International School for Advanced Studies

# The storage of semantic memories in the cortex: a computational study

Candidate: Vezha Boboeva

Advisor: Alessandro Treves

Thesis submitted for the degree of Doctor of Philosophy in Neuroscience

Trieste, 2017

# Contents

# List of Figures

## Abstract

The main object of this thesis is the design of structured distributed memories for the purpose of studying their storage and retrieval properties in large scale cortical auto-associative networks. For this, an autoassociative network of Potts units, coupled via tensor connections, has been proposed and analyzed as an effective model of an extensive cortical network with distinct short and long-range synaptic connections. Recently, we have clarified in what sense it can be regarded as an effective model. While the fully-connected (FC) and the very sparsely connected, that is, highly diluted (HD) limits of the model have thoroughly analyzed, the realistic case of the intermediate partial connectivity has been simply assumed to interpolate the FC and HD cases. In this thesis, we first study the storage capacity of Potts network with such intermediate connectivity. We corroborate the outcome of the analysis by showing that the resulting mean field equations are consistent with the FC and HD equations under the appropriate limits. The mean-field equations are only derived for *randomly diluted connectivity* (RD). Through simulations, we also study *symmetric dilution* (SD) and *state dependent random dilution* (SDRD). We find that the Potts network has a higher capacity for symmetric than for random dilution.

We then turn to the core question: how to use a model originally conceived for the storage of $p$ unrelated patterns of activity, in order to study semantic memory, which is organized in terms of the relations between the facts and the attributes of real-world knowledge. To proceed, we first formulate a mathematical model of generating patterns with correlations, as an extension of a hierarchical procedure for generating ultrametrically organized patterns. The model ascribes the correlations between patterns to the influence of underlying "factors"; if many factors act with comparable strength, their influences balance out and correlations are low; whereas if a few factors dominate, which in the model occurs for increasing values of a control parameter $\zeta$, correlations between memory patterns can become much stronger. We show that the extension allows for correlations between patterns that are neither trivial (as in the random case) nor a plain tree (as in the ultrametric case), but that are highly sensitive to the values of the correlation parameters that we define.

Next, we study the storage capacity of the Potts network when the patterns are correlated by way of our algorithm. We show that fewer correlated patterns can be stored and retrieved than random ones, and that the higher the degree of correlation, the lower the capacity. We find that the mean-field equations yielding the storage capacity are different from those obtained with

uncorrelated patterns through only an additional term in the noise, proportional to the number of learned patterns $p$ and to the difference between the average correlation between correlated patterns and independently generated patterns of the same sparsity.

Of particular interest is the role played by the parameter we have introduced, $\zeta$, which controls the strength of the influences of different factors (the "parents") in generating the memory patterns (the "children"). In particular, we find that for high values of $\zeta$, so that only a handful of parents are effective, the network exhibits *correlated retrieval*, in which the network, though not being able to retrieve the pattern cued, settles into a configuration of high overlap with another pattern. This behavior of the network can be interpreted as reflecting the *semantic* structure of the correlations, in which even after capacity collapse, what the network *can* still do is to recognize the strongest features associated with the pattern. This observation is better quantified using the mutual information between the pattern cued and the configuration the network settles into, after retrieval dynamics. This information is found to increase from zero to a non-zero value abruptly when increasing the parameter $\zeta$, akin to a phase transition. Two alternative phases are then identified, $\zeta < \zeta_c$, in which many factors are on equal footing and there is not much structure. In this phase, when the network fails to retrieve, it fails to retrieve any learned configuration. For $\zeta > \zeta_c$, memories form clusters, such that while the specifics of the cued pattern cannot be retrieved, some of the structure informing the cluster of memories can still be retrieved.

In a final short chapter, we attempt to understand the implications of having stored correlated memories on latching dynamics, the spontaneous behavior which has been proposed to be an emergent property, beyond the simple cued retrieval paradigm, of large cortical networks. Progress made in this direction, studying the Potts network, has so far focused on uncorrelated memories. Introducing correlations, we find a rich phase space of behaviors, from sequential retrieval of memories, to parallel retrieval of clusters of highly correlated memories and oscillations, depending on the various correlation parameters. The parameters of our algorithm may be found to emerge as critical control parameters, corresponding to the statistical features in human semantic memory most important in determining the dynamics of our trains of thoughts.

### List of articles published, in press and in preparation

[Kang et al., 2017] Kang, C.J., Naim, M., Boboeva, V. and Treves, A., 2017. Life on the edge: Latching dynamics in a potts neural network. Entropy, 19(9), p.468. (*published*)

[Naim et al., 2017] Naim*, M., Boboeva*, V., Kang, C.J. and Treves, A., 2017. Reducing a cortical network to a Potts model yields storage capacity estimates. arXiv preprint arXiv:1710.04897. (*submitted, in review*) (Chap. 3)

Boboeva, V. and Treves, A., 2018. Estimating the storage capacity for semantic memories in the cortex (*in preparation*) (Chaps. 4 and 5)

Boboeva, V. and Treves, A., 2018. Latching dynamics of semantic memories (*in preparation*) (Chap. 6)

# Acknowledgements

I am deeply grateful to my supervisor Alessandro. I would like to thank him for his encouragement, his support, for always being there. I would like to thank him for giving me the opportunity to work on this creative and challenging topic, and for allowing me to be part of this unusual group that is *limbo*, for the oftentimes surreal but unique moments.

I would like to dedicate this thesis to my admired and beloved grandfather, who left us before I finished it. The memory of his kindness and the legacy of his work will always remain. I thank him for the inspiration that he has been and that will continue to be. I thank my grandmother for her warmth, my mother, Anahita for her strength of character and will, Parisa and Pariya for their endless kindness, for being so much like my grandfather, for being my best friends. Geert and sweet Roya, who never ceases to amaze me.

I would like to thank *limbo*. And the friends I made in SISSA. A most special place goes to Zeynep, my limbo-sister, for all we have endured together. For the bond we have developed in these years. To Romain, whose noisy presence has made research over the past years less lonely, for the endless discussions on every topic one could think of. To Chol-Jun and Michi, for the fruitful collaboration and friendship we developed, in such a short time. To Sophie, Kristina, Sara, Simona, Katarina, Max and Davide. To past members of limbo for introducing us to the uncertainty that we would learn to face in the future: Federico, Ritwik, Yair and Eugenio. To Kiana for the countless laughs and crazy experiences, Arash for the countless debates, Nader, and Milad, truly one of a

kind, whose presence in Trieste have made me feel close to home. To Massimo and Georgette, for our shared music tastes, and Carlotta, who bore with being my officemate. To Perrine and Alissa, for the evenings out on warm summer Trieste nights.

I would like to thank my friends of a previous life: Kamelia, Parastoo, Saghar, Sima, Maryam, Sakina, Mikha, Pierre, Sylvestre and Robert with whom I have not always been in touch, but that always remain in my heart and my thoughts.

Last but not least, my greatest gratitude goes to Alberto, who has had to bear with me during these past years; for his kindness and generosity, his intellectual and emotional support, without him the past years would have been far less enjoyable.

# 1

# Introduction

One of the most fascinating aspects of the human brain is its ability to ascribe to and recognize meaning in objects, events and to more generally make sense of the world. Semantic memory, comprising our acquired knowledge about the world, has been argued to reside in the cortex, the newest addition of evolution to our brain. The cortex has been found to function according to principles that are functionally different from older memory systems, such as the hippocampus. Both systems, comprising long term memory, have been proposed to make independent contributions to memory storage and retrieval.

According to David Marr [Marr et al., 1991], the hippocampal network, the final constituent of the mediotemporal lobe memory stream, operates as a content-addressable memory. The CA3 region of the hippocampus has been described as an auto-associator that processes the retrieval of stored information from a partial cue [McNaughton and Morris, 1987, Rolls, 1989]. In this picture, representations pertaining to new information to be stored in the CA3 are equivalently arbitrary. Mammals have been hypothesized to have evolved circuits involving the dentate gyrus, specifically to assign to a new item, represented as a distributed pattern of activity of neocortical pyramidal neurons, a new arbitrary representation composed of CA3 pyramidal neurons [Treves and Rolls, 1992, Leutgeb et al., 2007]. Therefore, any two items, regardless of their similarity in content and degree of overlap in their neocortical representations, will tend to have a similar level of overlap in CA3 [Leutgeb et al., 2004].

Contrary to memory representations in the hippocampus, representations in the cortex tend to be more distributed and operate over a larger area comprising different areas, with a *graded* usage of modality specific to highly multimodal association areas [Braitenberg and Schüz, 1991]. A memory item is then composed of subparts, or *features*, each associated with a non-arbitrary, stable representation within a cortical area. Items with overlapping content are therefore represented in the cortex as spatio-temporal activity patterns with a corresponding strength of correlation. The intrinsic consistency of memories sharing overlapping anatomical substrates, in contrast to the arbitrariness of hippocampal memories, underlies the semantic structure of cortical, as opposed to hippocampal, representations. Meaning is then reflected in the complex set of association and abstraction or *correlation* between cortical representations (Fig. 1.1).

However, it is the statistical independence of memory patterns that has made available most of the mathematically sophisticated analyses. If they have been successfully used to describe the CA3 circuit for example, they are irrelevant to semantic memory, which has in the shared structure between memories its raison d'être. Some progress in this direction has been made by the fruitful body of research instigated by the Hopfield model. Though initially featuring uncorrelated patterns, extensions for the storage of correlated patterns were eventually made. One of the earliest attempts to introduce correlations was through an algorithm that arranged patterns on an ultrametric tree [Parga and Virasoro, 1987, Gutfreund, 1988]. It was found that storing correlated patterns reduces the storage capacity, i.e. the maximal number of activity patterns that can be stored and retrieved, when a standard Hebbian plasticity learning rule is used; however, making modifications to the learning rule, the storage capacity was found to be restored to some extent. Further attempts include, for example, the study of the retrieval properties of the Hopfield network when the memorized patterns are statistically correlated in pairs [Tamarit and Curado, 1991]. In this study, there is a finite correlation between the memories of each pair, but memories of different pairs are uncorrelated. In particular, they find two retrieval regimes: for low temperature, the network retrieves the stored patterns, while for higher temperature the network is able to recognize pairs, but it is not able to distinguish between its two patterns. Other methods involve the generation of memories through Markov chains and have been used, for instance, to study variants of the Hopfield model [Löwe, 1998].

However, are such simple schemes suitable statistical models of the organization of semantic memory? The object of most of the earlier theoretical studies (e.g ultrametrically organized patterns) were oversimplified models which fail to capture many of the relevant features of semantic memory. With a somewhat opposite approach, the parallel-distributed processing (PDP) framework, with a largely data driven basis, focused on computer simulations that could qualitatively reproduce results in agreement with patterns of deficits seen in the neuropsychological literature; no framework, however, has been proposed for more theoretical questions of a quantitative scope [Rumelhart et al., 1986, Farah and McClelland, 1991, Rogers et al., 2004, Plaut, 1995]. As a critical step in this direction, in this thesis, we formulate an algorithm designed to capture some of the key elements in the organization of semantic memory, before turning to the network storage of such semantically correlated patterns.

The analytical tools allowing for a complete analysis have been applied to fully connected or else partially connected networks, in which the average connectivity between the units vanishes. These models have been thoroughly analyzed and scaling relations have been found for the storage capacity as a function of the mean connectivity and the sparsity of the network. Remarkably, such scaling relation holds, when activity is sparse, for both limit cases of full connectivity and extremely sparse connectivity. Does it mean that it holds also for any connectivity in between, including realistic models of cortical connectivity?

From the point of view of plausibility, such studies of randomly wired networks fall short of describing some features of the anatomy of cortical connectivity. For example, it has been shown [Hellwig, 2000] that in layers II and III of mouse visual cortex the probability of connection falls from $50-80$ percent for directly adjacent neurons to $0-15$ percent at a distance of 500 micrometers. Building on such observations, the properties of an autoassociative network of threshold-linear units whose synaptic connectivity is spatially structured has also been investigated [Roudi and Treves, 2004].

Other studies however, have evidenced that at a larger scale, cortical connectivity is not randomly distributed, not even after allowing for a distance-dependent parameter. For example, it has been shown that in the prefrontal cortex of monkeys, patches of a few hundred microns make connections to and from other discrete patches of cortex of the same size [Pucak et al., 1996]. A patch is connected to about $15-20$ other patches in its proximity via grey matter connections, and to at least $15-20$ more distant patches connected via white matter connections.

Variant models of associative memory networks that implement this separation of scale between dense local connectivity and sparse long-range connectivity have been studied [O'Kane and Treves, 1992a, O'Kane and Treves, 1992b, Mari and Treves, 1998, Dubreuil and Brunel, 2016]. This study is in line with such an approach, in that it aims at describing each patch of cortex, a functional voxel of a few $mm^2$, comprising some roughly $10^5$ neurons, as one local network interacting through the B system, whose activity is coarsely subsumed into a Potts unit. The Potts network, aimed at describing the cortex, or a large part of it, is comprised of $N$ such units, constituting the A-system. We refer to [Naim et al., 2017] for a detailed analysis of the approximate thermodynamic and dynamic equivalence of the full multi-modular model and the Potts network.

The organization of this thesis is as follows. We will start by introducing the Potts network in Chap. 2. In Chap. 3, we study the storage capacity of the Potts network for different models of connectivity, a first step we take towards cortical plausibility. In Chap. 4, after a brief overview of the principal theories of semantic organization and some of the empirical evidence, we will introduce an original algorithm, designed to capture some of the resulting key concepts, in order to produce patterns of activity relevant to semantic memory. In Chap. 5, we will study the storage capacity of the Potts network with patterns correlated by way of our algorithm. Finally, in Chap. 6, we begin to assess how the correlated nature of semantic memories impacts on latching dynamics, which can be studied in Potts networks that include models of adaptation and inhibition; this brings us beyond simple associative cued retrieval, into more complex thought processes and into language.

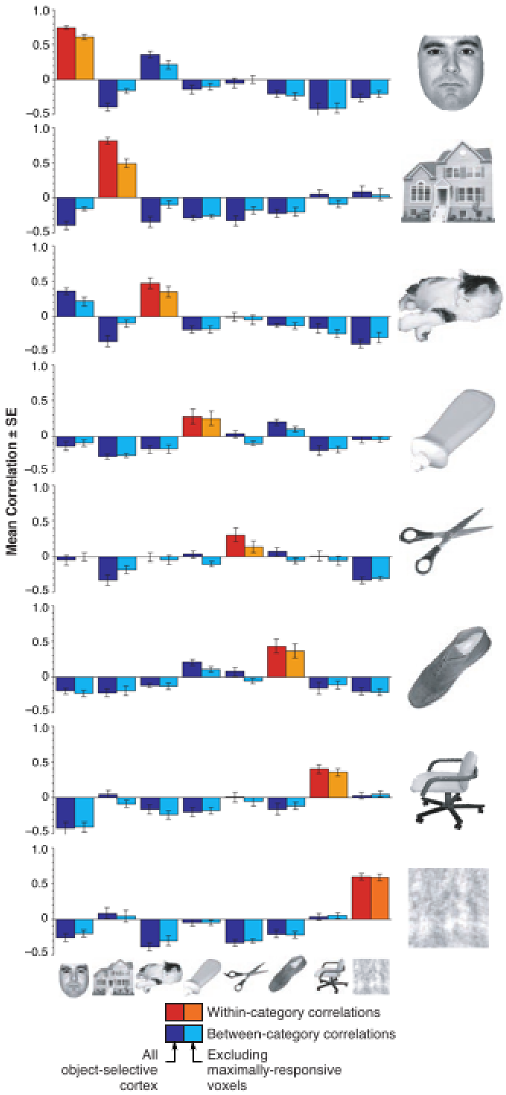


**Fig. 1.1: Correlation between neural representations as measured by fMRI.** Figure and caption from [Haxby et al., 2001]. The functional architecture of the object vision pathway in the human brain was investigated using functional magnetic resonance imaging to measure patterns of response in ventral temporal cortex while subjects viewed sets of faces, cats, man-made objects, and nonsense pictures. A distinct pattern of response was found for each stimulus set. The distinctiveness of the response to a given set was not due simply to the regions that responded maximally to that set, because the set being viewed also could be identified on the basis of the pattern of response when those regions were excluded from the analysis. Patterns of response that discriminated among all sets were found even within cortical regions that responded maximally to only one set. These results indicate that the representations of faces and objects in ventral temporal cortex are widely distributed and overlapping. Reported in the figure: mean within-set and between-set correlations (+−SE) between patterns of response across all subjects for all ventral temporal object-selective cortex (red and dark blue) and for ventral temporal cortex excluding the cortex that responded maximally to either of two sets of objects being compared (orange and light blue).

# 2

# Potts networks for semantic memory

## 2.1. The cerebral cortex

The cortex is the outer layer of the brain characterized by a folded shape, displaying ridges (gyri) and fissures (sulci), providing a greater surface area in the confined volume of the cranium. The cerebral cortex is entirely made of gray matter, consisting mainly of neuron bodies contrasting with the underlying white matter, which consists mainly of myelinated axons traveling to and from the cortex. One of the organizational features of the cortex is the division, parallel to the surface, into functional areas that serve various sensory, motor and cognitive functions. Another is the subdivision, perpendicular to the surface, into several layers that organize the input and output connectivity of the neurons. The cerebral cortex has been hypothesized to play key roles in memory, attention, perception, cognition, awareness, thought and language, among others.

A detailed study of the statistical neuroanatomy of the cortex by Braitenberg (see Fig. 2.1) has resulted in some key structural observations that can be drawn on. In the human brain, the number of neurons exceeds the number of input channels by at least 3 orders of magnitude, suggesting that the information capacity of the cortex is not related in any simple way to the capacity of the sensory channels. It may be that the number of neurons have been optimized to encode the number of situations and concepts the mammal comes to deal with in its lifetime.

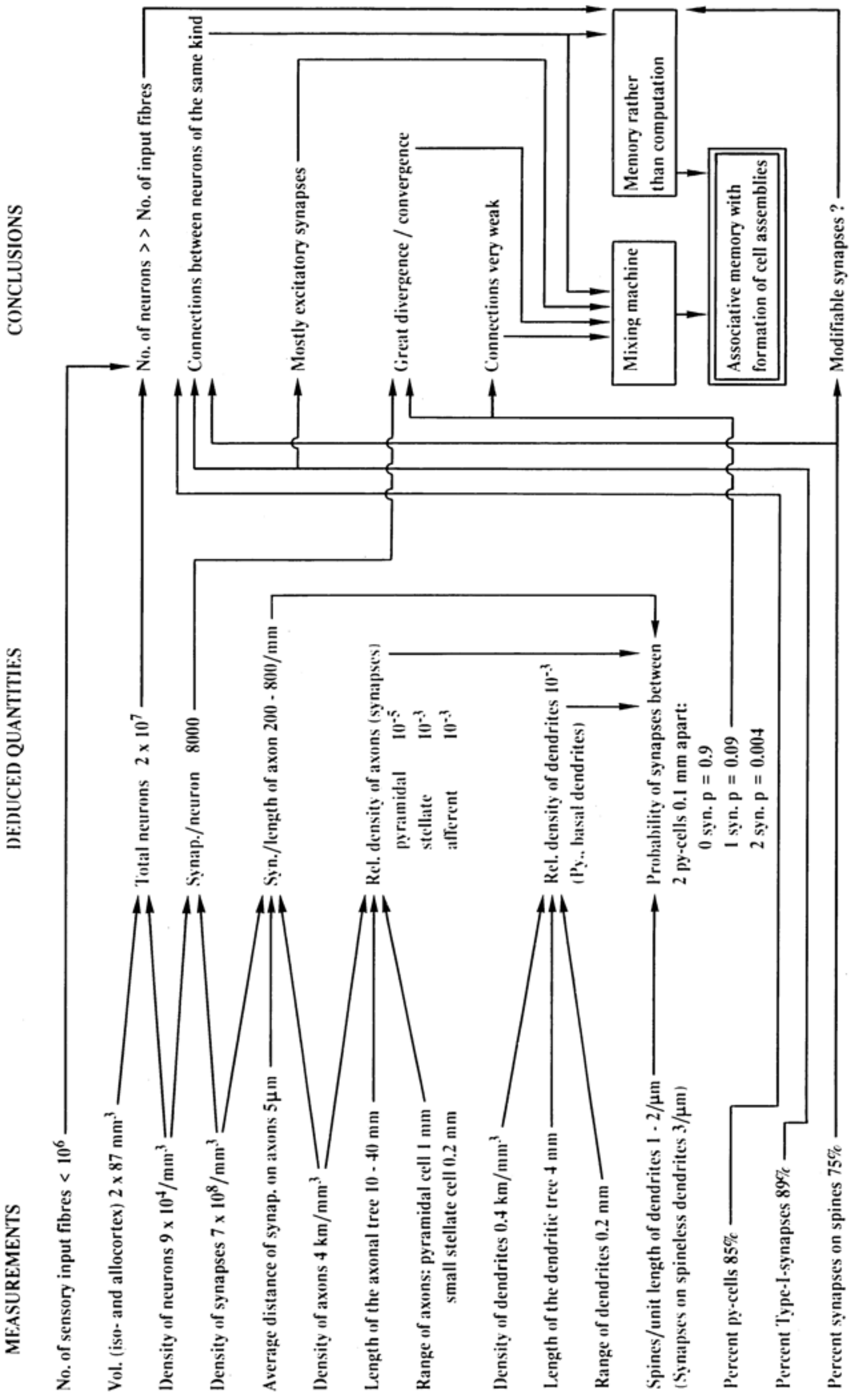

MEASUREMENTS
No. of sensory input fibres < $10^6$
Vol. (iso- and allocortex) 2 x 87 $mm^3$
Density of neurons 9 x $10^4/mm^3$
Density of synapses 7 x $10^8/mm^3$
Average distance of synap. on axons 5μm
Density of axons 4 $km/mm^3$
Length of the axonal tree 10 - 40 mm
Range of axons: pyramidal cell 1 mm
small stellate cell 0.2 mm
Density of dendrites 0.4 $km/mm^3$
Length of the dendritic tree 4 mm
Range of dendrites 0.2 mm
Spines/unit length of dendrites 1 - 2/μm
(Synapses on spineless dendrites 3/μm)
Percent py-cells 85%
Percent Type-I-synapses 89%
Percent synapses on spines 75%
DEDUCED QUANTITIES
Total neurons 2 x $10^7$
Synap./neuron 8000
Syn./length of axon 200 - 800/mm
Rel. density of axons (synapses)
pyramidal $10^{-5}$
stellate $10^{-3}$
afferent $10^{-3}$
Rel. density of dendrites $10^{-3}$
(Py., basal dendrites)
Probability of synapses between
2 py-cells 0.1 mm apart:
0 syn. p = 0.9
1 syn. p = 0.09
2 syn. p = 0.004
CONCLUSIONS
No. of neurons > > No. of input fibres
Connections between neurons of the same kind
Mostly excitatory synapses
Great divergence / convergence
Connections very weak
Modifiable synapses ?
Mixing machine
Memory rather than computation
Associative memory with formation of cell assemblies

**Fig. 2.1: Statistical neuroanatomy of the cortex by Braitenberg.** From [Braitenberg and Schüz, 1991]: Schematic summary of the structure of the reasoning. "The quantities measured are shown on the left, the theoretical conclusions on the right and the deductions which led to them in between. [...] the idea which fits the cortical network most admirably is that of an associative memory."

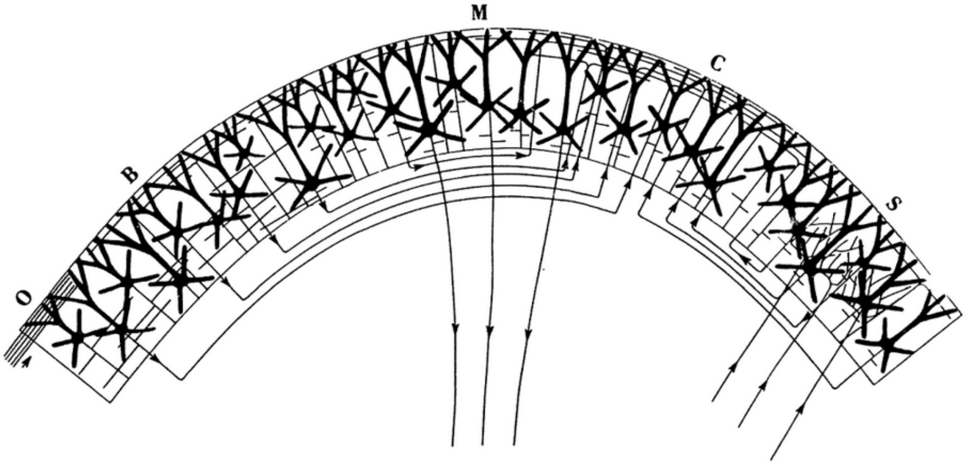


**Fig. 2.2: The A and B systems of Braitenberg.** Figure and caption from [Anderson, 1997]. An outline of some major cells and their pattern of innervation in the "skeleton cortex". There are some exceptional regions (O, the olfactory cortex, M, the motor cortex and S, the primary sensory areas) but the general case is the cortex with long-range, ametric subcortical connections and with short-range, metrically dependent, intracortical connections (A and B system in Braitenberg's terminology). B, divergence of the axons projecting into the white matter; C, convergence of fibers from the whole cortex onto a small region (A-system).

Moreover, the vast majority of neurons are interneurons, in the sense that they are neither directly under the influence of sensory input, nor involved in producing motor output. Even in layers where the input enters most prominently, such as in the fourth layer of sensory regions, "at least 4/5 of the afferent synapses of cortical neurons do not have an input fiber but presumably another cortical neuron as their presynaptic element". This suggests that the majority of information transmission occurs *between* neurons that communicate with one another, much more than they do with the outside world.

