Scuola Internazionale Superiore di Studi Avanzati - Trieste

Cognitive Neuroscience

Latching dynamics in Potts neural networks

Candidate: **Chol Jun Kang**

Supervisor: **Prof. Alessandro Treves**

Thesis submitted for the degree of Doctor Philosophiae

December 12, 2017

SISSA - Via Bonomea 265 - 34136 TRIESTE - ITALY

Scuola Internazionale Superiore di Studi Avanzati

Latching dynamics in Potts neural networks

Thesis submitted for the degree of Doctor Philosophiae
Academic Year 2017/2018

Candidate: **Chol Jun Kang**
Supervisor: **Prof. Alessandro Treves**

External examiner: **Prof. Oren Shriki**
External examiner: **Dr. Eleonora Russo**

SISSA - Via Bonomea 265, 34136 Trieste - Italy

December 12, 2017

Contents

List of Figures

Introduction

One purpose of Computational Neuroscience is to try to understand by using models how at least some parts in the brain work or how cognitive phenomena occur and are organized in terms of neuronal activity. The Hopfield model of a neural network, rooted in Statistical Physics, put forward by J. Hopfield in the 1980s, was one of the first attempts to explain how associative memory could work. It was successful in guiding experiments, e.g., in the hippocampus and primate inferotemporal cortex. However, some higher level cognitive functions that the brain accomplishes require, to be approached quantitaively, by more advanced models beyond simple cued retrieval. Thought processes, the faculty of language (in the narrow sense, as defined by [7]), confabulation, producing poetry, drawing, arithmetics, complex navigation and many types of creative activity could be possible examples. It is thought essential that recursion – *an ability to generate sequences of arbitrary length from a finite set of elements* – stands at the core of all these examples. In the early 2000s, Treves proposed to model global cortical dynamics with a Potts network [8], as will be explained more fully in the coming chapters. The model tries to capture recursive dynamics by first implementing Braitenberg's idea of the modularity of the cortex [2, 9, 10] into a Potts spin associative network model, earlier studied by Kanter [11] and by Bollé [12–15] in a statistical physics context; the proposal followed a number of studies of multi-modular cortical networks, with O'Kane and Fulvi Mari, in the 1990s, that had reached a kind of mathematical dead end [16–19]. The storage capacity and optimal Hebbian learning rule in the static Potts network (the version without adaptation) were found by Kropff and Treves in 2005 [20]. In the presence of time dependent thresholds, latching dynamics as a model of infinite recursion has been studied by Russo et al in 2008 [21]; latching can be defined as the successive retrieval of a sequence of stored memory patterns, which can be sustained for a certain length of time, or in some conditions indefinitely. Namboodiri and Pirmoradian also devoted their efforts to this line of work [21, 22].

This thesis consists of five chapters and is organized as follows.

In chapter 1, the concepts of neuron and anatomical structure of the cortex, focusing on the role of pyramidal cells, are presented, to motivate the Potts network as a realization of Braitenberg's perspective. The Hopfield model is also briefly introduced as warming up for the Potts network, and the necessity of sequential retrieval (latching) is discussed. In chapter 2, the Potts network as a mathematical model is defined, with the equations of motion that govern latching dynamics. This part corresponds to the modelling sections in [23]. The main body of chapter 3 is from [24]. In this chapter, the approximate equivalence between multi-modular and static Potts networks is discussed, where the specific multi-modular network can be

referred to as the one studied by O'Kane and Treves in the early 1990s [16,17]. The significance of the self-feedback term, called w-term, which was initially introduced by Russo and Treves in 2012 [5], becomes evident in the reduction from the modular level. The storage capacity is estimated by replica analysis in the presence of such w-term, which was missing in previous studies. The w-term on the one hand reinforces the strength of the global attractors in latching, but on the other hand it may end up ruining memory capacity. In chapter 4, which is essentially [23], latching dynamics is analysed in the slowly and fast adapting regime in terms of latching length, quality of retrieval, crossover in overlap and correlations of patterns.
The work presented in chapter 3 and 4 were done in collaboration with Vezha Boboeva and Michelangelo Naim. In chapter 5, *hetero*-associative learning is additonally introduced with a modified Hebbian rule for instructed sequential recalling, and its effects on latching and the correlation between successive patterns are discussed. This part comes from the work reported in [25]. Finally, in the concluding chapter we comment on the previous chapters and on future studies.

List of publications

1. C. J. Kang, M. Naim, V. Boboeva, A. Treves; "*Life on the edge: latching dynamics in a Potts neural network*", Entropy, 19, 468 (2017).

2. C. J. Kang, A. Treves; "*Instructed latching dynamics via hetero hebbian type learning rule in Potts model network*", in preparation

3. M. Naim, V. Boboeva, C. J. Kang, A. Treves; "*Reducing a cortical network to a Potts model yields storage capacity estimates*", submitted in Journal of Statistical Mechanics: *theory and experiment*; https://arxiv.org/pdf/1710.04897.pdf

Chapter 1

Background

1.1 Neurons and cortical anatomy

Nervous systems consist of neurons connected by synapses. To give approximate orders of magnitude, it is estimated that there are 10^{11} neurons and each can have about 10^5 synaptic connections with others in the human brain. These scales are already astronomical numbers, that can only be analysed by statistical tools. Pre and post synaptic neurons are linked by synapses, where neurotransmitters triggered by action potentials coming along the axon of the pre synaptic neuron are released, carrying information in chemical form to arrive at the dendrites or cell body of the post synaptic neuron (Fig.1.1).

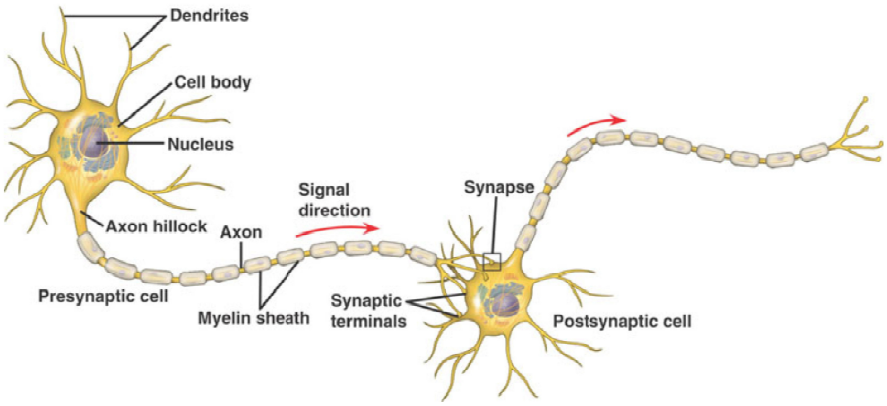


Figure 1.1: Synaptic connection of two neurons, reprinted from [1].

Among the best characterized neurons are *pyramidal* cells. They are present in the cerebral cortex, where they comprise about 80% of the neurons, in the hippocampus, amygdala and other areas of the brain. Although many types of neurons play important roles in functioning of the nervous system, our main topic throughout the thesis will be the pyramidal neurons of the cortex. Fig.1.2 shows examples of pyramidal neurons in temporal cortex (upper left), pyriform cortex (upper right), insula (lower left) and visual cortex (lower right), in the human brain.
In particular, frontal cortex, which will be the core region modelled by our Potts neural network (as will be explained later) consists of pyramidal neurons with a clear layered structure (Fig.1.3).

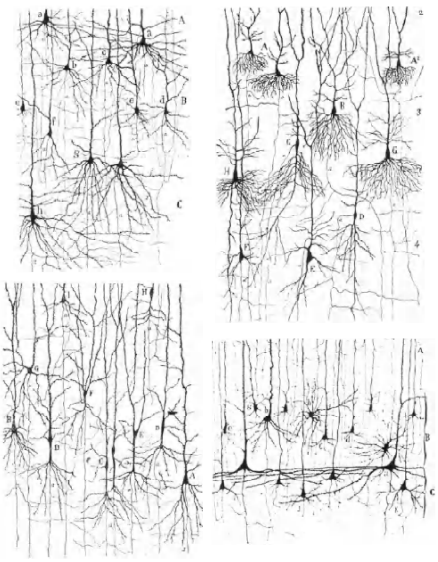


Figure 1.2: Pyramidal neurons in the human brain: temporal cortex (upper left), pyriform cortex (upper right), insula (lower left), visual cortex (lower right). The figure is taken from [2] and it was originally produced by Cajal (1911).

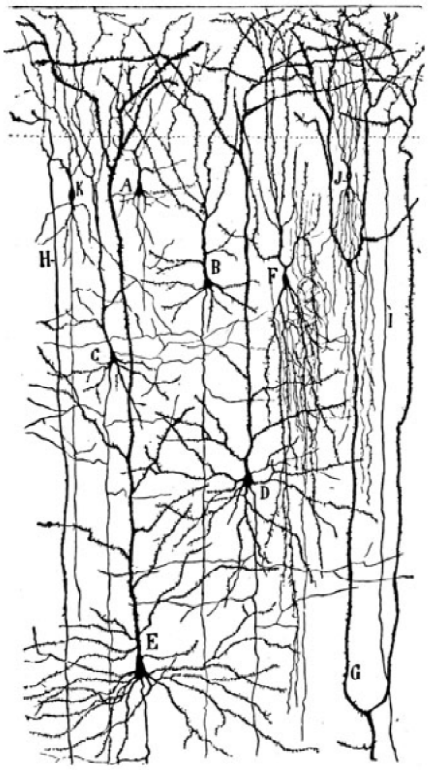


Figure 1.3: Scheme of frontal cortex in the human brain, from [3].

1.2 Cortex: Braitenberg's view

We summarize the view of Braitenberg on the cortex based on [2,26]. Pyramidal neurons in the cortex, and in particular in the frontal cortex, are connected with each other to form complicated networks. Their basal dendrites collect local collaterals, which do not usually exit the grey matter, coming from neighbouring cells, to form a local network (called the B-system in Braitenberg's terminology) that is specific to the local area. Some axons leave the grey matter and spread towards other cortical areas through the white matter. In the first or upper layer in the grey matter, the apical dendrites of pyramidal cells collect mainly these long-range cortico-cortical connections and form another network, called the A-system. There is no clearcut distinction, however, between the A and B-systems in their apical or basal position on the dendrites or the receiving cells; while there is in terms of local axons versus axons that travel through the white matter. They comprise, therefore, the global and local connections among pyramidal cells in the cortex, and it is assumed that the range spanned by the local network of one pyramidal cell is approximately of the same order of magnitude as the dimensions of its dendritic tree, up to a millimeter. The essential difference between local (B) and global (A) networks is that the B system can be considered to be *metric*, meaning that the probability of a local connection between two neurons strongly depends on their distance, while for the A system it is *ametric*, in that their distance is not directly relevant.

In this thesis we shall use multi-modular networks to model the A and B systems. The B system is modeled by the connections within a module and the A system consists of many modules that form a global network. It should be noted that the assumption of modules, with well-defined boundaries between them, is not realistic, and it is an additional ingredient, not necessarily implied by the distinction between the A and B systems *per se* As already mentioned in the introduction, considerable efforts to study extensive autoassociative networks have been carried out using multi modular models, by O'Kane, Fulvi Mari and Treves [16–19]. Later, Treves came up with an advanced model – the Potts model, which essentially treats the modules as multi flavour Potts variables. The whole network is connected among Potts spins by long-range synaptic connections.
Details will be defined and discussed in other chapters of the thesis.

1.3 The Hopfield neural network

Computational Neuroscience focuses on explaining at least in some aspects how the brain or its specific parts could work, by using simplified models. Since there are an enormous number of neurons in the brain, we could expect the outcome of collective behaviors of the neurons to be relevant for behaviour. Of course, it does not mean that all brain activities are the results of the coherence of a massive number of neurons. At least, however, if we put our interest on computations such as associative memory in cortex or hippocampus, we can say that in principle it is not possible to describe high level cognitive functions or brain activity exhibited as collective neuron behavior, even though we understand the information carried by

the activity of individual neurons.
This is not limited to the study of neuroscience, but it is rather accepted as a general principle when we deal with macroscopic properties of interacting elements, in many fields of science.

In fact, it has already been recognized and highlighted by P. W. Anderson in 1972 in his famous article "More is different" [27]. There are many examples. Suppose that we know the number of water molecules in a bucket. We know not only the initial conditions, coordinates and momenta of individual molecules, but also that they follow Newton's second law of motion. It is true that we can trace all the molecules at any time by integrating equations of motion. However, we can never explain the phase transition – gas to liquid or liquid to ice or vice versa with our information on molecules. It is because there are different fundamental laws at different scales of particles and quantitative change brings about qualitative differences. Sound waves in the crystals, ferromagnetism in magnets, superconductivity in Hg, and superfluidity in He^4 are the typical favourite examples by physicists. It should be emphasised that emergent behavior in many body systems has nothing to do with the microscopic details of the states of the elements.

Coming back to the brain, we are now convinced of the necessity of adopting models, methods and ideas developed in statistical physics for tackling theoretical problems in computational neuroscience. In our hands we have neurons, our microscopic constituents, connected by synapses.
In this regard, J. J. Hopfield was the first to introduce the statistical method in studying a large neural network quantitatively [28]. The classical Ising spin model in statistical physics is usually used as an metaphor. In the "Hopfield model" which

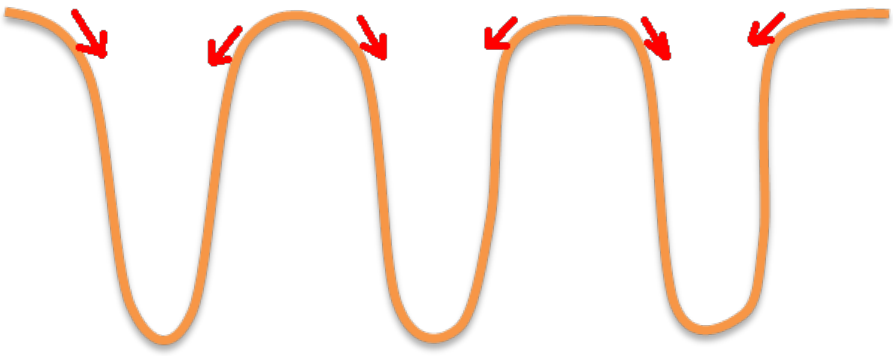


Figure 1.4: Cartoon of 1 dimensional surface of Energy function H in a Hopfield network. Deep wells represent attractors (stored patterns).

is a *standard model* for memory in computational neuroscience, a model neuron has only two (firing or quiescent) states, S_i^0, S_i^1,

$$S_i = \begin{cases} S_i^0 : & h_i < U_i \\ S_i^1 : & h_i > U_i \end{cases}, \tag{1.1}$$

where the input that each neuron experiences is $h_i = \sum_{i \neq j} J_{ij} S_j + h_i^{ext}$ and h_i^{ext} is the contribution external to the network.

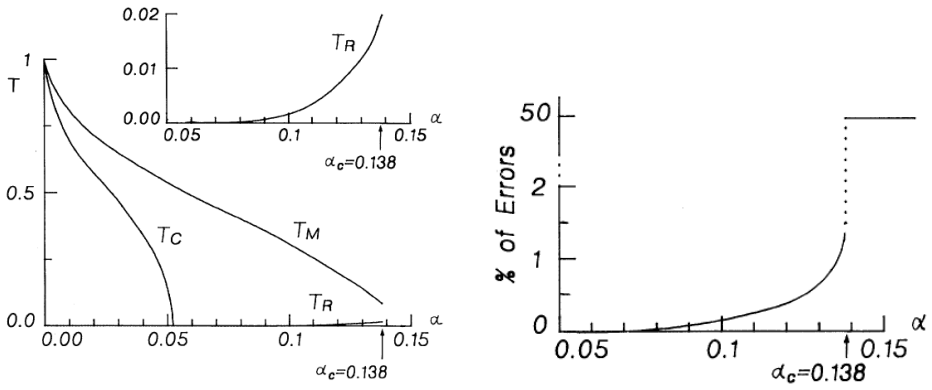
Although it seems to oversimplify the nature of actual neurons, it already reveals rich properties of neural networks just as in statistical physics the simple Ising model already captures the properties of phase transitions between different phases (paramagnetic and ferromagnetic), low energy excitations and other emergent behaviors. Synaptic connections between pre and post synaptic neurons i, j are imprinted following a Hebb rule

$$J_{ij} = \lambda \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \tag{1.2}$$

where p binary patterns $\{\xi_i^\mu | \mu = 1, \cdots, p; i = 1, \cdots, N\}$ are considered over N neurons. The basic assumption is that, at the stage of learning patterns, repeated and persistent stimulation of the pre synaptic neuron onto the post synaptic neuron reinforces the connection between them. Two neurons that are active are likely to be coupled together, which was expressed as "fire together, wire together" by Hebb. Furthermore this *synaptic plasticity* is additive with respect to the learned patterns.
With this learning rule the patterns that are stored become attractors or content-addressible memories of the global dynamics, in the Hopfield network. Mathematically, conditions that $J_{i,j} = J_{j,i}$ and $J_{i,i} = 0$ for arbitrary i guarantee the existence of those attractors, but relaxing the symmetry of $J_{i,j}$ does not fundamentally alter the picture that we have in mind.
The Hamiltonian (energy function) that governs such attractor dynamics can be defined

$$H = -\frac{1}{2} \sum_{i \neq j} J_{i,j} S_i S_j - \sum_i h_i^{ext} S_i. \tag{1.3}$$

Dynamic attractors in the energy landscape are also called autoassociative memories since any initial state of the network (cue) will lead to the successful retrieval of the closest stored pattern.



(a) (b)

Figure 1.5: Phase diagram in terms of $\alpha = p/N$ and temperature T (a); average fraction of errors in FM states at $T = 0$ (b). p is the total number of stored patterns and N the total number of neurons. Both figures are taken from [4]. See the text for T_R, T_M and T_C.

We present the results from [4] for the Hopfield model in the absence of an external field. In Fig.1.5a, the possible phases are presented in $\alpha - T$ phase diagram,

where $\alpha = p/N$, p is the total number of stored patterns and N the total number of neurons, considering the limits $p \to \infty$, $N \to \infty$. T_M is the temperature at which the first ferromagnetic (FM) solutions appear, T_c the temperature below which FM states become absolute minima. T_R is the temperature below which replica symmetry breaking occurs. In Fig.1.5b, the average percentage of errors is shown as a function of α. $\alpha_c = 0.138$ is the storage capacity at $T = 0$ above which retrieval fails for the Hopfield model. The replica method was applied and the plots were obtained by solving mean field equations. For more details, we refer to [29–33].

1.4 Beyond simple cued retrieval

As already mentioned in the introduction, one of our main questions is to understand how the human brain produces creative thoughts and behavior.

Indeed, systems neuroscience has mainly focused on the states induced, in particular in the cortex, by external inputs, be these states simple distributions of neuronal activity or more complex dynamical trajectories. It has largely eschewed the question of how such states can be combined into novel sequences that express, rather than the reaction to an external drive, spontaneous cortical dynamics. Yet, the generation of novel sequences of states drawn from even a finite set has been characterized as the infinitely recursive process deemed to underlie language productivity, as well as other forms of creative cognition [7]. If the individual states, whether fixed points or stereotyped trajectories, are conceptualized as dynamical attractors [34], the cortex can be thought of as engaging in a kind of chaotic saltatory dynamics between such attractors [35]. Attractor dynamics has indeed fascinated theorists, and a major body of work has shown how to make relevant for neuroscience the concepts and analytical tools developed within statistical physics, but the focus has been on compact, homogeneous neural networks [28–31, 36]. These have been regarded as simplified models of local cortical networks – as well as, e.g., of the CA3 hippocampal field – and have not been analysed in their potential saltatory dynamics, given that it would make no sense to consider local cortical networks as isolated systems. Even in the case of a ground-breaking investigation of putative spatial trajectory planning [37], the hippocampal activity that expressed it was thought not to be entirely endogeneous, but rather guided by external inputs, including those representing goals and path integration. Therefore, formal analyses of model networks endowed with attractor dynamics have been largely confined to the simple paradigm of cued retrieval from memory.

Attempts have been made to explore methodologies to study mechanisms beyond simple cued retrieval. The first primitive trial was by Abeles in 1982 [38]. He came up with a kind of feedforward network, which consists of multiple chains and each link in a chain contains a certain number of neurons that are intended to be synchronized. At every unit of time in the evolution of the network, each neuron in a link influences all of the neurons in the next link in the chain. The probability of the existence of the chain was studied as a function of the number of neurons

per chain. But still, the model is far from capturing dynamic processes for thoughts.

In 1986, Sompolinsky and Kanter proposed a network model in [39] that is capable of retrieving time series of patterns. They essentially added to the Hopfield model an asymmetric *hetero*-coupling term with a similar associative learning rule

$$J_{ij} = J_{ij}^1 + J_{ij}^2 = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu + \frac{\lambda}{N}\sum_{\mu=1}^{q} \xi_i^{\mu+1}\xi_j^\mu,$$

where $i \neq j$, $q < p$. λ is a parameter that controls the strength of the instructions with which we endow the network. They observed that at small λ, the sequence of instructed patterns are retrieved in time with fair qualities (worse than in the absence of λ, though), but when λ becomes strong enough, the sequence is smeared out quickly.

In 1987, Tsuda *et al.* [40] suggested another model of nonequilibrium neural network in which there are two building blocks (I,II). Block I is a recurrent network with positive and negative feedback, while block II has extra negative feedback that destabilizes the attractor states of the whole system. The recurrent network models essentially the ensemble of axon collaterals of pyramidal cells, while positive and negative feedback reflect the stellate and basket cells. The negative feedback that only exists in block II captures the properties of specific inhibitory neurons. The resulting network performs the dynamic process of recalling and its trajectory in phase space can be chaotic, depending on the system parameters. They called *attractor ruin* the state of destabilized attractors. For more details, we refer to [35, 40, 41].

In 1992, Herrmann *et al.* studied a Hopfield-like model with a provision for variable thresholds, by imposing their time dependence [42]. Concepts are stored hierarchically in semantic classes, and the authors focused on three kinds of transitions: semantic transitions within semantic class, episodic transition between different semantic classes, random transitions.

There are examples involved in drawing, confabulation, thought processes in general, and language, which are all considered to be largely independent of external stimuli, at their core, and to combine generativity with recursion [7, 43–48].
The choice for models that emulate recursive processes remains open. In this thesis, we use the adaptive Potts network proposed by Treves in 2005 [8] and explore interesting aspects of the latching. For the mathematical conditions under which latching exists, we refer to [49]. The model itself will be discussed in chapter 2 in detail.

Chapter 2

Potts neural networks

In this chapter, we introduce a Potts network in detail. In this thesis we sometimes differentiate the notion of adaptive Potts network from a static one, depending on whether or not it contains time dependence of thresholds for its units. In fact, we always mean adaptive network when we refer to a Potts network, except in chapter 3, where we derive it from a multi-modular model.

Potts neural networks, originally studied merely as a variant of mathematical or potentially applied interest [11–15], offer one approach to model spontaneous dynamics in extended cortical systems, in particular if simple mechanisms of temporal adaptation are taken into account [8]. They can be subject to rigorous analyses of e.g. their storage capacity [20], or of the mechanics of saltatory transitions between states [21] and are amenable to a description in terms of distinct 'thermodynamic' phases [5,50]. The dynamic modification of thresholds with timescales separate from that of retrieval, i.e., temporal adaptation, together with the correlation between cortical states, are key features characterizing cortical operations, and Potts network models may contribute to elucidate their roles. Adaptation and its role in semantic priming [51] have been linked to the instability manifested in schizophrenia [52].

The Potts description is admittedly an oversimplified effective model for an underlying two-level auto-associative memory network [17]. The even more drastically simplified model of latching dynamics considered by the Tsodyks group [53, 54], however, has afforded spectacular success in explaining the scaling laws obtained for free recall in experiments performed 50 years ago. The Potts model may be relevant to a wide set of behaviors and to related experimental measures, once the correspondence between model parameters and the quantities characterizing the underlying two-level network are elucidated. This correspondence will be the topic of chapter 3.

2.1 Potts units

Let us consider an attractor neural network model comprised of Potts units, as depicted in Fig.2.1. The rationale for the model is that each unit represents a local network of many neurons with its own attractor dynamics [28, 31], but in a simplified/integrated manner, regardless of detailed local dynamics. Local attractor states

are represented by $S+1$ Potts states: S active ones and one quiescent state (intended to describe a situation of no retrieval in the local network), σ_i^k, $k = 0, 1, \cdots, S$, with the constraint that $\sum_{k=0}^{S} \sigma_i^k \equiv 1$. We call this autoassociative network of Potts units a Potts network, and refer to studies of some of its properties [5, 8, 20, 21, 55].

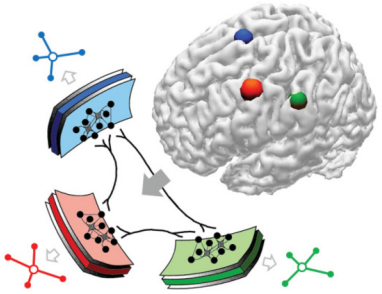


Figure 2.1: Global cortical model as a Potts neural network, redrawn from [5].

The 'synaptic' connection between two Potts units is in fact a tensor summarizing the effect of very many actual connections between neurons in the two local networks, but still, following the Hebbian learning rule [56], the connection weight between unit i in state k and unit j in state l can be written as [20]

$$J_{ij}^{kl} = \frac{c_{ij}}{Ca(1-a/S)} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu,k} - \frac{a}{S}\right)\left(\delta_{\xi_j^\mu,l} - \frac{a}{S}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{2.1}$$

where c_{ij} is 1 if two units i and j have a connection and 0 otherwise, C is the average number of connections per unit, a is the sparsity parameter, i.e. the fraction of active units in every stored global activity pattern ($\{\xi_i^\mu\}$, $\mu = 1, 2, \cdots, p$) and p is the number of stored patterns. The last two delta functions imply that the learned connection matrix does not affect the quiescent states. We will use the indices i, j for units, k, l for states and μ, ν for patterns. Units are updated in the following way:

$$\sigma_i^k = \frac{\exp\left(\beta r_i^k\right)}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U)\right]} \tag{2.2}$$

and

$$\sigma_i^0 = \frac{\exp\left[\beta(\theta_i^0 + U)\right]}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U)\right]}, \tag{2.3}$$

where r_i^k is the input to (active) state k of unit i integrated over a time scale τ_1, while U and θ_i^0 are, respectively, the constant and time-varying component of the effective overall threshold for unit i, which in practice act as inverse thresholds on its quiescent state. θ_i^0 varies with time constant τ_3, to describe local network adaptation and inhibitory effects. The stiffness of the local dynamics is parametrized by the inverse 'temperature' β (or T^{-1}), which is then distinct from the standard notion of thermodynamic noise. The input-output relations (2.2) and (2.3) ensure that

$$\sum_{k=0}^{S} \sigma_i^k = 1.$$

In addition to the overall threshold, θ_i^k is the threshold for unit i specific to state k, and it varies with time constant τ_2, representing adaptation of the individual neurons active in that state, i.e. their neural or even synaptic fatigue.

