ISAS - INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Recurrent networks in the brain and spatial memory: dynamics and storage capacity

Thesis submitted for the degree of
"Doctor Philosophiæ"

CANDIDATE

Francesco Paolo Battaglia

SUPERVISOR

Alessandro Treves

November, 1998

SISSA - SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

TRIESTE
Via Beirut 2-4

TRIESTE

Contents

1 Recurrent processing: an overview 7
- 1.1 Donald Hebb's idea . . . 7
- 1.2 The anatomical substrate of the cell assembly . . . 8
- 1.3 Models of the cell assembly . . . 10
- 1.4 The role of recurrent processing in brain function . . . 13
- 1.5 This thesis . . . 15

2 Recurrent processing in the hippocampus 17
- 2.1 Some anatomy . . . 18
- 2.2 Models of hippocampus in memory . . . 23
  - 2.2.1 Marr's first attempts . . . 23
  - 2.2.2 Recent models of the hippocampus as a memory . . . 24
- 2.3 The place cells system . . . 27
  - 2.3.1 Head direction cells . . . 28
  - 2.3.2 Path integration . . . 29
  - 2.3.3 Recurrent processing models for the place field system . . . 30
- 2.4 The formal analysis of hippocampal models . . . 33
  - 2.4.1 Spatial information encoding in an auto-associative memory . 35
  - 2.4.2 The continuum (space) nature of information does not affect information encoding efficiency . . . 36
  - 2.4.3 A definition of encoded space information . . . 39
  - 2.4.4 Simulations of the multi-chart network . . . 41

3 Analysis of the Multi-chart network 43
- 3.1 Introduction . . . 43
- 3.2 The single map network . . . 43
  - 3.2.1 The one-dimensional case . . . 45
  - 3.2.2 The two-dimensional case . . . 47
- 3.3 Storing more than one map . . . 48
  - 3.3.1 The fully connected network: "dot product" kernel . . . 49

3.3.2 Generic kernel: extremely diluted limit 51
3.3.3 Inhibition independent stability 56
3.3.4 The fully connected model 58
3.3.5 Sparser maps 60
3.4 information storage 61
3.5 Discussion 62

4 Speed of retrieval and stability of retrieval states **67**
4.1 Introduction 67
4.2 The stability-capacity conflict 70
4.3 Realistic inhibition may avoid the conflict 73
4.4 Simulations show stability and fast retrieval 75
4.5 Implications for recurrent processing in the brain 80

A Free energy for the "dot product" model **89**

B Generic kernel, extreme dilution **93**

C Replica free energy: generic kernel **97**

D Generic kernel: storable information calculation **101**

References **109**

List of Figures

2.1 Estimates of the number of principal cells in the hippocampal formation 20
2.2 Schematic representation of principal neurons shape in the hippocampus 21
2.3 The hippocampal circuit 22
2.4 Activity packet from experimental data 32
2.5 Maximal storage capacity for the Hopfield auto-associator 36
2.6 Maximal storable information for the Hopfield type auto-associator . 37
2.7 stable activity configurations in the simulated multi-chart auto-associator 42

3.1 Storage capacity for the 1-D multi-chart model 55
3.2 Storage capacity for the 2-D multi-chart model 56
3.3 1-D "activity peak" profile at maximal storage capacity 64
3.4 2-D "activity peak" profile at maximal storage capacity 65
3.5 Maximal storable information 66

4.1 Capacity of the threshold linear shunting inhibition model 82
4.2 Firing rates for the integrate and fire networ 83
4.3 Information time course for different values of synaptic time constant 84
4.4 Transient time constant and excitatory synaptic time constants . . . 85
4.5 Transient time constant plotted and somatic capacitance 86
4.6 Information time course plotted for the structureless network 87

B.1 The "potential" function $\mathcal{U}(u)$ 96

D.1 The r_{12} function 103
D.2 The I_2 function 107

Acknowledgments

First, I want to thank my family, for the support and the encouragement that made it possible for me to carry out my studies. I want to thank Daniela, who shared with me these years, the good times and those a little harder.

Alessandro Treves guided me through this strange world of neuroscience, providing good advice, good opportunities, and motivation.

I want to thank the people I met at SISSA, Stefano Panzeri, Lorenzo Cangiano, Andrea Benucci, all the people in the Cognitive Neuroscience sector, and all those I had fun with, and all my house-mates (a lot!) Marco Beato, Roberta Donato, Piero Parodi, Daniela Tropea and Marta Deganuto, in chronological order

I also thank Proff. Bruce McNaughton and Carol Barnes and all the people in Tucson for their kind hospitality last year in their lab, and especially Kati Gothard and Andy Fuglevand for the nice time I had staying with them.

Chapter 1

Recurrent processing: an overview

1.1 Donald Hebb's idea

The words "recurrent processing" have been invoked in many different contexts and with many different meanings, and seem to represent a fundamental ingredient of many theories of the brain, or of its subsystems. The origin of the concept is probably in Donald Hebb's idea of cell assemblies (Hebb, 1949), groups of neurons interconnected by a large number of excitatory synapses, which shape and determine the activity of the neurons itself. Recurrent excitatory synapses imply positive feedback: this feedback may be large enough to cause activity in the cell assembly to self-sustain, or may cooperate – and compete – with influences external to the cell assembly, in generating activity configurations for the assembly.

Probably Hebb's major contribution to the history of brain theory was the hypothesis that the *synaptic matrix*, the set of all the synaptic connections between pairs of neurons, encodes cell assemblies, and that synapses are *modifiable* by experience, in a very peculiar way. Hebb hypothesized that when two neurons are often coactive, the strength of the connection between them is increased by some sort of "growth process, or metabolic change". This has been one of the most influential ideas in neuroscience, both for theoretical and experimental research. The great majority of models of neural learning are in some sense "Hebbian": *learning rules* with few exceptions assume that synaptic strength changes are related to the correlation of activity in the pre-synaptic and in the post-synaptic cell. A lot of versions of the Hebb learning rule exist, probably claiming to be more different between each other than they actually are (for a review see e.g. Hertz, Krogh, & Palmer, 1991), and a large amount of research was carried out in the field of "learning theory", with approaches originating from computer science and statistical physics (see e.g. Sompolinsky, Tishby, & Seung, 1990).

The impact of Hebb's ideas on experimental neurophysiology was even more striking. An entire field of research was born, trying to find the physiological mechanism of the "growth process, or metabolic change". An *in vitro* analogue was eventually found (Bliss & Lomo, 1973), the Long-Term Potentiation (LTP) which is observed in hippocampal synapses after simultaneous synaptic stimulation of the pre-synaptic and the post-synaptic cells. The mechanism of LTP and the relationship between the LTP (and related phenomena) and learning (in the Hebb's sense) are very delicate issues, yet to be solved. In fact, LTP is usually induced by means of very artificial protocols, involving innaturally strong, long, or patterned stimuli, conditions that never occur physiologically. On the other hand, for example, blockade of what is thought to be the key molecular agent in LTP induction, the *NMDA* glutamate receptor has been shown to cause learning impairment in behaving animals (see e.g. Morris, Anderson, Lynch, & Baudry, 1986) although the evidence is far from being satisfactory (Barnes, 1995).

1.2 The anatomical substrate of the cell assembly

There is little doubt that recurrent connections between neurons dominate the brain, and particularly the *cortex*, the phylogenetically more recent and advanced part of the brain, including the *hippocampus* (archicortex), the *pyriform cortex* (olfactory cortex or paleocortex) and the *neocortex*, that is, the hemispheres (Braitenberg & Schüz, 1991).

As the hippocampus is one of the main objects of this thesis, and since it has distinctive anatomical features, we postpone the discussion of it to sec. 2.1, and we give here a few data regarding the neocortex, stressing the role of recurrent processing there.

The neocortex, like the hippocampus, has a *planar* organization; it is a sheet of tissue, about 2 mm thick, containing the cell bodies of the neural cells, the neurons, and covering the hemispheres. This thin, gray looking tissue (*gray matter*) overlies a lighter tissue, with no cell bodies, made mainly of fibers (*white matter*).

The neural circuitry looks very similar throughout the neocortex, which for this reason is also called *isocortex*. According to Braitenberg and Schüz (1991), there are three fundamental types of neural cells. *Pyramidal cells*, so called for the shape of their cell body (or *soma*), represent the majority of the cortical neurons (between 62% and 85% in the rat brain) and they are pre-synaptic to excitatory (or Type I, or *asymmetric* synapses, as they are called for the appearance of the synaptic membrane when examined at the electron microscope); the neurotransmitter at those terminal is usually glutamate. They have two distinct dendritic trees, one pointing upward, towards the exterior (apical dendrites), and the second one spreading in the lower

tiers of the cortex (basal dendrites). The axons can spread for long distances in the gray matter, as they also form the fibers connecting different cortical areas (through the white matter). Pyramidal cells also form the fibers exiting the cortex towards subcortical structures. They receive both excitatory and inhibitory synapses, the excitatory ones mainly on special dendritic excrescences called *spines*, the inhibitory ones on the cell body and on the proximal part of the dendritic tree. Other spiny cells, smaller and without the apical dendrites, are found in layer 4 of sensory cortices, in which they are the major recipients of the thalamic inputs: *spiny stellate neurons.*

There are cells with no spines on their dendrites, and in general with no axonal arborization outside the cell's proximal region. These are called *smooth* cells, and they are generally inhibitory, using GABA as neurotransmitter.

It is to be noted that genuine input and output connections are indeed a tiny fraction of the total of the connections between neurons in the cortex. They represent between one in 100 or one in 1000 among white matter fibers, and the proportion further decreases dramatically when one considers local, gray matter, projections (which probably account for one half of the synapses afferent to a single pyramidal cell).

The pattern of connections suggest an organization of the cortex in *columns*, representing the basic functional unit. The size of the column is determined by the typical range of the axonal arborization of pyramidal cells plus the length of the dendritic trees, altogether around 1-2mm. Pyramidal-pyramidal connections represent the vast majority of the synapses in the cortex, around 90% (Braitenberg & Schüz, 1991). In a column, there are about 10^5 neurons (let us consider here the pyramidal cells only, which are the majority of the cells and are probably the ones mainly responsible for actual computations), each one receiving $7 - 8 \times 10^3$ excitatory synapses (in the rat), of which about one half from neurons in the same column. This yields a *connectivity fraction* c, (that is, the average fraction of cells in the column making synapses with a given cell.) of ~ 0.03. This is a quite sparse connectivity, but still it implies that there is a lot of excitatory feedback within one cortical column: one can pass from a neuron to another one in a column, in one or two synaptic steps with very high probability. These recurrent connections have been invoked many times as a very important component in determining the functional behavior of the cortical modules, as we will see in the following of this thesis.

Each patch of cortex is further organized in *layers.* The layering organization is pretty much stereotyped across the whole neocortex (which because of this is also called *isocortex*). The conventional subdivision includes 6 layers, from the most superficial to the deepest. Roughly, Layer I has very few cell bodies in it, and is mainly composed of fibers parallel to the surface. In most parts of the cortex, layers II and III form a continuum, and they are rich in cell bodies. In sensory areas, layer IV is populated by many small cells, the spiny stellate cells, which we have mentioned be-

fore, constituting an important relay station for thalamic inputs. Layer V has many pyramidal cell, the largest ones in the cortex. Layer VI has again smaller cells.

The laminar pattern of connections within a column is rather stereotyped, and has been schematised by Douglas and Martin (1991b) (see also Douglas & Martin, 1998) as follows: layer IV spiny stellate cells receive inputs from thalamic nuclei (lateral geniculate nuclei in the case of primary visual cortex), which represent, though, only 10% of all the excitatory affering synapses. The majority of the excitatory input comes from the cortex itself, e.g. from other stellate cells (30%) and layer V-VI pyramidal cells (40%). Layer IV cells project to layer II-III pyramidal cells, which form a recurrent network among themselves and in turn project to layer V-V1 pyramidal cells, from which they receive feedback projections. Deep layers pyramidal cells are the column's output, sending projections in the white matter, as well as recurrent projections. Again, this sketchy circuit is dominated by intra-cortical feedback. The hypothesis Douglas and Martin make, along with other modelers as we will see in the following, is that cortical feedback is the main factor in determining response properties of neural modules, already at the level of primary sensory areas, as it is the case for orientation selectivity in visual area V1, a problem we will discuss a little more deeply in sec. 1.4.

1.3 Models of the cell assembly

The nature of the activity in cell assemblies is a matter of debate in the neuroscience community, one of the main points being probably which features of the activity encode information. For some authors, the precise temporal structures of the spike trains are important, and the correlation between spike timing in different cells is also extremely relevant. In this point of view, synchrony between spike from different cells carries information (Phillips & Singer, 1997) for example playing a fundamental role in solving the *binding problem* in visual cortex: cells encoding different features pertaining to the same object need to be tagged in some way, to allow a coherent representation of the object itself. A version of the Hebb's cell assembly theory including *time coding* was for example formulated by Abeles (Abeles, 1991). The opposite point of view denies the significance of the fine temporal structure of neural activity: cells would encode information only through their time averaged *firing rate* (*rate coding*) (Shadlen & Newsome, 1994; Amit, 1995). This was the approach taken by the first attempts at formalization of the cell assembly theory (Marr, 1971). Marr introduced the concept of *auto-associative memory*, that is, a memory that can be addressed through its content. When such a memory is given a corrupted, or partial, version of one of the items encoded in the connection structure, the positive feedback is capable of correcting the errors and yield a less degraded, or even perfect, version

of the encoded item.

Hopfield revived this idea, and suggested a toy neural implementation of an auto-associative memory (Hopfield, 1982), with neural "units" more resembling Ising spins as used in the theory of ferromagnetism. Hopfield work has been very important from the historical point of view, because it put into light the analogy between such neural systems and disordered magnetic systems. A part of the statistical mechanics community, more or less the same that in the same years was studying the prototypes of disordered statistical systems, the "spin glasses" (Mezard, Parisi, & Virasoro, 1988) got then interested into *neural networks*. The Hopfield net had a number of *attractor states* corresponding to the ferromagnetic states in a magnetic system. The attractor states, or memories, of the net were encoded in the synaptic matrix as a superimposition of Hebbian terms. The interference between memory items generated *disorder* in the system, such as it is present in a spin-glass. It became evident that the same techniques could be used in spin glass theory and in attractor neural network theory, for example the replica method, for averaging over quenched (static) disorder. This approach revealed to be very fruitful, leading to a complete statistical mechanics analysis of the Hopfield net (Amit, Gutfreund, & Sompolinsky, 1987), and successively of many related models (Amit, 1989). The main object of the statistical mechanics analysis of this class of models is the *storage capacity* that is, the number of possible items that can be stored before the interference between their encoding in the synaptic matrix disrupts the attractor, or "retrieval" states. An interesting related quantity is the *information capacity*, that is the amount of information that can be stored in the synaptic matrix and then retrieved through the attractor dynamics.

The Hopfield net was by no means a realistic representation of a real neural module. Neurons were represented by binary variables, assuming an active state (1) and a non-active state (-1), and the system is symmetric for permutation of 1 and -1 states. This has no analogue in a real neural system: in particular, in a Hopfield stable configuration each neuron has a probability 0.5 of being active, while only a small fraction of the neurons in a brain module is active at each moment (*sparse* representations). The problem was addressed by many modelers, who considered a modified Hopfield model in which the active state was represented by 1 and the inactive state by 0. In each attractor state, as encoded in the synaptic matrix, only a fraction a (the *sparseness*) of the units were active (Buhmann, Divko, & Schulten, 1989). Another problem with the Hopfield model was the binary representation of the activity of the neurons: Real neurons have a wide and graded dynamic range, and in physiological conditions they stay very far from the maximum activity, or saturation state. A slightly better modelization of neurons is the *threshold-linear* unit: firing rate is modeled as a continuous variable, which is a linear function of the input to the unit, if the input exceeds a threshold, otherwise it is zero. This is inspired by current-to-rate relationships as recorded from real neurons (Mason & Larkman, 1990), and

also as can be computed from detailed Hodgkin-Huxley type conductance models, including some kind of potassium conductance (Wang, 1998).

The statistical mechanics of the attractor network of threshold-linear units was studied by Treves (Treves, 1990; Treves & Rolls, 1991). This kind of models introduced the possibility of quantifying the additional information that could be stored and retrieved due to the graded nature of the response, and possibly of the encoded items.

As an outcome of these improvements of the Hopfield model, it became clear that the analogy between spin glasses and neural nets ("replica symmetry breaking", "frozen states", "slow dynamics", see e.g. Mezard et al., 1988), was more a feature of the Hopfield model itself. Treves (1991a) showed that the threshold-linear attractor net has no spin-glass state, and the replica-symmetric solution is stable in nearly all the parameter space. The analogy was lost, and part of the statistical mechanics community interest as well, but the techniques remained, and they are still a useful tool for investigation of biologically interesting problems, even in this thesis, see cap. 3.

The statistical analysis of models of the Hopfield class gives little hints about the dynamics of a real neural system. It focuses on the *static* properties of the network and it can characterize the equilibrium properties of it. Also, an incorrect interpretation of the Glauber-type dynamics that sometimes is used in simulations of such networks can lead to misleading conclusions about the dynamical properties of real systems, as we will discuss in cap. 4. The variable describing the state of a unit in these models, being discrete or continuous, represent a time-averaged firing rate. To address many issues about dynamics, it is on the other hand important to describe the system in terms of spike timings (this even if what is *encoding information* is the firing rate). Then one needs a model of the process that generates spikes. Probably the simplest possible model is the *integrate-and-fire* neuron: each cell is modeled as a capacitor which is charged by synaptic current and is discharged by a leakage conductance. When the capacitor ("membrane") potential reaches a certain threshold, a spike is emitted and the membrane potential reset. This model has a lot of problems, for example its current-to-rate relationship is highly non-linear, while both real neurons and more detailed conductance models show an approximately threshold-linear current-to-rate relationship. Still, it can be a good model for understanding the behavior of large networks of neurons (Amit & Brunel, 1996). This is the model used in the part of this thesis dealing with dynamics (cap. 4).

1.4 The role of recurrent processing in brain function

The conclusive evidence for attractor states, caused by the positive feedback generated by the recurrent connections, has not been found yet in any region of the brain. Nevertheless, the anatomical facts we have summarized in sec. 1.2 about the preponderance of the recurrent connections, and the computational power of feedback processing evidentiated by the very first theoretical speculations, as we have seen, gave rise to many hypotheses about the possible role of recurrent processing in explaining the function of several brain areas. These functions range from the relatively elementary processing performed in primary visual areas to higher cognitive functions; many of them relate to the hippocampus, and they will be discussed in cap. 2. Self-sustained activity in a recurrent, excitatory, auto-associative memory, network has been invoked as an explanation of *working memory* function. Working memory (Tanaka, 1992) is a temporary, short-term buffer, which stores the data used to carry on the task the subject is performing. One electrophysiological correlate of working memory is *delay activity* in tasks in which some variable has to be remembered in a delay period between the presentation of the variable itself and a behavioral response. A classical example is the experiment by Goldman-Rakic and coworkers (Kojima & Goldman-Rakic, 1984) in which a monkey had to remember the position in its visual field of a simple stimulus (a flash of light) and successively make a saccade in that direction, after a delay period. Simultaneous recording from cells in prefrontal cortex showed the persistent neural activation during the delay activity. The activation was selective, in that cells responded to a specific stimulus position only.

Other remarkable examples of delay activity are the experiments of Miyashita (Miyashita, 1988; Miyashita & Chang, 1988; Sakai & Miyashita, 1991) in which a monkey was shown a series of images (fractal geometric patterns, with no particular meaning); after each image, a delay period followed, then the presentation of a test image. The monkey had to perform a *delay match to sample* task, that is, it had to say, by means of some motor response, whether the test image was the same as the previous one or not. The activity of the neurons in the inferotemporal (IT) cortex, which is a higher visual area, involved in object recognition, was recorded during the delay period, which lasted for up to 16 seconds, as well as during the stimulus and the response phase. Sustained delay activity was found, and notably, the delay activity pattern was not identical to the stimulus activity. This would fit in the attractor network paradigm (Amit, 1995), in which the *internal representation* of an item is determined by the matrix of strengths of the recurrent collaterals, being an attractor of the dynamics induced by those connections. The recurrent connections are "learned", in the Hebbian sense, during the training phase, reflecting the corre-

lations in the activity generated by the stimuli in the area under consideration. The activity configuration elicited by the stimulus can be different from the internal representation, and can be corrected by the attractor dynamics during the delay period. If the monkey is trained with the images presented always in the same sequence, images presented in neighboring positions in the sequence may have correlated internal representations, with correlations decaying with the distance between images in the presentation sequence. It has been suggested (Griniasty, Tsodyks, & Amit, 1993) that this is due to Hebbian learning producing terms in the synaptic matrix proportional to the correlation between the activity generate in the neural module by subsequent stimuli in the presentation sequence. In this case the recurrent network works as auto-associative memory in the proper sense, as synaptic connections reflects the Hebbian traces of previous experiences, that can be recalled by presentation of the same item, or of a similar one. We have examples, from primary sensory areas, in which the recurrent synaptic matrix, and the attractor states generated, do not reflect experience, at least not in a direct way, but they *shape* the way the sensory stimulus is processed. We will only mention here the case of orientation selectivity in the primary visual cortex (V1): in V1 there are cells that selectively respond to bars oriented in a given direction, while they are minimally responsive bars oriented in the orthogonal direction. Orientation selectivity is absent from the response of the previous processing station, that is, the thalamic nucleus LGN. The early theory of orientation selectivity (Hubel & Wiesel, 1962) hypothesized that receptive fields of LGN cells projecting to an orientation selective cell in V1 are aligned, providing a greater excitation in the case of a stimulus aligned with the axis of the receptive fields distribution. According to this theory, the origin of orientation selectivity would be mainly feed-forward. More recently, other theories have been proposed, suggesting that the great deal of recurrent connections in the visual cortex, as we have discussed in sec. 1.2, is the main origin of orientation selectivity. This can be achieved if cells in V1 with similar preferred orientation (PO) are connected with excitatory synapses of strength decreasing with the difference of the POs. In the proper parameter range, this "recurrent V1" network, has attractor states in which only cells with PO pointing in a certain direction are active *even in the absence* of orientation selective inputs (Ben-Yishai, Bar-Or, & Sompolinsky, 1995). Thus, this theory would be capable of accounting for the independence from input contrast which has been observed in V1 (for a review see Sompolinsky & Shapley, 1997), fact which challenges the feed-forward theories of orientation selectivity. In this case, the origin and the function of the recurrent connections matrix is not the memory of some specific item, *they reflect some statistical feature of the environment* which was seen during early development in a critical period, or perhaps even statistical features of the spontaneous activity in the retina before birth (Shatz, 1996), and it is likely that the underlying mechanism is again some form of Hebbian synaptic plasticity. In the hippocampus we will see

a very similar case with the path-integration and head-direction cells systems, also including some form of continuum attractor network, which is one of the objects of this thesis.

1.5 This thesis

The objects of his work are two aspects of the theory of recurrent processing. In cap. 2 and 3 we deal with spatial processing in the hippocampus, analyzing with statistical mechanics methods the *multi-chart path-integrator*, a model for spatial computation in the rodent hippocampus. The model allows to draw parallels between the spatial-computational function of a recurrent network, such as the hippocampus, or some part of it, and information storage function, as it is usually considered in auto-associative memory theories. Cap. 2 provides some background and the rationale for the work, while cap. 3 deals with the technical treatment. Cap. 4 focuses on different aspects of the dynamical of auto-associative memories, that is the time-scale of the attractor dynamics, and the stability of the attractor states. While the study of these dynamics aspects, and particularly of the speed of processing issue, was inspired by considerations about the visual system, which is known to work extremely fast, we think that these are very important problems for the theory of processing in the brain in general. The approach undertaken in the work presented in cap. 4 aims to include more and more element of realism in the models under consideration, trying to single out the most determinant ones for the function of the system, and to figure out how they change its behavior.

Chapter 2

Recurrent processing in the hippocampus

The hippocampus is one of the most widely studied brain structure, for the distinctive position and architecture that makes it particularly easy to identify (and record from) the different regions within it, both *in vivo* and in slice preparations. As a result, a large amount of information is available about the anatomy and physiology of the hippocampus, at all levels. Long-Term Potentiation and Long-Term Depression, candidate phenomena for the role of cellular correlates of learning, were also found first in this brain structure. This was of particular interest considering what is known about the functional role of the hippocampus. In fact, this structure was recognized as involved in learning and memory. Scoville and Milner (1957) described a patient, known as H.M., who showed *anterograde amnesia* after bilateral hippocampal removal. This patient could not form new memories of his recent experiences, but he was still capable of remembering his experiences prior to surgery. Other patients, with similar syndromes, and similar deficits, were found in the following years, for example Zola-Morgan, Squire, and Amaral (1986) described a patient, R.B., whose only brain damage was a bilateral loss of the CA1 field of the hippocampus, assessed by post-mortem analysis. This patient showed anterograde amnesia like H.M. but less severe, suggesting the hypothesis that the memory impairment was related to the fraction of the hippocampus being lesioned.

The hippocampus, or better the *medial temporal region* was found to be particularly involved in *spatial memory*: Damage in this region in monkeys produced deficit in learning object-place memory tasks in which both the object seen and the place where it is seen are to be remembered (for review see Rolls, 1996a) although evidence exists that the surrounding cortical areas are more related to memory (with no particular spatial character) than the hippocampus itself.

A very large amount of evidence about the role played by the hippocampus in

processing and storing spatial information has been collected in rodents. Rats with lesioned hippocampus cannot learn navigational tasks which imply complex spatial computations: one classical example of that is the Morris water maze, in which the animal swims in a pool filled with opaque water and has to learn to find a submerged (and invisible) platform, relying only on remote visual cues. Hippocampally lesioned animals cannot learn this task, while they are able to learn a modified task in which the platform is made visible or is signalled by some local visual cue (Morris, 1981; Morris, Garrud, Rawlins, & O'Keefe, 1982). *In vivo* electrophysiological studies have shown that the main correlate of activity of principal cells in the hippocampus is the animal's position. These cells are therefore called *place cells* and represent one of the most studied phenomena in behavioral neurophysiology. Place cells and related models are one of the main focuses of this thesis, so we present a review of the main theoretical and experimental findings in sec. 2.3.

2.1 Some anatomy

The anatomical investigation of the hippocampus goes back in time to the Golgi studies of Ramon y Cajal (Ramón y Cajal, 1911) and Lorente de No (Lorente de Nó, 1933a, 1933b). Many reviews have been published on this topic, see e.g.(Amaral & Witter, 1995), (Amaral & Johnston, 1998). Here we just sketch the findings of greater importance for the following of this thesis, and we limit ourselves to the anatomy of the rat hippocampus.

A large anatomical unit known as the *hippocampal formation* is defined, formed by six cytoarchitectonically distinct regions: the dentate gyrus, the hippocampus proper, further subdivided in the CA1, CA2, and CA3 subfields, the subiculum, presubiculum, parasubiculum (these three latter known as the *subicular complex*), the entorhinal cortex. The reason to group together these regions is that they are linked by largely unidirectional synaptic projections forming a "loop", or a circuit. This is a feature which is seldom found in connections between neocortical areas which are mostly reciprocal or bidirectional. Other surrounding cortical areas in the temporal lobe, like the perirhinal cortex, does not share the same connectivity property with the rest of the hippocampal formation, having bidirectional connections with the entorhinal cortex.

Another possible criterion in grouping anatomical regions is the layering structure: the hippocampal proper and the dentate gyrus have a characteristic three layer structure, with one layer of principal neurons and fiber layers above and below. The entorhinal cortex has a structure more similar to the standard, six-layers structure of the neocortex, of which it is actually part. Pre- and para-subiculum have a distinctive layering structure, but they are usually included in the multilaminar structures.

In the rat, the hippocampus represents a significant fraction of the whole brain, also due to the relatively undeveloped neocortex. It is an elongated, banana-shaped formation, from the septal nuclei rostrally to the temporal cortex caudally. It is composed of two interlaced, C-shaped sheets, extending through the whole length, the dentate gyrus (DG), and the hippocampus proper or *Cornus Ammonis*, with the subfields (from the closest to DG to the farthest) CA3, CA2, CA1. The structures of the subicular complex and the entorhinal cortex (EC) surround this two sheets system.

The DG and CA sheets, have a similar layer structure. In DG, the principal cells (roughly, the glutamatergic (excitatory) cells which project outside DG) or granule cells are in a middle layer (*granule cell layer*).

In rat DG the number of granule cells has been estimated in a range from 0.6×10^6 to 2.2×10^6. Granule cells dendrites extend perpendicular to the granule cells layer, in what is called the *molecular layer* towards the exterior of the hippocampus. The axons leave from the cell body on the opposite side, in the *polimorphic layer*, and form the *mossy fibers*.

In CA the principal cells are called *pyramidal cells* from the shape of the cell bodies. They are distributed across a middle layer called the *pyramidal cell layer*. Their dendrites extend in both directions perpendicular to the pyramidal cell layer. The apical dendrites point to the inner part of the hippocampus, through the *stratum lucidum*, the *stratum radiatum* and the *stratum lacunosum-moleculare*. The basal dendrites, with a smaller total length point towards the exterior of the hippocampus, in the *stratum oriens*. Estimates of the number of pyramidal cells in CA are about 3.3×10^5 in the CA3 subfield and 4.2×10^5 in the CA1 subfield (see fig. 2.1). Fig. 2.2 shows a schematic diagram of a transverse section of the rat hippocampus, with drawings of the shape of the principal neurons, and the length of the dendritic trees, subdivided among the different layers.

Many other types of cells are present in DG and in CA: most of them are inhibitory GABA-ergic cells. They differ by morphology, connectivity pattern and positioning in the DG and CA layers. They are present in a much smaller number than principal cells (in DG about two orders of magnitude smaller), but they are fundamental in determining the dynamics of the activity in the hippocampus, for example in all the oscillatory collective phenomena that characterize local hippocampal EEG.

