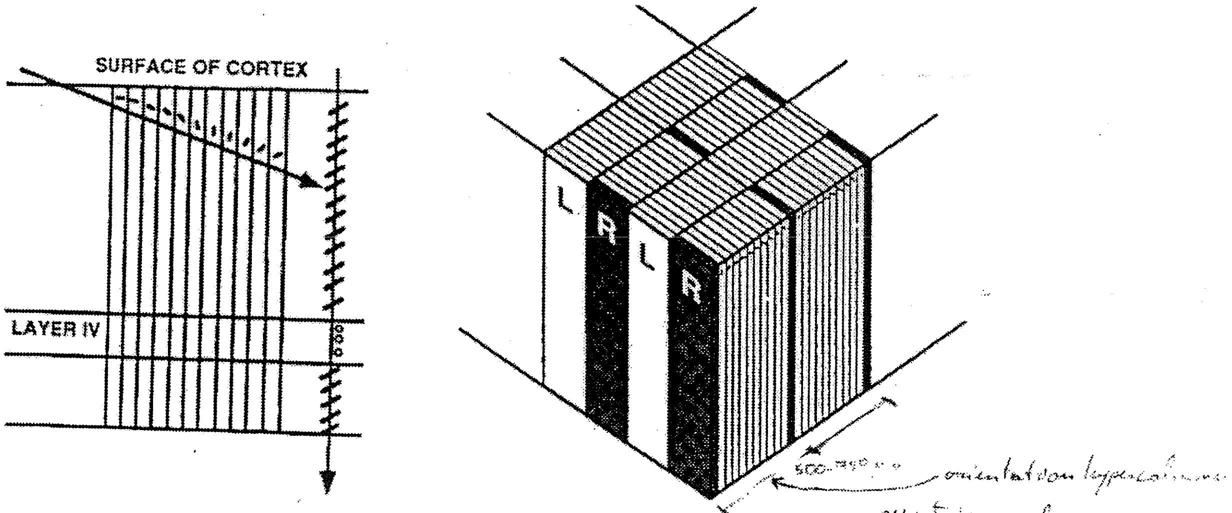

Fig. 9 *Left*. Organization of the striate cortex for orientation preferences, from the work of Hubel and Wiesel. For penetrations made perpendicular to the surface of the cortex the preferred orientation remains almost constant, except for layers IVa and IVc. For penetrations made almost parallel to the cortical surface, the preferred orientation rotates linearly with distance in sequences of 0.5–1.0 mm. Linearity is broken at longer intervals by reversals in the direction of rotation. *Right*. The Hubel and Wiesel model, generated from the inference that slightly rotated orientation preferences may be organized in stacks of parallel slabs, with spacing between slabs of ~700 µm. (From Blasdel, 1992*a*, with permission from Oxford University Press.)

Figure 1.2:

Columnar organization can be found also in other regions, like in the *motor* and in the *auditory* cortices, suggesting that **modularity** is common maybe to all the neocortex. In the scope of the present work our interest is focused on **association** areas, located in the parietal, temporal, and frontal lobes. Neurons of these areas cannot be classified according to their selectivity to simple sensorial stimuli. Lesion studies have shown that the damage to a portion of these higher areas produces impairment with tasks requiring complex behaviour. The *posterior parietal* cortex seems to be responsible for higher perceptual memory and processing (Fuster, 1997). There, neurons with similar properties are arranged in vertical columns extending across the layers; neurons of adjacent columns have different properties (Mountcastle et al., 1975; Mountcastle, 1995). No simple sensorial input seems to select activity in such regions. What all the columns in that area have in common is their activability in tasks involving the actions in, or the perception of, or the attention to the environment that surrounds the subject. A column could be active during fixation of gaze, or slow pursuit tracking, or reaching by an arm, or manipulation, or visual stimulation (Mountcastle, 1997). In the latter, the activating stimuli are much more complex than those that select columnar activity in the primary visual cortex, being for example the flow fields generated by the displacement of objects or by the movement of the subject itself.

Strictly related to the motion flow are the neurons in area **MT**. Albright et al. (1984) have found that neurons of monkey MT with similar motion direction preferences are arranged in vertical columns. These columns are organized in slabs (fig. 1.3) such that, moving from any column in a certain horizontal direction into the cortical sheet, a sequence of columns is found whose preferred motion directions constitute a discrete monotone succession of angles that describes a complete 360 degrees rotation in about 1mm of linear progression through the cortex. If from the same starting column one instead move horizontally in the direction orthogonal to the previous one, one finds a succession of columns whose favorite motion flows are just 180 degrees rotated each step of the sequence.

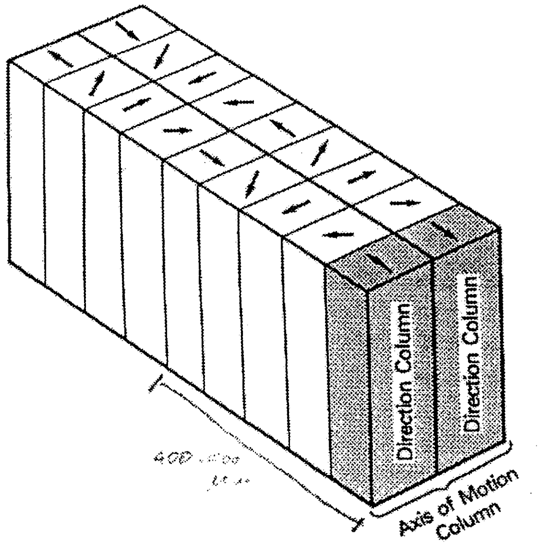

Fig. 13 Three-dimensional model of columnar organization for direction and axis of motion sensitivity of neurons in prestriate area MT in the macaque monkey. Vertical dimension represents depth in the cortex.The long axis of the figure represents two complete revolutions of axis of motion columns; the two directions of motion are represented in adjacent columns. When moving at right angles to the long axis, one encounters frequent 180° reversals in the preferred direction of motion, with no change in the preferred axis of motion. (From Albright *et al.*, 1984, with permission from the American Physiological Society.)

Figure 1.3:

In the nearby area **MSTd**, the columns are responsive to more complicated visual flow stimuli like contraction, expansion, rotation and translation. The flow selectivity has a non-trivial distribution onto the columns in this area (Graziano et al., 1994) such that each column is selective to a characteristic combination of the cited flows, that is a spiral, according to a continuous tuning curve. Besides, the visual fields of the neurons in MSTd area are very large and the neuronal response does not depend on the position of the reference object in its visual field.

The recognition of shapes and therefore of objects is mainly due to the **inferotemporal** cortex. In particular, electrophysiological studies (Tanaka, 1996) have shown that neurons in area TE respond selectively to visual stimuli that can be moderately complex (fig. 1.4) but also to some very rich of details like individual faces. TE neurons are thus responsive to non-elementary visual features. According to Tanaka (1996), TE neurons may be also sensible to orientation, size, and contrast polarity of their critical visual features, though being neutral to the object position in their large visual fields. Neurons recorded during electrode vertical penetration into TE are selective for the same feature or a slightly different one (Fujita et al., 1992). During oblique penetration, the monitored neurons respond to features that are slightly different as the electrode procedes for a maximum of about 400μm along the horizontal axis. Going further on, neurons are found that respond selectively to a completely different visual feature. This is the signature of the *columnar*, and actually *minicolumnar* (cf. section 1.2), organization of area TE in the inferotemporal cortex.

Columnar organization is evident also in **frontal association** cortex. Connections from the ipsilateral and contralateral areas are interdigitated into alternating columns that span the full thickness of the cortical sheet and whose width is 500μm (Goldman and Schwartz, 1982; Benes et al., 1988). The frontal lobe supports the highest levels of the hierarchy of motor memories. The *primary motor* cortex represents and mediates elementary motor acts. Above it, in the hierarchy, the *premotor* cortex encodes motor acts and programmes defined by goal, sequence or trajectory, rather than by specific movement or muscle group (Fuster, 1997). Higher in the motor hierarchy is the *prefrontal* cortex, that is commonly identified as the association cortex of the frontal lobe. Rich of connections with subcortical, limbic and neocortical areas, like the reciprocal connections with the association perceptual posterior

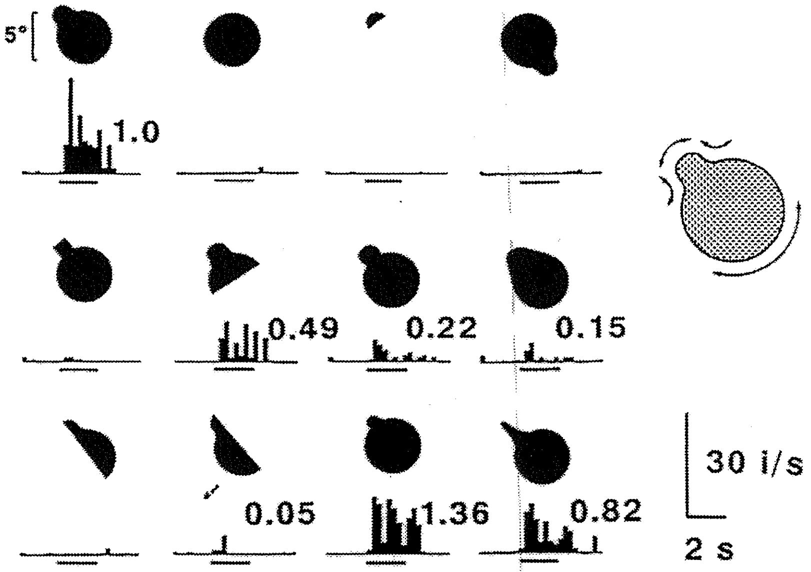

FIGURE 2. An example of further study of the selectivity after the reduction process was completed. This second cell is different from the cell whose responses are shown in Figure 1, but was also recorded from TE (from Kobatake & Tanaka, 1994).

Figure 1.4:

cortex, the prefrontal cortex may keep the representing schemas of goal-directed actions, usually named *plans*. In general, the activation of a motor representation in prefrontal cortex is more evident when the subject is learning a new motor task. After practice, it seems that the 'schemas' are relocated in motor areas that lie in lower levels of the motor hierarchy (Jenkins et al., 1994). The prefrontal cortex seems to support also other cognitive functions than pure motor planning. This is suggested, for example, by the results of some *delay task* experiments. A typical delay task is made of three steps: 1) a sensory cue; 2) the removal of the stimulus and a delay during which the subject must retain memory of the cue in order to execute the task correctly; 3) a motor response that must be the one appropriate to the given cue. During the second phase, in animal experiments, marked self-sustained neuronal activity is found in the prefrontal cortex. Also the posterior association area maintains high activity level since a perceptual stimulus is analyzed, and if the stimulus involves visual recognition also inferotemporal regions keep high activity. The prefrontal cortex is activated independently of the stimulus modality maybe due to its involvement in the planning of the movement to come, according to the learned task. In this view, the *working memory* is not peculiar of prefrontal cortex (Fuster, 1997). The latter indeed is often performing working memory processes because of the motorial components that usually are part of the experiments, and that not necessarily are acted. As expected, the co-operation between posterior and frontal association cortices is intensive, since both the presence of an external stimulus, that signals the beginning of a well defined task, and the planning of a motor action, that may be acted in the surrounding environment with possible involvement of objects, need a complex perceptual analysis (see e.g. Friedman and Goldman-Rakic (1994)). It is often hypothesized that the self-sustained activity is strictly related to the retrieval of old or new learned features in recurrent neuronal networks localized in the cited areas according to the kind/class/category of features they analyze.

Whether the columns are an epiphenomenon, due e.g. to developmental processes (Rakic, 1988; Rakic, 1972), or a convenient solution produced by natural selection is not entirely clear and the debate about this question is still open (Purves et al., 1992; Hevner et al., 1993). Nor it is clear what could be the possible advantages deriving from such an organization (Malach, 1994; Favorov and Kelly, 1994; Jacobs et al., 1991).

1.2 Minicolumns

The *column*, however defined, does not seem to be the ultimate elementary module, since there is anatomical evidence that cortical columns are made up of collections of parallel quasi-cylindrical constructions, the *minicolumns*, of densely connected neurons (Tommerdahl et al., 1993; Mountcastle, 1997; Tanaka, 1996).

As described in the previous section, small oblique displacements of the electrode in the somatosensory cortex reveal discontinuous shifts in the receptive fields. In area SI cylindrical regions about 50μm wide, with axis orthogonal to the cortical sheet, are found whose neurons have almost coincident receptive fields. Such cylinders are commonly called **minicolumns**. Neighboring minicolumns are usually grouped for having similar and overlapping receptive fields (Favorov and Whitsel, 1988; Favorov and Diamond, 1990), and are thus considered as constituents of the sensorial column selective for that part of the skin. The column is found to be about 350μm wide, being composed of about 80 minicolumns. As already stated, when the electrode exits a column to enter an adjacent one, the receptive field is found to have an abroupt marked shift on the skin surface. Each minicolumn may contain about 80-100 densely connected neurons. The presence of minicolumns has been verified in somatosensory SI also with the analysis of metabolism within single columns (Tommerdahl et al., 1993).

Also the small changes in visual stimuli selectivity found in area TE (cf. previous section) are attributed to the passage of the recording electrode from a minicolumn to an adjacent one (Tanaka, 1996).

These and other anatomical and physiological studies are currently suggesting to consider the minicolumn as the constructive element of the neocortical networks. The explanation of this phenomenon of diffuse modularity in mammalian neocortical structures is still lacking (at least to our knowledge).

1.3 Connections

It is often remarked (Braitenberg and Shutz, 1991; Young, 1993; Scannell et al., 1995) that the distribution of long-range projections is not governed by randomness except in the neighbourhood of the target region, where in any case experience is expected to model synaptic contacts. One function that short-range extramodular projections may subserve is the control of the average activity level, exerted by inhibition (Tommerdahl et al., 1993; Favorov and Kelly, 1994; Jacobs and Donoghue, 1991) against the spreading of excitation. The local intra-modular connectivity is relatively high (Braitenberg and Shutz, 1991), supporting the hypothesis that each module may serve, inter alia, as an autoassociative network (cf. sections 2.1 and 2.2). This hypothesis is also supported by the finding of persisting activity (Miyashita and Chang, 1988; Fuster, 1997) and of neuronal responses consistent with retrieval and recognition capabilities, especially in associative areas (Mountcastle, 1997; Tanaka, 1996; Van Hoesen, 1993).

1.3.1 Thalamo-cortical and cortico-thalamic projections

Old studies indicate that each cytoarchitectonic area of the cortex receives input from one and only one major thalamic nucleus, beside the input from the so-called 'non-specific' nuclei that project diffusely onto many cortical areas. Instead more recent studies indicate that each cortical area receives projections from several thalamic nuclei. Several studies have indicated that thalamic terminal fields are disjunctive in association cortex (Giguere and Goldman-Rakic, 1985; Jones et al., 1982), as they are in sensory regions (Jones et al., 1982; Jones, 1985). Besides, the cortical prefrontal neurons of layer VI that project to thalamus are found to form .5mm wide clusters, horizontally alternating along the cortical sheet with .5mm-wide neuronal clusters that are deprived of cortico-thalamic fibres. The reciprocal thalamocortical projections seem to spread more irregularly onto the cortical sheet, without selecting cortical modules to synapse into.

Any thalamic nucleus sends projections to several cortical areas. Nevertheless there is some evidence that thalamic neurons projecting to neocortex are organized in clusters, and different clusters seem to project to different cortical areas. It may be that any thalamic nucleus can activate a large network of cortical areas (Goldman-Rakic, 1988).

Thus, experimental evidence seems to suggest that the thalamo-cortico-thalamic system is organized in modules such that specific groups of cells in the thalamus project upon and receive input from specific columns of cortical cells.

1.3.2 Cortico-cortical connections and parallelism

For a long time, association areas have been treated as more or less homogeneous modality-unspecific neocortical regions. Recent studies seem to be in contrast with this view, and also suggest that at least from anatomical considerations associative areas can be subdivided in small areas in which a considerable degree of modal specificity survives (Goldman-Rakic, 1988; Cavada and Goldman-Rakic, 1986). In the parietal lobe these anatomical findings, in some cases supported by electrophysiological data (Hyvarinen, 1982; Mountcastle et al., 1984), suggest that small associative regions may be specialized for a sensorial modality more than for the others, as if they were interested in some aspects of the input signals that are markedly conveyed through a sensorial mode more than through the others. Besides, the projections from these quasi-specialized sub-areas to the prefrontal cortex terminate onto well defined and non overlapping regions, suggesting some modal preferences also in prefrontal associative areas.

On the other hand, a sub-division of the prefrontal cortex is suggested also by electrophysiological investigations. It seems that the prefrontal cortex can be divided in at least two parts that are responsible for different cognitive functions. The *dorsal* region of the prefrontal cortex may be specialized for working memory of spatial knowledge, while the *ventral* region may be more involved in non-spatial working memory tasks (Goldman-Rakic, 1987).

In support to the theory of parallel processing in prefrontal and parietal associative areas, there are also some anatomical findings on the connections of the two regions with the rest of the cortical and sub-cortical systems: "Double-label studies reveal that posterior parietal and dorsolateral prefrontal cortices project in common to virtually the same targets in over a dozen distinct cytoarchitectonic areas. [...] Moreover, the prefrontal and parietal axons within these 'third party' targets terminate in one of two characteristic modes: either as interdigitated , spatially alternating fiber columns or in complementary layers within a single column or set of columns."(Goldman-Rakic, 1988)

(Though being a very interesting problem, in the present work we do not consider the different modes of the axons in approaching the cortical targets, e.g. into different layers. We hope to be soon able to put these properties into our analytical model and to give them some functional interpretation.)

1.4 Organization and function of the columns

Assuming modularity and considering that inside a module the connectivity is sensibly higher than between different modules (Braitenberg and Shutz, 1991; Mountcastle, 1997), the question arises of how this multitude of units is organized to perform cognitive tasks. First among these tasks, the ability to retrieve from memory distributions of neuronal activity, as allowed by recurrent connectivity (Braitenberg and Shutz, 1991; Fuster, 1997). Fundamental in this view is the structure of the *long* *range connectivity.*

Different areas seem to detect different properties of the percept (see, e.g., Tanaka (1996), Mishkin et al. (1983)). Columns (or minicolumns?) may be the sites for the memory of the finest elements of the percept. (In this context, the term "percept" should be considered in its widest meaning, thus including also intellectual experiences.) Thus, many modular retrievals of single elementary features must be coordinated to result in a complete global pattern. The *global attractors* of the large network's dynamics must represent realistic correct combinations of local features, while "strange" combinations which have never been stored in memory should be suppressed.

Different "generalized" percepts (that from now on will be called *patterns*) may share some elementary features; this may also be a source for apparently abstract or even unconscious associations and analogies among initially different patterns (e.g. the *priming* effect).

One of the advantages of a modular structure of this kind could be the possibility of storing any feature that is common to many patterns in one of the modules, so that it can be used again in new patterns and consequently the whole network does not have to produce a completely new global activity distribution for every pattern.

Chapter 2

Modular autoassociators

The concept of autoassociative memory and the properties of neuronal autoassociators are introduced.

2.1 Autoassociative memory

By definition an *autoassociator* is a device able to retrieve a memorized representation of an item when cued with a portion, or a noisy version, of the 'perceptive' input from that item. For this reason, any autoassociative memory system is often referred to as *content addressable memory*, that is the stored information about an object can be extracted from the system memory by providing a fraction of the same informative content. Systems of this kind are of interest in many applicative disciplines, since they lend themselves as feature detectors, hand-written manuscripts readers, and for many other practical tasks in which it is necessary to verify whether an incomplete or noisy input pattern actually comes from a 'known' object and, in the case, to complete the representation of the object correctly.