2.2 Dynamics

The time evolution of the network is then governed by equations that include three distinct time constants:

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t) \tag{2.4}$$

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) \tag{2.5}$$

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t), \tag{2.6}$$

where the field that the unit i in state k experiences reads

$$h_i^k = \sum_{j \neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l + w \left(\sigma_i^k - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l \right). \tag{2.7}$$

The 'local feedback term' w is a parameter, first introduced in [5] that modulates the inherent stability of Potts states, i.e. that of local attractors in the underlying network model. It helps the network converge to an attractor faster by giving positive feedback to the most active states and so it effectively deepens their basins of attraction. Note that in this formulation, feedback is effectively spread over (at least) three time scales: w is positive feedback mediated by collective attractor effects at the neural activity time scale τ_1, θ_i^k is negative feedback mediated by fatigue at the slower time scale τ_2, while θ_i^0 is also negative, and it can be used to model both fast and slow inhibition; for analytical clarity, we consider the two options separately, as the 'slowly adapting regime', with $\tau_3 > \tau_2$, and the 'fast adapting regime', with $\tau_3 < \tau_1$. It would be easy, of course, to introduce additional time scales, for example by distinguishing a component of θ_i^0 that varies rapidly from one that varies slowly, but it would greatly complicate the observations presented in the following.

The overlap or correlation of the activity state of the network with the global memory pattern μ can be measured as

$$m_\mu = \frac{1}{Na\,(1 - a/S)} \sum_{j \neq i}^{N} \sum_{l \neq 0}^{S} \left(\delta_{\xi_j^\mu l} - \frac{a}{S} \right) \sigma_j^l. \tag{2.8}$$

Randomly correlated memory patterns are generated according to the following probability distribution

$$\begin{aligned} P(\xi_i^\mu = k) &= \frac{a}{S}, \\ P(\xi_i^\mu = 0) &= 1 - a, \end{aligned} \tag{2.9}$$

while an algorithm of generating correlated patterns is presented in Appendix A and further details can be found in [8], [57].

Chapter 3

The Potts network as a quantitative model of the cortex

In recent years considerable attention has been paid to the ambition to reconstruct and simulate in microscopic detail the structure of the human brain, possibly at the 1:1 scale, with outcomes that have been questioned [58]. A complementary perspective is that put forward by the late neuroanatomist Valentino von Braitenberg, who in many publications argued for the need to understand overarching principles of mammalian brain organization, even by recourse to dramatic simplification [2]. In this spirit, over 40 years ago Braitenberg proposed the notion of the *skeleton* cortex, that is comprised solely of its $\mathcal{N}$ pyramidal cells [9]. Since on their apical dendrites they receive predominantly synapses from axons that originate in the pyramidal cells of other cortical areas and travel through the white matter, while on their basal dendrites they receive mainly synapses from local axon collaterals, and the two systems, A(pical) and B(asal), can be estimated to include similar numbers of synapses per receiving cell, Braitenberg further detailed what could have later been called a *small world* scheme [26]. In such a scheme, the $\mathcal{N}$ pyramidal cells are allocated to $N = \sqrt{\mathcal{N}}$ modules, each including N cells, fully connected with each other. Each cell would further receive, on the A system, $N - 1$ connections from one cell drawn at random in each of the other modules. Therefore each cell gets $2(N - 1)$ connections from other pyramidal cells, the A and B systems are perfectly balanced, and the average minimal path length between any cell pair is just below 2. Of course, the modules are largely a fictional construct, apart from special cases, or at least their generality and character are quite controversial [59–61], but the distinction between long-range and local connections is real, and the simple model recapitulates a rough square-root scaling of both systems, with $N \sim 10^3 \div 10^5$, in skeleton cortices which in mammals range from ca. $\mathcal{N} \sim 10^6$ to ca. $\mathcal{N} \sim 10^{10}$.

The functional counterpart to the neuroanatomical scheme is the notion of Hebbian associative plasticity [56], considered as the key mechanism that modulates both long- and short-range connections between pyramidal cells. In such a view, autoassociative memory storage and retrieval are universal processes through which both local and global networks operate [2]. Cortical areas across species would then share these universal processes, whereas the information they express would be specific to the constellation of inputs each area receives, which the simplified

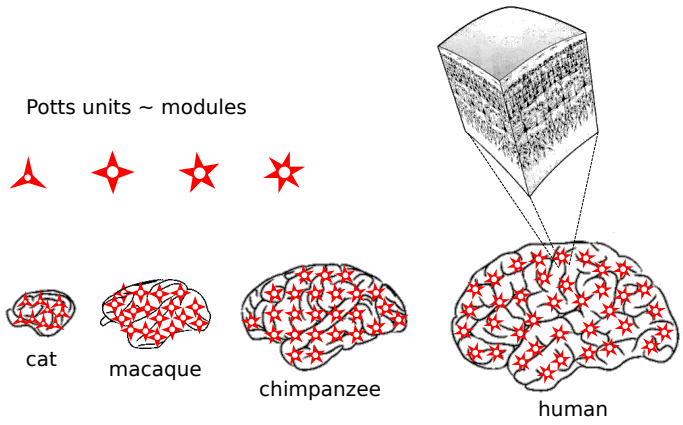


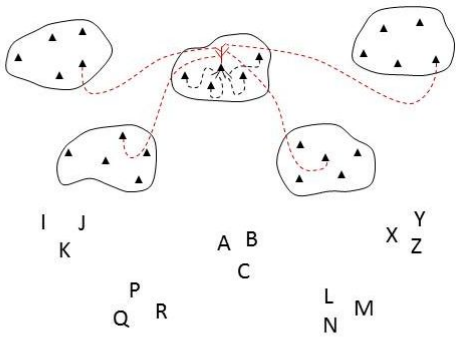
Figure 3.1: The Braitenberg model regards a skeleton cortex of $\mathcal{N}$ pyramidal cells as comprised of $\sqrt{\mathcal{N}}$ modules of $\sqrt{\mathcal{N}}$ cells each. The Potts model then reduces each module to a multi-state unit, where a state corresponds to a dynamical attractor of the local cortical module. How should the number of states per module, S, be thought to scale with $\mathcal{N}$?

skeleton model does not attempt to describe. Underlying the diversity of higher-order processes of which cortical cognition is comprised, there would be the common associative operation of multi-modular autoassociative memory.

The Hopfield model of a simple autoassociative memory network [28] has opened the path to a quantitative statistical understanding of how memory can be implemented at the network level, through thorough analyses of attractor neural networks. The initial analyses, with networks of binary units, then shifted towards networks with more of the properties seen in the cortex [62, 63].

As for connectivity, attempts to reproduce quantitative observations [64], given the apparent lack of specificity at the single cell level [65], in some cases have led to models without modules, but in which the probability of pyramidal-to-pyramidal connections depends on the distance between neurons, rapidly decreasing beyond a distance that conceptually corresponds to the radius of a module [66].

But has Braitenberg's suggested simplification, the skeleton of units with their A and B system, enabled the use of the powerful statistical-physics-derived analyses that had been successfully applied to the Hopfield model? Only up to a point. Studies of multi-modular network models including full connectivity within individual modules and sparse connectivity with other modules could only be approached in their most basic formulation, in which all modules participate in every memory, and their sparse connectivity is random [16, 17]; and attempts to articulate them further have led to analytical complexity [18, 19, 67, 68] or to the recourse to effectively local coding schemes [69], without yielding a plausible quantification of storage capacity. The Potts associative network, in contrast, has been fully analyzed in its original and sparsely coded versions [11–15, 20] and it has been argued to offer an ever further simplification of a cortical network than Braitenberg's [8], amenable to study also its latching dynamics [5]. The correspondence between Braitenberg's notion and the Potts model has not, however, been discussed. In this chapter, we do it with the



memory\module	1		m		N_m
1		P		M	X
	J	Q		M	
μ	K		B		Y
	I	Q	A		
		R		L	Z
p	K		C	N	

Figure 3.2: Toy model for a cortex comprised of modules, in which each pyramidal cell in the module receives sparse inputs from other modules on the apical dendrites-dashed line (in color) on top panel. Five modules, each of which contains five neurons and three features (local patterns/attractors) are presented for illlustration ($N_m = 5, p = 6, S = 3, a = 0.6$). A global memory patterns in the table can be thought of as comprised of features. Features have to be bound together by the tensor connections, in the Potts model, where sparse coding means that not all features pertain to every memory; the rest of the Potts units are in their quiescent state.

aim at establishing a clearer rationale for using the Potts model to study cortical processes.

3.1 From a multi-modular Hopfield network to a Potts network

3.1.1 Thermodynamic correspondence

Let us consider an underlying network of N_m modules ([16–19]), each comprised of N_u neurons, each of which is connected to all $N_u - 1$ other neurons within the same module, and to L other neurons distributed randomly throughout all the other modules. We make the critical "Hopfield" assumption [28] that both short- and long-range synaptic connections are symmetric. Each module can retrieve one of S local activity patterns, or *features*, that are learned with the corresponding short range connections. We index it with $k = 1, \ldots, S$. Furthermore, p global activity patterns, each consisting of combinations of aN_m features, are stored on the dilute long-range connections, as illustrated in Fig.3.2.

Let us make here the simplifying assumption that the firing rates, η, that represent a local pattern k within a module m, are identically and independently distributed across units, given by the distribution $P_\eta\left(\eta_{im}^k\right)$. A global pattern,

$\mu = 1, \ldots, p$, is a random combination $\{k_1^\mu, \ldots, k_m^\mu, \ldots, k_{N_m}^\mu\}$. Note as $\zeta \equiv pa/S$ the average number of global patterns represented by a specific local pattern, given global sparsity a, and assume it for simplicity to be an integer number. The total number of connections to a neuron is given by $C = L + N_u - 1$ and we define the fraction of long range connections as $\gamma = L/C$. We also impose, as in [6], that P_η satisfies $\langle \eta \rangle = \langle \eta^2 \rangle = a_u$, such that local representations are also sparse, with sparsity parameter a_u distinct from the global one a, both measures parametrizing, at different scales, sparse coding.

Using Hebbian covariance rules [70] in the multi-modular network, we have

$$J^{\text{short}}_{i_m,j_m} = \rho_s \frac{1}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^{k_m^\mu}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{k_m^\mu}_{j_m}}{a_u} - 1 \right) \tag{3.1}$$

$$J^{\text{long}}_{i_m,j_n} = \rho_l \frac{c_{i_m,j_n}}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^{\mu}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{\mu}_{j_n}}{a_u} - 1 \right) \tag{3.2}$$

where ρ_s and ρ_l are parameters that adjust the dimensions of short- and long-range connections, and can regulate their relative strength. The variable c_{i_m,j_n} is a binary variable

$$c_{i_m,j_n} = \begin{cases} 1 & \text{with probability} \quad \epsilon \\ 0 & \text{with probability} \quad (1-\epsilon) \end{cases} \tag{3.3}$$

where $\epsilon = L/N_u(N_m - 1)$.

In those cases in which an energy function can be defined, i.e., essentially, if $c_{i_m,j_n} = c_{j_n,i_m}$ the attractor states of the system, [31], correspond to the minima of a "free energy". The "Hamiltonian" of the multi-modular network, which is proportional to $N_u \times N_m$, is in those cases given by

$$\begin{aligned} \mathcal{H} &= -\frac{1}{2} \sum_m \sum_{i_m, j_m \neq i_m} J^{\text{short}}_{i_m,j_m} V_{i_m} V_{j_m} - \frac{1}{2} \sum_{m,n \neq m} \sum_{i_m,j_n} J^{\text{long}}_{i_m,j_n} V_{i_m} V_{j_n} \\ &= \mathcal{H}_s + \mathcal{H}_l \end{aligned} \tag{3.4}$$

where V_{i_m} can be threshold linear type defined as

$$V_{i_m}(\tau + \delta\tau) = \begin{cases} 0 & h_{i_m}(\tau) < T_{thr} \\ g \cdot (h_{i_m}(\tau) - T_{thr}) & h_{i_m}(\tau) > T_{thr}. \end{cases}$$

Threshold linear description of the unit is adopted in order to mimic the real firing operation of neurons ([6], [71], [72]). It is depicted in Fig.3.3. It is simple enough to treat the network in analytial way and complex enough to capture the real firing behaviour.

Estimating c_{i_m,j_n} with its mean ϵ, we can rewrite the second term as

$$\begin{aligned} \mathcal{H}_l &= -\sum_{m,n>m} \sum_{i_m,j_n} J^{\text{long}}_{i_m,j_m} V_{i_m} V_{j_n} \\ &= -\rho_l \sum_{m,n>m} \sum_{i_m,j_n} \frac{c_{i_m,j_n}}{C} \sum_{\mu=1}^{p} \left(\frac{\eta^{\mu}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{\mu}_{j_n}}{a_u} - 1 \right) V_{i_m} V_{j_n} \\ &\simeq -\rho_l \frac{\epsilon}{C} \sum_{m,n>m} \sum_{\mu} \sum_{i_m,j_n} \left(\frac{\eta^{\mu}_{i_m}}{a_u} - 1 \right) \left(\frac{\eta^{\mu}_{j_n}}{a_u} - 1 \right) V_{i_m} V_{j_n}. \end{aligned}$$

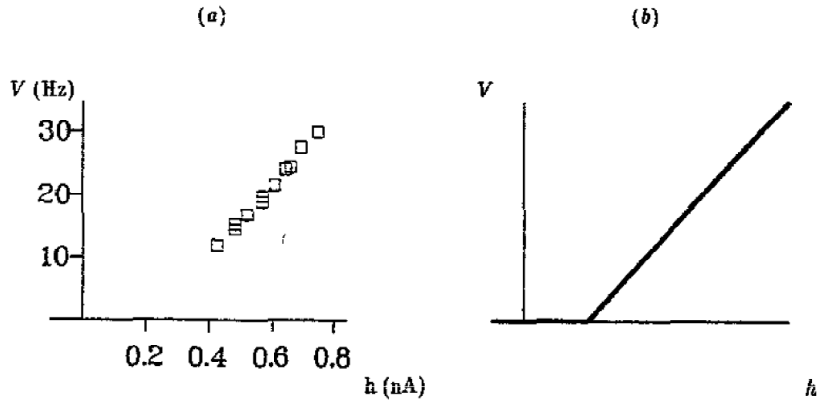


Figure 3.3: Threshold linear model: (a) Experimental data coming from a layer 2/3 pyramidal cell in rat visual cortex; (b) model. Plots are cited from [6].

For a given pattern μ the only contribution to $\eta_{i_m}^{\mu}$ is $\eta_{i_m}^{\xi_m^\mu}$. Let us now define the local correlation of the state of the network with each local memory pattern as

$$\sigma_m^{\xi_m^\mu} = \frac{1}{N_u}\sum_{i_m}\left(\frac{\eta_{i_m}^{\xi_m^\mu}}{a_u} - 1\right)V_{i_m} \tag{3.5}$$

where to avoid introducing additional dimensional parameters, we assume that the activity V_i of each model neuron is measured in such units, and suitably regulated by inhibition, that the local correlations are automatically normalized to reach a maximum value of 1. We then obtain

$$\begin{aligned}\mathcal{H}_l &= -\rho_l\frac{\epsilon N_u^2}{C}\sum_{m,n>m}\sum_{\mu}\sigma_m^{\xi_m^\mu}\sigma_n^{\xi_n^\mu}\\ &= -\rho_l\frac{\epsilon N_u^2}{C}\sum_{m,n>m}\sum_{\mu}\sum_{k}\sum_{l}\delta_{\xi_m^\mu k}\delta_{\xi_n^\mu l}\sigma_m^k\sigma_n^l\\ &= -N_u\sum_{m,n>m}\sum_{k,l}J_{mn}^{kl}\sigma_m^k\sigma_n^l\,,\end{aligned} \tag{3.6}$$

where we have introduced

$$J_{mn}^{kl} = \rho_l\frac{\epsilon N_u}{C}\sum_{\mu}\delta_{\xi_m^\mu k}\delta_{\xi_n^\mu l} = \rho_l\frac{\gamma}{N_m-1}\sum_{\mu}\delta_{\xi_m^\mu k}\delta_{\xi_n^\mu l}. \tag{3.7}$$

On the other hand, using (3.5), the first term can be rewritten as

$$\begin{aligned}\mathcal{H}_s &= -\sum_m\sum_{i_m,j_m>i_m}J_{i_m,j_m}^S V_{i_m}V_{j_m}\\ &\simeq -\rho_s\frac{\zeta}{C}\sum_m\sum_{i_m,j_m>i_m}\sum_{\xi=1}^{S}\left(\frac{\eta_{i_m}^{\xi}}{a_u}-1\right)\left(\frac{\eta_{j_m}^{\xi}}{a_u}-1\right)V_{i_m}V_{j_m}\\ &= -\rho_s\frac{\zeta}{C}\sum_m\sum_{\xi=1}^{S}\left\{\sum_{i_m,j_m}\left(\frac{\eta_{i_m}^{\xi}}{a_u}-1\right)\left(\frac{\eta_{j_m}^{\xi}}{a_u}-1\right)V_{i_m}V_{j_m} - \sum_{i_m}\left[\left(\frac{\eta_{i_m}^{\xi}}{a_u}-1\right)V_{i_m}\right]^2\right\}\\ &\simeq -\rho_s\frac{\zeta}{C}\sum_m\left\{N_u^2\sum_k\left(\sigma_m^k\right)^2 - \frac{S(1-a_u)}{a_u}\sum_{i_m}\left[V_{i_m}\right]^2\right\}.\end{aligned} \tag{3.8}$$

where we have noted the absence of self-interactions, and estimated with its mean $\zeta \equiv pa/S$ the number of contributions to the encoding of each local attractor state.

Putting together (3.6) and (3.8), where we neglect the last term in the $N_u \to \infty$ limit, and noting that $N_u/C \simeq 1-\gamma$, we have

$$\mathcal{H} \simeq -N_u \sum_{m,n>m} \sum_{k,l} J^{kl}_{mn} \sigma^k_m \sigma^l_n - N_u \rho_s \zeta (1-\gamma) \sum_m \sum_k \left(\sigma^k_m\right)^2 . \tag{3.9}$$

We have therefore expressed the Hamiltonian of a multi-modular Hopfield network in terms of *mesoscopic* parameters, the σ^k_m's, characterizing the state of each module in terms of its correlation with locally stored patterns. This could be regarded as (proportional to) the effective Hamiltonian of a reduced Potts model, if due attention is paid to entropy and temperature. Since the σ^k_m's are infinite (in the $N_m \to \infty$ limit) but infinitely fewer than the V_i's (in the $N_u \to \infty$ limit), the correct Potts Hamiltonian is akin to a free-energy for the full multimodular model, it should scale with N_m and not with $N_m \times N_u$, and it should include the proper entropy terms. One can write

$$\exp -\beta_{Potts} \mathcal{H}_{Potts}(\{\sigma^k_m\}) = \sum_{\{V_i\}} \exp -\beta \mathcal{H}(\{V_i\}|\{\sigma^k_m\}). \tag{3.10}$$

The correct scaling of the Potts Hamiltonian implies that an extra N_u factor present in the original Hamiltonian has to be reabsorbed in the effective inverse Potts temperature β_{Potts}, which then diverges in the thermodynamic limit. This means that the Potts network can be taken to operate at zero temperature, in relation to its interactions between modules. Within modules, however, the effects of a non-zero noise level in the underlying multi-modular network persist in the entropy terms. These can be estimated by suitable assumptions on the distribution of microscopic configurations that dominate the thermodynamic (mesoscopic) state of each module. One such assumption is that a module is mostly in states fragmented into competing *domains* of $n_0, n_1, \ldots, n_k, \ldots, n_S$ units, fully correlated with the corresponding local patterns, except for the first n_0, which are at a spontaneous activity level. This would imply that, dropping the module index m, $\sigma^k = n_k/N_u$, and the constraint $\sum_{k=0}^S \sigma^k = 1$ is automatically satisfied. The number of microscopic states characterized by the same $S+1$-plet $n_0, \ldots, n_k, \ldots, n_S$ is $N_u! / \prod_{k=0}^S n_k!$. The log of this number, which can be estimated as $-N_u \sum_{k=0}^S \sigma^k \ln \sigma^k$, has to be divided by β and then subtracted for each module from the original Hamiltonian, as the entropy term that comes from the microscopic free-energy. This becomes the effective Hamiltonian of the Potts network by further dividing by N_u, because a factor N_u has to be reabsorbed into β. Therefore one finds the additional entropy term in the reduced Hamiltonian

$$\beta \mathcal{H}^{\text{entropy}}_{Potts}(\{\sigma^k_m\}) = \sum_m \sum_{k=0}^S \sigma^k_m \ln \sigma^k_m. \tag{3.11}$$

The above shows that the original inverse temperature β retains its significance as a local parameter, that modulates the stiffness of each module or Potts units, even though the effective noise level in the long-range interactions between modules vanishes. The precise entropy formula depends also on the assumptions that all microscopic states be dynamically accessible from each other, which would have to be validated depending on the dynamics assumed to hold within each module. An alternative assumption is that individual units can in practice only be exchanged

between a fragment correlated with local pattern k and the pool n_0 of uncorrelated units. Under that assumption the entropy can be estimated from the log of the number $\prod_{k=1}^{S}(N_u!/n_0!n_k!)$, which yields

$$\beta\mathcal{H}'^{\text{entropy}}_{Potts}(\{\sigma_m^k\}) = \sum_m \sum_{k=1}^{S} \left\{ \sigma_m^k \ln \frac{\sigma_m^k}{\sigma_m^k + \sigma_m^0} + \sigma_m^0 \ln \frac{\sigma_m^0}{\sigma_m^k + \sigma_m^0} \right\} \tag{3.12}$$

as in [5].

Note that, in (3.9), the sparse connectivity between modules of the multi-modular network does not translate into a diluted Potts connectivity: each module, or Potts unit, receives inputs from each of the other $N_m - 1$ modules, or Potts units. One can consider cases in which, instead, there are only c_m connections per Potts unit, e.g. the *highly diluted* and *intermediate connectivity* considered in the storage capacity analysis below.

3.1.2 Parameters for the dynamics

These arguments indicate how the local attractors of each module can be reinterpreted as dynamical variables of a system of interacting Potts units. The correspondence cannot be worked out completely, however (and (3.9) is not fully equivalent to the Hamiltonian defined in [5]), if anything because the effects of inhibition cannot be included, given the inherent asymmetry of the interactions, in a Hamiltonian formulation. In the body of work on neural networks stimulated by the Hopfield model, some of the effects ascribed to inhibition have been regarded as incapsulated in the peculiar *Hebbian* learning rule that determines the contribution of each stored pattern to the synaptic matrix, with its subtractive terms. Similar subtractive terms can be argued on the same basis to take into account inhibitory effects at the module level, and they lead to replace the interaction

$$J_{mn}^{kl} = \rho_l \frac{\gamma}{N_m - 1} \sum_\mu \delta_{\xi_m^\mu k} \delta_{\xi_n^\mu l} \tag{3.13}$$

with

$$J'^{kl}_{mn} = \rho_l \frac{\gamma}{N_m - 1} \sum_\mu (\delta_{\xi_m^\mu k} - a/S)(\delta_{\xi_n^\mu l} - a/S), \tag{3.14}$$

the form which appears in [5]. The local feedback term there, parametrized by w, can be made to roughly correspond to the second term in (3.9) by imposing that $\rho_s \zeta (1-\gamma)/\rho_l \gamma = w/2$.

To extend further the approximate correspondence, beyond thermodynamics and into dynamics, we may assume that underlying the Potts network there is in fact a network of $N_m \times N_u$ integrate-and-fire model neurons, emulating the dynamical behaviour of pyramidal cells in the cortex, as considered by [73] and [74]. The simple assumptions concerning the connectivity and the synaptic efficacies are reflected in the fact that the inputs to any model neuron in the extended network are determined by globally defined quantities, namely the mean fields, which are weighted averages of quantities that measure, as a function of time, the effective fraction of synaptic conductances (g, in suitable units normalized to Δg) open on the membrane of any

cell of a given class, or cluster (G) by the action of all presynaptic cells of another given class, or cluster (F)

$$z_G^F(t) = \frac{1}{N_{local,F}} \sum_{\alpha \in F} \frac{g^\alpha(t)}{\Delta g_G^F}, \tag{3.15}$$

where g^α is the conductance of a specific synaptic input. The point is that among the clusters that have to be defined in the framework of Ref. [73], many cluster pairs (F,G), those that comprise pyramidal cells, share the same or a similar biophysical time constant, describing their conductance dynamics [73], i.e.

$$\frac{dz_G^F(t)}{dt} = -\frac{1}{\tau_G^F} z_G^F(t) + \nu_F(t - \Delta t), \tag{3.16}$$

where $\nu_F(t)$ is the firing rate. If τ_G^F is the same across distinct values for F and G, one can compare the equation for any such cluster pair to the first equation of (2.4), namely

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t).$$

Since r_i^k is the temporally integrated variable representing the activity of unit i in state k varying with the time scale of τ_1, it can be taken to correspond to the (integrated) activation of pyramidal cells in a module. One can conclude that τ_1 summarizes the time course of the conductances opened on pyramidal cells by the inputs from other pyramidal cells. It represents the inactivation of synaptic conductance and, like the firing rates are a function of the z, our overlap is a function of the r. Neglecting adaptation (θ_i^k), we can think of the correspondence as

$$h_i^k \sim \sum_{\alpha \in F} \nu^\alpha \to r_i^k \sim \sum_{\alpha \in F} z^\alpha \tag{3.17}$$

therefore r_i^k represents the state of the inputs to the integrate-and-fire neurons within a module, i.e., a Potts unit, and we can identify the constant τ_1 with the inactivation time constant for the synapses between pyramidal cells, τ_E^E, whereas inhibitory and adaptation effects will be represented by τ_2 and τ_3 in the Potts model.