The classical basic circuit of the hippocampus is an unidirectional sequence of pathways known as the tri-synaptic circuit (EC $\rightarrow$ DG $\rightarrow$ CA3 $\rightarrow$ CA1, see fig. 2.3) . The entorhinal cortex (EC), which receives, through the perirhinal and post-rhinal cortices, projections from *all* the sensory areas, and can therefore be seen as a sort of highest order sensory area, projects to the dentate gyrus and to CA3 via the *perforant path*, which originates from layer II of EC, pass through (perforates) the subiculum, and terminates on DG in the molecular layer, with a very orderly and topography pre-

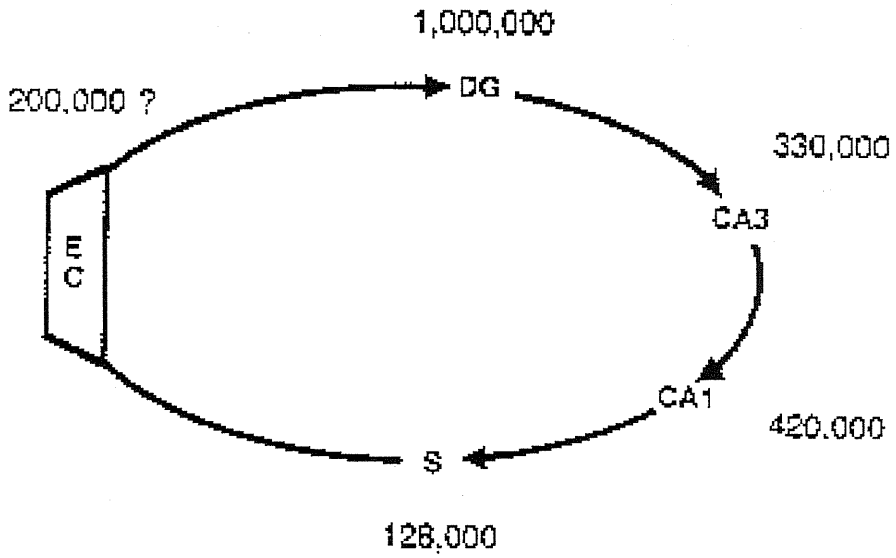

Figure 2.1: Estimates of the number of principal cells in different regions of the hippocampal formation, from Amaral et al. (1990): There are about 1×10^6 granule cells in DG, 3.3×10^5 pyramidal cells in CA3, 4.2×10^5 pyramidal cells in CA1, 1.28×10^5 in the subiculum (inserted in the hippocampal circuit in this figure) and about 2×10^5 cells in the layer II of the entorhinal cortex, projecting to DG.

serving connection pattern, which accounts for $\sim 85\%$ of all the synaptic connections in the molecular layer.

The granule cells, the principal cells in DG, project through the mossy fibers on the proximal dendrites of CA3 principal (pyramidal) cells, which therefore receives two distinct input systems. Mossy fibers show a very small degree of divergence and convergence. It has been estimated that each mossy fiber contacts approximately 14–28 CA3 pyramidal cells, conversely, each pyramidal cell receives about 50 contacts. As a matter of comparison, each cell in CA3 gets about 1.2×10^4 synaptic contacts from the perforant path from EC. On the other hand, the position on the proximal dendrites and the large size of the synapses suggest that their efficacy is particularly strong.

The pyramidal cells of CA3 project to all fields of the hippocampus proper. Most importantly they give rise to an extensive projection (*associational connections*) onto themselves and a big projection (*Schaffer collaterals*). These projections are very divergent: on average each pyramidal CA3 cell contacts as many as 30,000 to 60,000

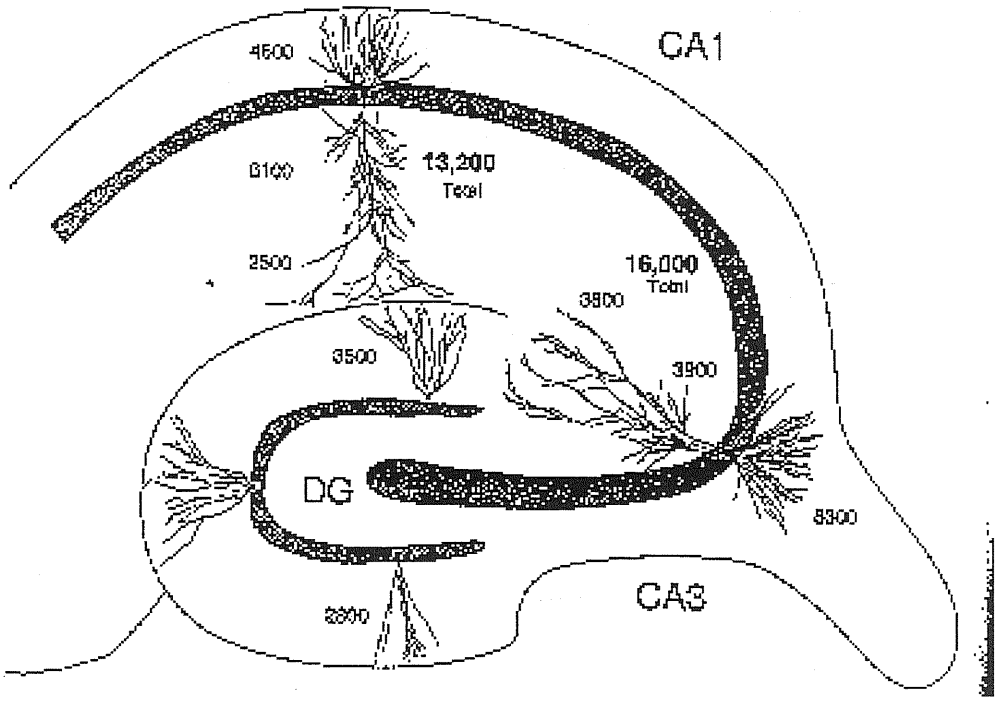

Figure 2.2: A transverse section of the rat hippocampus from Amaral and Johnston (1998), showing the shape of the principal neurons (granule cells in DG, pyramidal neurons in CA). The length of the dendritic trees (in μm) of the granule cells in the molecular layer and of the pyramidal cells in the stratum lacunosum-moleculare, stratum radiatum, and stratum oriens (and total lengths). Assuming a density of 1 synapse/μm, the figures given are also indirect estimates of the synaptic convergence onto hippocampal cells.

other pyramidal cells (among the 3×10^5 pyramidal cells in rat CA3), and with a certain geometric gradient (along the septo-temporal axis of DG one encounters from cells projecting to a narrower region to cells with a more widespread axonal arborization) span much of the extent of the hippocampus.

Due to the associational connections the CA3 region may act as a recurrent network, and for this reason it is seen by many modelers as "the heart" of the hippocampal system. The widespread connection pattern calls for unitary processing of information in the CA3 region, which therefore may form a single network from the functional point of view. CA1, in turn, projects to the deep layers of the entorhinal cortex, via the subiculum.

In the sketch of the hippocampus we have just drawn many connections are miss-

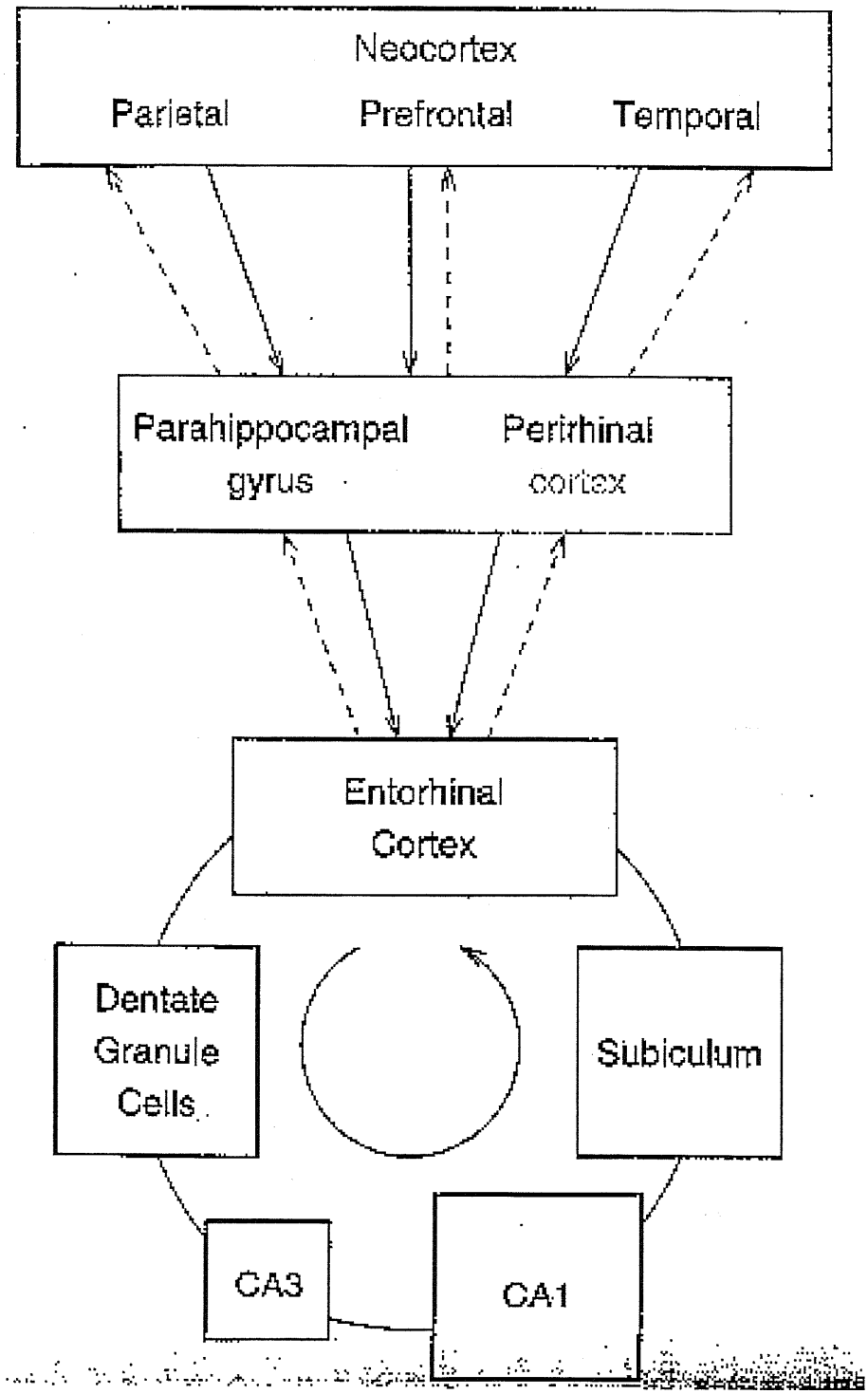

Figure 2.3: A representation of the hippocampal circuit, along with the main critical input-output pathways, from Rolls (1996b).

ing: several secondary projections between hippocampal regions are also present, as well as projections from and to many subcortical areas. We will not discuss these important feature, as they are outside the scope of this work.

2.2 Models of hippocampus in memory

The data available from hippocampal neuroanatomy, on the one side, and the knowledge from behavioral neurophysiology, electrophysiology, lesion studies etc. about the computational role of hippocampus on the other, call for a quantitative theory capable of putting in relation the two kinds of evidence, and of giving a plausible functional explanation of the processing performed in this brain structure. In principle, this approach may provide us with predictions, requirements and constraints about the organization and the function of the underlying neural system. It is possible to analyze in this sense the size of the neural populations, the connectivity structure, the features of the pattern of activity shown by the cells, and to investigate how these aspects of the system, or better the parameters determining their statistics, were tuned by the evolution to develop a system capable of performing the tasks it appears to be dedicated to. Moreover, the hippocampus seems to have rather different functions in different mammalian species, the most studied examples being rodents and primates (humans and non-humans), whereas the anatomy of the system seems to preserve the same gross features – at a statistical level, after the obvious rescaling in population sizes, connectivity etc, passing from the "small" rodent brain to the "large" primate brain. The theory may then try to explain how similar systems can be optimized for different tasks. Part of the work in this thesis addresses specifically this latter point.

2.2.1 Marr's first attempts

The class of models we are talking about have a common origin in the work of David Marr (Marr, 1970, 1971), who was the first to attempt to use statistical tools to draw conclusions about neural systems' function starting from neuroanatomical data. At the time he did his work, Marr had a limited amount of anatomical and neurophysiological evidence available, compared with what we have today, and could not use the powerful tools from statistical physics which have been developed in the last two decades. Nevertheless, most quantitative models of neural functions are inspired by his attempts, more or less openly.

Marr tried to explain the hippocampus (the *archicortex*) as a mainly *memory* system, as opposed to the *neocortex* capable of memorizing and classifying. Marr recognized the possible role of auto-associative memories, or in his terms, *free simple memories*, in correcting and completing neural representations, and also pointed out

that two relevant parameters in determining the performance of the system were the connectivity level and the *sparseness*, that is, how many cells are active in a typical representation. This two hypotheses have been made clearer in more recent years as we will see in the next section

2.2.2 Recent models of the hippocampus as a memory

Recent models of the hippocampus may be divided in two classes. The first one, inspired by neurophysiological findings from primates, tends to privilege the memory role of the hippocampus. The second one stems from the rodent literature, and is mostly concerned with the role of hippocampus in spatial computation. We will discuss both classes, then we will try to formulate a few suggestions about possible links between the two theories.

In primates, hippocampal damage impairs specifically, among others, *object-memory tasks*, that is, tasks in which the monkey had to remember the object seen *and* the place where the object was seen (Smith & Milner, 1981; Parkinson, Murray, & Mishkin, 1988; Gaffan, 1994). This suggests that the hippocampus participates in the memory of spatial context, or better, of the spatial relationships between objects. This kind of memory is formed very fast (one-shot learning), and amounts to "snapshots" of the scenes the animal is presented with. This kind of hypothesis is supported by the fact the the hippocampal formation receives inputs from all the higher sensory areas, in particular from the ventral visual stream (carrying information about objects) and the parietal, or dorsal visual stream, which is more specialized in spatial information (Milner & Goodale, 1995). In fact, the hippocampus itself may be the first area where full convergence of the different sensory streams is achieved, as some kind of topographical and modality segregation is preserved in the immediately previous stages, that is, the perirhinal and parahippocampal cortices, and the entorhinal cortex (Suzuki & Amaral, 1994). The deficit pattern following lesions to perirhinal and parahippocampal cortices is actually different from the one for hippocampal lesions, including *recognition* tasks like delayed non-matching to sample, object retention, and 8-pair concurrent discrimination) (Zola-Morgan, Squire, Amaral, & Suzuki, 1989; Zola-Morgan, Squire, & Ramus, 1994).

Electrophysiological recordings from the hippocampus provides further support for this hypothesis: in monkeys hippocampal cells can respond differently to stimuli presented in different positions in space, or even respond to a particular combination object-location, in a egocentric frame of reference or in different allocentric frames of reference (for a review see Rolls, 1996a).

In the modeling work of Alessandro Treves and Edmund Rolls, (Rolls, 1996b; Treves, Skaggs, & Barnes., 1996) the heart of the hippocampal system, the module in which ultimately all the sensory streams converge, is an auto-associative memory,

identified with the CA3 region for its extensive recurrent collaterals. The hippocampus is the locus of *episodic memory*, that is, the kind of memory which subsume many objects and relation between object in a snapshot of a episode (see also Cohen & Eichenbaum, 1993). The auto-associative memory implemented in CA3 could form spatial snapshot memories as well as episodic memories, with equal efficiency. The formal analysis was carried out based on recurrent networks of threshold-linear neurons (and related architectures) (Treves, 1990; Treves & Rolls, 1991). Memories are represented in the recurrent network as stable (and static) activity configurations characterized by the *sparseness* of the representation, defined as

$$a = \frac{(\sum_i r_i/n)^2}{\sum_i r_i^2/n} \tag{2.1}$$

where r_i are the cell firing rates. Note that if r_i can only assume the values 0 and r, (*binary* activity configurations) the sparseness is equal to the fraction of active units. As in Marr's intuition, the sparseness and the connectivity determine the memory *storage capacity*, that is, the maximum number of different items (configurations of active and inactive neurons) that may be stored in the net before disruption of the attractor retrieval states. The capacity is given by the formula:

$$p_{\max} \sim 0.2 \frac{C}{a \log(1/a)}$$

where C represents the average number of modifiable synapses afferent to a neuron. As it evident from this formula, sparse coding increases the storage capacity of the network, in terms of the number of storable items. The downside to this is that each item carries less and less information as it gets sparser. In fact, the information carried by each item is equal to the Shannon entropy of the corresponding activity configuration. For strictly binary configurations:

$$I_{\text{item}} = -(1-a)\log(1-a) - a \log a \tag{2.2}$$

and a similar amount is obtained for the non-binary configurations observed in the real system. The result of these two contrasting effects is that the total storable information, $I_{\text{tot}} = p_{\max} I_{\text{item}}$ is weakly increasing with sparser codes. By inserting in the formula reasonable estimates for rat CA3, that is, $a \sim 0.02$ and $C \sim 1.2 \times 10^4$, one obtains $p_{\max} \sim 30000$ different items. The corresponding I_{tot} is approximately 0.2-0.3 bits per synapses.

The analysis of (Treves & Rolls, 1991) is carried out for uncorrelated activity configurations, with no particular structure. To apply to the spatial memory case, the analysis needs to be extended, to consider items which are spatially organized. The

work we will present in cap. 3 while referred to a different model of the hippocampus, is an example of how to adapt the analysis to include the geometrical structure of the representations.

This model takes over the idea of the hippocampus as a circuit in which the information flows unidirectionally: CA3 is the key element in the circuit, which is modelized as a cascade of networks with Hebbian plasticity.

In particular the CA3 autoassociative net would receive inputs from the entorhinal cortex (perforant path) and from the dentate gyrus (mossy fibers). The dentate gyrus (which has more cells than the CA subfields), could contribute in orthogonalizing the represntations. From the analysis in (Treves & Rolls, 1992) the hypothesis was raised that the large number of weak synapses in the perforant path are not suited to efficiently store information in the CA3 network, as they are not capable to overcome the effect of the recurrent collaterals in determining the state of the network, while the strong mossy fibers synapses, with little convergence and divergence, are suited to effectively clamp the activity configuration in order to cause efficient storage of the newly learned item. On the other hand, the perforant path would be the main source of input for the retrieval of stored inputs, calling for the need for two input streams to CA3. An alternative theory of storage and retrieval is suggested by the experimental data of Hasselmo and Schnell (1994), who showed that synaptic transmission trough Schaffer collaterals in CA1 is 90% suppressed by acetilcholine (ACh), which leaves unaffected the perforant path. As Schaffer collaterals are formed by the same axons that form the recurrent collaterals in CA3, and as the hippocampus receives a cholinergic input from the septal nuclei, it is possible that this latter input is capable to switch the hippocampus from a "retrieval" mode, in which CA3 is dominated by the collaterals, to a "storage" mode, in which the recurrent dynamics is suppressed by acetilcholine, and the perforant path is able to encode new items. This hypothesis is also supported by the findings that while ACh suppresses synaptic transmission, it enhances LTP in the Schaffer collaterals.

The processing stage right after CA3 is the CA1 subfield, which is the circuit output stage to the entorhinal cortex. CA1 has a larger cell population than CA3 (which seems to be the narrower point in the hippocampal channel), so that it can re-expand the compressed representation in CA3. CA1 is modeled as a feedforward network, or an hetero-associator, with Hebbian synapses, which can contribute in increasing the information content of the hippocampal output. A quantitative analysis (Treves, 1995) suggested that the CA1 operation is optimized if the Schaffer collaterals (projection from CA3 to CA1) share the same synaptic plasticity as the associational collaterals (recurrent projections of CA3 onto itself).

2.3 The place cells system

The picture that arises from the experimental findings in the rat seems to be very different: a great deal of experimental work has addressed the role of the hippocampus in spatial memory *and* computation. Also, the notion of space cognition is very different between rodents and primates: monkeys have a very developed visual system, they can hold a complex representation of their environment, including the location of many objects they see and their spatial relationships, useful for planning of motion and action. Rats have a much more primitive visual system, and their cognition of space is probably very influenced by somatosensory and olfactory (local) cues. The main object of rats' representation of space is therefore likely to be their own position in space. This point of view is supported by the key finding that hippocampal principal cells recorded during behavioral tasks are mainly correlated with the animal's position. This was first observed for CA1 and CA3 pyramidal cells (O'Keefe & Dostrovski, 1971): each cells fires when the animal is in a certain region of space. These cells were therefore called *place cells*, and the regions of space in which they get active were called *place fields*. Place cells are a consistent representation of the animal's position. In fact, it has been shown (Wilson & McNaughton, 1993) that the position of the rat can be reconstructed, by means of a decoding algorithm, from the firing of an ensemble of cells. Many different kinds of information are to be integrated to generate such a representation of the animal's position such as place cells provide. *Kinesthetic* information has to be used, that is, information about the animal's motion, as well as visual (and other sensory) information. Moreover, local and remote visual cues have been proven to affect place cells in different ways. All these different inputs interact in a complex way shaping the behavior of the place cells system.

The place cells system (or the *navigational system* as it is the set of structures and functions allowing the rat to navigate, that is, to find its way in the environment) have been decomposed by many theorists (see e.g. Redish & Touretzky, 1997 in smaller subsystems dealing with the computations involved. The *path integration* system integrate the kinesthetic cues to compute the animal's trajectory (and the current position) with reference to some starting position. It needs to make use of information about direction of motion, roughly equivalent to head direction, which is what the *head direction subsystem* computes. The result of this computation must be corrected and completed with information about the cues present in the environment (*local view* subsystem) to yield a coherent representation of place (*place code*). The association between these subsystems and specific brain structures is undefined. Although several hypothesis have been formulated, it is well possible that each logic subsystem include many of the same brain structure, among the ones which take part in the spatial computation system. We now give a brief overview of the main experimental findings and theoretical suggestions about these subsystems, emphasizing on the possible role

of recurrent processing.

2.3.1 Head direction cells

To be able to compute the trajectory, the navigational system has to know about the direction the rat is heading to. It needs a compass system, which seems actually to exist in the brain in the form of *direction cells*, that is cells tuned to the orientation of the animal in space. Such cells have been found in three distinct subcortical brain areas: in the postsubiculum (PoS, or dorsal presubiculum, part of the hippocampal formation) (Ranck, 1984; Taube, Muller, & Ranck, 1990), in the anterior dorsal nucleus of the thalamus (AD) (Blair & Sharp, 1995; Knierim, Kudrimoti, & McNaughton, 1995; Taube & Muller, 1995), and in the lateral mammillary nuclei (LMN) (Leonhard, Stackman, & Taube, 1996). This three structures are interconnected and therefore they form a unitary system. A few suggestions about the functions of this system are given by the findings that lesions to AD disrupt head direction related firing in PoS (Goodridge & Taube, 1994), while postsubiculum lesion do not disrupt head direction related firing in AD (Goodridge & Taube, 1994; Taube, Goodridge, Golob, Dudchenko, & Stackman, 1996). The compass system remains coherent, even when the rat is placed in a different environment, and the differences between preferred orientation of cells remain constant. In the three brain regions, the head direction systems works in the dark as well (Taube et al., 1996), presumably driven by vestibular inputs (after vestibular lesion head direction selective firing is lost (Stackman & Taube, 1997)) In the three brain regions, the tuning curves of head direction cells are about 100° wide.

The fact that the head direction system is capable of maintaining direction related activity even in the dark, and when the animal is not moving (therefore virtually in absence of any input) suggested to theorists including attractor networks in their models of the head direction system. In fact, a system similar to what has been invoked to explain orientation selectivity in the primary visual cortex (see sec. 1.4), and to the model of place cells we analyze in cap. 3, may account for head direction selectivity: with some variation among different models, cells have excitatory reciprocal connections with cells with similar preferred orientation (so that the network is topologically arranged on a ring), and all the cells are inhibited by a non-structured population, which controls the average activity. Such a net, in the appropriate parameter regime, exhibits stable direction selective firing configurations (due to the attractor dynamics) even in the absence of any input. The vestibular inputs (related to angular velocity) have then to update the representation, by driving it clockwise or counterclockwise. This can be accomplished if the angular velocity input activates another set of synapses, unidirectional (non reciprocal) and therefore capable of displacing the activity configuration in a moving wave-like fashion. In the version by

Skaggs, Knierim, Kudrimoti, and McNaughton (1995) the head direction cells project to left and right *rotation cells*, which are also driven by vestibular inputs. The rotation cells in turn project to the head direction cells left or right of the ones they receive input from. When the rat moves, the rotation cells are activated and they cause the attractor representation to displace. Zhang (1996) proposed that head direction cells are subdivided in two populations, labeled as left and right, each one receiving vestibular input in correspondence with clockwise and anti-clockwise head rotations. The two populations are in turn asymmetrically connected to the left and right neighbors. When there is no vestibular input, the two populations are equally activated, and the situation is equivalent to having symmetric connections only, so that one has stable configurations. The vestibular input activates differentially the two populations revealing the asymmetry in the connections, and causing the activity configuration to move. Redish and Touretzky (1996) proposed a variant of this model, in which two attractor networks are considered, corresponding to PoS and AD, and the asymmetric connections are multiplicatively modulated by the angular velocity, in order to allow the network to track angular displacements at different angular velocities. A model of head direction cells also has to include a mechanism by which visual cues may "reset" the system: an excitatory selective input strong enough to move the activity configuration to point in the desired direction. The effect of the excitatory input may be to make the activity configuration slide to the new configuration, or to make it jump abruptly, depending on the angular distance between the old and the new configuration, as it was pointed out in the visual orientation selectivity theory context by Ben-Yishai et al. (1995).

2.3.2 Path integration

Path integration, or *dead reckoning*, is defined as the ability to come back to the starting point after some, even complex and tortuous, exploratory trajectory, in the absence of any visual landmark. More generally, path integration is the ability to integrate the information about self motion (body motion and head direction) to keep track of the vectorial (in two dimensions) displacement from the starting point. This system is integrated with the place field system, being one of its main logical inputs. The computations the path integration system has to perform are in some sense similar to the ones performed by the head direction system. *Velocity* inputs are to be integrated in time, in this case as well as in the head direction case, to track a trajectory, and if necessary corrected by visual inputs, this time in a two-dimensional space. The representation of position must be stable even in the absence of any input, and this suggests a role for attractor networks in some stage of the system. The identification of the different components of the path integration system with brain regions is still a matter of debate, but a number of hypotheses have been formulated,

all assuming the existence of an attractor system with symmetric connections, which keep the representation stable, and asymmetric (or *offset*) connections set which are in charge of moving it according to the velocity input.

The details change from one model to another: Samsonovich and McNaughton (1997) proposed that the attractor stage of the path integrator resides in the hippocampus, more specifically in the CA3 stage, that is, it coincides with the place cells system (the **P** stage, in the authors' terms). The **P** stage has a set of offset connections to and from the **I** stage, which contains a representation of place $\times$ velocity, that is, it has cells responding to particular combinations of place and motion. This stage may be identified with the subiculum, post-subiculum, and para-subiculum, where cells were found with *directional* place fields (the cells get activated only when the rat crosses the place filed in a particular direction), even in the case in which hippocampal place cells were non-directional (Taube, 1995; Sharp & Green, 1994). The **I** stage receives input from the head direction system modeled in the same way as in Skaggs et al. (1995), This model was criticized for the very complex wiring scheme required, and the complexity is even increased by the fact that in the hippocampus (the **P** stage) multiple arrangements of place cells are seen in different environments (see the next section), so that multiple wirings must exist between the **I** and the **P** stages. Samsonovich and McNaughton (1997) assume that the path integration connectivity is pre-wired, probably "learned" in some critical period during development (see also next section). The **P** stage also receives visual inputs from the **V** stage, that is, the cortex through the entorhinal cortex relay. The visual input can correct and update the representation in the **P** stage.

Sharp (1997) proposed in contrast that the path integration system be located in the subiculum, as place cells in the subiculum show approximately the same arrangement in different environments. Redish and Touretzky (1997) proposed that a loop between the parasubiculum, the subiculum and the superficial layers of the entorhinal cortex be responsible for path integration.

2.3.3 Recurrent processing models for the place field system

We have seen already that recurrent processing plays an important role in the models of the spatial navigation system. In the head direction and in the path integration systems auto-associative memories were used as a *buffer* to temporarily store the current direction/position, capable of updating it by integrating a velocity signal. The strongest arguments in favor of the auto-associative memory hypothesis are the fact that place fields are still observed when the rat is moving in the dark (Quirk, Muller, & Kubie, 1990), and the fact that if the light is lit off during the recording session, the place fields arrangement stays unchanged, even though more cells show place fields in the light than in the dark and place fields are generally more reliable in

the light than in the dark (Markus, Barnes, McNaughton, Gladden, & Skaggs, 1994). Thus, place cells can be generated in the absence of visual cues, updated by path integration only, and maintained in an input independent way.

Moreover, place fields may be seen from the initial entry in an environment (Hill, 1978), even though it may take 10-30 minutes for them to stabilize (Wilson & McNaughton, 1994). This was explained by assuming some kind of pre-wired connectivity structure: Zipser (1986) and Sharp (1991) supposed that the local view system is largely pre-formed. In contrast, Samsonovich and McNaughton (1997) assumed that the attractor structure responsible for path integration, and probably located in the hippocampus, is pre-wired. This model relies on the extensive recurrent connectivity which is observed at least in the CA3 subfield of the hippocampus. Synaptic connections between two cells encode *the distance* between the place field centers of the two cells, that is

$$J_{ij} \propto F(\vec{\eta}_i - \vec{\eta}_j) \tag{2.3}$$

where J_{ij} is the synaptic strength between cell i and cell j, which have place fields centered respectively in η_i and η_j, and $F(x)$ is a monotone decreasing function of its argument. The system is also thought to have non-specific inhibition, keeping the average activity level limited. As we will see in cap. 3, in a wide parameter regime this system has *place-related* stable states, *activity peaks*, or *activity packets*, as they were termed by Samsonovich and McNaughton (1997), that is states in which activity is confined to cells with place fields nearby (as it is actually the case for hippocampal place cells, see fig. 2.4). The activity packet can be moved by the path integration machinery as we have seen in sec. 2.3.2.

Muller and Stead (1996) formulated a similar model, in which recurrent connectivity was not taken as pre-wired, but was shaped by hebbian learning during experience, and reflects information about the possible routes between a start and a goal point in the environment (see also Blum & Abbott, 1996).