The concept of autoassociative memory is one of the main themes of neuroscience. It is widely accepted that the peripheral input stimuli to mammal nervous system produce peculiar distributions of activity in cortical structures subserving cognitive functions (cf. Chap.1). The recognition of a perceptual input as a familiar (that is, already experienced and learned) complex stimulus plausibly coincides with the reproduction of a characteristic neuronal activity distribution at the level of some cortical areas. Besides, the neural circuitry found in some cortical regions could in principle be able to support autoassociative process.

One of the first, and maybe the best known, analytical models of autoassociative neural networks is the Hopfield model. In the simplest version of that model, neurons are represented by binary units, thus considering only two possible states of activity of any neuron, that could be interpreted as the firing and the non-firing neuron states. The recurrent connectivity among the neurons is the key ingredient to provide the network with autoassociative abilities. While in the simple pattern associators the flux of the signal is unidirectional (fig.2.1), going from the input toward the output of the neural system so that the state of a neuron cannot influence the state of the afferent neurons, the recurrent networks are provided with backward connections that allow for recursive processing.

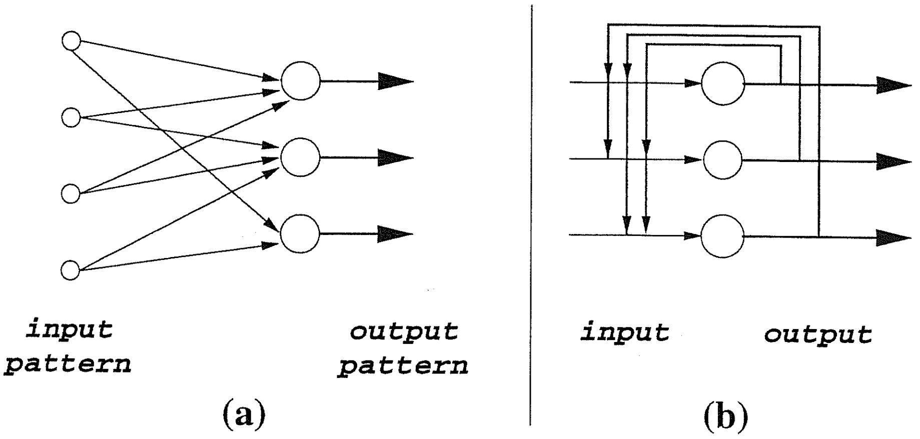

Figure 2.1: (a) Simple pattern associator: to any presented input the network associates one output pattern. (b) Recurrent network (autoassociator): once the neuronal activities are put in a specific initial condition and then let free to follow the dynamics determined by the collateral interactions, the activity distribution evolves toward the configuration that among the learned patterns is the most similar to the initial condition.

In the Hopfield model, the influence (the *field* h_i) of the afferent fibres on a given neuron (i) is represented by a weighted summation of their activity binary variables ($S_j = -1, 1$):

$$h_i = \sum_{j,j\neq i} J_{ij} S_j. \tag{2.1}$$

The 'weights' represent the synaptic strengths. Their values are determined by the correlation of the neuronal activities across the set of *learned* patterns according to the formula

$$J_{ij} = \frac{1}{N} \sum_{p=1}^{P} S_i^p S_j^p, \tag{2.2}$$

where N and P are the number of neurons and that of the learned patterns respectively, and S_i^p is the state of activity of neuron i when the network is imposed with the pattern of activity p. During the learning phase, that is the period in which the patterns are arbitrarily given to modify the synapses, the statistics of the patterns is supposed to attribute a neuron a given state of activation with probability 1/2 across the different patterns and independently from the other neurons. In this model it is assumed that the network is either in a learning phase or in an interrogation phase; in the latter the input is used as a cue possibly to retrieve a stored pattern, but cannot modify the synapses. The *hebbian* (covariance) learning rule (eq.2.2) is such that the synaptic weights are symmetric, that is the synapse of the projection of neuron i onto neuron j has the same strength that the one in the reciprocal connection $j \rightarrow i$. Besides, the network is fully connected. Evidently this model is quite a rude approximation of biological nets. Though numerous more elaborated evolutions of this model have been produced, it already shows the properties of autoassociative memory due to the existence of metastable states in one-to-one correspondence with the set of the learned patterns. Each of these metastable states is an attractor in the network dynamics, so that when the network is put in the basin of one of the attractors the dynamics drive the neurons to reproduce the distribution of activity characteristic of the learned pattern corresponding to the considered attractor. It could be said that, once put in an initial configuration of neuronal activities, the network evolves toward the most similar configuration among the learned ones.

The existence of the multiplicity of attractors, together with their basins, correctly corresponding to the learned configurations, may be corrupted when the network is requested to store too many patterns. Indeed, when the number of stored patterns approximates the number of neurons, the co-operative effect no longer dominates over the noise and the Hopfield network develops a very large number of attractors whose (meta)stable configurations have null overlap with every stored pattern. In this case, the network cannot work as an autoassociator since the probability that its dynamics is captured by one of the undesired attractors is very near to one.

The problem of the existence of an upper bound to the number of patterns that can be stored is common to all neuronal network models, though the nature of the phase beyond the upper bound may depend on the particular model (for example, models with analogue neuronal outputs do not present the marked spin-glass behaviour shown by the Hopfield network beyond maximal storage (Treves, 1991)). Supposing that the approximation of the continuous input evolution with a discrete sequence of representative patterns makes sense, the number of patterns that a neuronal network can store in its synapses is a fundamental static characteristic of the considered network. Since in many models the thermodynamic limit is necessary[1], the number of patterns that can be efficiently stored diverges; thus one usually defines its ratio with another diverging quantity (e.g., the number of neurons in the Hopfield model previously described) as the *storage capacity* of the consid-

[1] It means that in order to obtain correct results from the analytical treatment it is necessary to consider systems with very large number of neurons (infinite, in the ideal case). Practically, such 'thermodynamic' theories usually are good model for large finite systems.

ered neuronal network. When one tries to model a system of recurrently connected real neurons within the analytical framework of the attractor neural networks, it is necessary that the theoretical storage capacity is large enough to be biologically plausible. This is sometimes what decides a theoretical model to be viable or not.

2.2 Neocortical areas as large autoassociators

The neurophysiological data exposed in Chap.1 are interpreted in terms of neural network architecture and hypotheses are formulated about the subserved cognitive functions.

2.2.1 The cortical column as an autoassociator

As stated in Chap.1, the connectivity inside a *column* is very high compared to the average connectivity across the neocortex. This suggests to consider cortical areas as composed by a large number of *processing units.* Whether the fundamental unit is the column or the microcolumn is not crucial in the theoretical context that is to be introduced, and the term 'column' will be used in the following without specifically referring to one of the two possibilities.

The idea of modelling the neocortex as composed of a number of *compartments* is not new. In neuropsychological models of brain functions and impairments the presence of modular structures is very common, though usually no precise correspondence is provided between the *compartment* and the physiological *system* that in the real brain should subserve the functions theoretically ascribed to the former.

Also from the neuroanatomy some proposals of modularity have emerged. In Braitenberg (1978) the author suggested to consider the cortex as a set of $\sqrt{N}$ compartments each composed of $\sqrt{N}$ neurons, being N evidently the total number of cortical neurons. He justified his idea on the basis of connectivity statistics.

Indeed the possibility of a full connectivity among the cortical neurons is not to be considered at all. In a more plausible way the Braitenberg's model allows the neural information to propagate throughout the cortical areas. In that model every compartment sends an axon to every other compartment, apparently respecting the anatomy statistics about the number of white matter projections per pyramidal cell and the white matter volume. It is interesting that such simple approximate reasoning brings to a quantitative estimate of the compartment width that in the human brain roughly equals the columnar diameter.

Here we consider the column as a *module*, and a large number of its replicas is considered to form both the physiological and the functional architecture of neocortical areas. To be as near to the anatomical and electrophysiological data as possible, we restrict the scope of our model to associative neocortical areas (cf. Chap.1).

The prototypical column, for the densely recurrent structure of connections among its neurons, is an excellent candidate to have autoassociative memory abilities, so we consider the module as an efficient autoassociator of the kind of those described in section 2.1. In this view we are supported by numerous experimental findings of local self-sustained activity (Miyashita and Chang, 1988; Fuster and Jervey, 1982) and, above all, by the visual features responses found by Tanaka (1996) in the infero-temporal cortex of primates[2].

When the *long-range* cortical connections, that is the projections between different modules, are included in the model, the elementary autoassociators build up a *large autoassociator*. We adopt the hypothesis that each module analyzes a particular 'aspect' of the information content of the axonal signals arriving at the associative areas from primary and more peripheral cortices. It is not easy to provide examples

[2] In experiments Tanaka and his group have considered the firing rate of the neural cells as indicative of the 'recognition' of a feature, disregarding any temporal organization of the spikes. We also avoid the problem of temporal information in the 'static' analyses exposed in the following.

of such feature decomposition since we think that it may be much finer and obscure than common sense could suggest. There are easily comprehensible cases, such as the recognition of geometrical figures, the detection of particular visual features and the possible modality of use of the observed object. It seems more difficult to find examples in non perceptual contexts. The featural description here adopted is not approved by the whole scientific community (see e.g. McRae et al. (1997)).

The very high ratio between the local connectivity and the long-range connectivity brings us to consider the retrieval process of the large autoassociator as deriving from the co-operation of the modules in producing a meaningful composition of locally retrieved features. Then the problem arises of verifying the plausibility of such modelling of neocortical associative areas, namely to test the ability of such model networks in storing a large number of *global* patterns and correctly reproducing them in retrieval tasks.

2.3 Analysis of an 'isotropic' multimodular model

An analytical model of multimodular cortex is reported[3]. The exposition of calculations is avoided since quantitative information can be derived from the model to be introduced in Chap.3 that in the aim of the present work is to be considered more general. Nevertheless large space is given to some of the notions since they will be useful in the next chapter.

2.3.1 Modules and connectivity

A network with a large number of neurons is considered. They are grouped in M modules (fig.2.2), each containing N neurons. The *local* network, that is the set of neurons belonging to the same module, with their *intramodular* projections, is

[3] Most of the discussion in this section derives from O'Kane and Treves (1992a).

taken fully connected. Thus each neuron receives $N-1$ inputs from the neurons of the module it belongs to. From the other $M-1$ moduli, the neuron is assumed to receive L projections. According to data in Braitenberg and Shutz (1991), it is assumed that one half of the afferent axons to any neuron come from the ipsimodular neurons, the other one half coming from the rest of the network, that is through the *white matter.* Since both N and M are taken to be diverging in the thermodynamic limit, it follows that the long-range connectivity is very diluted, being $\frac{L}{MN}$ the probability that any neuron receives a projection from a given neuron of another module. Such a connectivity has been chosen in O'Kane and Treves (1992a) with the aim of approximating the neurobiological data according to which the local connectivity is largely denser than the long-range one.

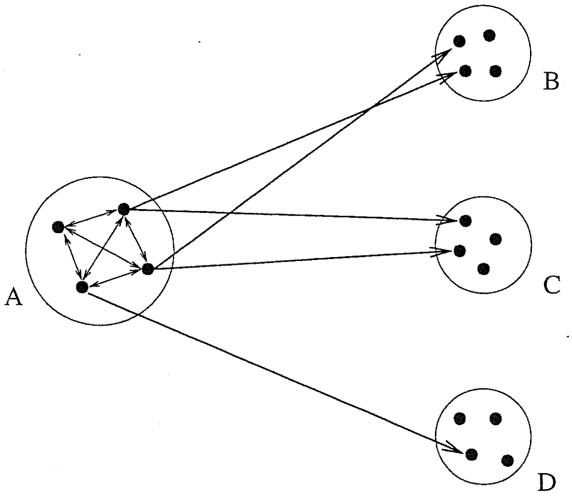

Figure 2.2: The structure of connections in the model of O'Kane and Treves (1992a). The black dots represent the neurons, the circles are the modules, the thin lines are the intramodular connections, the thick lines are the extramodular connections. For clarity we have drawn only the intramodular connections and the extramodular projections from the neurons in module A.

One of the assumptions that we will show to be crucial is about the distribution of the long-range, that is *extramodular*, axonal projections from any neuron: they are randomly distributed with constant probability across the neurons of all the other moduli. This is the reason why we call this model 'isotropic'.

Both the intramodular and the extramodular synapses are given weights according to a covariance rule. As shown in section 2.1, this rule assigns equal weights to reciprocal connections. Though at the level of the single module this means that any two neurons have symmetric coupling, in the long-range net actually no symmetry is required since the probability of having a bi-directional connection is vanishingly small due to the extreme dilution.

In O'Kane and Treves (1992a) the authors adopt a particular neuronal model in order to perform statistical mechanics calculations of the mean-field kind. They use the *threshold-linear* neuron model (Treves, 1990): the input signal is calculated as the weighted sum of the afferent neuronal outputs; if the input is below a characteristic threshold, the neuron is silent, if the input is above the threshold, the firing rate of the neuron is taken to be proportional to the difference between the input value and the threshold (fig.2.3). It has been shown (Treves, 1990) that (non-modular) networks of such neurons are able to perform the retrieval of stored configurations. In the analysis of our model (Chap.3) we will not make use of a detailed description of neuronal response, since this would preclude the possibility of obtaining compre-

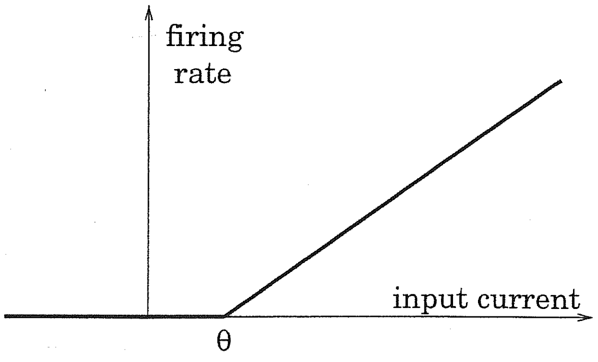

Figure 2.3: The input-output relation for the neuron model used in O'Kane and Treves (1992a). The threshold is indicated by θ. In abscissa, the value of the total input (arbitrary units). In ordinates, the spike firing rate.

hensible formulas due to the complexity of the more general framework. For this reason, we do not report the results of O'Kane and Treves (1992a) in their original mean-field fashion, preferring their signal-to-noise counterparts that are more easily comparable with our new results in Chap.3.

2.3.2 Pattern structure

The *global* pattern of activity, that is the distribution of the neuronal activity over the whole network ($N \cdot M$ neurons), is considered to be decomposed in M *features.* Namely, the flux of axonal information about what we may generically call *object*, arriving to the associative areas from more peripheral regions, is analyzed by the modules in a parallel manner. Each module takes care of the part of information carried by its afferents and that is supposed to be rich of details about one particular *aspect* of the object.

Different patterns may share some features. This means that in a module they may elicit the same distribution of neuronal activity, so that at the level of that module those patterns would be indistinguishable. In a sense this would be the consequence if two or more objects had some common 'characteristics'. In O'Kane and Treves (1992a) it is assumed that $\mu > 1$ patterns activate the same firing distribution in any given module. However, due to the high number of possible featural combinations, the probability that two or more patterns share common features in more than one module is vanishingly small (equal to zero in the thermodynamic limit).

Looked at as an isolate network, any module is supposed to store a large number D of local configurations, that are the features. Thus the total number of patterns stored in the large modular network is $P = \mu \cdot D$. It is also assumed that every pattern activate a precise feature in every module. Thus every module is always elicited and no state of quiescence is allowed. Besides, the activation of a feature in a module is taken to be independent from which features are activated in the other modules[4]. These also will reveal to be critical assumptions.

2.3.3 Metastable states

In O'Kane and Treves (1992a) it is shown that the proposed model network may be in several states with interesting cognitive interpretation.

One class of possible (meta)stable states includes the *retrieval states*, that is those states in which the network correctly sustains distributions of activity that reproduce the stored patterns. If the neurons are forced into a configuration of their activities that overlaps with one of the stored patterns, once let free the network is able to reconstruct all the corresponding original pattern of activity.

Beside the retrieval states, in fact some undesired states are present in large number. In each of these states locally, that is at the intramodular level, a good feature retrieval is performed, but different modules do not co-operate in producing a global retrieval. Since such state is a (meta)stable state, the network may settle in it and no longer evolve to a global retrieval state. If this happens, the network presents a meaningless patchwork of well defined but incoherent features. Namely, it is not able to recognize an object as a whole complex of co-ordinated characteristics. O'Kane and Treves (1992a) call these noxious states *memory glass*.

2.3.4 The phase diagram

O'Kane and Treves (1992a) calculated the critical storage of the network with mean-field statistical techniques. As explained at the end of subsection 2.3.1, we now show the corresponding results obtained with a signal-to-noise analysis, that is a

[4] Note that hypothesis is on the statistics of the pattern of signals arriving to the considered area, and not on the network dynamics.

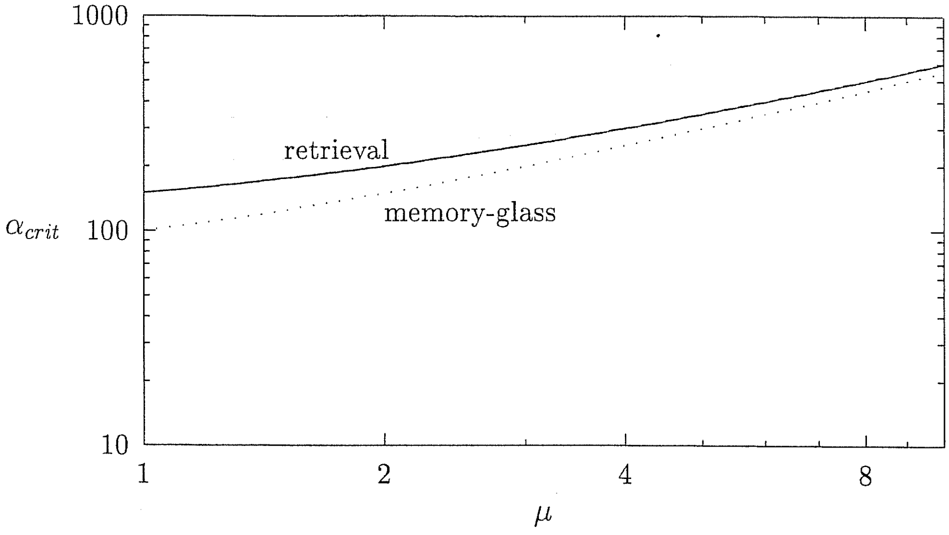

Figure 2.4: Plot of the critical storage of the network for *retrieval* states and *memory-glass* states. The useful range, that is the vertical distance between the two curves, is very small and decreases quickly as μ increases. This result is valid for very large networks with parameters $\gamma = .5$ and $a = .01$.

less precise and simpler method. This will allow us to calculate the capacity of the more complicated network of the next chapter with the further benefit of not requiring a detailed neuron model.

The storage capacity of the present network has different values if calculated either for retrieval states or for memory-glass states. As shown in fig.2.4, the retrieval is more robust, to the noise produced by the stored patterns, than memory glass (α_{crit} is defined as the ratio between the maximum number of patterns that can be stored before the considered retrieval or glassy states become unstable, and the number of connections per neuron, $C = L + N - 1$). Above the retrieval critical storage the network is unable to perform any kind of retrieval, neither global nor simply local. Below the memory-glass capacity the states of both global retrieval and memory glass are stable, so that the network with large probability falls in one of the numerous glassy states and does no longer evolve toward a complete global retrieval.

Thus the only possibility to make the network work as a good autoassociator consists in keeping the storage load in the range between the memory glass capacity curve and the retrieval capacity curve. Indeed, in that region the memory glass states are not stable and thus cannot prevent the network from flowing toward a good retrieval state.

Thus, in principle, if one does not require biological plausibility, the modular network described in this section may be a good autoassociator.