In addition, the global organization of the cortex provides further evidence against serial processing. The architecture of the cortex is that of a three-dimensional sheet of neurons in which a differential organization exists only in one direction, that along which cortical layers are stacked on top of each other. The input areas are arranged in parallel with the output layers, and not along the different layers. In the words of Braitenberg "the flow of information from the sensory to the motor areas, if there is such a thing, traverses the cortex in a direction of the cortical plane which is not distinguished by any special "wiring"".

Synapses between pyramidal cells, that constitute the majority of all synapses in the cortex, are excitatory and are made through dendritic spines. It has been argued that modifiable excitatory synapses constitute the main mechanism for learning, in which cell assemblies are established through the coincident activity of neurons strengthening their synaptic coupling: Hebbian learning. Evidence for this comes from STDP [Markram et al., 2012] as well as more recently electron microscopy of synapses [1].

Pyramidal cells, being the majority of cortical neurons, can be seen as the "skeleton cortex": with their long axons and high variability in size, they have been hypothesized to be the major neurons connecting different regions of the cortex together. A key feature of pyramidal cells is that their dendritic tree branches in two directions: basal dendrites collect input mainly from local axon collaterals, while apical dendrites, branching into the upper layers of the cortex, receive input largely from long-range cortico-cortical connections coming from other cortical regions.

Braitenberg and Schuz have elegantly synthesized this dual local and global characteristic of the cortex in terms of the A and B systems (referring to apical and basal dendrites) [Braitenberg and Schüz, 1991]. They suggest regarding the whole cortex as a memory machine, in which the B-systems encode a set of memories as local attractors and the A-system encodes global attractors, by virtue of long-range connections (see Fig. 2.2).

## 2.2. The cortex as a Potts network?

How can we study the statistical properties of model networks, organized with distinct local and global connections? Early attempts have found it challenging to deal simultaneously with the two levels of organization [O'Kane and Treves, 1992b]. In a recent contribution, however, [Naim et al., 2017], we elaborate on the correspondence between a multi-modular neural network and a coarse grained Potts network [Kanter, 1988], by grounding the Hamiltonian of the Potts network in the multi-modular one [2].

In the multi-modular model, units are taken to be threshold-linear, and they are fully connected

[1] Though multiple contacts between a pair of neurons is rare, sometimes a presynaptic neuron axon makes contact with multiple spines of the same postsynaptic neuron. Since spine head volumes belonging to a single dendritic tree have been found to be highly variable, the significant coincidence in spine head volumes of these spines residing on the same dendritic tree has been argued to provide evidence of co-activity [Bartol Jr et al., 2015].

[2] The Potts network is inspired from the standard Potts model of statistical physics. In the standard Potts model, states that can take $q$ values, corresponding to the vertices of a $q-1$ dimensional simplex. It differs from $n$-vector models, in which the state of each unit can take continuous values lying on a hypersphere – notable examples include the $XY$ model ($n=2$) and the Heisenberg model ($n=3$) – and reduces to the Ising model for $q=2$.

within a module, with Hebbian [Hebb, 2005] 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 traveled through the white matter. The Potts states relate then 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 roughly correspond, after suitable assumptions about inhibition, to the tensorial couplings between Potts units in the Potts Hamiltonian.

The Potts neural network [Kanter, 1988,Bollé et al., 1991,Bollé et al., 1992,Bollé et al., 1993], can be therefore regarded as an effective model of a global memory network comprised of local modules. Each of these modules is subsumed into a Potts unit. In the simplest version of the model, each Potts unit, which may be taken to represent the activity of a small patch of cortex, can be in $S$ equivalent states, each representing the local attractors of the full model. A quiescent state can be added to the $S$ active states to represent a situation of no local retrieval of any specific cell assembly. This model has been studied analytically with replica tools in [Kanter, 1988], where the author finds an $S^2$ behavior of the storage capacity for low values of $S$. Later, a logarithmic correction was derived in the high $S$ limit, such that the maximal number of patterns that can be stored and retrieved relative to the number of connections, $\alpha_c \approx \ln(S)^{-1}S^2$ [Kropff and Treves, 2005]. A slightly more realistic model, allowing for a quiescent state and therefore globally sparse representations, shows a higher capacity, scaling as $\alpha_c \approx \{a \ln(1/(a/S))\}^{-1}S^2$ in the $a/S \ll 1$ limit, where $a$ denotes the sparsity of the global representations [Kropff and Treves, 2005].

However, these analyses have been constrained to the limiting cases of fully connected and highly diluted networks. The latter case, in which the mean connectivity per unit $c_m$ is much smaller than the number of units, has been studied in [Kropff and Treves, 2005] and the storage capacity has been found to scale as $p_c \approx \dfrac{c_m S^2}{a\ln(1/a)}$ for very sparse networks, i.e. $a \ll 1$, in the thermodynamic limit, i.e. $N$, $p \to \infty$ with $p/N$ finite. All these analyses are valid only for independently generated memory patterns, that is, in the absence of any semantic structure.

In Chap. 3, in order to complement this previous work, we study the storage capacity of the Potts network with general "diluted" connectivity, through the self-consistent signal to noise analysis. This technique was first applied to the Hopfield network in cases where the replica technique was not applicable [Shiino and Fukai, 1993] and later to networks with threshold-linear units [Roudi and Treves, 2004], yielding results which reduce to the replica ones in the fully connected and highly diluted limits.

The aim of such a calculation is two-fold. The first is to obtain a generalization of the mean-field equations yielding the storage capacity for arbitrary diluted connectivity, recovering, at the fully-connected and highly-diluted limits, the same equations as those obtained in [Kropff and Treves, 2005]. The second is to later apply this same analysis to the case of *correlated patterns* that we report in Chap. 5. In the next section, we first introduce and define the model.

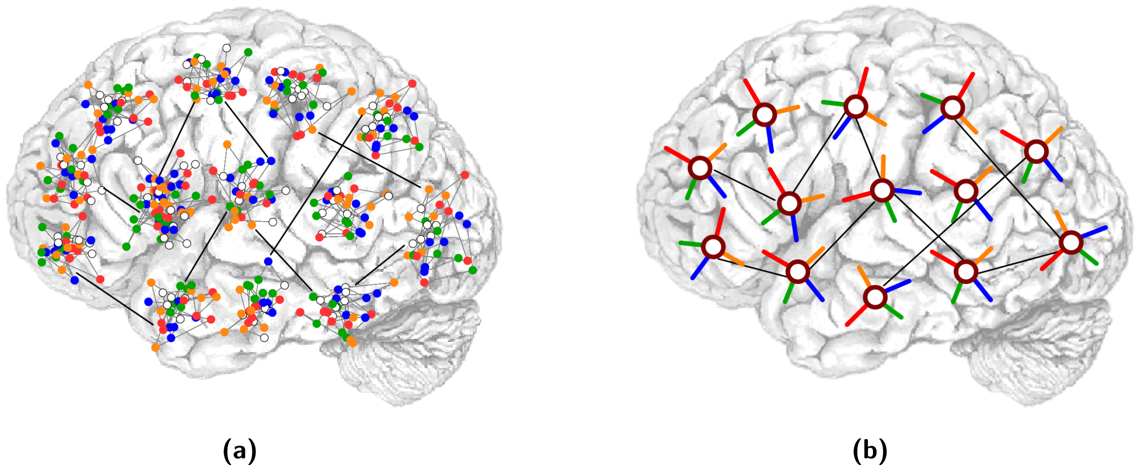


**Fig. 2.3: From local networks to Potts units.** The Potts network, here intended as a model of semantic memory, is a coarse description of the cortex in terms of local patches of dense connectivity, which store activity patterns corresponding to local attractors **(a)**. Each patch is a small local network characterized by high connectivity; *diluted* connections are instead present between units of different patches. The configuration of the individual patch is assmued to converge to a local attractor, synthetically captured by a Potts state. Each Potts unit, depicted in **(b)** can be in any of $S$ states, where green, orange, blue and red represent the active states ($S = 4$). The white circle at the center corresponds to the quiescent state, aimed at capturing a situation of no retrieval of the underlying local network.

## 2.3. Introducing the Potts network

The Potts neural network is a generalization of Hopfield's binary autoassociative network [Hopfield, 1982]. A Potts unit can be either in the *quiescent* state or else in one of the $S$ equivalent *active* states. By convention, we label these states with numbers from 0 to $S$, where $k = 0$ indicates the quiescent state and $k = 1...S$ the active ones, representing the possible local attractors (see Fig. (2.3)). Due to stochastic fluctuations, a unit can be, with a non-vanishing probability, in any of the $S+1$ states, so that the *activity* of unit $i$ is a distribution over $k = 0...S$, denoted by $\sigma_i^k$. By network state, or *configuration*, we refer to the collection of local states assigned to all units, $\{\sigma_i^k\}$, where $i \in \{1, \dots N\}$, $N$ being the number of units in the network.

*Couplings* between states of distinct units are defined, which are denoted by $J_{ij}$: they represent the strength with which connected units connected influence each other. In the case of the Hopfield network, the couplings $J_{ij}$ are just scalars. In the Potts network, these couplings are matrices $J_{ij}^{kl}$, which contain the strength of the coupling for the pair of units $i$ and $j$ being, respectively, in state $k$ and $l$.

Of crucial importance in the definition of the network model is the *learning rule*, which prescribes how the couplings in the model depend on a given training data set. In the model that is dealt with in this thesis, the training data set consists of a certain number $p$ of network configurations, denoted by $\{\bar{\xi}^\mu\}_{\mu=1}^p$. We refer to these configurations as *patterns*.

The way the patterns $\bar{\xi}^\mu$ are generated, i.e. their probability distribution, has effects on the "retrieval properties" of the network, i.e. the ability to retrieve with good accuracy one of the training patterns, if this is partially cued. A quantitative measure of this ability of the network is the *storage capacity*, the number of patterns the network is able to store and retrieve, relative to the number of synaptic connections per unit.

The learning rule according to which the patterns are used to build the synaptic connections between units is a Potts-adapted version of Hebbian learning:

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

where the factor $c_{ij}$ denotes the of the $(i, j)$-th entry of the adjacency matrix of the connectivity (graph), being equal to 1 if an edge exists from $j$ to $i$ and 0 otherwise. The constant $c_m$ is the average degree of this graph, i.e. the average number of connections at a given random node, so that one has $\langle c_{ij} \rangle = c_m/N$. The symbol $\delta$ here indicates the Kronecker $\delta$-function, which evaluates to 1 when the two indices are equal and 0 if they are different. The subtraction of the mean activity by state, $a/S$, ensures a higher storage capacity, as initially shown for the Hopfield network in [Tsodyks and Feigel'Man, 1988] and for the Potts neural network in [Kropff and Treves, 2005].

The fully connected network, in which $c_{ij} = 1$ for all pairs $(i, j)$ is the one which allows for a full-fledged analytic approach, by means of techniques borrowed from spin glass physics [Mézard et al., 1987]. It has been shown, as reviewed in [Amit, 1992], that such connectivity ensures that each of these configurations, if they are not too many, becomes a stable configuration, or an attractor of the energy function

$$H = -\frac{1}{2} \sum_{i,j \neq i}^{N} \sum_{k,l=1}^{S} J_{ij}^{kl} \sigma_i^k \sigma_j^l + U \sum_{i}^{N} \sum_{k}^{S} \sigma_i^k \,. \tag{2.2}$$

where $\sigma_i^k$ is referred to as the activity of unit $i$ in state $k$. The variable $\sigma_i^k$ can be interpreted as the probability with which the local network, synthesized into the Potts unit $i$, finds itself in the attractor $k$. This probability is given by the Boltzmann distribution with inverse temperature $\beta$:

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{e^{\beta U} + \sum\limits_{l=1}^{S} e^{\beta h_i^l}} \,, \tag{2.3}$$

where $h_i^k$, referred to as the *field* received by unit $i$ in state $k$, is determined by the activity of all the Potts units in a way that will specified in later sections. From Eq. (2.3), it follows that $\sum_{k=0}^{S} \sigma_i^k = 1$ at all times.

It is possible to define a stochastic dynamics on the Potts network. This is introduced by asynchronous updating of the units according to Eq. (2.3) [Russo and Treves, 2012] and is referred to as *retrieval dynamics*. Initializing the network in a partially degraded or incomplete pattern, this dynamics can drive the configuration of the Potts network into one of the global attractors. If this attractor corresponds to the pattern that has been cued, we then have pattern completion (retrieval). In the rest of the thesis, we will be concerned with the limit $\beta \to \infty$, in which the memory about the cued pattern is retained with maximum fidelity during the retrieval dynamics.

# 3

# Storage capacity of the Potts network with uncorrelated patterns

In this chapter we study the storage capacity of the Potts network, for the case of uncorrelated and identically distributed training patterns. Formally, this means that the $p$ patterns $\{\bar{\xi}^\mu\}$ are generated according to a probability distribution which is factorized into $p$ identical distributions of the individual patterns. With little abuse of notation, this writes

$$P(\bar{\xi}^1 \dots \bar{\xi}^p) = P(\bar{\xi}^1) \cdot \ldots \cdot P(\bar{\xi}^p) \ . \tag{3.1}$$

In turn, units in each pattern are also independent and identically distributed:

$$P(\bar{\xi}^\mu) \equiv P(\xi_1^\mu \dots \xi_N^\mu) = P(\xi_1^\mu) \cdot \ldots \cdot P(\xi_N^\mu) \ . \tag{3.2}$$

Moreover, every unit in each pattern is taken to be in the inactive one with probability $1 - a$, the remaining probability being shared uniformly by the $S$ active states.

$$\begin{cases} P\left(\xi_i^\mu = 0\right) = 1 - a \\ P\left(\xi_i^\mu = k\right) = \tilde{a} \equiv a/S \end{cases} \tag{3.3}$$

The probability of a unit being in any of the active states, $a$, is called the *sparsity* of the patterns. The sparsity obviously defines the average relative number of units being active in a given pattern.

## 3.1. Fully connected network

In this section, we report the main results regarding the storage capacity of the Potts network with full connectivity using the classic replica method. The full derivation can be found in [Naim et al., 2017] and [Kropff and Treves, 2005]. Let us consider the Hamiltonian defined in Eq. (2.2). The free energy writes

$$f = -\frac{1}{\beta}\Big\langle \ln Z \Big\rangle . \tag{3.4}$$

where $Z$ is the partition function of the system, and $\beta$ is the inverse temperature. The main idea behind the replica technique is the identity

$$\ln Z = \lim_{n\to 0} \frac{\Big\langle Z^n \Big\rangle - 1}{n} \tag{3.5}$$

known as the "replica trick", where the complicated problem of calculating the average of a logarithm of a disordered quantity can be simplified by applying this identity, reducing the problem to calculating $\langle Z^n \rangle$, where $n$ is assumed to be an integer.

Applying the replica technique ( [Sherrington and Kirkpatrick, 1975, Mézard et al., 1987]), following [Amit et al., 1985, Amit et al., 1987, Kanter, 1988], the free energy of $N$ Potts units in replica theory writes

$$f = -\frac{1}{\beta} \lim_{n\to 0} \lim_{N\to\infty} \frac{\Big\langle Z^n \Big\rangle - 1}{Nn} , \tag{3.6}$$

where $\langle \cdot \rangle$ is an average over the quenched disorder (represented by the randomness of the patterns defined by Eq. (3.3)). The replica symmetric free energy, under the assumption that the replicas are identical as in [Sherrington and Kirkpatrick, 1975], can be written as

$$\begin{aligned} f &= \frac{a\left(1-\tilde{a}\right)}{2}m^2 + \frac{\alpha}{2\beta}\left[\ln\left(a\left(1-\tilde{a}\right)\right) + \ln\left(1-\tilde{a}C\right) - \frac{\beta\tilde{a}q}{\left(1-\tilde{a}C\right)}\right] + \\ &+ \frac{\alpha\beta\tilde{a}^2}{2}\left(\tilde{r}\tilde{q} - rq\right) + \tilde{a}\tilde{q}\left[\frac{\alpha}{2} + SU\right] + \\ &- \frac{1}{\beta}\left\langle \int D\mathbf{z} \ln\left(1 + \sum_{l(\neq 0)} \exp\left[\beta\mathcal{H}_l^\xi\right]\right)\right\rangle \end{aligned} \tag{3.7}$$

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

$$\mathcal{H}_l^\xi = mv_{\xi l} - \frac{\alpha a\beta\left(r-\tilde{r}\right)}{2S^2}\left(1-\delta_{l0}\right) + \sum_{k=1}^{S}\sqrt{\frac{\alpha r P_k}{S\left(1-\tilde{a}\right)}}z_k v_{kl}\,. \tag{3.8}$$

$C$ 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 $\xi$ (note that $\mathcal{H}_0^\xi = 0$). No such interpretation can be given to $C$: it 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 Sect. 3.3, 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, $m, q$ and $r$:

$$\begin{aligned} m &= \frac{1}{a(1-\tilde{a})}\left\langle \int D^S z \sum_{l\neq 0} v_{\xi l}\frac{1}{1+\sum_{n\neq l}\exp\left[\beta\left(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right)\right]}\right\rangle \\ &\to \frac{1}{a\left(1-\tilde{a}\right)}\sum_{l\neq 0}\left\langle \int D^S z\, v_{\xi l}\prod_{n\neq l}\Theta\left[\mathcal{H}_l^\xi - \mathcal{H}_n^\xi\right]\right\rangle \end{aligned} \tag{3.9}$$

$$q \to \tilde{q} = \frac{1}{a}\sum_{l\neq 0}\left\langle \int D^S z \prod_{n\neq l}\Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\right\rangle \tag{3.10}$$

$$C = \frac{1}{\tilde{a}^2\sqrt{\alpha r}}\sum_{l\neq 0}\sum_{k}\left\langle \int D^S z\sqrt{\frac{P_k}{S\left(1-\tilde{a}\right)}}v_{kl}z_k\prod_{n\neq l}\Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\right\rangle \tag{3.11}$$

$$\tilde{r} \to r = \frac{q}{(1-\tilde{a}C)^2} \tag{3.12}$$

$$\beta\,(r-\tilde{r}) = 2\left(U\frac{S^2}{a\alpha} - \frac{C}{1-\tilde{a}C}\right) \tag{3.13}$$

where

$$\int Dz = \int dz \frac{\exp\left(-z^2/2\right)}{\sqrt{2\pi}}\,. \tag{3.14}$$

The differences between $r$ and $\tilde{r}$ and between $q$ and $\tilde{q}$ are of order $1/\beta$. From the last equation it can be seen that the threshold $U$ has the effect of changing the sign of $(r-\tilde{r})$ such that $\alpha \sim S^2/a$ with the variables $C$, $r$ and $\tilde{r}$ of order 1 with respect to $a$ and $S$.

The above averages can be calculated analytically and we refer to [Kropff and Treves, 2005] and [Naim et al., 2017] for their expressions. The resulting integrals are complicated but the highly sparse limit case $a \ll 1$ allows for an approximate simple expression of the storage capacity to be obtained. It is found that

$$\alpha_c \approx \frac{S^2}{4a\ln\left(\frac{2}{\tilde{a}}\sqrt{\ln\frac{1}{\tilde{a}}}\right)}\,. \tag{3.15}$$

While the $S^2/a$ behavior of the capacity can be intuitively understood as having its origin in the scaling of the number of synapses Eq. (2.1) as well as sparse coding, the logarithmic corrections may be understood from an information storage perspective, as reported in [Kropff and Treves, 2005]. The upper bound on the information carried by $p$ patterns comprised of $N$ units each can be computed using Shannon's information and Eq. (3.3)

$$I \leq pN\left\{-(1-a)\log_2(1-a) + a\log_2(S/a)\right\}\,, \tag{3.16}$$

In the limit of $a \ll 1$, the first term can be neglected. On the other hand, the number of synaptic variables is $N \cdot c_m \cdot S^2$, such that the amount of information per synapse can be approximated as

$$I_{syn} \leq \frac{\alpha a \log_2(S/a)}{S^2}\,, \tag{3.17}$$

This is consistent with the idea that the maximal information that can be stored in an auto-associative network is at most a fraction of a bit per synapse, and we retrieve the logarithmic correction given by Eq. (3.15).

## 3.2. Highly diluted network

In the previous section we summarized the steps taken to obtain the set of equations that determine the storage capacity of the fully connected 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 denote with $\lambda = \langle c_{ij} \rangle = c_m/N$ the dilution parameter. In general, $c_{ij}$ can be different from $c_{ji}$, leading to asymmetry in the connections between units.

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

$$P(c_{ij}, c_{ji}) = P(c_{ij})P(c_{ji}) \tag{3.18}$$

with

$$P(c_{ij}) = \lambda\delta(c_{ij} - 1) + (1 - \lambda)\delta(c_{ij})\,. \tag{3.19}$$

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

$$P(c_{ij}, c_{ji}) = \lambda\delta(c_{ij} - 1)\delta(c_{ji} - 1) + (1 - \lambda)\delta(c_{ij})\delta(c_{ji})\,. \tag{3.20}$$

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

$$P(c_{ij}^{kl}) = \lambda\delta(c_{ij}^{kl} - 1) + (1 - \lambda)\delta(c_{ij}^{kl})\,; \tag{3.21}$$

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 is the simplest to treat analytically. RD and SD are known in the literature as Erdos-Renyi graphs, respectively, directed and undirected. Many properties are known about such random graph models [Erdös and Rényi, 1960, Engel et al., 2004]. It is known that for $\lambda$ below a critical value, essentially all connected components of the graph are trees, while for $\lambda$ above this critical value, loops are present. In particular, a graph with $c_m < \log(N)$ (with $N \to \infty$) almost surely contains isolated vertices and is disconnected, while with $c_m > \log(N)$ it is almost surely 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 intermediate case of the next section, for which a complete analysis is necessary.

The capacity cannot be analyzed through the replica method, as the symmetry of interactions is a necessary condition for the existence of an energy function, and hence for the application of the thermodynamic formalism. We therefore apply the signal to noise analysis. The local field of unit $i$ in state $k$ writes

$$h_i^k = \sum_j \sum_l c_{ij} J_{ij}^{kl} \sigma_j^l - U\,(1 - \delta_{k,0}) \tag{3.22}$$

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

$$J_{ij}^{kl} = \frac{1}{c_m a(1-\tilde{a})} \sum_\mu v_{\xi_i^\mu k} v_{\xi_j^\mu l}\,. \tag{3.23}$$

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 towards a random direction, usually away from the cued memory pattern. The noise term writes

$$n_i^k = \frac{1}{c_m a(1-\tilde{a})} \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} c_{ij} \sum_l v_{\xi_i^\mu k} v_{\xi_j^\mu l}\, \sigma_j^l\,, \tag{3.24}$$

it 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:

$$\langle n_i^k \rangle_{P(\xi)} = \frac{1}{c_m a(1-\tilde{a})} \sum_{\mu>1}^{p} \sum_{j(\neq i)}^{N} c_{ij} \sum_l \langle v_{\xi_i^\mu,k} \rangle \langle v_{\xi_j^\mu,l} \sigma_j^l \rangle = 0\,. \tag{3.25}$$

The variance of the noise can be written in the following way:

$$\langle (n_i^k)^2 \rangle = \frac{1}{(c_m a(1-\tilde{a}))^2} \sum_{\mu>1}^{p} \sum_{j(\neq i)=1}^{N} \sum_l \sum_{\mu'>1}^{p} \sum_{j'(\neq i)=1}^{N} \sum_{l'} c_{ij} c_{ij'} \langle v_{\xi_i^\mu,k}\, v_{\xi_i^{\mu'},k} \rangle \langle v_{\xi_j^\mu,l}\, v_{\xi_{j'}^{\mu'},l'}\, \sigma_j^l \sigma_{j'}^{l'} \rangle\,, \tag{3.26}$$

where statistical independence between units has been used. For uncorrelated 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 pattern $\mu = 1$ as the one to be retrieved:

$$h_i^k = v_{\xi_i^1 k} m_i^1 + \sum_{\mu>1} v_{\xi_i^\mu k} m_i^\mu - U(1-\delta_{k0}). \tag{3.27}$$

where we define the local overlap $m_i$ as

$$m_i = \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^1 l} \sigma_j \, . \tag{3.28}$$

We now write

$$\sum_{\mu>1} v_{\xi_i^\mu, k} m_i^\mu \equiv \sum_{n=1}^{S} v_{n,k} \, \rho^n \, z_i^n \tag{3.29}$$

where $\rho$ 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. $\rho$ can be computed to find

$$\rho^n = \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} \tag{3.30}$$

where we have defined

$$q = \left\langle \frac{1}{Na} \sum_j \sum_l (\sigma_j^l)^2 \right\rangle . \tag{3.31}$$

The mean field then writes

$$h_i^k = v_{\xi_i^1 k} m + \sum_{n=1}^{S} v_{n,k} \sqrt{\frac{\alpha P_n}{(1-\tilde{a})S} q} z_n - U(1-\delta_{k0}) \, . \tag{3.32}$$

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, Eqs.(3.9) and (3.10). In the highly diluted limit, however, we do not obtain the last equation of the fully connected replica analysis, Eq. (3.12).

The difference between fully connected and diluted cases must vanish in the $\tilde{a} \ll 1$ limit, as shown in [Kropff and Treves, 2005, Derrida et al., 1987]. In this limit we have $x = U/\sqrt{\tilde{\alpha} q}$, $y = m/\sqrt{\tilde{\alpha} q}$, while Eqs.(3.9) and (3.10) remain identical. That is, one recovers the equations of the fully connected case, but with $C = 0$.