Another experimental fact proposed a fundamental ingredient for the model of Samsonovich and McNaughton (1997): in different environments completely different arrangements of place fields can be seen, that is the spatial relationship between the place fields centers of two given cells can be completely altered (see e.g. Bostock, Muller, & Ranck, 1991 and Markus et al., 1994). The place field arrangement can be changed partially by changing the behavioral task the animal has to perform (Markus et al., 1995), or even completely, after returning in a previously visited environment after distraction, in aged animals (Barnes, Suster, Shen, & McNaughton, 1997). If the place cells arrangement are indeed in large part determined by the recurrent connections in some auto-associative network and these are to be determined previous to experience, then *multiple* place fields arrangement must be pre-wired and co-exist at the same time in the synaptic matrix. This can be achieved in a way very

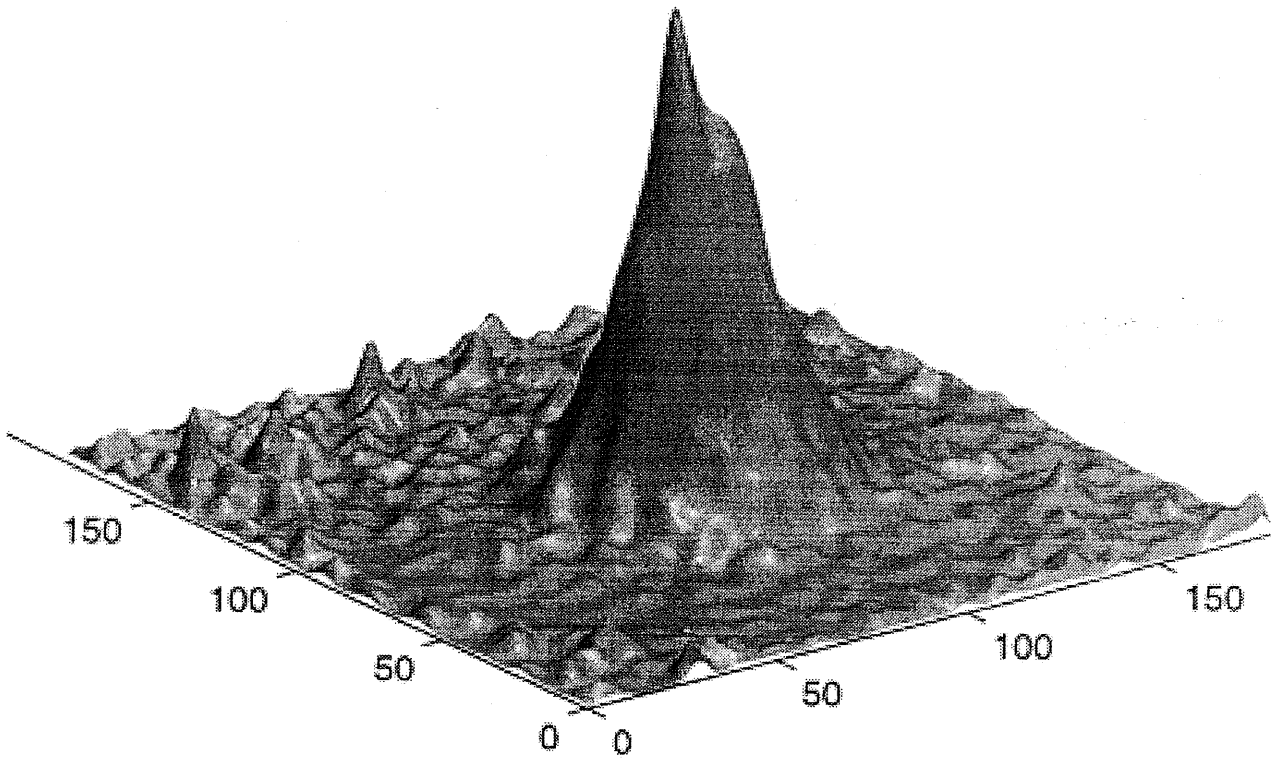

Figure 2.4: The *activity packet* as it was reconstructed by Samsonovich and McNaughton (1997) from the experimental data of Wilson and McNaughton (1993). The firing rate of about 100 hippocampal cells is plotted here, versus the position of the place field centers of the cells. The plot is the result of averaging across many snapshot of ensemble activity, with the space coordinates realigned in order to have the rat position always at the point (100,100).

similar to what is done in Hopfield-type models of auto-associative memory: different *charts*, as Samsonovich and McNaughton named the arrangements of place fields, are *superimposed* in the synaptic matrix, in a linear fashion. Many terms like the one in eq. 2.3 are summed up to form the synaptic matrix:

$$J_{ij} \propto \frac{1}{N} \sum_{\mu=1}^{p} F(\vec{\eta}_i^{\mu} - \vec{\eta}_j^{\mu}) \tag{2.4}$$

As in the case of Hopfield-type auto-associative memories, the superimposition of many, uncorrelated items (charts in this case) in the synaptic matrix cause interference, and as the number of stored items increases, interference eventually disrupts

the attractor states related to each of the items, in this case the chart states.

The analysis of the multiple chart-attractor network is one of the objects of this thesis, and the mathematical treatment is presented in cap. 3. In the following section we present the motivations and state the results in a non-technical way, referring the reader to the next chapter for technical details.

2.4 The formal analysis of hippocampal models: a possible synthesis

Up to here we saw two apparently very different theories of the hippocampal function, one "episodic memory" theory mainly inspired by the (human and non-human) primates literature, but with related experimental findings in the rat as well (see e.g. Eichenbaum, Kuperstein, Fagan, and Nagode (1987)), and one "spatial navigation" theory, based mainly on the neurophysiological evidences in rodents.. The main point of contact between these two points of view is the importance attributed to the recurrent connections present in the hippocampus, especially CA3. Mammalian species, from rodents to primates, seem to share the same anatomical features, such as the associational CA3 pathway, although there is not a comparable amount of quantitative evidence for the monkey as there is for the rat.

The role auto-associative memories play in these two theories is nevertheless very different: in the episodic memory theory, CA3 collects multi-modal sensory input and can form on the fly a coherent representation of them, containing also information about the spatial relationship between objects in the scene. Learning-related modifications are supposed to occur on the input pathways synapses *and* on the recurrent synapses, according to an Hebbian rule. Stored items can be later retrieved, maybe to be transferred in long-term storage sites in the neocortex as Marr (1971) proposed, for example during sleep. *Re-activation* of correlation between neural activity recorded from two cells during behavior was observed during sharp-wave sleep (Wilson & McNaughton, 1994), as a hint that activity configurations which occurred during behavior are being retrieved, perhaps due to some kind of attractor dynamics.

The performance of the auto-associative memory network is a subject of theoretical analysis: the number of stored items, the quality of retrieval, the total information stored in the network, the time required for the dynamics to approach the attractor states are all quantities of relevance for the function of the net, and ones that can be addressed by analytical and/or numerical means.

In the formal models, storage of patterns is modeled as Hebbian modifications that build up to form the synaptic connection matrix. If η_i^μ (for example, we may assume that the ηs can be 0 or 1) is the the activity level of unit i in the μ-th of p

stored patterns, the synaptic connection matrix reads, in the simpler models:

$$J_{ij} = \sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu \tag{2.5}$$

As we have anticipated, the main parameters determining the storage efficiency of the network are the sparseness (as defined for example by eq. 2.1) and the connectivity structure. The connectivity structures can be characterized be their degree of *dilution*: a *fully connected* network, in which every unit is connected to all the others, and an *extremely diluted* network, in which each unit receives connection from a vanishing fraction of the others, represent the two extremes of a range of possibilities. They also are two analytically addressable cases: more precisely, we define the extremely diluted network as the network in which the number of connections C a unit receives on average is

$$C \sim \log N$$

where N is the number of units in the network, when $N \to \infty$ (Derrida, Gardner, & Zippelius, 1987). If we define the *tree of ancestors* for unit i as the set of unit sending connections to unit i, plus the set of units sending connections to this latter set, and so on recursively, the trees of ancestors for units i and j have null intersection with probability 1 in this limit. Thus, the activity of unit i and unit j are uncorrelated, and the effect of interference reduces to a trivial Gaussian term, which makes this limit of very easy (and exact) solution. It is a little incorrect to define the network in this limit as a recurrent network, though: there is no feedback loop closing here, so virtually no "recurrent connections".

On the contrary, the fully connected network is dominated by feedback and this makes the analysis more difficult (see e.g. Amit et al., 1987, Treves, 1990), requiring the replica trick (or related tools) to average over the static noise due to the interference by the other stored items, when the network is in an attractor state correlated to one of the items. In the threshold linear units case, anyway, with sparse coding (that is, only a small fraction of the units are active) it is shown (Treves, 1991a) that the *replica symmetric* solution is stable and it is the valid one (we do not enter in the details of the replica theory, the reference book for these subjects is the one by Mezard et al. (1988)). In this case the *mean field equations* describing the stable states are very similar to the extremely diluted case (Treves & Rolls, 1991), with a Gaussian noise term, which gets normalized by the effect of the feedback, just like the transfer function gain does.

In this framework it is possible in both cases to calculate the network storage capacity, that is the number of items $p_{\max}$ which can be stored in the synaptic matrix,

preserving the retrieval states (attractor states correlated with one of the items). For the extremely diluted network, this is given in first approximation by the scaling law

$$p_{\max} \sim 0.2 \frac{C}{a \ln(1/a)}. \tag{2.6}$$

The fully connected network has in general a smaller storage capacity, but for very sparse coding, its storage capacity approaches the extremely diluted network capacity (see fig. 2.5). This means that sparse coding in some sense reduces the effect of feedback, as one could intuitively conclude from the observation that sparse coding makes less synapses active at a given time.

As we can see from eq. 2.6, the storage capacity diverges when $a \to 0$ that is, for extremely sparse coding. This does not mean that the efficiency of the network as a information storage device becomes infinitely good: as we already mentioned, the total storable information $I_{\text{tot}} = p_{\max} I_{\text{item}}$ is a slowly changing function of a. The total amount of information stored can be evaluated in a fraction of bits per synapses (Tsodyks & Feigelman, 1988), see fig. 2.6

2.4.1 Spatial information encoding in an auto-associative memory

In the place cells/path integration model by Samsonovich and McNaughton, the recurrent synaptic connections are pre-wired, perhaps during some kind of critical period in the early development. They are not supposed to change during exploration, not even in new environments. On the other hand, they provide a geometric structure, "sheets" where pieces of knowledge about different environment can be attached, constructing a "map" of the environment itself. The "map" also has path integration capabilities, as it is connected to a more complex apparatus, as we have seen. Nevertheless, the two models are based on the same mechanism: several items are stored in the synaptic matrix by linear superimposition and, as long as interference is not too strong, the network has attractor states related to each of the items. The multi-chart path integrator model is a model of spatial information processing, and unlike other models of its kind, it makes full use of the information storage capabilities of a recurrent matrix of Hebbian modifiable connections, as auto-associative memory models do. This is a interesting starting point for a comparative analysis of the function of recurrent processing in two contexts that were previously considered as unrelated, by quantitative and analytical means. We are interested in studying the performance, the storage capacity (how many *charts* we can store before interference disrupts the function of the network), the information capacity (as later defined) of the multi-chart path-integrator, and how these quantities vary with the relevant network parameters.

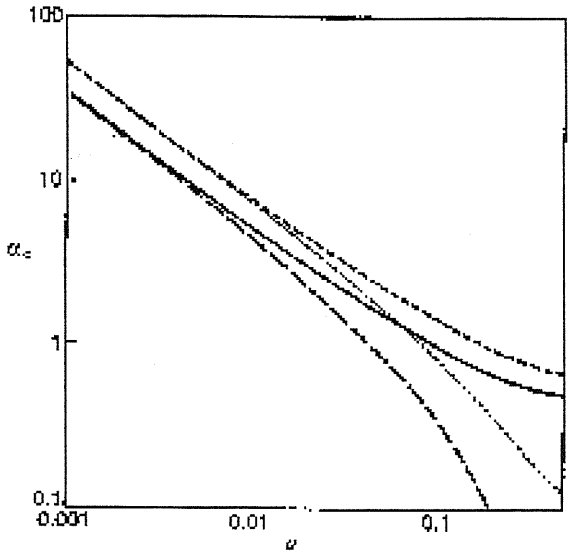

Figure 2.5: The maximal storage capacity for the Hopfield type auto-associator, as it was calculated by Treves and Rolls (1991). Here $\alpha_{\max} = p_{\max}/C$ is plotted here as a function of the sparseness a. The four lines are referred (from top to bottom in the right part of the plot) to a fully connected network with exponential patterns , binary patterns and to an extremely diluted network with exponential patterns and with binary patterns.

If we find that the relevant parameters are the same as for the Hopfield-type auto-associator (that is, again, the connectivity structure and the sparseness) and the way network performance depends on them is similar, then we can hypothesize that a system (as it could be CA3, or even a broader portion of the hippocampus and/or cortex) which is optimally tuned to operate as an Hopfield auto-associator, is also optimally tuned to work as a multi-chart auto-associator.

2.4.2 The continuum (space) nature of information does not affect information encoding efficiency

The statistical mechanics analysis of the multi-chart auto-associator was published in Battaglia and Treves (1998a), and it is presented in cap. 3.

First, we have to define sparseness: a close analog we could think of is the size of the activity peak (the space related attractor configuration). If we study the case in which the dynamics of the network evolves towards stable states with no external

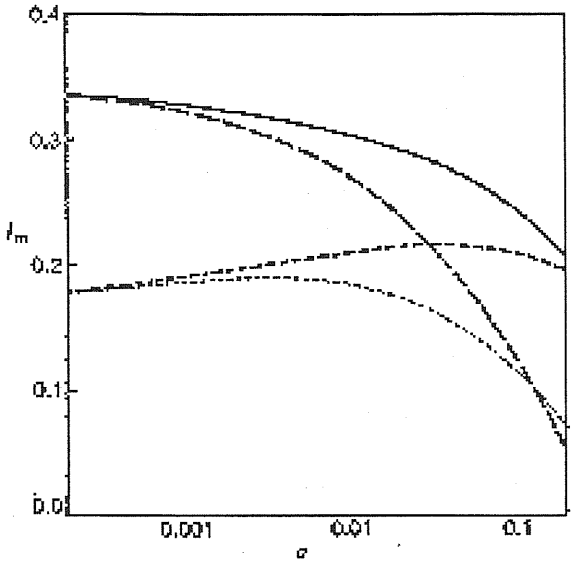

Figure 2.6: The maximal storable information for the Hopfield type auto-associator as it was calculated by Treves and Rolls (1991). The maximal storable information per synapse (in bits) is plotted here, as a function of the sparseness a. Lines as in fig. 2.5

inputs, the size of the place field can only depend on the spread of the connections, defined by the function $F(x)$ in eq. 2.4. Indeed, when all the other parameters are equal, the size of the place field must be proportional to the spread of place fields. we can therefore define the *map sparsity* as

$$a_m = k_d \frac{\rho^d}{|M|} \tag{2.7}$$

(see also eq. 3.48) where ρ is the typical range of the connections in the synaptic matrix, d is the dimensionality of the chart (so for the place fields on a place case it is $d = 2$). k_d is a proportionality factor, assumed to be independent of ρ between the "effective" size of the place fields and the range of the connections. $|M|$ is the size of the environment surface. To carry on the equivalence with the storage of "patterns" i.e. simple activity configurations, and assign an activity level to each unit, we may imagine subdividing the surface the chart is defined on, in portions occupying a fraction a_m of the whole chart surface. To each portion we can then associate an activity pattern, by assigning an elevated activity level to the units having place field centers within the region, and zero activity to the remaining units. This pattern

would be roughly equivalent to an "activity packet" in the sense of Samsonovich and McNaughton (1997). A number $1/a_m$ of such patterns, or activity packets, can completely cover the chart surface. We know from eq. 2.6 that we can store in an auto-associative memory a number $p_{max} = k_p C/(a_m \ln(1/a_m))$ of patterns of sparsity a_m in an auto-associative memory. (k_p is a proportionality factor. We could assume that the amount of interference noise a chart causes in a network is equivalent to the amount caused by $1/a_m$ patterns (covering the chart surface). If this was the case, then we would find a storage capacity *in terms of charts*, for the extremely diluted network:

$$c_{\max} = k\frac{C}{\ln(1/a_m)}. \tag{2.8}$$

Although the analogy is extremely lousy – it does not keep in consideration the fact that there are many more possible attractor states (actually, a continuum) than the set of patterns (or of activity packets) we considered – the formal treatment we developed and we will present in cap. 3, proves formula (2.8) to hold, at least to the leading order in $1/a_m$. The mathematical machinery yields the exact values for the k_d and the k proportionality factors, appearing in eqs. 2.7 and 2.8. k_d is about 3.6 for the 2-dimensional model and 4.5 for the uni-dimensional model. k is about 0.5. Thus, we can read eq. 2.8 as though each chart is substituted, for the sake of capacity calculations, by a set of $1/a_m$ patterns corresponding to activity packet with a radius of $(k_d)^{1/d}\rho$ in the chart space. k_d gives us the *effective* size of the activity packet. Indeed this effective size is comparable to what is obtained directly by computing the average shape of the activity packet (see fig. 3.4). It is also interesting to compare the average activity packet shape (averaged over the interference, or static, noise) in fig. 3.4 with the shape calculated by experimental data in fig. 2.4. It is to be noted that, if we assume that the place field centers of the units are uniformly distributed over the environment, the size of the activity packet coincides with the size of the typical place filed.

If we substitute in eq. 2.8 the figures we used in sec. 2.2.2 ($a_m \sim 0.02$ and $C \sim 1.2 \times 10^4$) we obtain a value of ~ 500 for $c_{\max}$. The estimate can be further refined by recognizing that not all the cells take part in all the charts. Our capacity calculation can be modified to take this new fact into account as follows: if a fraction a_c of cells has a place field in an environment, and each chart has a map sparseness a_m, then we can imagine constructing a larger chart by tiling $1/a_c$ charts, one in which all the cells have a place field in. The bigger chart has a sparsity $a_m a_c$. We can apply eq. 2.8 and find that up to

$$c^f_{\max} = k\frac{C}{\ln(1/(a_m a_c))}. \tag{2.9}$$

"big" charts, or

$$c^s_{\max} = k\frac{C}{a_c \ln(1/(a_m a_c))}. \tag{2.10}$$

"sparse" charts can be stored.

We can evaluate the fraction of cells having a place field in a given environment as $\sim 20\%$. If a typical place field spans, say, 1/20 of the environment, we get a value of $\sim$ 800 storable charts. From eqs. 2.8 and 2.10 it turns out that the number of storable charts *decreases* when a_m, that is, the size of the activity packet, decreases. We must bear in mind though, that a chart with smaller activity packets (or smaller place fields) is equivalent to an larger number of more distributed charts. Thus, the analogy with the Hopfield-type pattern auto-associator is still valid.

2.4.3 A definition of encoded space information

In the model of Samsonovich and McNaughton (1997) the multi-chart auto-associator is only marginally seen as an auto-associative *memory*: the recurrent connections are pre-wired, and they do not underlie any experience dependent change. The synaptic matrix does not encode any information about the experienced world. Nevertheless, it is still interesting to quantify the efficiency of the system in terms of stored (and retrievable) information. The first question is: What is the network storing information about? The function of the multi-chart auto-associator, the **P** stage in the spatial navigation system of (Samsonovich & McNaughton, 1997), is to receive inputs from a visual sensory stage and to organize them spatially, in a map, and to give the map path integration capabilities. The spatial organization of many different environments (or even different logical representations of the same environment), can be stored *in the feed-forward connections between the* **V** *stage and the* **P** *stage*, thanks to the multi-chart mechanism. The feature equivalent to the quality of retrieval in Hopfield-type auto-associators in this class of networks is therefore the *precision of the spatial localization* of the visual (or sensory) components of a scene.

The spatial navigation system locates sensory cues through an association between representations in the **V** stage and cells in the **P** stage. The precision of the localization of sensory cues is therefore the precision of the localization of the *place field centers* of the cells in the multi-chart auto-associator. Let us consider a very extreme situation in which the local view (containing local and distal landmarks etc.) A is seen from the position x_A. In the cortical **V** stage A is represented by the activation of the set of neurons q_A, which project *uniquely* onto the cell c_A. c_A is therefore the "spatial grandmother cell" of A. c_A has a place field centered in x_A. We assume here that the spatial selectivity properties of c_A are only determined by the recurrent collateral in the multi-chart stage **P**. The precision with which A can be associated

to the position x_A is therefore determined by the precision with which the place field center of c_A can be evaluated from the activity of the cell itself.

There are two sources of noise limiting this precision: there is the static noise caused by the traces in the synaptic matrix of the non-active charts, which will cause the activity packet to enlarge, and its profile to get fuzzy, then there is the effect of the irregular firing of the cells. Firing of hippocampal place cells is extremely irregular, probably well approximated by a Poisson process. This implies that there is an uncertainty associated with any estimate of the mean firing rate (which in our assumption is the variable carrying information) derived from a sample of activity in a limited time window. As we are now interested in *storage* of information in the synaptic matrix, we will only consider the first source of noise, leaving to the discussion section a few remarks about how the two sources relate to each other.

The activity of the cell can be monitored while the rat is wandering in the environment, covering ideally the whole surface. As the activity packet is displacing, the activity can be sampled infinitely many times with the rat in different locations. The information about the cell's place field center does not diverge while we collect more and more samples, though, since the activity values measured in nearby locations will be correlated, so they not yield independent information. In mathematical terms, we want to calculate the limit

$$I_{\text{chart}} = \lim_{S\to\infty} \frac{1}{C} I(\vec{x_A^{\mu}}, \{V_i^{\mu(k)}\}_{k=1\ldots S}) \tag{2.11}$$

of the mutual information between the place field of cell c_A in the μ-th chart, as it encoded in the synaptic structure, and a set of activity samples $\{V_A^{\mu(k)}\}$. As it is normalized in eq. 2.11, I_{chart} is the information per afferent synapse to cell c_A. To yield the total stored information per synapse, this quantity must be multiplied by the number of stored charts $p_{\max}$.

This information will depend in general on the size of the activity packet: the smaller the packet, the smaller the place field and the better the precision we can estimate the place field center with, and on the amount of static noise. The full calculation of I_{chart}, as defined in eq. 2.11, is not feasible analytically. We have developed an approximated procedure to calculate I_{chart} based on the idea of calculating the *information correlation distance* l_I defined as the distance between two positions of the rat such that the correspondgin activity packets give independent information. We calculate the quantity I_1^{chart}, that is, the information about the place field center we get from a single sample of activity, then we calculate the quantity $I_2^{\text{chart}}(l)$ which is the information from two samples of activity collected with the rat in two positions at a distance l. For $l = 0$ we will have $I_2^{\text{chart}}(l) = I_1^{\text{chart}}$ as the two samples will be identical. When l is very large, we will have $I_2^{\text{chart}}(l) = 2I_1^{\text{chart}}$. l_I is defined as the distance l for which $I_2^{\text{chart}}(l) = (2-\epsilon)I_1^{\text{chart}}$ according to some criterion ϵ.

Indeed, we can imagine to sample the activity values while the rat is spanning a lattice of position with lattice distance l_I so that the information we get from the samplings will be independent. This will yield a value for the information

$$I_{\text{chart}} \sim I_1 \frac{|M|}{l_I^d}, \tag{2.12}$$

This value must be multiplied by the maximum number of storable pattern $p_{\max}$, to obtain the total storable information per synapse. As even this approximate method involves very lengthy numerical computations, we were only able to evaluate this quantity for 1-dimensional charts. In this case we get a result of a fraction of bit per synapse, comparable to the maximal storable information for a Hopfield-type auto-associator.

2.4.4 Simulations of the multi-chart network

To give a flavor of the function of the multi-chart auto-associator, we present here the result of simulations of a fully connected network of 900 neurons. Due to the small size of the network, it was not possible to estimate with some precision the network storage capacity, so we only show here some stable activity configurations or attractors of the dynamics. The simulations were performed as follows: p charts were created by assigning to the i-th unit a place-field center η_i^μ on a 30x30 lattice of side 1, for the μ-th chart. This procedure was chosen instead of just choosing the place field centers randomly on a surface, because any inhomogeneity in the distribution of place field centers disrupts the continuum of attractors on the chart surface. Only attractors relative to activity packets concentred in chart regions with upwards fluctuations of the place field centers density are found to be stable, as it was pointed out by Tsodyks and Sejnowski (1994).

The connection matrix was formed as in eq. 2.4. The function $F(x)$ in eq. 2.4 was

$$F(x) = \exp\left(-\frac{x}{2\sigma^2}\right) \tag{2.13}$$

Neurons were represented by threshold linear-units, with activity given by

$$V_i = g[h_i - \theta]^+, \tag{2.14}$$

where $[x]^+ = \max(x, 0)$ and h_i is the synaptic input

$$h_i = \sum_{j \neq i} J_{ij} V_j \tag{2.15}$$

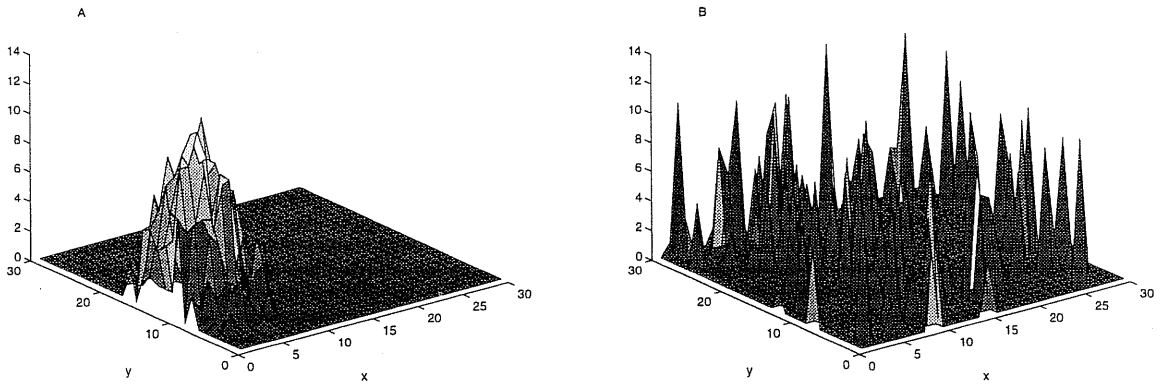

Figure 2.7: Simulated activity packets from the multi-chart auto-associator. With 900 cells, and a range of connections $\rho = 0.2$ one would expect $p_{\max} \sim 9$. Here we are in the low memory loading regime, as $p = 3$. A: The activity packet in the retrieved chart is a very clean, space related configuration. B: the activity configuration, plotted against the place filed center coordinates in a different chart. The configuration looks random, with no spatial selectivity.

The network was initialized in a configuration in which only units with place fields near a given point *in a given chart* were active. A Glauber type dynamics were simulated and the stable states were considered. The attractor states were of the activity packet form when plotted versus the (x, y) coordinates of the place field centers, in the chart being retrieved, see fig. 2.7A. The activity, plotted versus the place field centers coordinates in a different chart shows no spatial selectivity, see fig. 2.7B.

Chapter 3

Statistical mechanics analysis of the Multi-chart network

3.1 Introduction

In this chapter we present the mathematical analysis of the multi-chart auto-associator, as it was published in Battaglia and Treves (1998a) In sec. 3.2 the case of a single attractor chart stored is studied, then in sec. 3.3 the case of multiple stored charts is analyzed and the storage capacity is found, first for a simplified model and then for a more complex model which makes it possible to address the issue of sparsity of representations. In sec. 3.4 the storable information in a multi-chart network is calculated, making more precise the sense in which such a network is a store of information, and completing the parallel with auto-associative memories.

3.2 The single map network

As a first step, we consider the case of a single attractor map encoded in the synaptic structure, as it was proposed in (Tsodyks & Sejnowski, 1994). We focus here on the shape and properties of the attractor states, as a useful comparison for the following treatment of the multiple charts case.

The neurons are modelled as threshold linear units, with firing rate:

$$V_i = g[h_i - \theta]^+ = g(h_i - \theta)\Theta(h_i - \theta) \tag{3.1}$$

i.e. equal to zero if the content of the square brackets is negative. h represents the synaptic input current, coming from other cells in the same module, θ is a firing threshold, which may incorporate the effect of a subtractive inhibitory input, common

to all the cells, as it will be illustrated later on. The connectivity within the module is shaped by the selectivity of the units. If $\mathbf{r_i}$ is the position of the center of the place field of the i-th cell, in a manifold M, of size $|M|$, corresponding to the environment, the connection between cells i and j may be expressed as

$$J_{ij} = \frac{|M|}{N} K(|\mathbf{r}_i - \mathbf{r}_j|), \tag{3.2}$$

where K is a monotone decreasing function of its argument.

The synaptic input to the i-th cell is therefore given by

$$h_i = \sum_j J_{ij} V_j = \sum_j \frac{|M|}{N} K(|\mathbf{r}_i - \mathbf{r}_j|) V_j. \tag{3.3}$$

If the number N of cells is large, and the place fields centers (p.f.c.) are homogeneously distributed over the environment M (be it one or two-dimensional), we can replace the sum over the index j with an integration over the coordinates of the p.f.c:

$$h(\mathbf{r}) = \int_M d\mathbf{r}' K(|\mathbf{r} - \mathbf{r}'|) V(r'). \tag{3.4}$$

Note that the normalization in eq. 3.2 is chosen in order to keep the synaptic input to a given unit fixed when $|M|$ varies and the number of units is kept fixed, that is, the density of p.f.c.s $N/|M|$ varies (the $|M|$ factor will then compensate for the fewer input units within the range of substantial K strength). A fixed-point activity configuration must have the form

$$V(\mathbf{r}) = g\left[\int_M d\mathbf{r}' K(|\mathbf{r} - \mathbf{r}'|) V(\mathbf{r})' - \theta\right]^+. \tag{3.5}$$

We could write eq. 3.5 as

$$V(\mathbf{r}) = \begin{cases} g(\int_\Omega d\mathbf{r}' K(|\mathbf{r} - \mathbf{r}'|) V(\mathbf{r}') - \theta) & \mathbf{r} \in \Omega \\ 0 & \mathbf{r} \notin \Omega \end{cases} \tag{3.6}$$

where Ω is a domain for which there exists a solution of eq. 3.2 that is zero on the boundary.

If only solutions for which Ω is a convex domain are considered, the fact that $V(r)$ is zero on $\partial\Omega$ will ensure that units with p.f.c. outside Ω are under threshold,

therefore their activity is zero and solutions of eq. 3.2 are guaranteed to be solutions of eq. 3.5. The size and the shape of the domain Ω in which activity is different from zero is determined by eq. 3.2. As a first remark, we notice that it is independent from the value of the threshold θ. In fact, if V_θ is a solution of (3.2) with threshold θ, given the linearity of eq.3.2 within Ω,

$$V_{\theta'} = \frac{\theta'}{\theta} V_\theta$$

will be a solution of the same equation with θ' instead of θ, with the same null boundary conditions on Ω. Rescaling the threshold will then have the effect of rescaling the activity configuration by the same coefficient. This means that subtractive inhibition cannot shape, e.g. shrink or enlarge, this stable configuration, and therefore it is not relevant for a good part of the subsequent analysis. Some form of inhibition is nevertheless necessary to prevent the activity from exploding. Moreover, there are fluctuation modes which cannot be controlled by overall inhibition as they leave the total average activity constant. They will be treated in sec. 3.3.3. It is found that, at least in the one dimensional case, these modes do not affect stability in the single chart case.

In absence of an external input, any solution can be at most marginally stable, because a translation of the solution is again a solution of eq. 3.2. An external, "symmetry breaking" input, taken as small when compared to the contribution of recurrent synapses, is therefore implicit in the following analysis.

3.2.1 The one-dimensional case

The case of a recurrent network whose attractors reflect the geometry of a one-dimensional manifold, besides being a conceptual first step in approaching the 2-dimensional case, is relevant by itself, for example in modeling other brain systems showing direction selectivity, e.g. in head direction cells (Muller, Ranck, & Taube, 1996; Touretzky & Redish, 1996), and also for place fields on one-dimensional environments (Gothard, Skaggs, & McNaughton, 1996).

In this case eq. 3.2 reads:

$$\begin{aligned} V(r) &= g\left(\int_{-R}^{R} K(|r-r'|)V(r') - \theta\right) \\ V(R) &= V(-R) = 0. \end{aligned} \tag{3.7}$$

For several specific forms of the kernel K it is possible to solve explicitly eq. 3.7, yielding interesting conclusions. For example if:

$$K(|r-r'|) = e^{-|r-r'|}, \tag{3.8}$$

(see also (Tsodyks & Sejnowski, 1994)) differentiating eq. 3.7 twice yields:

$$V''(r) = -\gamma^2 V(r) + g\theta, \tag{3.9}$$

where $\gamma = \sqrt{2g-1}$.