2.4 May a modular autoassociator model neocortex?

An attempt of computationally modelling semantic memory has been done by Lauro-Grotto et al. (1997), who considered the retrieval abilities of a modular network in which each module was thought to be responsible of the storage and processing of a given semantic 'aspect' of the input signal. They analized the network analytically and numerically in the noiseless limit, and at a *low storing* level, that is no quenched noise was generated by stored patterns in the thermodynamical limit. They considered the thermodynamic limit only for the number of neurons, while the number of modules was kept small (in particular, in simulations the number of modules was between 2 and 5). Although the connectivity between modules was taken smaller than the intramodular one, the two connectivities were of the same order of magnitude. The network was in principle able to sustain states in which a fraction of the modules were in a non-retrieval state, that is a state of null overlap with every locally stored pattern. Nevertheless the authors found that the states with a small number of retrieving modules were strongly unstable: very often, when set in a stored configuration and then let free, if the imposed configuration has a small number of active retrieving modules, the network prefers to evolve toward 'meaningless configurations' (Lauro-Grotto et al., 1997). The authors considered this behaviour as a sign of non-plausibility of the model and of the necessity of introducing modifications.

In this section we try to understand whether the network presented in section 2.3 is a good candidate to model the neocortical areas supposed to function as autoassociators. The only property we are to discuss is the retrieval ability. We think that this kind of quality is not crucially dependent on the details of the adopted neuron prototype and that also the study of a non-dynamical approximate model like ours can give insights about brain functioning.

As stated in the previous section, that network is able to perform good retrieval. Even disregarding the absolute storage capacity, whose biological plausibility is difficult to test, the network is nevertheless an implausible model for real brain areas (O'Kane and Treves, 1992b). Indeed the network performs correctly only if the number of the stored patterns lies between the two curves in fig.2.4. This interval is too thin to be biologically plausible, and becomes even thinner the higher is the average number of patterns which a single feature is part of (fig.3.1). It is difficult to imagine how a biological system could select a number of patterns to learn with such a small percentile error.

From the results in O'Kane and Treves (1992a) it seems that the *local field*, that is the signals arriving to a neuron from the cells of the module it belongs to, competes with, and in some cases dominates over, the long-range component of the total neuron input. Thus the local module net cannot correctly discriminate between global states.

In the next chapter we will modify the network presented in the previous section with the aim, above all, of putting the memory-glass at a disadvantage in comparison with the retrieval states. The changes will also improve the biological plausibility of the model. Namely, making the net more similar to the biological system also improves its retrieval abilities. The main criticisms to the network described in O'Kane and Treves (1992a) from the point of view of the neurobiologist are:

- Modular activation: all the modules of the present model are involved in every pattern. Many studies, especially PET and similes, have shown that any pattern selectively elicits hot spots over the neocortical surface thus suggesting that only a (small) fraction of the modules take part in the pattern representation, the rest being in a quiescent state[5].

- Long-range projections: the model has been constructed distributing the axonal long-range projections at random among the modules. This is in contrast with some findings (e.g. by Braitenberg and Shutz (1991) in the rat) according to which every column interacts with a small number of other columns[6].

- Correlation of the activities: The model assumes that modules are activated independently one from the others across patterns. Implementing a modification to the model in agreement with the criticism exposed in the previous point, it becomes possible to consistently introduce correlation between the activation states of different modules following the plausible conjecture that two any modules are connected if they analize features that are significantly interdependent across the patterns statistics.

- Local connectivity: It is actually a minor item from a theoretical perspective, but in fact the local connectivity, though being much higher than the long-range one, is not complete. Thus the full intramodular connectivity used before

[5] To our knowledge, Lauro-Grotto hypothesized the reduction of the number of active modules in the model as a possible benefit to the network performance (Lauro-Grotto et al., 1997).

[6] In spite of the simplistic model described in Braitenberg (1978), long-range projections are known to be anisotropic, as stated in Braitenberg and Shutz (1991) pages 142 and 146.

should be substituted by a diluted one.

Chapter 3

A model for neocortical associative areas

A new modular network is proposed as a model of associative areas[1]. A qualitative description is followed by a mathematical analysis of the new model.

3.1 Two steps toward the suppression of memory glass

The first modification we operate on the model considered in O'Kane and Treves (1992a) is about the **average number of active modules** per stored pattern. In O'Kane and Treves (1992a) every pattern implies the activation of all modules, while, e.g. from PET studies, it is quite evident that each cognitive task increases the activity level of only a small subset of cortical hot spots (moduli?). The fact that each feature in any module is presented many times during learning implies that the modification of each local recurrent synapse, according to the covariance rule, occurs many times in the same direction; on the contrary, the long-range connections modify their own weights every time in a random direction, even during the presentation of a number of patterns that imply the activation of the same

[1] Most of the material in this chapter is taken from Fulvi Mari and Treves (1998).

feature in a module. It follows that the local component of the field perceived by a neuron is stronger than the long-range part to the extent the number of patterns that involve a feature in the module (=the multiplicity, μ) is large. This means that local currents come to dominate the dynamics of the single module as μ gets larger, thus giving memory-glass (MG) states almost the same stability of retrieval (R) states.

We introduce in the modified model the possibility for a module to be involved only in a small fraction (τ) of the memory patterns, its neurons being for the rest in a state of spontaneous firing. Now μ, still defined as the ratio between the number of stored global patterns and that of stored local features ($\mu = P/D$), no longer equals the number of patterns activating a given feature in a module. The latter quantity is now a binomial random variable with average value $\mu\tau$. This proviso improves the capacity of the whole network, but is not the resolutive change, since also the MG states are stable up to a higher limit, so that the ratio of the respective capacities stays unchanged in comparison with the model of O'Kane and Treves (1992a) (on a logarithmic scale, the distance between the curves remains the same, fig.3.1). However, note that in our model the modular sparseness τ allows to increase the ratio P/D, while retaining a moderate advantage of retrieval states over memory-glass states.

To pull down, as it were, the curve of the MG state, we may look for some modification that might help prevent the activation of incoherent combinations of features.

Examining the next point listed in section 2.4, we restrict the number of modules connected to a given one. To this end we introduce the communication **channels**. They may be seen as 'tubes' exiting a module, each one projecting onto a different module. The number of channels is a Poisson random variable with a small mean,

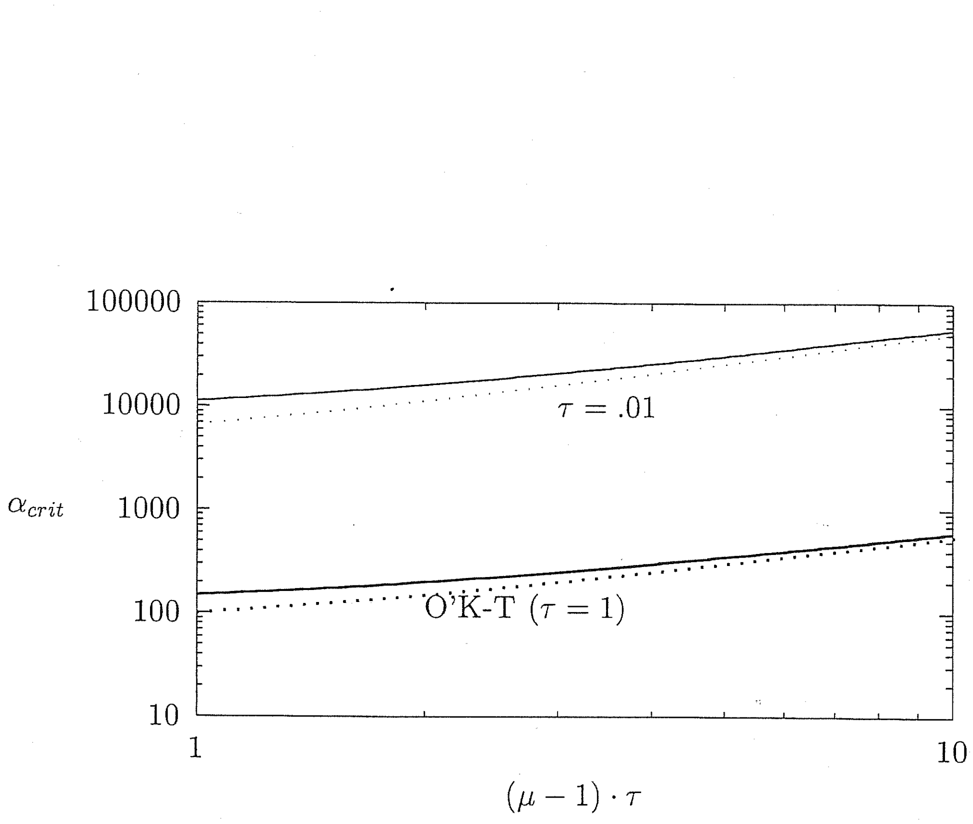

Figure 3.1: The retrieval and memory glass storage capacities vs feature multiplicity (α_{crit} is still defined as the ratio between the critical number of patterns that make the considered state unstable and the number $C = L + N - 1$).
From the top, the first two curves are respectively the retrieval and the memory glass critical capacities when $\tau = .01$. The other curves are the corresponding ones in the model described in O'Kane and Treves (1992a) (the new model reduces to that one if one sets $\tau = 1$). The improvement is evident, especially if one notes that for each value of μ the new model is positioned much more to the left (by a factor τ) than the old one. (The local connectivity parameter b, to be introduced later, is here equal to 1, that means full local connectivity.)

called s', that is supposed to be a small number, e.g. 2-7, in agreement with the neuroanatomical data (Braitenberg and Shutz, 1991). Any extra-modular projection of any module is now constrained to choose randomly one of the channels and to go through it. When the axon reaches the postsynaptic module, it can synapse onto any of the neurons therein. For simplicity, every channel is assumed to be *bi-directional*, that is: in the case of the channel that connects module m with module n, the axons of neurons in module m can pass through toward module n, and the axons from module n can pass through toward module m. Though the number of extra-modular fibres per neuron is untouched, now every module can 'see' only a few others. In fact, it would seem useless or even counterproductive to build up connections between modules that analyze independent features in natural patterns, so that these modules cannot exchange useful information, for instance, for a retrieval task, with each other.

We argue that the structure of the extra-modular connectivity is "semantically" meaningful: the fact that a module sends projections to another one underlies the fact that the aspects eliciting the two modules are correlated (e.g. semantically). Consequently the activation of the first module makes the probability of involvement of the second in the same pattern higher than chance (in the statistics of the patterns). We implement this idea in the model by creating a table of marginal conditional probabilities on the statistics of the patterns, given the structure of "communication channels" among modules. These probabilities are such as to favour (in comparison to chance) the activation of a module if another module, presynaptic to the former, is active. Analogously, if the presynaptic module is not active, then the postsynaptic is not active with larger probability than chance. Since the probability scheme introduced here is marginal, it does not give immediately a picture of the global statistics of the patterns of activity. To the verification of the self-consistency of the probability scheme and of the available ranges of parameters, we dedicate the rather technical Chapter 4.

Finally, we introduce the local connectivity dilution. This further improves the advantage of retrieval states over memory glass, since local field is weakened and long-range signals are thus more effective.

3.2 The mathematical model

We consider a neural network composed of a large number M of modules, each made up of a large number N of identical neurons. Inside any module the connection probability is finite, while the extramodular average connectivity is very diluted. In any module (m) of the network, the state (output) of any single neuron (i_m)

is assumed to be represented by a real non-negative number (V_{i_m}). Each neuron receives a projection from any other neuron of its same module with finite probability b, and from a small subset of the neurons belonging to the other modules (for a total L of extra-modular inputs, with L/N finite in the thermodynamic limit). It computes, independently from the other neurons, the sum of the afferent currents produced by its synapses. The total afferent current into a neuron is computed as the weighted sum of the outputs of the presynaptic neurons, where the weights are given by the (real-valued) synaptic strengths ($J^S_{i_m j_m}$ for intramodular connections and $J^L_{i_m j_n}$ for extramodular projections). The long-range (extra-modular) and short-range (intra-modular) contributions are summed as in the following formula, that defines the total input current (*field*):

$$h_{i_m} = \sum_{j_m \neq i_m} J^S_{i_m j_m} V_{j_m} + \sum_{n \neq m} \sum_{j_n} J^L_{i_m j_n} V_{j_n} + I_{i_m} \tag{3.1}$$

where the index n runs over the total number (M) of modules except the module of the neuron considered (m). I_{i_m} represents the non-specific inhibitory contribution whose role is to keep the network at an average activity level below the saturation; this inhibition depends on the average neuronal activity. Synaptic strengths are supposed to be determined by hebbian (covariance) learning of a large discrete set of quenched patterns. The quenched values of the synapses are set as

$$J^S_{i_m j_m} = \frac{1}{C} b_{i_m j_m} \sum_p \tau^p_m \left(\frac{\eta^p_{i_m}}{a} - 1\right)\left(\frac{\eta^p_{j_m}}{a} - 1\right) \tag{3.2}$$

and

$$J^L_{i_m j_n} = \frac{1}{\tau C} c_{i_m j_n} \sum_p \tau^p_m \tau^p_n \left(\frac{\eta^p_{i_m}}{a} - 1\right)\left(\frac{\eta^p_{j_n}}{a} - 1\right) \tag{3.3}$$

where p runs over the P patterns; $\{\eta^p_{i_m}\}$ are independent random variables representing the activities of the neurons during learning of pattern p (for the sake of simplicity they are taken in the following to be just binary, i.e. to assume only the value 0 or 1); a is the *sparseness* of the representation, that is to say, the probability for a $\eta^p_{i_m}$ to be equal to 1. The binary quenched random variable τ^p_m, which assumes value 1 with probability τ and 0 with probability $1-\tau$, models the activity state of module m during realization of pattern p: if the pattern in consideration requires the activation of a feature in that module, then $\tau^p_m = 1$, otherwise $\tau^p_m = 0$.

As can be easily recognized, the expressions 3.2 and 3.3 for the synapses are a generalization to a modular neural network of the well known formula for a network with sparse neuronal activity (Tsodyks and Feigel'man, 1988). In our model, we have added sparseness in the patterns of activation of the modules during global processing. The synapse between two neurons belonging to the same module can be modified only if this module is involved by the processes developed in the global net, e.g. following a sensory input. If, on the contrary, the presynaptic and post synaptic neurons belong to two different modules, the necessary condition to allow for modification of the synapse is that both modules are involved in the global processing.

The quenched random variable $b_{i_m j_m}$ determines the dilution of the connectivity inside a single module; its value is 1 with probability b, and 0 with probability $1-b$. That is, the probability for a neuron to receive axonal projection from any other neuron of the same module is taken equal to b.

The factor $\frac{1}{\tau}$ appearing in eq.3.3 ensures the correct normalization. When the contributions to the input of a unit from the other modules are summed, on average only a fraction τ of those terms are non-zero. The omission of the renormalizing factor $\frac{1}{\tau}$ would make the remote component of the field negligible compared to the local one when τ is much smaller than one (cf. section 3.7).

We assume solutions of the network equations according to which any module is either retrieving a feature ($\tau^p_m = 1$) or remaining in a state of spontaneous activity

$(\tau^p_m = 0)$. We also assume that the average activity in any module is of the same order during retrieval as in the spontaneous firing state; the firing rate of any neuron in a module in the spontaneous activity state is lower than that of a typical excited neuron, and higher than that of a quiescent neuron, in a retrieving module. The single unit activity level has then a uniform distribution over the neurons in a passive module, while the distribution in a retrieving module is bimodal (assuming binary patterns), with the same average.

The quenched random variable $c_{i_m j_n}$ assumes value 1 if neuron j_n in module n projects onto the dendrites of neuron i_m of module m; otherwise it is equal to zero. The quantities $\{c_{i_m j_n}\}$ hold the information about the anatomical structure of long-range connectivity of the considered multimodular net. We factorize

$$c_{i_m j_n} = s_{mn} \cdot k_{i_m j_n}. \tag{3.4}$$

The structure variable s_{mn} is equal to 1 if there exists a *communication channel* from module n to module m. With this we intend to model a functional structure such that axonal projections, from a module to another, exist only if these two modules are one after the other in an anatomical tree of cortical modules. The probability of the event $s_{mn} = 1$ is equal to $\frac{s'}{M}$, so that any module receives projections from a finite number (about s') of other modules, even in the thermodynamic limit; this is a model of the experimental observations described in Braitenberg and Shutz (1991). The variable $k_{i_m j_n}$ assumes value 1 with probability equal to $\frac{L}{s'N}$; otherwise it is zero. Hence the value of $k_{i_m j_n}$, given the existence of a communication channel from module n to module m $(s_{mn} = 1)$, reports the existence, or absence, of an axon from neuron j_n that synapses onto neuron i_m.

The output of the neuron receiving the total current h_{i_m} depends on the term $h_{i_m} - \theta_0$, where θ_0 is the net contribution of spontaneous activity and real thresholds coming from the non-linear spike dynamics. We do not specify a particular transfer function, but assume some of its basic properties, which become clear in the following.

The activity of neuron i_m inside the module m, involved in the correct retrieval of the local feature corresponding to the global pattern p, is proportional to $\eta^p_{i_m}$. Since we assume that any global pattern p has probability $\frac{\mu}{P} = \frac{1}{D}$ to elicit a given feature in a module (different patterns are independently assigned the corresponding features with a uniform distribution), it is necessary to introduce the functions

$$d_n : [1, 2, ..., P] \longrightarrow [1, 2, ..., D] \tag{3.5}$$

defining the quenched correspondence of a global pattern p to the feature $d_n(p)$ elicited in module n in the global activation of pattern p. So there are on average μ global patterns involving the retrieval of each feature d_m. For example, $\eta^{d_m}_{i_m}$ refers to the activity of neuron i_m in module m during the presentation of one of the (about μ) global patterns ($\{p_1, p_2, ..., p_\mu\}$) that involve the same feature d_m in module m ($d_m(p_\alpha) = d_m$ for $\alpha = 1, 2, ..., \mu$). It will be necessary in the following calculations to keep in mind that the probability for two patterns to elicit the same feature in a given module is vanishingly small ($\frac{1}{D}$) in the thermodynamic limit, in which we take $D \longrightarrow \infty$. Analogously, the probability that two patterns eliciting the same feature in a module involve the same feature also in another module is $\frac{1}{D}$ (these observations are useful in the calculation of some correlation functions).

Note that we are not bound to binary neurons: the real activities during retrieval are in our model only proportional to the binary pattern, but it has been already noticed (e.g. Treves (1990)) that the introduction of more graded distributions of activity levels does not bring substantial changes.

3.3 Parameters of the model

Beside μ, already defined, we make use of other network parameters indicating finite ratios of quantities that diverge in the thermodynamic limit. They are the *storage load* $\alpha = \frac{P}{C}$ and the fraction of long-range afferent projections onto any neuron $\gamma = \frac{L}{C}$, being $C = N - 1 + L$. (The average number of inputs per neuron is $b(N-1) + L$, and reduces to C when full local connectivity, $b = 1$, is considered.) The capacity limits cited in the first paragraph are actually espressed through α, being α_c the maximal value of α for which the system is in a stable state, that can be either a global *retrieval* state or a *memory glass*. For example, α_c^{MG} is the value of the ratio $\alpha = \frac{P}{C}$ over which the system cannot persist in a memory-glass state, and decays in a retrieval state or in a uniformly disordered (paramagnetic) state. Analogously, if $\frac{P}{C} > \alpha_c^R$ the system can not sustain a state of retrieval, and falls in the disordered state (see below). Given α_c, the maximum number of patterns storable in the network, above which the state considered (retrieval or memory-glass) ceases to exist, is then given by $P_c = \alpha_c \cdot C$. So the largest number of patterns that can be stored and then succesfully retrieved is proportional to the average number of synaptic contacts onto any neuron ($P_c^R = \alpha_c^R \cdot C$).