## 3.3. Network with partial connectivity

Let us consider now cases where the connectivity is not full but not too sparse either. 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 Eq. (3.27). However, with high enough connectivity one must revise Eq. (3.29): the mean field has to be computed in a more refined way, through a self-consistent method, that we present here.

At the root of the self-consistent signal to noise analysis (SCSNA), [Roudi and Treves, 2004, Kropff, 2009, Shiino and Fukai, 1993], there is the assumption that the noise term can be expressed as the sum of two terms, one proportional to the activity of unit $i$ and the other being a Gaussian random variable,

$$\sum_{\mu>1} v_{\xi_i^\mu,k} m_i^\mu = \gamma_i^k \sigma_i^k + \sum_{n=1}^{S} v_{n,k}\, \rho_i^n z_i^n\,; \tag{3.33}$$

$z_i^n$ are standard Gaussian variables, and $\gamma_i^k$ and $\rho_i^n$ are positive constants to be determined self-consistently. The first term, proportional to $\sigma_i^k$, 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

$$\sigma_i^k = \frac{e^{\beta h_i^k}}{\sum\limits_l e^{\beta h_i^l}} \equiv F^k\Big(\{y_i^l + \gamma_i^l \sigma_i^l\}_l\Big)\,. \tag{3.34}$$

where $y_i^l = v_{\xi_i^1,l} m_i^1 + \sum_n v_{n,l} \rho_i^n z_i^n - U(1-\delta_{l,0})$. The activity $\sigma_i^k$ is then determined self-consistently as the solution of Eq. (3.34):

$$\sigma_i^k = G^k\Big(\{y_i^l\}_l\Big)\,, \tag{3.35}$$

where $G^k$ are functions solving Eq. (3.34) for $\sigma_i^k$. However, Eq. (3.34) 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

$$G^k\Big(\{y_i^l\}_l\Big) \simeq F^k\Big(\{y_i^l + \gamma_i^l \langle \sigma_i^l \rangle\}_l\Big)\ . \tag{3.36}$$

The coefficients in the SCSNA ansatz, Eq. (3.34), $\gamma_i^k = \gamma$ and $\rho_i^k = \rho^k$ are found to be

$$\gamma = \frac{\alpha}{S}\lambda\frac{\Omega/S}{1-\Omega/S} \tag{3.37}$$

and

$$(\rho^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q\Big\{1 + 2\lambda\Psi + \lambda\Psi^2\Big\}\,. \tag{3.38}$$

where $\alpha = p/c_m$ and $\Omega$, $q$ and $\Psi$ are found to be

$$\Omega = \left\langle \frac{1}{N}\sum_j \sum_l \frac{\partial G_j^l}{\partial y^l} \right\rangle, \tag{3.39}$$

$$q = \left\langle \frac{1}{Na}\sum_{j,l} (G_j^l)^2 \right\rangle, \tag{3.40}$$

$$\Psi = \frac{\Omega/S}{1-\Omega/S}\,. \tag{3.41}$$

where $\langle\cdot\rangle$ indicates the average over all patterns. The derivation of Eqs.(3.39)-(3.41) is reported in detail in App. A. The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k}m + \frac{\alpha}{S}\lambda\Psi(1-\delta_{k,0}) + \sum_{n=1}^{S} v_{n,k} z^n \sqrt{\frac{\alpha P_n}{S(1-\tilde{a})} q\Big\{1 + 2\lambda\Psi + \lambda\Psi^2\Big\}} - U(1-\delta_{k,0})\,. \tag{3.42}$$

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

$$m = \frac{1}{a(1-\tilde{a})}\left\langle \int D^S z \sum_{l(\neq 0)} v_{\xi,l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi, \tag{3.43}$$

$$q = \frac{1}{a}\left\langle \int D^S z \sum_{l(\neq 0)} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi, \tag{3.44}$$

$$\Omega = \left\langle \int D^S z \sum_{l(\neq 0)} \sum_k z^k \frac{\partial z^k}{\partial y^l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi) \right\rangle_\xi . \tag{3.45}$$

Note the similarities to the equations (Eqs.(3.9)-(3.11)) 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, $\lambda = 1$, Eqs.(3.43)-(3.44) are apparently equivalent to their corresponding expressions from the replica calculation. With a little algebra, also the expression for $\Omega$ can be shown to be consistent with the replica result. Indeed, from the following identity,

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q (1+\Psi)^2 \, , \tag{3.46}$$

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

$$\rho^2 = \frac{\alpha P_n}{S(1-\tilde{a})} r (1-\tilde{a}C)^2 (1+\Psi)^2 \; . \tag{3.47}$$

By comparing this with Eq. (3.8), the mean field, we get an equivalent expression for $\Psi$,

$$\Psi = \frac{\tilde{a}C}{1-\tilde{a}C} \, . \tag{3.48}$$

From the original definition of $\Psi$ in Eq. (3.41), it follows that the order parameter $C$, obtained through the replica method, is equivalent to $\Omega$, up to a multiplicative constant that is the sparsity:

$$C = \Omega / a \, . \tag{3.49}$$

We can show that Eq. (3.45) coincides with Eq. (3.11). Moreover, by comparing the SCSNA result for $\gamma$ to the replica one, we must have

$$\frac{\alpha}{S} \Psi - U = - \frac{\alpha a \beta (r - \tilde{r})}{2S^2} \tag{3.50}$$

from which

$$\beta (r - \tilde{r}) = 2 \left( U \frac{S^2}{\alpha a} - \frac{C}{1-\tilde{a}C} \right) \, , \tag{3.51}$$

identical to Eq. (3.13).

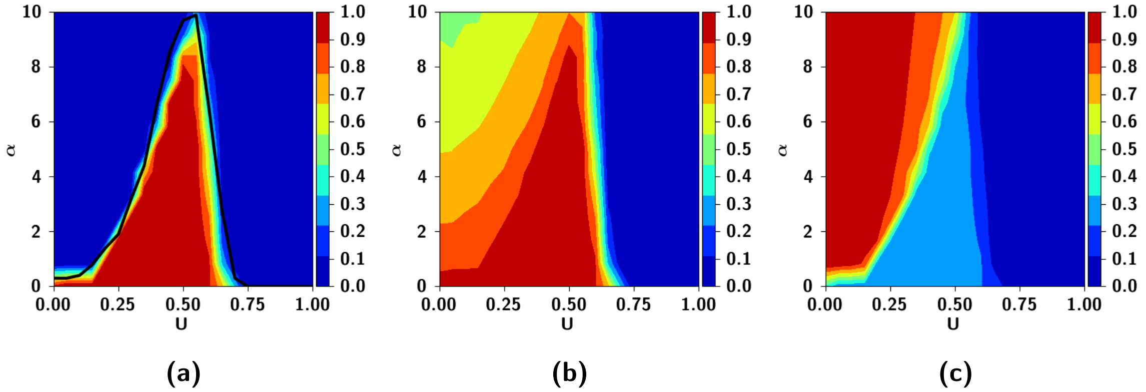


**Fig. 3.1: Storage load and threshold, phase diagrams. (a)** How often a fully connected Potts network retrieves memories, as a function of the threshold $U$ ($x$-axis) and the storage load $p$ ($y$-axis). Color represents the fraction of simulations in which the overlap between the configuration of network and a stored pattern is $\geq 0.9$. The solid lines are obtained by numerical solution of Eqs.(3.9)-(3.11). The optimal threshold, defined as one in which the network has highest capacity is found to be around $U = 0.5$. **(b)** Mean overlap of network configuration with pattern cued. For thresholds below the optimal, the network settles in a configuration with non-zero but low overlap with the pattern cued. This can be explained by **(c)**, where we see the sparsity of network configuration at the end of retrieval dynamics. The suboptimal threshold means that the network goes into "overheating", where many quiescent units become active. This results in the residual overlap with the cued pattern, seen in **(b)**. Network parameters are $N = 1000$, $S = 7$, $a = 0.25$, $\beta = 200$.

## 3.4. Simulation results

### 3.4.1 The effect of network parameters

In Fig. 3.1, we show the comparison between the analytical results presented above and numerical simulations, focusing on the effects of the threshold $U$ and the storage load $\alpha = p/c_m$. In all of the quantities of interest we find good agreement between analytical and numerical results.

In Fig. 3.1a we show the storage capacity. The maximum storage capacity $\alpha_c$ (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 $\gamma_i^k$ of the field, Eq. (3.22), 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 $\mu = 1$.

$$\gamma_i^k \equiv \sqrt{\langle (h_i^k)^2 \rangle - \langle h_i^k \rangle^2} = \sqrt{\frac{(p-1)a}{c_m S^2} + (\delta_{\xi_i^1,k} - \tilde{a})^2 \left( \frac{1}{c_m a} - \frac{1}{N} \right)}. \tag{3.52}$$

The optimal threshold $U_0$ is one that separates the two distributions, the fields of active units and the fields of quiescent units most effectively, such that the minimal number of units in either distribution reach the threshold to go in the wrong state

$$\frac{U_0 - \langle h_i^k|_{\xi_i^1=0} \rangle}{\gamma_i^k|_{\xi_i^1=0}} = -\frac{U_0 - \langle h_i^k|_{\xi_i^1\neq 0} \rangle}{\gamma_i^k|_{\xi_i^1\neq 0}} \tag{3.53}$$

$$U_0 = \frac{\gamma_i^k|_{\xi_i^1=0}}{\gamma_i^k|_{\xi_i^1=0} + \gamma_i^k|_{\xi_i^1\neq 0}} - \frac{a}{S}. \tag{3.54}$$

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. 3.1a. This result is also consistent with [Tsodyks and Feigel'Man, 1988], in which the authors have shown an enhanced storage capacity with low activity levels by the addition of a suitable threshold. Based on a similar signal to noise analysis in [Amit, 1992], this threshold is found to be $1/2 - a$, similar to what we have found. The addition of this threshold ensures that "the state of total inactivity becomes an attractor and has been proposed [Buhmann et al., 1989] as a cognitive identifier of non-recognition, the expectation being that stimuli which are too far from the memorized patterns flow to this unique, special attractor" [Amit, 1992].

In Fig. 3.1b, we plot the phase diagram of the average overlap of the network with the pattern cued. The portion of phase space $(U, p)$ for which this quantity is maximal corresponds perfectly to that for which the fraction of retrievals is also maximal, as shown in Fig. 3.1a. For lower values of the threshold $U$ and higher values of $p$, the average overlap is rather high and not zero. This can be understood by looking at Fig. 3.1c, where we see the phase diagram of the sparsity of the network, at the end of retrieval dynamics. Here we see that for low values of the threshold $U$ and high values of $p$, the network goes into "overheating", where many more units become active than they should be, hence lowering the mean overlap.

The two connectivity limit cases are illustrated in Fig. 3.2. In Fig. 3.2a, the dependence of the storage capacity $\alpha$ on the sparsity $a$ in the fully connected and diluted networks is shown, with

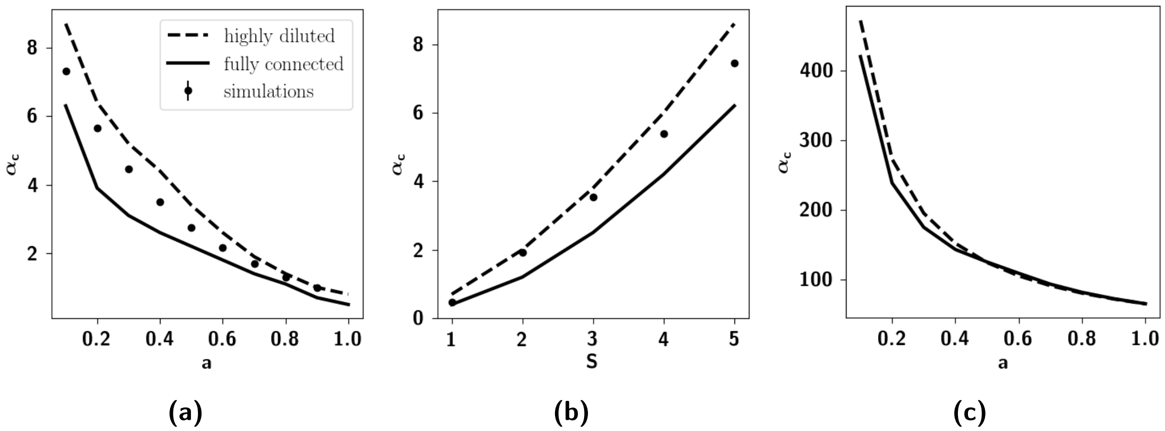


**Fig. 3.2: Storage capacity as a function of the sparsity and the number of Potts states.** **(a)** Storage capacity $\alpha_c$ as a function of the sparsity $a$, for $S = 5$. The capacity is a monotonically decreasing function of the sparsity $a$. Curves are obtained by numerical solution of Eqs.(3.9)-(3.11) for the two limit connectivity cases, while points correspond to simulations of a network with intermediate RD connectivity, set at $c_m/N = 0.1$. One can see that a network with this level of dilution is roughly intermediate between the two limit cases. **(b)** Storage capacity as a function of $S$ with same parameters as in **(a)** and with $a = 0.1$. The storage capacity scales quadratically with $S$. **(c)** $S = 50$, illustrating the highly sparse limit case for which the two curves coalesce. When not explicitly varied, simulations were performed with $N = 2000$, $c_m/N = 0.1$, $S = 5$, $a = 0.1$ and $\beta = 200$.

$U = 0.5$ and $S = 5$. The data points correspond to the simulations for the intermediate case of RD connectivity, with $\lambda = 0.1$. In Fig. 3.2b instead, $S$ is varied and in Fig. 3.2c $S = 50$, 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 Eqs.(3.9)-(3.11). Moreover, the storage capacity curve for the fully connected case in Fig. 3.2a matches very well with Fig. 2 of [Kropff and Treves, 2005]. Diluted curves are always above the fully connected ones in both Fig. 3.2a and Fig. 3.2b, as found in [Kropff and Treves, 2005].

### 3.4.2 The effect of the different connectivity models

In Fig. 3.3 we show the modulation of the storage capacity across the connectivity models introduced earlier. The RD and SDRD networks seem to have almost identical capacity. All models have the same capacity in 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.

Our results can be contrasted to the storage capacity with the same connectivity models (RD and SD; SDRD is not relevant) of the Hopfield model. For the Hopfield model, the effects of SD were investigated and it was found that the capacity decreases monotonically from the value $\simeq 0.4$ for highly diluted to the well-known value of $\alpha_c \simeq 0.14$ for the fully connected network [Sompolinsky, 1986]. In [Derrida et al., 1987], instead, the highly diluted limit of RD was studied and a value of $\alpha_c = 2/\pi \simeq 0.64$ was found. If we plausibly assume that the intermediate RD values interpolate those of the highly diluted $\alpha_c \simeq 0.64$ and fully connected $\alpha_c \simeq 0.14$ limit cases, the Hopfield network then has higher capacity for RD than for SD.

However, it is important to note that the overlap with which the network retrieves at $\alpha_c$, $m_c$, is not the same in the two models (RD and SD). In the highly diluted RD model [Derrida et al., 1987], the authors find that at zero temperature (which is the only case we consider), $m_c$ undergoes a second order phase transition with control parameter $\alpha$ such that $m_c \simeq \sqrt{3(\alpha_c - \alpha)}$: close to $\alpha_c$, $m_c$ is small, and not comparable to the values of $m_c$ for the highly diluted SD model [Sompolinsky, 1986] that we report in the left $y$-axis of Fig. 3.4: at $c_m/N \simeq 0.0024$, $m_c \simeq 0.64$. If we require the same precision of retrieval from the *RD* model, the above equation yielding the $m_c$ gives us a value $\alpha \simeq 0.5$, still higher than the SD value of 0.4. However, through simulations in Fig. 3.4 of the next section we have found that the network has a higher capacity ($> 0.6$) than the one predicted analytically (0.4).

When taking into consideration, for the Hopfield model, the increased capacity of the SD model with respect to what is predicted analytically, as well as the precision of retrieval, we find that both models behave qualitatively similarly. We clarify this in the next section by making the Potts-Hopfield correspondence exact.

### 3.4.3 Special case of the Potts network yields the Hopfield model

We can rewrite the Potts Hamiltonian, Eq. (2.2) with $S = 1$, $a = 0.5$ and $U = 0$ such that:

$$H = -\frac{1}{2}\sum_{i,j\neq i}^{N} J_{ij}\sigma_i\sigma_j \, , \tag{3.55}$$

$$J_{ij} = \frac{4}{c_m}\sum_{\mu=1}^{p}\left(\xi_i^\mu - \frac{1}{2}\right)\left(\xi_j^\mu - \frac{1}{2}\right) . \tag{3.56}$$

where $\sigma$ and $\xi$ take the values $\{0, 1\}$. We can rewrite the the latter quantities using the spin formulation $\{-1, +1\}$ using the transformation $2\sigma_i = s_i + 1$

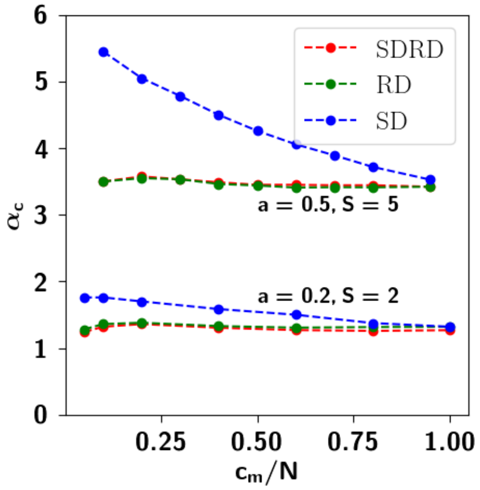


**Fig. 3.3: Storage capacity with different dilution graphs.** Storage capacity curves, obtained through simulations, as a function of the mean connectivity per unit $c_m/N$, for three different types of connectivity graphs, namely the random dilution (RD), symmetric dilution (SD) and state-dependent random dilution (SDRD). We find that SD has higher capacity than RD. The capacity for all three models coalesces at the fully connected limit, as the models become equivalent. Simulations carried out for two sets of parameters: ($N = 5000$, $S = 2$, $a = 0.2$) and ($N = 2000$, $S = 5$, $a = 0.5$). $U = 0.5$ and $\beta = 200$.

$$\tilde{H} = -\frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}s_i s_j - \frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}(s_i + s_j) - \frac{1}{8}\sum_{i,j\neq i}^{N} c_{ij}\tilde{J}_{ij}\,, \tag{3.57}$$

$$\tilde{J}_{ij} = \frac{1}{c_m}\sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu\,. \tag{3.58}$$

We note now that the first term in Eq. (3.57) is the Hopfield Hamiltonian for storing unbiased patterns, modulo a multiplicative term 1/4 [Amit et al., 1985]; at zero-temperature, however, an overall rescaling of the energies leaves the statistics of the system unchanged, so that we can consider the first term in Eq. (3.57) as exactly the Hopfield Hamiltonian. The last term is an additive constant that can be neglected, while the second term can be made to vanish by the addition of a unit dependent threshold term to Eq. (3.57)

$$\tilde{U}_i = \frac{1}{8}\sum_{j(\neq i)}^{N} (c_{ij} + c_{ji})\tilde{J}_{ij} \tag{3.59}$$

or equivalently, to Eq. (3.55) using the binary formulation

$$H = -\frac{1}{2}\sum_{i,j\neq i}^{N} J_{ij}\sigma_i\sigma_j + \sum_{i}^{N}\left(\frac{1}{4}\sum_{j(\neq i)}^{N}(c_{ij} + c_{ji})J_{ij}\right)\sigma_i \tag{3.60}$$

Considering $c_{ij}$ to be of the SD type such that $c_{ij} = c_{ji}$, this is the Hamiltonian considered by Sompolinsky [Sompolinsky, 1986]. The system with Hamiltonian given by Eq. (3.60) can be

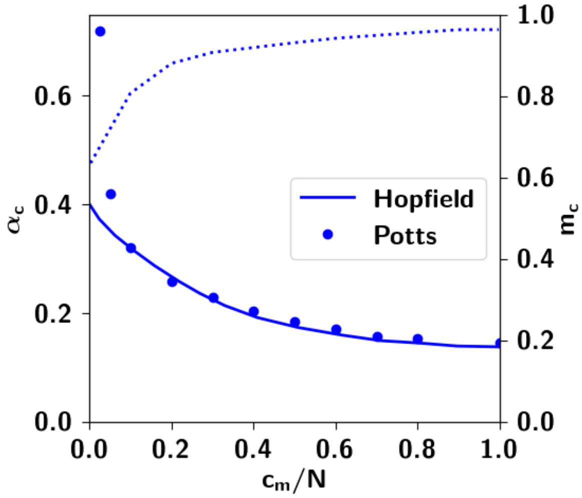


**Fig. 3.4: Special case of Potts network yields the Hopfield model.** Setting $S = 1$, $a = 0.5$ and the threshold to be unit dependent ($U = U_i$) the Hamiltonians of the two models become equivalent. Dots correspond to simulations of the Potts network with the latter parameters, while the uninterrupted line corresponds to analytical results obtained by Sompolinsky. The dashed line, to be read with the right $y$-axis, corresponds to the overlap at the critical capacity. For intermediate values of the connectivity, up to $c_m/N = 0.1$, our simulation results fit the analytical curve well, for higher levels of dilution, we find a greater capacity. Simulations performed with a network of $N = 2000$ units.

simulated by setting the parameters of the Potts network to $S = 1$, $a = 0.5$ and $U = U_i$ and the results compared to the analytical results derived in the latter study. We have carried out the simulations to reproduce these results, that we report in Fig. (3.4). Instead, considering $c_{ij}$ to be of the RD type yields the model studied by [Derrida et al., 1987] at the highly diluted limit.

The unit-dependent threshold that correlates with the learned patterns (and in our case with the diluted connectivity), for the equivalence of the two formulations of the Hamiltonians with spin and binary variables, was first found to be significant when the storage of biased patterns was considered [Tsodyks and Feigel'Man, 1988].

## 3.5. The storage capacity parameters

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? This a critical issue, if we are to make cortical sense of the distinct thermodynamic phases that can be analysed with the Potts model, and to develop informed conjectures about cortical phase transitions [Treves, 2005].

Let us consider a multi-modular Hopfield network of $N_m$ modules [O'Kane and Treves, 1992a, O'Kane and Treves, 1992b, Mari and Treves, 1998], each comprised of $N_u$ neurons, each of which is connected to all $N_u - 1$ other neurons within the same module, and to $C_A$ other neurons distributed randomly throughout all the other modules.

Equivalently [Naim et al., 2017], in the Potts network, if there are $N_m$ Potts variables, in the fully connected case, $N_m \cdot (N_m - 1) \cdot S^2/2$ connection variables are required (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, has only $N_m \cdot N_u \cdot C_A$ 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 $C_A \cdot N_u$.

In the simple Braitenberg model of mammalian cortical connectivity [Braitenberg, 1978b], which motivated the multi-modular network model [O'Kane and Treves, 1992a], $N_u \simeq N_m \sim 10^3 - 10^5$, as the total number 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 $1\,\text{mm}^2$ of cortical surface, also estimated to include $N_u \sim 10^5$ pyramidal cells [Braitenberg, 1978a]. A module, however, cannot be plausibly considered to be fully connected; the available measures suggest that, even at the shortest distance, the connection probability between pyramidal cells is at most of order 1/10. Therefore we can write, departing from the assumption $C_B = N_u - 1$ in the simplest version of Braitenberg's model, that $C_B \simeq 0.1 N_u$. If we were to keep the approximate equivalence $C_A \simeq C_B$, that would imply also $C_A \simeq 0.1 N_u$.

What about $c_m$ and $S$? What values would be compatible with associative storage? The number $S$ of local patterns on $N_u$ neurons receiving $C_B$ connections from each other can at most be, given sparsity $a_u$, of order $C_B/a_u$. If we assume that local storage tends to saturate this capacity bound (an assumption one may or may not consider), *and* we take $C_A \simeq C_B$, we have $S \cdot a_u \simeq C_B \simeq C_A$, but in turn we have, above, $C_A \cdot N_u > c_m \cdot S^2$, 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 [Kropff and Treves, 2005]

$$p_c \sim 0.15 \frac{c_m S^2}{a \ln(S/a)} \tag{3.61}$$

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 above considerations may sound rather vague. They neglect, *inter alia*, the large variability in the number of spines, hence probably in synapses, among cortical areas within the same species [Elston, 2000]. 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 Eq. (3.61). 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, and in keeping with the Frolov conjecture [Frolov et al., 1997], it is not different from any other associative memory network based on Hebb-like plasticity, 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 of order $N_u N_m$ bits of information each. The network can store and retrieve up to a number $p_c$ of them, which has been argued in [O'Kane and Treves, 1992b] to be limited by the *memory glass* problem to be of the same order of magnitude as the number $S$ of local attractors, itself limited to be (at most) of order $N_u$ or perhaps, as argued above, $\sqrt{N_u}$.

By glossing over the single-neuron resolution, the Potts model forfeits the locally extensive character of the information contained in each pattern, losing a factor $N_u / \ln S$, but it gains the factor $c_m S^2 / (2 N_u \ln S)$ in the number of patterns. Whether $S$ scales with $N_u$ or with $\sqrt{N_u}$ or in between, the upshot is more, but less informative, memories. Therefore, by focusing on long range interactions the Potts model misses out in information, but effectively circumvents the memory glass issue, which had plagued the earlier incarnation of the Braitenberg idea [Braitenberg, 1978a], and stores more patterns. How is that possible, if the Potts model is a reduced description of the underlying multi-modular model? The trick is likely in the Hebbian form of the tensor interactions, Eq. (2.1), which is *not* a straightforward reduction – it implies a fine inhibitory regulation that the multi-modular model had not attempted to achieve.