3.2 Storage capacity of the Potts network

3.2.1 Fully connected network

In the previous section, we have expressed the approximate equivalence between the Hamiltonian of a multi-modular Hopfield network and that of the Potts network. This means that we can study the retrieval properties of the Potts network, as an effective model of the full multi-modular network. In this section, we study the storage capacity of the Potts network with full connectivity using the classic replica method. Taking inspiration from [20] and [5], let us consider the Hamiltonian which is defined as:

$$\mathcal{H} = -\frac{1}{2}\sum_{i,j\neq i}^{N}\sum_{k,l=0}^{S} J_{ij}^{kl}\delta_{\sigma_i k}\delta_{\sigma_j l} + U\sum_{i}^{N}(1-\delta_{\sigma_i 0}) - \frac{w}{2}\sum_{i}^{N}\left[\sum_{k>0}\delta_{\sigma_i k}^2 - \frac{1}{S}(1-\delta_{\sigma_i 0})^2\right]. \quad (3.18)$$

The coupling between the state k in unit i and the state l in unit j is a Hebbian rule ([5, 15, 20, 28, 55])

$$\begin{cases} J_{ij}^{kl} = \frac{1}{Na(1-\tilde{a})}\sum_{\mu=1}^{p} v_{\xi_i^\mu k} v_{\xi_j^\nu l} \\ \\ v_{\xi_i^\mu k} = \left(\delta_{\xi_i^\mu k} - \tilde{a}\right)(1-\delta_{k0}) \end{cases} \quad (3.19)$$

where N is the total number of units in our Potts network (for clarity we drop henceforth the subscript N_m, except when discussing parameters in the end of this chapter), p is the number of stored random patterns, a is their sparsity, i.e., the fraction of active Potts units in each, and $\tilde{a} = a/S$. As mentioned above, U is the time-independent threshold acting on all units in the network, as in [20]. The main difference with the analysis in [20] is that here we have included the term proportional to w in (3.18). This self-reinforcement term pushes each unit into the more active of its states, thus providing positive feedback.

The patterns to be learned are drawn from the following probability distribution ([5, 20, 55])

$$\begin{cases} P(\xi_i^\mu = 0) = 1 - a \\ P(\xi_i^\mu = k) = \tilde{a} \equiv a/S\,. \end{cases} \quad (3.20)$$

Using the trivial property that $\delta_{i,j}^2 = \delta_{i,j}$ we can rewrite the Hamiltonian as

$$\begin{aligned} \mathcal{H} &= -\frac{1}{2Na(1-\tilde{a})}\sum_{\mu=1}^{p}\left(\sum_{i}^{N} v_{\xi_i^\mu \sigma_i}\right)^2 + \frac{1}{2Na(1-\tilde{a})}\sum_{i}^{N}\sum_{\mu=1}^{p} v_{\xi_i^\mu \sigma_i}^2 + \\ &+ \left(U - \frac{w(S-1)}{2S}\right)\sum_{i}^{N}\frac{v_{\xi_i^\mu \sigma_i}}{\delta_{\xi_i^\mu \sigma_i} - \tilde{a}}\,. \end{aligned}$$

In the following let us define

$$\tilde{U} = U - \frac{w(S-1)}{2S}\,. \quad (3.21)$$

We now apply the replica technique ([32,33,75]) to $\mathcal{H}$, following refs. [11,29–31,63]. The free energy of N Potts units in replica theory reads

$$f = -\frac{1}{\beta}\lim_{n\to 0}\lim_{N\to\infty}\frac{\left\langle Z^n\right\rangle - 1}{Nn}, \tag{3.22}$$

where $\langle\cdot\rangle$ is an average over the quenched disorder (in this case represented by the condensed patterns in our network), as in [31].

The partition function Z^n of n replicas can be written as

$$\begin{aligned}\left\langle Z^n\right\rangle &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}}\exp\left[-\beta\sum_{\gamma}^{n}H^\gamma\right]\right\rangle \\ &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}}\exp\left[\frac{\beta}{2Na\left(1-\tilde{a}\right)}\sum_{\mu\gamma}\left(\sum_{i}^{N}v_{\xi_i^\mu\sigma_i^\gamma}\right)^2-\frac{\beta}{2Na\left(1-\tilde{a}\right)}\sum_{i}^{N}\sum_{\mu\gamma}v^2_{\xi_i^\mu\sigma_i^\gamma}\right.\right. \\ &\quad \left.\left.-\beta\tilde{U}\sum_{i\gamma}\frac{v_{\xi_i^\mu\sigma_i^\gamma}}{\delta_{\xi_i^\mu\sigma_i^\gamma}-\tilde{a}}\right]\right\rangle.\end{aligned} \tag{3.23}$$

Using the Hubbard-Stratonovich transformation

$$\exp\left[\lambda a^2\right] = \int\frac{dx}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}+\sqrt{2\lambda}ax\right],$$

the first term in (3.23) can be written as

$$\exp\left[\frac{\beta}{2Na\left(1-\tilde{a}\right)}\left(\sum_{i}^{N}v_{\xi_i^\mu\sigma_i^\gamma}\right)^2\right] = \int\frac{dm_\mu^\gamma}{\sqrt{2\pi}}\exp\left[-\frac{\left(m_\mu^\gamma\right)^2}{2}+\sqrt{\frac{\beta}{Na\left(1-\tilde{a}\right)}}m_\mu^\gamma\sum_{i}^{N}v_{\xi_i^\mu\sigma_i^\gamma}\right].$$

The change of variable $m_\mu^\gamma \to m_\mu^\gamma\sqrt{\beta Na\left(1-\tilde{a}\right)}$, and neglecting the sub-leading terms in the $N\to\infty$ limit, gives us

$$\begin{aligned}\left\langle Z^n\right\rangle &= \left\langle \mathrm{Tr}_{\{\sigma^\gamma\}}\int\prod_{\mu\gamma}dm_\mu^\gamma\cdot\right. \\ &\quad \cdot\exp\beta N\left[\frac{a\left(1-\tilde{a}\right)}{2}\sum_{\mu\gamma}\left(m_\mu^\gamma\right)^2+\sum_{\mu\gamma}\frac{m_\mu^\gamma}{N}\sum_{i}^{N}v_{\xi_i^\mu\sigma_i^\gamma}-\frac{1}{2N^2a\left(1-\tilde{a}\right)}\sum_{i}^{N}\sum_{\mu\gamma}v^2_{\xi_i^\mu\sigma_i^\gamma}\right. \\ &\quad \left.\left.-\frac{1}{N}\tilde{U}\sum_{i\gamma}\frac{v_{\xi_i^\mu\sigma_i^\gamma}}{\delta_{\xi_i^\mu\sigma_i^\gamma-\tilde{a}}}\right]\right\rangle.\end{aligned} \tag{3.24}$$

Discriminating the condensed patterns (ν) from non condensed ones (μ) in the limit $p\to\infty$ and $N\to\infty$ with the fixed ratio $\alpha = p/N$,

$$\begin{aligned}\left\langle Z^n\right\rangle &= \mathrm{Tr}_{\{\sigma^\gamma\}}\int\prod_{\mu\gamma}dm_\mu^\gamma\int\prod_{\lambda\gamma}dq_{\gamma\lambda}dr_{\gamma\lambda}\cdot\exp\left\{-\frac{\beta N}{2}\sum_{\mu>s}\left[a\left(1-\tilde{a}\right)\sum_\gamma\left(m_\mu^\gamma\right)^2\right.\right.\\ &\left.-a\left(1-\tilde{a}\right)\beta\tilde{a}\sum_{\gamma\lambda}m_\mu^\gamma m_\mu^\lambda q_{\gamma\lambda}\right]-\frac{\alpha\beta\tilde{a}N}{2}\sum_{\gamma\gamma}q_{\gamma\gamma}-\beta Na\tilde{U}\sum_{\gamma\gamma}q_{\gamma\gamma}\\ &-\frac{N\alpha\beta^2}{2}\sum_{\gamma\lambda}r_{\gamma\lambda}\left(\tilde{a}^2q_{\gamma\lambda}-\frac{1}{NS\left(1-\tilde{a}\right)}\sum_{ik}P_kv_{k\sigma_i^\gamma}v_{k\sigma_i^\lambda}\right)\Bigg\}\cdot\left\langle\exp\beta N\left[\frac{a(1-\tilde{a})}{2}\right.\right.\\ &\left.\left.\sum_{\nu\gamma}^{\nu\le s}\left(m_\nu^\gamma\right)^2+\sum_{\nu\gamma}^{\nu\le s}\frac{m_\nu^\gamma}{N}\sum_i^N v_{\xi_i^\nu\sigma_i^\gamma}-\frac{1}{2N^2a(1-\tilde{a})}\sum_i^N\sum_{\nu\gamma}^{\nu\le s}v_{\xi_i^\nu\sigma_i^\gamma}^2\right]\right\rangle\end{aligned}\tag{3.25}$$

where we introduced $q_{\gamma\lambda}$, the overlap between different replicas, analogous to the Edwards-Anderson order parameter [76],

$$q_{\gamma\lambda}=\frac{1}{Na\tilde{a}\left(1-\tilde{a}\right)}\sum_{ik}P_kv_{k\sigma_i^\gamma}v_{k\sigma_i^\lambda}\,.\tag{3.26}$$

The saddle point equations are

$$\frac{\partial}{\partial m_\nu^\gamma}=0\longrightarrow m_\nu^\gamma=\left\langle\frac{1}{Na\left(1-\tilde{a}\right)}\sum_i\left\langle v_{\xi_i^\nu\sigma_i^\gamma}\right\rangle\right\rangle,\tag{3.27}$$

$$\frac{\partial}{\partial r_{\gamma\lambda}}=0\longrightarrow q_{\gamma\lambda}=\frac{1}{Na\tilde{a}\left(1-\tilde{a}\right)}\sum_i^N\left\langle\sum_kP_k\left\langle v_{k\sigma_i^\gamma}v_{k\sigma_i^\lambda}\right\rangle\right\rangle,\tag{3.28}$$

$$\frac{\partial}{\partial q_{\gamma\lambda}}=0\longrightarrow r_{\gamma\lambda}=\frac{S\left(1-\tilde{a}\right)}{\alpha}\sum_\mu\left\langle m_\mu^\gamma m_\nu^\lambda\right\rangle-\left[\frac{2S}{\alpha}\tilde{U}+1\right]\frac{\delta_{\gamma\lambda}}{\beta\tilde{a}}\,.\tag{3.29}$$

After performing the multidimensional Gaussian integrals over fluctuating (non condensed) patterns we have

$$\begin{aligned}\left\langle Z^n\right\rangle &= \int\prod_{\nu\gamma}^{\nu\in[1,\ldots,s]}dm_\nu^\gamma\int\prod_{\lambda\gamma}dq_{\gamma\lambda}dr_{\gamma\lambda}\cdot\\ &\cdot\exp N\Bigg\{-\beta\frac{a\left(1-\tilde{a}\right)}{2}\sum_{\nu\gamma}\left(m_\nu^\gamma\right)^2-\frac{\alpha}{2}\mathrm{Tr}\ln\left[a\left(1-\tilde{a}\right)\left(1-\beta\tilde{a}\mathbf{q}\right)\right]-\\ &\frac{\alpha\beta^2\tilde{a}^2}{2}\sum_{\gamma\lambda}r_{\gamma\lambda}q_{\gamma\lambda}-\beta\tilde{a}\left[\frac{\alpha}{2}+S\tilde{U}\right]\sum_{\gamma\gamma}q_{\gamma\gamma}+\left\langle\ln\mathrm{Tr}_{\{\sigma^\gamma\}}\exp\left[\beta\mathcal{H}_\sigma^\xi\right]\right\rangle_{\xi^v}\Bigg\},\end{aligned}\tag{3.30}$$

where

$$\mathcal{H}_\sigma^\xi=\sum_{\nu\gamma}m_\nu^\gamma v_{\xi^\nu\sigma^\gamma}+\frac{\alpha\beta}{2S\left(1-\tilde{a}\right)}\sum_{\gamma\lambda}r_{\gamma\lambda}\sum_kP_kv_{k\sigma^\gamma}v_{k\sigma^\lambda}\,.\tag{3.31}$$

Following (3.22),

$$\begin{aligned} f &= \lim_{n\to 0} f_n = \lim_{n\to 0}\left\{ \frac{a\left(1-\tilde{a}\right)}{2n}\sum_{\nu\gamma}\left(m_\nu^\gamma\right)^2 + \right. \\ &+ \frac{\alpha}{2n\beta}\mathrm{Tr}\ln\left[a\left(1-\tilde{a}\right)\left(1-\beta\tilde{a}\mathbf{q}\right)\right] + \frac{\alpha\beta\tilde{a}^2}{2n}\sum_{\gamma\lambda} r_{\gamma\lambda}q_{\gamma\lambda} \\ &+ \left.\frac{\tilde{a}}{n}\left[\frac{\alpha}{2}+S\tilde{U}\right]\sum_{\gamma\gamma} q_{\gamma\gamma} - \frac{1}{n\beta}\left\langle \ln \mathrm{Tr}_{\{\sigma^\gamma\}}\exp\left[\beta\mathcal{H}_\xi\right]\right\rangle_{\xi^v}\right\}. \end{aligned} \tag{3.32}$$

Furthermore, imposing the replica symmetry [32]

$$\begin{aligned} m_\gamma^\nu &= m \\ q_{\gamma\lambda} &= \begin{cases} q & \text{for} \quad \gamma \neq \lambda \\ \tilde{q} & \text{for} \quad \gamma = \lambda \end{cases} \\ r_{\gamma\lambda} &= \begin{cases} r & \text{for} \quad \gamma \neq \lambda \\ \tilde{r} & \text{for} \quad \gamma = \lambda, \end{cases} \end{aligned}$$

we finally obtain the replica symmetric free energy

$$\begin{aligned} f &= \frac{a\left(1-\tilde{a}\right)}{2}m^2 + \frac{\alpha}{2\beta}\left[\ln\left(a\left(1-\tilde{a}\right)\right) + \ln\left(1-\tilde{a}C\right) - \frac{\beta\tilde{a}q}{\left(1-\tilde{a}C\right)}\right] + \\ &+ \frac{\alpha\beta\tilde{a}^2}{2}\left(\tilde{r}\tilde{q}-rq\right) + \tilde{a}\tilde{q}\left[\frac{\alpha}{2}+S\tilde{U}\right] + \\ &- \frac{1}{\beta}\left\langle \int D\mathbf{z}\ln\left(1+\sum_{\mathbf{l}\neq 0}\exp\left[\beta\mathcal{H}_l^\xi\right]\right)\right\rangle, \end{aligned} \tag{3.33}$$

where $C = \beta\left(\tilde{q}-q\right)$ and

$$\mathcal{H}_l^\xi = m v_{\xi l} - \frac{\alpha a\beta\left(r-\tilde{r}\right)}{2S^2}\left(1-\delta_{l0}\right) + \sum_{k=1}^{S}\sqrt{\frac{\alpha r P_k}{S\left(1-\tilde{a}\right)}}z_k v_{kl}\,. \tag{3.34}$$

Detailed derivation of replica symmetric free energy is in Appendix B. C and $\mathcal{H}_l^\xi$ are both quantities that are typical of a replica analysis. $\mathcal{H}_l^\xi$ is the mean field with which the network affects state l in a given unit if it is in the same state as condensed pattern ξ (note that $\mathcal{H}_l^\xi = 0$). C measures the difference between $\tilde{q}$, the mean square activity in a given replica, and q, the coactivation between two different replicas. Note that in the zero temperature limit ($\beta \to \infty$), this difference goes to 0, such that C is always of order 1. It will be clarified in another section, through a separate analysis, that C is related to the derivative of the output of an average neuron with respect to variations in its mean field.

The self-consistent mean field equations in the limit of $\beta \to \infty$ are obtained by taking the derivatives of f with respect to the three replica symmetric variational parameters, m, q, r

$$\begin{aligned} m &= \frac{1}{a(1-\tilde{a})}\left\langle \int Dz \sum_{l\neq 0} v_{\xi l}\left[\frac{1}{1+\sum_{n\neq l}\exp\left[\beta\left(\mathcal{H}_l^\xi-\mathcal{H}_n^\xi\right)\right]}\right]\right\rangle \\ &\to \frac{1}{a(1-\tilde{a})}\sum_{l\neq 0}\left\langle \int Dz v_{\xi l}\prod_{n\neq l}\Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right]\right\rangle \end{aligned} \tag{3.35}$$

$$q \to \tilde{q} = \frac{1}{a}\sum_{l\neq 0}\left\langle \int Dz \prod_{n\neq l}\Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right]\right\rangle \tag{3.36}$$

$$C = \frac{1}{\tilde{a}^2\sqrt{\alpha r}}\sum_{l\neq 0}\sum_k\left\langle \int Dz\sqrt{\frac{P_k}{S(1-\tilde{a})}}v_{kl}z_k\prod_{n\neq l}\Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right]\right\rangle \tag{3.37}$$

$$\tilde{r} \to r = \frac{q}{(1-\tilde{a}C)^2} \tag{3.38}$$

$$\beta(r-\tilde{r}) = 2\left(\tilde{U}\frac{S^2}{a\alpha} - \frac{C}{1-\tilde{a}C}\right) \tag{3.39}$$

where

$$\int Dz = \int dz \frac{\exp(-z^2/2)}{\sqrt{2\pi}}. \tag{3.40}$$

The Θ function gives non-vanishing contribution only for $\mathcal{H}_l^\xi - \mathcal{H}_n^\xi > 0$, i.e.

$$\sum_{k>0}(v_{kl}-v_{kn})z_k > -m\sqrt{\frac{S^2(1-\tilde{a})}{\alpha a r}}(v_{\xi l}-v_{\xi n}) - \frac{\alpha a\beta(r-\tilde{r})}{2S^2}\sqrt{\frac{S^2(1-\tilde{a})}{\alpha a r}}(\delta_{n0}-\delta_{l0}).$$

Moreover, it is convenient to introduce two combinations of order parameters,

$$\begin{aligned} x &= \frac{\alpha a\beta(r-\tilde{r})}{2S^2}\sqrt{\frac{S^2(1-\tilde{a})}{\alpha a r}}, \\ y &= m\sqrt{\frac{S^2(1-\tilde{a})}{\alpha a r}}. \end{aligned}$$

At the saddle point, they become

$$\begin{aligned} x &= \frac{1}{\sqrt{q}+\tilde{a}C\sqrt{r}}\sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left[\tilde{U} - \tilde{\alpha}\frac{C}{2}\sqrt{\frac{r}{q}}\right], \\ y &= \sqrt{\frac{1-\tilde{a}}{\tilde{\alpha}}}\left(\frac{m}{\sqrt{q}+\tilde{a}C\sqrt{r}}\right), \end{aligned} \tag{3.41}$$

where $\tilde{\alpha} = \alpha a/S^2$. By computing the averages in (3.35) and (3.39), we get three equations that close the self consistent loop with (3.41),

$$\begin{aligned} q &= \frac{1-a}{\tilde{a}}\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\phi(z)^{S-1} \\ &+ \int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\phi(z+y)^{S-1} \\ &+ (S-1)\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\phi(z-y)\phi(z)^{S-2}, \end{aligned} \tag{3.42}$$

$$m = \frac{1}{1-\tilde{a}}\int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\phi(z+y)^{S-1} - q\frac{\tilde{a}}{1-\tilde{a}}, \tag{3.43}$$

$$
\begin{aligned}
C\sqrt{r} &= \frac{1}{\sqrt{\tilde{\alpha}\,(1-\tilde{a})}}\Bigg\{\frac{1-a}{\tilde{a}}\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\,(z)^{S-1} \\
&+ \int Dp\int_{-y(1-\tilde{a})+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\,(z+y)^{S-1} \\
&+ (S-1)\int Dp\int_{y\tilde{a}+x-i\sqrt{\tilde{a}}p}^{\infty} Dz\left(z+i\sqrt{\tilde{a}}p\right)\phi\,(z-y)\,\phi\,(z)^{S-2}\Bigg\},
\end{aligned}
\tag{3.44}
$$

where $\phi(z) = \left(1+\mathrm{erf}(z/\sqrt{2})\right)/2$. It is insghtful to consider the limit cases of (3.41)-(3.44). One such limit case is $\tilde{a} \ll 1$ and the resulting self-consistent equations are

$$x = \frac{1}{\sqrt{\tilde{\alpha}q}}\left(\tilde{U} - \frac{\tilde{\alpha}C}{2}\sqrt{\frac{r}{2}}\right) \tag{3.45}$$

$$y = \frac{m}{\sqrt{\tilde{\alpha}q}} \tag{3.46}$$

$$m = \phi(y-x) \tag{3.47}$$

$$q = \frac{1-a}{\tilde{a}}\phi(-x) + \phi(y-x) \tag{3.48}$$

$$C\sqrt{r} = \frac{1}{2\pi\tilde{a}}\left\{\frac{1-a}{\tilde{a}}\exp\left(-x^2/2\right) + \exp\left(-(y-x)^2/2\right)\right\}. \tag{3.49}$$

Detailed derivation is in Appendix C.

3.2.2 Highly diluted network

A more biologically plausible case is that of the *diluted* network where the number of connections per unit c_m is less than N. Specifically, we consider connections of the form $c_{ij}J_{ij}$, where J_{ij} is the usual symmetric matrix derived from Hebbian learning. c_{ij} equals 0 or 1 according to a given probability distribution and we note $\lambda = \langle c_{ij}\rangle/N = c_m/N$ the dilution parameter. In general, c_{ij} is different from c_{ji}, leading to asymmetry in the connections between units. In this case, the capacity cannot be analyzed through the replica method. We therefore apply the signal to noise analysis. The local field of unit i in state k writes

$$h_i^k = \sum_j \sum_l c_{ij} J_{ij}^{kl} \sigma_j^l - \tilde{U}\left(1-\delta_{k,0}\right) \tag{3.50}$$

where the coupling strength between two states of two different units is defined as

$$J_{ij}^{kl} = \frac{1}{c_m a(1-\tilde{a})}\sum_{\mu} v_{\xi_i^\mu k} v_{\xi_j^\mu l}\,. \tag{3.51}$$

In the highly diluted limit $c_m \sim \log(N)$ (cp. next section for more details), the assumption is that the field can be written simply as the sum of two terms, signal and noise. While the signal is what pushes the activity of the unit such that the network configuration converges to an attractor, the noise, or the crosstalk from all of the other patterns, is what deflects the network away from the cued memory pattern. The noise term writes

$$n_i^k \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_{l} v_{\xi_i^\mu k} v_{\xi_j^\mu l}\, \sigma_j^l \,,$$

that is, the contribution to the weights J_{ij}^{kl} by all non-condensed patterns. By virtue of the subtraction of the mean activity in each state $\tilde{a}$, the noise has vanishing average:

$$\langle n_i^k \rangle_{P(\xi)} \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} \sum_{l} \langle v_{\xi_i^\mu,k} \rangle \langle v_{\xi_j^\mu,l} \sigma_j^l \rangle = 0 \,.$$

Now let us examine the variance of the noise. This can be written in the following way:

$$\langle (n_i^k)^2 \rangle \propto \sum_{\mu>1}^{p} \sum_{j(\neq i)=1}^{N} \sum_{l} \sum_{\mu'>1}^{p} \sum_{j'(\neq i)=1}^{N} \sum_{l'} \langle v_{\xi_i^\mu,k}\, v_{\xi_i^{\mu'},k} \rangle \langle v_{\xi_j^\mu,l}\, v_{\xi_{j'}^{\mu'},l'}\, \sigma_j^l \sigma_{j'}^{l'} \rangle \,,$$

where statistical independence between units has been used. For randomly correlated patterns, all terms but $\mu = \mu'$ vanish. Having identified the non-zero term, we can proceed with the capacity analysis. We can express the field using the overlap parameter, and single out, without loss of generality, the first pattern as the one to be retrieved

$$h_i^k = v_{\xi_i^1 k} m_i^1 + \sum_{\mu>1} v_{\xi_i^\mu k} m_i^\mu - \tilde{U}(1-\delta_{k0}). \tag{3.52}$$

where we define the local overlap m_i as

$$m_i = \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^1 l} \sigma_j^l \,. \tag{3.53}$$

We now write

$$\sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu \equiv \sum_{n=1}^{S} v_{n,k}\, \rho^n\, z_i^n \tag{3.54}$$

where ρ is a positive constant and z_i^n is a standard Gaussian variable. Indeed in highly diluted networks the l.h.s., i.e. the contribution to the field from all of the non-condensed patterns $\mu > 1$, is approximately a normally distributed random variable, as it is the sum of a large number of uncorrelated quantities. ρ can be computed to find

$$\rho^n = \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} \tag{3.55}$$

where we have defined

$$q = \left\langle \frac{1}{Na} \sum_j \sum_l (\sigma_j^l)^2 \right\rangle. \tag{3.56}$$

The mean field then writes

$$h_i^k = v_{\xi_i^1 k} m + \sum_{n=1}^{S} v_{n,k} \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} z_n - \tilde{U}(1-\delta_{k0}) \,. \tag{3.57}$$

Averaging m_i and q over the connectivity and the distribution of the Gaussian noise z, and taking the $\beta \to \infty$ we get to the mean field equations that characterize the fixed points of the dynamics, (3.35) and (3.36). In the highly diluted limit

however, we do not obtain the last equation of the fully connected replica analysis, (3.38).

The difference between fully connected and diluted cases must vanish in the $\tilde{a} \ll 1$ limit, as shown in ([20], [77]). In this limit we have $x = \tilde{U}/\sqrt{\tilde{\alpha}q}$, $y = m/\sqrt{\tilde{\alpha}q}$ while (3.43) and (3.42) remain identical.

3.2.3 Network with intermediate connectivity

As in the previous section, we can express the field using the overlap parameter, and single out the contribution from the pattern to be retrieved, that we label as $\mu = 1$, as in (3.52). However, for high enough connectivity one must revise (3.54): the mean field has to be computed in a more refined way, through a self-consistent method, that we present here.

Given the high connectivity of the network, the probability distribution of the c_{ij} plays a crucial role. We will consider three different distributions. The first is refered to as *random dilution* (RD), which is

$$P(c_{ij}, c_{ji}) = P(c_{ij})P(c_{ji}) \tag{3.58}$$

with

$$P(c_{ij}) = \lambda\delta(c_{ij} - 1) + (1 - \lambda)\delta(c_{ij})\,. \tag{3.59}$$

The second is the *symmetric dilution* (SD), defined by

$$P(c_{ij}, c_{ji}) = \lambda\delta(c_{ij} - 1)\delta(c_{ji} - 1) + (1 - \lambda)\delta(c_{ij})\delta(c_{ji})\,. \tag{3.60}$$

The third is what we call *state dependent random dilution* (SDRD)–specific to the Potts network– in which

$$P(c_{ij}^{kl}) = \lambda\delta(c_{ij}^{kl} - 1) + (1 - \lambda)\delta(c_{ij}^{kl})\,; \tag{3.61}$$

we note that in this case the connectivity coefficients are state-dependent.

We have performed simulations with all three types of connectivity, but will focus the analysis onto the *RD* type, which exhibits higher capacity as shown in Fig.3.7. *RD* and *SD* are known in the literature as Erdos-Renyi graphs. Many properties are known about such random graph models [78], [79]. It is known that for λ below a critical value, essentially all connected components of the graph are trees, while for λ above this critical value, loops are present. In particular, a graph with $c_m < \log(N)$ will almost surely contain isolated vertices and be disconnected, while with $c_m > \log(N)$ it will almost surely be connected. $\log(N)$ is a threshold for the connectedness of the graph, distinguishing the highly diluted limit, for which a simplified analysis of the storage capacity is possible, from the present intermediate case, for which a complete analysis is necessary.