The parameter γ measures the fraction of connections that are long-range. If γ is very small, the network looks very similar to a set of M independent fully connected neuronal networks, that is, the global associative properties are negligible and the capacity of the large network coincides with that of single module. On the contrary, if γ is very near to 1, then the long-range interactions dominate over the local ones, and the large network behaves as a diluted associative network in which the molteplicity μ of any pattern becomes irrelevant. A model that describes neocortex is likely to fall somewhere in the middle of the range, since it is estimated that about one half of the projections that any neocortical pyramidal neuron receives come from outside the module the neuron belongs to, that is from the white matter.

3.4 The correlation of modular activity

The second main point of this model beside $\tau < 1$ is the introduction of the correlation between the set of activity patterns $\{\tau_m^p\}$ and that of communication channels $\{s_{mn}\}$, whose physical meaning has been described above. We require that when module n sends projections to module m and module n is active in a pattern, then in the same pattern module m is active with a probability higher than the unconditional one. Similarly, when module n is not active, module m is active with a probability smaller than chance. This preference of a postsynaptic module to be in the same state of the presynaptic module it is connected to is translated in the following table of conditional probabilities:

$$\begin{array}{l} \mathcal{P}(\tau_m^p = 1|\tau_n^p = 1, s_{mn} = 1) = t_1 \\ \mathcal{P}(\tau_m^p = 1|\tau_n^p = 0, s_{mn} = 1) = t_0 \\ \mathcal{P}(\tau_m^p = 1|\tau_n^p = 1, s_{mn} = 0) = \tau \\ \mathcal{P}(\tau_m^p = 1|\tau_n^p = 0, s_{mn} = 0) = \tau \end{array} \tag{3.6}$$

where $t_1 > \tau$, while $t_0 < \tau$ (τ is the unconditional probability for a module to be active, introduced above). So the activity of module m does not depend on that of module n if the latter does not project axons onto the former.

If the structure quenched variables are taken to be symmetric ($s_{mn} = s_{nm}$), the table above is a consistent probability scheme only if

$$(1 - t_1) \cdot \tau = t_0 \cdot (1 - \tau). \tag{3.7}$$

It can be shown that this equality implies that

$$\mathcal{P}(s_{mn} = 1|\tau_m^p = 1) = \mathcal{P}(s_{mn} = 1) = \frac{s'}{M}. \tag{3.8}$$

If we eliminate the reciprocity of the structure variables, the condition of eq. 3:7 is no longer necessary. This corresponds to the possibility of having

$$\mathcal{P}(\tau_m^p = 1 | s_{mn} = 1) \neq \mathcal{P}(\tau_n^p = 1 | s_{mn} = 1) = \tau. \tag{3.9}$$

We assume eq. 3.7 to be valid, and this implies that, averaging over the possibility to be active or not, a module receives non-zero contributions from a fraction τ of the modules it 'sees'. (For more about this point, cf. Chap. 4.)

During correct *retrieval* a fraction τ of the modules are active and reproduce the features corresponding to the global pattern retrieved by the whole large network. The remaining modules, inactive, present a uniform distribution of spontaneous firing. We mathematically represent this state by the expression of the neuronal output

$$V_{i_m} = B' \cdot \tau_m^p \cdot \eta_{i_m}^{d_m(p)} + A' \cdot (1 - \tau_m^p), \tag{3.10}$$

that is, in any module m that is active in pattern p ($\tau_m^p = 1$) the neuronal potentials follow a retrieval distribution given by the first term on the right-hand side of eq. 3.10. If module m is not activated by pattern p ($\tau_m^p = 0$), only the second term survives and the neurons of that module are in a spontaneous firing state. The real coefficients A' and B' should be set by solving the equations of the network, and p is the index of the global pattern that is retrieved. Actually it is not necessary to determine exactly the two coefficients, since in the final expression of the signal-to-noise ratio we are interested only in the ratio $\frac{A'}{B'}$, and this can be estimated to be about a from the sparseness requirement mentioned above.

In a *memory glass* state the fraction of active modules is arbitrary, and also the statistics introduced with eqs.3.6 is not considered to apply. It is easy to show (cf. appendix to this chapter) that the most robust memory-glass states (i.e. the ones with the highest capacity limit), even though less numerous, are those with a low fraction of active moduli and without correlations between activity and connectivity. The neuronal activity in a memory-glass state is expressed again as in eq. 3.10, with coefficients A and B that could be different from A' and B' of the retrieval case, but are assumed to keep $\frac{A}{B} = a$ all the same. Now no global pattern is reproduced by the network, since each of the many groups of moduli retrieves a fragment of a different pattern.

3.5 Signal-to-Noise analysis

We study the stability of the states of retrieval and memory glass with a simple signal-to-noise analysis. Considering the field h_{i_m} onto neuron i_m, as usual we isolate the specific term that tends to maintain the state stable, and call it *signal*, $\mathcal{S}$. The rest of the field is a combination of interference terms coming from the other patterns stored in the same synapses, that are the hebbian terms that have influenced the final values of single synapses. These terms are actually quenched random variables (close to normally distributed), whose values vary from a neuron to another, from a module to another, and changing the pattern (or feature) retrieved. The set of these interfering terms gives a net zero-mean contribution, and we call the total standard deviation *noise* $(\mathcal{N})$. There are three values that $\mathcal{S}$ may assume, corrisponding respectively to high, spontaneous, and suppressed firing activity of the neuron. In order to maintain the state stable, the noise has to be, roughly speaking, less than the distance between the signal corresponding to the high firing and that corresponding to suppressed firing.

We analize separately the retrieval states and the memory glass states. We do not report the calculations, that mainly involve handling sums and products of random variables and applying asympthotic theorems of probability theory. We show only the results, i.e. the conditions under which the states are stable; from these conditions, it is straightforward to extract the critical storage values in the two states studied (cf. Appendix about the memory glass capacity):

$$\begin{aligned} \left(\frac{S}{N}\right)_R &\sim \frac{b(1-\gamma)[1+(\mu-1)\tau]+\gamma\frac{t_1}{\tau}}{\sqrt{a\tau\alpha}\sqrt{b(1-\gamma)[1+(\mu-1)\tau]+\gamma\frac{t_1}{\tau}[\frac{a}{\tau}(1-t_1)+\frac{t_1}{\tau}]}} \\ \left(\frac{S}{N}\right)_{MG} &\sim \frac{b(1-\gamma)[1+(\mu-1)\tau]}{\sqrt{a\tau\alpha}\sqrt{b(1-\gamma)[1+(\mu-1)\tau]+\gamma\frac{t_1}{\tau}\frac{a}{\tau}}} \end{aligned} \tag{3.11}$$

where we have used the relation $\frac{A}{B} \sim a$.

As can be seen, our modifications act on both the components of the signal (intramodular and extramodular). In the case of the 'local' signal, we reduce the number of patterns that map on the same feature in a module by making modular activity sparser, and dilute the local connectivity. Thus, the local signal no longer dominates over the 'remote' signal: now the former is proportional to $[1 + (\mu - 1)\tau] \cdot b$, while in the old model (i.e. $\tau = 1$) this factor was μ. The ratio $\frac{P}{D}$ is however unchanged (equal to μ in both models).

The introduction of the activity correlation acts on the remote component of the field, bringing a factor $\frac{t_1}{\tau}$ instead of 1. Thus, strengthening the long-range interaction, the global pattern is favored. When the quantity $\frac{t_1}{\tau}$ becomes large, the capacity of memory-glass states is decreased, while the capacity of retrieval states is slightly modified.

3.6 Results

The critical storage load for the two states studied can be derived from espressions 3.11, and is shown in figure 3.2 vs the feature multiplicity μ. Following Braitenberg and Shutz (1991), we put $\gamma = .5$; the neuronal sparseness a and the modular sparseness τ have been taken both equal to .01, which seems in the order of what the little experimental evidence available would estimate. Figure 3.1 shows the improvement

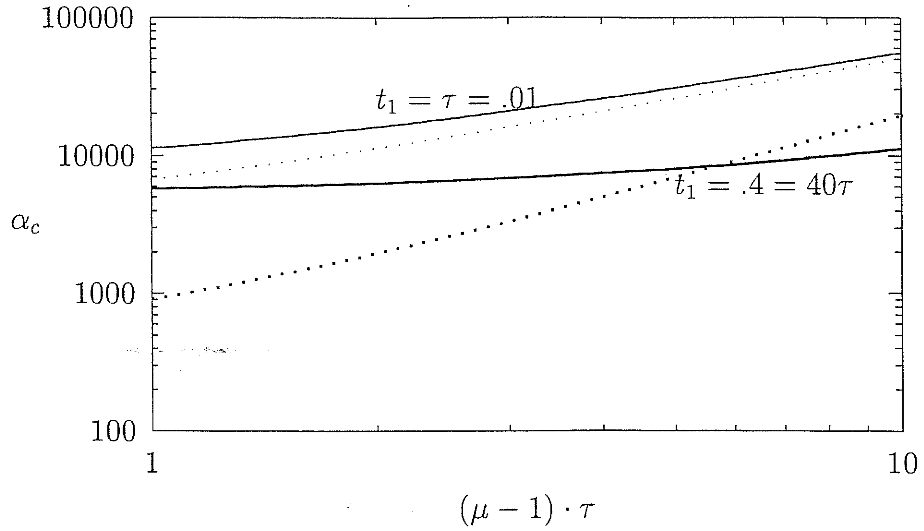

Figure 3.2: The critical capacities for two values of t_1. From the top: retrieval storage capacity and MG storage capacity for $t_1 = \tau = .01$; retrieval storage capacity and MG storage capacity for $t_1 = 40\tau = .4$. ($b = 1$ in both cases.) Note that the first two curves are the same as in fig.3.1.

obtained with the introduction of only modular sparseness, $\tau < 1$, keeping still the complete independence of patterns from long-range structure and full local connectivity. Note that the abscissa is not the same for the old and new model, since in the old one, one has to consider that $\tau = 1$; this gives our model an advantage over the old one, as explained in section 3.1.

It is possible to show that t_1 can assume values up to an upper bound (< 1) that depends on s' and τ. Beyond such a bound, the activity statistics here considered is no longer consistent. In the next, more mathematical chapter we report the theoretical and numerical study we have developed about this bound and other properties of the statistics considered. We anticipate that the study justifies values of $\frac{t_1}{\tau}$ up to 40 in the fairly plausible case of $s' = 4$ and $\tau = .01$.

Figure 3.2 shows the memory-glass suppression operated by the introduction of also the pattern-structure correlation, with $\frac{t_1}{\tau} = 40$. A crossover can be noted between the two capacities in the correlated case when the abscissa becomes large.

It is due to the fact that the most dangerous memory-glass states have a much smaller number of active modules than the retrieval states.

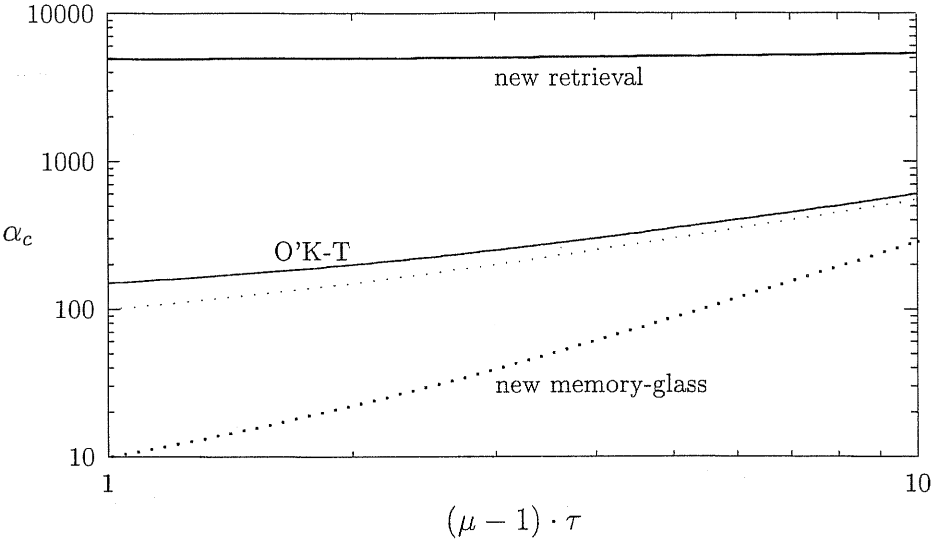

Figure 3.3: Comparison between the critical capacity of the original model in O'Kane and Treves (1992a) and the present model with $\tau = .01$, $t_1 = .4$, $b = .1$. The capacity of retrieval states has been largely increased while the memory-glass states have been strongly suppressed.

When also the local dilution is introduced ($b < 1$), the critical capacities of the new model become as plotted in fig. 3.3, where they can be directly compared with the corresponding ones of the original model of O'Kane and Treves (1992a).

To stress the importance of the modular activity sparseness and correlation, in fig. 3.4 we compare the performance of the new model without sparseness and correlation, that is the model of O'Kane and Treves (1992a) added of only local connectivity dilution, with the new model of fig. 3.3. Note the marked improvement given by the introduction of sparseness ($\tau < 1$) and correlation ($t_1 > \tau$), considering also that in the case without sparseness the label of the abscissae axis must be read as '$\mu - 1$'.

Finally, in figure 3.5 we compare the model with sparseness, correlation and local dilution with the model in which the correlation is not included ($t_1 = \tau$).

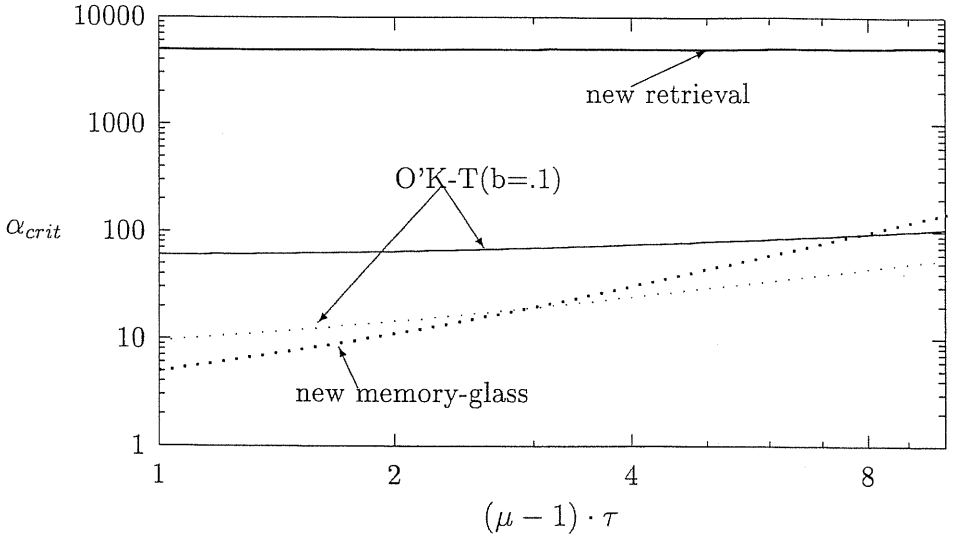

Figure 3.4: Comparison between the critical capacities of the new model with τ = .01, t_1 = .4, b = .1 and those of the same model deprived of the modular activity interdependence and sparseness (that is, $t_1 = \tau = 1$), here called 'O'K-T(b=.1)'.

In conclusion, introducing three modifications in the model considered by O'Kane and Treves (1992a) we have been able to increase the storage load limit of retrieval states and to strongly destabilize the memory-glass states that marred the memory operation of the original model. The modifications, a sparse rather than complete activation of cortical modules, the correlation between the patterns of activation and the underlying connectivity, and the dilute local connectivity are consistent with available evidence. Once the modular activity sparseness is introduced, the improvements separately provided by correlation and local dilution are of comparable magnitude, and together co-operate in strongly suppressing the noxious states while leaving the correct retrieval states almost unaltered. The larger the ratio $\frac{t_1}{\tau}$, the better the suppression of the undesired states. Thus, the question arises as to whether the correlated statistics is consistent, and, if so, how large $\frac{t_1}{\tau}$ can be.

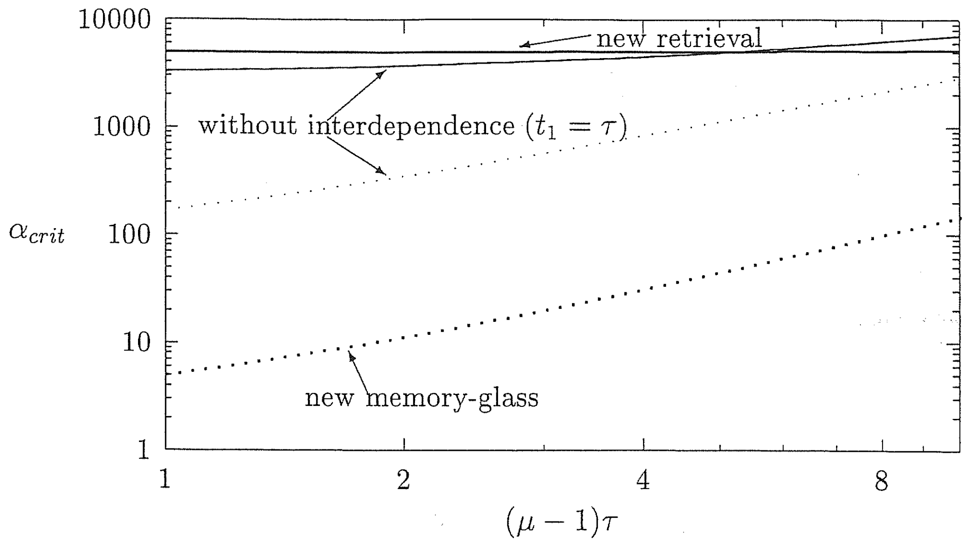

Figure 3.5: Comparison between the critical storage of the new model with $\tau = .01$, $t_1 = .4$, $b = .1$ and those of the model with modular activity sparseness and local dilution only (that is, $t_1 = \tau = .01$ and $b = .1$).

3.7 Comments on the synaptic rule

According to the synaptic rules of the model (eqs. 3.2 and 3.3), the increment of the weight due to one learning step (=1 pattern) must be larger in the long-range synapses than in the others, because the event of a synaptic contribution is rarer due to the product $\tau_m \tau_n$ that appears in eq.3.3. Indeed, we have assumed that the synapses of long-range axons are modified only when both the presynaptic and the postsynaptic modules are active, and this event is obviously rarer than the activation of any single module. The deletion of the factor $\frac{1}{\tau}$, responsible of the differentiation for the synapse learning speeds, allows the local currents to dominate, with the consequent reinforcement of memory-glass state. How large actually the compensation factor must be is not clear. We have chosen the factor $1/\tau$ before introducing correlation. One may object that after having introduced the correlation it should have been changed in t_1. However this scaling is not equivalent to the introduction of the correlation, since a simple constant factor would be linear in the long-range signal term (eq.3.11) and would be squared in the long-range noise variance, under the square root of the denominator. On the contrary, the factor arising from the correlation is linear both in the long-range signal and in the long-range noise variance. This allows a marked improvement of the signal-to-noise ratio that could not be obtained with a multiplicative constant in front of the long-range synaptic weight (eq.3.3).

Appendix: Memory glass storage capacity

Since the memory-glass states do not reproduce global patterns, also the modular activity statistics may be different from that imposed by the set of learned global patterns. Accordingly, the signal-to-noise ratio in memory-glass states can be shown to be

$$\left(\frac{S}{N}\right)_R \sim \frac{b(1-\gamma)[1+(\mu-1)\tau]}{\sqrt{a\tau\alpha}\sqrt{b(1-\gamma)[1+(\mu-1)\tau]+\gamma\frac{t_1}{\tau}[\frac{a}{\tau}(1-\tilde{\tau})+\frac{\tilde{\tau}}{\tau}]}} \tag{3.12}$$

where $\tilde{\tau}$ is the probability that any module is active when the network is in a memory-glass state. It is easy to show that when a and τ are similar in magnitude and very small, the memory-glass states with the highest signal-to-noise ratio are those in which few modules are active, that is ideally $\tilde{\tau} = 0$. Thus, putting $\tilde{\tau} = 0$ in the previous equation, the second espression in eq. 3.11 is obtained. This condition may be relaxed if one assumes that external inputs to the network have a finite fraction $\tilde{\tau}$ of active modules. However, we continue to deal with the worst case of $\tilde{\tau} = 0$.