This argument can be expanded and made more precise by considering a more plausible scenario with correlated memories, the object of the next chapter.

# 4

# Generating correlated representations

## 4.1. Some insights from the organization of semantic memory

The field of neuropsychology has been a stepping stone to the study of the organization of the semantic system. Two of the key disorders that have offered a wealth of insights are semantic dementia and herpes simplex encephalitis. One of the observations in patients of herpes encephalitis has been a number of category specific deficits they display in a variety of tasks [Lambon Ralph et al., 2007]. A major double dissociation of historical importance has been the observation of a differential deficit in recognizing living and non-living entities, first observed in four herpes encephalitis patients [Warrington and Shallice, 1984]. This initial discovery, taken at face value, seemed to support the view that semantic knowledge is explicitly stored in independent systems according to category or domain.

However, Shallice and colleagues had a different interpretation: they proposed [Warrington and Shallice, 1984] the sensory-functional hypothesis, in that living entities usually rely more on sensory properties while nonliving entities rely more on functional properties. This account suggested that conceptual knowledge is distributed across functionally and neuroanatomically distinct systems, that may be modality specific, i.e. dedicated to the storage and processing of specific types of information (visual, motor, etc.). The seemingly category-specific deficit would then presumably be a consequence of predominant damage to one of these systems. Among others, one line of evidence for this account came from a herpes encephalitis patient displaying deficits in functional features compared to sensory features, independent of category [Borgo and Shallice, 2003, Shallice and Cooper, 2011]. However, it is argued that evidence problematic for this account comes from observations that category-specific deficits are not necessarily associated with a difficulty in the analysis of either sensory or functional properties, as in patient EW [Caramazza and Shelton, 1998].

An alternative commonly held perspective is the domain-specific view by Caramazza and colleagues, in which they argue that evolutionary pressures have resulted in specialized neural systems for concepts in a number of domains, such as animals, fruit and vegetables, conspecifics, and tools. The authors claim that the relevant adaptations might include dedicated neural circuits or specialized cognitive processes for processing information about animals and plants. The evidence for their theory comes from patients displaying fractionations within the living things category [Caramazza and Shelton, 1998]. However, it has been evidenced that the disruption of conceptual knowledge does not always follow domain boundaries. For instance, some patients with a deficit for living things also have trouble with musical instruments, gemstones, or food; on the other hand, other patients with a deficit for nonliving things show poor performance on parts of the body [Silveri and Gainotti, 1988].

While evidence from neuropsychology remains tentative, from which it is difficult to draw any definite conclusion, arguably the most interesting and revealing result is the fact that "most patients with category-specific semantic deficits show a graded impairment rather than an all-or-none dissociation" [Tyler and Moss, 2001]. That is, few patients are within the normal range in their "preserved" category or modality across the semantic tasks on which they are tested. Clear dissociations are outliers in that the few patients whose performance is consistently within the normal range for the preserved category are the exception rather than the rule [Patterson and Plaut, 2009]. How can graded impairments arise from a lesion to a distinct, dedicated neural system for a specific category/domain or modality?

One possibility is that the neural circuits lie in proximity of each other within the brain, in such a way that lesions affect more than one system, but to differing extents. However, this argument remains untestable unless precise predictions are made, based on neuroanatomic grounds. So far though, the evidence on the correlation of specific areas with specific functions has not reached wide consensus [Tyler and Moss, 2001]. The theories just described fit into one of the two broad interpretations of category-specific deficits, the *neural structure principle*.

A second class of theories, the *correlated structure principle*, purports the organization of semantic memory as reflecting the statistical co-occurrence of object properties. One theory in this class was the OUCH - the Organized Unitary Content Hypothesis [Caramazza et al., 1990]. In this account, objects in conceptual space would tend to cluster, and if damage occurs on such a site, a category-specific deficit ensues [Capitani et al., 2003]. Yet others emphasize the co-occurrence of feature types that facilitate categorical knowledge and identification [EilingYee and Thompson-Schill, 2013].

According to the *conceptual structure account* of Tyler and colleagues [Tyler and Moss, 2001], categories have different internal structures based on feature correlations and distinctiveness. They argue that highly correlated features support categorical knowledge as a whole, whereas feature distinctiveness allows for the accurate discrimination between objects. Similarly, Cree and colleagues [Cree and McRae, 2003] claim that category-specific deficits can be explained through the sharedness and the distinctness of the features associated to the concepts within categories. In the same spirit Humphrey and colleagues [Humphreys and Forde, 2001] point out that living things share more common features than do nonliving things and that this greater similarity in features among living entities may create a "crowding" effect resulting in low discriminability, accounting for the typically disproportionate deficit in living object identification. The featural representation approach has, so far, been successful in explaining several findings related to semantic memory, such as similarity priming, feature verification, categorization and conceptual combination [McRae et al., 2005,McRae et al., 1997].

The *graded* property of semantic disorders is an observation of fundamental value that points at the importance of looking beyond approximate categories and boxes, into smaller, finer-grained properties and into concept features. As such, we draw on some of the key concepts [Vinson and Vigliocco, 2008] as a basis for our model of semantic memory:

- sharedness of features across concepts, features that are common to more than one entity (e.g., "live" applies to all living things)
- distinctiveness of features (i.e., features that are unique to a specific entity e.g., "hear" is unique to ears among body parts),
- correlation (i.e., co-occurrence of features among concepts e.g., entities that have a "tail" are likely to have "four legs").

In this chapter, after a brief elaboration on *uncorrelated patterns* used as model neural representations for the studies of Chap.3, we will first describe a procedure for generating *ultrametrically organized* patterns, as a model of generating activity patterns that belong to distinct categories. We will show that such patterns have limited correlation and fail to reproduce the "fuzzy" categories of semantic memory. Based on this limitation, we will extend and generalize the algorithm, the idea of which was first formulated in [Treves, 2005]. We will show that this generalization allows to generate representations of arbitrary scope and correlation. The activity patterns thus generated are a simple concrete way to go beyond hierarchical trees, toward models which envisage multiple influences.

## 4.2. Uncorrelated patterns

The initial studies of the capacity estimates of the Potts network [Kropff and Treves, 2005] as well as the analyses in the previous Chap. 3 featured patterns that were uncorrelated. Uncorrelated patterns are generated by drawing Potts states for different units of different patterns identically and independently from Eq. (3.3). This means that Eqs.(3.2) and (3.1) hold. For any two patterns $\mu \neq \nu$, $C_0$ is the fraction of quiescent units they share, $C_{as}$ is the fraction of active units that are in the same state and $C_{ad}$ the fraction of active units which are in different states. Finally $C_{a0}$ is the number of units quiescent in $\mu$ and active in $\nu$:

$$C_0^{\mu\nu} = \frac{1}{Na} \sum_{i=1}^{N} \delta_{\xi_i^\mu, \xi_i^\nu} \delta_{\xi_i^\nu, 0} \tag{4.1}$$

$$C_{as}^{\mu\nu} = \frac{1}{N(1-a)} \sum_{i=1}^{N} \delta_{\xi_i^\mu, \xi_i^\nu} (1 - \delta_{\xi_i^\nu, 0}) \tag{4.2}$$

$$C_{ad}^{\mu\nu} = \frac{1}{Na} \sum_{i=1}^{N} (1 - \delta_{\xi_i^\mu, \xi_i^\nu})(1 - \delta_{\xi_i^\mu, 0})(1 - \delta_{\xi_i^\nu, 0}) \tag{4.3}$$

$$C_{a0}^{\mu\nu} = \frac{1}{Na} \sum_{i=1}^{N} (1 - \delta_{\xi_i^\mu, \xi_i^\nu}) \delta_{\xi_i^\mu, 0} (1 - \delta_{\xi_i^\nu, 0}) \tag{4.4}$$

The distributions of these correlation values are straightforward and given by binomial distributions with different success probabilities:

$$P(C_0) = N(1-a)\, B\Big((N(1-a)\, C_0; N, (1-a)^2\Big) \tag{4.5}$$

$$P(C_{as}) = Na\,B\Big(Na\,C_{as}; N, \frac{a^2}{S}\Big) \tag{4.6}$$

$$P(C_{ad}) = Na\,B\Big(Na\,C_{ad}; N, \frac{(S-1)a^2}{S}\Big) \tag{4.7}$$

$$P(C_{a0}) = Na\,B\Big(Na\,C_{a0}; N, a(1-a)\Big) \tag{4.8}$$

where $B(k; N, p) \equiv \binom{N}{k} p^k (1-p)^{N-k}$.

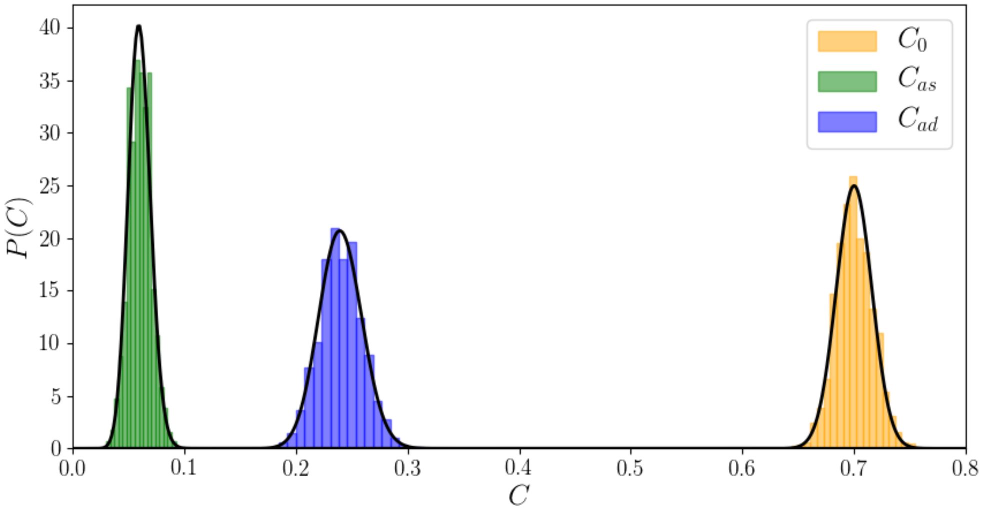


**Fig. 4.1: Pdfs of the patternwise correlations with uncorrelated patterns.**Probability density functions of $P(C_0)$, $P(C_{as})$, and $P(C_{ad})$ with uncorrelated patterns. While bars indicate results from simulations, solid lines represent the density functions. Here $N = 2000$, $a = 0.3$ and $S = 5$.

## 4.3. Single parents and ultrametrically organized children

The interest in ultrametrically organized patterns was largely due to the discovery of an ultrametric hierarchy of the free energy minima in the formal solution of the Sherrington-Kirkpatrick model of a spin glass [Mézard et al., 1987]. Subsequently a selectionist (as opposed to an empiricist *tabula rasa*) hypothesis of learning was proposed: the initial configuration of the network is a complex landscape with an abundance of valleys typical of spin glasses and learning consists of the progressive pruning or smoothening of this landscape [Toulouse et al., 1987]. It was then found that spin-glass ultrametricity is too sensitive to the effects of a stored pattern, but the appeal of its clarity had already set independent researchers to explore the possibility of storing ultrametrically organized patterns. In particular, the Hopfield model of neural networks was extended to allow for the storage and retrieval of hierarchically correlated patterns [Gutfreund, 1988].

In this study [Gutfreund, 1988], a set of random patterns, called parents are characterized by independent units, active with probability $a$

$$P(\xi_i^\pi) = a\,\delta(\xi_i^\pi - 1) + (1-a)\delta(\xi_i^\pi)\,, \tag{4.9}$$

where $\xi_i^\pi$ denote the activity of unit $i$ of parent $\pi$ and $0 < a < 1$ is the sparsity of the parents. In the next step, "child" patterns are drawn from the following distribution

$$P(\xi_i^{\pi\mu}) = \Big\{a + b(\xi_i^\pi - a)\Big\}\delta(\xi_i^{\pi\mu} - 1) + \Big\{1 - a - b(\xi_i^\pi - a)\Big\}\delta(\xi_i^{\pi\mu})\,, \tag{4.10}$$

where $\xi_i^{\pi\mu}$ denotes the activity of unit $i$ of child $\mu$ branching from parent $\pi$. $0 < b < 1$ parametrizes to what degree children are biased toward their (single) parent. For $b = 0$, child patterns become uncorrelated with no dependence on the parent, while for $b = 1$ the child patterns become identical to their single parent. Given the distributions above, we can compute the average activity of parents and child patterns [1]:

$$\langle \xi^\pi \rangle = a \tag{4.11}$$

$$\langle \xi^{\pi\mu} \rangle = a \tag{4.12}$$

as well as child-parent correlations:

$$\langle \xi^{\pi\mu} \xi^{\pi'} \rangle = \begin{cases} a^2 + ba - ba^2 & \pi = \pi' \\ a^2 & \pi \neq \pi' \end{cases} \tag{4.13}$$

As expected, children of the same branch have higher similarity to their own parent ($\pi = \pi'$), than to a parent of another branch ($\pi \neq \pi'$). We can also compute the correlation between two children of the same parent ($\pi = \pi'$) and that of two children belonging to distinct parents ($\pi \neq \pi'$)

$$\langle \xi^{\pi\mu} \xi^{\pi'\mu'} \rangle = \begin{cases} a^2 + a(1-a)b^2 & \pi = \pi' \\ a^2 & \pi \neq \pi' \end{cases} \tag{4.14}$$

[1] The state of each unit $i$ is drawn identically from the same distribution, such that we can drop the index $i$.

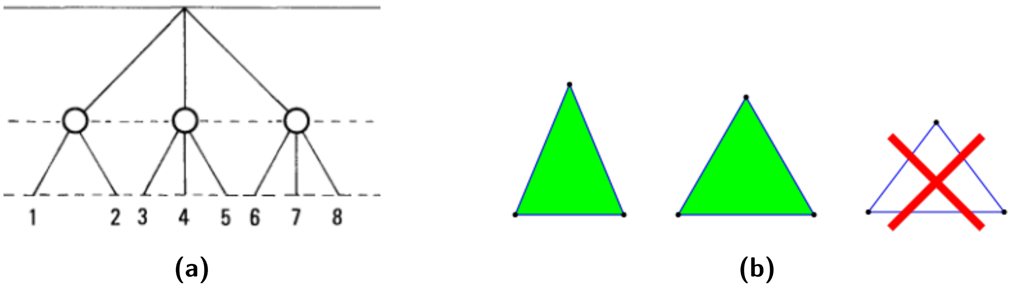


**Fig. 4.2: Ultrametric relations and strong triangle inequality.** **(a)** A tree, reproduced from [Mézard et al., 1987]. $1$, $2$, $3$, $4$, $5$, and $6$ are at the same level of the hierarchy. If we consider the nodes $1$, $3$ and $6$, they are each at a distance of $2$ of each other, the distance being defined as the distance to the nearest common branching point. If we consider nodes $3$, $4$ and $5$, then they are each at a distance of $1$ of each other, such that we get again an equilateral triangle. If we consider $1$, $2$ and $3$, then $d_{12} = 1$ while $d_{13} = d_{23} = 2$, such that we get an isosceles triangle with two long edges. The alternative, an isosceles triangle with two short edges is impossible to realize, there are *no intermediate points* between $1$ and $3$ or $2$ and $3$, as shown in **(b)**.

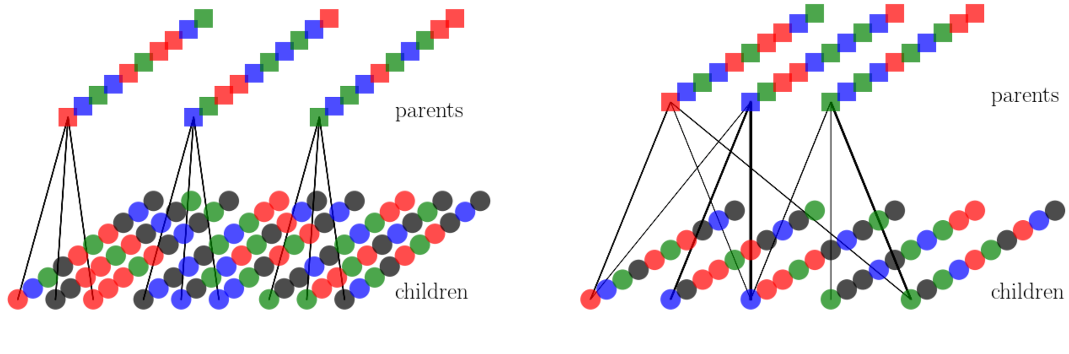
It trivially follows that

$$\langle \xi^{\pi\mu}\xi^{\pi\mu'}\rangle - \langle \xi^{\pi\mu}\xi^{\pi'\mu'}\rangle = a(1-a)b^2 \,. \tag{4.15}$$

This is one of the characteristics of this algorithm: it is possible to properly define a distance $d$ (see Sect. 4.6) such that three patterns $(x, y, z = \xi^{\pi\mu}, \xi^{\pi\mu'}, \xi^{\pi'\mu'})$ *at the same level of the hierarchy* can be seen to satisfy the *strong triangle inequality* $d(x, z) \leq max(d(x, y), d(y, z))$. An alternative definition can be formulated in terms of a tree: starting from two nodes at the same level of the hierarchy, the number of steps one has to go up the hierarchy to find the first common forefather. As illustrated in Fig. 4.2a, with such a definition, triplets of patterns can only be in one of the two triangle relations: equilateral and isosceles with two long edges, in other words, an ultrametric space has no node intermediate between any two nodes (Fig. 4.2b).

From the point of view of semantics, this is an implausible situation: if one considers the categories as the single archetypal parent from which all concepts descend, it becomes clear that such an ultrametric structure is unsuitable in describing all the semantic relations in which the ultrametric inequality is not satisfied: for example when a concept finds itself simply "in between" two equally distant concepts.

On the other hand, the very meaning of a concept can be thought of as the set of features that are associated to it. It may then be more sensible to consider the features characterizing a concept as its building blocks, hence its parents. In the following, what we will describe is an algorithm first



**(a)** **(b)**

**Fig. 4.3: Schematic representations of hierarchical versus multi-parent pattern generation algorithms.** **(a)** The workings of a hierarchical algorithm with $3$ parents and $3$ child patterns per parent. Colors correspond to active Potts states while black denotes the quiescent states. $S = 3$. **(b)** The workings of the multi-parent algorithm with $\Pi = 3$ parents and $p_{par} = 3$ child patterns per parent and $5$ total child patterns with number of parents $2, 1, 3, 1, 2$. Black arrows and their thickness denote strength of input. The main difference with the hierarchical algorithm is that each child pattern can receive input from multiple parents. If each parent is to represent a feature and each child a concept, the algorithm entails the generation of a concept from multiple features. Quantities such as number of features, their sharedness, their distinctiveness and correlation between concepts can be mapped onto variables in our model, such as the number of parents, the number of common parents, patternwise correlation.

sketched in [Treves, 2005] in which each child pattern is generated from multiple parents (features), more specifically a random subset of the total group of parents.

## 4.4. Multiple parents and non-trivially organized children

In the previous section, we elaborated on the observation that parents, playing the role of categories, were insufficient to capture the correlational structure of semantic memory. On the other hand, insights from experimental studies, mentioned at the beginning of this chapter, pointed to the role that a number of *quantitative* variables, such as the number, the sharedness and distinctiveness of features and correlations, play in describing semantic memory. How can we incorporate such observations into our model of semantic memory? In a simplified scenario, one may consider features as the parents from which the concept children are to descend. We can then map quantities such as the number of features, the sharedness, the distinctiveness to variables in our model, respectively the number of parents, the number of common parents, and the distinctiveness of each parent.

The multi-parent pattern generation algorithm works in three stages. In the first stage, a set of $\Pi$ random patterns are generated to act as parents. In the second stage, each of the $\Pi$ parents are assigned to $p_{par}$ randomly chosen children. Then, a "child" pattern is generated: each pattern, receiving the influence of its parents, 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 each child pattern. A schematic representation can be seen in Fig. 4.3b. In this section we will give a quantitative description of this procedure.

### 4.4.1 The algorithm operating on simple binary units

One relevant quantity that helps to mathematically describe each child pattern is the number of parents acting on it. Each parent is assigned $p_{par}$ children out of a total of $p$. The probability distribution that a given child has $n_p$ parents, out of a total pool of $\Pi$ is given by a binomial, with $f = p_{par}/p$

$$P(n_p) = \binom{\Pi}{n_p} f^{\,n_p}(1-f)^{\Pi-n_p} \tag{4.16}$$

The algorithm draws, for the input $x_i^{\pi\to\mu}$ from unit $i$ of parent $\pi$ to unit $i$ of pattern $\mu$, a uniformly distributed random number in the interval $(0,1]$ with probability $a_p$ and zero with probability $1-a_p$ such that we can write:

$$P(x_i^{\pi\to\mu}) = a_p\,U_{(0,1]}(x_i^{\pi}) + (1-a_p)\delta(x_i^{\pi}) \tag{4.17}$$

$a_p$ is analogous to the $a$ parameter in Eq. (4.9). If $a_p \sim 0$ then a child pattern is very unlikely to receive, on a particular unit, the contribution from one of its parents. On the other hand, if $a_p \sim 1$ then all parents influencing a child contribute to its field, whichever the unit. $U_{(0,1]}$ denotes the uniform distribution, such that input from parents is graded, contrary to the previous section.

Importantly, we have made the choice of non-sparse parents, but sparse input from parents, aimed at decorrelating units, while conserving correlations between patterns. This choice will prove to be crucial in Sect. 3.3, where statistical independence between units will lead to a vanishing mean noise, using only a simple covariance rule. For $S = 1$, this means that the patterns generated by the algorithm are uncorrelated, but the importance of having non-sparse parents with sparse input from them becomes important when dealing with more than one Potts state. Nevertheless, in this section, we proceed with the calculation with $S = 1$ as it provides warm-up exercise to then treat genuine Potts units.

The main difference with respect to the single-parent algorithm is that here, one must compute the total field $h_i^\mu$ that a unit $i$ of pattern $\mu$ receives from all parents

$$h_i^\mu = \sum_{\pi=1}^{\Pi} x_i^{\pi\to\mu}\, \mathbb{I}_{\Omega_\mu}(\pi) + \epsilon\,. \tag{4.18}$$

where $\Omega_\mu$ is the set of all parents acting on pattern $\mu$ and where we have that $|\Omega_\mu| = n_p(\mu)$. $\mathbb{I}_{\Omega_\mu}(\pi)$ is the indicator function that is 1 if parent $\pi$ is assigned to pattern $\mu$ and 0 otherwise. $\epsilon$ is a small random input ($\epsilon \ll a_p$) allowing for there to be some input when $a_p \ll 1$.