When applying the self-consistent signal to noise analysis (SCSNA), [66, 80, 81] the noise term is assumed to be a sum of two terms

$$\sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu = \gamma_i^k \sigma_i^k + \sum_{n=1}^{S} v_{n,k}\, \rho_i^n z_i^n \tag{3.62}$$

where z_i^n are standard Gaussian variables, and γ_i^k and ρ_i^n are positive constants to be determined self-consistently. The first term, proportional to σ_i^k, represents the noise resulting from the activity of unit i on itself, after having reverberated in the loops of the network; the second term contains the noise which propagates from units other than i. The activation function writes

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{\sum_l e^{\beta h_i^l}} \equiv F^k\Big(\big\{y_i^l + \gamma_i^l \sigma_i^l\big\}_l\Big) . \tag{3.63}$$

where $y_i^l = v_{\xi_i^1,l} m_i^1 + \sum_n v_{n,l} \rho_i^n z_i^n - U(1-\delta_{l,0})$. One would need to find σ_i^k as

$$\sigma_i^k = G^k\Big(\big\{y_i^l\big\}_l\Big) , \tag{3.64}$$

where G^k are functions solving (3.63) for σ_i^k. However, (3.63) cannot be solved explicitly. Instead we make the assumption that $\{\sigma_i^l\}$ enters the fields $\{h_i^l\}$ only through their mean value $\langle\sigma_i^l\rangle$, so that we write

$$G^k\Big(\big\{y_i^l\big\}_l\Big) \simeq F^k\Big(\big\{y_i^l + \gamma_i^l \langle\sigma_i^l\rangle\big\}_l\Big) . \tag{3.65}$$

In Appendix D there are the details of the calculation that yield $\gamma_i^k = \gamma$ and $\rho_i^k = \rho^k$.

$$\gamma = \frac{\alpha}{S} \lambda \frac{\Omega/S}{1-\Omega/S} \tag{3.66}$$

where $\alpha = p/c_m$, $\langle\cdot\rangle$ indicates the average over all patterns and where we have defined

$$\Omega = \left\langle \frac{1}{N} \sum_{j_1} \sum_{l_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} \right\rangle . \tag{3.67}$$

From the variance of the noise term one reads

$$(\rho^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q\Big\{1 + 2\lambda\Psi + \lambda\Psi^2\Big\} , \tag{3.68}$$

where we have defined

$$q = \left\langle \frac{1}{Na} \sum_{j,l} (G_j^l)^2 \right\rangle \tag{3.69}$$

and

$$\Psi = \frac{\Omega/S}{1-\Omega/S} . \tag{3.70}$$

The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k} m + \frac{\alpha}{S} \lambda\Psi(1-\delta_{k,0}) + \sum_{n=1}^{S} v_{n,k} z^n \sqrt{\frac{\alpha P_n}{S(1-\tilde{a})} q\Big\{1 + 2\lambda\Psi + \lambda\Psi^2\Big\}} - \tilde{U}(1-\delta_{k,0}) . \tag{3.71}$$

Taking the average over the non-condensed patterns (the average over the Gaussian noise z), followed by the average over the condensed pattern $\mu = 1$ (denoted by $\langle\cdot\rangle_\xi$), in the limit $\beta \to \infty$, we get the self-consistent equations satisfied by the order parameters

$$m = \frac{1}{a(1-\tilde{a})}\left\langle \int D^S z \sum_{l(\neq 0)} v_{\xi,l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi , \tag{3.72}$$

$$q = \frac{1}{a}\left\langle \int D^S z \sum_{l(\neq 0)} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi , \tag{3.73}$$

$$\Omega = \left\langle \int D^S z \sum_{l(\neq 0)} \sum_k z^k \frac{\partial z^k}{\partial y^l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi . \tag{3.74}$$

where in the last equation for Ω, integration by parts has been used. Note the similarities to the equations ((3.35)-(3.37)) obtained through the replica method for the fully connected case. The equations just found constitute a generalization to $\lambda < 1$. In particular, in the highly diluted limit $\lambda \to 0$, we get $\gamma \to 0$ and $(\rho^n)^2 \to \frac{\alpha P_n}{(1-\tilde{a})S} q$, which are the results obtained in the previous section; in the fully connected case, $\lambda = 1$, the correspondence between the m and q variables is obvious, while for Ω it can be shown with some algebraic manipulation. Indeed, from the following identity,

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q(1+\Psi)^2 , \tag{3.75}$$

by using the replica variable $r = q/(1-\tilde{a}C)^2$ we get

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} r(1-\tilde{a}C)^2(1+\Psi)^2 . \tag{3.76}$$

By comparing this with (3.34), the mean field, we get an equivalent expression for Ψ,

$$\Psi = \frac{\tilde{a}C}{1-\tilde{a}C} . \tag{3.77}$$

From the original definition of Ψ in (D.9), it follows that the order parameter C, obtained through the replica method, is equivalent to Ω, up to a multiplicative constant:

$$C = \Omega/a . \tag{3.78}$$

We can show that (3.74) coincides with (3.37). Moreover, by comparing the SCSNA result for γ to the replica one, we must have

$$\frac{\alpha}{S}\Psi - \tilde{U} = -\frac{\alpha a \beta (r - \tilde{r})}{2S^2} \tag{3.79}$$

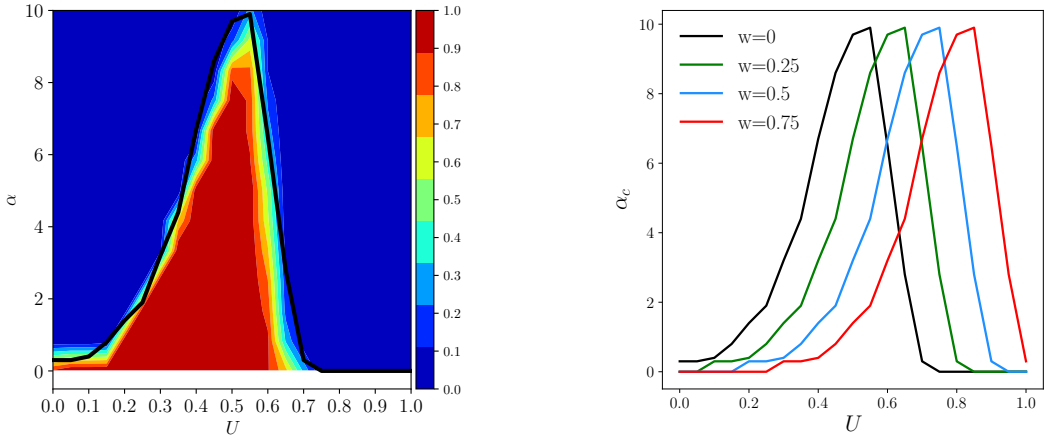
from which

$$\beta(r-\tilde{r}) = 2\left(\tilde{U}\frac{S^2}{\alpha a} - \frac{C}{1-\tilde{a}C}\right) , \tag{3.80}$$

identical to (3.39).

3.3 Simulation results

Do computer simulations confirm the analyses above? Starting with the effect of setting the overall threshold, we show, in Fig.3.4a, retrieval performance as a function of the threshold, both through simulations and by solving (3.41).



(a) (b)

Figure 3.4: (a) How often a fully connected Potts network retrieves memories, as a function of the threshold U and the number of stored memories p, with $N = 1000$, $S = 7$, $a = 0.25$, $\beta = 200$. Color represents the fraction of simulations in which the overlap between the activity state of network and a stored pattern is ≥ 0.9. The solid lines are obtained by numerical solution of (3.41)-(3.44). (b) The dependence of α_c on U for different values of w. While for the optimal threshold U a non-zero value of w is detrimental to the capacity, for higher than optimal thresholds it can lead to a lower effective threshold $\tilde{U}$, enhancing capacity.

It is clear that the simulations agree very well with numerical results. The maximum storage capacity α_c (where $\alpha \equiv p/c_m$, or $\alpha \equiv p/N$ for a fully connected Potts network) is found at approximately $U = 0.5$, as can also be shown through a simple signal to noise analysis. It is possible to compute approximately the standard deviation γ_i^k of the field, (3.52) with respect to the distribution of all the patterns, as well as as the connectivity c_{ij}, by making the assumption that all units are aligned with a specific pattern to be retrieved $\sigma_j^l = \xi_j^1$. We further discriminate units that are in active states $\xi_i^1 \neq 0$ from those that are in the quiescent states $\xi_i^1 = 0$ in the retrieved pattern $\mu = 1$.

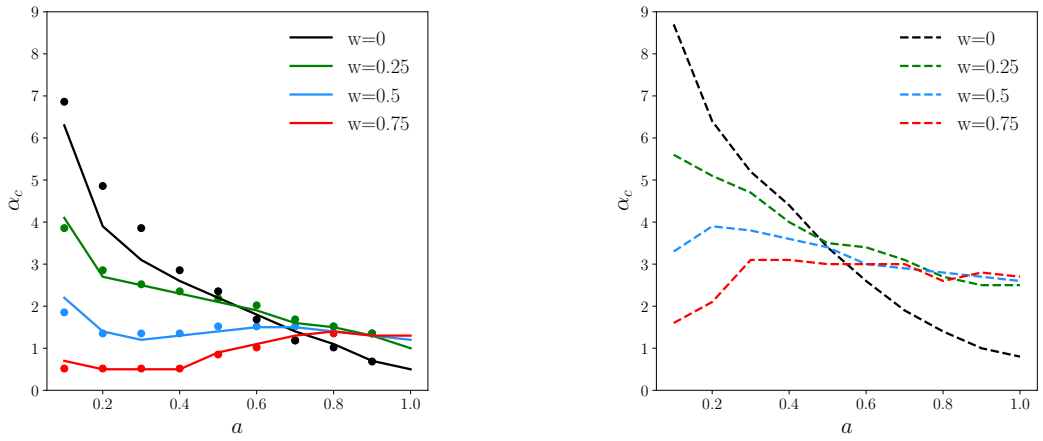
$$\gamma_i^k \equiv \sqrt{\langle (h_i^k)^2 \rangle - \langle h_i^k \rangle^2} = \sqrt{\frac{(p-1)a}{c_m S^2} + (\delta_{\xi_i^1,k} - \tilde{a})^2 \left(\frac{1}{c_m a} - \frac{1}{N} \right)}. \tag{3.81}$$

The optimal threshold U_0 is one that separates the two distributions, optimally such that the minimal number of units in either distribution reach the threshold to go in the wrong state

$$\frac{U_0 - \langle h_i^k |_{\xi_i^1=0} \rangle}{\gamma_i^k |_{\xi_i^1=0}} = - \frac{U_0 - \langle h_i^k |_{\xi_i^1 \neq 0} \rangle}{\gamma_i^k |_{\xi_i^1 \neq 0}}$$

$$U_0 = \frac{\gamma_i^k|_{\xi_i^1=0}}{\gamma_i^k|_{\xi_i^1=0} + \gamma_i^k|_{\xi_i^1\neq 0}} - \frac{a}{S}. \tag{3.82}$$

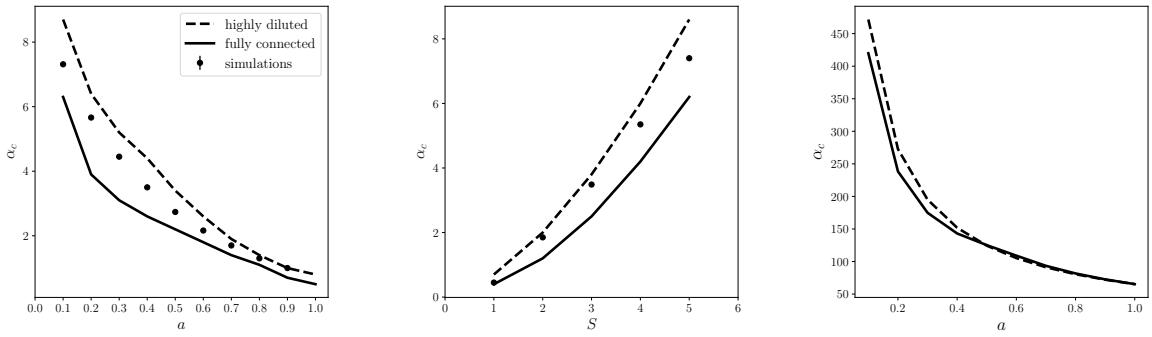
We can see that $U_0 \longrightarrow 1/2-\tilde{a}$ for $\gamma_i^k|_{\xi_i^1=0} \sim \gamma_i^k|_{\xi_i^1\neq 0}$, consistent with the replica analysis and simulations in Fig.3.4a. Given such an optimal value for U, Fig.3.4b shows that the effect of the feedback term w on the storage capacity, being purely subtractive, is just to shift to the right the optimal value.



(a) Fully connected network (b) Diluted network

Figure 3.5: Storage capacity α_c as a function of sparsity a for different values of w for both fully connected (a) and diluted (b) networks as obtained by numerical solution of (3.41)-(3.44). (a) also includes points from simulations. The parameters are $S = 5$, $U = 0.5$, $\beta = 200$.

Fig.3.5 illustrates the same effect of the feedback term, by setting $U = 0.5$ and charting the storage capacity as a function of the sparsity a for different values of w, for both fully connected (a) and diluted networks (b). In both cases, α_c decreases monotonically with increasing w, for low a, when $U = 0.5$ is close to optimal. Increasing a, one reaches a region where $U = 0.5$ is set too high, and therefore α_c benefits from a non-zero w, even though its exact value is not critical. For very high sparsity parameter (non-sparse coding) all curves except $w = 0$ seem to coalesce. The envelope of the different curves represents optimal threshold setting that takes feedback into account, and as a function of a it shows, both for fully connected and diluted networks the decreasing trend familiar from the analysis of simpler memory networks [82].



(a) (b) (c)

Figure 3.6: (a) Storage capacity α_c as a function of the sparsity a. Dots correspond to simulations of a network with $N = 2000$, $c_m/N = 0.1$, $S = 5$, and $\beta = 200$ while curves are obtained by numerical solution of (3.41)-(3.44). (b) Storage capacity as a function of S with same parameters as in (a) and with $a = 0.1$. (c) $S = 50$, illustrating the $\tilde{a} \ll 1$ limit case.

The two connectivity limit cases are illustrated in Fig.3.6, which shows, in (a), the dependence of the storage capacity α on the sparsity a in the fully connected and diluted networks with $U = 0.5$, $w = 0$ and $S = 5$. In Fig.3.6b instead, S is varied and in Fig.3.6c $S = 50$, corresponding to the highly sparse limit $\tilde{a} \ll 1$. While for $S = 5$ the two curves are distinct, for the highly sparse network with $S = 50$ the two curves coalesce. The curves are obtained by numerically solving (3.41)-(3.44). Moreover, the storage capacity curve for the fully connected case in (a) matches very well with Fig.2 of [20]. Diluted curves are always above the fully connected ones in both (a) and (b), as found in [20].

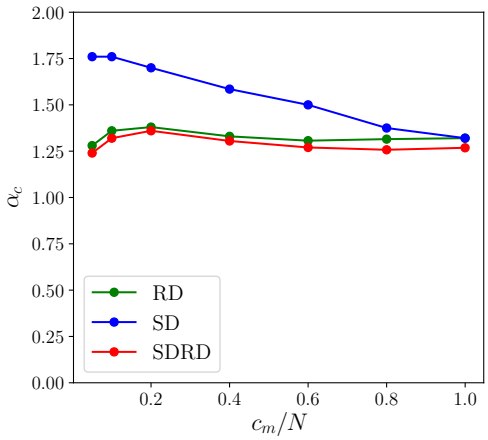


Figure 3.7: Storage capacity curves, obtained through simulations, as a function of the mean connectivity per unit c_m/N for three different types of connectivity matrices c_{ij}. Network parameters are $S = 2$, $a = 0.1$, $U = 0.5$ and $\beta = 200$.

Finally, we show in Fig.3.7 the little change in storage capacity across the connec-

tivity models introduced earlier. Contrary to the Hopfield network, SD has higher capacity than RD. Both RD and SDRD on the other hand seem to have almost identical capacity. All models have the same capacity at the fully connected case, as they should. Note in particular the very limited decrease of $\alpha_c = p/c_m$ with c_m/N increasing up to almost full connectivity, with all three models, in contrast with what one observes in the Hopfield network. This is because coding is relatively sparse, at $a = 0.1$, and made effectively even sparser by $S = 2$, so that $\tilde{a} = 0.05$.

In the end, the storage capacity of the Potts network is primarily a function of a few parameters, c_m, S and a, that suffice to broadly characterize the model, with minor adjustments due to other factors. How can these parameters be considered to reflect cortically relevant quantities?

The Potts network, if there are N_m Potts variables, requires, in the fully connected case, $N_m \cdot (N_m - 1) \cdot S^2/2$ connection variables (since weights are taken to be symmetric we have to divide by 2). In the diluted case, we would have $N_m \cdot c_m \cdot S^2$ variables (the factor 2 is no longer relevant, at least for $c_m \to 0$). The multi-modular Hopfield network, as shown in Sect.3.1, has only $N_m \cdot N_u \cdot L$ long-range synaptic weights. This diluted connectivity between modules is summarily represented in the Potts network by the tensorial weights. Therefore, the number of Potts weights cannot be larger than the total number of underlying synaptic weights it represents. Then $c_m \cdot S^2$ cannot be larger than $L \cdot N_u$.

In the simple Braitenberg model of mammalian cortical connectivity [26], which motivated the multi-modular network model [17], $N_u \simeq N_m \sim 10^3 - 10^5$, as the total number of of pyramidal cells ranges from $\sim 10^6$ in a small mammalian brain to $\sim 10^{10}$ in a large one. In a large, e.g. human cortex, a module may be taken to correspond to roughly $1\,\mathrm{mm}^2$ of cortical surface, also estimated to include $N_u \sim 10^5$ pyramidal cells [10]. A module, however, cannot be plausibly considered to be fully connected; one can imagine instead a short-distance connection probability of the order of $1/10$, and a number of short-range connection similar to the one, L, of long-range ones, yielding $L \simeq 0.1 N_u$.

What about c_m and S? What values would be compatible with associative storage? If there are S patterns on N_u neurons, there would only be $S \cdot N_u \cdot a_u$ variables available in order to determine local synaptic weights. It is reasonable then to take $S \cdot N_u \cdot a_u > L \cdot N_u$, but in turn we have $L \cdot N_u > c_m \cdot S^2$, hence

$$c_m \cdot S < N_u \cdot a_u$$

which would lead, if we take again $a_u \sim 0.1$, to c_m and S to be at most of order $10^1 - 10^2$ over mammalian cortices of different scale, essentially scaling like the fourth root of the total number of pyramidal cells, which appears like a plausible, if rough, modelling assumption. We could take these range of values, together with the approximate formula (see [20] and Fig.3.6b)

$$p_c \sim 0.15 \frac{c_m S^2}{a \ln(S/a)} \tag{3.83}$$

to yield estimates of the actual capacity the cortex of a given species. The major factor that such estimates do not take into account, however, is the correlation

among the memory patterns. All the analyses reported here apply to randomly assigned memory patterns. The case of correlations will be treated elsewhere in [57].

The above considerations may sound rather vague. They capture, however, the quantitative change of perspective afforded by the coarse graining inherent in the Potts model. We can simplify the argument by neglecting sparse coding as well as the exact value of the numerical pre-factor k (which is around 0.15 in (3.83)). The Potts model uses $N_m c_m S^2$ weights to store up to $k c_m S^2 / \ln S$ memory patterns, each containing of order $N_m \ln S$ bits of information, therefore storing up to k bits per weight. In this respect, it is not different from any other associative memory network, including the multi-modular model which it effectively represents. In the multi-modular model, however, (in its simplest version) the $2kN_u^2 N_m$ bits available are allocated to memory patterns that are specified in single-neuron detail, and hence contain in principle $N_u N_m$ bits of information each. The network can store and retrieve up to a number p_c of them, which has been argued in [16] to be limited by the *memory glass* problem to be of the same order of magnitude as the number of local attractors, itself limited to be of order N_u. By losing the single-neuron resolution, the Potts model forfeits the locally extensive character of the information contained in each pattern, but it gains essentially a factor $S/\ln S$ (scaling approximately as $\sqrt{N_u}$) in the number of patterns. Many more, but less informative, memories. This argument can be expanded and made more precise by considering, again, a more plausible scenario with correlated memories.

Chapter 4

Latching dynamics

In this chapter, we address the question of when robust latching occurs, as a model of spontaneous sequence generation, with extensive computer simulations, mostly focused on latching between randomly correlated patterns. We consider first the slowly adapting regime ($\tau_1 \ll \tau_2 \ll \tau_3$) in which active states (τ_2) adapt slower than activity propagation to other units (τ_1), while inhibitory feedback is restricted to an even slower timescale, τ_3. Next we contrast with it the fast adapting regime ($\tau_3 \ll \tau_1 \ll \tau_2$) in which, instead, inhibitory feedback is immediate, relative to the other two time scales.

The critical parameters at play are the number of patterns, p, the number of active states, S, and the number of connections per unit, C, and we also look at the effect of the feedback term w. The other parameters, including T, τ_1, τ_2, and τ_3, are kept fixed during simulations, after having chosen *a priori* values that can lead to robust latching dynamics in the two regimes.

4.1 Slowly adapting regime

In the slowly adapting regime, over a (short) time of order τ_1 the network, if suitably cued, may reach one of the global attractors, and stay there for a while; whereupon, after an adaptation time of order τ_2, it may latch to another attractor, or else activity may die [5]. But how distinct is the convergence to the new attractor? One may assess this as the difference between the two highest overlaps the network activity has, at time t, with any of the memory patterns, $m_1(t) - m_2(t)$: ideally, $m_1 \simeq 1$ and m_2 is small, so their difference approaches unity. A summary measure of memory pattern discrimination can be defined as $d_{12} \equiv \langle \int dt(m_1(t) - m_2(t)) \rangle_{\text{initial cue}}$, where of course the identity of patterns 1 and 2 changes over the sequence.

As discussed in [5], by looking at the latching length, how long a simulation runs before, if ever, the network falls into the global quiescent state, one can distinguish several 'phases'. Depending on the parameters, the dynamics exhibit finite or infinite latching behaviour, or no latching at all. Typically, when increasing the storage load p the latching sequence is prolonged and eventually extends indefinitely, but at the same time its distinctiveness decreases, since memory patterns cannot be individually retrieved beyond the storage capacity; and even before, each acquires neighbouring patterns, in the finite and more crowded pattern space, with which it

is too correlated to be well discriminated.

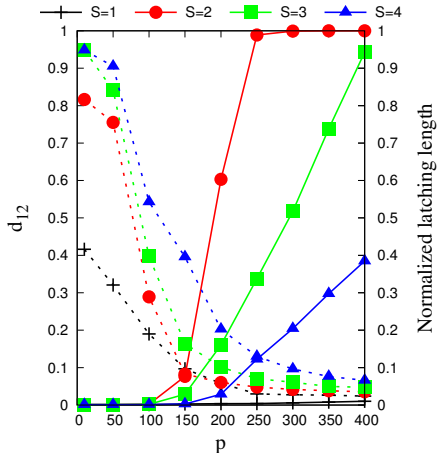


Figure 4.1: Trade-off between latching sequence length (solid lines) and retrieval discrimination (dashed lines). Different colors indicate different S values, while $C = 400$ throughout. The latching length l is in time steps (not in the number of transitions), normalized by the time of the simulation, $N_{update} = 6 \cdot 10^5$.

In Fig.4.1 we see that for each $S = (2, 3, 4)$, as p is increased beyond a certain value, latching dynamics rapidly picks up and extends eventually through the whole simulation, but in parallel its discriminative ability decreases and almost vanishes – the p-range where d_{12} is large is in fact when there is no latching, and d_{12} only measures the quality of the initial cued retrieval. For $S = 1$ no significant latching sequence is seen, whereas for higher values, at fixed p, its distinctiveness increases with S but its length decreases from the peak value at $S = 2$.

Since the latching length l is not itself sufficient to characterize latching and has to be complemented by discriminative ability, we find it convenient to quantify the overall *quality* of latching with a new quantity Q defined as

$$Q = d_{12} \cdot l \cdot \eta, \tag{4.1}$$

where η is introduced to exclude cases in which the network gets stuck in the initial cued pattern, so that no latching occurs, however high d_{12} and l are:

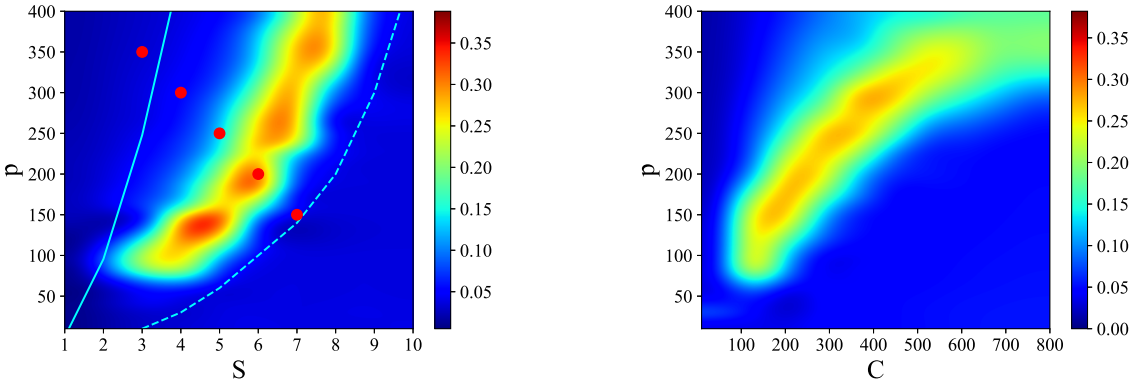
$$\eta = \begin{cases} 1: & \text{if at least one transition to a second memory pattern occurs} \\ 0: & \text{otherwise.} \end{cases} \tag{4.2}$$

Q is therefore a positive real number between 0 and 1, and we report its color-coded value to delineate the relevant phases in phase space.