Solutions vanishing at $-R$ and R (and not vanishing in $]-R, R[$), have the form:

$$V(r) = A\cos(\gamma r) + \frac{g\theta}{2g-1}, \tag{3.10}$$

with

$$A = -\frac{g\theta}{(2g-1)\cos(\gamma R)}. \tag{3.11}$$

The value of R for which (3.10) is a solution of eq. 3.7 is determined by the integral equation itself: for example, by evaluating $V'(R)$ or $V'(-R)$ from eq. 3.7 we get:

$$V'(-R) = -V'(R) = g\theta. \tag{3.12}$$

Substituting (3.10) and (3.11) in eq. 3.12 we have:

$$\tan(\gamma R) = -\gamma$$

so that

$$R = \frac{\tan^{-1}(-\gamma) + n\pi}{\gamma}.$$

Requiring R to be positive and $V(x)$ to be positive for $-R < r < R$, leads to choose

$$R = \frac{-\tan^{-1}(\gamma) + \pi}{\gamma}, \tag{3.13}$$

note that $A > 0$.

R is then a monotone decreasing function of γ, and therefore of the gain g.

This is also true for other forms of the connection kernel K. As an example, consider the kernel:

$$K(r - r') = \cos(r - r') \tag{3.14}$$

by a similar treatment it is shown that a solution is obtained with

$$R = \frac{1}{g}. \tag{3.15}$$

The kernel

$$K(r-r') = \Theta(1-|r-r'|)(1-|r-r'|) \tag{3.16}$$

will result in a peak of activity of semi-width

$$R = \frac{\pi}{\sqrt{2g}}. \tag{3.17}$$

Equations of the type (3.5) have more solutions in addition to the ones considered above, representing a single activity peak. For example, if we consider an infinite environment, periodic solutions will be present as well, representing a row of activity peaks separated by regions of zero activity. These solution can be checked to be unstable if we model inhibition as an homogeneous term acting on all cells in the same way and depending on the average activity. Intuitively, if we perturb the solution by infinitesimally displacing one of the peaks, it will tend to collapse with the neighbor which has come closer.

3.2.2 The two-dimensional case

To model the place cells network in the hippocampus we need to extend this result to a two dimensional environment. The equation for the neural activity will be:

$$V(\mathbf{r}) = g\left[\int_M d\mathbf{r}' K(|\mathbf{r}-\mathbf{r}'|)V(\mathbf{r}') - \theta\right]^+. \tag{3.18}$$

The generalization to 2-D is straightforward if for the kernel $K(|\mathbf{r}-\mathbf{r}'|)$ we consider the one with Fourier transform is

$$\hat{K}(\mathbf{p}) = \frac{2}{1+\mathbf{p}^2}, \tag{3.19}$$

(the two-dimensional analog of the kernel of eq. 3.8) that is, a kernel resembling the propagator of a Klein-Gordon field in Euclidean space. The fact that this kernel is divergent for $(\mathbf{r}-\mathbf{r}') \to 0$ does not give rise to particular problems, since, in the continuum limit of eq. 3.4, the contribution to the field h coming from the nearby points will stay finite, and in fact two units will be assigned p.f.c's so close to each other to yield a overwhelmingly high connection only with a small probability. Let us look for a solution with circular symmetry such that activity $V(\mathbf{r})$ is zero outside the circle of radius R, $\mathcal{C}(R)$. If we apply the Laplacian operator on both sides of

$$V(\mathbf{r}) = g\int_{\mathcal{C}(R)} d\mathbf{r}' K(\mathbf{r}-\mathbf{r}')V(\mathbf{r}') - \theta \tag{3.20}$$

we obtain:

$$\nabla^2 V(\mathbf{r}') = -\gamma^2 V(\mathbf{r}') + g\theta \tag{3.21}$$

(again, $\gamma^2 = 2g - 1$), which in polar coordinates reads:

$$V''(r) + \frac{1}{r}V'(r) = -\gamma^2 V(r) + g\theta. \tag{3.22}$$

The solution is

$$V(r) = AJ_0(\gamma r) + \frac{g\theta}{2g-1}. \tag{3.23}$$

J_0 is the Bessel function of order 0. For the solution to vanish on the boundary of $\mathcal{C}(R)$ one must take:

$$A = \frac{g\theta}{(2g-1)J_0(\gamma R)}.$$

The other condition that determines R may be found by substituting (3.23) in eq. 3.20. Here again, $R(g)$ is a monotone decreasing function.

As in the one-dimensional case, solutions with a non-connected (or even non-convex) support can be seen not to be stable.

3.3 Storing more than one map

Let us imagine now that the p.f.c's for each cell are drawn with uniform distribution on the environment manifold M, and connections are formed according to (3.2). Several "space representations" may be created by drawing again at random the r.p.c. of each cell from the same distribution. The connection between each pair of cells will then be the sum of a number of terms of the form (3.2), one for every "space representation", or "map", or "chart". With $p = \alpha N$ maps, and the r.p.c. of the i-th cell in the μ-th map indicated by $\mathbf{x}_i^{(\mu)}$:

$$J_{ij} = \sum_{\mu=1}^{p} \frac{|M|}{N} K(|\mathbf{r}_i^{(\mu)} - \mathbf{r}_j^{(\mu)}|). \tag{3.24}$$

The question that immediately arises is: what is the capacity of this network, that is, how many maps can we store, so that stable activity configurations, corresponding

to some region in the environment described by one map, like the ones described by the solutions of eq. 3.2, are present? The problem resembles the classic attractor neural network problem (Amit, 1989), with threshold linear units. A standard treatment has been developed (Treves, 1990) allowing to calculate the capacity of a network of threshold linear units with patterns drawn from a given distribution and stored by means of a hebbian rule. The treatment is very simplified in the extreme dilution limit (Derrida et al., 1987; Treves, 1991b). In the next sections it will be shown how this treatment can be extended to the map case, first for one particular form of the kernel K, leading to the solution of the capacity problem for a fully connected network; in the following, the solution is extended to more general kernels, first in the diluted limit, then for the fully connected network.

Another related question is: how much information is the synaptic recurrent structure encoding, and in which sense is the synaptic structure a store of information? The aim is to develop a full parallel between the multi-chart network and autoassociative networks, and if possible to characterize the parameters constraining the performance of this system.

3.3.1 The fully connected network: "dot product" kernel

Let us consider a manifold M with periodic boundary conditions, that is, a circle in one dimension and a torus in two dimensions. The p.f.c. of a cell $\mathbf{r}_i$ can then be described by a 2-dimensional unit vector $\vec{\eta}_i$ for the one-dimensional case and by a pair of unit vectors $\vec{\eta}_i^{1,2}$ for the two dimensional case. Suppose now that the contribution from the μ-th map to the connection between cell i and cell j is given by:

$$K(|\mathbf{r}_i^{(\mu)} - \mathbf{r}_j^{(\mu)}|) = \sum_{l=1}^{d} (\vec{\eta}_i^{l(\mu)} \cdot \vec{\eta}_j^{l(\mu)} + 1), \tag{3.25}$$

so that

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \sum_{l=1}^{d} (\vec{\eta}_i^{l(\mu)} \cdot \vec{\eta}_j^{l(\mu)} + 1), \tag{3.26}$$

where d is the dimensionality.

$p = \alpha N$ is the number of stored charts. Eq. 3.25 describes an excitatory, very wide spread form for the kernel (3.2) (the contribution to the connectivity is zero only if the r.p.c.s of the two cells are at the farthest points apart, i.e. at 180°). This spread of connectivity would lead to configurations of activity that are large in the r.p.c. space, that translated in auto-associative memory language would be very "unsparse", i.e. very distributed representations. It is therefore plausible that this will severely limit

the capacity of the net. In any case, the form of (3.25), factorizable in one term depending on $\vec{\eta}_i$ and one term depending on $\vec{\eta}_j$, after incorporating the constant part in a function $b^0(x)$, makes it possible to perform the free energy calculation through Gaussian transformations as in (Treves, 1990). A similar model has been studied in (Samsonovich & McNaughton, 1997) with McCulloch-Pitts neurons.

A Hamiltonian useful to describe the thermodynamics of such a system is

$$H = -\frac{1}{2}\sum_{i,j(\neq i)} J_{ij}V_iV_j - NB\left(\sum_i \frac{V_i}{N}\right) - \sum_l\sum_i\sum_\mu \mathbf{s}^{l(\mu)}\cdot\vec{\eta}_i^{l(\mu)}V_i \tag{3.27}$$

where $B(x) = \int^x b(y)dy$, and $b(x)$ is a function describing an uniform inhibition term depending on the average activity in the net. $\mathbf{s}^{l(\mu)}$ is a symmetry breaking field, pointing in a direction in the $\mu - th$ map space. The mean field free-energy in the replica symmetric approximation can be calculated (the partition function is calculated as the trace over a measure that implements the threshold-linear transfer function, see (Treves, 1990)). The presence of a phase with spatially specific activity correlated with one map will be signaled by solutions of the mean field equations with a non-zero value for the order parameter

$$\mathbf{x}^{l(\mu)} = \frac{1}{Nd}\sum_{i=1}^{N}\vec{\eta}_i^{l(\mu)}V_i \tag{3.28}$$

which plays the role of the overlap in an auto-associative memory. This parameter has the meaning of a population vector (Georgopoulos, Kettner, & Schwartz, 1988), that is, the animal position is indicated by an average over p.f.c.s of the cells weighted by cells activity.

The set of resulting mean field equations can be reduced to a set of two equations, eqs. A.7 and A.8, in two variables, the "non-specific" signal-to-noise ratio, w, and the "specific", space related signal-to-noise ratio v. The details of the calculation are reported in Appendix A.

The critical value α_c indicating the storage capacity of the network is the maximum value for which eq. A.7 still admits solutions corresponding to space related activity (non-zero v) and may be found numerically. At this value α_c the system undergoes a first-order phase-transition towards a state in which no space-related activity is possible. Eq. A.8 gives the range of gain values for which there exist solutions at a given $\alpha \leq \alpha_c$ (Treves, 1990).

In this model there is no possibility for modulating the spread of connections in the chart-space. As we anticipated, the activity configurations that one obtains are very wide, with a large fraction of units active at the same time. Cells will have very large place fields, covering a large part of the environment (of the order of roughly one half

for the one dimensional case, and roughly one quarter for the two-dimensional case). As one would infer from the analysis of autoassociative memories storing patterns, for example binary, these "unsparse" representations of space will lead to a very small capacity of the net.

For the model defined on the one-dimensional circle the capacity found is $\alpha_c \sim 0.03$. At this value the system undergoes a first order transition. As α increases beyond α_c, $\mathbf{x}$ jumps discontinuously from a finite value to zero.

The capacity for the diluted analogue of this model (see (Treves, 1991b), Appendix A and section 3.3.2) is given by the equation

$$E_1(w, \mathbf{v}) \equiv [(1+\delta)A_2]^2 - \alpha A_3 = 0. \tag{3.29}$$

Remember that in this case $p = \alpha c N$ where c is the connectivity fraction parameter, see section 3.3.2. In this case $\alpha_c \sim 0.25$. At α_c the transition is second order, with the "spatial overlap" $\mathbf{x}$ approaching continuously zero, verified at least with the precision at which it was possible to to solve numerically eq. 3.29. For the 2-D case, storage capacities are $\alpha_c \sim 0.0008$ for the fully connected network and $\alpha \sim 0.44$ for the diluted network.

To get a larger capacity, and to provide a possible comparison with the experimental data from the hippocampus, in which the tuning of place fields is generally narrow, we must extend our treatment to more general kernels, and this will be done in the following two sections.

3.3.2 Generic kernel: extremely diluted limit

Consider a network in which every threshold-linear unit, whose activity is denoted by V_j, senses a field:

$$h_i = \frac{1}{c} \sum_{j \neq i} C_{ij} J_{ij} V_j, \tag{3.30}$$

where J_{ij} is given by eq. 3.24. From now on the kernel K is defined as

$$\begin{aligned} K(\vec{r} - \vec{r}') &= \hat{K}(\vec{r} - \vec{r}') - \bar{K} \\ \bar{K} &= \left\langle\left\langle \hat{K}(\vec{r} - \vec{r}') \right\rangle\right\rangle \end{aligned} \tag{3.31}$$

for any $\vec{r}$, where $\langle\langle \ldots \rangle\rangle$ means averaging over $\vec{r}$. With this notation, whatever the original kernel $\hat{K}$, K is the subtracted kernel which averages to zero. The manifold M is taken with periodic boundary condition (that is a circle in one dimension and a torus in the two dimensional case).

C_{ij} is a "dilution matrix"

$$C_{ij} = \begin{cases} 1 & \text{with prob. } c, \\ 0 & \text{with prob. } 1-c \end{cases} \tag{3.32}$$

and $Nc/\log N \to 0$ as $N \to \infty$. In the thermodynamic limit $N \to \infty$ the activity of any two neurons V_i and V_j will be uncorrelated (Derrida et al., 1987). A number of charts $p = \alpha cN$ is stored. Looking for solutions with one "condensed" map, that is, solutions in which activity is confined to units having p.f.c. for a given chart in a certain neighborhood, it is possible to write the field h_i as the sum of two contributions, a "signal", due to the condensed map and a "noise" term, $\rho z - z$ being a random variable with Gaussian distribution and variance one – due to all the other, uncondensed, maps, namely, in the continuum limit, labeling units with the position $\vec{r}^1$ of their p.f.c. in the condensed map,

$$h(\vec{r}^1) = g \int_M d\vec{r}^{1\prime} K(\vec{r}^1 - \vec{r}^{1\prime}) V(\vec{r}^{1\prime}) + \rho z; \tag{3.33}$$

the noise will have a variance

$$\rho^2 = \alpha y |M|^2 \langle\langle K^2(\vec{r} - \vec{r}')\rangle\rangle, \tag{3.34}$$

where

$$y = \frac{1}{N} \sum_{i=1}^{N} \langle V_i^2 \rangle. \tag{3.35}$$

The fixed point equation for the average activity profile $x^1(\vec{r})$ is

$$x^1(\vec{r}) = g \int^{+} Dz (h(\vec{r}) - \theta). \tag{3.36}$$

where again Dz is the Gaussian measure, and

$$h(\vec{r}) = \int d\vec{r}' K(\vec{r} - \vec{r}') x^1(\vec{r}') + b(x) - \rho z \tag{3.37}$$

and

$$x = \int \frac{d\vec{r}}{|M|} x^1(\vec{r}) \tag{3.38}$$

is the average overall activity. The average squared activity (entering the noise term) will read:

$$y = g^2 \int \frac{d\vec{r}}{|M|} \int^{+} Dz(h(\vec{r}) - \theta)^2. \tag{3.39}$$

The fixed point equations may be solved introducing the rescaled variables

$$w = \frac{b(x) - \theta}{\rho} \tag{3.40}$$

$$v(\vec{r}) = \frac{x^1(\vec{r})}{\rho}. \tag{3.41}$$

The fixed point equation for $v(\vec{r})$ is

$$v(\vec{r}) = g\mathcal{N}\left(\int d\vec{r}' K(\vec{r} - \vec{r}')v(\vec{r}') + w\right) \tag{3.42}$$

where

$$\mathcal{N}(x) = x\Phi(x) + \sigma(x), \tag{3.43}$$

($\Phi(x)$ and $\sigma(x)$ are defined in eq. A.15 and eq. A.16) is a "smeared threshold linear function", monotonically increasing, with

$$\lim_{x \to -\infty} \mathcal{N}(x) = 0$$

and

$$\lim_{x \to +\infty} \mathcal{N}(x)/x = 1.$$

In terms of w and $v(\vec{r})$, y reads:

$$y = \rho^2 g^2 \int \frac{d\vec{r}}{|M|} \mathcal{M}\left(\int d\vec{r}' K(\vec{r} - \vec{r}')v(\vec{r}') + w\right) \tag{3.44}$$

where

$$\mathcal{M}(x) = (1 + x^2)\Phi(x) + x\sigma(x). \tag{3.45}$$

By substituting eq. 3.44 in eq. 3.34, we obtain

$$\frac{1}{\alpha} = g^2 |M| \langle\langle K \rangle\rangle \int d\vec{r} \mathcal{M}\left(\int d\vec{r}' K(\vec{r} - \vec{r}')v(\vec{r}') + w\right). \tag{3.46}$$

If we can solve eq. 3.42 and find $v(\vec{r})$ as a function of w and g, a solution is found corresponding to a value of α given by eq. 3.46. To find the critical value of α, we have to maximize α over w and g.

The mathematical solution of eq. 3.42 is treated in Appendix B.

With this model, we can modulate the spread of connections by acting on $K(\vec{r}-\vec{r}')$ or alternatively, by varying the size of the environment. The results are depicted in fig. 3.1 for the 1-D circular environment and in fig. 3.2 for the 2-D toroidal environment (upper curves). Examples of the solutions of eq. 3.42 are displayed in fig. 3.3 for the 1-D environment and in fig. 3.4 for the 2-D environment.

We note that, as the environment gets larger in comparison to the spread of connections (therefore, to the size of the activity peak), the capacity decreases approximately as

$$\alpha_c \sim -1/\log(a_m) \tag{3.47}$$

where a_m is the *map sparsity* and it is equal to:

$$a_m = \frac{k_d}{|M|} \tag{3.48}$$

where k_d is a factor roughly equal to ~ 4.5 for the 1-D model and ~ 3.6 for the 2-D model.

That is, the sparser the coding, the less the capacity. This is, at first glance, in contrast with what is known from the theory of auto-associative networks, in which sparser representations usually lead to larger storage capacities.

For comparison, keeping the formalism of (Treves, 1990), for threshold-linear networks with hebbian learning rule, encoding memory patterns $\{r_i\}_{i=1...N}$ with sparsity a defined as

$$a_p = \frac{\langle\langle r\rangle\rangle^2}{\langle\langle r^2\rangle\rangle}$$

(for binary patterns this is equal to the fraction of active units), and for small a, the capacity is given by

$$\alpha_p \sim \frac{1}{a_p \log(1/a_p)}. \tag{3.49}$$

The apparent paradox (larger capacity with sparser patterns, smaller with sparser charts) is solved as one recognizes that each chart can be seen as a collection of configurations of activity relative to different points in space covering, as in a tiling,

the whole environment. Each configuration is roughly equivalent to a pattern in the usual sense. Intuitively, and in sense that will made clearer below, a chart is equivalent, in terms of "use of synaptic resources" to a number proportional to a_m^{-1} of patterns of sparsity a_m.

The proportionality coefficient, or equivalently, the distance at which different configurations are to be considered to establish a correct analogy, will be dealt with in Appendix D.

These considerations and the comparison of eq. 3.47 and eq. 3.49 make clear that α_c is the exact analogue of the pattern autoassociators' α_p.

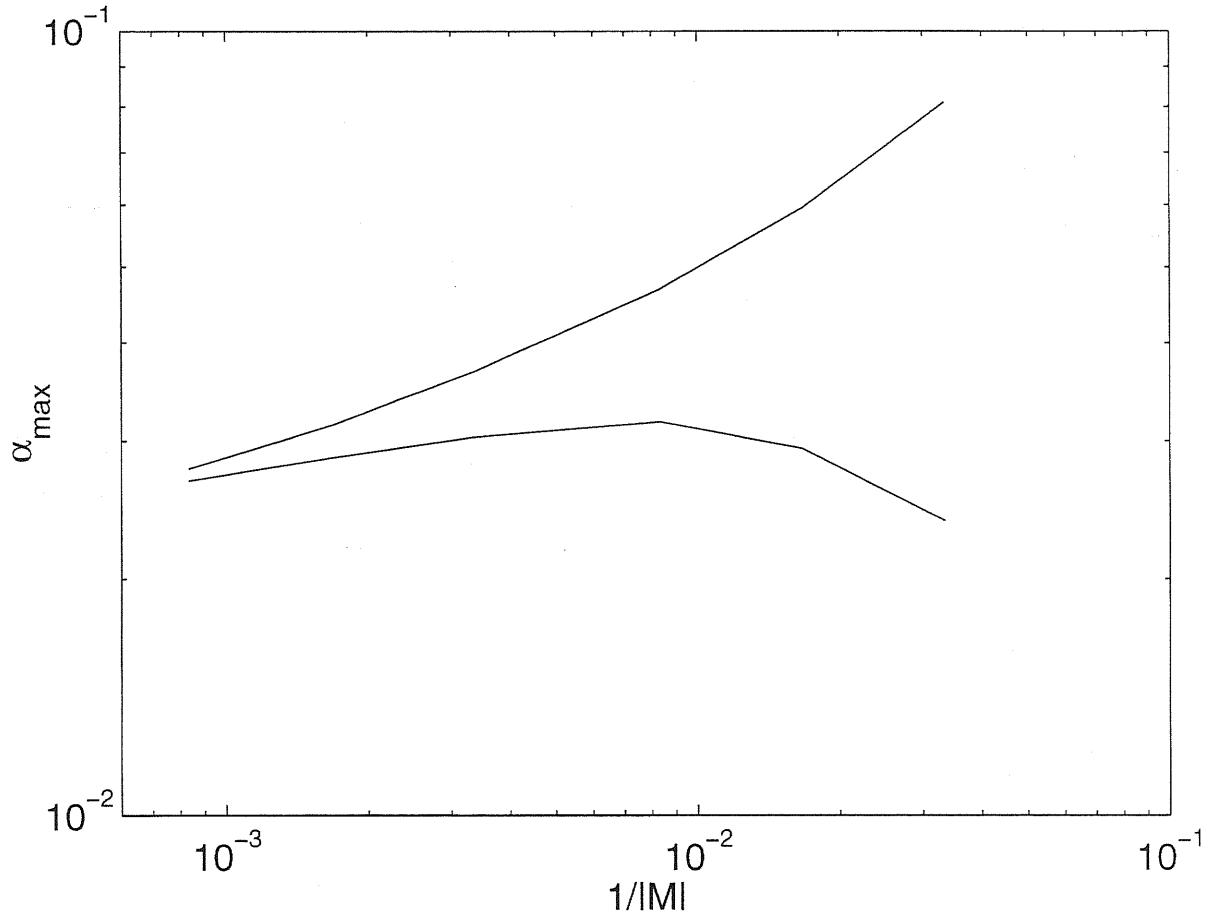

Figure 3.1: The storage capacity plotted as a function of the "map sparsity" a_m, for the 1-D model, for the extremely diluted (upper curve) and the fully connected (lower curve) limit.

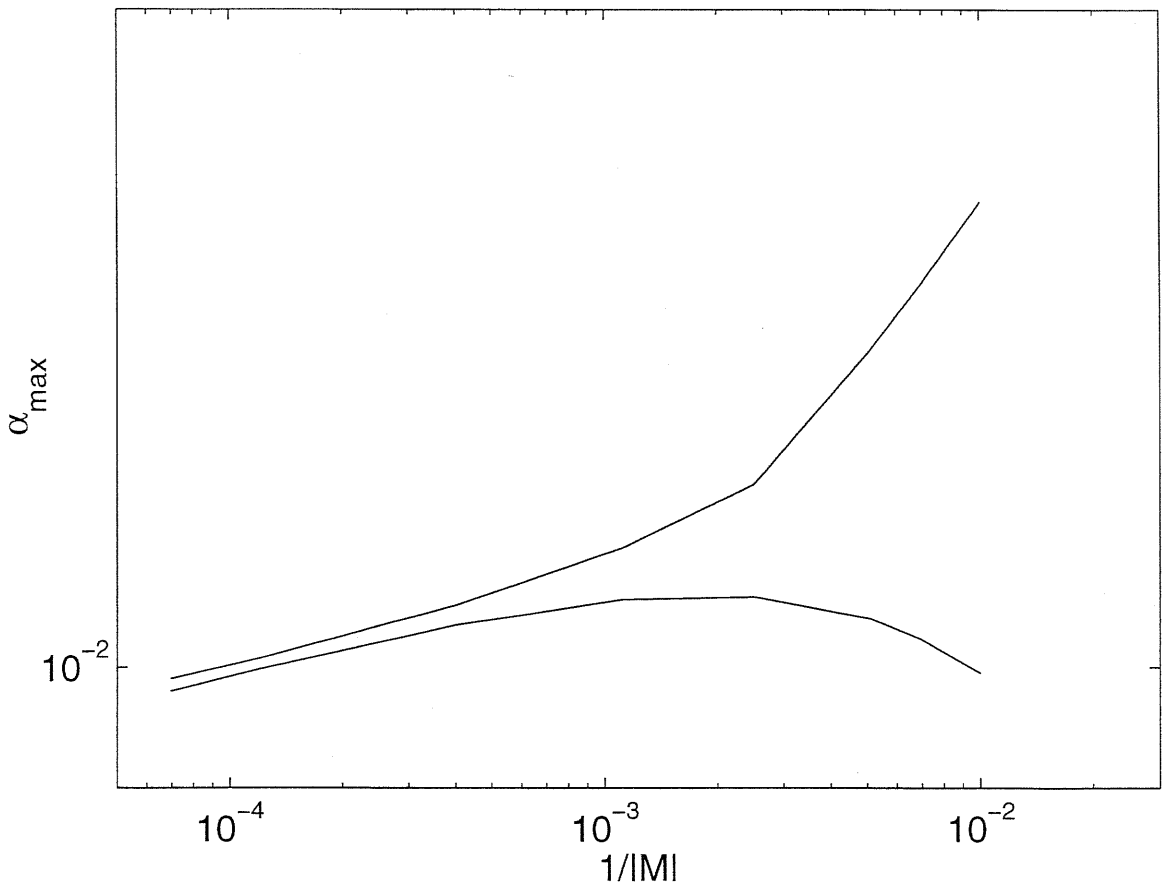

Figure 3.2: Same as fig. 3.1, for the 2-D model. The capacity is smaller than for the 1-D model for the same a_m.

3.3.3 Inhibition independent stability

The dynamical stability of the solutions of eq. 3.42 is in general determined by the precise functional form chosen for the inhibition, which we assumed a function of the average overall activity in the net. Nevertheless, there are fluctuations modes which leave the average activity unaltered. Stability against these modes is therefore unaffected by the inhibition and may be checked already for a general model. Let us consider a "synchronous" dynamics, that is, all the neurons are updated simultaneously at each time step. The evolution operator for the variables $V(r,t)$ and $\rho(t)$ is:

$$V(r,t+1) = g\rho(t)\mathcal{N}\left(\int_M \frac{dr'}{|M|}K(r-r')\frac{V(r',t)}{\rho(t)} + \frac{b(x(t))}{\rho(t)}\right) \quad (3.50)$$

$$\rho^2(t+1) = g^2\alpha|M|\rho^2(t)\langle\langle K^2\rangle\rangle \times \int_M dr\mathcal{M}\left(\int_M \frac{dr'}{|M|}K(r-r')\frac{V(r',t)}{\rho(t)} + \frac{b(x(t))}{\rho(t)}\right). \quad (3.51)$$

This evolution operator has as its fixed points $V_0(r) = \rho_0 v_0(r)$ and ρ_0 where $v_0(r)$ and ρ_0 are the solutions of eq. 3.42, 3.34, and 3.44, i.e. the stable states of our system.

We can linearize the evolution operator around $(V_0(r), \rho_0)$ and look for fluctuation modes (eigenvectors) $(\delta V(r), \delta\rho)$ with

$$\int_M dr\,\delta V(r) = 0 \quad (3.52)$$

We obtain the following equations:

$$\lambda\delta V(r) = g\Phi(u_0(r))\left[\int_M dr' K(r-r')\delta V(r')\right] + g\sigma(u_0(r))\delta\rho \quad (3.53)$$

$$\lambda\delta\rho = \left(1 - \frac{1}{2}g\alpha|M|\langle\langle K^2\rangle\rangle\int_M dr u_0(r)v_0(r)\right)\delta\rho + \frac{1}{2}g\alpha|M|\langle\langle K^2\rangle\rangle\int_M dr u_0(r)\delta V(r), \quad (3.54)$$

where

$$u_0(r) = \mathcal{N}^{-1}\left(\frac{v_0(r)}{g}\right).$$

Inserting eq. 3.53 in eq. 3.52:

$$\delta\rho = -\frac{\int_M dr'\Phi(u_0(r))\left[\int_M dr K(r-r')\delta V(r')\right]}{\int_M dr\sigma(u_0(r))} \quad (3.55)$$

Eq. 3.55 can be inserted again in eq. 3.53, obtaining a closed integral equation in δV. Unfortunately, this equation is very difficult to solve, but we can derive a stability condition by making an ansatz on the form of the eigenfunction $\delta V(r)$. More precisely, let us concentrate on the 1-D case. We look for solutions with even symmetry (we know there must be an eigenfunction with odd symmetry, and eigenvalue equal to 1, corresponding to a coherent displacement of the activity peak). This kind of solution

corresponds to spreading and shrinking of the activity peak. Let us assume that the even eigenfunction with the highest eigenvalue (the most unstable) has only two nodes (an even eigenfunction must have at least two nodes because of eq. 3.52), at r_0 and $-r_0$. Let us take the sign of the eigenfunction $\delta V(r)$ such that $\delta V(0) > 0$. From eqs. 3.52 and 3.55 we see that

$$\delta\rho < 0.$$

Now, from eq. 3.54:

$$\lambda = \left(1 - \frac{1}{2}g\alpha|M|\langle\langle K^2\rangle\rangle \int_M dr u_0(r)v_0(r)\right) + \frac{1}{2}g\alpha|M|\langle\langle K^2\rangle\rangle \int_M dr u_0(r)\frac{\delta V(r)}{\delta\rho}, \tag{3.56}$$

and we recognize that

$$\int_M dr u_0(r)\frac{\delta V(r)}{\delta\rho} < 0.$$

Thus,

$$\lambda < 1 - \frac{\Gamma}{2} \tag{3.57}$$

with

$$\Gamma = g\alpha|M|\langle\langle K^2\rangle\rangle \int_M dr u_0(r)v_0(r). \tag{3.58}$$

Thus, if the ansatz we formulated holds, we have a stability condition $\Gamma > 0$, which is found to be fulfilled for all the solutions we found relative to maximal storage capacity. This implies that the storage capacity result is not affected by instability of the solutions, provided of course that an appropriate form for inhibition is chosen. This stability result is also related to the correlation in the static noise for two solutions centered at different p.f.c.s, as we will show in App. D.

It can also be shown that by taking the $\alpha \to 0$ limit (i.e. the single chart case), one always has $\Gamma > 0$ since it is $v_0(r) = 0$ when $u_0(r) < 0$.