The fields of all units of all patterns are drawn from the same distribution. In App. B, the full derivation of the probability distribution for the field $h_i^\mu$ is reported in detail. We note in passing that such a distribution has a non-trivial expression and, to our knowledge, it can only be evaluated numerically. However, a very simple analytic expression can be given for the moments of the distribution of $h_i^\mu$, as shown in Fig. 4.4b.

$$\langle h \rangle = n_p\, \frac{a_p}{2} \tag{4.19}$$

$$\sigma_h = \sqrt{n_p\, a_p \left(\frac{1}{3} - \frac{a_p}{4}\right)}\,. \tag{4.20}$$

In Fig. 4.4a, where we see that the analytical results match tightly those from implementation of the algorithm. As a last step, a fraction $a$ of the units within a given pattern having fields above a threshold $h_m$ are set to become active. The threshold $h_m$ is then implicitly given in terms of the cumulative distribution function

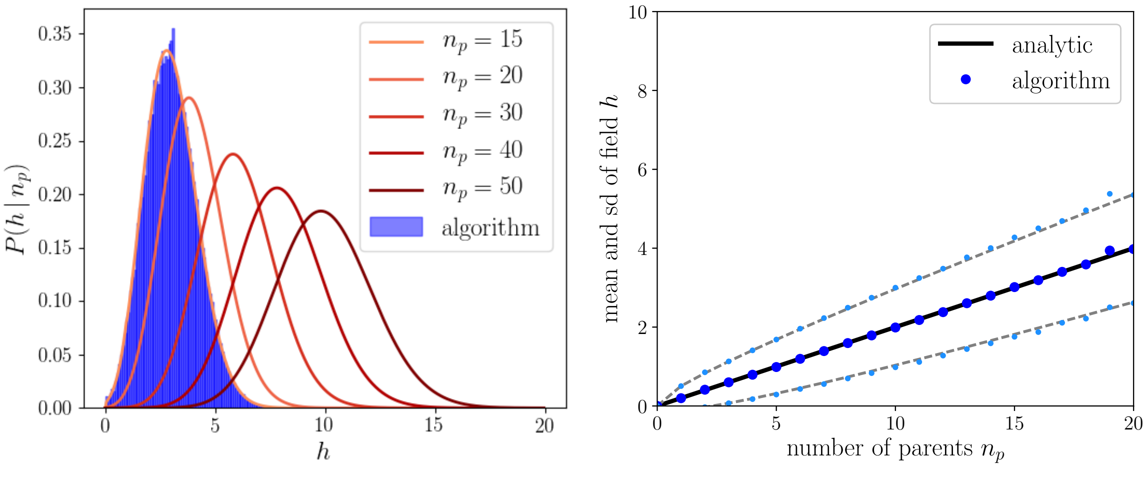
$$P(h' < h_m \,|\, n_p) = 1 - a\,. \tag{4.21}$$

For a given child pattern $\mu$ with number of parents $n_p$, we can now define the probability that it will be activated, given the field that it receives. If this field is greater than the threshold $h_m$, it will become activated:

$$P(\xi_i^\mu = 1 \,|\, h_i^\mu) = \Theta(h_i^\mu - h_m)\,. \tag{4.22}$$

### 4.4.2 The algorithm operating on genuine Potts units

With genuine Potts states, the main difference with respect to the previous results is that the input from a parent $\pi$ to the field of a pattern $\mu$ can be to any one of $S$ states with equal probability. This means that only a subset $\Omega^k$ of the total parents will contribute to state $k$. We denote the number



**(a)** **(b)**

**Fig. 4.4: Field from a number of parents with $S = 1$.** **(a)** Solid lines correspond to the analytical distributions of the field, Eq. (B.8), in blue is the distribution of the fields produced by a simulation of the algorithm for $n_p = 15$. The parameters are $N = 2000$, $S = 1$, $a_p = 0.4$, $n_p = 15 \dots 50$ and $\Pi = 100$. **(b)** The mean and standard deviation of the field as a function of the number of parents.

of parents in the subset as $|\Omega^k| = n^k$. The joint distribution of number of parents by state is

$$P(n^1, ..., n^S) = \frac{n_p!}{S^{n_p} \prod_{k=1}^{S} n^k!} \,. \tag{4.23}$$

such that the constraint $\sum_{k=1}^{S} n^k = n_p$ is satisfied. We can then write the field of unit $i$ in state $k$ of pattern $\mu$:

$$h^{\mu}_{i,k} = \sum_{\pi=1}^{\Pi} x^{\pi \to \mu}_{i,k} \mathbb{I}_{\Omega^k_\mu}(\pi) + \epsilon \,. \tag{4.24}$$

Then, the algorithm is such that it selects, unit by unit, the state receiving the maximal input. Following some calculations shown in App. C, we can compute the distribution of the fields of those states having received maximal input $H$ (Fig. 4.5a). We can then compute, exactly as before, the threshold above which the unit becomes activated.

$$P(H' < H_m \,|\, n_p) = \int_{-\infty}^{H_m} P(H' \,|\, n_p)\, dh' = 1 - a \,. \tag{4.25}$$

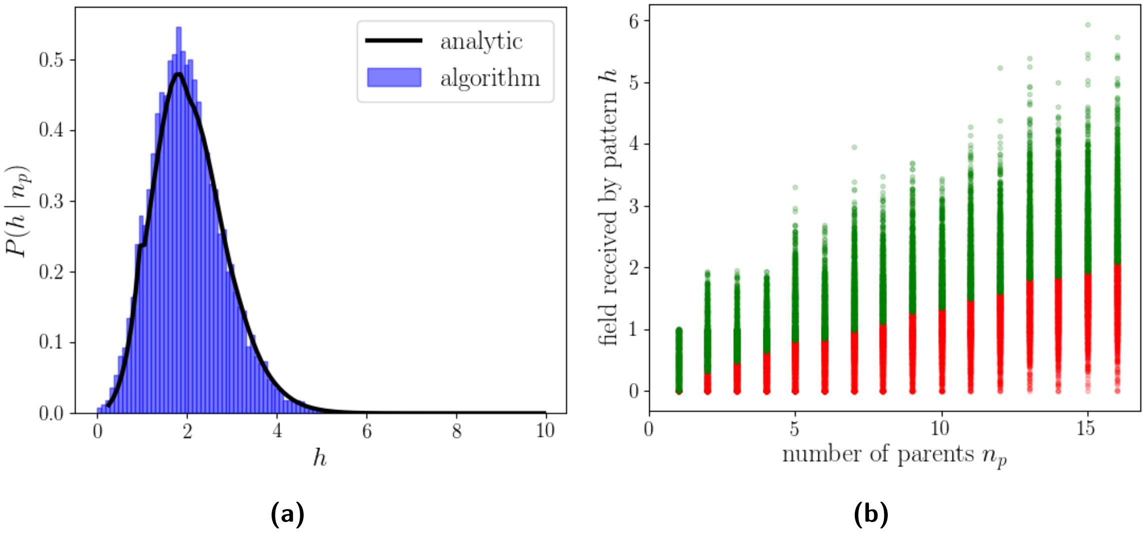


**Fig. 4.5: Field from a number of parents with $S = 2$.** **(a)** Distribution of the maximal fields for $S = 2$ and $n_p = 30$. In blue is the distribution of the fields produced by the algorithm and the black line is Eq. (C.7). **(b)** The $x$-axis corresponds to patterns with different number of parents and the $y$-axis to the fields of the units in that pattern. Red point correspond to units that are set to quiescent and green to those that are activated. The boundary between the green and the red corresponds to $h_m$, the minimum field required for a unit to be set to active. Parameters are $N = 2000$, $S = 2$, $a_p = 0.4$ and $\Pi = 100$.

Having obtained the minimal field $H_m$ required to activate a unit (Fig. 4.5b), we now need only the distribution of the field given the number of parents in that state $P(h^k|n^k)$, which is none other than Eq. (B.8) (replacing $n_p$ with $n^k$). We finally get to the distribution of activity across units and states, given the field received

$$P(\xi^\mu_{ik} = 1 \,|\, h^\mu_{ik}) = \Theta(h^\mu_{ik} - H_m)\,. \tag{4.26}$$

Given the algorithm we have just described, the main mechanism determining the state of a unit in a given pattern is the extent to which the units belonging to parents affecting a child are in the same state. If parent units are all aligned, this entails a high number of parents by state $n^k$, making the unit receive a higher mean field in a single state, making it more probable to become activated. On the other hand, lower alignment between parents will entail the field received by a child unit to be spread among the different states, and make it less probable for the child unit to find itself among those with maximal fields, as given by Eq. (C.7).

In the description above, we have aimed at describing the mechanism through which individual

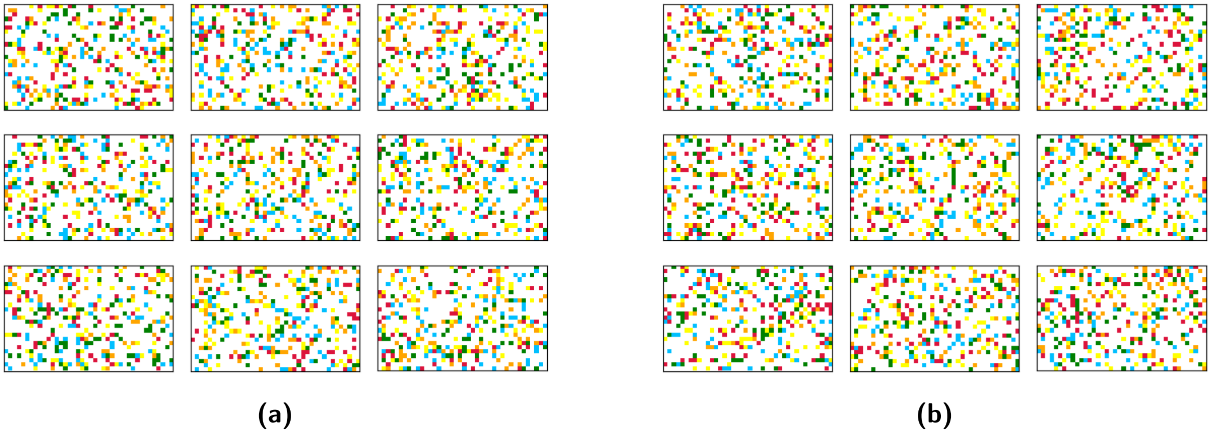


**Fig. 4.6: Sample patterns for Potts networks. (a)** 9 sample patterns generated independently following Eq. (3.3). Each subplot corresponds to one pattern. Each colored square to one unit, the colors indicating active states and white indicating the quiescent state. **(b)** Same as **(a)** but generated from one run of the multi-parent algorithm. The uniform sampling of the different Potts states is such that Eq. (3.3) holds. By design, also Eq. (3.2) holds (see text) while (3.1) does not hold anymore. Correlation parameters are $a_p = 1$, $f = 0.05$ and $\Pi = 150$ common parents. Parameters are $S = 5$ and $a = 0.3$.

child patterns are generated. At this level of description, in order to determine whether or not a unit of a pattern will become activated, the only relevant parameter is the number of parents and their degree of alignment in Potts space. From the point of view of an individual child pattern then, all parents are equivalent and can be considered as identical and independently distributed, a property we have thoroughly exploited in the previous sections. In the next section, we turn to the correlations between patterns. We study how the correlations generated by the algorithm differ from those obtained by a random procedure, as in Sect. 4.2 In particular, are they dominated by the number of parents that a pair of child patterns have in common? Is this a plausible model for semantic memory?

### 4.4.3 Resulting patterns and their correlations

In Fig. 4.6a and Fig. 4.6b we can see sample patterns generated randomly and with the algorithm from a common set of $\Pi$ parents, respectively. Patterns generated by the algorithm sample different active states uniformly, such that Eq. (3.3) still holds. However, the joint distribution $P(\bar{\xi}^1 \dots \bar{\xi}^p)$ is not factorizable anymore, as in Eq. (3.1). However, due to the way the algorithm works, units are still identical and independent (see Sect. 4.4). In Sect. 4.2 we defined the patternwise correlation as the fraction of units that are co-active and in the same state in both patterns [2]

$$C_{\mu\nu} = \frac{1}{Na} \sum_{i}^{N} \sum_{k}^{S} \delta_{\xi_i^\mu, k} \delta_{\xi_i^\nu, k} \, . \tag{4.27}$$

Analogously, we define the unitwise correlation as the fraction of patterns in which two units are co-active and in the same state

$$C_{ij} = \frac{1}{pa} \sum_{\mu}^{p} \sum_{k}^{S} \delta_{\xi_i^\mu, k} \delta_{\xi_j^\mu, k} \, . \tag{4.28}$$

In Fig. 4.7 we can see the distributions of $C_{\mu\nu}$ and $C_{ij}$ for nine different combinations of the parameters $a_p$ and $f$. The distributions are very sensitive to the specific values of the parameters. For low values of $a_p$ and $f$, pairs of Potts units have uncorrelated activity when averaged across patterns, in the sense that the distribution $C_{ij}$ has zero covariance. Pairs of patterns, instead, are correlated with a distribution $C_{\mu\nu}$ of non-zero covariance, that is positively skewed. On the other hand, high values of $a_p$ and $f$ seem to make both distributions more positively skewed. Low values of $a_p$ and high values of $f$ result in both distributions becoming more and more normal, while high values of $a_p$ and low values of $f$ result in a normally distributed correlation between units and a highly skewed multi-modal distribution between patterns.

To assess these observations more systematically, in Fig. 4.9a we can see boxplots of the $C_{\mu\nu}$ distributions for different values of $a_p$ keeping $f = 0.05$ fixed. While the mean correlation seems to be unaffected by increasing $a_p$, the standard deviation and the skewness increase.

In 4.9b, conversely, we can see boxplots of $C_{\mu\nu}$ distributions for different values of $f$ keeping $a_p = 0.4$ fixed. Increasing $f$ increases the mean of the correlation distribution, without affecting the standard deviation and the skewness. These effects can be understood intuitively because of the different roles that these parameters play in the algorithm. $a_p$ is the parameter that increases the probability that a child parent receives input from a parent unit, increasing the overall similarity of a child to its parents. This means that those children that have common parents will be more similar and more highly correlated, giving rise to the higher values in the distribution. $f$, on the other hand, is ratio of the pool of children affected by one parent to the total number of children. Increasing the pool of children that are to be affected by a given parent, means that all children share, on average, more parents, leading to a shift in the overall distribution. This effect is seen in Fig. 4.8, where we

[2] We drop the subscript "$as$" in order to simplify the notation

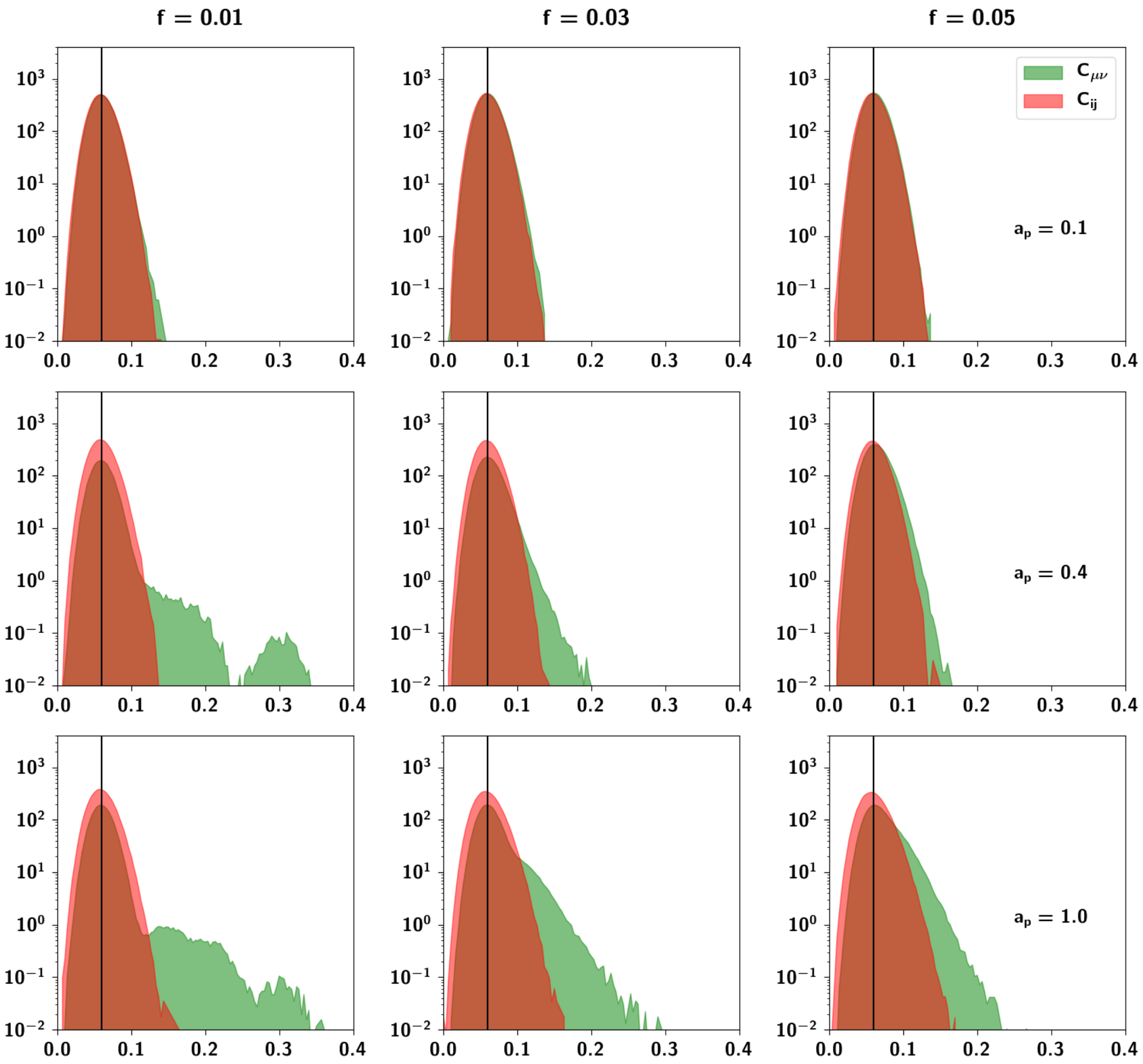


**Fig. 4.7: Pdfs of pairwise correlation between patterns and units of the multi-parent algorithm.** Probability density function of correlations between units (in red) and between patterns (in green) and three different values of $f$, corresponding to an average of $1.5$, $4.5$ and $7.5$ parents per pattern. The black vertical line corresponds to the average correlation with uncorrelated patterns distributed independently according to Eq. (3.3). Parameters are $S = 5$, $a = 0.3$, $a_p = 1.0$ and $\Pi = 150$. The algorithm produces correlations between patterns with high variability relative to the correlation between units, in line with candidate patterns for semantic memory. The green distribution can be compared to Fig. 4.1, i.e. the patternwise correlation distributions obtained with independently generated patterns of the same sparsity and number of Potts states. Note that the algorithm is sensitive to the parameters and produces a wide variety of correlation between patterns.

plot, as a function of the number of common parents, in the left and right y-axes respectively, the fraction of pairs of patterns and the mean correlation between those pairs of patterns. This is shown in more detail in Fig. 4.10, in which pairs of patterns are decomposed into different distributions sharing an increasing number of parents. It can be seen that the correlation distribution shifts to the right with increasing number of common parents. While the mean correlation seems to be determined by the number of parents two children have in common, it is not the only factor in play. Children have variable number of parents, and the number of total, *unshared* parents, can lower the mean correlation between pairs of children. The two parameters $a_p$ and $f$ therefore play different roles in generating the correlations.

### 4.4.4 The ultrametric limit

It is interesting to note a limit case of the algorithm. If $\langle n_p \rangle_\mu = \Pi f \sim 1$ as in Fig. 4.7 ($f = 0.01$, $\Pi = 150$), on average, most children will have a single parent, which effectively produces ultrametric patterns. Indeed, for these parameters, since the number of total parents $\Pi = 150$ is smaller than the total number of children generated, $p = 1000$, many children share a given single parent. The mean value of their correlation, at $a/S$, is the same as the mean correlation between uncorrelated patterns, as is predicted by Eq. (4.14). Note that the distribution is multimodal. The values forming the second mode of the distribution consist of the correlation between children belonging to the same (single) parent.

### 4.4.5 The random limit

Another limit is the random limit in which $a_p \ll 1$. In this case, most units will not receive input from their respective parents, regardless of how many they are, and the unit will align itself in the direction of a random Potts state given by the input $\epsilon$. In this way, it is possible to parametrically generate patterns ranging from independent ($a_p \ll 1$) to ultrametric ($a_p = 1$, $f\Pi = 1$).

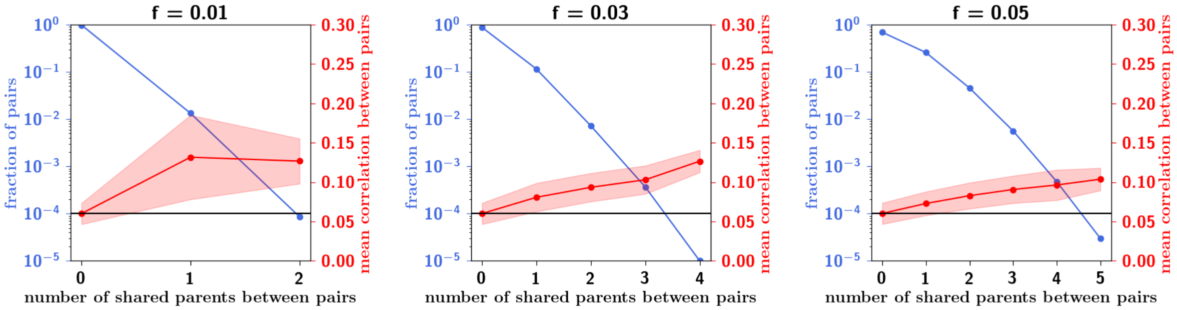


**Fig. 4.8: More common parents lead to higher patternwise correlation.** Fraction of pairs of patterns (left $y$-axis) and mean correlation between those pairs (right $y$-axis) as a function of number of common parents for three different values of $f$. The black horizontal line corresponds to the average correlation with uncorrelated patterns distributed according to Eq. (3.3). Those pairs of patterns having more common parents (features) are more highly correlated, on average. The other correlation parameters are $a_p = 0.4$ and $\Pi = 150$.

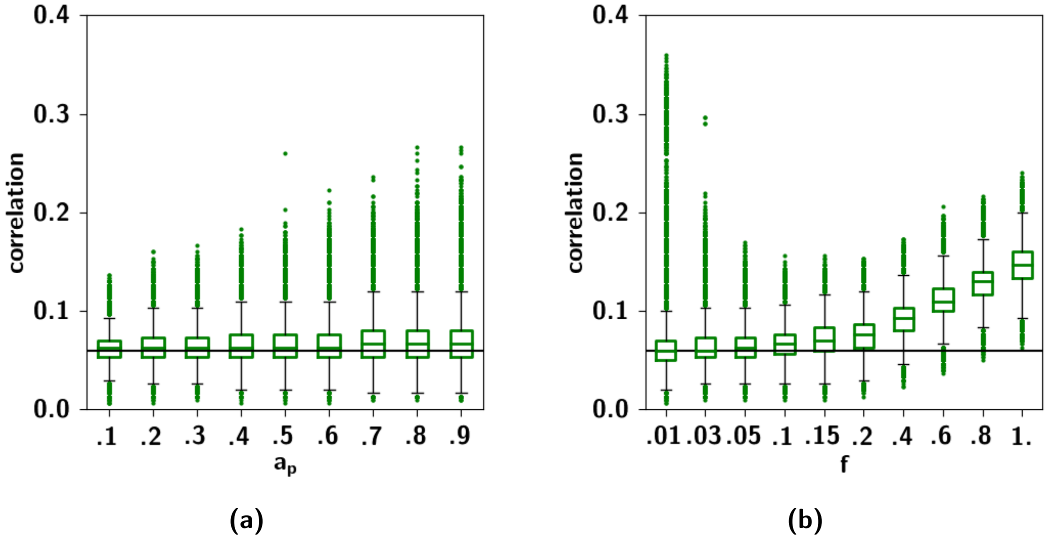


**Fig. 4.9: Boxplots of patternwise correlations. (a)** Boxplots of $C^{\mu\nu}$ for different values of $a_p$, with $f = 0.05$ fixed. **(b)** Boxplots of $C^{\mu\nu}$ for different values of $f$ with $a_p = 0.4$ fixed. The two figures display the different roles that the parameters $a_p$ and $f$ play in generating the correlations. Increasing $a_p$ increases the overall similarity of a child to the parent from which it receives input, such that those children having common parents will be more highly correlated, as evidenced by the increasing skewness of the distributions. In contrast, $f$ is the ratio of the pool of children attributed to one parent over the total number of children. Increasing $f$ leads to an increase in the mean number of common parents, such that all children will be more correlated, as evidenced by the shift in the overall distribution. The black vertical line corresponds to the average correlation with uncorrelated patterns distributed according to Eq. (3.3). Other parameters are $a = 0.3$, $S = 5$ and $\Pi = 150$.

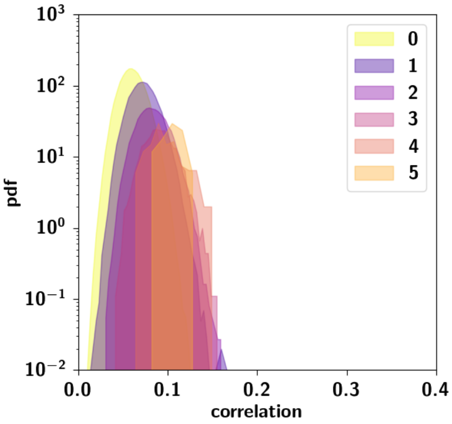


**Fig. 4.10: Distributions of patternwise correlations with increasing number of shared parents.** This figure is another depiction of the correlation distribution of Fig. 4.7 with $f = 0.05$ and $a_p = 0.4$. Here, all patternwise correlation values are decomposed into those that have an increasing number of common parents. It can be seen that the corresponding distributions shift to the right, corroborating the results of Fig. 4.8.

## 4.5. Correlation in a sample of 60 nouns

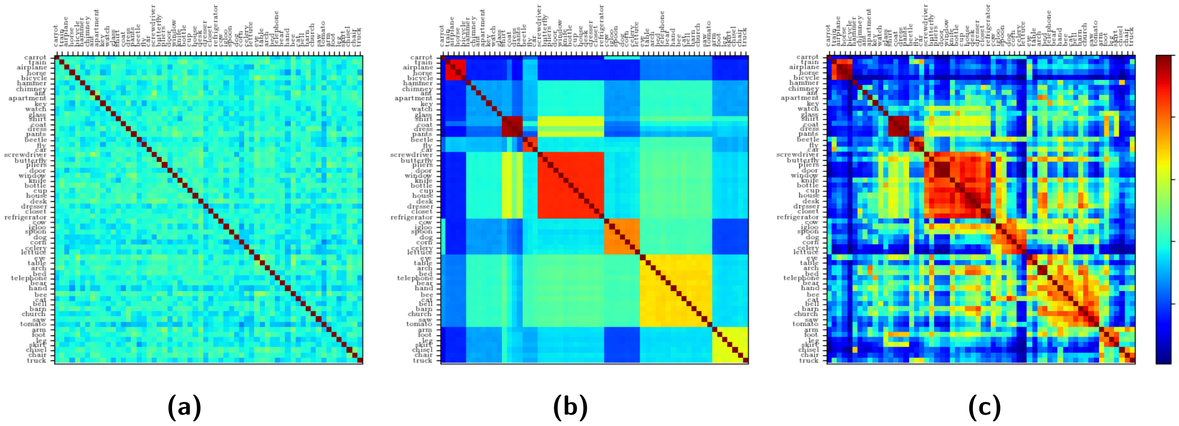


**Fig. 4.11: Correlation matrices of nouns, under different hypotheses. (a)** Null-hypothesis correlation matrix between nouns, obtained through shuffling, feature by feature, the feature weights across all nouns, resulting in a matrix with a near perfect homogeneity in correlation values and devoid of any structure. **(b)** Hierarchical hypothesis correlation matrix, obtained through shuffling, feature by feature, the feature weights only among nouns belonging to eight clusters that we define after applying a clustering algorithm to the original correlation matrix (see **(c)**). The seven clusters correspond to the seven warm colored blocks along the diagonal, as well as one light blue, corresponding to nouns that did not belong to any cluster. **(c)** Original correlation matrix.