Thus, *low quality* latching with small Q may result from either small d_{12} or short l, or both. The parameters that determine Q which we focus on are S, C and p, after having suitably chosen all the other parameters, which are kept fixed. Their default values in the slowly adapting regime are $N = 1000$, $a = 0.25$, $U = 0.1$, $T = 0.09$, $w = 0.8$, $\tau_1 = 3.3$, $\tau_2 = 100.0$, $\tau_3 = 10^6$, unless explicitly noted otherwise. If activity does not die out before, simulations are terminated after $N_{update} = 6 \cdot 10^5$ steps, the total number of updates of the entire Potts network, and are repeated with different

cued patterns. Re the values of S, C and p, we use the following notation, for simplicity:

$$Q = Q(S,C,p) = \begin{cases} Q(S,p): & C = 150 \quad \text{fixed}, \\ Q(C,p): & S = 5 \quad\;\; \text{fixed}. \end{cases} \tag{4.3}$$



(a) (b)

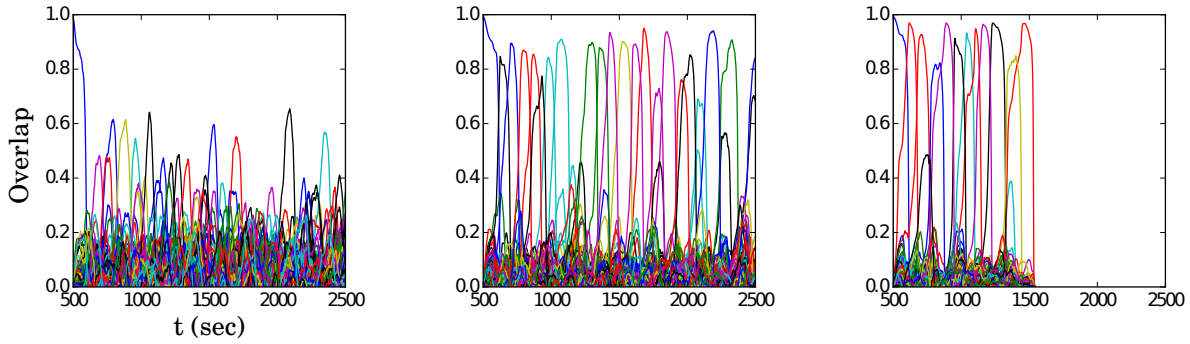
Figure 4.2: Phase space for $Q(S,p)$ in (a) and $Q(C,p)$ in (b) with randomly correlated patterns in the slowly adapting regime. The parameters are $C = 150$ and $S = 5$, if kept fixed, and $w = 0.8$. The red spots in (a) mark the parameter values used in the following analyses.

Fig.4.2 shows that there are narrow regions in the $S-p$ and $C-p$ planes, which we call *bands*, where relatively *high quality* latching occurs. The values of p with the 'best' latching scale almost quadratically in S, and sublinearly in C. Moreover, one notices that below certain values of S and C no latching is seen, i.e. the band effectively ends at $S \sim 2$, $p \sim 90$ in Fig.4.2a and at $C \sim 50$, $p \sim 70$ in Fig.4.2b. Importantly, the band in Fig.4.2a is confined in the area delimited by the cyan solid and dashed curves above and below it. The dashed curve is for the onset of latching, i.e. the phase transition to finite latching [5], while the solid curve above is the storage capacity curve in a diluted network, given by the approximate relation

$$p_c \simeq \frac{CS^2}{4a \ln \frac{2S}{a\sqrt{\ln \frac{S}{a}}}}, \tag{4.4}$$

beyond which retrieval fails [5]. It should also be noted that overall Q values are not large, in fact well below 0.5 throughout both $S-p$ and $C-p$ planes. The reason is, again, in the conflicting requirements of persistent latching, favoured by dense storage, high p, and good retrieval, allowed instead only at low storage loads (in practice, relatively low p/S^2 and p/C values).

In Fig.4.3 we show representative latching dynamics at three selected points in the (S,p) plane, in terms of the time evolution of the overlap of the states with the stored activity patterns (see (2.8)). The three points, marked in red, span across



(a) (5,250) (b) (6,200) (c) (7,150)

Figure 4.3: Latching behaviour for (S, p) equal to, respectively, (5,250), (6,200), and (7,150) in Fig.4.2a.

the band in Fig.4.2a, and we see that latching is indefinite but noisy in the example at (5,250), which is apparently too close to storage capacity, while memory retrieval is good at (7,150) but the sequence of states ends abruptly, as the network is in the phase of finite latching [5]. The two trends are representative of the two sides of the band, while in the middle, at (6,200), one finds a reasonable trade-off, with relatively good retrieval combined with protracted latching.

We use two statistical measures, the *asymmetry* of the transition probability matrix and Shannon's *information entropy* [22, 55, 83] to characterize the essential features of the dynamics in different parameter regions. For that, we take all five red points from Fig.4.2a, such that they cut across the latching band in the $S-p$ plane, and extend further upwards. We first compile a transition probability (or rather, frequency) matrix M from all distinct transitions observed along many latching sequences generated with the *same set* of stored patterns, as in [55]. The dimension of the matrix M is $(p+1)\times(p+1)$, as it includes all possible transitions between p patterns *plus* the global quiescent state. M is constructed from the transitions between states having both overlaps above a given threshold value, e.g. 0.5, in a data set of $1,000$ latching sequences, by accumulating their frequency between any two patterns into each element of the matrix and then normalizing to 1 row by row, so that $M_{\mu,\nu}$ reflects the probability of a transition from pattern μ to ν. A, the degree of asymmetry of M, is defined as

$$A = \frac{||M - M^T||}{||M||}, \tag{4.5}$$

where M^T is the transpose matrix of M and $||M|| = \sum_{\mu,\nu}|M_{\mu,\nu}|$. Note that A is small for unconstrained bi-directional dynamics and large for simpler stereotyped flows among global patterns, attaining its maximum value $A = 2$ for strictly uni-directional transitions. Note also that if the average had been taken over *different* realizations of the memory patterns, given sufficient statistics A would obviously vanish.

Another measure we apply to the transition matrix M is Shannon's information entropy, defined as

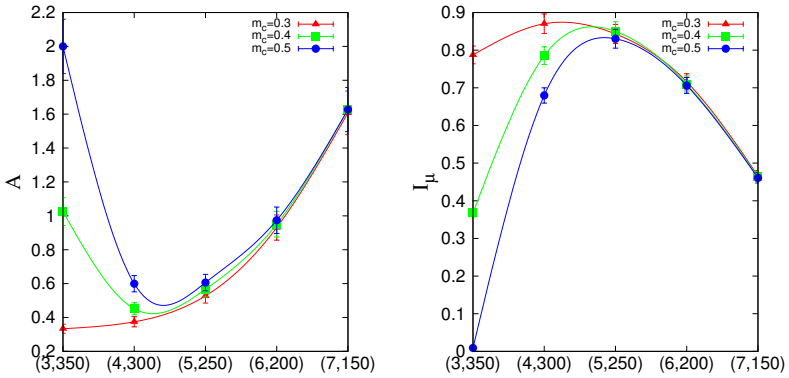
$$I_\mu = \left\langle \frac{1}{\log_2(p+1)} \sum_{\nu=1}^{p+1} M_{\mu,\nu} \log_2 \left(\frac{1}{M_{\mu,\nu}} \right) \right\rangle_\mu . \tag{4.6}$$

I_μ takes positive real values from 0 (deterministic, all transitions from one state are to a single other state) to 1 (completely random), since it is normalized by $\log_2(p+1)$, which corresponds to a completely random case.

We use these two measures, A and I_μ, on the points, marked red in Fig.4.2a

$$(3,350) - (4,300) - (5,250) - (6,200) - (7,150)$$

that lie on a segment going through the latching band observed in the slowly adapting regime.



(a) (b)

Figure 4.4: (a) Asymmetry A of the transition matrix and (b) Shannon's information entropy, I_μ along the (3,350)–(4,300)–(5,250)–(6,200)–(7,150) parameter series from Fig.4.2. Different curves correspond to different thresholds for the overlap of the two states between which the network is defined to have a transition. The error bars report the standard deviation of either quantity for each of 1,000 sequences.

If we focus on transitions between states reaching at least a threshold overlap of 0.5, Fig.4.4 appears to show two complementary, almost opposite U-shaped curves as the two measures, asymmetry and entropy, are applied to the 5 points along the segment. One branch of each U shape extends over the range that includes the high-Q latching band: these are the right branches of the two curves, in which asymmetry decreases from a large value $A \simeq 1.6$ at (7,150) to a smaller one $A \simeq 0.6$ at (5,250), while concurrently the entropy increases from $I_\mu < 0.5$ at (7,150) to $I_\mu > 0.8$ at (5,250). As Fig.4.3 indicates, at (7,150) latching sequences are distinct but very short, and few entries are filled in the transition matrix: generally either $M_{\mu\nu} = 0$ or $M_{\nu\mu} = 0$, so that asymmetry is high and entropy relatively low. This holds irrespective of the number of sequences that are averaged over. The opposite happens at (5,250), where many transitions are observed, and in filling the transition matrix they approach the random limit. The point with the highest Q-value, (6,200), is characterized by intermediate values of asymmetry and entropy which, we have previously observed, may be seen as a signature of complex dynamics [55].

Extending the range upwards, it seems as if the asymmetry, with threshold 0.5, were to eventually increase again, reaching its maximum $A = 2$ at (3,350), with a decreasing entropy, vanishing at the same point (3,350). These left branches are, however, dependent on the threshold values used, as Fig.4.4 shows, and do not imply that transitions become more deterministic, because in this region there are simply fewer and fewer distinct transitions, discernible above the noise (Fig.4.3). The left branches merely reflect the increasing arbitrariness with which one can identify significant correlations with memory states in the rambling dynamics observed at higher storage loads.

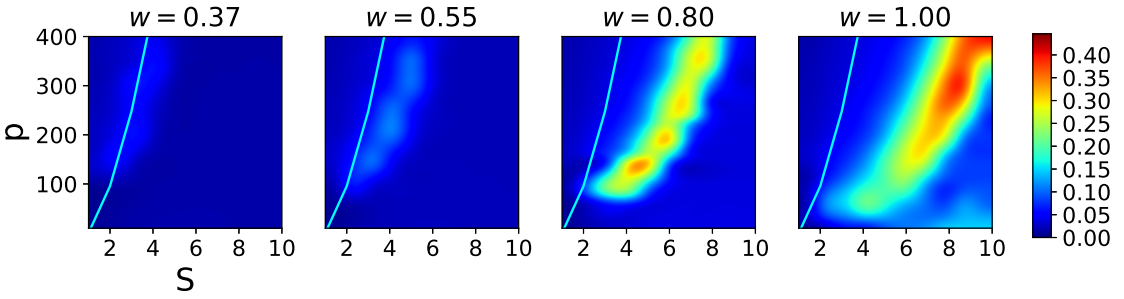
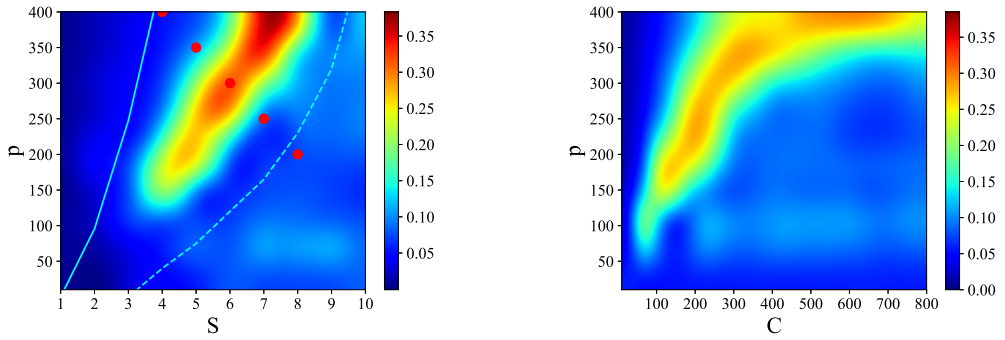


Figure 4.5: Latching quality $Q(S, p)$ with increasing local feedback, w =0,37, 0.55, 0.8, and 1.0 in the slowly adapting regime. Randomly correlated patterns are used, with $C = 150$ as in Fig.4.2a.

In Fig.4.5 we see that the effect of the local feedback term, w, is first to enable latching sequences of reasonable quality, and then to also shift the latching band to higher values of S, effectively pushing this behaviour away from the storage capacity curve representing the retrieval capability of the Potts associative network. Hence, if one were to regard S as a structural parameter of the network, and w as a parameter that can be tuned, there is an optimal range of w values which allows good quality latching for higher storage. This argument has to be revised, however, by considering also the threshold U, since increasing w can be shown to be functionally equivalent, in terms of storage capacity, to decreasing U [24]. Also for U, in fact, one can find an optimal range for associative retrieval to occur, in the simple Potts network with no adaptation and with $w = 0$ [20]. This near equivalence between U and $-w$ does not hold anymore in the fast adapting regime, to which we turn next.

4.2 Fast adapting regime

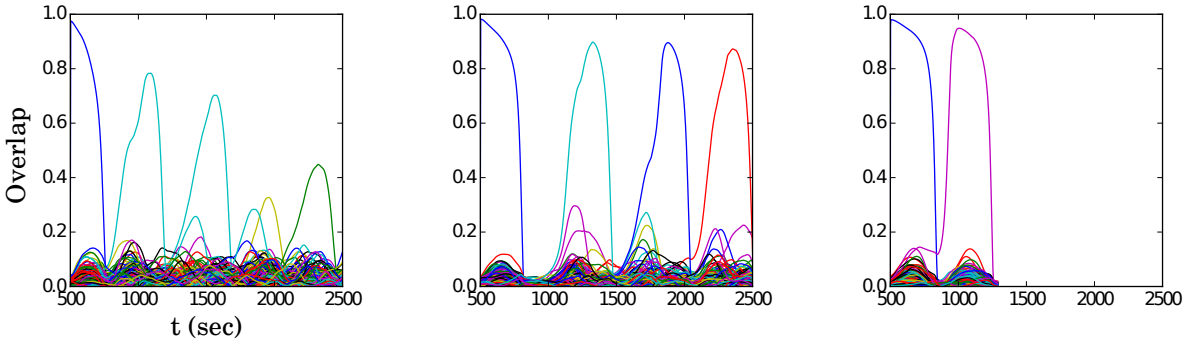
We characterize the fast adapting regime by the alternative ordering of time scales $\tau_3 < \tau_1 \ll \tau_2$, such that the mean activity in each Potts unit is rapidly regulated by fast inhibition, at the time scale τ_3. (2.6) stipulates that $\sum_{k=1}^{S} \sigma_i^k(t)$, the total activity of each unit, is followed almost immediately, or more precisely at speed τ_3^{-1}, by the generic threshold $\theta_i^0(t)$. Extensive simulations, with the same parameters as for the slowly adapting regime, except for $w = 1.37$, $\tau_1 = 20$, $\tau_2 = 200$ and $\tau_3 = 10$, show that, similarly to the slowly adapting regime, there are latching bands in the $Q(S, p)$ and $Q(C, p)$ planes, see Fig.4.6. With these parameters, in particular the larger value chosen for the feedback term w, the bands occupy a similar position as



(a) (b)

Figure 4.6: Phase space for $Q(S,p)$ in (a) and $Q(C,p)$ in (b) with randomly correlated patterns in the fast adapting regime. The parameters are identical to those in the slowly adapting regime, with the exception of $w = 1.37$, $\tau_1 = 20$, $\tau_2 = 200$, $\tau_3 = 10$. The red spots in (a) mark, again, the parameter values used in the Figures below.

in the slowly adapting regime. Again, they appear to vanish below certain values of S and C, more precisely around $S \sim 3$, $p \sim 120$ in Fig.4.6a and around $C \sim 50$, $p \sim 90$ in Fig.4.6b, and to scale subquadratically in S and sublinearly in C. The band in the $S-p$ plane is again confined by the storage capacity (solid cyan curve) and by the onset of (finite) latching (dashed curve). The storage capacity curve, which is independent of threshold adaptation, follows the same (4.4).



(a) (5,350) (b) (6,300) (c) (7,250)

Figure 4.7: Latching behaviour for (S,p) equal to, respectively, (5,350), (6,300), and (7,250) in Fig.4.6a.

Examples of latching behaviour outside and inside the band are presented in Fig.4.7, at the same values for S but shifted by $\Delta p = 100$, i.e. at the "red" points (5,350), (6,300), and (7,250) in the $S-p$ plane. Again, we see from Fig.4.6a that (5,350) lies just above the band, while (6,300) is right on the centre. To the right of the band, e.g. at (7,250), the transitions are distinct but latching dies out very soon, while on the left, e.g. at (5,350), the progressively reduced overlaps are a

manifestation of increasingly noisy retrieval dynamics. In all three examples we observe that latching steps proceed slowly, even slower than the doubled time scale $\tau_2 = 200$ would have led to predict. This appears to be because often a significant time elapses between the decay of the overlap of the network with one pattern and the emergence of a new one.

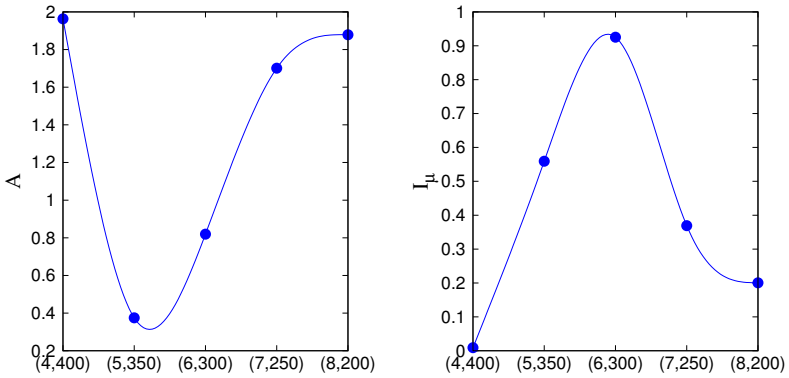


Figure 4.8: (a) Asymmetry A of the transition matrix and (b) Shannon's information entropy, I_μ along the (4,400)–(5,350)–(6,300)–(7,250)–(8,200) parameter series from Fig.4.6a, using only a threshold 0.5 for the overlaps before and after each transition.

Fig.4.8 shows the asymmetry and entropy measures, A and I_μ, along the points

$$(4,400) - (5,350) - (6,300) - (7,250) - (8,200),$$

in Fig.4.6a, where, again, we have chosen a series shifted by $\Delta p = 100$ upwards in order to centre it better on the high quality latching band. Only an overlap threshold of 0.5 is considered. What one can see, in contrast with the slowly adapting regime, is that now the two measures are not quite complementary. The point $(6,300)$ that lies inside the band, very much at its quality peak, shows again an intermediate value for the asymmetry, but the highest value, given the overlap threshold, for the entropy. The discrepancy may be ascribed to the different prevailing type of latching transition observed in the fast adapting regime, Fig.4.7. As discussed in [21], in a Potts network latching transitions with a high cross-over, which can only occur between memory patterns with a certain degree of correlation, can be distinguished from those with a vanishing cross-over, which are much more random. In the fast adapting regime, as indicated by the examples in Fig.4.7, all transitions tend to be of the latter type. A more careful analysis indicates, in fact, that they are quasi-random, in that they avoid a memory pattern in which largely the same Potts units are active as in the preceding pattern. In fact, the value of the entropy at (6,300) implies that on average from each of the 300 memory patterns there are transitions to at least 190 other patterns (to 190 if they were equiprobable, in practice many more); therefore only the few patterns which happen to be more (spatially) correlated are avoided.

Towards the left, the curves do not vary much depending on the threshold chosen

for the overlaps, but the asymmetry eventually becomes maximal and the entropy vanishes simply because sequences of robustly retrieved patterns do not last long, so in this particular case it would take more than 1000 sequences to accumulate sufficient statistics.

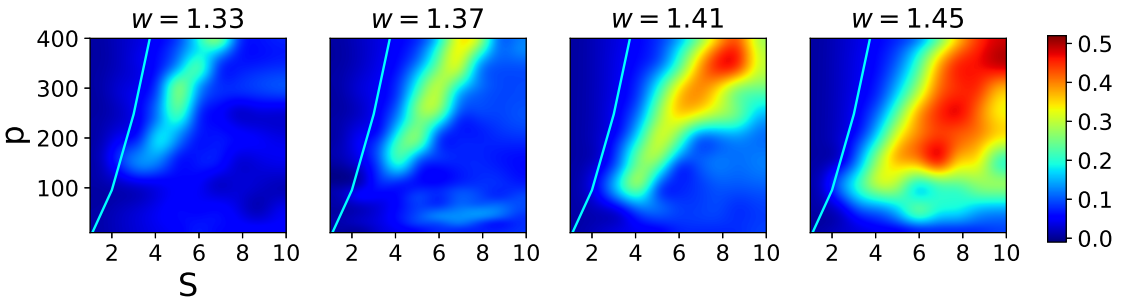


Figure 4.9: Latching quality $Q(S,p)$ with increasing local feedback, w =1.33, 1.37, 1.41, and 1.45 in the fast adapting regime. Randomly correlated patterns are used, with C = 150 as in Fig.4.6a.

The effects of increasing the w term in the fast adapting regime are shown in Fig.4.9, where one notices two main features. First, there is heightened sensitivity to the exact value of w, so that relatively close data points at w =1.33, 1.37, 1.41, and 1.45 yield rather different pictures. Second, although again increasing w shifts the latching band rightward, by far the main effect is a widening of the band itself. This is because in the presence of rapid feedback inhibition a larger w term ceases to be functionally similar to a lower threshold, which in the slowly adapting regime was leading in turn to noisier dynamics and eventually indiscernible transitions. In the fast adapting regime, the increased positive feedback can be rapidly compensated by inhibitory feedback, so that in the high-storage region overlaps remain large, until they are suppressed by storage capacity constraints (the cyan curve, which remains at approximately the same distance from the larger and larger latching band).

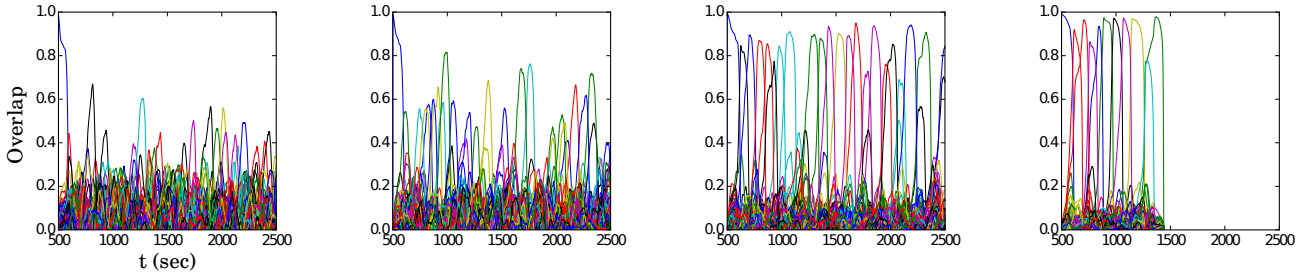
We now turn to more explicit comparison of the transition dynamics in two regimes.

4.3 Comparison of the two regimes

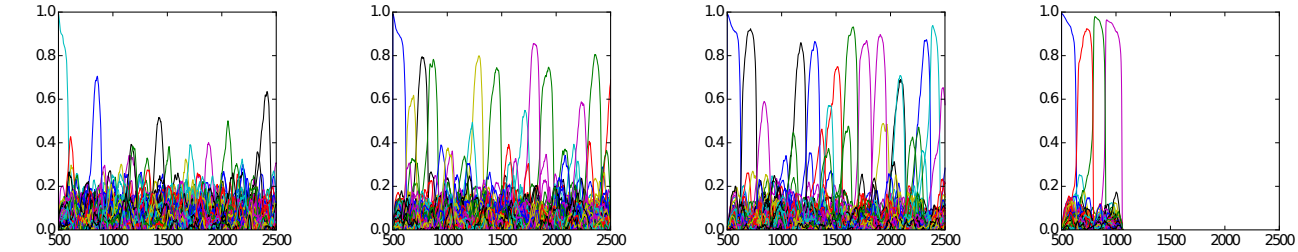
To look more closely at latching dynamics in the slowly and fast adapting regimes, we take the following points from Figs.4.2a, 4.6a, which allow us to cut through the bands at two different storage levels

$$\begin{cases} p = 200 & S = (4,5,6,7) \\ p = 400 & S = (6,7,8,9) \end{cases} \tag{4.7}$$

Fig.4.10 shows in different colors the overlaps of the state of the network with the global patterns, for sample sequences along the points (4.7), in the slowly adapting regime. For both p = 200 and 400, latching length is observed to decrease with S, unlike the discrimination between patterns, as measured by d_{12}, in agreement with Fig.4.1. Note that the two rows in the figure are similar, indicating that the shift Δp = 200 is approximately compensated by the rightward shift ΔS = 2.

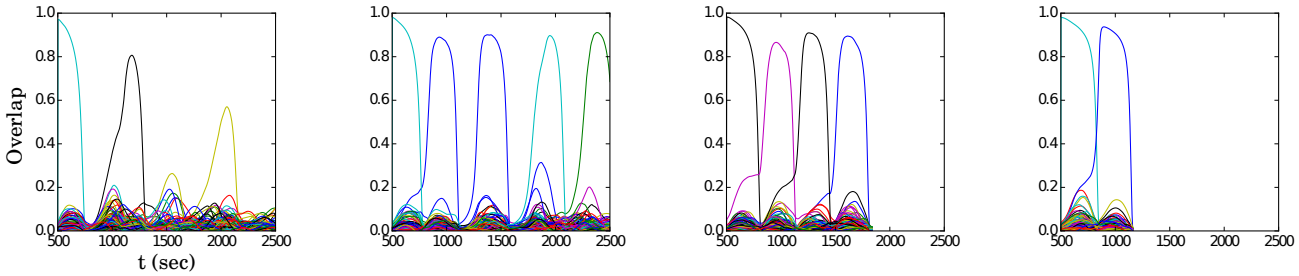


(a) p=200, S=4 (b) p=200, S=5 (c) p=200, S=6 (d) p=200, S=7

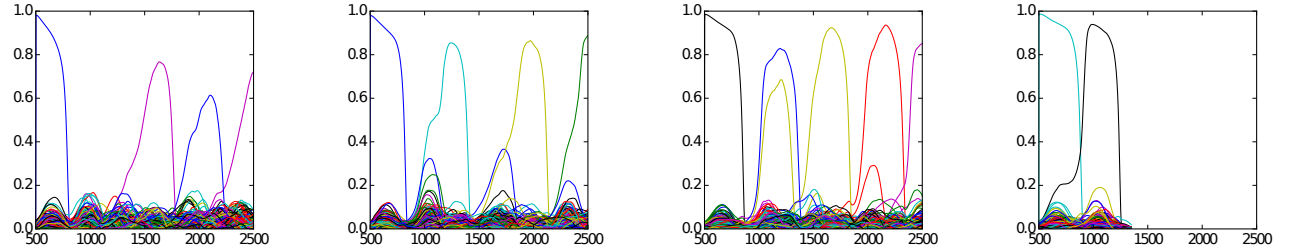


(e) p=400, S=6 (f) p=400, S=7 (g) p=400, S=8 (h) p=400, S=9

Figure 4.10: Latching behaviour in the slowly adapting regime. A sample of points (4.7) from Fig.4.2a.