Chapter 4

Bounds to the correlation between modules

A class of families of marginal probabilities on sets of discrete random variables is studied and a necessary and sufficient condition for the consistency of the given marginals is provided. Use is made of theorems from Markov Random Fields and from Probability Distributions with Given Marginals.

Then, the specific consistency problem from the theory on neocortical modular activity described in the previous chapter is faced. The above-mentioned condition for consistency of marginals provides a method for the verification of the hypotheses assumed by that model, according to which connected modules of neurons are simultaneously active with probability higher than chance, and inter-modular connections are very diluted. The results of the verification strongly support the neuroscientific theory[1]

4.1 Introduction and summary

The probability distribution of a proper subset from the set of random variables of a stochastic system is usually called the *marginal probability* or simply the *marginal*

[1] The material for this chapter is taken from Fulvi Mari (1998).

and from elementary probability it can be obtained by integration of the joint probability distribution function over the values of the RVs that are not involved in the marginal to be calculated.

When the complete p.d.f. is not known and use is made of an Ansatz for its marginals, the problem of verifying the self-consistency of such marginals arises. In other words, it is necessary to demonstrate the existence of at least one joint p.d.f. whose marginals coincide with the conjectured ones. Otherwise, the probabilistic model would be unrealizable: no set of RVs that follow the required statistics could be defined. In the case of a physical theory, such inconsistency would mean that no real system could be described by that theory (as long as Probability Theory is able to model physical systems).

The problem of the consistency of a set of marginals belongs to the mathematical theory of Probability Distributions with Given Marginals. Some works about this subject can be found in the mathematics literature, the most remarkable ones being, maybe, the papers by Vorobev (Vorobev, 1962) in a combinatorial context, and by Kellerer (Kellerer, 1964) in a more functional measure-theoretic fashion.

In the present chapter I face the problem of marginal consistency over sets of discrete RVs whose correlations are randomly generated, so that the system may be represented as a random graph whose nodes can be each in one of two possible states.

In section 4.2, I propose a general systematic method for verifying the consistency of a family of marginals of a certain class, defined over stochastic systems of discrete RVs, making use of results from the theory of Markov Random Fields (MRFs) and of Vorobev's theorem.

From section 4.3 to 4.9, I deal with a specific marginal problem to which I partially apply the method. The neuroscientific origin of this problem is given by the analytical neuronal model of the previous chapter.

In section 4.3, I define the problem in terms of a probabilistic model, putting in evidence the underlying question of probability distribution with given marginals.

In section 4.4 I show the inapplicability of the already existing theorems on consistency by Vorobev and Kellerer.

In section 4.5, I derive some conditions on the parameters of the model from an assumption of self-averaging.

In section 4.6, approximating the method introduced in section 4.2 through an Ansatz, I create a fictitious stochastic dynamics on the configuration space of the system, with the aim of obtaining an asymptotic distribution of the activities that satisfies the marginals and which, at the same time, can be studied through numerical simulations.

In section 4.7 an analytical approach to the dynamical system of section 4.6 is exposed. In fact, this analytical study of the asymptotic distribution allows the verification of the consistency of the given marginals only in a particular limit, due to mathematical difficulties in 'mere' calculations connected with the non-mean-field behavior of the statistical system.

In section 4.8 an analogy with diluted Ising ferromagnets is proposed.

Then, in section 4.9 I describe and discuss the results of the numerical simulations, that are in favor of the hypothesis of consistency of the given family of marginals thus supporting the underlying neuroscientific theory.

In section 4.10, I sum up the conclusions and some directions for ongoing investigations.

4.2 Mathematical basis of the method

Let $\mathcal{S}$ be a finite set of n discrete random variables, and $\mathcal{M}$ a set of marginals such that any marginal is defined either over a couple of RVs or on a single RV from $\mathcal{S}$. For any marginal $\mathcal{P}_A \in \mathcal{M}$ over a RV or a couple of RVs, called A, in $\mathcal{S}$, I call A the *support* of $\mathcal{P}_A$. The main hypothesis is that the marginals of all the possible couples in $\mathcal{S}$ are present in $\mathcal{M}$.

Compatibility: The marginals in $\mathcal{M}$ have to be compatible, that is the marginalization of every couple of them on the intersection of their supports must coincide. Otherwise, $\mathcal{M}$ is a set of inconsistent marginals. For example, let $F(x_1, x_2)$ and $G(x_2, x_3)$ be two of the marginals in $\mathcal{M}$; then it is necessary that

$$\sum_{x_1} F(x_1, x_2) = \sum_{x_3} G(x_2, x_3)$$

for any value of x_2. In the rest I assume the compatibility of the marginals in $\mathcal{M}$.

Construction of the MRF: Given a variable $x_i \in \mathcal{S}$, one may consider all the couple marginals in $\mathcal{M}$ whose supports contain x_i. This subfamily is denoted $\mathcal{N}_i$, and Y_i is the set of RVs that belong to the supports of $\mathcal{N}_i$. The RVs in $Y_i - \{x_i\}$ are the *neighbors* of x_i from the point of view of the MRF theory. The marginals in $\mathcal{N}_i$ are compatible and their supports constitute a regular complex (Aleksandrov, 1956). Thus Vorobev's theorem (Vorobev, 1962) can be applied: *At least one joint probability distribution of all the variables in Y_i exists such that its marginals over the supports of $\mathcal{N}_i$ coincide with the given marginals in $\mathcal{N}_i$.* Any distribution with this property is called an *extension* of $\mathcal{N}_i$.

Usually there are several extensions. Suppose to have two extensions $\mathcal{P}^{(1)}_{\mathcal{N}_i}$ and $\mathcal{P}^{(2)}_{\mathcal{N}_i}$ of $\mathcal{N}_i$. Then for any $x_m \in Y_i - \{x_i\}$:

$$\sum_{x'} \mathcal{P}^{(1)}_{\mathcal{N}_i}(x_i, x_m, x') = \sum_{x'} \mathcal{P}^{(2)}_{\mathcal{N}_i}(x_i, x_m, x'), \qquad (4.1)$$

where x' is a collective symbol for $\{x_j \in Y_i - \{x_i, x_m\}\}$ and the sums are over the values they can assume. It follows that the two local distributions can differ only by a function $F_i = \mathcal{P}^{(1)}_{\mathcal{N}_i} - \mathcal{P}^{(2)}_{\mathcal{N}_i}$ such that

$$\sum_{x'} F_i(x_i, x_m, x') \equiv 0 \tag{4.2}$$

for every RV $x_m \in Y_i - \{x_i\}$. Thus, knowing one particular extension $\mathcal{P}_i$ of $\mathcal{N}_i$, all the other extensions of $\mathcal{N}_i$ can be obtained from $\mathcal{P}_i$ by adding functions F_i defined on Y_i satisfying the property of eq.4.2.

Considering only binary RVs ($x_k = 0, 1$), the function F_i can always be written as

$$F_i(x) = \sum_{\{m,n \neq i\}} (1 - 2x_m)(1 - 2x_n)\psi_{mn}(x) \tag{4.3}$$

where x represents all the variables in Y_i, the sum is over the indices of the RVs in $Y_i - \{x_i\}$, and ψ_{mn} may depend on all the RVs in $Y_i - \{x_i, x_m\}$. Thus one can choose a convenient extension of $\mathcal{N}_i$ and then add to it the generic function in eq.4.3. As a convenient solution I propose the conditional independency distribution

$$\mathcal{P}^{(0)}_i(x) = \mathcal{P}(x_i) \prod_A \mathcal{P}_A(x_A | x_i), \tag{4.4}$$

where the product is over all the supports of $\mathcal{N}_i$, $\mathcal{P}_A$ is derived from the marginal in $\mathcal{N}_i$ with support A, and x_A is the other RV that belongs to the support A together with x_i. The distribution in eq.4.4 has the property that its marginalization over any variable in $Y_i - \{x_i\}$ is still a conditional (with respect to x_i) independency distribution.

Thus, the most general joint p.d.f. on binary RVs in Y_i that satisfies the marginals in $\mathcal{N}_i$ can be written as

$$\mathcal{P}_i(x_{Y_i}) = \mathcal{P}^{(0)}_i(x_{Y_i}) + \sum_{\{m,n \neq i\}} (1 - 2x_m)(1 - 2x_n)\psi_{mn}(x_{Y_i - \{x_m, x_n\}}), \tag{4.5}$$

where x_R is the collective symbol for all the RVs in the set $R \subseteq \mathcal{S}$. Since ψ_{mn} additively modifies the probabilities of the events related to Y_i, it cannot be completely arbitrary, so that its values can move only in bounded intervals.

For simplicity I indicate the local solutions by Ψ_i only, implicitly keeping the arbitrariness, provided by eq.4.5, that represents degrees of freedom possibly useful to the solution of the marginal problem.

From the parametrized joint distribution Ψ_i, it is easy to extract the conditional probability $\Phi_i(x_i|\{x_k \in Y_i, k \neq i\})$. (For simplicity, I assume that no event has null probability.)

Performing this construction for every variable x_m in $\mathcal{S}$, one obtains the field $\{\Phi_i, i = 1, \ldots, n\}$ candidate to be a MRF.

Derivation of the Gibbs state: From the theorem of equivalence between MRFs and Gibbs states (Kemeny et al., 1976; Hammersley and Clifford, 1971) it follows that, given an MRF, a Gibbs field that generates the MRF exists with the canonical neighbor potential

$$\begin{aligned} V_X(\eta) &= \sum_{U \subseteq X} (-1)^{|X-U|} \ln \Phi_i(\eta^U) \quad \forall i \in X \in \mathcal{C} \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \forall X \notin \mathcal{C}, \end{aligned} \tag{4.6}$$

where η is an event (that is, a realization of the RVs in X), $\mathcal{C}$ is the set of all the cliques of the MRF, U is a subset of the clique X, η^U is the vector such that $\eta^U_k = \eta_k$ when $x_k \in U$ and $\eta^U_k = 0$ otherwise, and i indicates any variable (=node) in X.

Thus the Gibbsian distribution $\mathcal{G}_S$ can be constructed[2], whose marginals possibly generate the MRF.

[2] The hamiltonian of the Gibbs distribution is

$$H(\omega) = \frac{1}{\beta} \sum_Q V_Q(\omega)$$

where Q is any subset of $\mathcal{S}$, ω is any configuration of the system, and $\frac{1}{\beta}$ is the 'temperature'.

Verification: At this point; possibly making use of the above-mentioned degrees of freedom, one has to verify whether all the marginals in $\mathcal{M}$ can be obtained marginalizing $\mathcal{G}_S$ (obviously, it is equivalent and more convenient to check the set $\{\Psi_j, j = 1, \ldots, n\}$). If they can, then the family $\mathcal{M}$ is consistent and $\mathcal{G}_S$ gives the solution (or the set of solutions, since there could remain some free parameters). If they cannot, then the given family $\mathcal{M}$ of marginals is inconsistent. In a single proposition:

The family $\mathcal{M}$ of compatible marginals is a consistent set of marginals if and only if the constructed Gibbs distribution $\mathcal{G}_S$ is compatible with all those marginals.

This verification rule might be formalized and shown in a more mathematical treatment, and its hypotheses a little relaxed. I will expose the technicalities and a more detailed description of the method in a forthcoming paper.

The only problem now is provided by the practical calculations, where in the case of large systems Statistical Mechanics of Gibbs ensembles may be of some help. Indeed, sometimes calculations are not easily performed, as in the case analyzed in the rest of this chapter, where the difficulty in managing with a large set of randomly correlated random variables does not allow one to apply the complete procedure presented in this section.

4.3 The randomly interacting system

In the rest of this chapter I define the neuroscientific problem of marginals in more general terms and find some results addressing its solution. In order to imagine the following in a neuronal context, the reader should identify the nodes of the random graph I will define with their activities as the modules and the modular activities respectively. The presence of a channel between modules m and n will be represented by the binary connection variable s_{mn} being equal to 1 (instead of 0).

Consider a set of binary RVs $\{\tau_m, m = 1, \ldots, M\}$, with large M; the configurations assumed by this set trial by trial are points in $\{0,1\}^M$. Let the average "activity" $< \frac{1}{M}\sum_m \tau_m >$ be equal to τ. (Pointed brackets indicate the ensemble average.)

The RVs $\{\tau_m\}$ are not independent. I consider the case in which the joint p.d.f. $\mathcal{P}(\{\tau_m\})$ is unknown and an Ansatz is proposed on the pairwise marginals. To simplify the formulation, it is useful to give an intuitive representation of the system from the beginning.

Consider a random graph $G(M, s)$, where s is the probability for any edge to be present; the adjacency matrix (s_{mn}) is symmetric, that is the graph is undirected, and there is no simple loop $(s_{mm} = 0)$. Each node represents one of the RVs; the presence of the edge indicates a marked dependence in the activities of the two nodes, while two non-adjacent variables are nearly independent (in this case the dependence in a couple is defined as the non factorizability of their marginal p.d.f.[3]). To be more precise, the marginal distribution of the activities in a randomly chosen pair of nodes is given by the following probability table:

$$\begin{aligned} \mathcal{P}(\tau_m = 1, \tau_n = 1 | s_{mn}) &= \tau t_1 s_{mn} + \tau^2 (1 - s_{mn}) \\ \mathcal{P}(\tau_m = 1, \tau_n = 0 | s_{mn}) &= (1-\tau) t_0 s_{mn} + \qquad (4.7) \\ &+ (1-\tau)\tau(1 - s_{mn}) \end{aligned}$$

where t_1 is a probability larger than τ and t_0 is smaller than τ. According to eqs.4.7, the probability that two connected nodes are in the same state of activity (0 or 1), averaged over the connected couples, is greater than chance (positive dependence). Actually, if one looks at the same pair (m, n) during trials, the probabilities of the

[3] This definition is convenient also from the neurophysiological point of view since in experiments usually only a small number of modules can be simultaneously monitored

activities of the two nodes are

$$\begin{aligned}\mathcal{P}(\tau_m = 1, \tau_n = 1|S) &= f^{11}_{mn}\tau t_1 s_{mn} + h^{11}_{mn}\tau^2(1 - s_{mn})\\ \mathcal{P}(\tau_m = 1, \tau_n = 0|S) &= f^{01}_{mn}(1-\tau)t_0 s_{mn} + h^{01}_{mn}(1-\tau)\tau(1 - s_{mn})\end{aligned} \tag{4.8}$$

where S represents the knowledge of the adjacency matrix, and $f^{11}_{mn}, h^{11}_{mn}, f^{01}_{mn}, h^{01}_{mn}$ are positive *structure factors* that take into account the fluctuations of the marginals across the node couples and depend only on the quenched structure of the graph. The structure factors are normalized as follows:

$$\begin{aligned}&\tfrac{2}{sM(M-1)}\textstyle\sum_{(m,n)} f^{11}_{mn}s_{mn} = 1,\\ &\tfrac{2}{sM(M-1)}\textstyle\sum_{(m,n)} f^{01}_{mn}s_{mn} = 1,\\ &\tfrac{2}{(1-s)M(M-1)}\textstyle\sum_{(m,n)} h^{11}_{mn}(1 - s_{mn}) = 1,\\ &\tfrac{2}{(1-s)M(M-1)}\textstyle\sum_{(m,n)} h^{01}_{mn}(1 - s_{mn}) = 1,\end{aligned} \tag{4.9}$$

where

$$s = \frac{2}{M(M-1)}\sum_{(m,n)} s_{mn}. \tag{4.10}$$

So, the average of eqs.4.8 over the connected and unconnected couples gives back eqs.4.7; for example,

$$\begin{aligned}&\tfrac{2}{sM(M-1)}\textstyle\sum_{(m,n)} s_{mn}\mathcal{P}(\tau_m = 1, \tau_n = 1|S) =\\ &= \mathcal{P}(\tau_m = 1, \tau_n = 1|s_{mn} = 1) = \tau t_1.\end{aligned} \tag{4.11}$$

I also assume that the marginal distribution of a single node does not depend on the structure of connections, that is

$$\mathcal{P}(\tau_m = 1|S) = \tau. \tag{4.12}$$

Indeed, I want to study the existence of distributions with marginals deviating from independence without introducing an *a priori* position-dependent distribution of single node activity.

From eqs.4.8 and 4.12:

$$\begin{aligned}\mathcal{P}(\tau_m = 1|S) = [f^{11}_{mn}\tau t_1 + f^{01}_{mn}(1-\tau)t_0]s_{mn}+\\ +[h^{11}_{mn}\tau^2 + h^{01}_{mn}(1-\tau)\tau](1-s_{mn}) = \tau.\end{aligned} \tag{4.13}$$

Then it must be that

$$\begin{aligned}f^{11}_{mn}\tau t_1 + f^{01}_{mn}(1-\tau)t_0 = \tau,\\ h^{11}_{mn}\tau + h^{01}_{mn}(1-\tau) = 1.\end{aligned} \tag{4.14}$$

Averaging the first one over the connected pairs, I obtain the necessary condition

$$(1-\tau)t_0 = \tau(1-t_1). \tag{4.15}$$

This may be seen as an "equilibrium" condition; infact, if one averages eqs.4.8 over the connected couples and over the unconnected ones respectively obtains:

$$\begin{aligned}\mathcal{P}(\tau_m = 1|s_{mn} = 1) = \tau t_1 + (1-\tau)t_0,\\ \mathcal{P}(\tau_m = 1|s_{mn} = 0) = \tau.\end{aligned} \tag{4.16}$$

Thus, the relation in eq.4.15 implies that the presence, or absence, of a connection between node m and any other node does not condition, by itself, the activity of node m (on average). This is consistent with eq.4.12.

Relation 4.15 may also be seen as a "detailed balance": if each node represents the state of a dynamical system and this can move only from one state to an adjacent one at each time step, then the probability of observing the evolution of the system from state m of "class 1" ($\tau_m = 1$) to state n of "class 0" ($\tau_n = 0$) is equal to the probability of observing the reverse kind of transition.

One of the aims of this chapter is to show that the marginals in 4.8 over the random graph can be realized if the parameters of this stochastic system satisfy appropriate conditions. In particular, I show that, for fixed s and τ, there exists an upper bound to t_1 above which the network can no longer support the desired statistics.

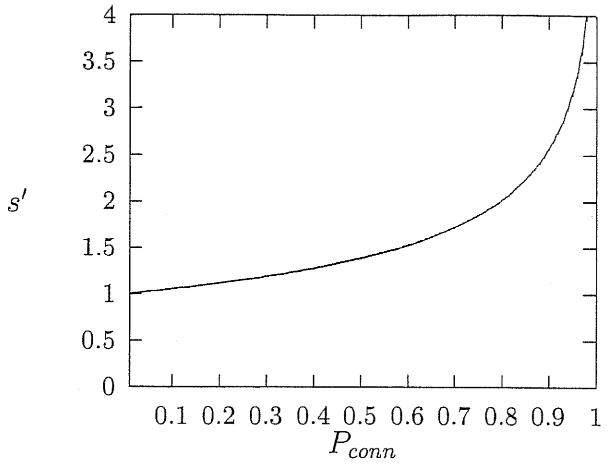

Figure 4.1: Average number of channels per node versus the probability for two nodes to be reachable one from the other through a path of the graph.

Trivially, the problem is solvable if $t_1 = \tau$, that is with independent RVs. I concentrate on the case with $t_1 > \tau$; the case with $t_1 < \tau$ is similarly treatable.