In the previous section, we studied the correlation properties of the patterns generated by the multi-parent algorithm, intended to model patterns of semantic memory. In an ideal scenario, these could be contrasted to patterns of multi-voxel activity associated to a semantic retrieval task. In a slightly less ideal scenario, it is not implausible to assume that some of the correlational structure of semantic memory may be reflected in language.

In [Mitchell et al., 2008], the authors trained a computational model that *predicts* the fMRI neural activation associated with words. Their model is trained with a combination of data from a trillion-word text corpus and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for other concrete nouns in the text corpus, with highly significant accuracy over 60 nouns.

Based on the significant accuracy with which the model, trained partially with corpus data, predicted fMRI activation, we took the $N = 60$ nouns used in the above study. We computed the pairwise correlation between these nouns, as measured by a set of intermediate or surrogate features, such as the co-occurrence with specific verbs within a sentence in the corpus. We denote by $f_{ik}$ the co-occurrence frequency of noun $i$ with verb $k$. Then, in order to express each concept as a feature vector, we normalized each co-occurrence frequency to obtain feature weights such that

$$w_{ij} = \frac{f_{ij}}{\sqrt{\sum\limits_{k=1}^{M} f_{ik}^2}} \tag{4.29}$$

In this way each noun is expressed as a vector of $M = 25$ features such that $n_i = (w_{i1}, w_{i2}, ..., w_{iM})$ is normalized to 1. We can then compute the correlation between nouns $i$ and $j$ as

$$C_{ij} = \sum_{k=1}^{M} w_{ik} w_{jk} \tag{4.30}$$

In Fig. 4.11c, we report the correlation matrix of the nouns. In Fig. 4.11b, we compute the correlation between the nouns, *after* having shuffled the feature weights across nouns within eight distinct clusters obtained through a clustering algorithm. One of the eight clusters (second, light blue, along the diagonal) groups those nouns that did not have a significant correlation with any of the other nouns, such that some of the off-block values are *higher* than within block values. For all of the other blocks though, the ultrametric structure is evident. All in-block values are higher than off-block values. It is apparent from the comparison with Fig. 4.11c, that the ultrametricity assumption fails to represent the non-negligible number of off-block values with high correlation, nor account for off-

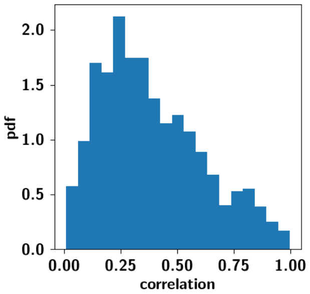


**Fig. 4.12: Correlation distribution between nouns.** Pairwise correlation between a set of $N = 60$ nouns in a corpus, as measured by their co-occurrence with a set of intermediate features (verbs). The distribution consists of the same values reported as a matrix in Fig. 4.11c.

block entries with correlation values in between that of nouns belonging to two distinct clusters. In Fig. 4.11a instead, we shuffle the feature weights across all nouns, resulting in a homogenous matrix with very little variability. This simple example illustrates the strong and implausible assumption of a hierarchical hypothesis. The correlation distribution can be seen in Fig. 4.12.

## 4.6. Ultrametric content

A further characterization of the resulting patterns is in terms of a distance. A distance measure can be derived from the correlation following the same procedure as [Treves, 1997]. We first define a so-called "confusion" matrix

$$P(\mu|\nu) = \frac{C_{\mu\nu}}{\sum_{\mu=0}^{p} C_{\mu\nu}} . \tag{4.31}$$

where $C_{\mu\nu}$ is an element of the correlation matrix and where $P$, the confusion matrix, is obtained by normalizing each element of the correlation matrix appropriately. Next, we symmetrize the above function to obtain

$$d(\mu,\nu) = -\log\left(\frac{P(\nu|\mu)P(\mu|\nu)}{P(\mu|\mu)P(\nu|\nu)}\right) , \tag{4.32}$$

a quasi-distance, in the sense that it satisfies only the reflective and symmetric properties, $d(\mu,\mu) = 0$ and $d(\mu,\nu) = d(\nu,\mu)$. The triangular inequality $d(\mu,\nu) + d(\mu,\rho) \leq d(\mu,\rho)$ does not necessarily hold. It can be made to hold by raising $d$ to a sufficiently small power $d \to d^{1/p}$, called the "trivialization" of $d$, as has been explained in detail in [Treves, 1997]. Using this procedure, distances between triplets of patterns $\{\mu,\nu,\rho\}$ can be computed. If we note by $d_{min}$ the edge of minimal length, $d_{max}$ the edge of maximal length and $d_{med}$ the edge of intermediate length, then we can plot, in a two-dimensional graph, the ratios $\delta_1 = d_{min}/d_{max}$ and $\delta_2 = d_{med}/d_{max}$. In Fig. 4.13, we report these plots for

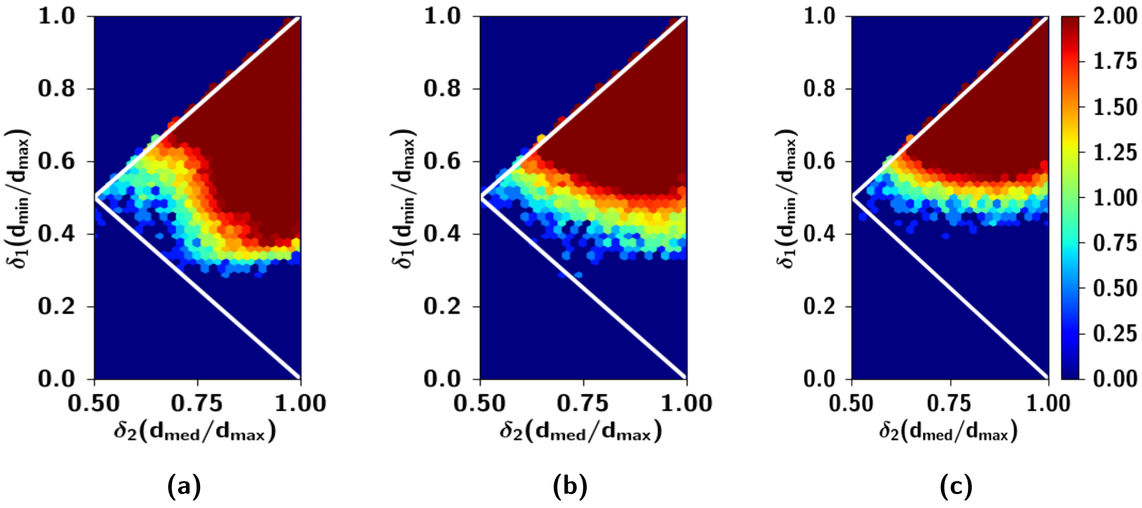


**Fig. 4.13: Distance relations between triplets of correlated patterns.** Two-dimensional histograms of $\delta_1 = d_{min}/d_{max}$ against $\delta_2 = d_{med}/d_{max}$ for three values of the correlation parameters $f = 0.01$, $0.03$ and $0.05$. The other parameters are $a_p = 1.0$ and $\Pi = 150$.

patterns generated with the multi-parent algorithm, for three different values of the parameter $f$, for which we also report the correlation distributions in Fig. 4.7. As can be anticipated from those distributions, increasing $f$ strengthens the overall correlation between patterns and reduces the distances between them.

Triplets that satisfy the triangular inequality lie above the line $\delta_1 = 1 - \delta_2$, while triplets that satisfy the ultrametric inequality lie on the vertical line where $\delta_2 = 1$. Among these, triplets that are equilateral triangles lie at the point $\delta_1 = \delta_2 = 1$. To measure the overall closeness of the cloud of triplets to the fully ultrametric limit one can define the *ultrametric content*

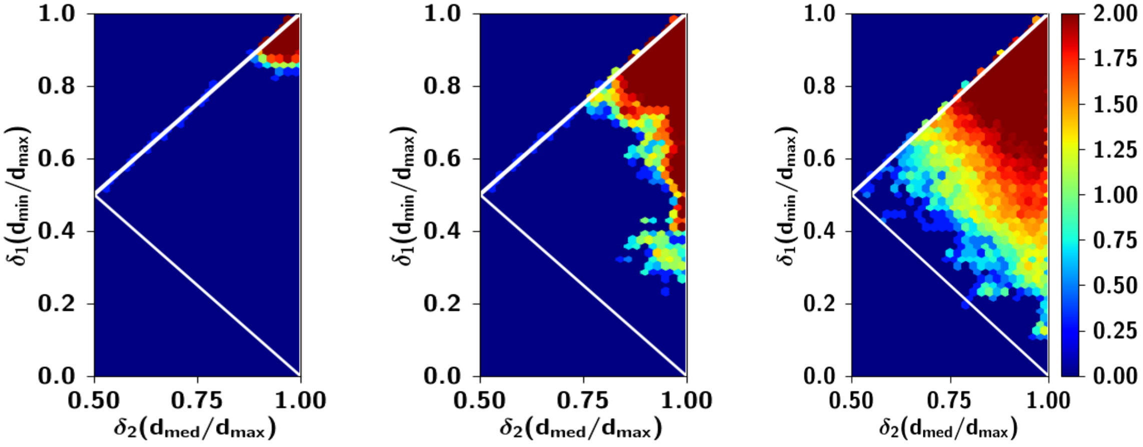
$$\lambda_{um} = \left\langle \frac{\log \delta_1 - \log \delta_2}{\log \delta_1 + \log \delta_2} \right\rangle \tag{4.33}$$

where $\langle \cdot \rangle$ denotes the mean over all triplets. This quantity does not depend on the trivialization of $d$ and it ranges from 0 (for triplets forming isosceles triangles with two short sides) to 1 (for a fully ultrametric set: equilateral triangles and isosceles triangles with two long sides).

Can we compare distance relations thus obtained with those of nouns, under different hypotheses?

In Fig. 4.14a, we can see the ratios $\delta_1$ against $\delta_2$ for triplets in Fig. 4.11a. The little fluctuation in the correlation is reflected in the distance relation between triplets of nouns, where they all tend to cluster toward the value $\delta_1 = \delta_2 = 1$. The ultrametric content, as defined by Eq. (4.33) is $\lambda_{um} = 0.4$.

In Fig. 4.14b instead, we can see that nouns largely categorize into clusters, so that triplets span two regions: $\delta_1 \sim \delta_2$ for triplets that belong to the same cluster, and the ridge $\delta_1 = 1$ for triplets that belong to distinct clusters. The ultrametric content is found to be $\lambda_{um} = 0.61$, still far from the limit value $\lambda = 1$, but significantly higher than when nearly all metricity is destroyed, as achieved with the shuffling of Fig. 4.14a. In Fig. 4.14c instead, we can see the distance scatter obtained from the original correlation matrix between the nouns: these values span much of the area below $\delta_1 = \delta_2$. Notably, the triplets do not only lie around the vertical line $\delta_2 = 1$, but span the right triangle, implying triplets forming all kinds of triangles, ranging from isosceles with two long sides to isosceles with two short sides on the line $\delta_1 = \delta_2$, with an intermediate ultrametric content of $\lambda_{um} \sim 0.52$.



(a)

(b)

(c)

**Fig. 4.14: Distance relations between triplets of nouns, under different hypotheses. (a)** Two-dimensional histograms of $\delta_1 = d_{min}/d_{max}$ against $\delta_2 = d_{med}/d_{max}$ for triplets of 60 nouns taken from corpus after full reshuffling of feature weights. $\lambda_{um} \sim 0.4$ **(b)** same as **(a)**, after within-cluster reshuffling (the ultrametric hypothesis). $\lambda_{um} \sim 0.61$ **(c)** Distance relations obtained from the original correlation matrix. $\lambda_{um} \sim 0.52$.

## 4.7. Semantic distinctiveness

How important is each individual feature, across all nouns? We can compute a simple measure of semantic "distinctiveness", by simply summing the feature weights of all nouns $s_j = \sum_i^N w_{ij}$.

In Fig. 4.15, we report the summed weights of the $M = 50$ features across all the nouns considered, sorted and plotted on a semi-log scale. Remarkably, given the very small dataset used, it can be approximated to a good extent by an exponential law. The suboptimal fit may conceivably be the result of limited and unbalanced sampling. Indeed the words, the nouns or the verbs were not chosen with comparable frequency. This measure is therefore only approximative, in that it is an aggregate measure of "distinctiveness"; however at the resolution of the description of the model, it is sufficient as such. Our measure is related to the measure called "semantic relevance" used by Sartori and colleagues [Sartori and Lombardi, 2004] as well as "semantic differential" used by Osgood [Osgood, 1964]. The difference with the latter measure, however, is that ours is a cumulative measure across all of the nouns derived from co-occurrence statistics in a corpus, while semantic differential refers to a scale in which individuals rate the connotative meaning of objects, events, and concepts [3].

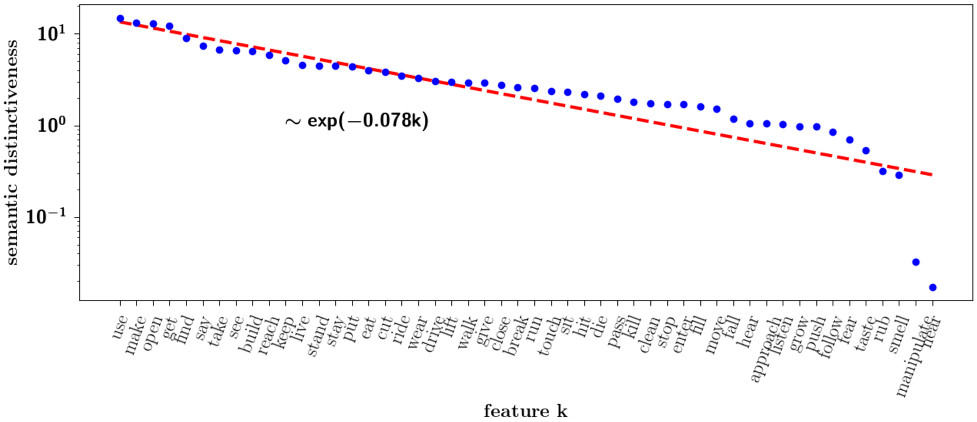


**Fig. 4.15: Semantic distinctiveness of features.** The $x$-axis enumerates all of the features used to compute the correlation between the nouns, sorted according to their summed weights across all nouns (reported on a semi-logarithmic $y$-axis).

To take into account this observation, we consider a more refined model in which the parents in our algorithm (the features), ranked from 1 to $\Pi$, have the strength of their inputs damped exponentially with rate $\zeta$, such that Eq. (4.24) is revised in the following way

$$h_{i,k}^{\mu} = \sum_{\pi=1}^{\Pi} x_{i,k}^{\pi \to \mu} \mathbb{I}_{\Omega_{\mu}^{k}}(\pi) \exp(-\zeta\pi) + \epsilon \,. \tag{4.34}$$

[3] It is likely that the two measures are related, however, in order to avoid confusion, we do not use the same term to refer to our measure.

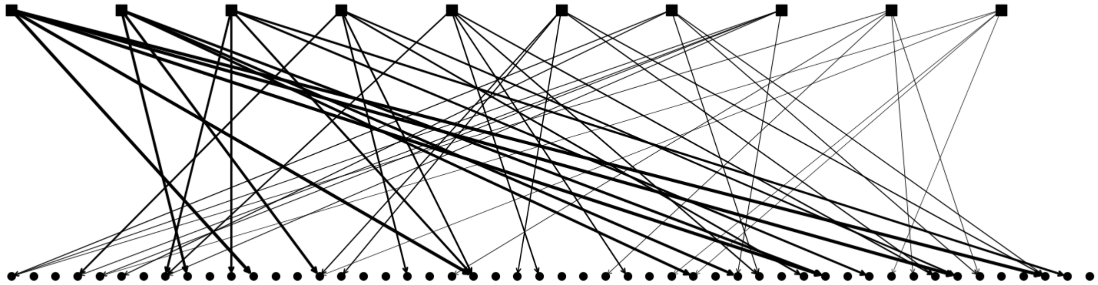


**Fig. 4.16: Multi-parent algorithm incorporating semantic distinctiveness.** One sample representation of parent-child relations. The squares on the top row represent parents, while the circles at the bottom row represent children. Black lines represent input from the parents to the children. The strength with which each parent affects its children is proportional to $\exp(-\zeta\pi)$, where $\pi$ indexes the parents, as explained in the text. The parameters are $\Pi = 10$ parents, $p_{par} = 5$ children per parent and $p = 50$ total children. For simplicity, each square and each circle represent parents and children as entities, without the individual units.

where $x_i^{\pi\to\mu}$ is the input from parent $\pi$ to child pattern $\mu$, $\Omega_\mu$ is the set of all parents acting on pattern $\mu$ and $\mathbb{I}_{\Omega_\mu}(\pi)$ is the indicator function that is 1 if parent $\pi$ is assigned to pattern $\mu$ and 0 otherwise. The limit $\zeta \to 0$ corresponds to the algorithm described in the previous sections such that we recover Eq. (4.18). In this way, we introduce a parameter, $\zeta$, which can be related to the slope seen in distinctiveness distributions observed in real data, such as the one in Fig. 4.15.

In Fig. 4.16 we can see a schematic representation of this new algorithm. In previous sections, we investigated the role played by $a_p$, the sparsity of input that a child-unit receives from a parent-unit. This input takes graded values in the range $(0, 1)$ across different units, leading to variability of input across units of a pattern. The parameter $\zeta$, though also affecting the strength of input, plays a different role, as it affects the global strength with which each parent affects its children, leading to variability of input across all units of different patterns. A high value of $\zeta$ contributes to highly unbalanced input from all parents influencing a child pattern, such that most units of a pattern align with the most powerful parent, or the most "distinctive" feature.

How are the correlations affected by $\zeta$? In Fig. 4.17 we report the distributions for three different values of $\zeta$. While for low values of $\zeta$, i.e. parents homogeneous in their strengths, the pattern-wise correlation is unaffected (see Fig. 4.7), increasing $\zeta$, we see the emergence of a tail of highly correlated patterns. For small $f$, this has the effect of smearing the bi-modal distribution, while for larger $f$, the already existing tail becomes fatter.

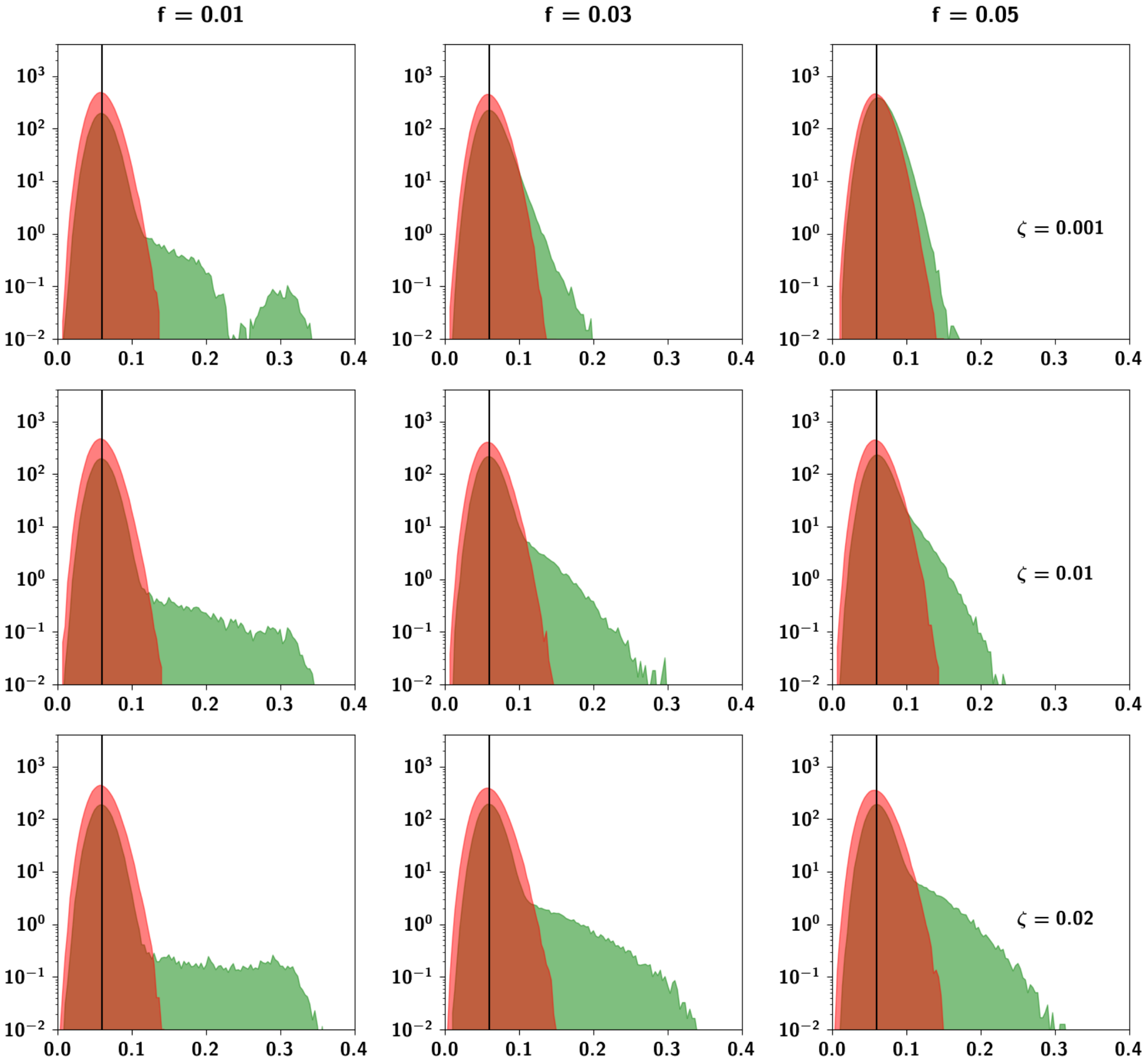


**Fig. 4.17: Pdfs of pairwise correlation between patterns and units of the multi-parent algorithm.** Probability density function of correlations between units (in red) and between patterns (in green) for three different values of the parameter $\zeta$ and $f$, keeping $a_p = 0.4$ constant. For the low value of $\zeta = 0.001$, this figure is reproduces the middle panel of Fig. 4.7. For higher values of $\zeta$, where the parents become highly heterogeneous, we see the emergence of very high correlations.

## 4.8. Discussion

Tree-like models have for a long time remained an important way of thinking about semantic organization, prompted by early studies of category specific deficits. Subsequently, it was clarified that other arrangements of distributed representations could also lead to such deficits [Farah and McClelland, 1991], but to date, a plausible model of generating such representations, amenable to further studies, has been largely lacking. To make progress in this direction, we have given a quantitative description of a novel algorithm, with a crucial conceptual difference from that of previous attempts: our algorithm allows for child patterns to receive input from multiple parents. The parents, can be thought of as semantic features, or else semantic category generators, and lead to correlations that are highly sensitive to the values of the parameters.

We have made many simplifications, also at variance with earlier attempts. For one thing, we have adopted an algorithm that defines the activity of parents over *all* units of subsets of patterns, a difference with [Treves, 2005]. In the latter algorithm, it was the inverse: parents were defined over subsets of units of all patterns. This detail made a difference in that the latter algorithm produced patterns that tended to correlate *units*. The motivation behind the latter algorithm was one of plausibility: features are usually associated to different modalities such as sensory, motor, visual etc. and are known to be represented in patches of cortex that are sometimes anatomically segregated. Such motivation remains valid, and further versions of the pattern generation algorithm should be explored. Such attempts can be taken by for example defining the parents over more realistic distributions of patterns and units.

Moreover, we have considered parents as comprising only active states, and the *input* from parents to be sparse. This approximation is far from being realistic, but has the advantage of decorrelating units, allowing for the mean field equations determining the storage capacity to take a simple form, as we show in Chap. 5. Notwithstanding all of the simplifications, the resulting correlations are shown to be sensitive to the values of the correlation parameters and are rich enough to allow for interesting further work.

The architecture of our algorithm bears some resemblance to restricted Boltzmann machines (RBMs), in that the parents (representing the features) play the same role as the hidden variables, while the children (the nouns, or concepts) play the role of the visible variables. The parameters which define the rules with which children are generated from the parents are related to the weights of the links, in the RBM, between visible and hidden units. One interesting property of RBMs is that their weights can be trained in such a way that the hidden variables extract relevant features from data in an unsupervised way. After the training, the network can be used as a generative model to produce new data: this is what also our algorithm does. The part missing in our generative model, or at least in the way we use it, is the training part: given the structure of the connections, it might be included in our scheme. This might be an interesting direction in which to extend our algorithm.

# 5

# Storage capacity of the Potts network with correlated patterns

After having studied in some detail the algorithm which generates correlated patterns, we can turn to study the storage capacity and how it is affected by the correlations. We have carried out numerical simulations with the standard "Hebbian" or closely related "covariance" learning rule for Potts networks [Kropff and Treves, 2005] with the learning rule in Eq. (2.1), and have observed that the storage capacity is diminished in the case of correlated patterns, a result that has been obtained analytically by others [Löwe, 1998, Engel, 1990], albeit for different sources of correlations.