(a) p=200, S=4 (b) p=200, S=5 (c) p=200, S=6 (d) p=200, S=7

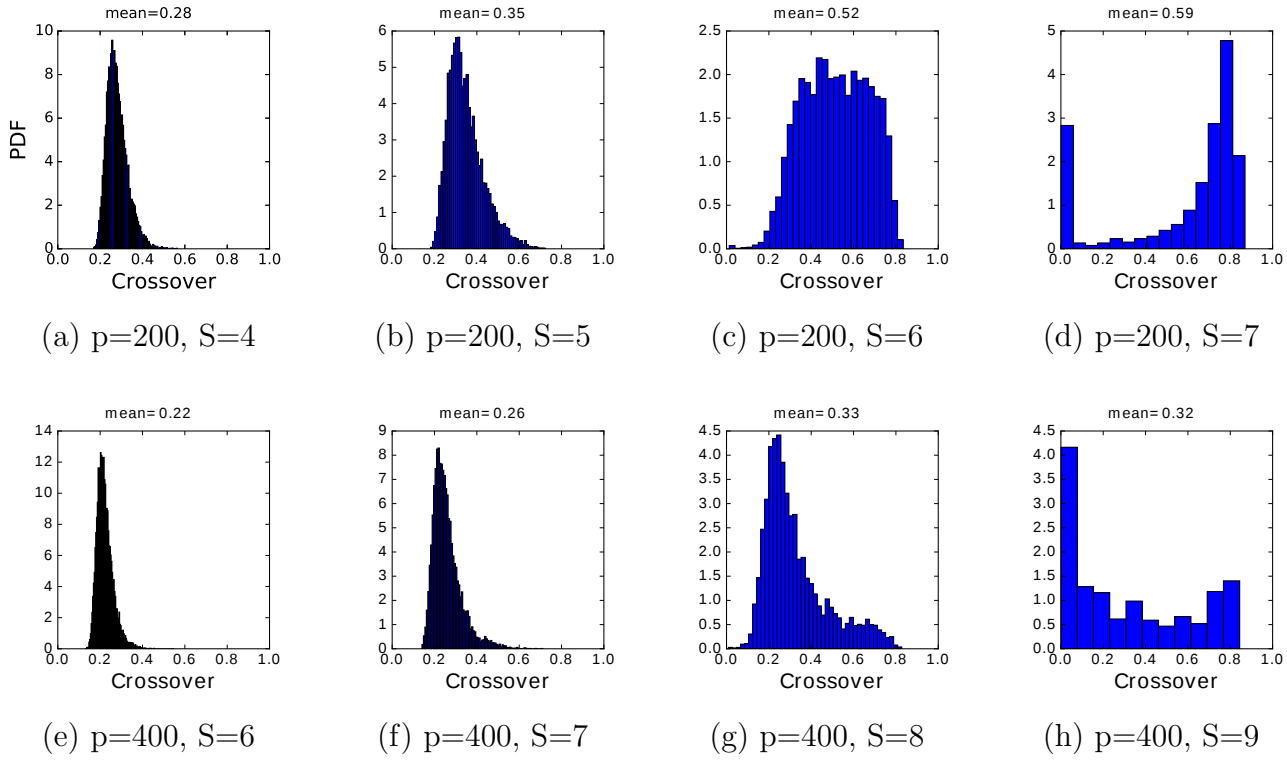


(e) p=400, S=6 (f) p=400, S=7 (g) p=400, S=8 (h) p=400, S=9

Figure 4.11: Latching behaviour in the fast adapting regime. A sample of points (4.7) from Fig.4.6a.

The fast adapting regime shows the same trends, again one sees in Fig.4.11 the approximate compensation between the two shifts $\Delta p = 200$ and $\Delta S = 2$, but latching appears in general less noisy.

The main difference between the two regimes, however, is in the distribution of crossover values, those when the network has equal overlap with the preceding and the following pattern: their distribution (PDF, or probability density function) is shown in Figs.4.12 and 4.13.



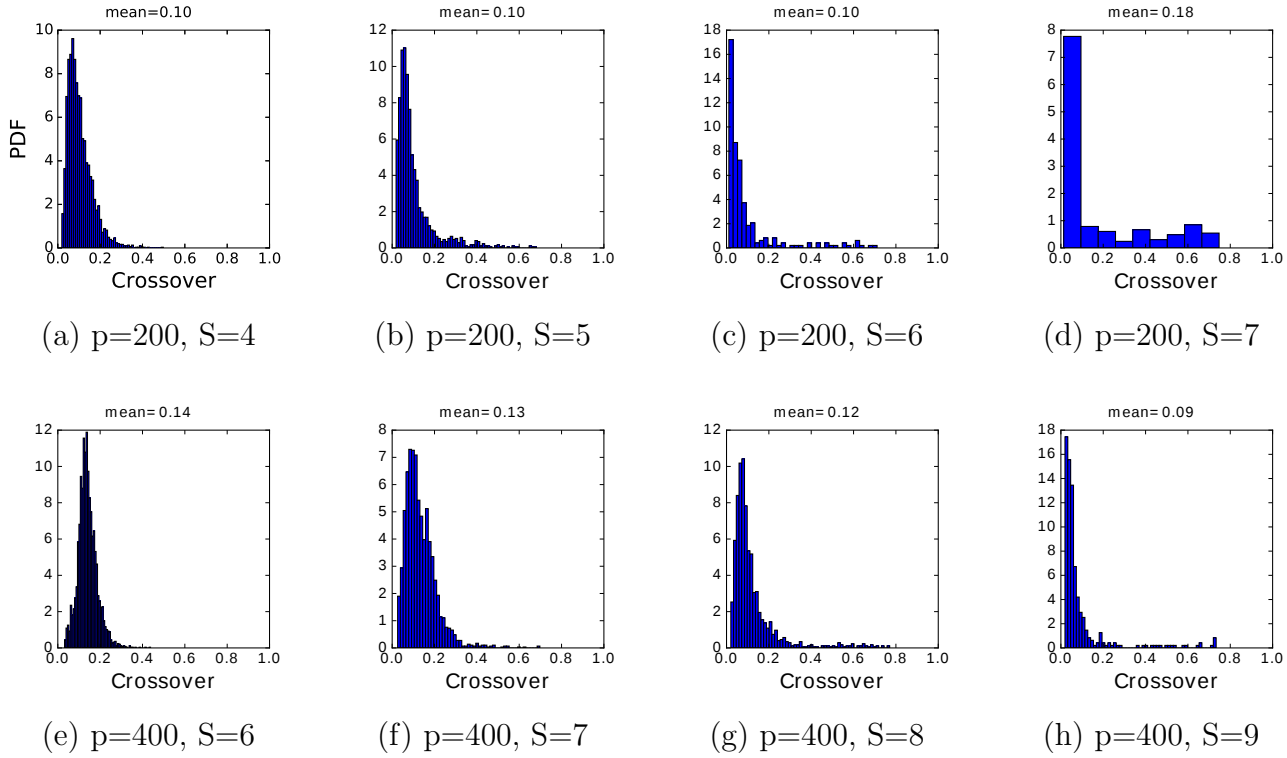
(a) p=200, S=4 (b) p=200, S=5 (c) p=200, S=6 (d) p=200, S=7

(e) p=400, S=6 (f) p=400, S=7 (g) p=400, S=8 (h) p=400, S=9

Figure 4.12: Probability density function (PDF) of crossover values in the slowly adapting regime.

We see that in the fast adapting regime most transitions occur at very low crossover, i.e. the correlation with the preceding memory has to decay almost to zero before the next memory pattern can be activated. Only in regions of the (S, p) plane where latching sequences are very short, a few transitions only, we begin to see a small fraction of them with crossover values above 0.2. In most cases, the inhibitory feedback conveyed by the variable θ_i^0 is so fast as not to allow transitions to be carried through by positive correlations, i.e. by the subset of Potts units which are in the same active state in the preceding and successive pattern. The choice of the next pattern is not completely random, as indicated by the relative entropy values still below unity, but is determined essentially by negative selection, as mentioned above: the next pattern tends to have few active Potts units that coincide with those active in the preceding pattern.

In the slowly adapting regime, instead, due to the slow variation of the non-specific threshold, active Potts units can remain active, but they are *encouraged* by the variables θ_i^k to switch between active states if they have been in the same for too



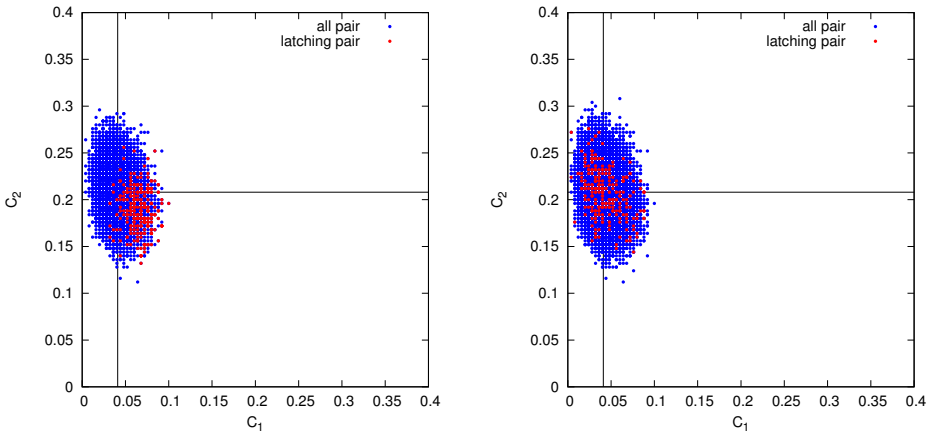
(a) p=200, S=4 (b) p=200, S=5 (c) p=200, S=6 (d) p=200, S=7

(e) p=400, S=6 (f) p=400, S=7 (g) p=400, S=8 (h) p=400, S=9

Figure 4.13: Probability density function (PDF) of crossover values in the fast adapting regime.

long. This can produce, particularly in the center of the latching band, sequences of patterns succeeding each other at high crossover, as shown by the distribution in Fig.4.12c. Even when latching is very noisy and approaches randomness, as in panels Fig.4.12a,e, crossover values are consistently above 0.2, indicating a preference for patterns insisting on the same set of active Potts units, unlike the fast adapting regime. Finally, when the number of states S is too large or, equivalently, that of patterns p too low, we observe some transitions with minimal crossover and a majority with very large crossover, as if occurring only with those patterns that were already partially retrieved when the network had still the largest overlap with the preceding pattern; but the main observation is that there are very few transitions at all, so that to plot a probability density distribution we need to use wide bins, in panels Fig.4.12d,h (and in Fig.4.13d).

This difference between the two regimes is confirmed by an analysis of the correlations between successive patterns in latching sequences. In the Potts network, at least two types of spatial correlation between patterns are relevant: how many active Potts units the two patterns share, and how many of these units are active and in the same state. We quantify them with C_1, the fraction of the units active in one pattern that are active also in the other, *and* in the same state; and with C_2, the fraction that are also active, but in a different state. In a large set of randomly determined patterns, the mean values are $\langle C_1 \rangle = a/S$ and $\langle C_2 \rangle = a(S-1)/S$. The full distribution, among all pairs, is scattered around these mean values. But do transitions occur between any pair of patterns?



(a) Slowly adapting regime (b) Fast adapting regime

Figure 4.14: Scatterplots of the fractions C_1 and C_2 of Potts units active in one pattern that are active also in another, and in the same state or, respectively, in another active state. The panels show the full distribution between any pattern pair, in the slowly (a) and fast adapting (b) regimes, in blue; and the distribution between successive patterns in latching transitions, in red. The blue distribution for the fast adapting regime (for which $a = 0.25$, $S = 6$, $p = 300$ and $w = 1.32$) is similar to the one for the slowly adapting regime (for which again $a = 0.25$, $S = 6$, but $p = 200$ and $w = 0.65$), except that it is slightly wider, because of the higher storage load; while the red distributions are markedly different. Vertical lines indicate mean values.

Fig.4.14 shows that relative to the full distribution, in blue, transitions tend to occur, in the slowly adapting regime on the left, only between patterns with C_1 *above* and C_2 *below* (or at most around) their average values. Thus when the network has retrieved a memory representation it looks for correlated ones, as it were, where to jump. In the fast adapting regime this is not the case: transitions are almost random, except there appears to be a slight tendency to avoid those with C_1 well above its mean value. Note that the values of p and w are different in the two panels, and are chosen so as to be in roughly equivalent positions within the respective latching bands.

The analysis of the crossover points, therefore, affords insight into the rather different transition dynamics prevailing in the fast and slowly adapting regimes, in particular in the center of their latching bands; suggesting that in a more realistic cortical model, which combines both types of activity regulation, there should still be a significant component of 'slow adaptation' for interesting sequences of correlated patterns to emerge. The preceding simulations, however, were all carried out with randomly correlated patterns, in which the occasional high or low correlation of a pair is merely the result of a statistical fluctuation. Does the insight carry over to a more stuctured model of the correlations among memory patterns? This is what we ask next.

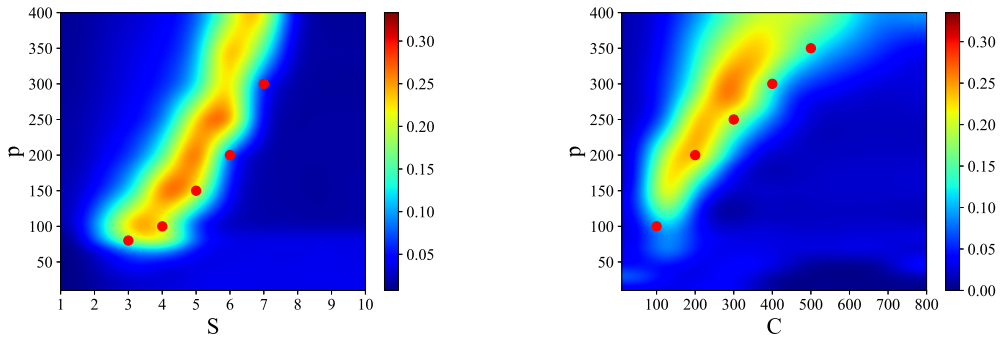
4.4 Analysis with correlated patterns

Correlated patterns were generated according to the algorithm mentioned by [8] and discussed in detail in [57]. The multi-parent pattern generation algorithm works in three stages. In the first step, a total set of Π random patterns are generated to act as parents. In the second step, each of the total set of parents are assigned to p_{par} randomly chosen children. Then a "child" pattern is generated: each pattern, receiving the influence of its parents with a probability a_p, aligns itself, unit by unit, in the direction of the largest field. In the third and final step, a fraction a of the units with the highest fields is set to become active. In this way, child patterns with a sparsity a are generated. In addition, another parameter ζ can be defined, according to which the field received by a child pattern is weighted with a factor $\exp(-\zeta k)$ where the index k runs through all parents. This is meant to express a non-homogeneous input from parents.

It is clear that such patterns however, cannot be considered as independent and identically distributed, as in (2.9), because their activity is drawn from a common pool of parents. In fact, they are correlated, in the sense that those children receiving congruent input from a larger number of common parents will tend to be more similar. All of these observations are studied in more detail in [57], and here we only focus on how correlations affect the phase diagrams. In the following simulations, the parameters pertaining to the patterns are $a_p = 0.4$, $\Pi = 100$, $\zeta = 0.1$ while p_{par}/p, the probability that a pattern be influenced by a parent is kept constant to 0.277.

Simulations with correlated patterns were carried out across the same $S-p$ and $C-p$ planes in phase space, in the slowly adapting regime, as shown in Fig.4.15. We focused on the slowly adapting regime based on the results of the crossover anal-

ysis. All other simulation parameters were kept at the values used with randomly correlated patterns.



(a) (b)

Figure 4.15: Phase space, cut across the $Q(S, p)$ plane in (a) and $Q(C, p)$ in (b), with correlated patterns in the slowly adapting regime. Red dots represent the quality peaks in the the same planes, with randomly correlated patterns. The parameters are $C = 150$ and $S = 5$, if kept fixed, and $w = 0.8$.

We see from the figure that the presence of non-random correlations among the memory patterns, albeit weak, shifts the bands to the left and upward in phase space, keeping approximately the dependence of the viable storage load p on S and C, but at somewhat higher values. It is as if more memories could 'fit', if correlated, into the same latching dynamics.

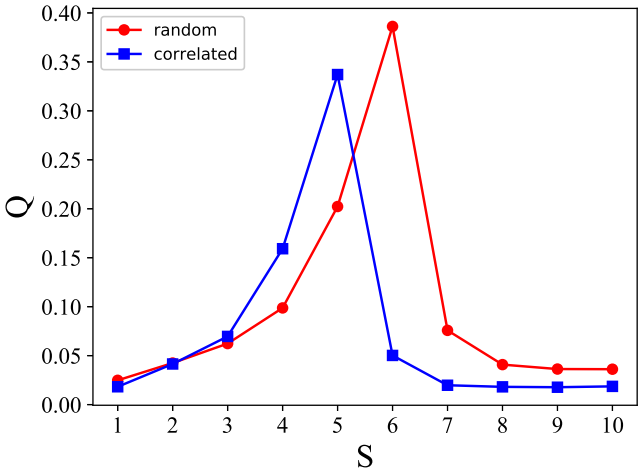


Figure 4.16: Comparison of $S - p$ phase spaces along $p = 200$ with random (red dotted line) and correlated (blue dotted) patterns in the slow adapting regime.

Fig.4.16 shows the $S - p$ plane cut along $p = 200$, to better compare the cases with correlated (blue) and random (red) patterns. It is apparent that there is a

leftward shift, in the case of correlated patterns, from the red curve applying to the random case, but the dependence on S remains very similar.

Chapter 5

Trying to give latching instructions

In the previous chapter, we have studied spontaneous latching dynamics. Although the analysis was largely limited to the case of randomly correlated memory patterns, we have seen in Fig.4.14 that, particularly along the critical line where infinite latching arises and is of good quality, in the slowly adapting regime, it is the (random) positive fluctuations in the degree of correlation between memory patterns that determine which transitions occur. One may ask to what extent one may superimpose on such spontaneous dynamics explicit *instructions*, that is, a list of transitions that the network is instructed or encouraged to go through, by encoding them in the connection weights. Following [39], in this chapter we introduce therefore a *hetero*-associative additional component to the previously purely *auto*-associative Hebbian weights, and we focus on how effectively are these instructed transitions, encoded in the learning rule, followed during latching, and on how much these instructions alter the spontaneous behaviour of the network, i.e., the one determined by the correlational structure of its memories.

5.1 Associative learning rule

The tensor connection between unit i in state k and unit j in state l is generalized by adding a *hetero*-associative component of strength λ, to

$$J_{ij}^{kl} = \frac{c_{ij}}{Ca(1-a/S)} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu,k} - \frac{a}{S} \right) \left\{ \left(\delta_{\xi_j^\mu,l} - \frac{a}{S} \right) + \lambda \sum_{d=1}^{D} \left(\delta_{\xi_j^{\nu_d},l} - \frac{a}{S} \right) \right\} (1-\delta_{k0})(1-\delta_{l0}), \tag{5.1}$$

where the λ term effectively guides or instructs each pattern in the direction of D other patterns. At each stage in the latching sequence, the network may follow one of the D instructions, or proceed of its own to a different transition (or the sequence may stop)

We provide the network with a table in which each memory pattern ($\{\xi^\mu|_{\mu=1,2,\cdots,p}\}$) is associated with its own set of D instructed patterns ($\{\xi^\nu|_{\nu=1,2,\cdots,D}\}$) that are selected randomly among the p that are stored auto-associatively.

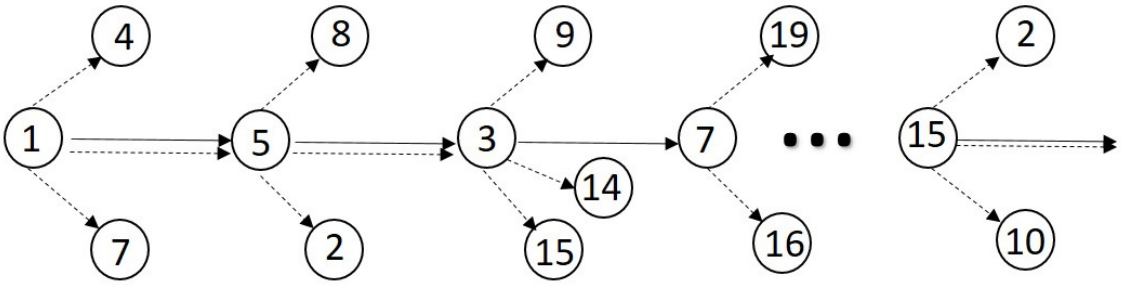


Figure 5.1: An example of latching sequence (1-5-3-7-···-15-···) and the corresponding instructions ((4, 5, 7) to 1, (2, 3, 8) to 5, (9, 14, 15) to 3, ···). Instructed transitions are denoted by dashed lines, while solid lines denote those, instructed or not, occurring in the latching sequence.

An example of latching partially governed by instructions is shown in Fig.5.1. A latching sequence 1-5-3-7-···-15-··· is indicated by solid lines while the patterns *hetero*-associated to each pattern in the sequence are denoted by dashed lines. In the example, patterns (4,5,7) are associated to pattern 1 and latching proceeds towards pattern 5. For patterns 1, 5, 7, 15, dynamics flows along the associated patterns. Only for pattern 3 spontaneous latching occurs.

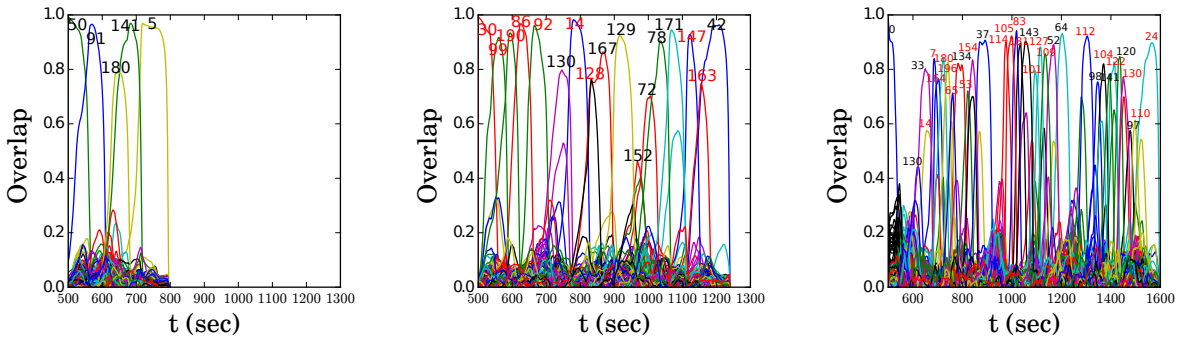
It should be noted that the unambiguous identification of which patterns occur at distinct stages in the sequence is only possible when latching is of sufficient quality. If not, it may for example happen that an instructed transition does occur, but is *masked* by a spontaneous transition to a different pattern, occurring simultaneously and with slightly larger overlap. It is therefore appropriate to first assess the quality of latching dynamics, in the presence of instructions.

5.2 The effect of *hetero*-associative instructions on latching dynamics

As mentioned in previous chapters, the character of latching dynamics in a Potts network may be quantified in terms of several different measures, including the latching length, l, the difference between the two highest overlaps, d_{12}, and the crossover in overlap between two successive patterns, m_{cross} [5, 21–23]. The quality of latching, Q, combines the first two of these measures to give a visual impression of where robust latching occurs in phase space.

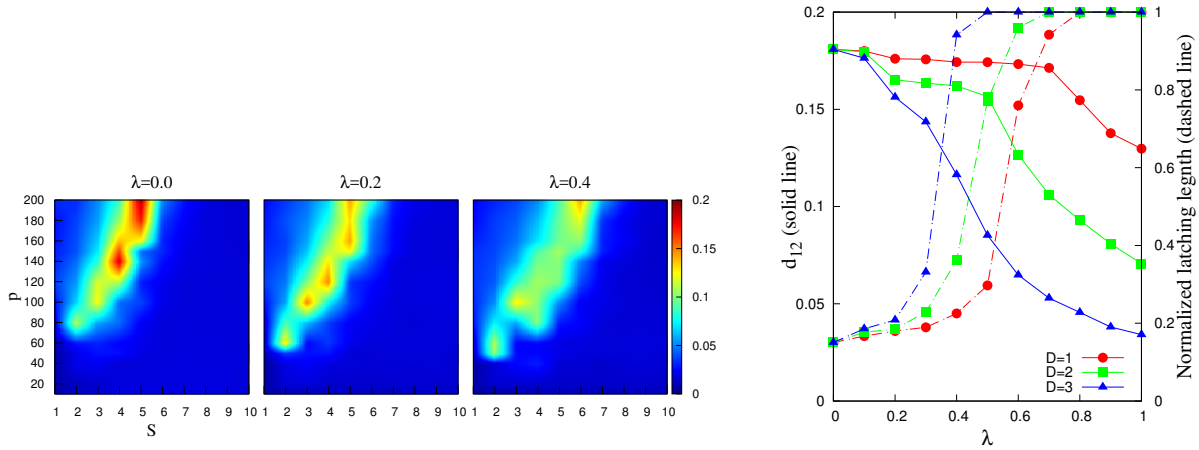
We first address the question of what is the effect on latching behavior, in terms of the above quantities, when *hetero*-associations supplement the original *auto*-associative learning rule, with relative strength λ.

We keep the parameters $N = 600$, $C = 90$, $p = 200$, $S = 7$, $a = 0.25$, $U = 0.1$, $\beta = 12.5$, $w = 0.45$, $\tau_1 = 3.3$, $\tau_2 = 100.0$ and $\tau_3 = 10^6$, corresponding to the slowly adapting regime, throughout the chapter. Simulations are terminated after $6 \cdot 10^5$ updates and repeated over 1000 cued patterns. To see the influence of instructions on latching, we focus on the $D = 2$ case, where each pattern is *hetero*-associated with two other patterns at the learning stage.



(a) $\lambda = 0.0$ (b) $\lambda = 0.1$ (c) $\lambda = 0.3$

Figure 5.2: Retrieval dynamics with $\lambda = 0$, 0.1 and 0.3 in (a), (b) and (c). Numbers indicate the patterns with the highest overlap that compose the retrieved sequence, and those in red denote instructed patterns. In these examples, $D = 2$.