3.3.4 The fully connected model

The treatment of the model with the fully connected network and a kernel K for connection weights satisfying the condition (3.31) will use the replica trick to average

over the disorder (the realizations of the $\vec{r}$'s) and will eventually lead to a non-linear integral equation for the average activity profile in the space of the "condensed map" very similar to eq. 3.42. Let the Hamiltonian of the system be

$$H = -\frac{1}{2}\sum_{i,j(\neq i)} J_{ij}^{c} V_i V_j - NB\left(\sum_i \frac{V_i}{N}\right) - \sum_i \sum_\mu \mathrm{s}^{l(\mu)} \cdot r_i^{(\mu)} V_i \tag{3.59}$$

where now the J_{ij}^{c} are given by (3.24) with a generic kernel

$$K(\vec{r} - \vec{r}') = \hat{K}(\vec{r} - \vec{r}') - \bar{K} \tag{3.60}$$

where, again,

$$\bar{K} = \langle\langle \hat{K}(\vec{r} - \vec{r}')\rangle\rangle.$$

The free energy calculation is sketched in Appendix C. Again, the stable states of the system are governed by mean field equations. The mean field equation, eq. C.16 is an integral equation in the functional order parameter $v(\vec{r})$, the average space profile of activity.

If we are able to solve eq.C.16 and find $v^{\sigma}(\vec{r})$ as a function of w and g', by substituting eq. C.17 and eq. C.18 in eq. C.11 we have an equation that gives us the value of α corresponding to that pair (g', w). α_c is then the maximum of α over the possible values of (g', w).

To solve eq. C.16, it is easy to verify that if $\tilde{v}(\vec{r})$ is a solution of

$$\tilde{v}(\vec{r}) = g'\mathcal{N}\left(\int_M d\vec{r}'\, \hat{K}(\vec{r} - \vec{r}')\tilde{v}(\vec{r}') + \hat{w}\right) \tag{3.61}$$

with

$$\hat{w} = w - \bar{K}\int_M d\vec{r}\, \tilde{v}(\vec{r})$$

that is, the same equations as eqs. B.2 and B.3, then

$$v(\vec{r}) = \int_M d\vec{r}' \left[L(\vec{r} - \vec{r}') - \bar{L}\right] \tilde{v}(\vec{r}')$$

is a solution of eq. C.16. $\tilde{v}$ can therefore be interpreted as the average activity profile, apart from a constant. Eq. 3.61 can be solved with the same procedure used for eq. 3.42, and the maximum value of α can be found by maximizing over g' and $\hat{w}$.

The results for 1-D and 2-D environment are depicted in fig. 3.1 and 3.4 (lower curves). As we may expect from pattern autoassociator theory, the capacity is much

lower than for the diluted model, due to an increased interference between different charts.. As the sparsity $a \sim 1/|M|$ gets smaller, the capacities of the two models get closer, both being proportional to $\frac{1}{\log(|M|/k_d)}$. Reducing the sparsity parameter of space representations has therefore the effect of minimizing the difference between nets with sparse and full connectivity.

3.3.5 Sparser maps

A possible extension of this treatment is inspired from the experimental finding that, in general, not all the cells have place cells in a given environment. Ref. (Wilson & McNaughton, 1993) e.g. reported that $\sim 28 - 45\%$ of pyramidal cells of CA1 have a place field in a certain environment. We would like to see how this fact could affect the performance of the multi-chart auto-associator. It is then natural to introduce a new sparsity parameter, the *chart sparsity* a_c indicating the fraction of cells which participate in a chart. We will show that, for the capacity calculation, a_c^{-1} "sparse" charts are equivalent to a single "full" chart, of size a_c^{-1}. We will present the argument for the diluted case, the fully connected case is completely analogous.

Let m_i^μ be equal to 1 if cell i participates in chart μ, that is with probability a_c. Thus, the synaptic coupling J_{ij} will read

$$J_{ij} = \sum_{\mu=1}^{p} \frac{|M|}{a_c N} K(\mathbf{x}_i^{(\mu)} - \mathbf{x}_j^{(\mu)}) m_i^\mu m_j^\mu. \tag{3.62}$$

Let us consider a solution with one condensed map: cells participating with p.f.c. in r in that map will have a space related signal-to-noise ratio

$$v(\vec{r}) = g\mathcal{N}\left(\int d\vec{r}' K(\vec{r} - \vec{r}') v(\vec{r}') + w\right) \tag{3.63}$$

for all the neurons *not* participating in the condensed map we will have

$$v = \mathcal{N}(w). \tag{3.64}$$

The noise will have a variance

$$\rho^2 = \alpha a_c y \left(\frac{|M|}{a_c}\right)^2 \langle\langle K^2(\vec{r} - \vec{r}^\mu)\rangle\rangle, \tag{3.65}$$

that is a_c times the value we would get for the same number of "full" charts with size $|M|/a_c$. and now

$$y = \rho^2 g^2 \left\{ a_c \int_M \frac{d\vec{r}}{|M|} \mathcal{M}\left(\int d\vec{r}' K(\vec{r} - \vec{r}') v(\vec{r}') + w\right) + (1 - a_c)\mathcal{M}(w) \right\}. \tag{3.66}$$

By comparing eq. 3.44 and eq. 3.66, and remembering that for $\vec{r}$ far from the activity peak, $v(\vec{r}) \sim \mathcal{N}(w)$ we realize that this y value is approximately equivalent to the y value we would get for "full" charts of size $|M|/a_c$.

Inserting eq. 3.65 and eq. 3.66 in eq. 3.46, one finds, for the maximal capacity:

$$\alpha_{c\,(\text{sparse charts})} \sim \frac{1}{a_c \log(\frac{|M|}{k_d a_c})} =\sim \frac{1}{a_c \log(\frac{1}{a_m a_c})}. \tag{3.67}$$

As we anticipated one may interpret this result as follows: the capacity is the same as if we had taken a_c^{-1} "sparse" charts, including $\sim N$ cells, and put them side by side to form one single "full" chart. If we have started with αC "sparse" charts we now have have $a_c \alpha C$ "full charts". From eq. 3.46 we see that we can store at most $\alpha_{c\,(\text{full charts})} C$ full charts and

$$\alpha_{c\,(\text{full charts})} \sim \frac{1}{\log(\frac{1}{a_m a_c})}$$

and this explains eq. 3.67. Therefore, this network is *as efficient* in terms of spatial information storage as the one operating with full charts.

3.4 information storage

Like a pattern auto-associator, the chart auto-associator is an information storing network. The cognitive role of such a module could be to provide a spatial context to information of non-spatial nature contained in other modules, which connect with the multi-chart module. Each chart represents a different spatial organization, possibly related to a different environmental/behavioral condition. Within each chart, a cell is bound to a particular position in space, thus being the means for attaching some piece of knowledge to a particular point in space, through inter-module connections. To give a very extreme, unrealistic, but perhaps useful, example, let us assume that each cell encodes a particular discrete item, or the memory of some events happened somewhere in the environment, as in a "grandmother cell" fashion, encoding "the grandmother sitting in the armchair in the dining room". The encoding of the "grandmother" may be accomplished by some set of afferents from other modules. The multi-chart associator can then attach a spatial location to that memory of the "grandmother". The spatial location encoded is ideally represented for each cell by its p.f.c.

In this sense, the information encoded in the network, which can be extracted by measures of the activity of the units, is the information about the spatial tuning of the units, that is their p.f.c.s.

To restate this concept in a formal way, we look for

$$I_s = \lim_{S\to\infty} \frac{1}{CN} \sum_{\mu} \sum_{i} I(\vec{r_i^{\mu}}, \{V_i^{\mu(k)}\}_{k=1...S}) \tag{3.68}$$

that is the information per synapse that can be extracted from S different observations of activity of the cells with the animal is in S different positions , and the system in activity states related to chart μ. This quantity does not diverge as $S \to \infty$, since repeated observations of activity with the animal in nearby positions do not yield independent information, because of correlations between activity configurations, correlations which decrease with the distance at which the configurations are sampled.

The full calculation of this quantity involves a functional integration over the distribution of noise affecting cell activity as the animal is moving, and exploring the whole environment. In Appendix D we suggest a procedure to approximate this quantity based on an "information correlation length" l_I such that samples corresponding to animal positions at a distance l_I yield approximately independent information.

I_s is the amount of spatial information which is stored in the module. It is the exact analogue of the stored information for pattern auto-associators (Treves, 1990). As for storage capacity, it is to be found numerically, by maximization over w and g.

As for the capacity one can find the solution which maximizes I_s. The resulting $I_{\max}$ is a function of the size of the relative spread of connections $a = 1/|M|$, and it amounts to a fraction of bit per synapse (see fig. 3.5).

As with pattern auto-associators, the information stored increases with sparser representations. The increase is more marked for the fully connected network. For very sparse representations the performance of the fully connected model approaches the extreme dilution limit.

3.5 Discussion

We have studied the multi-chart threshold linear associator, as a spatial information encoding and storage module. We have given the solution for the dot-product kernel model, then we have introduced a formalism in which the generic kernel problem is soluble.

The second treatment has the advantage of providing a form for the average activity peak profile, which can be compared with the experimental data (see for example (Samsonovich & McNaughton, 1997), fig.1).

We have shown that the non-linear integral mean field equation (eq. 3.42) can be solved at least for one class of connection kernels $K(r - r')$.

The storage capacity for both models has been found. We note that the capacity for the dot-product model is compatible with the wide kernel (non sparse) limit of the generic model in one and two dimensions in the fully connected and in the diluted condition.

The generic kernel treatment makes it possible to manipulate the most relevant parameter for storage efficiency, i.e. the spread of connections. It is shown that this parameter plays a very similar role as sparsity for pattern auto-associators. In the multi-chart case, moreover, the effective sparsity of the *stable configurations* is determined also by the value of the gain parameter g, as shown analytically for the noiseless case. Nevertheless, the capacity of the network depends on the spread of connection parameter $a_c = k_d/|M|$ through a relation which is the exact analogue of the relation between sparsity and capacity for the pattern auto-associator, at least in the very sparse limit.

We have only considered here the capacity problem for one form of the connection kernel, although the treatment we propose is applicable, at least, to the other kernels considered for the noiseless case. Our hypothesis is that a similar law for sparsity is to be found as eq. 3.47, at least in the high sparsity limit, for more general forms of the kernel.

We have then shown that the capacity scales in such a way that the information stored is not changed when only a fraction of the cells participate in each chart. In this case the firing of a cell carries information not only about the position of its p.f.c. in the chart environment, but also about *which* environment the cell has a place-field in. This information adds up, so that $1/a_c$ charts can be assembled in a single larger chart of size $1/a_c$ times larger.

We have introduced a definition of stored information for the multi-chart memory network, which measures the number of effective different locations which can be discriminated by such a net: representations of places at a distance less than l_I are confused, because of the finite width of the activity peaks, and because of the static noise.

l_I does not vary much when $|M|$ varies. This is consistent with the fact that the storage capacity is well fitted by eq. 3.47 with $k_d \sim 4.5$. l_I turns out to be ~ 3.5 for the 1-D model, with the arbitrary value for f of 0.95. l_I is therefore similar to the "radius" of the activity peak which should correspond to the "pattern" in the parallel between the chart auto-associator and the pattern auto-associator.

It was not possible to carry over the calculation of r_{12} and I_2 in the 2-D model as it turns out to be too computationally demanding. Therefore we are not able to show the values of the storable information The fact that the storage capacity follows eq. 3.47 also in this case is an indirect hint of a behavior very similar to what is found in 1-D.

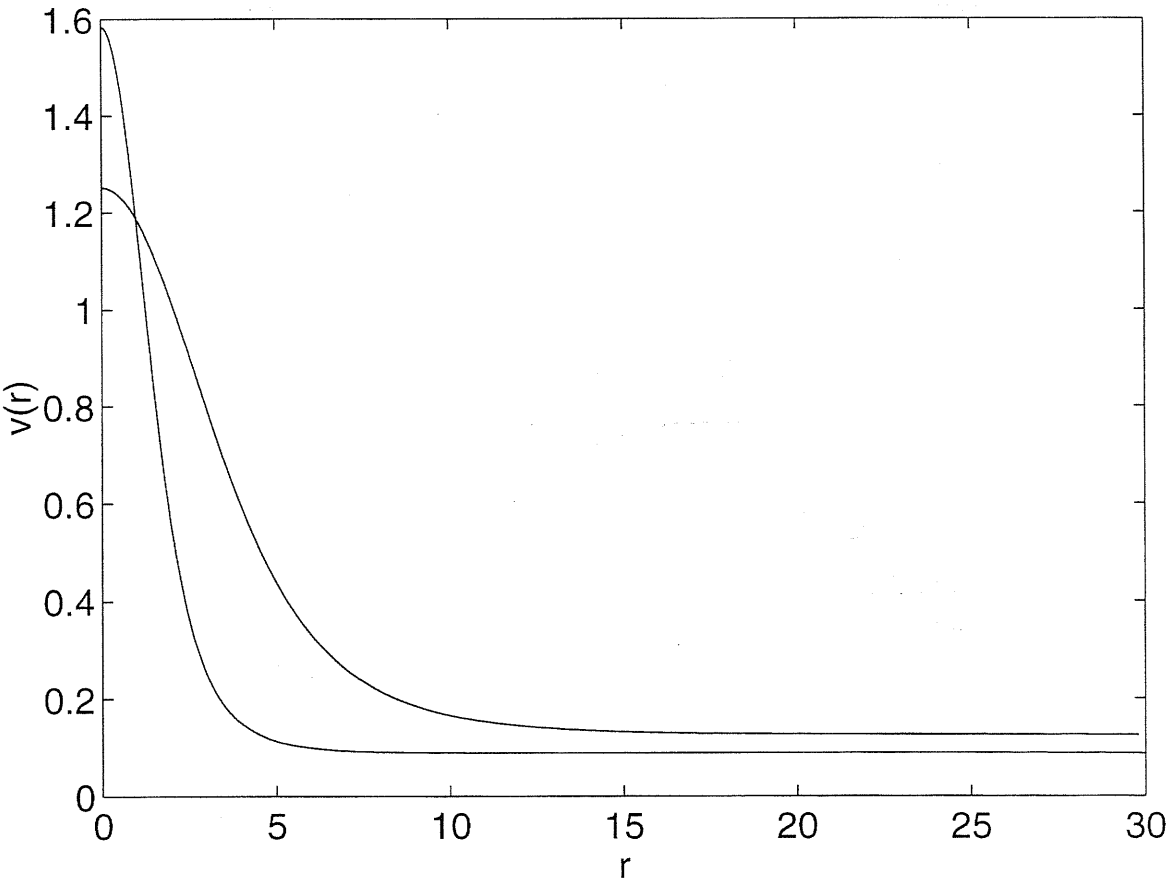

Figure 3.3: The "activity peak" profile corresponding to the solution of eq. 3.42 at the maximal storage level at $|M| = 30$ and $|M| = 15$. The second case is plotted expanded to match the environment size of the first one and to show the effect of more widespread connections.

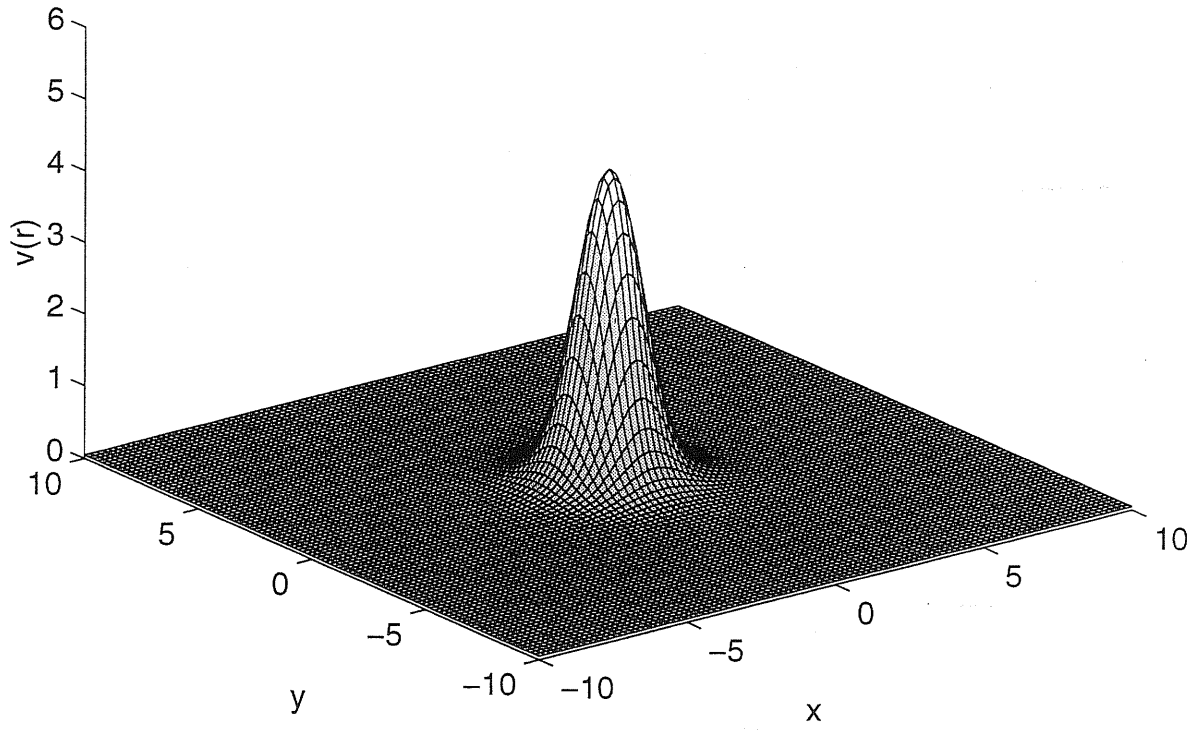

Figure 3.4: The maximal storage activity peak profile in 2-D at $|M| = 400$.

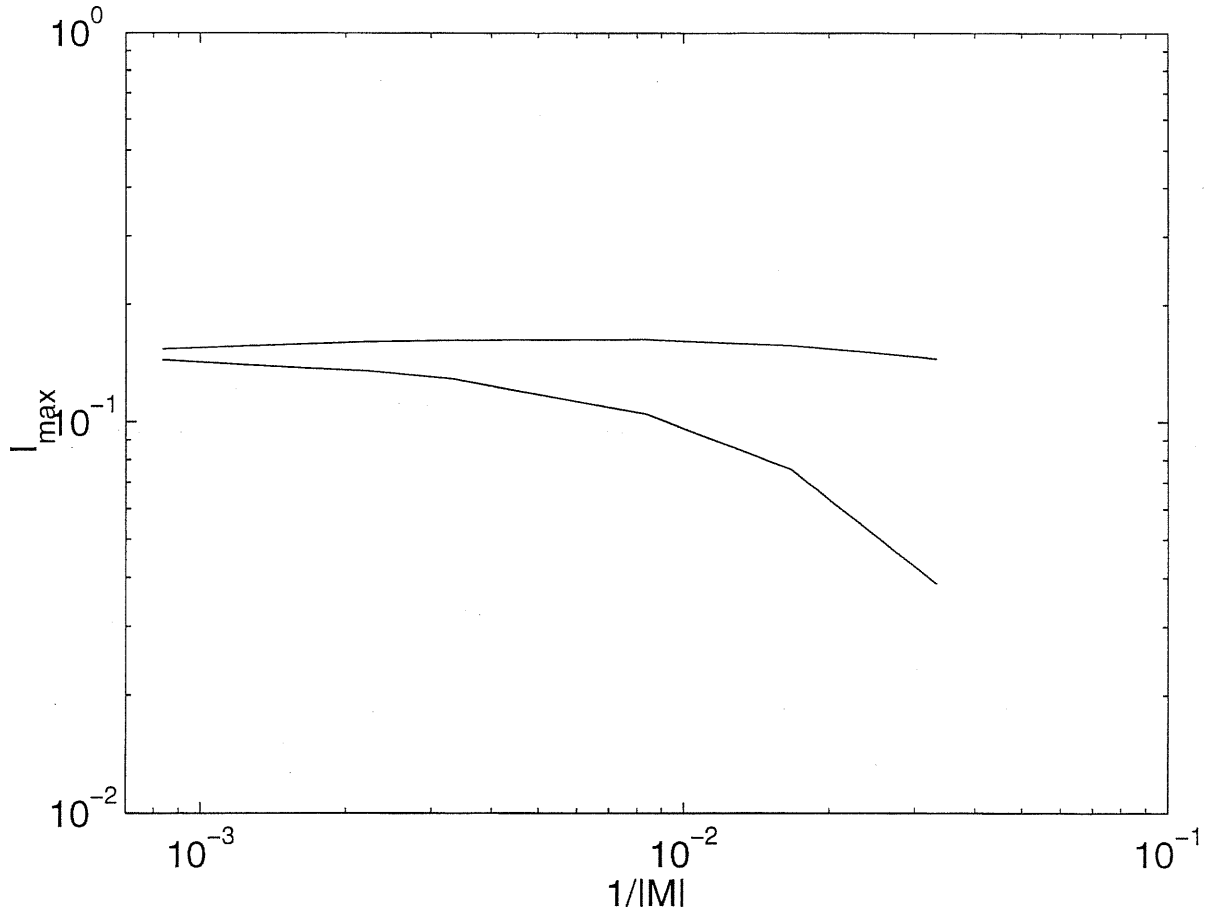

Figure 3.5: The maximal storable information per synapse, as a function of $1/|M|$.

Chapter 4

Speed of retrieval and stability of retrieval states in an integrate-and-fire network

4.1 Introduction

Information processing in the brain is extremely fast. Sensory systems can produce responses involving very complex computations in a very short time, in the range of tens, or hundreds of milliseconds. A very striking example is the visual system: complex scene analysis and object recognition, which require processing by higher visual areas, can be performed in very short times. A demonstration of this is given by the experiments of Thorpe, Fize, and Marlot (1996). Human subjects were shown images of complex scenes, and they had to figure out whether there was an animal in the scene or not, and react consequently. This is a very abstract task, as the kind of animal, the view, the position of the animal in the image, the background it should appear against, were not specified. Thorpe et al. showed that signs of a correct detection (or even of an erroneous one) of an animal in the image show up in the subject's EEG trace after about 150 msec from image onset.

It is likely that visual information has to flow through 8-10 synaptic connections from the retinal ganglion cells to the higher visual area that perform the abstract computation. This purports 15-20 msec available for each processing stage. In this short time, the neural module has to carry out the computation it is devoted to, and it has to present the outcome to the next layer in a readable form. This represent an important constraint for any theory of processing in the brain: not only the model has to prove itself capable of solving a given problem, but it also has to prove it can do it fast enough, once the parameters and the time scales of the real brain hardware

are taken into account.

15-20 msec are indeed, in some sense, the elementary time scale of neurons, that is, the time it takes to fire a spike, or more properly, the *inter-spike interval*, for cells (not uncommon especially in higher sensory areas) with a typical firing rate of 50-60 Hz. Even cells with higher firing frequencies will emit a few action potentials in that time. Also the time scales of synaptic conductances and dendritic integration are comparable to this interval of time. Thus, the constraints coming from function of consideration require that each neural module does its work, which is complex and likely involves a large number of units cooperating, in the time a single unit can fire a few spikes, or transmit a few synaptic elementary bits of information. This seemed to some authors (see e.g. Thorpe & Imbert, 1989) to rule out a possible role for recurrent processing in systems, like vision, that have to perform so fast. The recurrent processing, this is the argument, involves *relaxation*, that is approaching an equilibrium state, an attractor. The typical dynamics of symbolic neural units considered in earlier, statistical physics flavored models, and in models from the *Connectionist* school, and even implemented in simulations is *Glauber dynamics*, or some more or less close relative. In this class of dynamics at each *discrete* time step, the activity value of each unit is modified according to some update rule. The somewhat tricky point is now to identify to which time scale a discrete time step has to correspond to, if sensible conclusions about the collective phenomena time scales are to be drawn. The inter-spike interval, or the synaptic conductances and dendritic integration time constants seemed to be viable choices. On the other hand, it turns out that approach to equilibrium in attractor networks (auto-associative memories etc.) is achieved in at least a few time steps, and this would be far too long, given the speed of processing requirements and the time interval that would correspond to the time step. The conclusion one could draw from this picture is that processing must be prevalently feed-forward, with information flowing through an uni-directional chain of synaptic relays.

Moreover, rate coding appears to be unfeasible for fast information transmission: a reliable estimate of the firing rate of one cell requires a fairly large number N_s of subsequent inter-spike intervals to be read, since the error on the estimate is proportional to $1/\sqrt{N_s}$. Thorpe and Imbert proposed a different coding scheme, based on the *order of arrival* of spikes from different cells. Many other schemes, involving the detailed time structure of the activity, have been proposed.

What we attempt here is to re-examine this issue, starting with the assumption that Glauber-type, discrete time dynamics are not enough to realistically describe dynamics of neural assemblies. We need a formal model capable of representing with some degree of accuracy what happens at the level of single spikes. The simplest suitable model is the integrate-and-fire neuron (Lapique, 1907; Eccles, 1957) with conductance-based synaptic transmission (Eccles, 1964). This latter feature amounts

to assuming that an action potential event in the pre-synaptic cell causes a *conductance change* in the post-synaptic cell (as opposed, for example, to an injection of current), the synaptic conductance successively following its own dynamics, typically inactivating, or decaying, for example with an exponential time behavior. Treves (1993) analyzed this family of models with analytical tools and yielded an analytical formula for the time constants of the exponentially decaying transient modes, through which firing activity in the network approaches the firing at steady state. There are many different modes, each of which has a time constant with a real part, describing the rate of decay of the mode, and an imaginary part, specifying the frequency of the oscillations accompanying the decay. An important family of transients has the real part of the time constant determined by the rate of inactivation of the synaptic conductances opened by activity on the recurrent collaterals.

Since such rate of inactivation in the brain is typically short of the order of 10-20 msec, even when taking into account the dendritic spread not included explicitly in the integrate-and-fire description (Hestrin, Nicoll, Perkel, & Sah, 1990; Colquhoun, Jonas, & Sakmann, 1992; McBain & Dingledine, 1992), a prediction arising from the analysis is that the contribution of recurrent collaterals to the retrieval of a memory representation may take place in a relatively short time, over a few tens of msec, independently of the prevailing firing rates and of the membrane time constants, however defined, of the neurons in the population (Treves, Rolls, & Tovee, 1996). The analysis has however remained incomplete, because it only describes the modes close to steady state, and not the full dynamics from an arbitrary initial state, and because it is unable to tell to what extent each individual mode will be activated when the activity evolves from any initial state. These limitations can be overcome by computer simulations of the same network model considered by the analytical treatment.

A second aspect which has to be addressed by models that aim to be applicable to the real brain, is that of the *stability* of the steady states which are taken to correspond to memory retrieval. As with any steady state in the dynamics of a system of many units, there are very many possible sources of instability. One example is the instability of the steady states in which the firing of different units is asynchronous, to synchronization among groups of units (Tsodyks, Mitcov, & Sompolinsky, 1993; Deppisch, Bauer, & Schillen, 1993; Abbott & van Vreeswijk, 1994; Hansel, Mato, & Meunier, 1995). The asynchronous stable state is an important assumption for all the conclusions from equilibrium statistical mechanics about the stable states of a network, described in terms of firing rate to hold and be of relevance for real neural systems. A more basic potential instability, however, arises out of the fact that the Hebbian modifiable connections which are thought to mediate associative memory in the brain are those between pyramidal excitatory cells. Therefore a recurrent autoassociative memory is in itself a positive feedback circuit, and unless its activity can

be tightly controlled by appropriate inhibition, it will tend to explode. Although the stability of realistic networks of excitatory and inhibitory units has been studied already (e.g., Abbott, 1991), it was not in the context of auto-associative memories. In following section we show that there is in such networks a fundamental conflict between stability to excitatory explosion and storage capacity. In the next section, we show that the conflict can be avoided by inhibition which is predominantly multiplicative in nature. Then we go back to the issue of the time scales for retrieval, with simulations that support and qualify the analytical predictions. The last sectoin fo this chapter discusses the implications of these results for the operations of associative memories in the brain. This work was published in paper form in Battaglia and Treves (1998b).

4.2 The stability-capacity conflict

A full analysis of the stability of asynchronous steady firing states must be carried out using appropriately detailed models, but the requirements for stability against excitatory runaway reverberations can be discussed, to start with, using a simple two-variable model. In such a model, two variables ν_E and ν_I describe the average firing rates of excitatory and inhibitory units, which approach their steady state values with time constants τ_E and τ_I. The steady state values are determined by these average firing rates and by the level of afferent inputs. If we assume that, above threshold, the dependence is approximately linear, the dynamical system can be written (Wilson & Cowan, 1972)

$$\tau_E \dot{\nu_E} = -\nu_E + J_E^E \nu_E - J_E^I \nu_I + \nu_E^{aff} \quad (4.1)$$

$$\tau_I \dot{\nu_I} = -\nu_I + J_I^E \nu_E - J_I^I \nu_I + \nu_I^{aff} \quad (4.2)$$

where the J's are the adimensional effective couplings (signs are chosen so that they are all positive in value) between the dynamical variables, as they emerge, essentially, from averaging synaptic strengths across pairs of active units, and ν^{aff} are constant terms, which depend on the afferent input activity and are proportional to fixed point rates. They ensure that equilibrium rates are not zero, even if the network does not receive any input, reflecting the capability of the network of self-sustain its activity. If this system of equations has a fixed point, its stability requires that

$$\mathrm{Tr} = (J_E^E - 1)/\tau_E - (J_I^I + 1)/\tau_I < 0 \quad (4.3)$$

$$\mathrm{Det} = -(J_E^E - 1)(J_I^I + 1) + J_I^E J_E^I > 0. \quad (4.4)$$

Both inequalities can obviously be satisfied for arbitrary values of the mean excitatory-excitatory coupling among active units, J_E^E, provided inhibitory couplings are strong

enough to control excitation. If we want, however, on the basis of this simple two-variable model, to ensure the stability of a real auto-associator, both inequalities must be satisfied with *ample margins*. The reason is that exactly which units are active will be highly variable, and therefore the effective value of J_E^E at any moment in time will fluctuate substantially. It is easy to realize, then, that for values of the four mean couplings much larger than 1, the determinant appearing in the second condition will be of the order of such large value, squared. Now, the fixed point firing rates are

$$\nu_E^{fp} = \frac{(J_I^I + 1)\nu_E^{aff} - J_E^I \nu_I^{aff}}{\text{Det}} \tag{4.5}$$

$$\nu_I^{fp} = \frac{J_I^E \nu_E^{aff} - (J_E^E - 1)\nu_I^{aff}}{\text{Det}}, \tag{4.6}$$

which means that if the couplings are large, under conditions of robust stability the mean excitatory firing rate at the fixed point will be much lower than the one determined by afferent inputs alone, $\nu_E^{fp} << \nu_E^{aff}$. This is however incompatible with the the effective operation of the network as content addressable memory, since it makes recurrent processing minor with respect to the feed-forward relay of the cue. In fact, when we tried to simulate memory retrieval with large couplings, and at the same time we insisted on the condition that local intrinsic inputs dominate over external afferent inputs (a condition intended to mimick the observed cortical anatomy Abeles, 1991), we always run into large oscillations (Simmen, Treves, & Rolls, 1996), due to even transient imbalances between local excitation and inhibition, which resulted in large fluctuations in the effective couplings, and prevented the network from reaching a steady retrieval state. Only by using as a cue the nearly complete memory pattern we could effect proper retrieval, but then recurrent connections played only a minor role. Therefore, to obtain robust stable fixed points we had to restrict ourselves to smaller effective couplings, in particular to values of J_E^E not much above 1. In that case, since the excitatory self-coupling always appears in the combination $(J_E^E - 1)$, its potentially devastating influence on the stability of the fixed point will be reduced, and at the same time conditions will exist under which even small cues will be sufficient to initiate retrieval. Keeping the excitatory self-coupling low conflicts, however, with ensuring a large storage capacity, as shown next.