The probability for two nodes to be reachable, one from the other, through a path in $G(M, s)$ is easily found to be as in figure 4.1 versus the average number of connections per node ($=s'=s\cdot(M-1)$, the analytical relation being $s' = -\frac{\ln(1-p)}{p}$; cfr Appendix A). In the range interesting for neuroscience ($2 < s' < 10$) such probability is very near to 1. If t_1 were equal to 1, the large majority of the nodes would be in the same state (0 or 1), which is not acceptable. Thus, t_1 must be smaller than 1.

Is it sufficient to keep t_1 smaller than 1? One is brought to conjecture that the higher the correlation inside the set of nodes (that is the higher s'), the lower the upper bound to t_1. This idea seems to be confirmed by the results obtained in this chapter.

4.4 Existing theorems

The theorem of Kellerer (Kellerer, 1964; Kellerer, 1991) gives a necessary and sufficient condition for the consistency of a finite set of marginals, but in the present context it is not useful. In fact, since I am dealing with a large number of randomly correlated RVs, the hypotheses of Kellerer's theorem cannot be verified for each specific realization of the random graph representing the correlations. On the other hand, I could verify the hypotheses *in probability* when considering the thermodynamic limit ($M \to \infty$), but then I would have an infinite set of marginals and the cited theorem could not be applied. Besides, the theorem requires the verification of a condition over a large set of test functions, and this would also be a difficult task.

The crucial hypothesis in Vorobev's theorem (Vorobev, 1962) is the regularity of the complex of the subsets of RVs that are the arguments of the given marginals. Roughly speaking, the set of marginals is consistent (or "extendable" in Vorobev's terminology) if it is not possible to describe a cycle in the cited complex passing from one subset to another one only if these have not empty intersection. Obviously, this is not the case in the present model, since every couple of nodes is argument of a marginal. So, Vorobev's theorem cannot be applied in the present context. However, it should be noted that most of the couple marginals are almost factorizable, and this could be an hint toward an extension of Vorobev's theorem to the present model.

4.5 Conditions from self-averaging

A first condition on the parameters is obtained from the reasonable assumption that in the thermodynamic limit the average activity on the graph is not affected by the condition that one particular node is active (actually, this is equivalent to requesting self-averaging of the average activity $\frac{1}{M}\sum\limits_{m}\tau_m$; cf. Appendix B):

$$\begin{aligned} &< \tfrac{1}{M-1}\sum_{m\neq n}\tau_m|\tau_n = 1 >= \\ &= \tfrac{1}{M-1}\sum_{m\neq n}\mathcal{P}(\tau_m = 1|\tau_n = 1, S) \longrightarrow \tau \end{aligned} \tag{4.17}$$

for any n. Then the average over the nodes must be

$$\tfrac{1}{M}\sum_{n}\tfrac{1}{M-1}\sum_{m\neq n}\mathcal{P}(\tau_m = 1|\tau_n = 1, S) \longrightarrow \tau. \tag{4.18}$$

Using eqs.4.8, it follows

$$s\cdot(t_1-\tau) = 0. \tag{4.19}$$

If s is finite when $M\to\infty$, t_1 cannot be greater than τ. Thus, the stochastic system I consider in this work makes sense only if $\lim\limits_{M\to\infty} s = 0$. Indeed, it is interesting to study the statistical properties of the model without conditioning them through adequate tuning of the input patterns statistics: the main point is how the structure of connections can reflect the correlations among the parallel contributions of the input signals, keeping the average activity constant.

If I allowed for anticorrelation between unconnected nodes, that is

$$\begin{aligned} \mathcal{P}(\tau_m = 1, \tau_n = 1|s_{mn}) &= \tau t_1 s_{mn} + \tau x(1-s_{mn}) \\ \mathcal{P}(\tau_m = 1, \tau_n = 0|s_{mn}) &= (1-\tau)t_0 s_{mn} + \\ &+ (1-\tau)x'(1-s_{mn}) \end{aligned} \tag{4.20}$$

with $x<\tau$ and $x' = \tau\frac{1-x}{1-\tau} > \tau$, the previous argument of self-averaging would give the bound (see Appendix B):

$$x = \frac{\tau - st_1}{1-s}. \tag{4.21}$$

This implies that t_1 must never be larger than $\frac{\tau}{s}$. In particular, t_1 must be equal to τ if $s=1$ (while x loses physical meaning).

Thus, marginals in 4.20 can satisfy the request of self-averaging also with finite s. In this chapter I discuss only the case with $s = O(\frac{1}{M})$, and hence with $x=\tau$, that seems to be the more interesting at least for the neuroscientific problem. Perhaps useful results for the finite s case could be obtained by introducing correlations in the set of connections $\{s_{mn}\}$ also, such that the probability for the connection (i, j) to exist is larger than chance if the connections (i, k) and (k, j) exist, for any k; this would also be interesting from the point of view of neuroscience since there is appreciable evidence that any two small neocortical areas are more probably connected with each other if another area is connected to both (Jouve et al., 1998).

4.6 A Fictitious Dynamics from a MRF

I introduce a stochastic dynamics of the activity of interacting nodes to construct a probability distribution whose marginals are those in eqs.4.7, when s' is finite. I want to underscore that this dynamics is purely fictitious: no particular real physical process is assumed to underlie the stochastic system. The desired p.d.f. is given by the asymptotic distribution of the dynamical process, so demonstrating that at least one solution to the marginal problem exists.

First, guessing an Ansatz, I consider the marginals in eq.4.8 with $f^{11}_{mn} = f^{01}_{mn} = 1$ for any connected pair (m, n), and insert them into eq.4.4 in order to obtain the conditional probability distribution of the activities of all the nodes directly connected to another one, the condition being the activity of the latter:

$$\begin{aligned} \mathcal{P}(\{\tau_n, \forall n | s_{mn} = 1\} | \tau_m = 1) &= t_1^{n_+}(1 - t_1)^{n_-}, \\ \mathcal{P}(\{\tau_n, \forall n | s_{mn} = 1\} | \tau_m = 0) &= t_0^{n_+}(1 - t_0)^{n_-}, \end{aligned} \tag{4.22}$$

where n_+ and n_- are the numbers of adjacent nodes in activity level 1 and, respectively, 0. The assumption of conditional independence in eqs.4.22 as a first approximation is partially justified by the absence, with probability equal to 1, of direct connections among the $n_+ + n_-$ nodes connected with module m, and can be considered as the zero-th order approximation of eq.4.5 for small ψ_{mn}. This approximation is adopted due to the difficulties in the analytical treatment (see the following), and in fact is a limit to the application of the method that instead requires the most general local marginal.

For a pair of unconnected nodes (m, n) that share a common neighbor, eqs.4.22 give

$$h_{mn}^{11} = \tau \left(\frac{t_1}{\tau}\right)^2 + (1-\tau)\left(\frac{t_0}{\tau}\right)^2 . \tag{4.23}$$

At this point, the p.d.f. for a node given the activities of all its neighbors can be found through Bayesian inversion. After some algebra, this can be written as

$$\mathcal{P}(\tau_m = 1|\{\tau_n, \forall n | s_{mn} = 1\}) = \frac{1}{1 + \exp\{-\beta(n_+ - \lambda n_- + \nu)\}} \tag{4.24}$$

being

$$\beta = \ln\frac{t_1}{t_0}, \qquad \beta\lambda = \ln\frac{1-t_0}{1-t_1}, \qquad \beta\nu = \ln\frac{\tau}{1-\tau}. \tag{4.25}$$

Equation 4.24 defines a MRF over the graph if $\mathcal{M}$ is consistent.

Then I use eq.4.24 to generate the discrete-time stochastic asynchronous dynamics:

At each time step

- choose a node randomly,

- update its activity with probability given by eq.4.24.

Actually, numerical simulations have shown that the system is more stable (in simulations M must be finite) and reaches equilibrium more easily if the term ν in eq.4.24, which is mainly responsible for keeping the average activity in the net at the desired level, is modified as the instantaneous average activity deviates from τ. Thus, in place of eq.4.24 I consider the following expression:

$$\mathcal{P}(\tau_m = 1|\{\tau_n, \forall n \neq m) = \frac{1}{1 + \exp\{-\beta(n_+ - \lambda n_- + \nu\frac{1}{\tau M}\sum_m \tau_m)\}}. \tag{4.26}$$

So the attractor with average activity τ is strengthened. When, at equilibrium, the average activity is equal to τ, expression 4.26 coincides with expression 4.24.

4.7 Analytical approach

The Markov dynamics generated by eq.4.26 satisfies detailed balance in the thermodynamic limit. Hence, during time evolution, the system tends to an equilibrium state. It can be shown that the corresponding asymptotic p.d.f. over the configurations is Boltzmann-like with Hamiltonian

$$H = -\frac{1-\lambda}{2}\sum_{i,j} s_{ij}\tau_i\tau_j + \lambda\sum_{i,j} s_{ij}\tau_i(1-\tau_j) - \frac{\nu}{2\tau M}\left(\sum_i \tau_i\right)^2 . \tag{4.27}$$

Applied to the present case, the theorem of section 4.2 confirms, through eq.4.6, the form of the Hamiltonian in eq.4.27 (except for the strengthening factor of ν), since the contributions to H by the cliques with more than two nodes are negligible in the thermodynamic limit. However, the equivalence theorem cannot substitute the construction of a dynamics since the latter is needed for the simulations until an analytical solution of the statistical canonical model with H is found (from which to extract the relation between t_1^{max} and s'). It seems interesting that the use of the characteristics of a MRF as updating laws for an asynchronous dynamics drives the system toward an equilibrium distribution that coincides with the global distribution provided by the equivalence theorem.

The Boltzmann distribution with the Hamiltonian of eq.4.27 is the candidate for the distribution with the desired marginals I am looking for. Mathematical tools of the statistical mechanics of the Boltzmann-Gibbs ensemble may be used, in principle, to study the equilibrium properties of the system. Unfortunately, the application of the mean-field techniques to the calculation of the partition function soon finds a serious mathematical obstacle whose physical correlate is the high variability of the field (acting on a node) across nodes.

Then, since the system does not lend itself to a mean-field analysis, I try another approach. If t_1 were equal to τ the system would not deviate from independence and the only effective term of the Hamiltonian in eq.4.27 in the partition function would be

$$H_0 = -\frac{\nu}{2\tau M}\left(\sum_i \tau_i\right)^2 . \tag{4.28}$$

Such a reduced Hamiltonian is easily treatable with mean-field techniques. Thus, I perform a perturbative expansion for $t_1 \simeq \tau$. Defining $\epsilon = t_1 - t_0$ as the small parameter, the constants entering the "perturbed" Hamiltonian are:

$$\begin{aligned} t_1 &= \tau + \epsilon(1-\tau), \\ t_0 &= \tau - \epsilon\tau, \\ \beta &= \tfrac{\epsilon}{\tau} + O(\epsilon^2), \\ \beta\lambda &= \tfrac{\epsilon}{1-\tau} + O(\epsilon^2). \end{aligned} \tag{4.29}$$

Writing H as $H_0 + H_1$ and

$$Z_0 = \sum_{\{\tau_i\}} \exp\left(-\beta H_0(\{\tau_i\})\right), \tag{4.30}$$

up to the second order in ϵ I have

$$Z \simeq Z_0\left[1 - <\beta H_1>_0 + \frac{1}{2} < (\beta H_1)^2 >_0\right] \tag{4.31}$$

where $< A >_0$ is the average of A in the ensemble with Hamiltonian H_0 ("unperturbed" system).

Then, the free-energy is (up to $O(\epsilon^3)$)

$$<< \ln Z >> \simeq \ln Z_0 - << \langle\beta H_1\rangle_0 >> + \frac{1}{2} << \left[\left\langle(\beta H_1)^2\right\rangle_0 - \left(\langle\beta H_1\rangle_0\right)^2\right] >>, \tag{4.32}$$

where $<< \cdot >>$ indicates the average over the quenched variables.

To verify if the statistics of the system have the desired marginals, I define the "observables"

$$\begin{aligned} \tilde{t}_1 &= \tfrac{1}{\tau M s'} \sum_{i,j} s_{ij}\tau_i\tau_j, \\ \tilde{\tau} &= \tfrac{1}{M} \sum_i \tau_i, \end{aligned} \tag{4.33}$$

The averages of these quantities can be derived from eq.4.32, and are respectively equal to $t_1 + O(\epsilon^2)$ and $\tau + O(\epsilon^2)$. This means that the canonical model, in the approximation used in the mathematical treatment, obeys a p.d.f. whose marginals are those of eqs.4.8.

This result strongly supports the belief that the particular marginal problem analyzed in this work (with s' finite) has at least one solution and, consequently, that the Ansatz for the marginal used to solve the original neurobiological problem constitutes a meaningful model. Unfortunately, the perturbative expansion up to order ϵ^2 does not provide an upper bound to t_1. As already stated, any increase in s' makes the system more strongly correlated; this should reduce the upper bound to t_1 monotonically until, for s' large enough, the statistics should not appreciably deviate from independence ($t_1 = \tau$).

4.8 Analogy with the Ising ferromagnet

One could think about the correlated system in question as an Ising spin model in which each spin interacts ferromagnetically with a small number (about s') of other spins randomly chosen at the beginning. Let us define "cluster centered around a spin" the set of all the spins connected to that one; this relation is quenched. It can be shown that, at least in an approximate calculation (e.g., large temperature), as intuition suggests, the spins of a cluster tend to be oriented parallel to the center of that cluster even if the temperature is well above the Curie point and the global magnetization is zero. In a sense this phenomenon reminds of Weiss domains but in the present model there is no topology: the spins belonging to the same cluster may be scattered throughout the system and may interact with several different clusters. If s is finite, the ferromagnetic model is purely mean-field and the phenomenon of cluster polarization disappears (s_{ij} are homogeneously distributed).

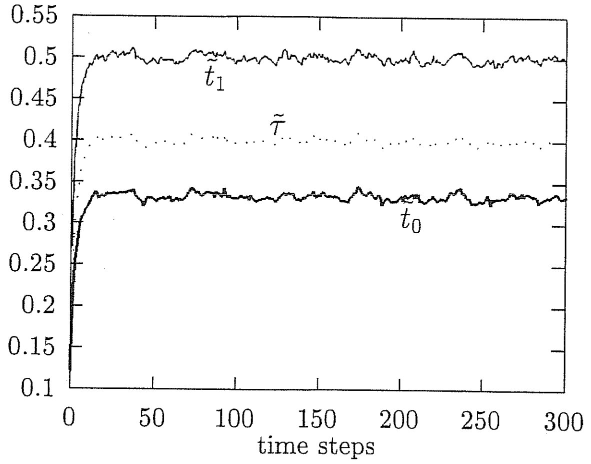

Figure 4.2: Typical output of a simulation with $\tilde{t}_1$ and τ in the useful region ($s' = 5$, $\tau = .4$, $t_1 = .5$, $M = 40,000$).

4.9 Simulation results

I have implemented the dynamics in numerical simulations, where s' has been taken between 2 and 8 since this appears to be the interesting region to study. Having a small s' allows one to simulate very large networks up to more than 100,000 nodes, on a modest PC.

The main quantities observed in the simulations have been $\tilde{t}_1$, $\tilde{\tau}$, $\tilde{t}_0$. According to these "order parameters" (actually, the last one is obtainable from the other two), the system soon reaches equilibrium (if the selected parameters t_1 and τ are in due ranges; t_0 is then fixed by the relation 4.15). The oscillations of the order parameters derive from the finiteness of the simulated net and are in good agreement with the prediction of the variances (see Appendix B).

In figure 4.2 a typical output of the simulation when t_1 is in the useful range (see below) is shown. At first the dynamics drive the system activities toward the equilibrium distribution. Then, the order parameters oscillate around their averages: $\tilde{t}_1 \simeq 0.5000 \pm 0.0043$, $\tilde{\tau} \simeq 0.4001 \pm 0.0037$, $\tilde{t}_0 \simeq 0.3335 \pm 0.0038$, with standard

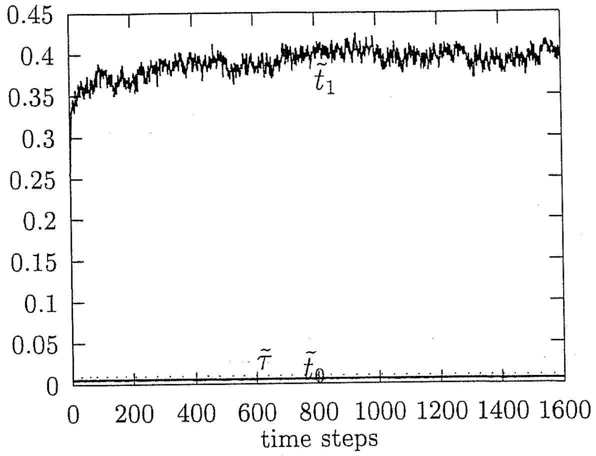

Figure 4.3: Output of a simulation with biologically plausible parameters: $t_1 = .4$, $\tau = .01$, $s' = 4$, $M = 10^5$. The quantities $\tilde{\tau}$ and $\tilde{t}_0$ are very small compared with $\tilde{t}_1$ ($\frac{t_1}{\tau} = 40$).

deviations that are compatible with the estimates based on the pattern statistics (Appendix B) and that decrease as the number of nodes increases. The relation in eq.4.15 is fully respected, that is:

$$(1-\tilde{\tau})\tilde{t}_0 = \tilde{\tau}(1-\tilde{t}_1). \tag{4.34}$$

In figure 4.3 it is shown the output of a simulation with parameters t_1 and τ biologically plausible. The order parameters assume the following values: $\tilde{t}_1 \simeq 0.3956 \pm 0.0100$, $\tilde{\tau} \simeq 0.0100 \pm 0.0002$, $\tilde{t}_0 \simeq 0.0060 \pm 0.0002$. The standard deviations are compatible with the estimates and relation in eq.4.15 is fully respected.

In these conditions the system does not switch among different equilibrium values of the order parameters, possibly indicating that the dynamics cannot switch to different metastates, at least at the level of the measured quantities. This has been verified by testing the system dynamics for a very long time and for several quenched structures of connections.

The other important result from simulations is the estimation of the upper bound t_1^{max} to t_1, that is the value of t_1 beyond which the system is no longer able to sustain

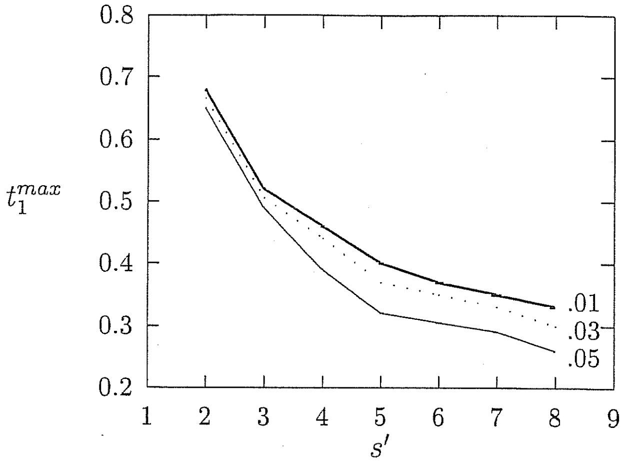

Figure 4.4: Estimation, from simulations, of the upper bounds to t_1 as functions of s', for three values of τ (.01, .03, .05)

the distribution with the desired marginals. If t_1 is increased above its upper bound, the average values of $\tilde{t}_1$, $\tilde{\tau}$, $\tilde{t}_0$ no longer equal t_1, τ, t_0, being $\tilde{t}_1$ and $\tilde{\tau}$ lower than t_1 and τ respectively. In this out-region case the relation in eq.4.15 is not respected.