(a) (b)

Figure 5.3: (a) Phase space of $Q(S, p)$ with hetero coupling strength $\lambda = 0.0$, 0.2 and 0.4. $D = 2$; (b) λ dependence of d_{12} (solid line) and l (dashed line), with $S = 7$ and $p = 200$. Red, green and blue stand for $D = 1, 2, 3$.

We show examples of latching behavior with and without the λ term in Fig.5.2a,b,c. Fig.5.2b shows that adding a small *hetero*-associative component to the connection weights has the main effect of lengthening the latching sequence, which however also becomes less distinct. In the next panel, with larger λ, retrieval quality has deteriorated much further, and one begins to notice collective instabilities, or bursts of nearly simultaneously retrieved patterns, that stand in marked contrast to the relatively clean sequence of the purely spontaneous latching in Fig.5.2a.

Numbers on top of the largest overlaps comprise the sequence, and those in red indicate instructed transitions. Even before a quantitative analysis, the panel suggest that tripling the λ value does not succeed in eliminating spontaneous transitions. In fact, we show below that the opposite is the case.

The bright regions, or *bands*, where relatively high-Q latching occurs are shown, for different values of λ (0.0, 0.2 and 0.4) in Fig.5.3a. The number of *hetero*-associative instructions at each stage is still $D = 2$. As already discussed in the previous chapter and [23], the area right to the band, with relatively large S and small p, shows good *quality* retrieval, measured by relatively high d_{12} (and m_{cross}), but short latching length. Instead, in the area left to the band, with relatively small S and large p, latching extends indefinitely but is very noisy, and the values of d_{12} and m_{cross} become very low. The band decreases gradually in peak values as λ grows, as illustrated in Fig.5.3a.

To afford a closer look at phase space, a point (S=7, p=200) is chosen and the values of d_{12} (solid line) and l (dashed line) are shown as a function of λ for $D = 1, 2, 3$ (red, green, blue) in Fig.5.3b. For all three D values retrieval quality as measured by d_{12} gradually deteriorates, while latching duration rapidly reaches the length of the simulations, with increasing λ. As a function of D, the λ value offering the best compromise between d_{12} and l shifts to the left with more instructed options, indicating that in the large D limit only very gentle instructions (small λ values) can be effective.

We can interpret these observations in following way. For a given value of D, a strong *hetero*-associative coupling λ enhances the network tendency to latch, resulting in prolonged sequences, but it also disrupts *auto*-associative retrieval, making the process noisier, and d_{12} lower. These effects are amplified for larger D values, since it becomes easier to latch in one of many instructed directions, but noise is also larger and it becomes difficult to retrieve any clean pattern. As a result, λ and D produce similar effects, in the sense that they both degrade latching quality while increasing latching length.

5.3 Instructed versus spontaneous latching transitions

How often does the Potts network follow the instructions it is given?

To measure the fraction of transitions that comply with the instructions given at the learning stage, we introduce a quantity f as

$$f = \frac{T_{instruct}}{T_{tot}}, \tag{5.2}$$

where $T_{instruct}$ is the number of transitions, i.e. pairs of successively retrieved patterns, with overlap above 0.5, that follow the instructions, and T_{tot} is the total number of pairs of successive patterns in the latching sequence. f is 1 if the network completely follows the instructions it is given, and 0 if it never does.

For convenience, we introduce some abbreviations; FP denotes a pair of patterns that follows the instructions, SP a spontaneous transition, LP a generic latching pair, spontaneous or instructed, and AP any possible pair, whether occurring in a latching sequence or not.

Simulations are run with the same parameters as in the previous section.

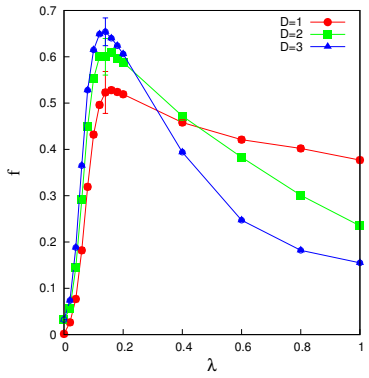


Figure 5.4: λ dependence of f for $D = 1, 2, 3$ (red, green, blue).

The fraction of FPs, f, is shown for $D = 1, 2, 3$ (red, green, blue) in Fig.5.4. From the figure we see that the network initially follows the instructions to an extent proportional to λ, but it quickly reaches a maximum degree of compliance, around $\lambda = 0.15$, at values $f > 0.5$ (which increase mildly with D). For larger values of λ, the compliance f drops, al the more rapidly the larger is D.

This may be because the network is effectively accompanied towards an instructed pattern only with a gentle fillip, i.e., at small enough λ, whereas larger values of λ push the network with a shove that perhaps drives it to the instructed pattern, but then often also past it and onward to an immediate further transition, that steps beyond the instructed path. For larger values of D, the concurrent shove in several directions accelerates the decrease in compliance f.

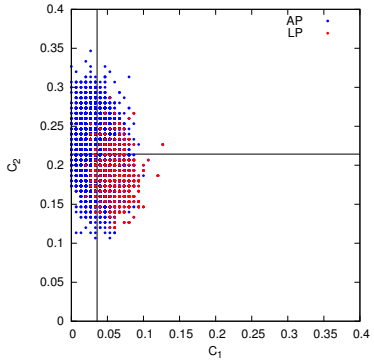
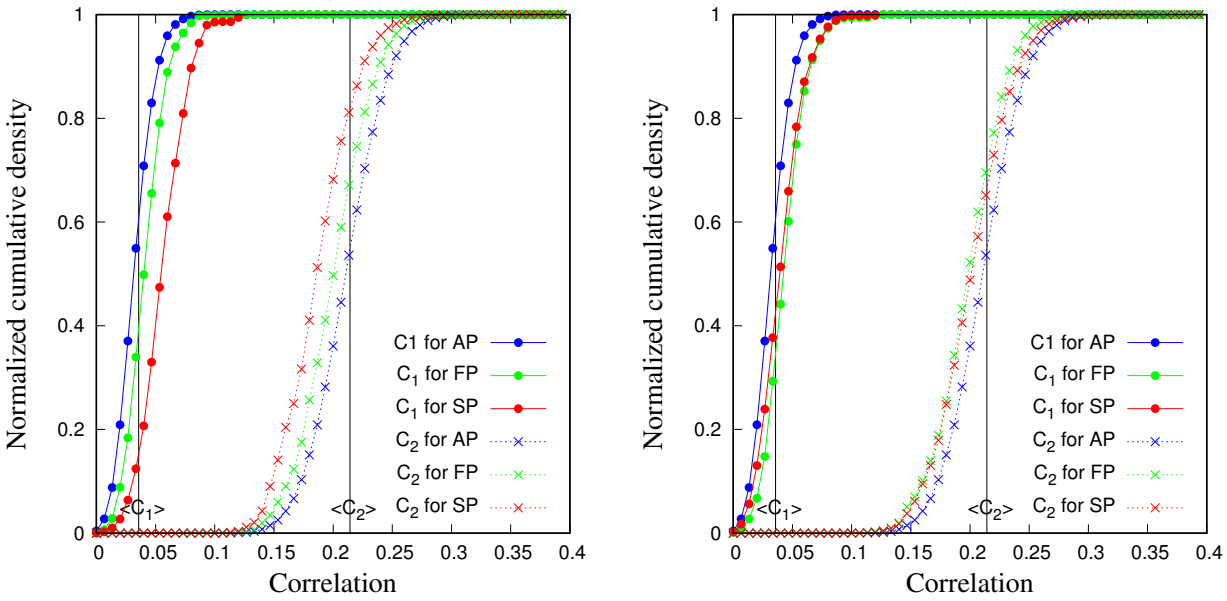


Figure 5.5: C_1-C_2 correlation scatterplot for $D = 2$, $\lambda = 0$. Blue dots are for APs and red dots for LPs. $S = 7$ and $p = 200$.



(a) $D = 2$ (b) $D = 10$

Figure 5.6: Cumulative density of pattern pairs (AP in blue, FP in green, SP in red) for increasing values of correlation, as measured by C_1 and C_2. Solid lines with circular dots are for C_1 and dashed lines with cross dots are for C_2. $\lambda = 0.1$, $S = 7$ and $p = 200$.

The correlations between latching pairs in the absence of the *hetero*-associative term are shown in Fig.5.5. C_1 and C_2 are the fractions of units that are active in the same (different) states in the two patterns of a pair. As discussed in [5], [23], latching occurs mostly for positively correlated pattern pairs, i.e., when C_1 is larger than its mean value, while C_2 is smaller. Indeed, in the scatterplot LPs (red dots) are printed on top of APs (blue dots) around the bottom right portion of the distribution of the latter in the $C_1 - C_2$ plane. The vertical and horizontal lines which cross the figure represent $\langle C_1 \rangle \equiv a/S$ and $\langle C_2 \rangle \equiv a(S-1)/S$, the average values of C_1 and C_2.

As soon as we introduce the λ coupling term, the group of red dots begins to diffuse towards the centre of the blue distribution, but in order to understand the change we need to separate FPs from SPs, among the full set of LPs. In Fig.5.6, the cumulative density of APs, FPs and SPs (in blue, green and red) are shown with respect to C_1 (solid lines with circular dots) and C_2 (dashed lines with cross dots). $\lambda = 0.1$, $S = 7$, $p = 200$ and both $D = 2$ and $D = 10$ cases are considered. The number of APs, FPs and SPs are of course normalized by their own total numbers in the latching sequence.

In Fig.5.6a, for $D = 2$, there is a clear separation of FPs and SPs from APs along both C_1 and C_2 axes. Moreover, we see that SP (red) is distributed towards larger C_1 and lower C_2 values, as already shown in Fig.5.5. FPs (green) are closer to APs (blue) in their scatter of correlation values, but still their cumulative density does not coincide with theirs. This is because at $D = 2$ there two options to follow at each stage and, even though the instructions are imparted at random, when the network does follow one of the two it tends to choose one with the correlations that more

resemble those of SPs. The separation between SPs and APs is less marked when D is high ($D = 10$ in Fig.5.6b), because at fixed λ increasing D increases the noise, and spontaneous transitions tend to occur more randomly; but even more imperceptible is the separation between SPs and FPs since, among the many options, the network apparently picks the instruction that moves it in a direction it would take anyway, spontaneously.

Conclusion

In the thesis, we studied the Potts associative memory network, a model for semantic memory storage in the cortex and possibly for recursive dynamics.

After two introductory chapters, in chapter 3 we elaborated on the correspondence between a multi-modular neural network and a coarse grained Potts network, by grounding the Hamiltonian of the Potts model in the multi-modular one. Units are taken to be threshold-linear, in the multi-modular model, and they are fully connected within a module, with Hebbian synaptic weights. Sparse connectivity links units that belong to different modules, via synapses that in the cortex impinge primarily on the apical dendrites, after their axons have travelled through the white matter. We related Potts states to the overlap or correlation between the activity state in a module and the local memory patterns, i.e., to weighted combinations of the activity of its threshold-linear units. The long range interactions between the modules then roughly correspond, after suitable assumptions about inhibition, to the tensorial couplings between Potts units in the Potts Hamiltonian. It becomes apparent how the w-term, which was initially introduced by [5] to model positive state-specific feedback on Potts units, arises from the short range interactions of the multi-modular Hamiltonian. Keeping the w-term in the Potts Hamiltonian, we applied the replica method to derive analytically the storage capacity for the fully connected Potts model. A simplified derivation was applied also to the highly diluted connectivity network, while the case with intermediate connectivity was studied by a self-consistent signal-to-noise analysis. The intermediate results smoothly interpolate the limit cases of fully and high diluted networks, but the two limit cases themselves are in fact very similar in capacity, if measured by $\alpha \equiv p/c_m$, in the sparse coding limit $a \to 0$, a limit which is approached very rapidly in the Potts model, because the relevant parameter is in fact $\tilde{a} \equiv a/S$. The effect of the w-term is effectively, in the vicinity of the memory states, reduced to altering the threshold, which leads to the storage capacity being suppressed by this term, if the threshold was close to its optimal value, to a more pronounced extent in the sparse coding regime. If one assumes that the threshold is set close to its optimal value *after* taking the feedback term into account, the value w becomes irrelevant for the storage capacity, while it still affects network dynamics, as briefly mentioned in [23] and chapter 4. Future studies are needed to more thoroughly characterize the dynamical properties of the Potts network, as they are modulated by the strength of the w-term.

In chapter 4 we have found the region, in the Potts network phase space spanned by the number of Potts states S, the number of connections per unit C and the storage load p, where latching dynamics occur, and we have described their character,

comparing and contrasting the slowly and fast adapting regimes. In relation to [8], where the possibility of such a latching region was pointed out on the basis of limited simulations, we have now a firmer basis to extrapolate to regions of parameter space of relevance to the human cortex, possibly a step toward quantitatively studying human specific capacities, including creative behaviour. A common hallmark in both regimes is that good quality latching occupies a band which scales almost quadratically in the $p-S$ plane, while it is sublinear in the $p-C$ plane. These bands are bounded by the storage capacity line, above, and by the boundary between no latching and finite latching, below. If, as discussed in the previous chapter, we were to take $C \approx 10^2$ and $S \approx 10^2$ as the orders of magnitude of interest for the human brain, we would conclude that the relevant storage load, or semantic depth, is in the region $p \approx 10^5$, in both regimes. At the center of the band in the slowly adapting regime, asymmetry and entropy take intermediate values, pointing at maximally complex and potentially useful dynamics, intermediate between the deterministic and the random extremes. High crossover values indicate that many transitions occur between highly correlated patterns. Using correlated patterns shifts the position of the band in phase space, but preserving the features observed with random patterns, still in the slowly adapting regime. In the fast adapting regime, instead, in the center of the band, which can be made wider and more robust, the entropy is higher, and correspondingly only low crossover transitions are observed, indicating that the network latches most of the time from one pattern to any other among the many with which it is weakly or anti-correlated, avoiding only those few with which it is highly correlated. Therefore, we can conclude that the fast adapting regime, modelling rapid inhibitory feedback, offers a robust framework for latching dynamics, but of an essentially random, not very useful nature; whereas in the slowly adapting regime, modelling slow inhibition or local fatigue, correlations can drive latching transitions, potentially enabling semantic content in a stream of thoughts or linguistic productions, but with fragile dynamics, living at the very edge between memory overload and sequence termination because of the inability of the network to jump forward. This suggests the opportunity of considering models that integrate both fast and slowly adapting dynamics in their non-specific thresholds, so as to combine the useful features of both regimes. In the end, we acknowledge the inherent limitation of considering a simple homogeneous Potts network, with no differentiation among its units and no internal structure. In order to make contact with cognitive processes, of any kind, this limitation has to be overcome, as perhaps attempted, with one first step among many possible ones, by arranging Potts units on a ring [84]. Nevertheless, even in its crudest form the Potts network with its latching dynamics can be used to explore e.g. novel theories as to the evolutionary origin of complex cognition [85]. It establishes a quantitative framework to understand phase transitions [5], complementary to the perspective offered by other modelling approaches to sequence generation in cortical networks [86]. At the most abstract level, it can be considered an implementation of a fuzzy logic system [87, 88], but with the critical advantage that its parameters can eventually be related to cortical parameters.

In chapter 5, finally, we have assessed the possibility of adding to the spontaneous dynamics expressed by the Potts network considered in previous chapters, a set of

specific instructions, i.e., transitions that the network is encouraged to take when it is in or near one of its memory states. The instructions are encoded, following a suggestion by Kanter and Sompolinsky, in a *hetero*-associative term parametrized by a factor λ. Another important parameter is the number of instructed transitions per memory pattern, D. The main conclusion of the study is that combining what is effectively a supervised learning of transitions with the spontaneous expression of latching sequences works only to a limited extent. As either λ or D grow in value, latching quality deteriorates, and in fact large values of either parameter end up adding noise to the dynamics. Further, the network follows the instructions most of the time only over a λ range that narrows down around $\lambda = 0.15$ as D increases, while the transitions are increasingly indistinct. The ultimate reason for the difficulty of imparting instructions, in the model, is that these are arbitrary, while latching dynamics in the Potts network, especially in the critical band studied above, in the slowly adaptive regime, favors transitions between correlated pairs of patterns. If D is large, the network can choose among many options the ones that are more correlated to its current state, but even then the presence of all the other options, with a sufficient λ factor, generates noise.

In conclusion, the Potts network offers an interesting simple model of complex spontaneous dynamical behaviour, that it is difficult to harness to externally-determined goals via supervised learning. To explore the capability of the model to approach concrete problems where latching dynamics may be relevant, it is critical of course to include structure in the so far unstructured homogeneous network, and to allow the dynamics to harmoniously reflect such structure, whether explicit or implicit, without attempting to force it to follow a prescribed course.

Appendix A

Generation of correlated patterns

In this appendix we sketch one way of generating correlated patterns and for more details we refer to [57].

A conventional way of generating the correlated patterns is mentioned in [89] by Gutfreund.

Random assortments of patterns which are called parents are taken following probability distribution

$$P(\xi_i^\pi) = a\delta(\xi_i^\pi - 1) + (1-a)\delta(\xi_i^\pi), \tag{A.1}$$

where a is sparsity of the pattern.

A "child" pattern μ is descended from a parent π following the distribution.

$$P(\xi_i^{\pi\mu}) = \{a + b\,(\xi_i^\pi - a)\}\,\delta(\xi_i^{\pi\mu} - 1) + \{1 - a - b\,(\xi_i^\pi - a)\}\,\delta(\xi_i^{\pi\mu}), \tag{A.2}$$

where b takes the value between 0 and 1 and measures kinds of influence to be born out from their single parent. For instance, children patterns become randomly correlated with each other regardless of their parent when $b = 0$, but they are correlated in a same degree when $b = 1$. Probability distributions above are identical for any unit i, we drop the index for convenience.

The average activity of parent and child patterns is computed as

$$\langle \xi^\pi \rangle = a,$$

$$\langle \xi^{\pi\mu} \rangle = a.$$

Children have more similarity to their parent than to other parents.

$$\langle \xi^{\pi\mu}\xi^{\pi'\mu} \rangle = \begin{cases} a^2 + ba - ba^2 : & \pi = \pi' \\ a^2 : & \pi \neq \pi'. \end{cases}$$

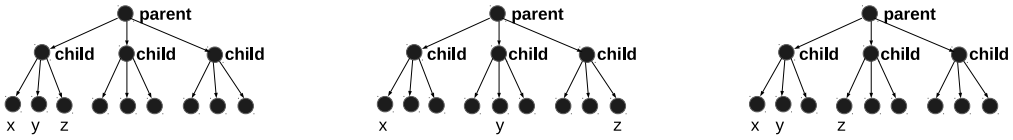
The correlation between children from different parents reads

$$\langle \xi^{\pi\mu}\xi^{\pi'\mu'} \rangle = \begin{cases} a^2 + a(1-a)b^2 : & \pi = \pi' \\ a^2 : & \pi \neq \pi'. \end{cases}$$

"Children" descended from a same "parent" have a higher correlation than they have when "children" descend from different "parents',' and this is the main aspect we would like to point to.

$$\langle \xi^{\pi\mu}\xi^{\pi'\mu'}\rangle - \langle \xi^{\pi\mu}\xi^{\pi\mu'}\rangle = ab^2(1-a).$$

It is instructive to consider an example. In Fig.(A.1), three nodes (children) x, y, z at the same level of the hierarchy satisfy the ultrametric inequality $d(x,z) \leq max(d(x,y),d(y,z))$. d refers to the distance to the nearest common forefather. When the three nodes are altogether in the same branch or all in branches different from each other, they become equidistant from one another. The situation in which two belong to the same branch and the third to the other yields always an isoscles triangle with two long sides (see A.2).



(a) (b) (c)

Figure A.1: Diagrammatic view of correlated parent pattern generation from single parent.

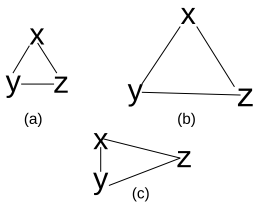


Figure A.2: Correlation distances between children generated by single parent pattern generation algorithm.

This algorithm to generate correlated patterns by single parent hierarchy seems simple and insightful. However, in reality it is likely to happen that there are cases with two short sides and a long side which in principle can not be realized.

For example, in Fig.A.3, "whale" is correlated to "tiger" more than "shark", but in fact, it has loosely speaking, similar correlation with both "tiger" and "shark", which is the typical case that can not be resolved within this frame of algorithm. One of the solution is to have several randomly chosen parents contributing to each pattern with a certain weight, as suggested by Treves ([8]). What is expected is to have larger correlations between patterns sharing more parents, reflecting what is semantically relevant between similar concepts.

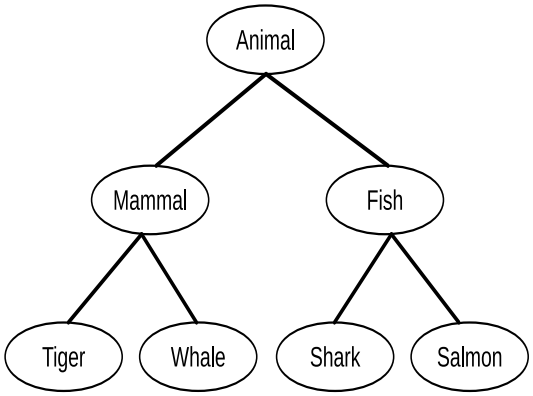


Figure A.3: Example of single parent pattern generation algorithm.

A new algorithm of multi parent pattern generation is proposed. Factors Π which are intended to be the parents for the correlated patterns are generated as the random subsets of patterns. "Children" patterns are generated from the factors. Each child pattern listens instructions from multiple factors with some noise, to eventually decide to align itself in the direction to which the factor with the largest field points, keeping that a fraction a of the units in a given pattern are set to be active.

Consider the simplest case with $\zeta = 0$ and $S = 1$. We define the occupation number n_f as the number of parents acting on each child. The probability that a child pattern is assigned to a single parent follows a Bernoulli trial with probability $\frac{p_{\text{fact}}}{p}$. Therefore, the probability that a given child pattern is generated from n_f parents out of Π reads

$$P(\hat{n}_{f_\mu} = n_{f_\mu}) = B\left[n_f; \Pi, \frac{p_{\text{fact}}}{p}\right]. \tag{A.3}$$

a_{pf} is another parameter tuning the degree of correlation between children patterns. $a_{pf} \sim 0$ corresponds to the case in which children patterns are never affected by the parents. Whereas $a_{pf} \sim 1$ implies that all the parent patterns try to influence each child pattern.

The probability distribution of the field h with n_f becomes

$$P(h|n_f) = (1-a_{pf})^{n_f}\,\delta(h) + \sum_{k=1}^{n_f}\sum_{j=0}^{k} \frac{(-1)^j n_f! a_{pf}^k (1-a_{pf})^{n_f-k}}{(n_f-k)!(k-j)!j!(k-1)!}(h-j)^{k-1}\theta(h-j). \tag{A.4}$$

The first term implies the situation in which all the n_f factors contribute zero field with probability $(1-a_{pf})^{n_f}$. The field of each unit in the child pattern is drawn by this distribution.

A fraction a of the units in the pattern are set to be active. Units experiencing the fields that satisfies $h > h_m$ are determined to be active.

$$P\left(h' < h_m | n_f\right) = 1 - a.$$

For the child pattern μ, we define

$$P(\xi_i^\mu = 1|h_i^\mu) = \theta\left(h_i^\mu - h_m\right). \tag{A.5}$$

It is trivial to generalize the results to arbitrary S, in the same way. Having obtained h_m, we have

$$P\left(\xi_{ik}^\mu = 1|h_{ik}^\mu\right) = \theta\left(h_{ik}^\mu - h_m\right), \tag{A.6}$$

where k denotes the active and inactive states of Potts unit. h_{ik}^μ follows the distribution $P(h_{ik}|n_k)$, where n_k is the number of parent patterns acting on the state k of pattern μ.