Consider a simple auto-associator in which the weights of the connections among the units are determined by a linear sum of Hebbian-modification terms, as e.g. in the Hopfield model (Hopfield, 1982). If the units represent excitatory cells and the weights ultimately correspond to conductances, one may assume that all such a memory structure is superimposed on a baseline connection weight, large enough as to keep positive even the individual weights which happen to undergo more negative

modifications [1]. Therefore one may write for the weight between units i and j

$$w_{ij} = w^0 + \frac{1}{C}\sum_{\mu=1}^{p}(\frac{\eta_i^\mu}{<\eta>} - 1)(\frac{\eta_j^\mu}{<\eta>} - 1) \tag{4.7}$$

where η_i^μ is the firing rate of unit i in the μ'th memory pattern, $<\eta>$ is the average firing rate, the network stores p patterns with equal strength, and C is the number of inputs per unit. The specific (covariance) form of the Hebbian term and the normalization factor are inessential to the argument that follows, and were chosen for consistency with previous analyses (Treves & Rolls, 1991). The minimum connection weights will be those between pairs in which the pre- and post-synaptic unit happen to be anti-correlated across all patterns, i.e. whenever one of the two is firing, for example at a typical elevated rate η^*, the other is quiescent. Then the condition which ensures that the underlying conductance remains positive, even in such cases, reads

$$w^0 \geq \frac{p}{C}\frac{\eta^*}{<\eta>}. \tag{4.8}$$

On the other hand, the effective excitatory self-coupling, i.e. the effect that the average excitatory firing rate exerts on each excitatory unit, is given by summing conductances across input lines and multiplying by the gain γ characterizing the unit's input-output transform in a linear range above firing threshold,

$$J_E^E = \gamma C w^0. \tag{4.9}$$

Note that the Hebbian terms average to zero when summing across the C inputs. Previous analyses (Treves, 1990; Treves & Rolls, 1991) have shown that for the network to be able to retrieve memory patterns, the gain has to be sufficiently strong, as expressed by the condition

$$\gamma \geq \frac{a}{(1-a)}, \tag{4.10}$$

where $0 < a < 1$ is the *sparseness* of the firing patterns, defines as $a =<\eta>^2 / <\eta^2>$ (Treves & Rolls, 1991). Putting now together the condition that the effective excitatory self-coupling be at most of order 1 with the last 3 equations one realizes why stability conflicts with storage capacity:

$$O(1) \approx J_E^E = \gamma C w^0 \geq p\frac{\eta^*}{<\eta>}\frac{a}{(1-a)}, \tag{4.11}$$

[1] This assumption is made for the sake of clarity. In the simulations that follow, we use an equivalent formulation, which is however less transparent to the analysis

that is, in this case, to be stable at retrieval, the network must not store more than a number of memory patterns

$$p_{max} \simeq \frac{<\eta>}{\eta^*}\frac{(1-a)}{a} = O(1)! \tag{4.12}$$

that is, more than a handful of patterns. In simulations that followed these very specifications, we found it difficult to obtain retrieval in nets storing more than 2-3 patterns, whatever their size. The conflict arises out of requiring *simultaneously* dynamical stability and effective retrieval ability and biological plausibility (in that the memory is stored on the connections between excitatory units, and in that each conductance must be a positive quantity). It does not arise in storage capacity analyses based on simplified formal models (Amit et al., 1987; Treves & Rolls, 1991) if one treats connection weights as real variables that can have either sign, and can change in sign as more memories are stored.

It is important to note that recurrent auto-associative memory models based on an alternative simple "learning rule", the so-called Willshaw models (Willshaw & Buneman, 1969), although assuming only positive (or zero) weights among excitatory units, still suffer from similar limitations. That class of models, however, is more difficult to treat analytically (Golomb, Rubin, & Sompolinsky, 1990), and does not lend itself to such a simple discussion of the conflict; moreover, what is limited is not simply p, the number of memories that can be stored (which can be well above 2-3, (Amit & Brunel, 1996)), but the total amount of information that can be stored and retrieved, which is proportional to p but also decreases the sparser are memory patterns (and the more information need be provided with the cue).

4.3 Realistic inhibition may avoid the conflict

A seemingly innocuous assumption that was made in writing down Eqs. 4.2 is that excitatory firing rates depend linearly not just on themselves but also, through a separate linear term, on inhibitory rates. This is equivalent to considering what is sometimes called subtractive inhibition. Purely subtractive inhibition is a convenient model for $GABA_B$ inhibition, that acts through K^+ channels of limited total conductance, primarily by hyperpolarizing the receiving cell (Connors, Malenka, & Silva, 1988). If co-located on dendrites along with excitatory inputs, $GABA_B$ can be thought of as providing an additional term which is negative in sign and hence subtractive, and occurs on a slower time scale (Hablitz & Thalmann, 1987).

$GABA_A$ inhibition, which is responsible for fast inhibitory control of the activity level of recurrent networks (Wong, 1987), is sometimes referred to as multiplicative (or, rather, divisive) in nature. This is because it acts via Cl^- channels of relatively

large total conductance (Connors et al., 1988) and inversion potential not far below the resting potential; hence its effect is more shunting than hyperpolarizing. If located on proximal dendritic branches or on the soma (Andersen, Eccles, & Böyning, 1964), it can be modeled to a first approximation as producing a division of the current resulting from more distal inputs (Abbott, 1991).

Purely multiplicative inhibition acting on excitatory cells would lead to substitute the first of Eqs.4.2 with

$$\tau_E \dot{\nu_E} = -\nu_E + J_E^E(\nu_I)\nu_E + \nu_E^{aff} \tag{4.13}$$

i.e., the excitatory self-coupling is now a function of the average firing rate of inhibitory units (the second equation can be modified as well, but this is easily seen to be irrelevant for the present discussion). To the extent that afferent inputs are absent or negligible, at the fixed point the self-coupling takes the value 1, thereby automatically ensuring stability, at least in the sense of Eqs.4.4 (since the terms in $J_E^E - 1$ disappear from the inequalities). Real inhibition, of course, is not purely multiplicative, however the situation holding in this limit clarifies that under appropriate conditions (if inhibition is multiplicative to a sufficient degree) the stability of recurrent networks against runaway excitation is automatically guaranteed.

As for the upper limit on storage capacity, we have checked, by repeating previous analyses (Treves & Rolls, 1991) of recurrent associative memories of threshold-linear units with a gain γ now dependent on the average inhibitory rate, that the same exact equations determine the storage capacity. Such a result stems from the fact that, by acting on the gain, inhibition now keeps the effective J_E^E entering the stability analysis close to 1, but it leaves identical, as the analytical treatment shows, the capacity equations. This confirms that the form of inhibition used has no effect on such absolute limit (a limit which with subtractive inhibition was far beyond what could be achieved in practice). We have also carried out simulations of a simple network model with 3000-5000 threshold-linear units as used in the analytical calculation, at several sparseness values. We estimated storage capacity from the simulations by progressively increasing memory load, and determining the critical level at which no retrieval of any stored pattern was possible. Results are shown in Fig. 4.1, and confirm the analytical prediction, which is the exact reproduction of previous analyses with subtractive inhibition (Treves & Rolls, 1991). Note that a value of the storage parameter $\alpha = 0.3$, for example, corresponds to 900 stored patterns.

We have then carried out simulations of a more detailed network model with spiking units and conductance-based synaptic action, both in order to understand whether realistic inhibition still allows retrieval of more than 2-3 patterns (the limit we had on similar simulations with purely subtractive inhibition), and, once disposed of this limitation, in order to address anew, in a realistic context, the issue of the time scales for recurrent memory retrieval.

4.4 Simulations show stability and fast retrieval

The simulated network consisted of $N_{ex} = 800$ excitatory units and $N_{in} = 200$ inhibitory ones. Each integrate-and-fire unit represents a neuron as a single branch, compartmented dendrite through which the cell receives all its input, and a point-like soma, where spikes are generated. Though very simple, the compartmental model is still computationally demanding, and severely limits the size of the network that we could implement on a Linux workstation. The current flowing from each compartment to the external medium is written

$$I(t) = g_{leak}(V(t) - V^0) + \sum_j g_j(t)(V(t) - V_j), \tag{4.14}$$

where g_{leak} is a constant, passive leakage conductance, V^0 the membrane resting potential, $g_j(t)$ the value of the j-th synapse conductance at time t, and V_j the reversal potential of the j-th synapse. $V(t)$ is the potential in the compartment at time t. Synaptic conductances have an exponential decay time behavior, obeying the equation

$$\frac{dg_j}{dt} = -\frac{g_j}{\tau_j} + \Delta g_j \sum_k \delta(t - t_k^j), \tag{4.15}$$

where τ_j is the synaptic decay time constant, and Δg_j is the amount the conductance is increased when the presynaptic unit fires a spike. Δg_j thus represents the (unidirectional) coupling strength between the pre-synaptic and the post-synaptic cell. t_k^j is the time at which the pre-synaptic unit fires its k-th spike.

For each time step of $1ms$, the cable equation for the dendrite is integrated (MacGregor, 1987) with a finer time resolution of $0.1ms$ and the somatic potential is compared with the spiking threshold V^{thr}. When this is exceeded, post-synaptic conductances are updated and the somatic potential is reset to the after-hyperpolarization value V^{ahp} throughout the neuron.

Connections from excitatory to inhibitory, from inhibitory to excitatory, and between inhibitory units are taken to be homogeneous, that is, all of the same strength. Synaptic parameters depend only on the type of pre-synaptic and post-synaptic unit. The connectivity level is 0.25, between populations and 0.5 within the inhibitory population, that is, each unit synapses onto a fraction of the units of the receiving population, chosen at random. The excitatory units, in contrast, are all connected to each other. This very high connectivity, out of the actual anatomical range, is necessary because of the small size of the simulated network, to produce sufficient statistical averaging in the synaptic input to each unit.

Excitatory-to-excitatory connections encode in their strength p memorized patterns of activity η_i^μ, consisting of binary words with sparseness (in this simple binary

case the fraction of 1s, or active cells in the pattern) $a = 0.1$. Encoding is implemented through a modified Hebb rule. In contrast with Eq. 4.7, which includes a baseline weight, all conductances are initially set to zero and then, for each pattern, the synapse from the i-th to the j-th unit is modified by a covariance term

$$\Delta g = \frac{g_{EE}}{C_{EE}} \left(\frac{\eta_i^\mu}{a} - 1 \right) \left(\frac{\eta_j^\mu}{a} - 1 \right), \tag{4.16}$$

If the conductance becomes negative, it is reset to zero. Memories are therefore stored through a "random walk with one reflecting barrier" procedure. The barrier acts as a "forgetting" mechanism (Parisi, 1986), as whenever the conductance value bumps into the barrier, it loses memory about the previously presented patterns. As there is no upper boundary, the average value of excitatory connection strengths grows with the number of memory items learned. The network is tested at low memory loading ($p = 10$). A systematic study of the storage capacity of the net would not be very meaningful because of the small size of the network.

The excitatory synapses impinge on the distal end compartment of the post-synaptic dendrite, and they have a positive reversal potential (referred to resting membrane potential). Inhibitory synapses are distributed uniformly along the dendritic body, and they have a reversal potential equal to the resting membrane potential (except for the simulations in Fig. 4.2). Inhibition is therefore predominantly shunting, with a geometry very similar to the one considered in Abbott (1991), leading to a mainly multiplicative effect on the post-synaptic firing rate. Table 1 summarizes the parameters used for the simulations.

Once the connection matrix is constructed, a test of the retrieval dynamics was performed according to the following protocol: The network is activated by injecting a current in a random fraction $a = 0.1$ of the units (Fig. 4.2, panel A). The excitatory and the inhibitory population become diffusely active. Notice that units active in the memory pattern being tested are on average slightly more active than the other units. This is explained by the fact that they have on average a slightly stronger excitatory input, because the memory being tested contributes a positive term in the random walk construction of the connection strengths. Since p is not too large even a single term makes a difference (Amit & Brunel, 1996).

After 100 msec, the random current is replaced with a *cue* current, injected in a fraction $a + \varrho(1 - a)$ of the units active in the pattern being tested and in a fraction $a(1 - \varrho)$ of the units inactive in the pattern. In this way, the cue is again a binary word with sparseness $a = 0.1$, and ϱ is the the average correlation between pattern and cue, which in the runs shown in the figures was set at $\varrho = 0.3$.

The cue current lasts for 300 msec. The average firing rate for the "1" units is much higher than for the "0"ones. When the cue current is removed, the "1" units

Quantity	Symbol	Value
# Exc. Cells	N_E	800
# Inh. Cells	N_I	200
Corruption level	ρ	0.3
Random activity period	t_{init}	100 (msec)
Cue period	t_{cue}	300 (msec)
Retrieval period	t_{retr}	200 (msec)
Sampling time window	t_{win}	30 (msec)
# Dendritic compartments	N_{cmp}	10
Dendritic compartment leakage conductance	G_0^d	6.28×10^{-12} (S)
Somatic compartment leakage conductance	G_0^s	5×10^{-9} (S)
Dendritic-dendritic axial conductance	G_0^{dd}	2.25×10^{-7} (S)
Exc. Somatic capacitance	$C_{soma,E}$	$0.5 - 4 \times 10^{-10}$ (F)
Inh. somatic capacitance	$C_{soma,I}$	5×10^{-12} (F)
Cue current	I_{cue}	0.25 (nA)
Firing threshold potential (exc.)	Θ_E	32 (mV)
Firing threshold potential (inh.)	Θ_I	25 (mV)
After-spike hyperpolarization potential	V_{ahp}	-15 (mV)
Excitatory-excitatory connectivity level	C_{EE}	1
Excitatory-inhibitory connectivity level	C_{EI}	0.25
Inhibitory-excitatory connectivity level	C_{IE}	0.25
Inhibitory-inhibitory connectivity level	C_{II}	0.5
"Unitary" excitatory-excitatory synaptic conductance [see (16)]	g_{EE}	$5 \times 10^{-8}(S)$†
Excitatory-inhibitory synaptic conductance	g_{EE}	$4 \times 10^{-9}(S)$
Inhibitory-excitatory synaptic conductance	g_{EE}	$2 \times 10^{-8}(S)$
Inhibitory-inhibitory synaptic conductance	g_{EE}	$9 \times 10^{-10}(S)$
Excitatory-inhibitory synaptic time constant	τ_{EI}	1 (msec)
Excitatory synaptic equilibrium (reversal) potential	V_E	65 (mV)
Inhibitory synaptic equilibrium (reversal) potential	V_I	0 (mV)
Excitatory-excitatory synaptic time constant	τ_{EE}	5 - 40 (msec)
Excitatory-inhibitory synaptic time constant	τ_{EI}	1 (msec)
Inhibitory-excitatory synaptic time constant	τ_{IE}	1 (msec)
Inhibitory-inhibitory synaptic time constant	τ_{II}	1 (msec)

Table 4.1: Parameters used for integrate-and-fire simulations. Ranges are indicated for quantities which varied within runs. Potential values are referred to membrane resting potential. † The excitatory-excitatory synaptic conductance is scaled when the synaptic time constant is varied, to preserve the total charge transmitted during a synaptic event (see text). The value given is the one used for $\tau_{EE} = 20$ (msec). Simulations in Fig. 2 are an exception as concerns inhibitory reversal potential (see caption of the figure) and t_{retr}, which is set at 500 (msec).

sag briefly but then recover and stay steadily active, while activity in the others decays at zero firing or at a very low level. The memory pattern has therefore been successfully retrieved. To test the specific effect produced by the type of inhibition, we performed the stepwise manipulation shown in Fig. 4.2. First (panel B), all inhibitory connections onto excitatory cells were moved to the end of the dendritic tree, colocalized with excitatory inputs. This made them somewhat less "multiplicative", and also weaker. The result is that inhibition becomes unable to suppress the firing of excitatory units which should be quiescent, and the network fails to retrieve correctly (the residual difference between "1" and "0" units being due to the finite-p effect mentioned above). To make inhibition stronger again while maintaining its subtractive character, the equilibrium potential of inhibitory synapses was lowered in panels C-F in steps of 10 mV. The result is that inhibition tends to suppress activity across excitatory units, without ever allowing the retrieval state to re-emerge after removing the cue. This manipulation then indicates that altering the form of inhibition makes the network cross its capacity limit. Since even the first form, with the inputs spread along the dendritic tree, is far from being purely multiplicative, this capacity limit is well below the upper limit predicted by non-dynamical calculations.

The simulations were repeated varying the neural and synaptic parameters, namely the excitatory synaptic time constant (changing at the same time the synaptic conductance to keep the strength of the connection invaried) and the somatic capacitance, in order to vary the firing rate. The inhibitory synaptic time constant was kept smaller than the excitatory time constant, in order to speed up the stabilizing effect of recurrent inhibition.

To assess to quality of retrieval we have taken the same information theoretical measure used when recording from behaving animals (Rolls, Treves, & Tovee, 1997; Treves et al., 1996): The retrieval protocol is repeated for up to 30 "trials" for each stored memory. 10 randomly selected excitatory units are "recorded", i.e., sampled for the number of spikes they fire in a time window of 30 ms. The window slides with a step of 5 msec spanning the entire simulated time course. The firing rate vector thus constructed at any time step of each trial is then *decoded*. This is done (Rolls et al., 1997) by matching it with the $p = 10$ mean firing rate vectors produced at the same time step when testing the retrieval of each of the memories, and finding the closest match. The result of decoding all the trials is a probability table $P(s'|s)$ containing the likelihood that when testing for memory s the activity of the sample of units was decoded as matching the average vector from pattern s'. The *mutual information* between the actual and decoded pattern

$$I(s,s') = \sum_s \frac{1}{p} \sum_{s'} P(s'|s) \log_2 \frac{P(s'|s)}{P(s')} \tag{4.17}$$

was calculated and then corrected for limited sampling (Treves & Panzeri, 1995;

Panzeri & Treves, 1996). To reduce fluctuations, results were averaged, at each time step, over a number of samples of recorded units from the same run. The resulting quantity is a sensitive measure of how well the activity of the network in the time window can be used to discriminate which cue was presented, and unlike simpler measures (such as the correlation of the firing vector with the underlying memory pattern) can be used with identical procedures in simulations and in recording experiments.

In fig.4.3 we show the time course of the information for different values of the excitatory time constant. The mutual information stays close to zero during the random activity period (the small baseline is a remnant of the finite size error after the correction), and when the cue is presented it rises steadily to an equilibrium value, which depends on the correlation between the cue and the pattern, with a time course well fitted by a saturating exponential. This appears to be consistent with the linearized analysis for transients (Treves, 1993), and indicates that the transient modes that are activated in this condition belong to a single family, i.e. they share the same real part of the time constant. The time constant from the exponential fit is in a close-to-linear relationship with the synaptic (inactivation) time constant, as shown in fig. 4.4, with in the case shown a best-fit proportionality coefficient of 2.538.

Varying the firing rate does not appear to have a comparable effect on the transient time constant: fig. 4.5 plots the transient time constant relative to different values of somatic capacitance, corresponding to firing rates ranging from $\sim$ 15 to $\sim$ 100 Hz.

When the cue is removed, the information rises again very rapidly to a higher equilibrium value, as the network is no longer constrained by the noisy cue, indicating that the network is acting as an "error corrector" during this later phase. The second transient is very rapid indeed, and it is in fact masked by an artifact induced by the finite size of the time window used to measure information (the artifact is that, during the time window, what the measure reflects is actually a weighted sum of the lower value before cue removal and the higher value that is reached in a very short time). In fact, if one shrinks the sample window size, this linear raise shortens correspondingly (not shown). Although the actual time structure of this transients is still to be clarified, it seems clear that it follows a very different mode in this path to equilibrium. The final approach to the retrieval attractor is thus essentially immediate. In fact, Tsodyks and Sejnowski (1995) showed that in integrate-and-fire networks in a "balanced" regime, dominated by fluctuations, the network can change state very rapidly as a reaction to a change in input currents. In the regime studied in that work, excitatory and inhibitory input are precisely balanced so that their absolute value is much larger than their algebraic sum. Input fluctuations measured in Excitatory Post-Synaptic Potentials (EPSP) are much larger than the voltage difference between spike threshold and after-spike hyperpolarization. That is, fluctuations in the input can very easily drive the neuron to threshold, so that

dynamics cannot be described in the mean-field framework. The mean-field solution of the dynamics is consistent with the time-scales behavior we obtain for the first transient (cue onset), while the second transient (cue removal) in our simulations is much faster, suggesting the possibility that the dynamics regime at that point be more similar to that of Tsodyks and Sejnowski (1995), although the whole issue deserves further investigation.

Finally, in fig. 4.6 we show the information behavior of the network when the excitatory collaterals are made information-less, or memoryless, by giving them all the same strength. A finite, small amount of information is seen in the cue phase only, at a much smaller level than for the structured network, and it falls to zero as the cue is removed. This demonstrates that selective activity, and in particular the capability of this network to retrieve memory patterns, depends crucially on the information encoded on its collaterals.

4.5 Implications for recurrent processing in the brain

The more effective control that shunting inhibition may exert on runaway recurrent excitation, compared with subtractive inhibition, is an intuitive principle, that has informed direct experimental studies (Wong, 1987). What has been shown here is how shunting inhibition, in particular, may help avoid a specific conflict between stability and extensive memory storage that would otherwise prevent the applicability of the abstract concept of a recurrent auto-associator to actual recurrent networks in the brain.

An attempt to demonstrate the large conductance changes that may underlie shunting inhibition (Douglas & Martin, 1991a) has not confirmed the expectation; however it is unclear to what extent the model used (the striate cortex of anaesthetized cats) is relevant to the conditions we considered of massively reverberating excitation.

Having ensured the possibility of stable asynchronous firing attractor states, simulations of a model network with spiking units and synaptic conductances have been used to confirm and extend earlier analytical results on the time required for memory retrieval mediated by recurrent processing to occur. The time course of the initial approach to the attractor state is, as in the analytical treatment, a saturating exponential, or a mixture of exponentially relaxing transient modes with similar (real part of the) time constant. This retrieval time constant is a linear function of the time constant for the inactivation of excitatory synaptic conductances, and depends only mildly on prevailing firing rates or on neuronal time scales (as determined, e.g., by membrane capacitance).

In practice, the contribution of recurrent processing, in this particular instance of an auto-associator, can be dominant already within a few tens of msec (with the parameters of fig. 4.3, within 2.5 times of the synaptic time constant, which can be thought of as being in the 10 msec range (Colquhoun et al., 1992)). This leads to the conclusion that at least *local* recurrent processing can be fast, and that it is wrong to exclude its relevance in cases in which neuronal activity is found to acquire its selectivity within a few tens of msec of its onset (Thorpe & Imbert, 1989; Treves et al., 1996).

This result lends credibility to the hypothesis that recurrent auto-association may be an ubiquitous function of local recurrent circuits throughout neocortex, as well as possibly the main function of recurrent connections in the hippocampal CA3 region (Treves & Rolls, 1994). At the same time, it raises the possibility of a direct manipulation of the time for such a function to be executed, by acting on the inactivation kinetics of synaptic AMPA channels.

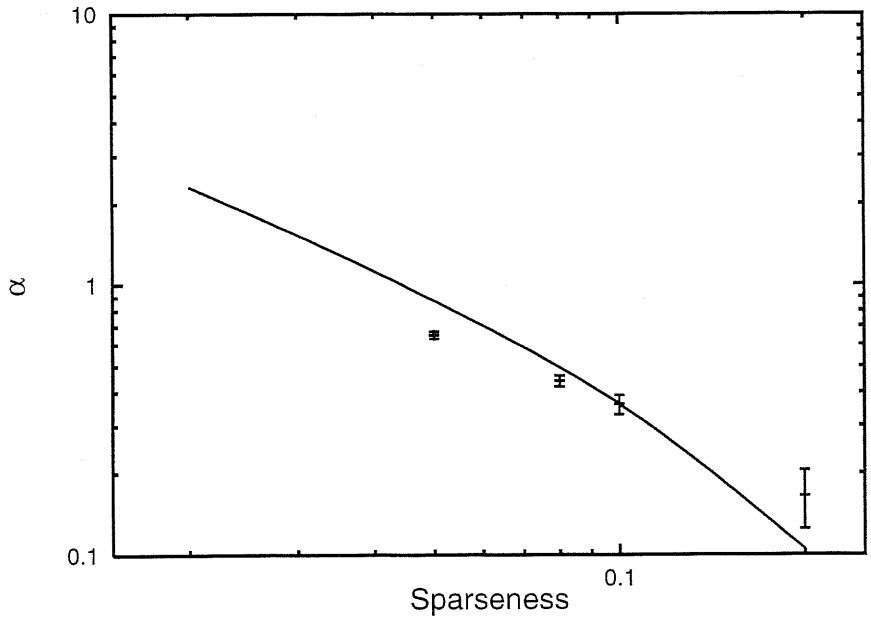

Figure 4.1: Simulation results for the capacity of a network of 3000 threshold-linear neurons (5000 for $a = 0.05$) are compared with the theoretical prediction (full line) at different values of the sparseness a. The prediction arises from equations identical to those found by Treves (1990)

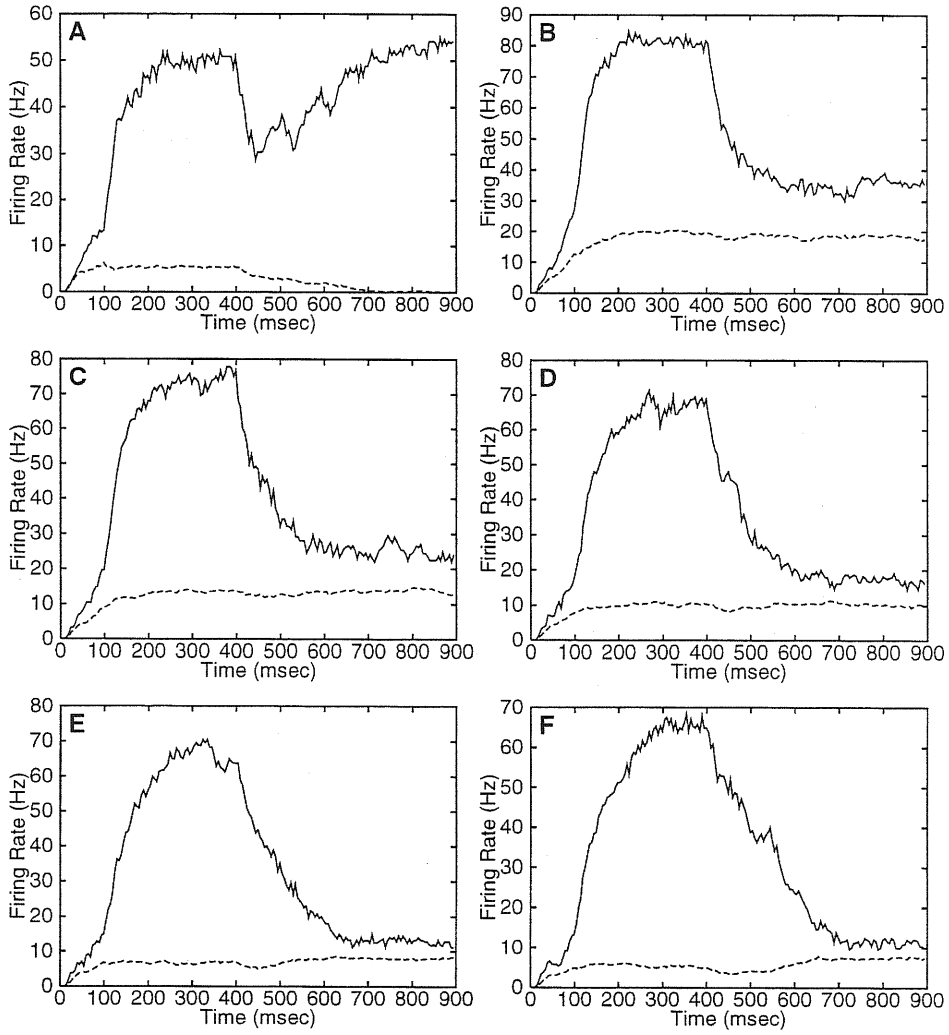

Figure 4.2: Firing rates computed with a time window of 30 msec are plotted for excitatory units for different geometries and reversal potential V_I. Units are divided between the "1" population (upper trace), active in the recalled memory, and the "0" population (lower trace), that was silent in the recalled memory. (A): $V_I = 0mV$ with respect to membrane equilibrium potential and inhibitory synapses are distributed along the dendritic body. In this condition inhibition acts to some extent multiplicatively on the firing rate. Efficient retrieval of the memory is shown by sustained activity in the "1" population and complete activity suppression in the "0" population after the cue has been removed. (B-E): Inhibitory synapses are located on the edge of the dendritic cable. Reversal potential V_I is $0mV$ (with respect to equilibrium) (B), $-10mV$ (C), $-20mV$ (D), $-30mV$ (E) and $-40mV$ (F). Whatever the reversal potential, the two populations are never satisfactorily discriminated.

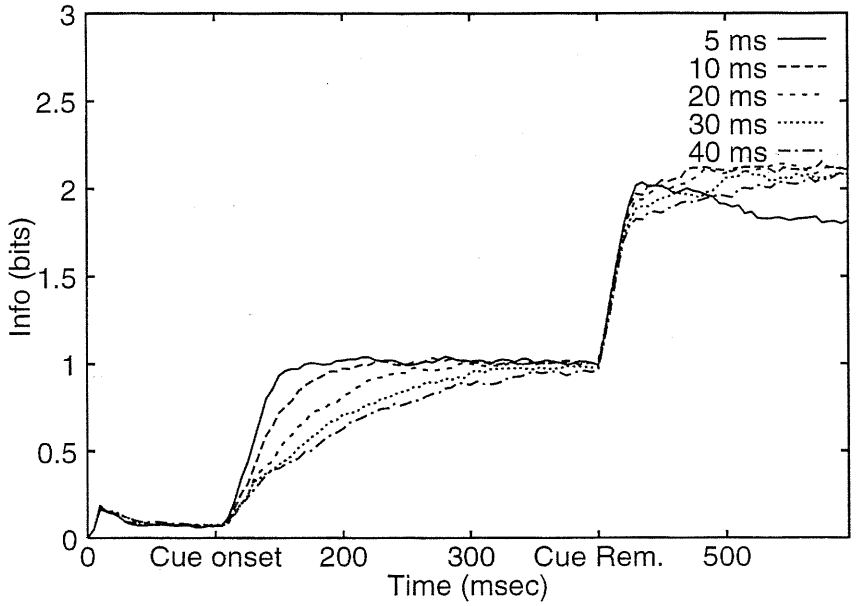

Figure 4.3: Information time course for different values of synaptic time constant. The transient corresponding to cue onset is well fitted by an exponential function. The raise is faster with shorter synaptic time constant.