Figure 4.4 shows the dependence of t_1^{max} on s' for three different values of τ (around the biologically interesting values). It confirms that the more correlated the system, that is the larger s', the smaller the useful range available to t_1. This result, obviously, might not be universal; other dynamics, that is other Ansatz on the probability of eq.4.5, might provide the system with a larger useful range for t_1. So, the bound depicted in figure 4.4 might not be the necessary one implied by the general assumption about the marginals in eq.4.7. Further investigation is going on to find an analytical prediction on the bound and to verify its generality.

Simulations also show the necessity of the structural factors, at least when the MRF is defined by eq.4.26. For example, it is very clear from computer outputs that the correlation between two unconnected nodes that share a common neighbor is, for many such pairs of nodes, sensibly larger than in the case of independence (e.g.[4],

[4] The indicated standard deviations are not the experimental ones (overestimates) but those

$< \tau_m \tau_n >= .0055 \pm .0007$ instead of $.0025 \pm .0005$, being $\tau = .05$, $t_1 = .3$, $s' = 4$, $M = 20,000$; or $< \tau_m \tau_n >= .0044 \pm .0005$ instead of $.0025 \pm .0005$, being $\tau = .05$, $t_1 = .2$, $s' = 4$, $M = 100,000$). The values are in good agreement with the value $\tau t_1^2 + (1 - \tau) t_0^2$ from eq.4.23, that is what one expects from a generating dynamics as the one here implemented and from the consideration that the probability for a pair of nodes to have more than one common neighbor in the diluted graph is negligible, and the contribution given by longer connecting paths is not dominant. The ensemble average of the activity of any node is in good agreement with the fixed value of τ. The correlation of the activities of connected nodes is most usually in excellent agreement with the expected value $t_1 \tau$.

4.10 Conclusions

In the first part of this chapter I have presented a necessary and sufficient condition for the consistency of a class of families of marginal distributions defined on finite sets of discrete random variables. The main assumption is the presence, in the given family of marginals, of all possible pairs of random variables from the given stochastic system. Using mainly the theorem of equivalence between Markov Random Fields and Gibbs states, I have shown that a family of marginals is consistent if and only if such marginals equal the corresponding marginalizations of a Gibbs distribution constructed according to a precise rule.

The second part of the chapter has been dedicated to the study of a particular system from theoretical neuroscience.

In previous chapters it has been shown that one of the ingredients to construct a plausible model of the dishomogeneous neuronal networks subserving memory retrieval in the associative areas of primate neocortex is the introduction of corre-

inferred from a binomial distribution over trials with the mean equal to the indicated experimental mean.

lations among the activities of connected modules of recurrent networks. Modules, consisting of densely connected neurons, can be each in one of two possible states (either active or non-active). Any two modules are more likely to be in the same state of activity if connected by a channel of communication through which, as an abstraction, neuronal projections are allowed to pass. This biological system is modeled as a set of binary random variables (corresponding to modular activities) that are randomly correlated, thus constituting a graph where the presence of an edge between two nodes is to be interpreted as the existence of prominent interdependence between the two corresponding RVs.

The positive results obtained in the work (Fulvi Mari and Treves, 1998) are valid only if one verifies the hypothesis used there about the marginal distributions of the activities of pairs of connected and unconnected units (=modules). In particular, the set of marginals has to be confirmed as a consistent family.

Theorems already existing (to my knowledge) about probability distributions with given marginals cannot be applied because of the large number of RVs together with the randomness in the structure of correlations.

I have formulated the problem in terms of probability distributions on random graphs. I have demonstrated that the stochastic system cannot be realized if any node is connected with finite probability to any other node while the global average activity is fixed. Then, with the idea of making use of the condition derived in the first part of the chapter, I have produced a fictitious dynamics whose asymptotic distribution has the desired marginals, thus showing that the set of given marginals is consistent. For small correlations I have analytically accounted for the equilibrium properties of the system. For arbitrary correlations I have performed numerical simulations whose outputs confirm (within the statistical fluctuations) the analytical results more generally and provide the upper bound to the node activity correlation as a function of the average activity for some values of the average number of neighbors per node. Both approaches have strongly suggested that the particular marginals family is consistent, thus supporting the underlying neuronal theory. An analogy with a magnetic system is also proposed to emphasize the main phenomenon.

Further investigations are going on to find out an analytical relation between the upper bound to the correlation and the number of neighbors that fits the numerical results.

I am also studying possible refinements of the result of section 4.2 and a more extensive use of Markov Random Fields theory to the problem of marginal consistency of arbitrary families of marginals.

Appendix A: Connectivity of the random graph

Be p the probability for any node, different from node m, to be connected to node m through a path belonging to the graph. Then

$$\begin{aligned} 1-p &= \sum_{k=0}^{M-1} \begin{pmatrix} M-1 \\ k \end{pmatrix} s^k(1-s)^{M-1-k}(1-p)^k = \\ &= \left(1-\tfrac{ps'}{M-1}\right)^{M-1} \to e^{-ps'}, \end{aligned} \tag{4.35}$$

in the thermodynamic limit. It follows

$$s' = -\frac{\ln(1-p)}{p}. \tag{4.36}$$

Appendix B: Self-averaging

As already stated, the average over the patterns of the average activity in the network is

$$< \frac{1}{M}\sum_m \tau_m >= \tau. \tag{4.37}$$

The fluctuation around this mean can be calculated:

$$\begin{aligned}\sigma^2 &=< \tfrac{1}{M}\sum_m(\tau_m-\tau)\cdot\tfrac{1}{M}\sum_n(\tau_n-\tau) >= \tfrac{1}{M^2}\sum_{m,n} < (\tau_m-\tau)(\tau_n-\tau) >= \\ &= \tfrac{\tau(1-\tau)}{M} + \tfrac{2}{M^2}\sum_{(m,n)} < \tau_m\tau_n > -\tfrac{4\tau}{M^2}\sum_{m,n} < \tau_m > +\tfrac{M-1}{M}\tau^2 = \\ &= \tfrac{\tau(1-\tau)}{M}\tfrac{2}{M^2}\sum_{(m,n)}[f^{11}_{mn}s_{mn}t_1\tau + h^{11}_{mn}(1-s_{mn})x\tau] - \tfrac{2(M-1)\tau}{M^2}\sum_m < \tau_m > +\tfrac{M-1}{M}\tau^2.\end{aligned} \tag{4.38}$$

Using 4.9:

$$\sigma^2 = \frac{\tau(1-\tau)}{M} + \frac{M-1}{M}st_1\tau + \frac{M-1}{M}(1-s)x\tau - \frac{M-1}{M}\tau^2. \tag{4.39}$$

In the thermodynamic limit ($M \to \infty$):

$$\sigma^2 \longrightarrow s\tau(t_1 - x) + \tau(x-\tau). \tag{4.40}$$

Assuming self-averaging means to assume that the right-hand-side of eq.4.40 vanishes. This implies:

$$x = \frac{\tau - st_1}{1-s}, \tag{4.41}$$

that is easily seen to be not larger than τ.

In the simulations I have taken $s = \frac{s'}{M-1}$, with s' finite, and, according to eq.4.41, $x = \tau$. Thus, the fluctuation is given by:

$$\sigma^2 = \frac{\tau(1-\tau)}{M} + \frac{M-1}{M}s\tau(t_1-\tau). \tag{4.42}$$

In a very similar way, fluctuation of the parameter $\tilde{t}_1$ can be estimated:

$$< \tilde{t}_1 >= t_1; \tag{4.43}$$

$$< (\tilde{t}_1 - t_1)^2 >\simeq \frac{2}{M}\frac{t_1}{\tau}(t_1 + \frac{1}{s'}), \tag{4.44}$$

having assumed the approximation

$$\begin{aligned}&\mathcal{P}(\tau_i = 1, \tau_j = 1, \tau_k = 1, \tau_l = 1 | s_{mn} = 1, s_{kl} = 1) \simeq \\ &\simeq \mathcal{P}(\tau_i = 1, \tau_j = 1 | s_{mn} = 1)\cdot\mathcal{P}(\tau_k = 1, \tau_l = 1 | s_{kl} = 1),\end{aligned} \tag{4.45}$$

being i, j, k, l four different site indices. Estimate in 4.42 is in excellent agreement with simulation data, while that in 4.44 is slightly an overestimate due maybe to the approximation 4.45.

Chapter 5

Neuropsychology and the model: features, modular retrieval, and category-specific impairment

Some kinds of 'modular' structures often appear in neuropsychology, especially in the connectionists' attempts at modeling neurological disease (e.g.,Rumelhart and McClelland (1986). Recent theories have been formulated, in particular, in order to account for the apparently incongruent results from neuropsychological investigations on semantic category-specific impairments (Gonnerman et al., 1997; Devlin et al., 1998). Even though their connected *units* tend to abstract from the neurophysiological correlates keeping only some salient characteristics of distributed neuronal systems, we found a deep analogy between such theories and our model.

5.1 Features and category-specific impairments

Among the category-specific deficits the most studied is the double dissociation in the ability of the patients at recognizing artifacts and natural kinds. Typically, a patient suffering such a deficit presents either a difficulty in naming items among artifacts more than among natural kinds, or, more frequently, the opposite pattern. One of the first reported cases of category-specific semantic impairments is that of patient VER, described in Warrington and McCarthy (1983). VER, after having suffered a left hemisphere infarct, presented a peculiarly selective impairment in assigning items to their correct categories: in a matching to sample test she scored 96% correct for the category *flowers*, 86% correct for *animals*, and the remarkably lower 63% for objects. Thus, VER was more accurate with 'biological' items than with inanimate objects. One year later, Warrington and Shallice (1984) described also some patients that were presenting the pattern of deficits opposite to VER's one, that is a selective difficulty with living things rather than with non-living things. Since then many other analogous cases of selective semantic impairments have been reported. Initially, many investigators interpreted the difference between these dual patterns of deficits as a consequence of the differences in the exact locations of the lesions on the cortices, having assumed that the structures responsible for specific categories are spatially distributed on distinct areas of the human cortex (Pietrini et al., 1988). However the living/non-living categorization soon failed to account for some pathological cases. Indeed, some patients seemed to perform with categories like *body parts* similarly to how they did with *artifacts* more than with *natural kinds*. Analogously, the degree of the ability of some patients with *musical instruments* was nearer to that shown with *natural kinds*. For these cases the previous theory would have predicted the converse behaviour. Warrington and McCarthy (1987) suggested that the correct discrimination to be considered is that between *functional* and *perceptual* properties of the items. They argued that living things are more often discriminated through their perceptual properties while the distinction between artifacts is mostly based on their functional properties. This observation justifies the performances in knowledge tests with musical instruments and body parts: musical instruments, though being artifacts, are distinguished one from the other more on the base of their visual features than on that of their functional features (their functions are very similar); ȧnalogously, body parts are identified more often through their functions than through their visual aspects.

5.2 A topographic computational model

According to the theory exposed in the previous section, different cortical 'structures' should be responsible for the representations of different features and should be located in different parts of the cortex.

These hypothetical characteristics of the semantic system have been implemented in a computational model by Farah and McClelland (1991). They adopted the assumption common in the field of Parallel Distributed Processing studies according to which the knowledge of an item corresponds to the ability of reproducing a specific distribution of activities across a network of neuron-like units. These units are usually grouped in pools modeling different cortical structures responsible for different tasks. In the model proposed in Farah and McClelland (1991) the units are thought of as each representing some 'aspect' or property of any item presented to the peripherical sensorial system. In particular, two classes of units are considered: those representing visual aspects of the items, and those representing functional characteristics of the same items.

The units are grouped in three pools (fig.5.1): two of them model the peripherical verbal and visual systems, whose activity patterns can either be determined by an imposed input or result from the elaboration that has followed a previous, possibly incomplete, input. The third pool models the semantic system and consists of a set of units representing functional properties and a set of units representing visual, that is *perceptive*, aspects of the items. The simulated net has 24 name units, 24 picture units, 60 visual semantic units, and 20 functional semantics units. There are bidirectional connections between the units belonging both to the same

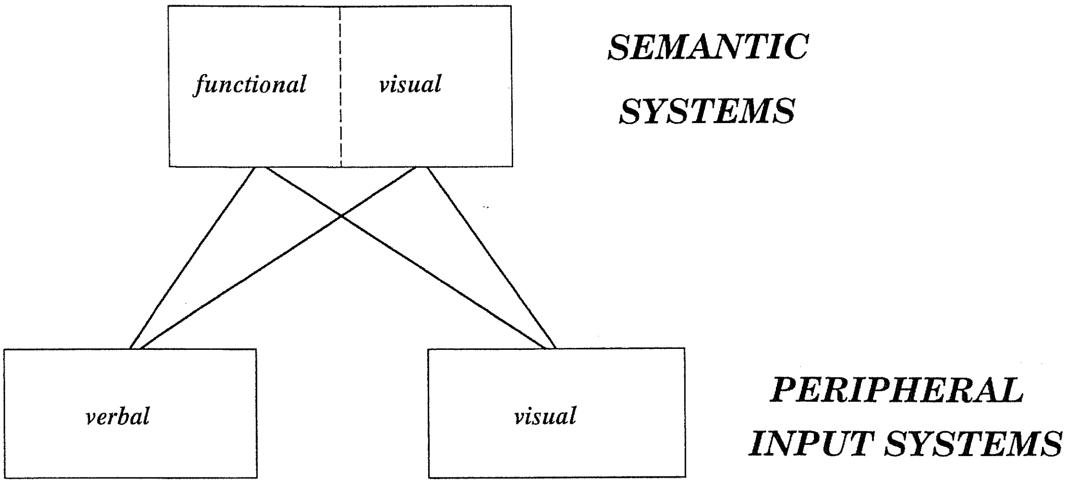

Figure 5.1: Schema of the model introduced by Farah and McClelland (1991). All units belonging to the same pool can be connected. All units belonging to distinct pools can be connected with the exception of the verbal input and visual input pools that cannot communicate directly. (Redrawn from Farah and McClelland (1991).)

pool and to distinct pools, except that there are not direct connections between the verbal and the visual input groups. Any unit can be in one of two states of activity, and connected units interact according to a synchronous dynamics[1]. In Farah and McClelland (1991) the network learned 20 randomly generated patterns of activity, representing 10 living and 10 non-living things. The representation of each item included all the 48 name and picture units; instead on the semantic level the living things were represented with an average of 16.1 visual and 2.1 functional active units, while the non-living things were represented with an average of 9.4 visual and 6.7 functional active units. These numbers were set following statistics, reported in the same article, according to which in common language living things are attributed perceptual and functional features in the ratio 7.7:1, while for non-living objects this ratio is 1.4:1. These figures are in agreement with the suggestion given in Warrington and McCarthy (1987), that is: *living things rely much more on*

[1] In our model (Fulvi Mari and Treves, 1998; Fulvi Mari, 1998) each module (that is, unit) is able to settle in one of many states of activation (local feature retrieval).

perceptual features than on functional features, while non-living things rely almost equally on both. Then Farah and McClelland (1991) simulated cortical lesions onto the semantic system by damaging (more exactly, eliminating) units. Damage to semantic units caused a deficit of the network in producing the correct name output when presented with a visual input. They found the easily predictable result that selective lesions to perceptual semantic units provoke impairments especially with living things, while lesions to functional semantic units bring a deficit more marked with non-living things. They also found that when the degree of damage in the perceptual semantic system was large enough, also the performance with non-living things is impaired. Analogously, extensive damage to functional semantic units led to impairments with living things. This phenomenon is due to the recurrent connectivity inside the semantic systems and the consequent interaction of the two subgroups: this allows the network to produce the correct name output in the absence of lesions, but also implies that the dynamics of the whole semantic system is distorted even when only one of the two subgroups is damaged. Thus the model in Farah and McClelland (1991) was constructed to support the idea that in patients impaired more heavily with living [non-living] things recognition the lesions are mostly localized in neocortical areas that are responsible for memory and processing of perceptual [functional] features.

This apparently satisfying model of semantic category-specific deficits is criticized by some authors as failing to predict the pattern of deficit in Alzheimer disease patients.

5.3 The intriguing case of Alzheimer Disease

There are few studies on category-specific deficits in patients affected by Alzheimer disease. In Silveri et al. (1991) the authors report the results of tests on a group of mild to moderate AD subjects, indicating a pattern of impairment of the kind of category-specific deficits with on average more difficulty with living things than with non-living things. Giustolisi et al. (1993) found that three AD patients in the early stage of the disease behaved as those described by Silveri et al. (1991), but six months later two of them showed the same degree of impairment with living and non-living things, thus suggesting the occurrence of variations in the selectivity of the deficit with the progress in time of the disease.

The main objection to the model described in the previous section is that although AD causes category-specific semantic impairments, it is characterized by a damage widely diffused over the brain, in particular over cortical associative areas, and not specially localized in any area. Thus, the main hypothesis, of topographical difference, in the distribution of perceptual and functional 'units' in the neocortex in Farah and McClelland (1991) is no longer valid, and that model is unable to exhaustively explain the behavioral consequences of AD. Indeed even if the semantic units were actually distributed as in Farah and McClelland (1991), the widespread AD lesions would equally damage the two groups. Nonetheless, category-specific impairments are evident in some AD patients.

5.4 A non-topographic model for category-specific deficits

In agreement with the objection reported in the previous section, Gonnerman et al. (1997) have formulated an alternative theory to predict the AD semantic impairments together with those due to cortical focal injury. They introduce in the previous model the concepts of *intercorrelation* between semantic features that are simultaneously activated in many of the object representations, and of *distinguishing features* by which members within a category are best discriminated. The basic assumptions they extrapolate from experimental data are the following:

- Concepts are represented as distributed patterns of activation over semantic features throughout the network;
- Perceptual features are more numerous than others;
- Living things tend to have a higher ratio of perceptual to functional features then do artifacts;
- Living things have more intercorrelations among features than do artifacts;
- The distinguishing features for biological kinds tend to be perceptual rather than functional, while the opposite pattern is true for artifacts;
- Living things tend to share more features across items than do artifacts.

About the biological kinds they state that, while damage may remove certain connections, the high level of interconnectivity allows the remaining links to compensate for the loss. As the disease progresses, however, and damage accumulates, a critical point is reached beyond which the performance of the net rapidly decrease.

Due to the weaker intercorrelation, artifacts are differently affected by the progression of the damage. Since any of the feature is scarcely supported by the others as a consequence of the low intercorrelation, the decrease of the performance with artifacts is more 'linear' (fig,5.2), so that in a first phase of the disease the performance with artifacts becomes worse than that with biological kinds, but after the abrupt decrease the performance with biological kinds crossovers below the one with artifacts. Since biological items usually are represented by more correlated features and any feature is activated by many natural items, each feature has a low probability to be very informative. As a consequence, in biological representations it is less probable that a main distinguishing feature is damaged than in the case

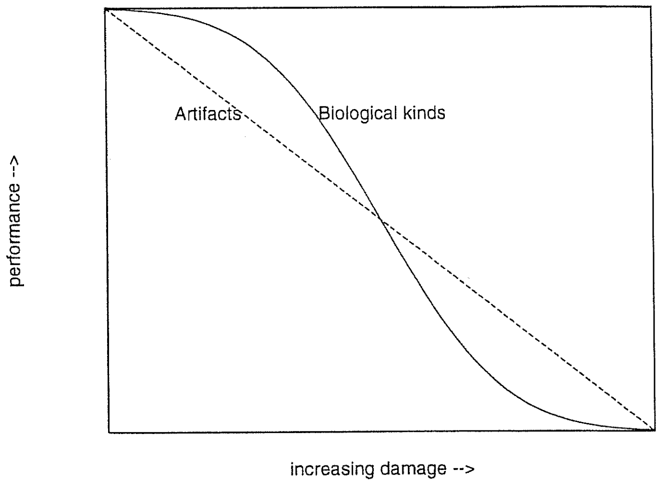

Figure 5.2: Qualitative evolution of the performance with biological kinds and artifacts according to Gonnerman et al. (1997). (Redrawn.)

of the artifacts, for which individual functional features usually constitute fundamental discriminators. Beyond the critical point cited above entire living things categories are damaged, while for artifacts the damage to a feature continues to impair only single objects rather than categories. The prediction thus depends on the distribution of the damage over the neocortex on a small scale, and it could also happen that a single biological item cannot be recognized while other things of the same category are, but it is a low probability event. The authors do not exclude the importance of function localizations, saying that it could also happen that a cortical area mainly responsible for perceptual features (temporal lobe) is more heavily damaged than the areas with functional features (frontoparietal regions), thus, e.g., initially causing a worse performance with living things than with artifacts. This observation protects their theory against some exceptions verified in screening the AD group.