Appendix B

Derivation of the replica symmetric free energy

In this appendix, we derive the replica symmetric free energy (3.33), starting from

$$f = \lim_{n\to 0} f_n = \lim_{n\to 0}\left\{\frac{a\left(1-\tilde{a}\right)}{2n}\sum_{\nu\gamma}\left(m_\nu^\gamma\right)^2 + \frac{\alpha}{2n\beta}\operatorname{Tr}\ln\left[a\left(1-\tilde{a}\right)\left(\mathbb{1}-\beta\tilde{a}\mathbf{q}\right)\right] + \frac{\alpha\beta\tilde{a}^2}{2n}\sum_{\gamma\lambda} r_{\gamma\lambda}q_{\gamma\lambda}\right.$$
$$\left. + \frac{\tilde{a}}{n}\left[\frac{\alpha}{2} + S\left(U - \frac{w\left(S-1\right)}{2S}\right)\right]\sum_{\gamma\gamma} q_{\gamma\gamma} - \frac{1}{n\beta}\left\langle \ln \operatorname{Tr}_{\{\sigma^\gamma\}} \exp\left[\beta\hat{H}_\xi\right]\right\rangle_{\xi^v}\right\} \tag{B.1}$$

by using

$$m_\gamma^\nu = m$$
$$q_{\gamma\lambda} = \begin{cases} q & \gamma \neq \lambda \\ \tilde{q} & \gamma = \lambda \end{cases}$$
$$r_{\gamma\lambda} = \begin{cases} r & \gamma \neq \lambda \\ \tilde{r} & \gamma = \lambda. \end{cases}$$

The terms in (B.1) are evaluated as follows;

- $\frac{a(1-\tilde{a})}{2n} m^2 n = \frac{a(1-\tilde{a})}{2} m^2$;
- $\frac{\alpha}{2n\beta}\operatorname{Tr}\ln\left[a\left(1-\tilde{a}\right)\left(\mathbb{1}-\beta\tilde{a}\mathbf{q}\right)\right]$;
  this matrix has $n-1$ eigenvalues of $a\left(1-\tilde{a}\right)\left[1-\left(\tilde{q}-q\right)\beta\tilde{a}\right]$ and one eigenvalue of $a\left(1-\tilde{a}\right)\left[1-\beta\tilde{a}\tilde{q}-\left(n-1\right)\beta\tilde{a}q\right]$. With the definition $C \equiv \beta\left(\tilde{q}-q\right)$ and the relation $\ln\left(1+x\right) \sim x$ for small x, it becomes

$$\begin{aligned}\frac{\alpha}{2n\beta}\operatorname{Tr}\ln\left[a\left(1-\tilde{a}\right)\left(\mathbb{1}-\beta\tilde{a}\mathbf{q}\right)\right] &= +\frac{\alpha}{2n\beta}\left\{\left(n-1\right)\ln\left[a\left(1-\tilde{a}\right)\left(1-\tilde{a}C\right)\right]\right. \\ &\quad \left. + \ln\left[a\left(1-\tilde{a}\right)\left[1-\beta\tilde{a}\tilde{q}-\left(n-1\right)\beta\tilde{a}q\right]\right]\right\} \\ &= +\frac{\alpha}{2n\beta}\left\{n\ln\left[a\left(1-\tilde{a}\right)\left(1-\tilde{a}C\right)\right]\right. \\ &\quad \left. + \ln\left[1-\frac{n\beta\tilde{a}q}{1-\tilde{a}C}\right]\right\} \\ &\xrightarrow{n\to 0} + \frac{\alpha}{2\beta}\left[\ln\left(a\left(1-\tilde{a}\right)\right) + \ln\left(1-\tilde{a}C\right) - \frac{\beta\tilde{a}q}{\left(1-\tilde{a}C\right)}\right]\end{aligned}$$

- the third term

$$\begin{aligned}\frac{\alpha\beta\tilde{a}^2}{2n}\sum_{\gamma\lambda} r_{\gamma\lambda}q_{\gamma\lambda} &= \frac{\alpha\beta\tilde{a}^2}{2n}\left(\sum_{\gamma=\lambda} r_{\gamma\lambda}q_{\gamma\lambda} + \sum_{\gamma\neq\lambda} r_{\gamma\lambda}q_{\gamma\lambda}\right)\\ &= \frac{\alpha\beta\tilde{a}^2}{2n}\left(\tilde{r}\tilde{q}n + n\left(n-1\right)rq\right)\\ &\xrightarrow{n\to 0} \frac{\alpha\beta\tilde{a}^2}{2}\left(\tilde{r}\tilde{q} - rq\right)\ ;\end{aligned}$$

- $\frac{\tilde{a}}{n}\left[\frac{\alpha}{2} + S\left(U - \frac{w(S-1)}{2S}\right)\right]\sum_{\gamma\gamma} q_{\gamma\gamma} = \tilde{a}\tilde{q}\left[\frac{\alpha}{2} + S\left(U - \frac{w(S-1)}{2S}\right)\right]$;

- the exponent of the Hamiltonian $\hat{H}_\xi$ in the last term is expanded using the Hubbard Stratonovich transform as

$$\begin{aligned}\exp\left[\hat{H}_\xi\right] &= \exp\left[nmv_{\xi\sigma} + \frac{\alpha\beta}{2S\left(1-\tilde{a}\right)}\left(n\tilde{r} + n\left(n-1\right)r\right)\sum_k P_k v_{k\sigma}^2\right]\\ &= \exp\left[nmv_{\xi\sigma} + \frac{\alpha\beta a}{2S^2}n\left(\tilde{r}-r\right)\left(1-\delta_{\sigma 0}\right) - \frac{\alpha\beta n^2 r}{2S\left(1-\tilde{a}\right)}\sum_k P_k v_{k\sigma}^2\right]\\ &= \int dz_k \exp\left[-\frac{z_k^2}{2} + nmv_{\xi\sigma} + \frac{\alpha\beta a}{2S^2}n\left(\tilde{r}-r\right)\left(1-\delta_{\sigma 0}\right) + n\sum_k\sqrt{\frac{\alpha\beta P_k}{2S\left(1-\tilde{a}\right)}}z_k v_{k\sigma}\right],\end{aligned}$$

  and therefore

$$-\tfrac{1}{n\beta}\left\langle\!\left\langle \ln \mathrm{Tr}_{\{\sigma^\gamma\}} \exp\left[\beta\hat{H}_\xi\right]\right\rangle\!\right\rangle_{\xi^v} = -\tfrac{1}{\beta}\left\langle\!\left\langle \int Dz \ln\left(1 + \textstyle\sum_{\sigma\neq 0}\exp\left[\beta\hat{H}_\sigma^\xi\right]\right)\right\rangle\!\right\rangle.$$

Finally, we get the replica symmetric free energy

$$\begin{aligned}f =& \frac{a\left(1-\tilde{a}\right)}{2}m^2 + \frac{\alpha}{2\beta}\left[\ln\left(a\left(1-\tilde{a}\right)\right) + \ln\left(1-\tilde{a}C\right) - \frac{\beta\tilde{a}q}{\left(1-\tilde{a}C\right)}\right] + \frac{\alpha\beta\tilde{a}^2}{2}\left(\tilde{r}\tilde{q} - rq\right)\\ &+ \tilde{a}\tilde{q}\left[\frac{\alpha}{2} + S\left(U - \frac{w\left(S-1\right)}{2S}\right)\right] - \frac{1}{\beta}\left\langle \int Dz \ln\left(1 + \sum_{\sigma\neq 0}\exp\left[\beta\hat{H}_\sigma^\xi\right]\right)\right\rangle.\end{aligned} \tag{B.2}$$

Appendix C

Saddle point equations in limit case

We consider (3.41)-(3.44) in the limit of $\tilde{a} \ll 1$.
Using

$$\int Dw = \int \frac{dw}{\sqrt{2\pi}} \exp\left(-w^2/2\right) = 1,$$

$$q \approx \frac{1-a}{\tilde{a}} \int_x^\infty Dz\phi^S(z) + \int_{x-y}^\infty Dz\phi^S(x+y) + (S-1) \int Dw \int_x^\infty Dz\phi(z-y)\phi^(S-1)(z).$$

Notice that $\phi(z)$ is like a Heaviside step function for real z values. Moreover,

$$\begin{cases} \phi(z) &= \left(1 + \mathrm{Erf}(\frac{z}{\sqrt{2}})\right)/2 = \left(1 + \frac{1}{\sqrt{2\pi}} \int_0^{z/\sqrt{2}} \exp\left(-t^2/2\right) dt\right)/2 \\ d\phi &= Dz \\ \phi^S &\sim O(1). \end{cases}$$

The first term in q then becomes

$$\frac{1-a}{\tilde{a}} \int_x^\infty Dz\phi^S(z) \approx \frac{1-a}{\tilde{a}} \int_x^\infty d\phi(z) = \frac{1-a}{\tilde{a}}(1-\phi(x)) = \frac{1-a}{\tilde{a}}\phi(-x)$$

and the second term is

$$\int_{x-y}^\infty Dz\phi^S(x+y) \approx \int_{x-y}^\infty d\phi \approx \phi(y-x).$$

The last term can be neglected since it is much smaller than the first two terms. Therefore, we get expressions for q and m

$$q = \frac{1-a}{\tilde{a}}\phi(-x) + \phi(y-x) \tag{C.1}$$

and

$$m = \int_{x-y}^\infty Dz\phi^S(z+y) \approx 1 - \phi(x-y) = \phi(y-x). \tag{C.2}$$

$C\sqrt{r}$ is treated in the same way, as

$$C\sqrt{r} \approx \frac{1}{\sqrt{\tilde{a}}} \left\{ \frac{1-a}{\tilde{a}} \int_x^\infty Dzz \cdot \phi^S(z) + \int_{x-y} Dzz \cdot \phi^S(z+y) + (S-1) \int_x^\infty Dzz \cdot \phi(z-y)\phi^(S-1) \right\}.$$

The first term in curly bracket above is

$$\frac{1-a}{\tilde{a}}\int_x^\infty Dzz\cdot\phi^S(z) \approx \frac{1-a}{\tilde{a}}\int_x^\infty Dz\cdot z = \frac{1-a}{\tilde{a}}\int_x^\infty \frac{dz}{\sqrt{2\pi}}\exp\left(-z^2/2\right)\cdot z = \frac{1-a}{\sqrt{2\pi}\tilde{a}}\exp\left(-x^2/2\right)$$

while for the second term,

$$\int_{x-y} Dz\cdot z = \frac{1}{\sqrt{2\pi}}\int dz\exp\left(-z^2/2\right)\cdot z = \frac{1}{\sqrt{2\pi}}\exp\left(-(y-x)^2/2\right). \tag{C.3}$$

We finally get the expression for $C\sqrt{r}$

$$C\sqrt{r} \approx \frac{1}{\sqrt{2\pi\tilde{a}}}\left\{\frac{1-a}{\tilde{a}}\exp\left(-x^2/2\right) + \exp -(y-x)^2/2\right\}. \tag{C.4}$$

Appendix D

Self consistent signal to noise analysis

There are $p-1 \gg 1$ terms in (D.1), so that the ansatz remains valid also when taking one of these many contributions out.

$$\sum_{\nu>1} v_{\xi_i^\nu,k} m_i^\nu = v_{\xi_i^\mu,k} m_i^\mu + \sum_{\nu\neq 1,\mu} v_{\xi_i^\nu,k} m_i^\nu = v_{\xi_i^\mu,k} m_i^\mu + \gamma_i^k \langle \sigma_i^k \rangle + \sum_n v_{n,k}\, \rho_i^n z_i^n \ , \tag{D.1}$$

where γ_i^k and ρ_i^n are independent of μ. The contribution from the non-condensed pattern $\mu \neq 1$ is assumed to be small, so that we can expand G_i^k to first order in $v_{\xi_i^\mu,k} m_i^\mu$:

$$\begin{aligned} \sigma_j^l &= G^l\left[\left\{ v_{\xi_j^1,k} m_j^1 + \sum_n v_{n,k} \rho_j^n z_j^n - U(1-\delta_{k,0}) \right\}_{k=0}^{S}\right] \\ &\quad + \sum_n v_{\xi_j^\mu,n}\, m_j^\mu \frac{\partial G^l}{\partial y^n}\left[\left\{ v_{\xi_j^1,k} m_i^1 + \sum_n v_{n,k} \rho_j^n z_j^n - U(1-\delta_{k,0}) \right\}\right] . \end{aligned} \tag{D.2}$$

Reinserting the expansion into the r.h.s of (3.53) we recognize a relation of the form

$$m_i^\mu = L_i^\mu + \sum_j K_{ij}^\mu m_j^\mu \tag{D.3}$$

where

$$\begin{aligned} K_{ij}^\mu &\equiv \frac{1}{c_m a(1-\tilde{a})} \sum_{l,n} c_{ij} v_{\xi_j^\mu,l} v_{\xi_j^\mu,n} \frac{\partial G_j^l}{\partial y^n} \ , \\ L_i^\mu &\equiv \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^\mu,l} G_j^l \ . \end{aligned}$$

The overlap m_i^μ can be found by iterating (D.3),

$$m_i^\mu = L_i^\mu + \sum_{j_1} L_{j_1}^\mu \left\{ K_{ij_1}^\mu + \sum_{j_2} K_{ij_2}^\mu K_{j_2 j_1}^\mu + \sum_{j_2} \sum_{j_3} K_{ij_2}^\mu K_{j_2 j_3}^\mu K_{j_3 j_1}^\mu + \ldots \right\} . \tag{D.4}$$

Therefore, the noise term can be written explicitly as

$$\begin{aligned} \sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu &= \sum_n v_{n,k} \sum_{\mu>1} \left\{ \sum_j \sum_l \frac{1}{c_m a(1-\tilde{a})} c_{ij} \delta_{\xi_i^\mu,n} v_{\xi_j^\mu,l} G_j^l + \right. \\ &\quad + \left. \sum_{j_1} \sum_j \sum_l \frac{1}{c_m a(1-\tilde{a})} c_{j_1 j} \delta_{\xi_i^\mu,n} v_{\xi_j^\mu,l} G_j^l \left(\sum_{l_1,n_1} \frac{1}{c_m a(1-\tilde{a})} c_{ij_1} v_{\xi_{j_1}^\mu,l_1} v_{\xi_{j_1}^\mu,n_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{n_1}} + \ldots \right) \right\} . \end{aligned}$$

In order to obtain the expression for γ_i^k, in (D.1) we consider only the terms with $j = i$ and $l = k$, and take the average over the connectivity and the patterns:

$$\begin{aligned}\gamma_i^k &= \frac{\alpha}{S}\lambda\left\langle \frac{1}{S}\frac{1}{N}\sum_{j_1}\sum_{l_1}\frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} + ... \right\rangle \\ &= \frac{\alpha}{S}\lambda\left\{\Omega/S + (\Omega/S)^2 + ...\right\} \\ &= \frac{\alpha}{S}\lambda\frac{\Omega/S}{1-\Omega/S}\end{aligned} \tag{D.5}$$

where we use the fact that $c_{ii} = 0$, $\alpha = p/c_m$, $\langle\cdot\rangle$ indicates the average over all patterns and where we have defined

$$\Omega = \left\langle \frac{1}{N}\sum_{j_1}\sum_{l_1}\frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} \right\rangle. \tag{D.6}$$

By virtue of the statistical independence of units, the average over the non-condensed patterns for the $i \neq j$ terms vanishes. From the variance of the noise term one reads

$$(\rho_i^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})}q\left\{1 + 2\lambda\Psi + \lambda\Psi^2\right\}, \tag{D.7}$$

where

$$q = \left\langle \frac{1}{Na}\sum_{j,l}(G_j^l)^2 \right\rangle \tag{D.8}$$

and

$$\Psi = \frac{\Omega/S}{1-\Omega/S}\ . \tag{D.9}$$

The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k}m + \frac{\alpha}{S}\lambda\Psi(1-\delta_{k,0}) + \sum_n v_{n,k}z^n\sqrt{\frac{\alpha P_n}{S(1-\tilde{a})}q\left\{1 + 2\lambda\Psi + \lambda\Psi^2\right\}} - \tilde{U}(1-\delta_{k,0})\,. \tag{D.10}$$

Acknowledgement

Foremost, I would like to express my sincere gratitude to my advisor Prof. Alessandro Treves for his tireless encouragements, discussions and constant supports. His door was always open for me and it must be emphasised that this work could not have been done without his insightful guidance. Thanks you Vezha Boboeva and Michelangelo Naim for the enthusiatic collaborations, with you, we have accomplished wonderful jobs. I would like to thank Prof. John Nicholls for delievering informal lectures and discussions on the basics of neuroscience.
Special thanks goes to Prof. Jin U Kang for his great cares, supports and helps in many aspects, most importantly, for giving opportunity to come to Trieste
. I also have to convey my gratitude to Prof. Fernanado Quevedo and Prof. Stefano Ruffo, the directors of ICTP and SISSA, who have beening supporting and trying to help us in many ways. Thanks to Prof. Sandro Scandolo, Dr. Nicola Seriani and Dr. Natasha Stojic, I do not forget your constant supports. I should not also forget to acknowledge Prof. Massimo Capone, Prof. Giuseppe Santoro and Dr. Laura Fanfarillo for their systamatic guidance in early stage of the research.
I am grateful to Prof. Yong Hae Ko, Prof. Chang Ho Choi, Prof Hak Chol Pak, Prof. Kuk Chol Ri, Prof. Sung Jin O, Dr. Chol Won Ri, Dr. Myong Chol Pak and Thae Hyok Kim, Kum Hyok Jong for their endless supports, cares and encouragements. Of course, I am not going to miss the names; Ok Song An, Kwang Hyok Jong and Ui Ri Mun, with you we have spent wonderful days. I would like to thank Francesco Grandi, Lorenzo Privitera, Tommaso Zanca, Simone Notarinicola, Caterina De Franco, Mariam Rushisvilli, Maja Berovic and Seher Karkuzu, we studied, discussed, enjoyed together.
I am grateful to Adriano Amaricci, Leonardo Romor and Alberto Saritori for useful discussions in programming.
Thanks to my officemates, Dr. Sophie Rosay, Katarina Marjanovic and Zeynap Kaya for their kindness and helps, with you I really enjoyed the time. Especially, I thank Sophie and John for critical reading and pointing out the irrelevant expressions in the manuscript. I finally acknowledge all the professors in Cognitive neuroscience section in SISSA and members in *limbo*.

Bibliography

[1] Sophie Rosay. *A Statistical Mechanics approach to the modelling and analysis of place-cell activity*. PhD thesis, Paris, Ecole Normale Supérieure, 2014.

[2] Valentino Braitenberg and Almut Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. 2013.

[3] Santiago Ramón y Cajal. *Histologie du système nerveux de l'homme & des vertébrés*. A. Maloine, 1909.

[4] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985.

[5] Eleonora Russo and Alessandro Treves. Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E*, 85(5):051920, 2012.

[6] Alessandro Treves and Edmund T Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991.

[7] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[8] Alessandro Treves. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology*, 22(3-4):276–291, 2005.

[9] Valentino Braitenberg. Thoughts on the cerebral cortex. *Journal of theoretical biology*, 46(2):421–447, 1974.

[10] Valentino Braitenberg. Cell assemblies in the cerebral cortex. In *Theoretical approaches to complex systems*, pages 171–188. Springer, 1978.

[11] Ido Kanter. Potts-glass models of neural networks. *Physical Review A*, 37(7):2739, 1988.

[12] Désiré Bollé, Patrick Dupont, and Jort van Mourik. Stability properties of potts neural networks with biased patterns and low loading. *Journal of Physics A: Mathematical and General*, 24(5):1065, 1991.

[13] Désiré Bollé, Patrick Dupont, and J Huyghebaert. Thermodynamic properties of the q-state potts-glass neural network. *Physical Review A*, 45(6):4194, 1992.

[14] Désiré Bollé, B Vinck, and VA Zagrebnov. On the parallel dynamics of the q-state potts and q-ising neural networks. *Journal of statistical physics*, 70(5):1099–1119, 1993.

[15] Désiré Bollé, Roland Cools, Patrick Dupont, and J Huyghebaert. Mean-field theory for the q-state potts-glass neural network with biased patterns. *Journal of Physics A: Mathematical and General*, 26(3):549, 1993.

[16] Dominic O'kane and Alessandro Treves. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems*, 3(4):379–384, 1992.

[17] Dominic O'Kane and Alessandro Treves. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25(19):5055, 1992.

[18] Carlo Fulvi Mari and Alessandro Treves. Modeling neocortical areas with a modular neural network. *Biosystems*, 48(1):47–55, 1998.

[19] Carlo Fulvi Mari. Extremely dilute modular neuronal networks: Neocortical memory retrieval dynamics. *Journal of Computational Neuroscience*, 17(1):57–79, 2004.

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

[21] Eleonora Russo, Vijay MK Namboodiri, Alessandro Treves, and Emilio Kropff. Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, 10(1):015008, 2008.

[22] Eleonora Russo, Sahar Pirmoradian, and Alessandro Treves. Associative latching dynamics vs. syntax. In *Advances in Cognitive Neurodynamics (II)*, pages 111–115. Springer, Dordrecht, 2011.

[23] Chol Jun Kang, Michelangelo Naim, Vezha Boboeva, and Alessandro Treves. Life on the edge: Latching dynamics in a potts neural network. *Entropy*, 19(9), 2017.

[24] Michelangelo Naim, Vezha Boboeva, Chol Jun Kang, and Alessandro Treves. Reducing a cortical network to a potts model yields storage capacity estimates. *in preparation*, 2017.

[25] Chol Jun Kang, Naim, and Alessandro Treves. Instructed latching dynamics via hetero hebbian type learning rule in potts model network. *in preparation*, 2017.

[26] Valentino Braitenberg. *Cortical architectionics: General and areal*, volume 18. In M. A. B. Brazier & H. Petsche (Eds.),Architectonics of the cerebral cortex, New York, 1978.

[27] Philip W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[28] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[29] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[30] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

[31] Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press, 1992.

[32] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[33] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Co Inc, 1987.

[34] Daniel J Amit. The hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and brain sciences*, 18(4):631–631, 1995.

[35] Ichiro Tsuda. Dynamic link of memorychaotic memory map in nonequilibrium neural networks. *Neural networks*, 5(2):313–326, 1992.

[36] Edmund T Rolls, Alessandro Treves, and Edmund T Rolls. *Neural networks and brain function*, volume 572. Oxford university press Oxford, 1998.

[37] Brad E Pfeiffer and David J Foster. Hippocampal place cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74, 2013.

[38] Moshe Abeles. *Local cortical circuits: an electrophysiological study*, volume 6. Springer Science & Business Media, 2012.

[39] H. Sompolinsky and I. Kanter. Temporal association in asymmetric neural networks. *Phys. Rev. Lett.*, 57:2861–2864, 1986.

[40] Ichiro Tsuda. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences*, 24(5):793810, 2001.

[41] Ichiro Tsuda, Edger Koerner, and Hiroshi Shimizu. Memory dynamics in asynchronous neural networks. *Progress of Theoretical Physics*, 78(1):51–71, 1987.

[42] M. Herrmann, E. Ruppin, and M. Usher. A neural model of the dynamic activation of memory. *Biological Cybernetics*, 68(5):455–463, 1993.

[43] Paul W Burgess. Confabulation and the control of recollection. *Memory*, 4(4):359–412, 1996.

[44] Russell Epstein. The neural-cognitive basis of the jamesian stream of thought. *Consciousness and Cognition*, 9(4):550–575, 2000.

[45] Moshe Abeles. Time is precious. *Science*, 304(5670):523–524, 2004.

[46] Friedemann Pulvermüller. Brain mechanisms linking language and action. *Nature reviews. Neuroscience*, 6(7):576, 2005.

[47] Tomer Shmiel, Rotem Drori, Oren Shmiel, Yoram Ben-Shaul, Zoltan Nadasdy, Moshe Shemesh, Mina Teicher, and Moshe Abeles. Temporally precise cortical firing patterns are associated with distinct action segments. *Journal of neurophysiology*, 96(5):2645–2652, 2006.

[48] Ronen Sosnik, Moshe Shemesh, and Moshe Abeles. The point of no return in planar hand movements: an indication of the existence of high level motion primitives. *Cognitive neurodynamics*, 1(4):341–358, 2007.

[49] Carlos Aguilar, Pascal Chossat, Martin Krupa, and Frdric Lavigne. Latching dynamics in neural networks with synaptic depression. *PLOS ONE*, 12(8):1–29, 08 2017.

[50] Mohammad-Farshad Abdollah-nia, Mohammadkarim Saeedghalati, and Abdolhossein Abbassian. Optimal region of latching activity in an adaptive potts model for networks of neurons. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(02):P02018, 2012.

[51] Itamar Lerner, Shlomo Bentin, and Oren Shriki. Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive science*, 36(8):1339–1382, 2012.

[52] Itamar Lerner, Shlomo Bentin, and Oren Shriki. Excessive attractor instability accounts for semantic priming in schizophrenia. *PLoS One*, 7(7):e40663, 2012.

[53] Sandro Romani, Itai Pinkoviezky, Alon Rubin, and Misha Tsodyks. Scaling laws of associative memory retrieval. *Neural computation*, 25(10):2523–2544, 2013.

[54] Stefano Recanatesi, Mikhail Katkov, Sandro Romani, and Misha Tsodyks. Neural network model of memory retrieval. *Frontiers in computational neuroscience*, 9, 2015.

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

[56] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[57] Vezha Boboeva and Alessandro Treves. The storage capacity of the potts network with correlated patterns. *in preparation*, 2017.

[58] Leonid Schneider. Human brain project: bureaucratic success despite scientific failure. https://forbetterscience.com/2017/02/22/human-brain-project-bureaucratic-success-despite-scientific-failure/.

[59] Vernon B Mountcastle. The columnar organization of the neocortex. *Brain: a journal of neurology*, 120(4):701–722, 1997.

[60] Pasko Rakic. Confusing cortical columns. *Proceedings of the National Academy of Sciences*, 105(34):12099–12100, 2008.

[61] Jon H Kaas. Evolution of columns, modules, and domains in the neocortex of primates. *Proceedings of the National Academy of Sciences*, 109(Supplement 1):10655–10660, 2012.

[62] Alessandro Treves. Graded-response neurons and information encodings in autoassociative memories. *Physical Review A*, 42(4):2418, 1990.

[63] Daniel J Amit and Nicolas Brunel. Learning internal representations in an attractor neural network with analogue neurons. *Network: Computation in Neural Systems*, 6(3):359–388, 1995.

[64] Bernhard Hellwig. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological cybernetics*, 82(2):111–121, 2000.

[65] Nir Kalisman, Gilad Silberberg, and Henry Markram. The neocortical microcircuit as a tabula rasa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):880–885, 2005.

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

[67] R Lauro-Grotto, S Reich, and Miguel A Virasoro. The computational role of conscious processing in a model of semantic memory. 1997.

[68] Nir Levy, David Horn, and Eytan Ruppin. Associative memory in a multimodular network. *Neural Computation*, 11(7):1717–1737, 1999.

[69] Alexis M Dubreuil and Nicolas Brunel. Storing structured sparse memories in a multi-modular cortical network model. *Journal of computational neuroscience*, 40(2):157–175, 2016.

[70] MV Tsodyks and MV Feigel'Man. The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)*, 6(2):101, 1988.

[71] A Treves. Threshold-linear formal neurons in auto-associative nets. *Journal of Physics A: Mathematical and General*, 23(12):2631, 1990.

[72] A Treves. Dilution and sparse coding in threshold-linear nets. *Journal of Physics A: Mathematical and General*, 24(1):327.

[73] Alessandro Treves. Mean-field analysis of neuronal spike dynamics. *Network: Computation in Neural Systems*, 4(3):259–284, 1993.

[74] Francesco P Battaglia and Alessandro Treves. Stable and rapid recurrent processing in realistic autoassociative memories. *Neural Computation*, 10(2):431–450, 1998.

[75] Tamás Geszti. *Physical models of neural networks*. World Scientific, 1990.

[76] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[77] Bernard Derrida, Elizabeth Gardner, and Anne Zippelius. An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4(2):167, 1987.

[78] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[79] Andreas Engel, Rémi Monasson, and Alexander K Hartmann. On large deviation properties of erdös–rényi random graphs. *Journal of Statistical Physics*, 117(3-4):387–426, 2004.

[80] Emilio Kropff. Full solution for the storage of correlated memories in an autoassociative memory. *Computational Modelling in Behavioural Neuroscience: Closing the Gap Between Neurophysiology and Behaviour*, 2:225, 2009.

[81] Masatoshi Shiino and Tomoki Fukai. Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Physical Review E*, 48(2):867, 1993.

[82] Alessandro Treves and Edmund T Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991.

[83] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[84] Sanming Song, Hongxun Yao, and Alessandro Treves. A modular latching chain. *Cognitive neurodynamics*, 8(1):37–46, 2014.

[85] Daniele Amati and Tim Shallice. On the emergence of modern humans. *Cognition*, 103(3):358–385, 2007.

[86] Kanaka Rajan, ChristopherD. Harvey, and DavidW. Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128 – 142, 2016.

[87] Yizhang Jiang, Fu-Lai Chung, Hisao Ishibuchi, Zhaohong Deng, and Shitong Wang. Multitask tsk fuzzy system modeling by mining intertask common hidden structure. *IEEE transactions on cybernetics*, 45(3):548–561, 2015.

[88] Chiu-Chuan Tu and Chia-Feng Juang. Recurrent type-2 fuzzy neural network using haar wavelet energy and entropy features for speech detection in noisy environments. *Expert systems with applications*, 39(3):2479–2488, 2012.

[89] Hanoch Gutfreund. Neural networks with hierarchically correlated patterns. *Physical Review A*, 37(2):570, 1988.