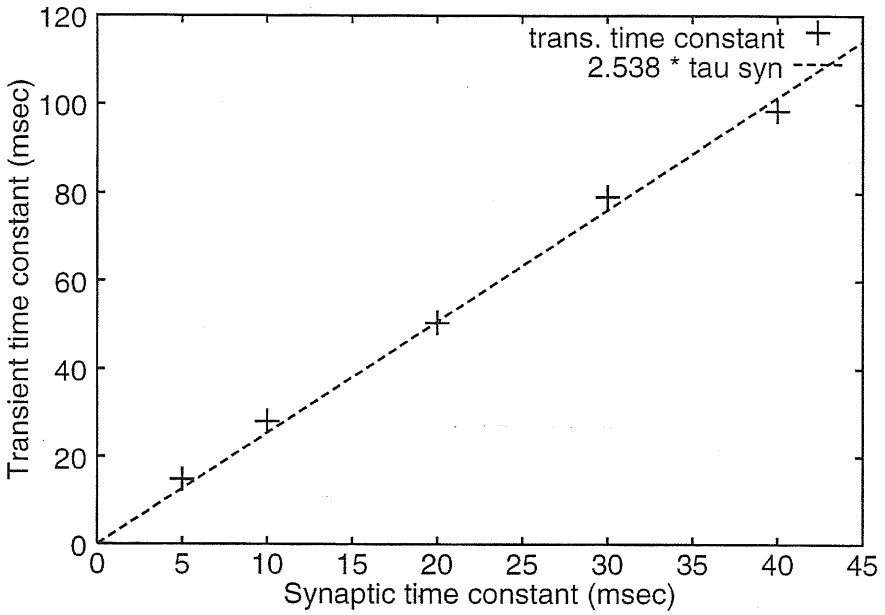

Figure 4.4: Transient time constant plotted against excitatory synaptic time constants. The firing rates were the same in each case, since conductance values were rescaled in order to equalize the charge entering into the cell through the synapse. The best linear fit line is shown. The slope of the fitted line is 2.538.

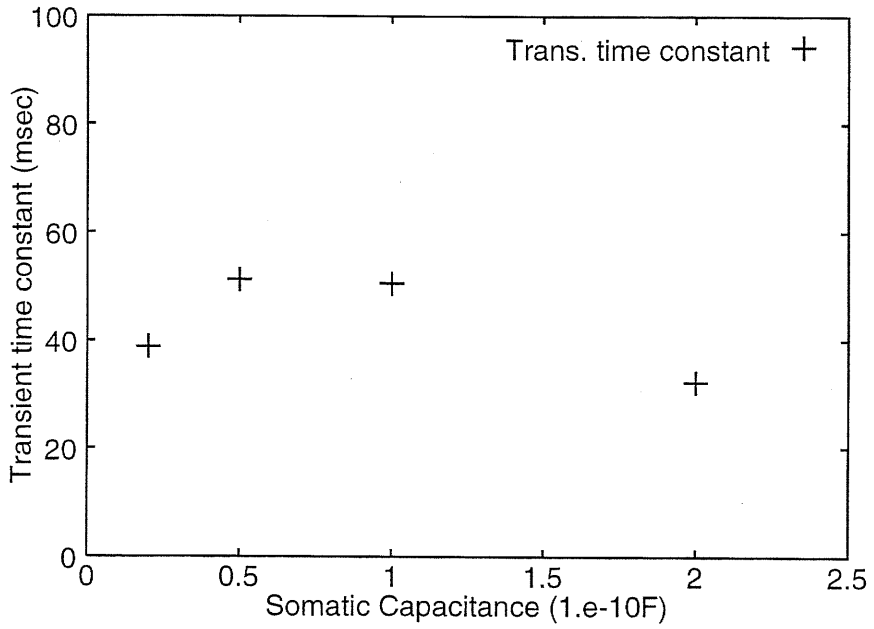

Figure 4.5: Transient time constant plotted for different values of somatic capacitance. Firing rates during the cue phase ranged correspondingly from 15 to 100 Hz. No clear dependence of the information time course is apparent when firing rates are varied in this way.

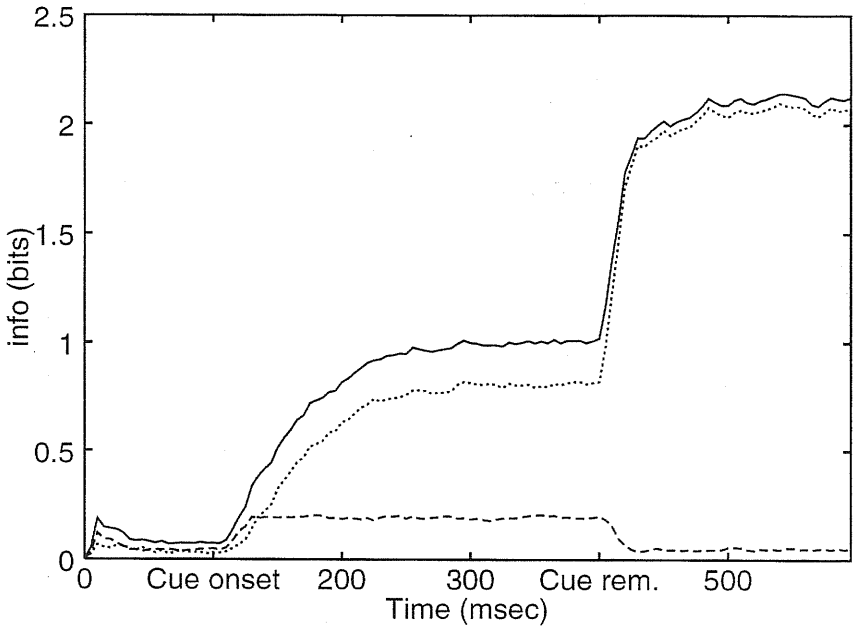

Figure 4.6: Information time course plotted for the structureless network compared with time course for the network structured as in previous figures. During the cue phase, information reaches just a fraction of the steady state value in the structured case. After the cue is removed, information decays to zero reflecting the absence of self-sustained activity.

Appendix A

Replica symmetric free energy for the "dot product" kernel model

The replica symmetry free-energy reads

$$f = -T\left\langle\left\langle \int Dz \ln \mathrm{Tr}(h, h_2)\right\rangle\right\rangle - \frac{1}{2}\sum_{(\sigma),l}|\mathbf{x}^{(\sigma),l}|^2 - B(x) - \sum_{(\sigma),l}(|\mathbf{s}^{(\sigma),l}|x + \mathbf{s}^{(\sigma),l}x^{(\sigma),l}) - \sum_{(\sigma),l}\mathbf{t}^{(\sigma),l}\mathbf{x}^{(\sigma),l} - tx - r_0y_0 + r_1y_1 + \frac{\alpha d}{2\beta}\left(\ln[1 - T_0\beta(y_0 - y_1)] - \frac{\beta y_1}{1 - T_0\beta(y_0 - y_1)}\right) \tag{A.1}$$

very much like eq.19 in (Treves, 1990) and with the same meaning for symbols, except that the population vector $\mathbf{x}^{(\sigma),l}$ plays the role of the overlap x^σ, the vector Lagrange multiplier $\mathbf{t}^{(\sigma),l}$ appears instead of its scalar counterpart t^σ and the dimensionality d appears multiplying the last term. h and h_2 are

$$h = -t - \sum_{(\sigma),l}\mathbf{t}^{(\sigma),l}\cdot\vec{\eta}^{(\sigma),l} - z(-2Tr_1)^{1/2} \tag{A.2}$$

$$h_2 = r_1 - r_0. \tag{A.3}$$

$\langle\langle \ldots \rangle\rangle$ means averaging over the distribution of p.f.c.'s $\vec{\eta}$. T is the noise level in the thermodynamic analysis. T_0 is defined here as:

$$\left\langle\left\langle(\mathbf{t}_1^l\cdot\vec{\eta}^{(\mu),l})(\mathbf{t}_2^l\cdot\vec{\eta}^{(\mu),l})\right\rangle\right\rangle = \frac{T_0}{d}\mathbf{t}_1^l\cdot\mathbf{t}_2^l \tag{A.4}$$

and it is found to be equal to 1/2 in 1-D and to 1 for the 2-D torus.

The saddle point equations can be found from this equations, and t and $\mathbf{t}^{(\sigma),l}$ can be eliminated, in the same way as in (Treves, 1990). Carrying on the calculation the $T = 0$ equations eventually reduce to two equations in the two variables (in the case of a single "condensed" map):

$$w = \frac{[b(x) - \theta]}{\rho} \tag{A.5}$$

$$\mathbf{v}^l = \frac{(\mathbf{x}^l + \mathbf{s}^l)}{\rho}. \tag{A.6}$$

Take for simplicity $|\mathbf{v}^l| = v$ (while the direction is set by $\mathbf{v}^l \propto \mathbf{s}^l$). The two equations read:

$$E_1(w, \mathbf{v}) \equiv (A_1 + \delta A_2)^2 - \alpha A_3 = 0 \tag{A.7}$$

$$E_2(w, \mathbf{v}) \equiv$$

$$(A_1 + \delta A_2)\left(\frac{1}{gT_0(1+\delta)} + \alpha - A_2\right) - \alpha A_2 = 0 \tag{A.8}$$

where $\delta = |\mathbf{s}^1|/|\mathbf{x}^1|$ is the relative importance of the external field and:

$$A_1(w, v) = \frac{1}{v^2 T_0}\left\langle\left\langle \mathbf{v}^l \cdot \vec{\eta}^l \int^+ Dz(w + \sum_l \mathbf{v}^l \cdot \vec{\eta}^l - z)\right\rangle\right\rangle - \left\langle\left\langle \int^+ Dz\right\rangle\right\rangle \tag{A.9}$$

$$A_2(w, v) = \frac{1}{v^2 T_0}\left\langle\left\langle \mathbf{v}^{(1)} \cdot \vec{\eta}^{(1)} \int^+ Dz(w + \sum_l \mathbf{v}^l \cdot \vec{\eta}^l - z)\right\rangle\right\rangle \tag{A.10}$$

$$A_3(w, v) = \left\langle\left\langle \int^+ Dz(w + \sum_l \mathbf{v}^l \cdot \vec{\eta}^l - z)^2\right\rangle\right\rangle \tag{A.11}$$

Dz is the Gaussian measure $(2\pi)^{-1/2} e^{-z^2/2} dz$. The + sign on the integral means that integration extremes are chosen such that $(w + \sum_l \mathbf{v}^l \cdot \vec{\eta}^l - z) > 0$.

When the quenched average on the η's is performed, A_1, A_2, A_3 reduce to (for the d-dimensional torus $\mathcal{C}^d$):

$$A_1(w, v) = \frac{1}{(2\pi)^d v T_0} \int d\theta^l (\sum_l \cos\theta^l) \times$$

$$[(w + v\sum_l \cos\theta^l - vT_0)\Phi(w + v\sum_l \cos\theta^l) +$$

$$(w + v\sum_l \cos\theta^l)\sigma(w + v\sum_l \cos\theta^l)] \tag{A.12}$$

$$A_2(w,v) = \frac{1}{(2\pi)^d v T_0}\int d\theta^l(\sum_l \cos\theta^l)\times$$
$$[(w+v\sum_l \cos\theta^l)\Phi(w+v\sum_l \cos\theta^l)+$$
$$(w+v\sum_l \cos\theta^l)\sigma(w+v\sum_l \cos\theta^l)] \quad (A.13)$$

$$A_3(w,v) = \frac{1}{(2\pi)^d}\int d\theta^l[1+(w+v\sum_l \cos\theta^l)^2]\Phi(w+v\sum_l \cos\theta^l)+$$
$$(w+v\sum_l \cos\theta^l)\sigma(w+v\sum_l \cos\theta^l) \quad (A.14)$$

where

$$\Phi(x) = \int_{-\infty}^{x}\frac{dz}{\sqrt{2\pi}}e^{\frac{-z^2}{2}} \quad (A.15)$$

$$\sigma(x) = \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}. \quad (A.16)$$

Appendix B

Generic kernel, extreme dilution

Let us consider the one-dimensional case first, and consider the kernel

$$K(\vec{r}-\vec{r}') = \hat{K}(\vec{r}-\vec{r}') - \frac{2}{|M|} = e^{-|\vec{r}-\vec{r}'|} - \frac{2}{|M|}. \tag{B.1}$$

Eq. 3.42 can be written

$$v(\vec{r}) = g\mathcal{N}\left(\int d\vec{r}'\hat{K}(\vec{r}-\vec{r}')v(\vec{r}') + \hat{w}\right) \tag{B.2}$$

where

$$\hat{w} = w - \frac{2}{|M|}\int d\vec{r}'v(\vec{r}'). \tag{B.3}$$

For the purpose of finding α_c, maximizing with respect to $\hat{w}$ is equivalent to maximizing with respect to w.

To solve eq.3.42, the transformation

$$u(\vec{r}) = \mathcal{N}^{-1}\left(\frac{v(\vec{r})}{g}\right) \tag{B.4}$$

is used, which results in

$$u(\vec{r}) = g\int d\vec{r}'\hat{K}(\vec{r}-\vec{r}')\mathcal{N}[u(\vec{r}')] + \hat{w}. \tag{B.5}$$

By differentiating twice we get

$$u''(\vec{r}) = -2g\mathcal{N}[u(\vec{r})] + u(\vec{r}) - \hat{w} = -\frac{d}{du}\mathcal{U}[u(\vec{r})] \tag{B.6}$$

where

$$\mathcal{U} = \int^{u} du' \left(2g\mathcal{N}(u') - u' + \hat{w}\right). \tag{B.7}$$

The differential equation (B.6) is *locally* equivalent to the non-linear integral equation (B.5) Equation (B.6) must be solved numerically. As in the single map case, not all the solutions of the differential equation (B.6) are solution of the integral equation (B.5). Solutions of (B.6) are solution of (B.5), strictly speaking, only in the case $M \equiv \mathbb{R}^d$. Nevertheless, we force the equivalence since, also in the case of limited environments, with periodic boundary conditions, possible pathologies are not important for solutions with activity concentrated far from the boundaries.

In order to classify the solutions of eq. B.6 it is useful to study the "potential function" $\mathcal{U}$. If w is negative and large enough in absolute value, $\mathcal{U}(u)$ has a maximum and a minimum at the two roots of equation

$$\frac{d}{du}\mathcal{U}(u) = 2g\mathcal{N}(u) - u + \hat{w} = 0, \tag{B.8}$$

or, in terms of v:

$$v = g\mathcal{N}(2v + \hat{w}), \tag{B.9}$$

corresponding to constant solutions of eq. 3.42. We look for solutions representing a single, symmetric peak of activity centered in $r = 0$. We therefore need to solve the Cauchy problem given by eq. B.6 with the initial conditions:

$$u(0) = u_0 \tag{B.10}$$
$$u'(0) = 0 \tag{B.11}$$

From fig. B.1 it is clear that if $u_0 > u^*$ the solution will escape to $-\infty$ for r tending to infinity. This will correspond to v tending asymptotically to 0, and this solution cannot be a solution for the integral equation (3.42) as the asymptotic value must be a root of (B.9).

The solutions of the problem with $u_0 < u^*$ are periodic, corresponding to multiple peaks of activity, and they are discarded as unstable with the same arguments holding for the single map case. There is also the constant solution

$$u(r) = u_{min}, \tag{B.12}$$

which obviously will not correspond to space related activity. The solution corresponding to the single activity peak can only be the one with $u_0 = u^*$. It tends asymptotically to u_{max}. This solution can be found numerically and inserted in

eq.3.46 to find the value of α associated with the pair $(g, \hat{w})$. The solution will only be present for values of $\hat{w}$ for which $\mathcal{U}(u)$ has the extremal points u_{max} and u_{min}, that is:

$$\hat{w} < \hat{w}^* \tag{B.13}$$

where w^* is equal to $-2g\mathcal{N}(u^*) + u^*$ and u^* is the root of the equation:

$$\Phi(u) = \frac{1}{2g} \tag{B.14}$$

obtained by derivating twice $\mathcal{U}$, and this shows that eq. B.5 cannot have solutions for $g < 1/2$, as in the single map case.

In the two dimensional case, we can consider the kernel

$$K(\vec{r} - \vec{r}') = \hat{K}(\vec{r} - \vec{r}') - \frac{2}{|M|} \tag{B.15}$$

where $\hat{K}$ is the kernel having Fourier transform:

$$\hat{K}(\mathbf{p}) = \frac{2}{1 + \mathbf{p}^2} \tag{B.16}$$

The solution is worked out in the same way with the transformation (B.4) and application of Laplacian. If we consider solutions with circular symmetry and pass to polar coordinates (r, ϕ), the equation for r reads:

$$u''(r) + \frac{1}{r}u'(r) = -2g\mathcal{N}[u(r)] + u(r) - w \tag{B.17}$$

We still have a single peak solution with tends asymptotically to u_{max}, but in this case we cannot rely on the $\mathcal{U}$ function argument to find the initial condition at $r = 0$, which has to be found numerically.

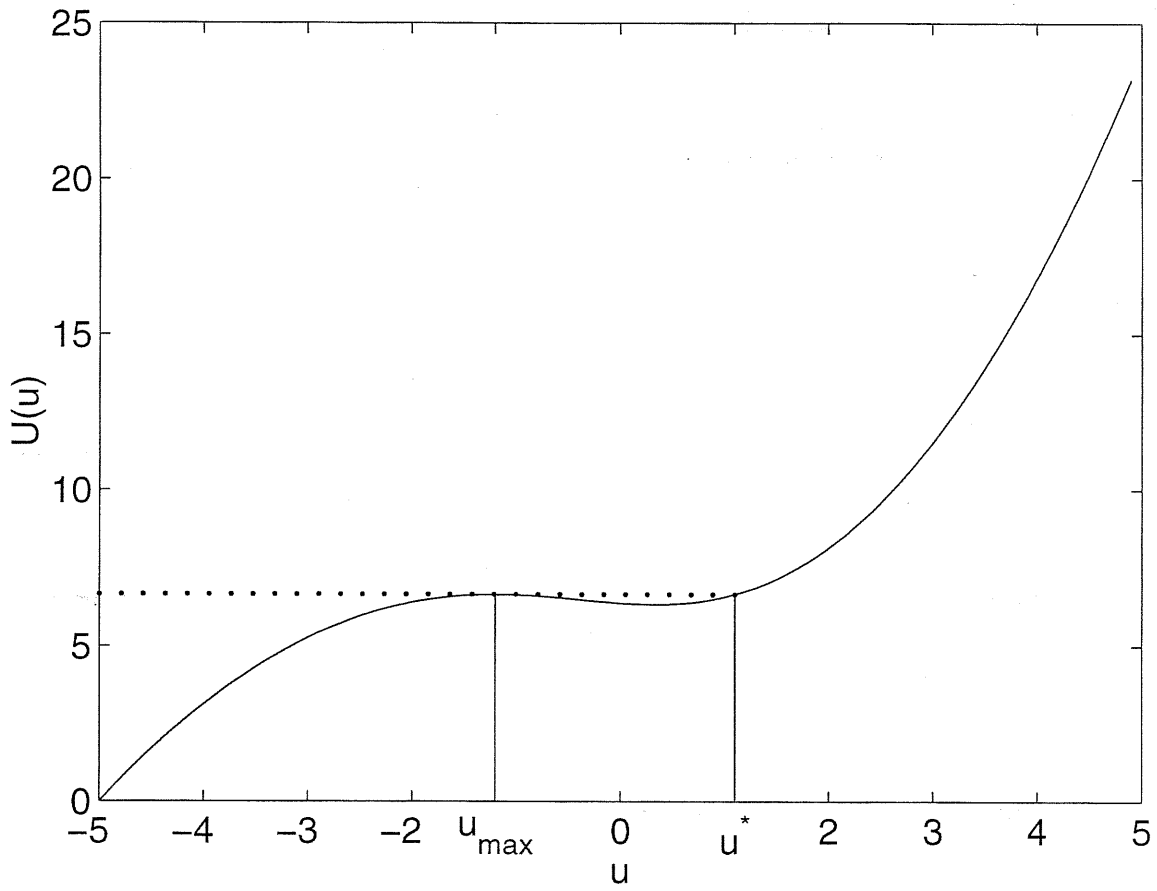

Figure B.1: The "potential" function $\mathcal{U}(u)$ defined by eq. B.7 and entering the differential equation eq. B.6. Solutions with $u'(0) = 0$ and $u(0) = u_0$, with $u_{max} < u_0 < u^*$ are oscillating. The solution with $u_0 = u^*$ is the one we seek, asymptotically approaching u_{max} as $r \to \infty$.

Appendix C

Replica free energy calculation for the generic kernel

Again we will consider an environment M with periodic boundary conditions. We assume that there exists a kernel L such that

$$\int d\vec{r}'' L(\vec{r} - \vec{r}'') L(\vec{r}'' - \vec{r}') = |M| K(\vec{r} - \vec{r}'). \tag{C.1}$$

Instead of the vector order parameter $\mathbf{x}^\mu$ we used for the dot-product kernel case (or of the scalar overlap x^μ of (Treves, 1990)) we can use the functional order parameter

$$x^\mu(\vec{r}) = \frac{1}{N} \sum_i [L(\vec{r} - \vec{r}_i^{\,\mu})] V_i \tag{C.2}$$

in terms of which the interaction part of the Hamiltonian (3.59) reads

$$\begin{aligned}
&\frac{1}{2} \sum_i \sum_{j \neq i} J_{ij} V_i V_j = \\
&\frac{|M|}{2N} \sum_\mu \sum_i \sum_j \left[K(\vec{r}_i^{\,\mu} - \vec{r}_j^{\,\mu}) - \bar{K} \right] V_i V_j - \frac{\alpha |M|}{2} (K(0) - \bar{K}) \sum_i V_i^2 = \\
&\frac{1}{2} N \sum_\mu \int d\vec{r} [x^\mu(\vec{r})]^2 - \frac{\alpha |M|}{2} (K(0) - \bar{K}) \sum_i V_i^2
\end{aligned} \tag{C.3}$$

Introducing the "square root" kernel L allows us to perform the standard Gaussian transformation manipulation and to carry out the mean field free energy calculation

in the replica symmetry approximation:

$$f = -T\left\langle\left\langle\int Dz \ln \mathrm{Tr}(h,h_2)\right\rangle\right\rangle - \frac{1}{2}\sum_\sigma \int_M d\vec{r}[x^\sigma(\vec{r})]^2 - \frac{\alpha|M|}{2}(K(0)-\bar{K})y_0 + B(x) -$$
$$\sum_\sigma \int_M d\vec{r}\, t^\sigma(\vec{r})x^\sigma(\vec{r}) - tx - r_0y_0 + r_1y_1 +$$
$$\frac{\alpha}{2\beta}\sum_{\mathbf{p}}\left(\ln[1-T_0(\mathbf{p})\beta(y_0-y_1)] - \frac{\beta y_1}{1-T_0(\mathbf{p})\beta(y_0-y_1)}\right) \tag{C.4}$$

where now $T_0(\mathbf{p})$ is the Fourier transform of the kernel $|M|\hat{K}$

$$T_0(p) = |M|\int_M d\vec{r}e^{-i\mathbf{p}\vec{r}}K(\vec{r}). \tag{C.5}$$

We now have

$$h = b(x) + \sum_\sigma \int_M d\vec{r}' x^\sigma(\vec{r}')\left[L(\vec{r}^\sigma - \vec{r}') - \bar{L}\right] - z(-2tr_1)^{1/2} \tag{C.6}$$
$$h_2 = -r_0 + r_1. \tag{C.7}$$

The $T = 0$ mean field equations are much like in (Treves, 1990) apart from the $x^\sigma(\vec{r})$ equation which reads:

$$x^\sigma(\vec{r}) = g'\left\langle\left\langle\left[L(\vec{r}^\sigma - \vec{r}') - \bar{L}\right]\int^+ Dz\left\{\int_M d\vec{r}'\left[L(\vec{r}^\sigma - \vec{r}') - \bar{L}\right]x^\sigma(\vec{r}') + b(x) - \theta - \rho z\right\}\right\rangle\right\rangle \tag{C.8}$$

where now the + sign on the integral means that the limits of integration over z are chosen such that

$$\int d\vec{r}'\left[L(\vec{r}^\sigma - \vec{r}') - \bar{L}\right]x^\sigma(\vec{r}') + b(x) - \theta > 0. \tag{C.9}$$

g' is a renormalized gain, which takes into account the effect of static noise, defined by:

$$(g')^{-1} = g^{-1} - \alpha\sum_p T_0(p)\frac{\bar{\Psi}}{1-T_0(p)\bar{\Psi}} \tag{C.10}$$

where $\bar{\Psi}$ is given by eq. C.18.

The noise variance ρ^2 is given by

$$\rho^2 = -2Tr_1 = \alpha \sum_{\mathbf{p}} \frac{[T_0(p)]^2 y_0}{\left[1 - T_0(p)\bar{\Psi}\right]^2} \tag{C.11}$$

where

$$y_0 = (g')^2 \left\langle \left\langle \int^+ Dz \left\{ \int_M d\vec{r}' \left[L(\vec{r}^{\,\sigma} - \vec{r}') - \bar{L} \right] x^\sigma(\vec{r}') + b(x) - \theta \right\}^2 \right\rangle \right\rangle \tag{C.12}$$

and

$$\bar{\Psi} = g' \left\langle \left\langle \int^+ Dz \right\rangle \right\rangle . \tag{C.13}$$

We now pass to the rescaled variables

$$v^\sigma(\vec{r}) = \frac{x^\sigma(\vec{r})}{\rho} \tag{C.14}$$

$$w = \frac{b(x)}{\rho} \tag{C.15}$$

obtaining

$$v^\sigma(\vec{r}) = g' \int_M d\vec{r}^{\,\sigma} \left[L(\vec{r}^{\,\sigma} - \vec{r}) - \bar{L} \right] \mathcal{N} \left(w + \int_M d\vec{r}' \left[L(\vec{r}^{\,\sigma} - \vec{r}') - \bar{L} \right] v^\sigma(\vec{r}') \right) \tag{C.16}$$

$$\frac{y_0}{\rho^2} = (g')^2 \int_M \frac{d\vec{r}^{\,\sigma}}{|M|} \mathcal{M} \left(w + \int_M d\vec{r}' \left[L(\vec{r}^{\,\sigma} - \vec{r}') - \bar{L} \right] v^\sigma(\vec{r}') \right) \tag{C.17}$$

$$\bar{\Psi} = \int_M \frac{d\vec{r}^{\,\sigma}}{|M|} \Phi \left(w + \int_M d\vec{r}' \left[L(\vec{r}^{\,\sigma} - \vec{r}') - \bar{L} \right] \right) v^\sigma(\vec{r}). \tag{C.18}$$

Appendix D

Generic kernel: storable information calculation

First, the information per synapse we get from a single observation of activity, with the animal in a certain position times the number of stored charts is

$$I_1 = \alpha \int \frac{d\vec{r}}{|M|} \left\{ \int_{-\infty}^{u(\vec{r})} \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \log \left(\frac{e^{-\frac{z^2}{2}}}{\int \frac{d\vec{r'}}{|M|} e^{-\frac{(z-u(\vec{r})+u(\vec{r'})^2)}{2}}} \right) + [1-\phi(u(\vec{r}))] \log \left(\frac{[1-\phi(u(\vec{r}))]}{\int \frac{d\vec{r'}}{M} \left[1-\phi(u(\vec{r'}))\right]} \right) \right\}. \quad \text{(D.1)}$$

Next, we wish to calculate the joint information from two measures of activity, from the same cells, from all charts, while the rat is in two different locations, at a distance ϵ. These two measures are correlated random variables: let

$$V_1 = [h_1 - \rho z_1]^+$$

be the activity of a cell measured while the rat is in position 1, and

$$V_2 = [h_2 - \rho z_2]^+$$

be the activity of the same cell while the rat is in position 2.

The two noise variables are distributed according to a joint bivariate gaussian distribution:

$$p(z_1, z_2) = \frac{1}{2\pi\sqrt{1-r_{12}^2}} \exp\left(-\frac{1}{2(1-r_{12}^2)} (z_1^2 + z_2^2 - 2r_{12} z_1 z_2) \right). \quad \text{(D.2)}$$

The correlation coefficient r_{12} is a function of the distance ϵ, implicitly defined through the equation

$$\rho^2 r_{12}(\epsilon) = \alpha |M| \langle\langle K^2 \rangle\rangle y_{12}(\epsilon) \tag{D.3}$$

where $y_{12}$is defined as

$$y_{12}(\epsilon) = \frac{1}{N} \sum_i \langle V_{i,1} V_{i,2} \rangle \tag{D.4}$$

and assuming periodic boundary conditions:

$$y_{12}(\epsilon) = \rho^2 g^2 \int \frac{dr}{|M|} \int^{++} Dz_{12} \times \tag{D.5}$$

$$\left(\int \frac{dr'}{|M|} K(r - r') v(r') + w - z_1 \right)$$

$$\left(\int \frac{dr''}{|M|} K(r - r'') v(r'' + \epsilon) + w - z_2 \right) \tag{D.6}$$

or,

$$y_{12}(\epsilon) = \rho^2 g^2 \int \frac{dr}{|M|} \int^{++} Dz_{12} u(r) u(r + \epsilon) \tag{D.7}$$

where $u(r)$ is defined by eq. B.4. The integration measure for the noise variable is defined as

$$\int^{++} Dz_{12} = \int_{u(r)-z_1>0, u(r+\epsilon)-z_2>0} dz_1 dz_2 p(z_1, z_2). \tag{D.8}$$

Inserting eq. D.7 in eq. D.3 we yield:

$$r_{12} = \alpha |M| \langle\langle K^2 \rangle\rangle g^2 \int dr \mathcal{Q}(u(r), u(r + \epsilon), r_{12}) \tag{D.9}$$

where

$$\mathcal{Q}(x, y, r_{12}) = \int^{++} Dz_{12} (x - z_1)(y - z_2).$$

Eq. D.9 can be solved numerically, an example is provided in fig. D.1, but a few features can be explored analytically, in the neighborhood of $\epsilon = 0$. $r_{12} = 1, \epsilon = 0$ is a solution, but now consider what happens when ϵ increases.