In Chap. 3 and App. A, we discussed the application of the SCSNA technique to the Potts network with uncorrelated patterns. In the following sections we extend this analysis to the case of correlated patterns and try to get estimates of the storage capacity accounting for these correlations in an effective way. In this case, the variability of the noise has to be re-examined. We saw in Chap. 3 that the variance of the noise can be approximately written in the following way:

$$\langle (n_i^k)^2 \rangle = \frac{1}{(c_m a(1-\tilde{a}))^2} \sum_{\mu>1}^{p} \sum_{j(\neq i)=1}^{N} \sum_{l} \sum_{\mu'>1}^{p} \sum_{j'(\neq i)=1}^{N} \sum_{l'} c_{ij} c_{ij'} \langle v_{\xi_i^\mu,k}\, v_{\xi_i^{\mu'},k} \rangle \langle v_{\xi_j^\mu,l}\, v_{\xi_{j'}^{\mu'},l'}\, \sigma_j^l \sigma_{j'}^{l'} \rangle \,, \qquad (5.1)$$

where statistical independence between units is implicitly used. While in the case of uncorrelated patterns, all terms but $\mu = \mu', j = j'$ and $l = l'$ vanish, with correlated patterns this is not the case.

In this latter case, the additional terms $\mu \neq \mu', j = j'$ and $l = l'$ must be considered. Given the statistical independence of units, however, all other terms are zero.

## 5.1. Self-consistent signal to noise analysis

The calculation is for the most part identical to the one for uncorrelated patterns, we therefore refer to App. A. Following the same procedure as in Chap. 3, we can compute $\gamma_i^k = \gamma$ and $\rho_i^k = \rho$. The expression for $\gamma$ is exactly the same as with uncorrelated patterns

$$\gamma = \frac{\alpha}{S}\lambda\frac{\Omega/S}{1-\Omega/S}\,. \tag{5.2}$$

For the calculation of $\rho$ however, considering the additional terms, one finds

$$(\rho^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})}q\Big\{1+\frac{p\,\overline{C}_{as}}{S(1-\tilde{a})}\Big(\overline{C}_{as}-\tilde{a}\Big)\Big\}\Big\{1+2\lambda\Psi+\lambda\Psi^2\Big\}\,. \tag{5.3}$$

where $\alpha = p/c_m$ as before, and where $C_{as}$, defined in Sect. 4.2, is the fraction of units that are in the same Potts state, normalized by $a$. For uncorrelated patterns, this quantity is on average $\tilde{a}$, such that the second quantity in the first curly brackets is zero, and we recover the same quantity as Eq. (3.38). Note that this additional term scales with $p$ and is larger for higher-than-random ($\tilde{a}$) mean correlation between patterns. In Fig. 4.7 and more synthetically in Fig. 4.9 the patternwise correlation distribution can be seen for different values of the correlation parameters. This increase is the source of the decrease in capacity, with the most dramatic decrease being predicted by increasing the parameter $f$, affecting the mean of the patternwise correlation distribution (see Fig. (4.9b)). $\Omega$, $q$ and $\Psi$ are found to be, as in Chap. 3,

$$\Omega = \left\langle \frac{1}{N}\sum_j\sum_l \frac{\partial G_j^l}{\partial y^l}\right\rangle, \tag{5.4}$$

$$q = \left\langle \frac{1}{Na}\sum_{j,l}(G_j^l)^2\right\rangle, \tag{5.5}$$

$$\Psi = \frac{\Omega/S}{1-\Omega/S}\,. \tag{5.6}$$

where $\langle\cdot\rangle$ indicates the average over all patterns. The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k}m + \frac{\alpha}{S}\lambda\Psi(1-\delta_{k,0}) + \sum_{n=1}^{S} v_{n,k}z^n\sqrt{\frac{\alpha P_n}{S(1-\tilde{a})}q\Big\{1+\frac{p\,C_{as}}{S(1-\tilde{a})}\Big(C_{as}-\tilde{a}\Big)\Big\}\Big\{1+2\lambda\Psi+\lambda\Psi^2\Big\}} - U(1-\delta_{k,0})\,. \tag{5.7}$$

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

$$m = \frac{1}{a(1-\tilde{a})}\Bigg\langle \int D^S z \sum_{l(\neq 0)} v_{\xi,l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\Bigg\rangle_\xi\,, \tag{5.8}$$

$$q = \frac{1}{a}\Bigg\langle \int D^S z \sum_{l(\neq 0)} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\Bigg\rangle_\xi\,, \tag{5.9}$$

$$\Omega = \Bigg\langle \int D^S z \sum_{l(\neq 0)} \sum_{k} z^k \frac{\partial z^k}{\partial y^l} \prod_{n(\neq l)} \Theta(\mathcal{H}_l^\xi - \mathcal{H}_n^\xi)\Bigg\rangle_\xi\,. \tag{5.10}$$

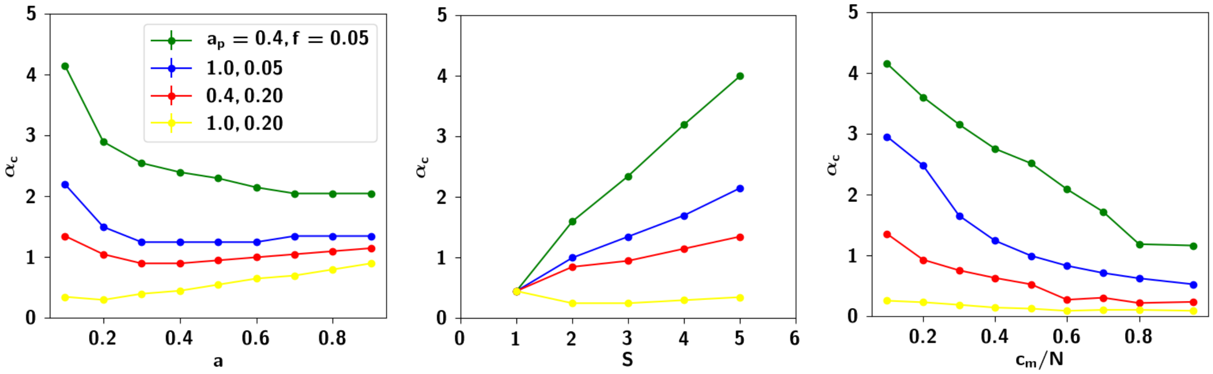
The equations can be solved numerically as before; however, the values of $C_{as}$ must be extracted from simulations, as the exact dependence of $C_{as}$ on the various correlation parameters $(a_p, f, \zeta)$ cannot be derived.

## 5.2. Simulation Results

### 5.2.1 Network parameters $a$, $S$ and $c_m$

In Fig. 5.1 we show the storage capacity for correlated patterns, for different values of the correlation parameters. As can be seen, increasing both of these parameters is detrimental to the capacity. The decrease in capacity brought on by the correlations is due to the variance of the noise, that has an additional term, scaling with $p$. However, there seems to be a qualitative change in the capacity behavior with $a$: increasing correlations, the capacity first decreases and soon reaches a constant level, and then *increases* with $a$ ($a_p = 1, f = 0.2$), as can be seen in Fig. 5.1a. In Fig. 5.1b instead, we can see the capacity as a function of the the number of Potts states $S$. For $S = 1$, as the algorithm produces uncorrelated patterns, the capacity remains the same, regardless of the correlation parameters. For higher values of $S$, on the other hand, the capacity strongly depends on the values of the correlation parameters $f$ and $a_p$. Another interesting point is the behavior of the capacity as a function of $c_m$. The simulations of Fig. 5.1c have been carried out with RD. In Chap. 3, we saw that with uncorrelated patterns, increased connectivity does not impact the capacity of the network as shown in Fig. 3.3. With correlations, this seems not to be the case: increasing the connectivity affects the ability of the network to retrieve.

The increased variability of the noise with correlated patterns, scaling with $p$, may require a higher threshold $U$ in order for the network to not go into overactivity. The analysis of computing the optimal threshold, carried out in Sect. 3.4.1 for uncorrelated patterns, may not hold anymore. To check for this possibility, we carried out the same simulations as in Fig. 3.1 for correlated patterns with parameters $a_p = 0.4$ and $f = 0.05$.



(a)

(b)

(c)

**Fig. 5.1: Storage capacity of correlated patterns. (a)** Storage capacity $\alpha_c$ as a function of the sparsity $a$ for different values of the correlation parameters $a_p$ and $f$. The storage capacity is defined as the critical storage at which half of all cued patterns are retrieved with overlap of $0.7$ and above. Increasing $a_p$ and $f$ are generally both detrimental to the capacity. While for low values of $a_p$ and $f$ (green curve) the capacity behaves qualitatively like that of uncorrelated patterns, for higher values of the parameters, even the qualitative behavior is different. For intermediate values of $a_p$ and $f$ after an initial decrease in capacity, the curves remain quasi constant with increasing $a$. For high values of $a_p$ and $f$ though, increasing the sparsity seems to even increase the capacity. **(b)** $\alpha_c$ as a function of the number of Potts states $S$ for $a = 0.1$. **(c)** $\alpha_c$ as a function of the connectivity $c_m$ for RD with the same parameters as above but $S = 5$ and $a = 0.1$. The capacity decreases as a function of increasing connectivity. This can be contrasted to uncorrelated patterns, where for this model it is found that the capacity remains quasi-constant with increasing $c_m$, at least for the parameters for which simulations were carried out. When not explicitly varied, parameters are $N = 2000$, $c_m = 200$, $a = 0.1$, $S = 5$, $U = 0.5$, $\beta = 200$, $\zeta = 10^{-6}$ and $\Pi = 150$.

### 5.2.2 The threshold $U$

In Fig. 5.2 we show the phase diagram of the fraction of retrievals in the $(U, p)$ space. One can see that the optimal threshold computed in Sect. 3.4.1 is not optimal in the case of correlated patterns, but nevertheless, increasing the threshold does not dramatically increase the capacity.

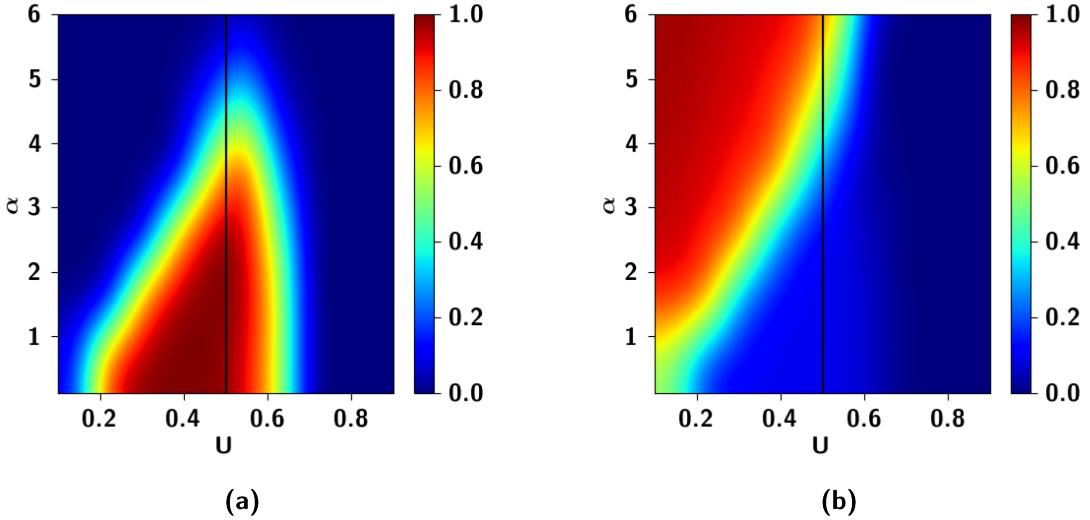


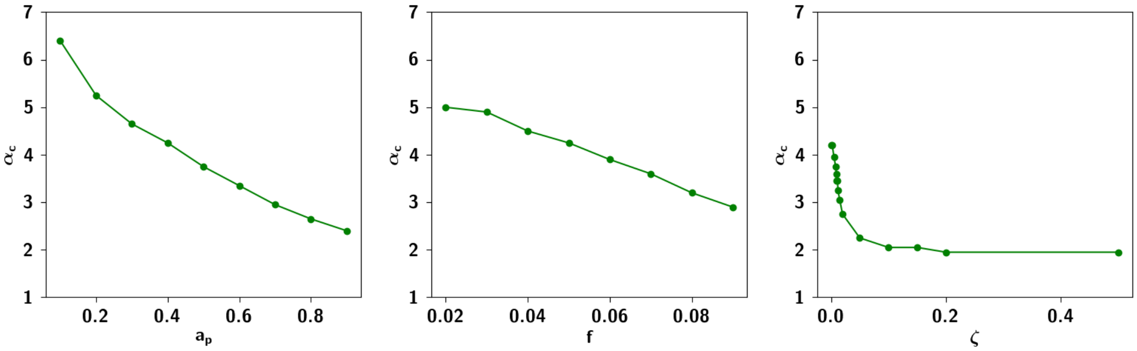
**Fig. 5.2: Storage load and threshold, phase diagram for correlated patterns. (a)** Phase diagram of the fraction of successful retrievals, defined as those trials where the cued pattern is retrieved with an overlap higher than 0.7. **(b)** Sparsity of the network, at the end of the retrieval dynamics. Network parameters are $N = 2000$, $c_m = 200$, $a = 0.1$, $S = 5$. Correlation parameters are $a_p = 0.4$, $f = 0.05$, $\zeta = 10^{-6}$ and $\Pi = 150$.

### 5.2.3 Correlation parameters $f$, $a_p$ and $\zeta$

In Fig. 5.3 we can see the storage capacity as a function of the three different correlation parameters $a_p$, $f$ and $\zeta$. We can see that increasing each of these parameters decreases capacity, albeit in very different manners. $\alpha_c$ decreases approximately linearly with increasing $a_p$, as shown in Fig. 5.3a. As we saw in the Chap. 4, $a_p$ is the degree to which children are similar to each of their individual parents. Increasing this parameter increases the similarity between those children receiving input from the same parents, increasing their overall similarity and therefore decreasing their discriminability. The dependence of $\alpha_c$ on $f$ is shown in Fig. 5.3b and it is somewhat similar to the effect of $a_p$.

On the other hand, $\alpha_c$ decreases dramatically with increasing $\zeta$. High values of $\zeta$ correspond to only a handful of parents out of the total giving significant weight to the activity of the children. Intuitively one might expect a behavior akin to that of small $f$. Its effect, however, is different:

when $f$ is small, the mean number of parents is affected. Increasing $\zeta$, the mean number of parents stays the same, only that for any given child, only the strongest are effective. This means that the strongest parent can still affect many children, increasing the correlation between a higher number of children, as opposed to a scenario in which $f$ is small and any one parent affects a small number of children. For very high $\zeta$, the strongest parents dominate the activity to an extent that those children affected by the strongest parents tend to become increasingly identical, but a small fluctuation in the activity away from the cued pattern will result in the trial of the simulation to be discarded as a successful cued retrieval, and instead counted as a correlated retrieval, as we show in more in detail in the next section.



(a) (b) (c)

**Fig. 5.3: Storage capacity of correlated patterns, correlation parameters.** Storage capacity curves as a function of various correlation parameters. **(a)** $\alpha_c$ as a function of $a_p$. **(b)** $\alpha_c$ as a function of $f$. **(c)** $\alpha_c$ a sa function of $\zeta$. When not explicitly varied, the correlation parameters are $a_p = 0.4$, $f = 0.05$, $\zeta = 10^{-6}$ and $\Pi = 150$. Network parameters are $N = 2000$, $c_m = 200$, $a = 0.1$, $S = 5$.

### 5.2.4 Correlated retrieval

In the previous section we saw that correlations decrease the capacity of the network to retrieve. But when the network fails to retrieve the cue, what is the configuration that it settles into? Correlations cause basins of attraction to get closer in configuration space and possibly to merge, raising the possibility that other patterns, similar to the ones cued, are retrieved. We carried out simulations with correlated patterns for different values of $\zeta$. On the left $y$-axis of Fig. 5.4a, we can see the fraction retrieved with overlap above $m^* = 0.9$ as a function of increasing $\alpha$. It becomes apparent that with low values of $\zeta$ the fraction retrieved falls abruptly after the capacity limit is reached, while for larger values of $\zeta$, the fraction retrieved falls more slowly, but starting at smaller values of the loading $\alpha$.

On the right $y$-axis of the same figure, we can see the fraction of retrieved patterns, with overlap above $m^* = 0.9$, different from the ones cued. In this figure, this fraction remains identically zero. In Fig. 5.4b however, we plot the fraction retrieved with overlap above $m^* = 0.7$. Reducing the threshold for successful retrieval $m^*$, we observe *correlated retrieval.* The fraction of patterns retrieved, but not cued, increases and reaches 1. This effect happens abruptly with $\zeta$, in the sense that while for $\zeta = 0.01$ the network does not retrieve at all, already for $\zeta = 0.02$, the fraction retrieved of a second, correlated pattern reaches 1. The network can retrieve patterns only with an approximate upper precision, that we can attempt to roughly estimate by analyzing the overlap, the order parameter with which we measure "successful retrieval". The overlap writes:

$$m^{\mu} = \frac{1}{Na(1-\tilde{a})} \sum_{i=1}^{N} \sum_{k=1}^{S} (\delta_{\xi_i^{\mu},k} - \tilde{a}) \sigma_i^k \tag{5.11}$$

where $\mu$ denotes the pattern to be retrieved ($m = O(1)$) and $\sigma_i^k$ the activity in state $k$ of unit $i$ of the network. During retrieval dynamics, under the effect of noise (Eq. (5.3)), units switch from cued states to other states. We can, however, make the approximation that transitions occur only between active and quiescent states and neglect the potential transitions occurring between one active state and another active state [1].

The sums over $\delta_{\xi_i^{\mu},k}\, \sigma_i^k$ yield at most $Na$. In this case all units that are active in the pattern remain active in the network, and in the same state. If among those that are active in the pattern, a fraction $b$ become quiescent, then the sums over $\delta_{\xi_i^{\mu},k}\, \sigma_i^k$ yield $Na(1-b)$. This is one way the overlap is diminished. The other possibility is "overheating", meaning that among those that are quiescent in the pattern, a fraction $c$ become active: the sums over $\tilde{a}\sigma_i^k$ yield $N(1-a)c$, so we can

[1] This approximation is based on the mean values of the fields in the direction of the different states: units that should be in a given active state receive both the signal $1 - a/S$ as well as noise, while an active state that should not be active in the pattern receives only noise of mean zero. The threshold, that we set here to 0.5 is in between, such that transitions occur more frequently between active and quiescent states.

rewrite the overlap using the two quantities $b$ and $c$[2] :

$$m(b,c) = \frac{1}{(1-\tilde{a})}\left\{1 - b - \frac{1}{S}(a + (1-a)c)\right\} \tag{5.12}$$

To compute approximately the upper precision of retrieval[3], we are in the situation in which all units become active ($b = 0$ and $c = 1$) such that we get $m_u \equiv m(b = 0, c = 1) = \dfrac{(1-1/S)}{(1-a/S)}$. With the parameters of the simulations ($a = 0.1$, $S = 5$), $m_u \sim 0.82$. In Fig. 5.4a and Fig. 5.4b, we plot the fraction of cued retrieval and correlated retrieval for two different criteria of successful retrieval, $m^* = 0.9$ and 0.7, respectively *above* and *below* the upper precision of retrieval. It is apparent from the figure that decreasing the criterion from $m^* = 0.9$ to 0.7, the fraction of correlated retrieval goes from 0 to 1, consistent with the value of 0.82 we estimated.

Note that the resulting overlap ($m$) after retrieval dynamics is a random variable with a distribution $P(m)$ from which the fraction retrieved can be defined as $f_r = \int\limits_{m^*}^{1} dm\, P(m)$ and is a cumulative quantity that depends on $m^*$. It may then be useful to use a measure that is independent of the exact values with which we define $m^*$. One such measure is the mutual information between the cued pattern and the configuration the network settles into. Information measures have the advantage of being computed using the probability distributions themselves, and as logarithmic measures, are highly sensitive to the higher order moments of the distributions.

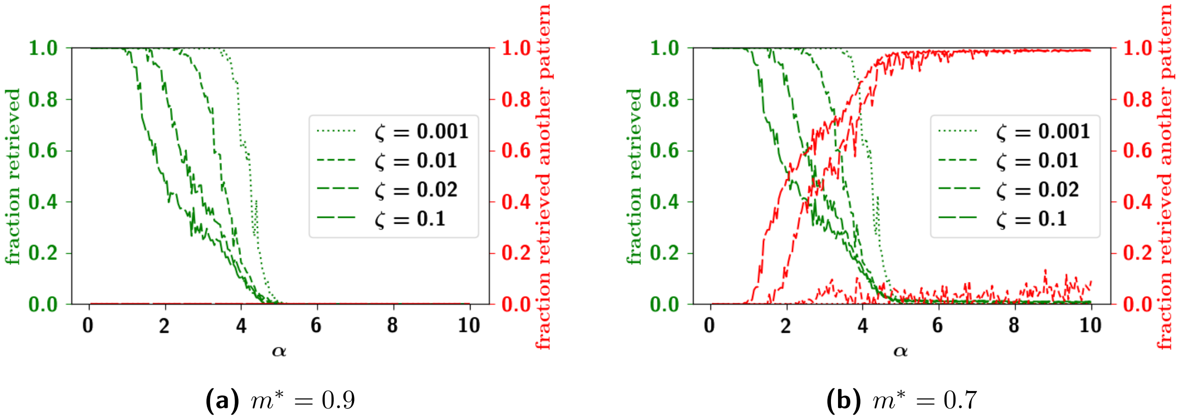
## 5.3. Information

The mutual information between the pattern cued (c) and the configuration in which the network settles (r) writes:

$$I(c,r) = \sum_{k,l=0}^{S} C^{kl}(c,r) \log_2\left(\frac{C^{kl}(c,r)}{C^k(c)C^l(r)}\right) \tag{5.13}$$

[2] Looking at the limit cases of the average overlap: $m(b = c = 0) = 1.$ and $m(b = c = 1) = -\frac{1}{S(1-a/S)}$. Indeed if all units that should be active stay active and all those that stay quiescent stay quiescent, then the network activity corresponds to a pattern. In contrast, the inverse situation yields a low, negative value of the overlap, that is practically never seen in simulations, for reasonable values of the parameters. The most common situation is "overheating", when $b = 0$ and $c = 1$.

[3] The crucial observation is that beyond a certain value of the loading $p$, the network is not able to keep the sparsity at the constant level of the patterns, as shown in Fig. 5.5: the sparsity ramps until reaching the maximal value possible of 1.



**(a)** $m^* = 0.9$

**(b)** $m^* = 0.7$

**Fig. 5.4: Cued and correlated retrieval. (a)** Left $y$-axis: the fraction of retrieved patterns (with overlap higher than $0.9$), as a function of the storage load $\alpha$, for different values of $\zeta$. While for low values of $\zeta$, the fraction retrieved falls abruptly after a critical value of the loading, for higher values of $\zeta$, the capacity falls more slowly, though starting at smaller values of the storage load $\alpha$. **(b)** Left $y$-axis: the fraction of retrieved patterns (with overlap higher than $0.7$), as a function of the storage load $\alpha$. Right $y$-axis: fraction of retrieved patterns, not corresponding to those cued. Note that with the more stringent criterion for successful retrieval as in **(a)**, there is no correlated retrieval. Network parameters are $N = 2000$, $c_m = 200$, $S = 5$, $a = 0.1$, $U = 0.5$, $\beta = 200$. Correlation parameters are $a_p = 0.4$, $f = 0.05$ and $\zeta = 10^{-6}$.

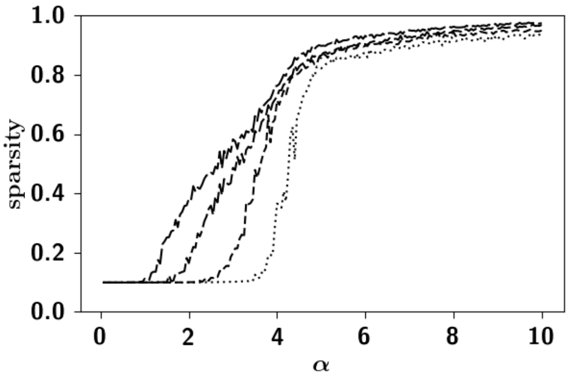


**Fig. 5.5: "Overheating" of the network.** The sparsity of the network, as a function of the storage load of the network, for four different values of the parameter $\zeta$. For low values of $\zeta$, the sparsity increases abruptly when capacity is reached. For higher values of $\zeta$, the sparsity increases more slowly, but starting at lower values of the loading $\alpha$. Parameters are identical to those of Fig. 5.4.

where

$$C^{kl}(c,r) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i^c,k}\sigma_i^l\,, \tag{5.14}$$

$$C^{k}(c) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i^c,k}\,, \tag{5.15}$$

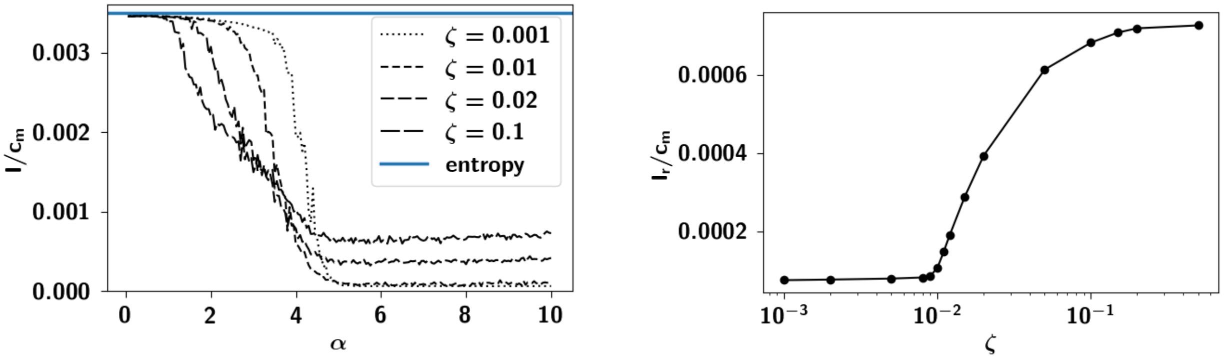
$$C^{l}(r) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^l\,. \tag{5.16}$$

The maximum value of this quantity is attained when the cued pattern is also the one retrieved: $c = r$. In this case the mutual information reduces to

$$I(c) = \sum_{k=0}^{S} C^k(c)\log_2\left(\frac{1}{C^k(c)}\right) = \Big\{ -(1-a)\log_2(1-a) + a\log_2(S/a)\Big\}\,, \tag{5.17}$$

that we recognize to be the entropy of the cued pattern. In Fig. 5.6a we can see the mutual information as a function of the loading $\alpha$ for different values of the parameter $\zeta$, averaged across cued retrieval of many patterns. The effects of $\zeta$ on the mutual information is analogous to the effect observed for the fraction retrieved. The mutual information has a sharp fall-off upon increasing $\alpha$, which is more and more abrupt as $\zeta$ decreases. For small values of $\alpha$, the mutual information which is a plateau does not depend on $\zeta$: the value of the mutual information at this plateau corresponds to the entropy.