Devlin et al. (1998) tried to give these qualitative ideas a more quantitative support by implementing a connectionist model whose dynamics was numerically sim-

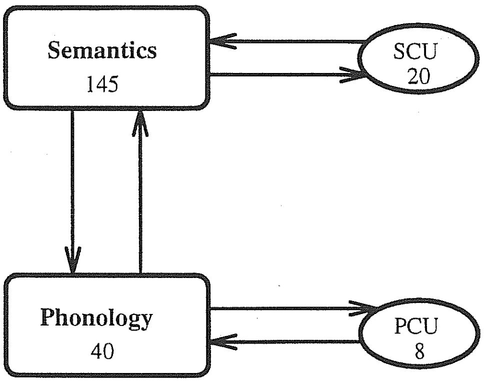

Figure 5.3: Scheme of the connectionist network simulated by Devlin et al. (1998). The figure in a block reports the number of units in that layer. (Redrawn.)

ulated. Their model is an attractor network consisting of 4 layers (fig.5.3). There is no connection joining units belonging to the same layer. The semantic layer contains units that represent the perceptual features (88) and units that represent functional features (57), while the activity distribution over the phonological units represent the presented or produced word thus providing an input/output interface. The semantic layer is fully connected to the Semantic Clean-up Units (SCU); this allows for recursive process at the semantic level (cf. Chap.2 about recurrent networks abilities). For the same reason, the phonologic layer is coupled with a set of Perceptual Clean-up Units (PCU). They obtained a list of features from a group of undergraduate students who had been asked to ascribe functional and perceptual characteristics to 60 given items. The resulting statistics did not coincide with those in Farah and McClelland (1991), but had the same trend suggested by Warrington and McCarthy (1987) about the ratios of perceptual to functional features in living and non-living things. They also verified their hypotheses on the intercorrelations between features, finding that features related to biological kinds occurred simultaneously more often than those related to artifacts. For this reason and for the large number of items that shared features, each biological feature retained on average less 'information' about the represented item than artifact features did. Usually, among the features representing an artifact, the number of 'informative' features (that is, the features that resulted to have an 'informativeness index' above an arbitrarily fixed threshold) was larger than for biological kinds. "Consequently random damage is more likely to strike an informative feature in an artifact object than in a natural kind object and thus increase the likelihood of misnaming that item."(Devlin et al., 1998)

The network was trained until it was able to produce the correct output word when driven into the semantic representation of an item, and to retrieve the correct semantic representation when presented with an input word. Then, they simulated AD damages by progressively removing a number of randomly chosen connections between the semantic pool and its clean-up pool. The majority of the simulations displayed a mild selective impairment with artifacts in a brief initial period of the damage progression, but quite soon the performance with biological kinds became on average the worse. Only some simulations showed a more marked differential behavior. The authors stress the qualitative agreement with the proposed theory. There was also quite a large variability of the behaviour of the net across repeated simulations. According to the authors, this explains the apparently contradictory results obtained by a few patients in neuropsychological semantic tests.

5.5 The modular neuronal model with selective semantic deficits

In neuropsychology literature, as in the papers reported above, the concepts of *feature* and of *units* supporting their representation are quite common, and often the approximate positions of these hypothetical units on the neocortical surface is also indicated, using a rough correspondence between cortical areas and cognitive functions obtained with neuropsychological investigations. Nevertheless, in that literature a more precise relation between these units and neocortical neuronal structures is never indicated.

At the level of neocortical associative areas, we propose the identification of the units with the *modules*, whose existence and structure has been well studied (see Chapter 1). Considering modules as small autoassociators, as described in Chapters 2 and 3, we make the hypothesis that each of them analyzes one particular 'aspect' of the information coming from perceptual cortices, as a shape, a visual characteristic of an object, a modality of use of an object, and so on. If the input arriving to a module equals, or is very similar to, a significant part of a learned pattern, the local net of the module 'recognizes' it and settles in the corresponding retrieval state for some time. (We have not investigated strictly dynamical aspects of the process, having limited our study to a 'static' analysis.) Actually, the *module* may be seen as a generalization of the connectionist *unit*, since the latter is usually allowed to settle in one of only two possible states (an either 'yes' or 'no' answer), while the module, once active, can 'choose' from a large set of possible output states. The interaction between modules, that is the long-range projections of their neurons, provide the ability of the whole net to compose the complete representation of an item through the retrieval of the correct feature (or a silent state if no feature is elicited) in every module. When all the aspects of the object are correctly represented, that is each module recognizes the particular aspect that it analyzes, it can be said that the whole modular network has recognized the object. If, e.g. due to damage, some modules lose their ability to retrieve the features, the items whose representation patterns differ by only the aspects analyzed by the damaged modules will be confused and, thus, misnamed and inappropriately identified.

In this view, we probe our model (Fulvi Mari and Treves, 1998) at reproducing the semantic deficits due to damage to connections and/or to breakdown of entire modules. We adopt the same hypotheses on semantic statistics as in Devlin et al. (1998). We can implement the larger number of correlated *units* that analyze features belonging to biological kinds by providing the correspondent modules with more neighbours, that is by assigning to those modules a number of connected modules larger than the average, s'. The higher correlation between biological features is reproduced with a t_{1p} larger than the average t_1. (See chapters 2,3, and 4 for the meaning of the symbols.)

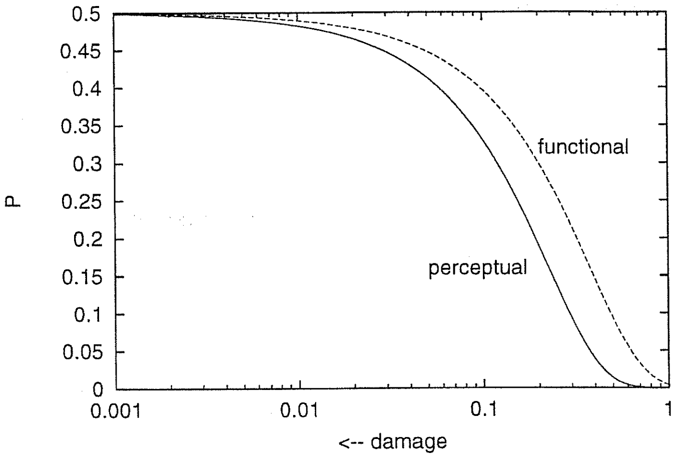

Figure 5.4: Error probability as predicted by the model in Fulvi Mari and Treves (1998), with $\gamma = .5$,$b = .1$,$A = aB$,$a = \tau = .01$,$t_1(funct) = .2$, $t_1(perc) = .4$, $\mu - 1 = 100$, $s'(funct)) = 2$, $s'(perc) = 6$, $s' = 4$, $\alpha = 100$.

To model the damage, we assume that the activity of the neurons belonging to a module is less selective as the disease worsens. For simplicity, we assume that the degree of the damage is homogeneous throughout the network. This should also be the case of more interest in the view of non-focal damages analyzed in Gonnerman et al. (1997) and Devlin et al. (1998). If we only implement the hypothesis of intercorrelation between features, we obtain (fig.5.4) that the probability that a module retrieves a wrong feature or becomes inactive monotonically increases as the injury increases, but on average the performance of a 'natural kinds' module is always better than that of an 'artifacts' module. Thus, the only hypothesis of intercorrelation is not sufficient to reproduce the crossover in our model at the level of the single module. However, it is necessary to consider the global information and to estimate the inability in identifying the pattern as a function of the probability of failure by any single module to obtain a reliable prediction from our model. When a module fails, the probability of the failure of a module connected to the former is plausibly larger than chance, thus apparently putting in disadvantage biological features in comparison with the artifact ones. On the other hand, in the presence of a few errors, it is more difficult to identify an item when featural interdependence is low (artifacts) than when it is high (biological kinds), implying that the failure of a biological feature may have lighter consequences in the recognition of the corresponding item than it happens with artifact features. This sort of 'redundancy' error correction may revert to a dramatic 'co-operation in erring' when the number of failing modules reaches a critical value. A complete model that aims to describe the AD deficits must include all these aspects and cannot be based on only single module predictions. Work is going on in this direction, to verify analitically the plausibility of the model in Gonnerman et al. (1997) and Devlin et al. (1998) with our neuronal model.

Acknowledgements

Above all I thank God the Father and His Son Jesus Christ, whose help has been and is fundamental in my research as in all aspects of my life.

I also wish to thank my supervisor Dr. Alessandro Treves, with whom part of the present work has been developed.

I thank Prof. Nestor Parga for his useful comments and suggestions on the first version of the manuscript.

I cannot forget to thank my mother Anna Maria and Elisa for their love and patience.

Bibliography

Albright, T., Desimone, R. and Gross, C. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque, *J. Neurophysiol.* **51**: 16–31.

Aleksandrov, P. (1956). *Combinatorial Topology*, Md., Graylock Press, Baltimore.

Benes, F., Majocha, R. and Marotta, C. (1988). A modular arrangement of neuronal processes in human cortex: disruption with aging in Alzheimer's disease, *J. Geriatr. Psychiatry Neurol.* **1**: 3–10.

Braitenberg, V. (1978). Cortical architectonics: general and areal, *in* M. Brazier and H. Petsche (eds), *Architectonics of the Cerebral Cortex*, Raven Press, New York, pp. 443–465.

Braitenberg, V. and Shutz, A. (1991). *Anatomy of the Cortex: statistics and geometry*, Springer-Verlag, Berlin Heidelberg.

Cavada, C. and Goldman-Rakic, P. (1986). Subdivisions of area 7 in rhesus monkey exhibit selective patterns of connectivity with limbic, visual and somatosensory cortical areas, *Soc. Neurosci. Abst.* **12**: 262.

Devlin, J., Gonnerman, L., Andersen, E. and Seidenberg, M. (1998). Category specific semantic deficits in focal and widespread brain damage. A computational account, *J. Cogn. Neurosci.* **10**(1): 77–94.

Farah, M. and McClelland, J. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity, *J. Exp. Psych.: Gen.* **120**: 339–357.

Favorov, O. and Diamond, M. (1990). Demonstration of discrete place-defined columns -segregates- in the cat SI, *J. Comp. Neurol.* **298**: 97–112.

Favorov, O. and Kelly, D. (1994). Minicolumnar organization within somatosensory cortical segregates: I.development of afferent connections, *Cerebral Cortex* **4**: 408–427.

Favorov, O. and Whitsel, B. (1988). Spatial organization of the peripheral input to area 1 cell columns. I.The detection of 'segregates', *Brain Res.* **472**: 25–42.

Friedman, H. and Goldman-Rakic, P. (1994). Coactivation of prefrontal cortex and inferior parietal cortex in working memory tasks revealed by 2DG functional mapping in the rhesus monkey, *J. Neurosci.* **14**: 2775–2788.

Fujita, I., Tanaka, K., Ito, M. and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex, *Nature* **360**: 343–346.

Fulvi Mari, C. (1998). Markov fields and probability distributions with given marginals. A general method and a problem from theoretical neuroscience, *Phys. Rev. E* (submitted).

Fulvi Mari, C. and Treves, A. (1998). Modeling neocortical areas with a modular neural network, *Biosystems* (to appear).

Fuster, J. and Jervey, J. (1982). Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task, *J. Neurosci.* **2**: 361–375.

Fuster, J. M. (1997). Network memory, *Trends in Neuroscience* **20**(10): 451–459.

Giguere, M. and Goldman-Rakic, P. (1985). Disjùnctive distribution of mediodorsal thalamic afferents in the prefrontal cortex of rhesus monkey, *Soc. Neurosci. Abst.* **11**: 677.

Giustolisi, L., Bartolomeo, P., Daniele, A., Marra, C. and Gainotti, G. (1993). Category specific semantic impairment for living things in the early stages of Alzheimer's desease: Further evidence from a study on single cases, *J. Clin. Exp. Neuropsych.* **15**: 403.

Goldman, P. and Schwartz, M. (1982). Interdigitation of contralateral and ipsilateral columnar projections to frontal association cortex in primates, *Science* **216**: 755–757.

Goldman-Rakic, P. (1987). Circuitry of the prefrontal cortex and the regulation of behavior by representational memory, *in* F. Plum and V. Mountcastle (eds), *Handbook of Physiology, The Nervous System, V ed.*, American Physiological Society, Bethesda, MD, pp. 373–417.

Goldman-Rakic, P. (1988). Changing concepts of cortical connectivity: parallel distributed cortical networks, *in* P. Rakic and W. Singer (eds), *Neurobiology of neocortex*, John Wiley, pp. 177–202.

Gonnerman, L., Andersen, E., Devlin, J., Kempler, D. and Seidenberg, M. (1997). Double dissociation of semantic categories in Alzheimer's disease, *Brain and Language* **57**: 254–279.

Graziano, M., Andersen, R. and Snowden, R. (1994). Tuning of MST neurons to spiral motions, *J. Neurosci.* **14**: 54–67.

Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices, *unpublished* .

Hevner, R., Aboitiz, F. and Illing, R.-B. (1993). More modules, *Trends in Neuroscience* **16**(5): 178–180.

Hyvarinen, J. (1982). *The Parietal Cortex of Monkey and Man*, Springer-Verlag, Berlin, Heidelberg.

Jacobs, K. and Donoghue, J. (1991). Reshaping the cortical motor map by unmasking latent intracortical connections, *Science* **251**: 944–947.

Jacobs, R., Jordan, M. and Barto, A. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks, *Cognitive Science* **15**: 219–250.

Jenkins, I., Brooks, D., Nixon, P., Frackowiak, R. and Passingham, R. (1994). Motor sequence learning: a study with positron emission tomography, *J. Neurosci.* **14**: 3775–3790.

Jones, E. (1985). *The Thalamus*, Plenum Press, New York.

Jones, E., Friedman, D. and Hendry, H. (1982). Thalamic basis of place- and modality-specific columns in monkey somatosensory cortex: a correlative anatomical and physiological study, *J. Neurophysiol.* **48**: 545–568.

Jouve, B., Rosenstiehl, P. and Imbert, M. (1998). A mathematical approach to the connectivity between the cortical visual areas of the macaque monkey, *Cerebral Cortex* **8**: 28–39.

Kaas, J. H. (1987). The organization of neocortex in mammals: implications for theories of brain function, *Ann. Rev. Psychol.* **38**: 129–151.

Kellerer, H. (1964). Verteilungsfunktionen mit gegebenen marginalverteilungen, *Z. Wahrscheinlichkeitstheorie* **3**: 247–270.

Kellerer,·H. (1991). Indecomposable marginal problems, *in* G. Dall'Aglio, S. Kotz and G. Salinetti (eds), *Advances in Probability Distributions with Given Marginals*, Kluwer, Dordrecht.

Kemeny, J., Snell, J., Knapp, A. and Griffeath, D. (1976). *Denumerable Markov Chains (2nd ed.)*, Springer-Verlag, New York.

Lauro-Grotto, R., Reich, S. and Virasoro, M. (1997). The computational role of conscious processing in a model of semantic memory, *Cognition, Computation and Consciousness*, Oxford Univ. Press, Oxford, U.K.

Malach, R. (1994). Cortical columns as devices for maximizing neuronal diversity, *Trends in Neuroscience* **17**(3): 101–104.

McRae, K., de Sa, V. and Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning, *J. Exp. Psych.: Gen.* **126**(2): 99–130.

Mishkin, M., Ungerleider, L. and Macko, K. (1983). Object vision and spatial vision: two cortical pathways, *Trends in Neuroscience* **6**: 414–417.

Miyashita, Y. and Chang, H. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex, *Nature* **331**: 68–70.

Mountcastle, V. (1995). The parietal system and some higher brain functions, *Cerebral Cortex* **5**: 377–390.

Mountcastle, V. B. (1997). The columnar organization of the neocortex, *Brain* **120**: 701–722.

Mountcastle, V., Lynch, J., Georgopoulos, A., Sakata, H. and Acuna, C. (1975). Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space, *J. Neurophysiol.* **38**: 871–908.

Mountcastle, V., Motter, B., Steinmetz, M. and Duffy, C. (1984). Looking and seeing: the visual functions on the parietal lobe, *in* G. Edelman, G. W.E. and C. W.M. (eds), *Dynamical Aspects of Neocortical Function*, John Wiley & Sons, New York, pp. 159–193.

O'Kane, D. and Treves, A. (1992a). Short- and long-range connections in autoassociative memory, *J. Phys. A: Math. Gen.* **25**: 5055–5069.

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

Pietrini, V., Nertempi, P., Vaglia, A., Revello, M., Pinna, V. and Ferro-Milone, F. (1988). Recovery from herpes simplex encephalitis: Selective impairment of specific semantic categories with neuroradiological correlation, *J. Neurology, Neurosurgery, and Psychiatry* **51**: 1284–1293.

Purves, D., Riddle, D. and LaMantia, A.-S. (1992). Iterated patterns of brain circuitry (or how the cortex gets its spots), *Trends in Neuroscience* **15**(10): 362–368.

Rakic, P. (1972). Mode of cell migration to the superficial layers of fetal monkey neocortex, *J. Comp. Neurol.* **145**: 61–83.

Rakic, P. (1988). Intrinsic and extrinsic determinants of neocortical parcellation: a radial unit model, *in* P. Rakic and W. Singer (eds), *Neurobiology of neocortex*, John Wiley, pp. 5–27.

Rumelhart, D. and McClelland, J. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Massachussetts.

Scannell, J., Blakemore, C. and Young, M. (1995). Analysis of connectivity in the cat cerebral cortex, *The Journal of Neuroscience* **15**(2): 1463–1483.

Silveri, M.-C., Daniele, A., Giustolisi, L. and Gainotti, G. (1991). Dissociation between knowledge of living and nonliving things in dementia of Alzheimer type, *Neurology* **41**: 545–546.

Tanaka, K. (1996). Representation of visual features of objects in the inferotemporal cortex, *Neural Networks* **9**(8): 1459–1475.

Tommerdahl, M., Favorov, O., Whitsel, B., Nakhle, B. and Gonchar, Y. (1993). Minicolumnar activation patterns in cat and monkey S1, *Cerebral Cortex* **3**: 399–411.

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

Treves, A. (1991). Are spin-glass effects relevant to understanding realistic autoassociative networks?, *J. Phys. A* **24**: 2645–2654.

Tsodyks, M. and Feigel'man, M. (1988). The enhanced storage capacity in neural networks with low activity level, *Europhys. Lett.* **6**: 101–105.

Van Hoesen, G. W. (1993). The modern concept of association cortex, *Curr. Opin. Neurob.* **3**: 150–154.

Vorobev, N. (1962). Consistent families of measures and their extension, *Theory Prob. Appl.* **VII**(2): 147.

Warrington, E. and McCarthy, R. (1983). Category specific access dysphasia, *Brain* **106**: 859–878.

Warrington, E. and McCarthy, R. (1987). Categories of knowledge: Further fractionations and an attempted integration, *Brain* **110**: 1273–1296.

Warrington, E. and Shallice, T. (1984). Category specific semantic impairments, *Brain* **107**: 829–853.

Young, M. P. (1993). The organization of neural systems in the primate cerebral cortex, *Proc. R. Soc. Lond. B* **252**: 13–18.