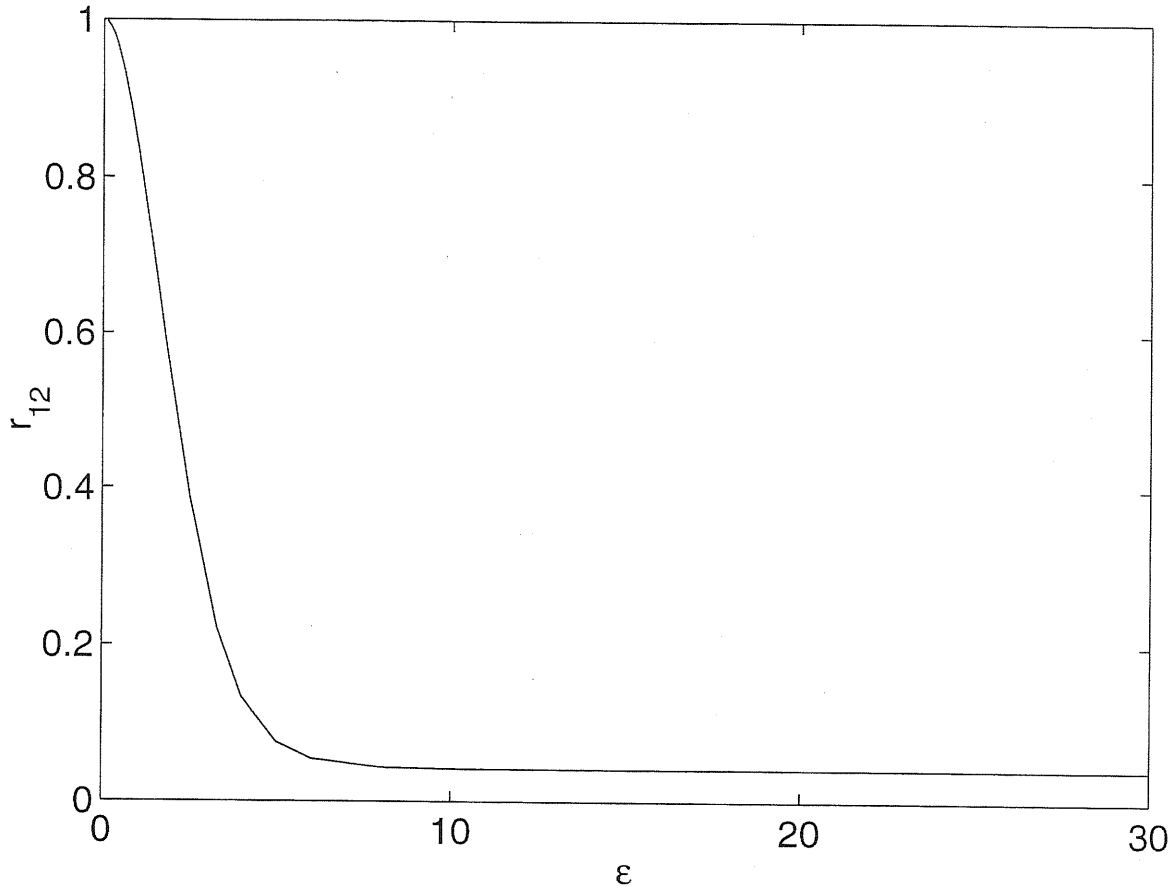

Figure D.1: The r_{12} function plotted as a function of the distance between the two p.f.c.s, ϵ in the $|M| = 30$ case.

The derivatives of

$$\mathcal{D}(r_{12}, \epsilon) = \alpha \langle \langle K^2 \rangle \rangle g^2 \int \frac{dr}{|M|} \mathcal{Q}(u(r), u(r+\epsilon), r_{12}) - r_{12} \tag{D.10}$$

with respect to ϵ and r_{12} must be taken into consideration. One has:

$$\frac{\partial}{\partial \epsilon} \mathcal{D}(r_{12} = 1, \epsilon = 0) = \alpha \langle \langle K^2 \rangle \rangle g^2 \int \frac{dr}{|M|} \times$$
$$\frac{\partial}{\partial y} \mathcal{Q}(x = u(r), y = u(r+\epsilon), 1) u'(r) = 0, \tag{D.11}$$

$$\frac{\partial^2}{\partial \epsilon^2}\mathcal{D}(r_{12}=1,\epsilon=0)<0 \tag{D.12}$$

and

$$\begin{aligned}&\frac{\partial}{\partial r_{12}}\mathcal{D}(r_12\to 1,\epsilon=0)=\\&\alpha\langle\langle K^2\rangle\rangle g^2\int\frac{dr}{|M|}\Phi(u(r))-1=\\&\int\frac{dr}{|M|}\Phi(u(r))\left(\int\frac{dr}{|M|}\mathcal{M}(u(r))\right)^{-1}-1\end{aligned} \tag{D.13}$$

From eq. D.12 it turns out that when the derivative in eq. D.13 is greater than zero, the solution $r_{12}=1$ disappears as one moves from $\epsilon=1$, but another solution is still present so that

$$\lim_{\epsilon\to 0^+} r_{12}(\epsilon)<1. \tag{D.14}$$

Note that the condition

$$\frac{\partial}{\partial r_{12}}\mathcal{D}(r_{12}\to 1,\epsilon=0)>0$$

is equivalent to

$$\Gamma=g\alpha|M|\langle\langle K^2\rangle\rangle\int_M dr\,u(r)v(r)<0, \tag{D.15}$$

and the quantity Γ enters in the stability analysis considerations we sketched in sec. 3.3.3, at least for the 1-D case. Solutions with $\Gamma>0$ are stable against inhibition orthogonal fluctuations, so that it is likely that the possible pathology implied by eq. D.14 reflects an instability of the solution. We have always found numerically that for solution corresponding to the maximal storage capacity and information, $\Gamma>0$.

Once we know the joint probability distribution for z_1 and z_2, we can calculate the information we can extract about the p.f.c. of a cell from two measurements of activity, while the rat is standing in two positions at a distance ϵ, from all charts.

$$
\begin{aligned}
I_2(\epsilon) = \quad & \alpha \int_M \frac{dr}{|M|} \Bigg\{ \int^{++} Dz_{12} \Big[-\frac{1}{2(1-r_{12}^2)} (z_1^2 + z_2^2 - 2r_{12} z_1 z_2) - \\
& \log\Big(\int_M \frac{dr'}{|M|} \exp\Big(-\frac{1}{2(1-r_{12}^2)} [(u(r') - u(r) + z_1)^2 + (u(r'+\epsilon) - u(r+\epsilon) + z_2)^2 \\
& -2r_{12}(u(r') - u(r) + z_1)(u(r'+\epsilon) - u(r+\epsilon) + z_2)] \Big) \Big) \Big] \\
& +2 \int^{+-} Dz_{12} \Big[\log\Big(\int_{z_2' > u(r+\epsilon)} \frac{dz_2'}{2\pi} \exp\Big(-\frac{1}{2(1-r_{12}^2)} (z_1^2 + z_2'^2 - 2r_{12} z_1 z_2') \Big) \Big) \\
& - \log\Big(\int_M \frac{dr'}{|M|} \int_{z_2' > u(r'+\epsilon)} dz_2' \exp\Big(-\frac{1}{2(1-r_{12}^2)} [(u(r') - u(r) + z_1)^2 + (u(r'+\epsilon) - u(r + \\
& -2r_{12}(u(r') - u(r) + z_1)(u(r'+\epsilon) - u(r+\epsilon) + z_2')] \Big) \Big) \Big] \\
& + \int^{--} Dz_{12} \Big[\log\Big(\int_{z_1' > u(r), z_2' > u(r+\epsilon)} \frac{dz_1' dz_2'}{2\pi\sqrt{1-r_{12}^2}} \exp\Big(-\frac{1}{2(1-r_{12}^2)} (z_1'^2 + z_2'^2 - 2r_{12} z_1' z_2') \Big) \\
& - \log\Big(\int_M \frac{dr'}{|M|} \int_{z_1' > u(r'), z_2' > u(r'+\epsilon)} \frac{dz_1' dz_2'}{2\pi\sqrt{1-r_{12}^2}} \exp\Big(-\frac{1}{2(1-r_{12}^2)} \\
& [(u(r') - u(r) + z_1')^2 + (u(r'+\epsilon) - u(r+\epsilon) + z_2')^2 \\
& -2r_{12}(u(r') - u(r) + z_1')(u(r'+\epsilon) - u(r+\epsilon) + z_2')] \Big) \Big) \Big] \Bigg\}.
\end{aligned}
$$

The minus signs $(-)$ beside the integration signs mean that respectively the first, or the second condition determining the integration intervals in eq. D.8 are reversed. The first term in the sum accounts for the contribution coming from measurement in which both activity values are positive. The second term is the contribution from measurements in which one value is zero and the other is positive. The third term comes from measurements in which both values are zero.

For $\epsilon = 0$, $I_2 = I_1$, since the two measures are identical.

For large ϵ one has $I_2 \sim 2I_1$, because the noise decorrelates and because in general the two measures will give non-zero results in distinct regions of the environment. The behavior of I_2 as a function of ϵ is exemplified in figure D.2. We define as "information correlation length" the value l_I of ϵ for which

$$I_2 - I_1 = f I_1, \tag{D.17}$$

where f is a fixed fraction, say 0.95. Note that this quantity We may say that measurements of activity with the rat in two positions at a distance l_I give independent information.

This allows us to define as the *stored information* I_s the quantity

$$I_s = I_1 \frac{|M|}{l_I^d}, \tag{D.18}$$

that is, sampling the activity of a cell $\frac{|M|}{l_I^d}$ times, with the animal spanning a lattice with size l_I, we may effectively add up the information amounts we get from each single sample, as if they were independent.

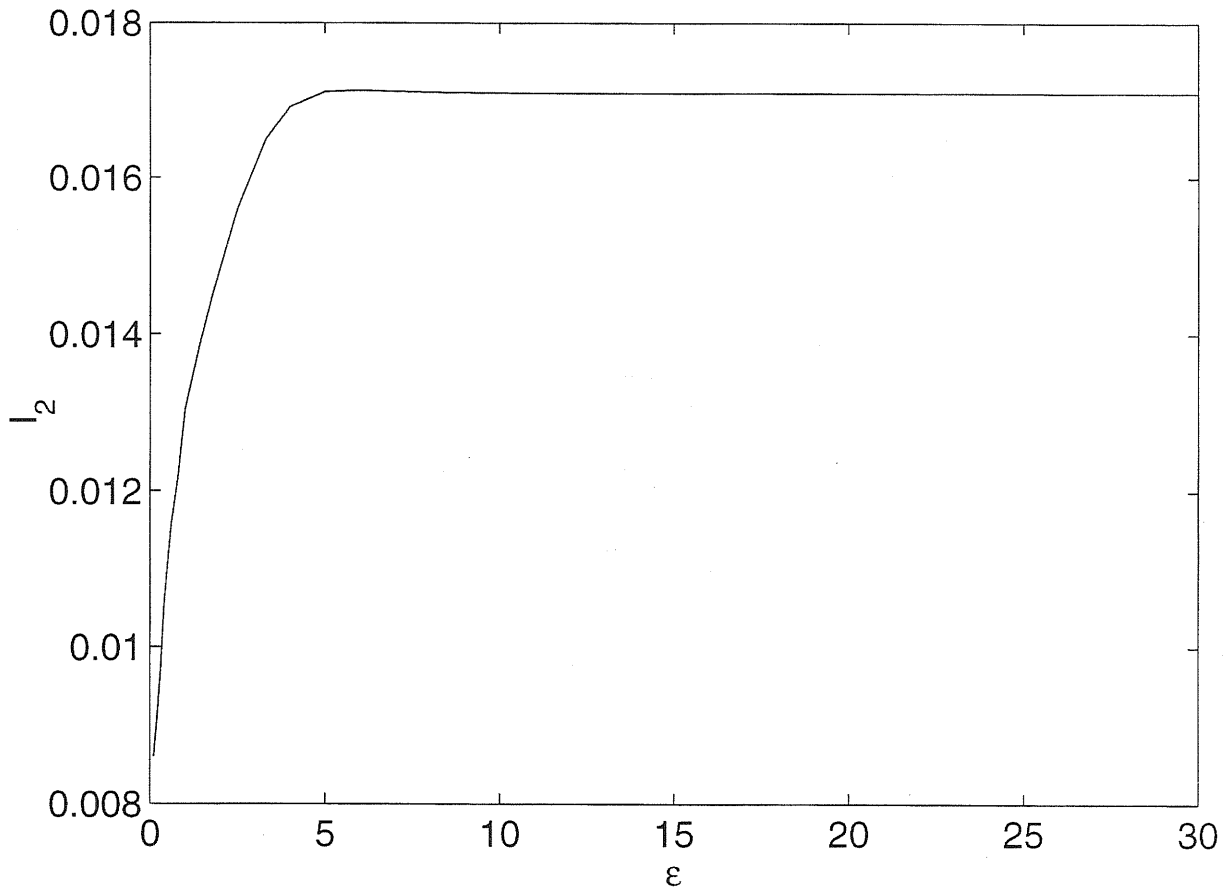

Figure D.2: The I_2 function plotted as a function of the distance between the two p.f.c.s, ϵ in the $|M| = 30$ case. Note that l_I, with $f = 0.95$ (see eq. D.17) would be approximately 3.5. This is seen not to change much when $|M|$ is varying (not shown).

References

Abbott, L. (1991). Realistic synaptic input for model neural networks. *Network*, *2*, 245–258.

Abbott, L., & van Vreeswijk, C. (1994). When inhibition not excitation synchronizes neural firing. *J. Comput. Neurosci.*, *1*, 313–321.

Abeles, M. (1991). *Corticonics - neural circuits of the cerebral cortex.* Cambridge University Press.

Amaral, D., Ishizuka, N., & Claiborne, B. (1990). Neurons, numbers and the hippocampal network. *Progress in brain research*, *83*, 1–11.

Amaral, D., & Johnston, D. (1998). Hippocampus. In G. Shepherd (Ed.), *The synaptic organization of the brain* (2 ed., pp. 417–458). New York: Oxford University Press.

Amaral, D., & Witter, M. (1995). Hippocampal formation. In G. Paxinos (Ed.), *The rat nervous system* (pp. 443–493). San Diego: Academic Press.

Amit, D. (1989). *Modeling brain function.* New York: Cambridge University Press.

Amit, D. (1995). The hebbian paradigm reintegrated: local reverberation as internal representation. *Behav. Brain Sci.*, *18*, 617–657.

Amit, D., & Brunel, N. (1996). Model of a global spontaneous activity and local structured activity during delay periods in cerebral cortex. *Cerebral Cortex*, *7*, 237–252.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Ann. Phys. (N.Y.)*, *173*, 30–67.

Andersen, P., Eccles, J., & Böyning, Y. (1964). Location of post synaptic inhibitory synapses on hippocampal pyramids. *J. Neurophysiol*, *27*, 592–607.

Barnes, C. (1995). Involvment of ltp in memory: Are we "searching under the street light? *Neuron*, *15*, 751–754.

Barnes, C., Suster, M., Shen, J., & McNaughton, B. (1997). Multistability of cognitive maps in the hippocampus of old rats. *Nature*, *388*(6639), 272–5.

Battaglia, F., & Treves, A. (1998a). Attractor neural networks storing multiple space representation: a model for hippocampal place fields. *Phys. Rev. E.* (in press)

Battaglia, F., & Treves, A. (1998b). Stable and rapid recurrent processing in realistic auto-associative memories. *Neural Computation*, *10*, 431–450.

Ben-Yishai, R., Bar-Or, R., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.*, *92*, 3844–3848.

Blair, H., & Sharp, P. (1995). Anticipatory head direction signals in anterior thalamus: Evidence for a thalamocortical circuit that integrates angular head motion to compute head direction. *J. Neurosci.*, *15*, 6260–6270.

Bliss, T., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol. Lond.*, *232*, 331–356.

Blum, K., & Abbott, L. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Computation*, *8*(1), 85–93.

Bostock, E., Muller, R., & Ranck, J. (1991). Experience dependent modifications of hippocampal place cells firing. *Hippocampus*, *1*, 193–206.

Braitenberg, V., & Schüz, A. (1991). *Anatomy of the cortex - statisics and geometry.* Berlin: Springer-Verlag.

Buhmann, J., Divko, R., & Schulten, K. (1989). Associative memory with high information content. *Phys. Rev. A*, *39*, 2689–2692.

Cohen, N., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Colquhoun, D., Jonas, P., & Sakmann, B. (1992). Action of brief pulses of glutamate on ampa/kainate receptors in patches from different neurones of rat hippocampal slices. *J. Physiol.*, *458*, 261–287.

Connors, B., Malenka, R., & Silva, L. (1988). Two inhibitory post-synaptic potentials, and gaba_A and gaba_B receptor-mediated responses in neocortex of rat and cat. *J. Physiol*, *406*, 443–468.

Deppisch, J., Bauer, H., & Schillen, T. (1993). Alternating oscillators and stochastic states in a network of spiking neurons. *Network*, *4*, 243–257.

Derrida, B., Gardner, E., & Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *Europhysics Letters*, *4*(2), 167–173.

Douglas, R., & Martin, K. (1991a). A functional microcircuit for cat visual cortex. *J. Physiol (London)*, *440*, 735–769.

Douglas, R., & Martin, K. (1991b). A functional microcircuit for the cat visual cortex. *J. Physiol. (Lond.)*, *440*, 735–769.

Douglas, R., & Martin, K. (1998). Neocortex. In G. Shepherd (Ed.), *The synaptic organization of the brain* (Fourth ed., pp. 459–509). New York: Oxford University Press.

Eccles, J. (1957). *The physiology of nerve cells.* Baltimore: John Hopkins U.P.

Eccles, J. (1964). *The physiology of synapses.* New York: Academic Press.

Eichenbaum, H., Kuperstein, M., Fagan, A., & Nagode, J. (1987). Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task. *J. Neurosci.*, *7*(3), 716–732.

Gaffan, D. (1994). Scene-specific memory for objects: a model of episodic memory impairment in monkeys with fornix transection. *J. Cogn. Neurosci.*, *6*, 305–320.

Georgopoulos, A., Kettner, R., & Schwartz, A. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. ii. coding of the direction of movement by a neuronal population. *Journal of Neuroscience*, *8*(8), 2928–37.

Golomb, D., Rubin, N., & Sompolinsky, H. (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A.*, *41*, 1843–1854.

Goodridge, J., & Taube, J. (1994). The effects of lesions of the postsubiculum on head direction cell firing in the anterior thalamic nuclei. *Soc. Neurosci. Abs.*, *20*, 805.

Gothard, K., Skaggs, W., & McNaughton, B. (1996). Dynamics of mismatch correction in the hippocampal ensemble code for space: interaction between path integration and environmental cues. *Journal of Neuroscience*, *16*(24), 8027–40.

Griniasty, M., Tsodyks, M., & Amit, D. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors in neural networks. *Neural Computation, 5*, 1–17.

Hablitz, J., & Thalmann, R. (1987). Conductance changes underlying a late synaptic hyper-polarization in hippocampal CA3 neurons. *J. Neurophysiol, 58*, 160–179.

Hansel, D., Mato, G., & Meunier, C. (1995). Synchrony in excitatory neural networks. *Neur. Comp., 7*, 307–337.

Hasselmo, M., & Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modeling and brain slice physiology. *J. Neurosci, 14*, 3898–3914.

Hebb, D. (1949). *The organization of behaviour.* New York: Wiley.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory fo neural computation.* Reading, MA: Addison-Wesley.

Hestrin, S., Nicoll, R., Perkel, D., & Sah, P. (1990). Analysis of excitatory synaptic action in pyramidal cells using whole-cells recording from rat hippocampal slices. *J. Physiol., 422*, 203–225.

Hill, A. (1978). First occurrence of hippocampal spatial firing in a new environment. *Exp. Neurol., 74*, 204–217.

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

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture of the cat's visual cortex. *J. Physiol. (Lond.), 160*, 106–154.

Knierim, J., Kudrimoti, H., & McNaughton, B. (1995). Place cells, head direction cells, and the learning of landmark stability. *J. Neurosci., 15*, 1648–1659.

Kojima, S., & Goldman-Rakic, P. (1984). Functional analysis of spatially discriminative neurons in prefrontal cortex of rhesus monkey. *Brain Research, 291*, 229–240.

Lapique, L. (1907). Recherches qualitatives sur l'excitation électrique de nerfs traités comme une polarisation. *J. Physiol. Pathol. Gen., 9*, 620–635.

Leonhard, C., Stackman, R., & Taube, J. (1996). Head direction cells recorded from the lateral mammillary nuclei in rats. *Soc. Neurosci. Abs., 22*, 1873.

Lorente de Nó, R. (1933a). Studies on the structure of the cerebral cortex. *J. Psychol. Neurol.*, *45*, 381–438.

Lorente de Nó, R. (1933b). Studies on the structure of the cerebral cortex. II. continuation of the study of the ammonic system. *J. Psychol. Neurol.*, *46*, 113–177.

MacGregor, R. (1987). *Neural and brain modeling.* Academic Press.

Markus, E., Barnes, C., McNaughton, B., Gladden, V., & Skaggs, W. (1994). Spatial information content and reliability of hippocampal ca1 neurons: effects of visual input. *Hippocampus*, *4*(4), 410–21.

Markus, E., Qin, Y., Leonard, B., Skaggs, W., McNaughton, B., & Barnes, C. (1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience*, *15*(11), 7079–94.

Marr, D. (1970). A theory for cerebral neocortex. *Proc. Roy. Soc. Lond. B*, *176*, 161–234.

Marr, D. (1971). Simple memory: a theory for the archicortex. *Phil. Trans. Roy. Soc. Lond. B*, *262*, 24–81.

Mason, A., & Larkman, A. (1990). Correlations between morphology and electrophysiology of pyramidal neurones in slices of rat visual cortex: II. electrophysiology. *J. Neurosci.*, *10*, 1415–1428.

McBain, C., & Dingledine, R. (1992). Dual-component miniature excitatory synaptic currents in rat hippocampal CA3 pyramidal neurons. *J. Neurophysiol.*, *68*, 16–27.

Mezard, M., Parisi, G., & Virasoro, M. (1988). *Spin glass theory and beyond.* Singapore: World Scientific.

Milner, D., & Goodale, M. (1995). *The visual brain in action.* Oxford: Oxford University Press.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–820.

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

Morris, R. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*, 239–260.

Morris, R., Anderson, E., Lynch, G., & Baudry, M. (1986). Selective impairment of learning and blockade of long-term potentiation by an n-methyl-d-aspartate receptor antagonist, ap5. *Nature, 319*, 774–776.

Morris, R., Garrud, P., Rawlins, J., & O'Keefe. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature, 297*, 681–683.

Muller, R., Ranck, J., & Taube, J. (1996). Head direction cells: properties and functional significance. *Current Opinion in Neurobiology, 6*(2), 196–206.

Muller, R., & Stead, M. (1996). Hippocampal place cells connected by hebbian synapses can solve spatial problems. *Hippocampus, 6*, 709–719.

O'Keefe, J., & Dostrovski, J. (1971). The hippocampus as a spatial map. preliminary evidence form unit activity in the freely moving rat. *Brain Res., 34*, 171–175.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network, 7*, 87–107.

Parisi, G. (1986). A memory which forgets. *J. Phys. A, 19*, L617.

Parkinson, J., Murray, E., & Mishkin, M. (1988). A selective mnemonic role for the hippocampus in monkeys: memory for the location of objects. *J. Neurosci., 8*, 4059–4067.

Phillips, W., & Singer, W. (1997). In search of common foundations for cortical computation. *Behavioral and brain sciences, 20*, 657–722.

Quirk, G., Muller, R., & Kubie, J. (1990). The firing of hippocampal place cells in the dark depends on rat's recent experience. *J. Neurosci., 10*, 2006–2017.

Ramón y Cajal, S. (1911). *Histologie du systeme nerveux de l'homme et des verterbres.* Paris: Maloine.

Ranck, J., Jr. (1984). Head-direction cells in the deep cell layers of the dorsal presubiculum. *Soc. Neurosci. Abs, 10*, 599.

Redish, A., & Touretzky, D. (1996). Modeling interactions of the rat's place and head direction systems. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 61–71). MIT press.

Redish, A., & Touretzky, D. (1997). Cognitive maps beyond the hippocampus. *Hippocampus*, *7*, 15–35.

Rolls, E. (1996a). The representation of space in primate hippocampus, and episodic memory. In T. Ono, B. McNaughton, S. Molotchnikoff, E. Rolls, & H. Nishijo (Eds.), *Perception, memory and emotion: frontier in neuroscience* (pp. 375–400). Amsterdam: Elsevier.

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

Rolls, E., Treves, A., & Tovee, M. (1997). The representational capacity if the distributed encoding of information provided by population of neurons in the primate temporal visual cortex. *Exp. Brain. Res.*, *114*, 149–162.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, *354*, 108–109.

Samsonovich, A., & McNaughton, B. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*(15), 5900–20.

Scoville, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry*, *20*, 11–21.

Shadlen, M., & Newsome, W. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, *4*, 569–579.

Sharp, P. (1991). Computer simulations of hippocampal place cells. *Psychobiology*, *19*(2), 103–115.

Sharp, P. (1997). Subicular cells generate similar spatial firing patterns in two geometrically and visually distinctive environments: comparison with hippocampal place cells. *J. Neurosci.*, *85*, 71–92.

Sharp, P., & Green, C. (1994). Spatial correlates of firing patterns of single cells in the subiculum of freely moving rat. *J. Neurosci.*, *14*, 2339–2356.

Shatz, C. (1996). Emergence of order in visual system development. *Proc. Natl. Acad. Sci. USA*, *93*, 602–608.

Simmen, M., Treves, A., & Rolls, E. (1996). On the dynamics of a network of spiking neurons. In F. Eekman & B. J.M. (Eds.), *Computations and neuronal systems: Proceedings of CNS95*. Boston: Kluwer Academic Publishers.

Skaggs, W., Knierim, J., Kudrimoti, H., & McNaughton, B. (1995). A model of the neural basis of the rat's sense of direction. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems 7* (pp. 173–180). Cambridge, MA: MIT press.

Smith, M., & Milner, B. (1981). The role of the right hippocampus in the recall of spatial location. *Neuropsychologia, 19*, 781–793.

Sompolinsky, H., & Shapley, R. (1997). New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology, 7*, 514–522.

Sompolinsky, H., Tishby, N., & Seung, H. (1990). Learning from examples in large neural networks. *Phys. Rev. Lett., 65*, 1683–1686.

Stackman, R., & Taube, J. (1997). Firing properties of head direction cells in the rat anterior thalamic nucleus: dependence on vestibular input. *J. Neurosci., 17*, 4349–4358.

Suzuki, W., & Amaral, D. (1994). Topographic organization of the reciprocal connections between the monkey entorhinal cortex and the perirhinal and parahippocampal cortices. *J. Neurosci., 14*, 1856–1877.

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology, 2*, 502–505.

Taube, J. (1995). Pleace cells recorded in the parasubiculum of the freely moving rat. *Hippocampus, 5*, 569–583.

Taube, J., Goodridge, J., Golob, E., Dudchenko, P., & Stackman, R. (1996). Processing the head direction signal: A review and commentary. *Brain Res. Bull., 40*(5/6), 477–486.

Taube, J., & Muller, R. (1995). Head direction cell activity in the anterior thalamic nuclei, but not the postsubiculum, predicts the animal's future directional heading. *Soc. Neurosci. Abs., 21*, 946.

Taube, J., Muller, R., & Ranck, J., Jr. (1990). Head-direction cells recorded from the post-subiculum in freely moving rats. i. description and quantitative analysis. *J. Neurosci., 10*, 420–435.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Thorpe, S., & Imbert, M. (1989). In Pfeifer, Schreter, & Fogelman-Soulié (Eds.), *Connectionism in perspective.*

Touretzky, D., & Redish, A. (1996). Theory of rodent navigation based on interacting representations of space. *Hippocampus*, *6*(3), 247–70.

Treves, A. (1990). Graded-response neurons and information encodings in auto-associative memories. *Physical Review A*, *42*, 2418–2430.

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

Treves, A. (1991b). Dilution and sparse coding in threshold-linear nets. *Journal of Physics A: Math. Gen.*, *24*, 327–335.

Treves, A. (1993). Mean-field analysis of neuronal spike dynamics. *Network*, *4*, 259–284.

Treves, A. (1995). Quantitative estimate of the information relayed by the Schaffer collaterals. *J. Comput. Neurosci.*, *2*, 259–272.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neur. Comp.*, *7*, 399–407.

Treves, A., & Rolls, E. (1991). What determines the capacity of auto-associative memories in the brain? *Network*, *2*, 371–97.

Treves, A., & Rolls, E. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, *2*, 189–200.

Treves, A., & Rolls, E. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–391.

Treves, A., Rolls, E., & Tovee, M. (1996). In V. Torre & F. Conti (Eds.), *Neurobiology: Proceedings of the international school of biophysics, XXIII course, may 1995* (pp. 371–382). New York: NATO Advanced Study Institute, Plenum.

Treves, A., Skaggs, W., & Barnes. (1996). How much of the hippocampus can be explained by functional constraints? *Hippocampus*, *6*, 666–674.

Tsodyks, M., & Feigelman, M. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett*, *46*, 101.

Tsodyks, M., Mitcov, I., & Sompolinsky, H. (1993). Patterns of synchrony in inhomogeneous networks of oscillators with pulse interactions. *Phys. Rev. Lett.*, *71*, 1280–1283.

Tsodyks, M., & Sejnowski, T. (1994). Associative memory and hippocampal place cells. *Proceedings of the third workshop on Neural Networks: from Biology to High Energy Physics - International Journal of Neural Systems, 6 Supp.*, 81–86.

Tsodyks, M., & Sejnowski, T. (1995). Rapid state switching in balanced cortical network models. *Network*, *6*, 111–124.

Wang, X. (1998). Calcium coding and adaptive temporal computation in cortical pyramidal neurons. *J. Neurophysiol.*, *79*, 1549–1566.

Willshaw, D., & Buneman, O. (1969). Non-holographic associative memory. *Nature*, *222*, 960–962.

Wilson, H., & Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, *12*, 1–20.

Wilson, M., & McNaughton, B. (1993). , erratum appears in science 1994 apr 1;264(5155):16]. *Science*, *261*(5124), 1055–8.

Wilson, M., & McNaughton, B. (1994). Dynamics of the hippocampal ensemble code for space. *Science*, *265*(5172), 676–9.

Wong, R. and. (1987). Inhibitory control of local excitatory circuits in the guinea-pig hippocampus. *J. Physiol.*, *338*, 611–629.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head direction cell ensemble: A theory. *J. Neurosci.*, *16*, 2112–2126.

Zipser, D. (1986). Biologically plausible models of place recognition and goal location. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 432–470). MIT press.

Zola-Morgan, S., Squire, L., & Amaral, D. (1986). Human amnesia and the medial temporal lesion: enduring memory impairment following a bilateral lesion limited to field ca1 of the hippocampus. *J. Neurosci*, *6*, 2950–2967.

Zola-Morgan, S., Squire, L., Amaral, D., & Suzuki, W. (1989). Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *J. Neurosci.*, *9*, 4355–4370.

Zola-Morgan, S., Squire, L., & Ramus, S. (1994). Severity of memory impairment in monkeys as a function of locus and extent of damage within the medial temporal lobe memory system. *Hippocampus*, *4*, 483–95.