Perhaps the most interesting observation is the *residual information*, the remaining constant information, after capacity collapse. In Fig. 5.6b, this residual information is plotted as a function of the parameter $\zeta$, and it can be seen that it increases sharply at $\zeta_c \sim 0.02$, before saturating at a given value. This effect is reminiscent of a phase transition with control parameter $\zeta$ and where the information plays the role of the order parameter. Below the critical value of $\zeta_c$, once the capacity is depassed, there is no more retrievable information at all. Above $\zeta_c$ however, the network retrieves some information about the cued pattern. This non-zero residual information is tightly linked to correlated retrieval of the previous section, and the saturation of the information is linked to the upper precision of the overlap that we approximately computed.



**(a)** **(b)**

**Fig. 5.6: Information (a)** Information per synapse as a function of the storage load $\alpha$, for different values of the parameter $\zeta$. For low values of $\zeta$, the information decays abruptly, at a value of the storage load $\alpha$, while for larger values of $\zeta$, we observe a more gradual decay, starting at lower values of the storage load. For high *enough* values of $\zeta$ however, the information does not go to zero, but rather saturates at a given constant value. We call this *residual information*. In **(b)**, we plot this residual information as a function of $\zeta$. The sharp increase of this residual information at $\zeta_c \sim 0.02$ suggests a phase transition giving rise to two regimes. In the first, with $\zeta < \zeta_c$, the residual information is approximately zero. In the second, $\zeta > \zeta_c$, this information is non-zero and reflects that the network, though not being able to retrieve the *fine* structure of the memory cued, still manages to retrieve the *gross* structure. Simulations carried out with parameters identical to those of Fig. 5.4.

## 5.4. Discussion

The findings of this chapter have interesting interpretations for both the question of encoding and that of retrieval within the more general framework of learning. There is now a growing consensual view of the cortex as a slow memory system that uses overlapping distributed representations to represent the general statistical structure of the environment.

How does the cortex extract and encode the general statistical structure of the ensemble of stimuli that it receives? Far from having a definitive answer to this question, it has been suggested that the interaction between the hippocampus and the cortex is a crucial element in the consolidation of memories. The general idea is that memories are first stored in the hippocampal system via synaptic changes and that these support the reinstatement of recent memories in the neocortex. Neocortical synapses are slightly modified on each reinstatement and the gradual, neocortical changes accumulating over time encode remote memory. This organization hypothetically allows the hippocampal system to rapidly encode new items without disrupting this structure, and allows the cortex to slowly integrate memories in a structured way from ensembles of experiences that are otherwise not necessarily organized in any way. This view is in particular supported by evidence that damage to the hippocampal system results in recent memory disruption but leaves remote memory intact.

However, early modelling attempts typically resorted to backpropagation to account for the structured learning of the cortex [McClelland et al., 1995]. Backpropagation has been widely challenged on the basis that it lacks a plausible biological mechanism [4] [Crick, 1989, Zipser, 1988]. While hippocampal learning in these accounts was taken to fit the framework of learning unrelated patterns of activity, it remains unclear how to model neocortical learning. Our account offers a plausible framework for neocortical learning, in which semantic structure is extracted progressively from the statistics of parents and encoded in the cortex via Hebbian learning.

Taken at face value, the diminished capacity accompanied by the emergence of correlated retrieval suggests that this ability for generalization comes at the cost of losing the resolution with which we can retrieve the individual memories. However, this result as such is incomplete, and must be taken also in relation to the differential role of other memory structures and in particular the hippocampus in retrieval. For example, in humans, it has been shown that the ventral hippocampus projects directly to the medial prefrontal cortex, providing an immediate route for hippocampal representations to arrive to the prefrontal cortex, suggesting a model of bidirectional hippocampus/prefrontal cortex interactions that support context-dependent memory retrieval [Preston and Eichenbaum, 2013].

It should be noted that correlated retrieval does not imply the complete loss of information about the cued memory item, as evidenced by the residual information. The residual information is specifically the semantic component of the cued memory, in that even when the *specifics* about it is compromised, the *gross* information about it is still retrievable.

[4] From [O'Reilly and Rudy, 2001]: "Specifically, backpropagation requires that an error value is propagated backwards from the dendrite of a receiving neuron, across the synapse, into the axon terminal of the sending neuron, down the axon of this neuron, and then integrated and multiplied by some kind of derivative, and then propagated back out of its dendrites."

# 6

# Latching dynamics

In the previous chapters we have studied the properties of the Potts network, operating solely under retrieval dynamics. It has been noted that incorporating time-dependent thresholds to the dynamics of units that depend on their activity, one observes a destabilization of the memory attractors, that can lead to the dynamic activation of other memories [Horn and Usher, 1989,Herrmann et al., 1993]. Such time-dependent thresholds have been intended as models of adaptation and inhibition and have been used to study the Potts network [Treves, 2005,Kropff and Treves, 2007,Russo et al., 2008]. We revise the simple updating rule, Eq. (2.3) to

$$\sigma_i^k = \frac{\exp\left(\beta r_i^k\right)}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U)\right]} \tag{6.1}$$

and

$$\sigma_i^0 = \frac{\exp\left[\beta(\theta_i^0 + U)\right]}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta(\theta_i^0 + U)\right]}, \tag{6.2}$$

where $r_i^k$ is the input to (active) state $k$ of unit $i$ integrated over a time scale $\tau_1$, while $U$ and $\theta_i^0$ 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. $\theta_i^0$ varies with time constant $\tau_3$, to describe local network adaptation and inhibitory effects. The stiffness of the local dynamics is parametrized by the inverse "temperature" $\beta$ (or $T^{-1}$), which is then distinct from the standard notion of thermodynamic noise. The input-output relations Eq. (6.1) and Eq. (6.2) ensure that

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

In addition to the overall threshold, $\theta_i^k$ is the threshold for unit $i$ specific to state $k$, and it varies with time constant $\tau_2$, representing adaptation of the individual neurons active in that state, i.e., their neural or even synaptic fatigue. The time evolution of the network is then governed by equations that include three distinct time constants:

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t), \tag{6.3}$$

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t), \tag{6.4}$$

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t), \tag{6.5}$$

where the field that the unit $i$ in state $k$ experiences is

$$h_i^k = \sum_{j \neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l + w \left( \sigma_i^k - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l \right). \tag{6.6}$$

The *local feedback term* $w$ is a parameter, first introduced in [Russo and Treves, 2012], 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 $\tau_1$, $\theta_i^k$ is negative feedback mediated by fatigue at the slower time scale $\tau_2$, while $\theta_i^0$ is also negative, and it can be used to model both fast ($GABA_A$) and slow ($GABA_B$) 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 $\theta_i^0$ that varies rapidly from one that varies slowly, but it would greatly complicate the observations presented in the following.

Under such dynamics, the main behavior of the network is that of spontaneous hopping of the network from one attractor to another, guided, among others, in the slow regime, by correlations between the different memory attractors, as shown in Fig. 6.1. One of the properties studied in [Russo and Treves, 2012] was the length of the latching sequences as a function of various network parameters. In particular, phase boundaries were found in the $w - T$ plane, marking the onset of different latching regimes, no latching, finite latching and infinite latching.

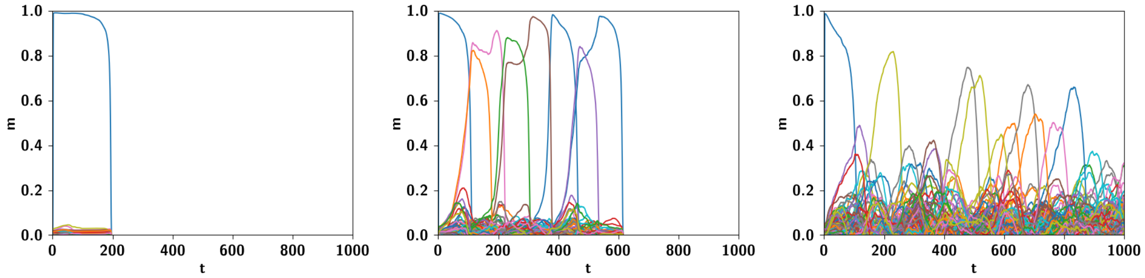
More recently, [Kang et al., 2017] found bands in $p - S$ and $p - C$ planes, where lengthy latching sequences co-exist together with good retrieval of each individual attractor visited by the network. In Fig. 6.3 we show phase diagrams of a quantity, the normalized latching length, multiplied by the average overlap of the network with the attractors visited. Co-existence of high sequence length with good retrieval is possible in a confined area of the phase space of parameters. An example of a latching sequence can be seen in Fig. 6.1. Increasing the number of learned patterns, from $p = 50$ to $p = 90$ to $p = 200$, the length of the latching sequence increases, but eventually to the detriment of the quality of retrieval. [Russo and Treves, 2012] and more recently we have obtained detailed phase diagrams of the statistical properties of latching dynamics, extending their simulations to patterns correlated by the way of our algorithm [Kang et al., 2017]. However, apart from phase diagrams, simple plots of the evolution of the overlaps suggest a rich and qualitatively different behavior of the latching network, depending on the correlation parameters, that need a more comprehensive analysis by introduction of order parameters, beyond the aggregate measures we studied. It should be noted that latching-like behavior with correlations have been studied in binary networks [Herrmann et al., 1993], but for a different structure of correlations.

## 6.1. Latching with uncorrelated patterns

In Sect. 3.3, we computed the standard deviation of the noise received by unit $i$ in state $k$, Eq. (3.38). Under the assumption of good retrieval we have, approximately that $q \sim 1$, and $\Omega \sim 0$, such that $\Psi \sim 0$. We can then write

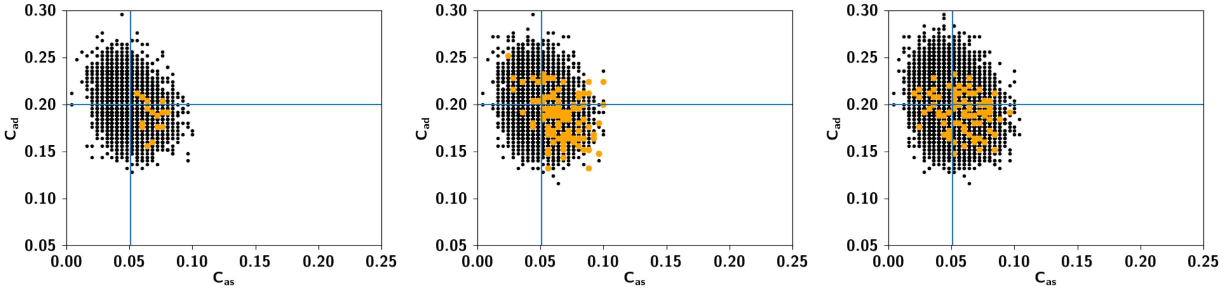
$$\rho \sim \sqrt{\frac{(p-1)a}{c_m S^2}} \, . \tag{6.7}$$

This approximate relation holds under the assumption of uncorrelated patterns. We can define a "modified" signal, one that incorporates the field in the direction of another pattern $\rho$, the one that



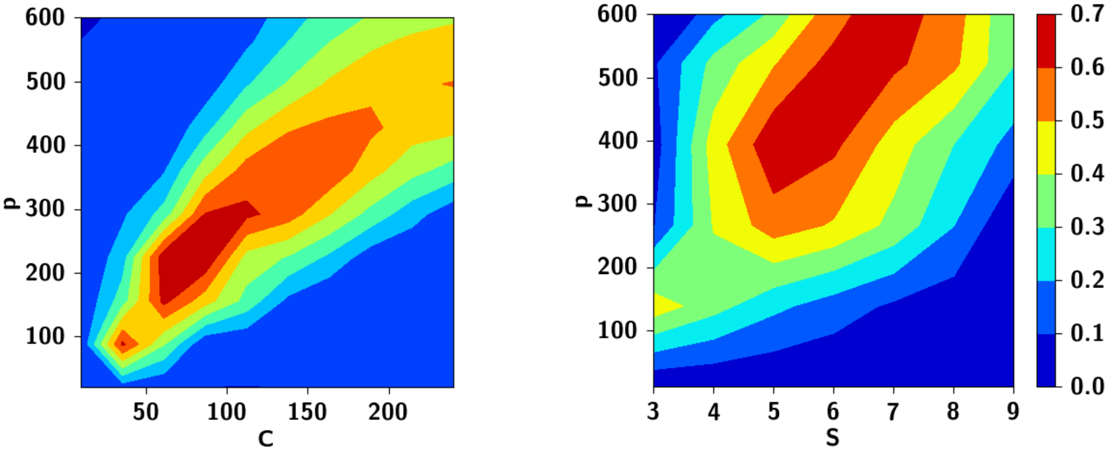
**(a)** $p = 50$ **(b)** $p = 90$ **(c)** $p = 200$

**Fig. 6.1: Latching sequences of uncorrelated patterns.** Three sample latching sequences with uncorrelated patterns of a network in the slow latching regime ($\tau_1 = 3.33$, $\tau_2 = 100$ and $\tau_3 = 10^6$). The $x$-axis corresponds to time, as measured by units of network updates. The $y$-axis corresponds to the overlap. Different colors correspond to different patterns. Increasing $p$, one observes different latching regimes. For too low $p$, in the *no latching regime*, there is only retrieval and the network cannot latch onto another pattern. Increasing $p$, one reaches the *finite latching regime*, where one observes a finite sequence of well retrieved patterns. Increasing $p$ even further, one reaches the *infinite latching regime*, where sequences are indefinitely long but where the network cannot retrieve any of them very well. Network parameters are $N = 1000$, $S = 5$, $a = 0.25$, $c_m = 150$, $U = 0.1$, $\beta = 11$, $w = 0.8$.



**(a)** $p = 100$ **(b)** $p = 150$ **(c)** $p = 200$

**Fig. 6.2: Scatter of correlation with uncorrelated patterns.** $C_{as}$ versus $C_{ad}$ for all pairs of patterns (black points) and only those pairs with a latching transition (yellow points). We consider a latching transition between two patterns to have occurred if one pattern overcomes another with an overlap of $0.5$ or higher. The blue vertical and horizontal lines correspond to, respectively the average $C_{as}$ and $C_{ad}$. For low values of loading $p$, the transitions span the lower right quadrant where $C_{as}$ is higher than average, and $C_{ad}$ is lower than average. As loading increases, transitions start to span the space of correlations more evenly.



(a)

(b)

**Fig. 6.3: Phase diagrams of retrieval and latching co-existence.** Phase diagram of a measure of latching ability multiplied with retrieval ability. Warmer colors correspond to regions in which the network is able to both *retrieve* patterns well (with high correlation with a memory item) and *latch* onto others, forming long sequences of transitions.

is most highly correlated with the one cued, $\nu$:

$$s_i^k = \frac{1}{Ca(1-\tilde{a})}\left\{ v_{\xi_i^\nu,k}(1-\tilde{a}) \sum_{j(\neq i)}^{N} \sum_{l=1}^{S} c_{ij}\delta_{\xi_j^\nu,l} + v_{\xi_i^\rho,k} \sum_{j(\neq i)}^{N} \sum_{l=1}^{S} c_{ij} v_{\xi_j^\rho,\xi_j^\nu}\, \delta_{\xi_j^\nu,l} \right\} \tag{6.8}$$

This can be rewritten using the correlation measures we defined in Sect. 4.2:

$$s_i^k \sim v_{\xi_i^\nu,k} + \frac{v_{\xi_i^\rho,k}}{(1-\tilde{a})}\left[ C_{as}(1-\tilde{a}) + C_{ad}(-\tilde{a}) + C_{a0}(-\tilde{a}) \right], \tag{6.9}$$

where $C_{as}$, $C_{ad}$, and $C_{a0}$ have been defined in Sect. 4.2. Normalizing these values with the expectation value for uncorrelated patterns, we define:

$$\begin{aligned} \Gamma_{as} &= \frac{C_{as}}{a/S} \\ \Gamma_{ad} &= \frac{C_{ad}}{a(S-1)/S} \\ \Gamma_{a0} &= \frac{C_{a0}}{(1-a)}, \end{aligned} \tag{6.10}$$

such that we can rewrite the signal

$$s_i^k \sim v_{\xi_i^\nu,k} + \frac{v_{\xi_i^\rho,k}}{(1-\tilde{a})}\left[(\Gamma_{as}-\Gamma_{a0})\frac{a(1-a)}{S} + (\Gamma_{as}-\Gamma_{ad})\frac{a^2(S-1)}{S^2}\right] \tag{6.11}$$

Let us consider those units that are inactive in the current attractor $\nu$ and active in the upcoming attractor $\rho$. We then have:

$$s \sim \tilde{a}\{-1 + (\Gamma_{as}-\Gamma_{a0})(1-a) + (\Gamma_{as}-\Gamma_{ad})\tilde{a}(S-1)\} \tag{6.12}$$

It turns out that with uncorrelated patterns, the increased field in the direction of a second, highly correlated pattern is what leads the transitions, as can be seen in Fig. 6.2. In this figure, a scatterplot $C_{as}$ versus $C_{ad}$ for all pairs of patterns (black points) and only those pairs with a latching transition (yellow points) can be seen. For low values of loading $p$, the transitions span the lower right quadrant where $C_{as}$ is higher than average, and $C_{ad}$ is lower than average. As loading increases, transitions start to span the space of correlations more evenly. [Russo et al., 2008], [Russo and Treves, 2012] studied the onset of latching in the $w-T$ plane as well as a detailed study of the dependence of latching length as a function of network parameters. In Chap. 5, we have shown the capacity of the network to be lower with correlated patterns. We have shown recently [Kang et al., 2017] however, that the network can also latch in the presence of correlations, but it remains unclear whether the latching sequences are in any way different from those of uncorrelated patterns.

## 6.2. Latching with correlated patterns

In Sect. 5.1 we saw that with correlations, the variability of the noise scales differently. In the regime of good retrieval, we can write

$$\rho \sim \sqrt{\frac{(p-1)a}{c_m S^2}\left\{1 + \frac{p\overline{C}_{as}}{S(1-\tilde{a})}\left(\overline{C}_{as}-\tilde{a}\right)\right\}}\,. \tag{6.13}$$

The signal writes as

$$s \sim -\tilde{a} + (\Gamma_{as}-\Gamma_{a0})\frac{a(1-a)}{S} + (\Gamma_{as}-\Gamma_{ad})\frac{a^2(S-1)}{S^2}\,. \tag{6.14}$$

### 6.2.1 The effect of $f$

These approximate values of the modified signal and the noise indicate a quantitative change with respect to uncorrelated patterns, but the figures reported to App. D suggest some qualitative differences in the kinds of sequences generated by the network. In the figure, we observe latching sequences for three different values of the loading $p$, truncated to 1000 network updates. An observation proper to this value of $f$ is *parallel retrieval* of patterns. In the same figure, a scatterplot of the $C_{as}$ versus $C_{ad}$ correlations suggest that parallel retrieval occurs with patterns that are highly correlated with one another. For this value of $f$, however, the corresponding average number of parents per pattern is $n_p \sim 1.5$ and many patterns have single parents and share this parent, as in a hierarchical setting. The high clustering of the correlation entails parallel retrieval of those patterns being in the same cluster: the network does not discriminate between the different patterns in the same cluster. The same sequence, truncated to 5000 network updates can be seen in Fig. 6.4a.

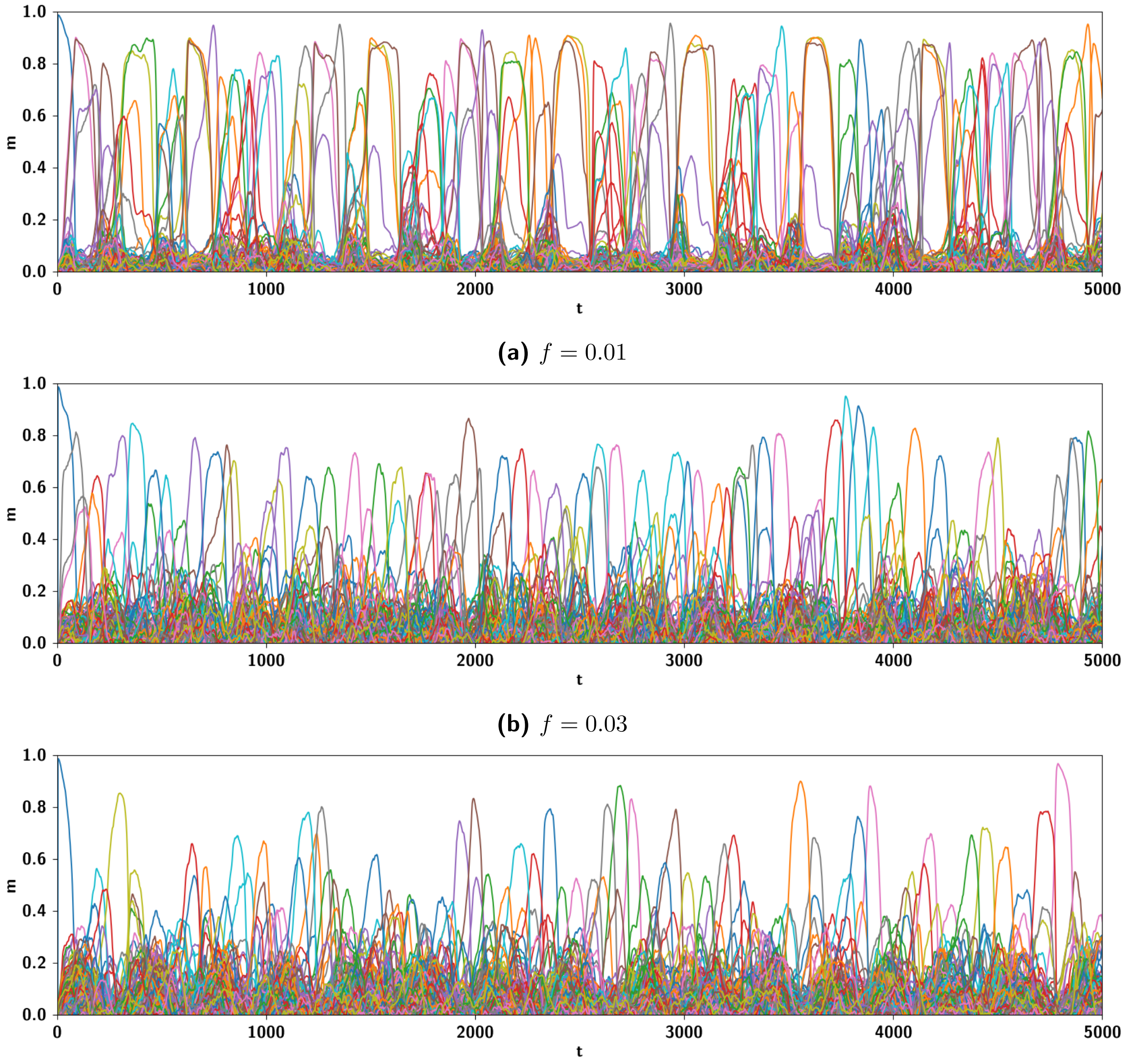
The multi-parent algorithm benefits then from additional motivation, from the perspective of latching dynamics. Allowing for $f$, and with it the average number of parents $n_p$ to increase, one does not observe this behavior anymore, as shown in the same figure. The network now latches *between* different patterns, and those that are highly correlated with one another, as shown in the scatterplots in App. D. Increasing this parameter $f$ even further, the storage capacity is compromised, and with it the *quality* of latching [Kang et al., 2017].

### 6.2.2 The effect of $a_p$

In Fig. 6.5 we report latching sequences for three different values of $a_p$, for higher loading $p = 200$. While for lower values of $a_p = 0.1$, we have a sequential activation of memories (though with lower quality, due to the higher loading), for higher values of $a_p$, especially with $a_p = 1.0$, we have the appearance of oscillations: the same cluster of patterns is periodically activated.

### 6.2.3 The effect of $\zeta$

In Fig. 6.6, we report latching sequences for three different values of $\zeta$, chosen to be before and after the cusp of transition shown in Fig. 5.6b. For lower values of $\zeta = 0.01$, we see a activation of clusters, or parallel retrieval, but with a more variable distribution of the overlaps than what one observes in Fig.6.4a. At the cusp, $\zeta = 0.015$, we observe a very interesting behavior: the network

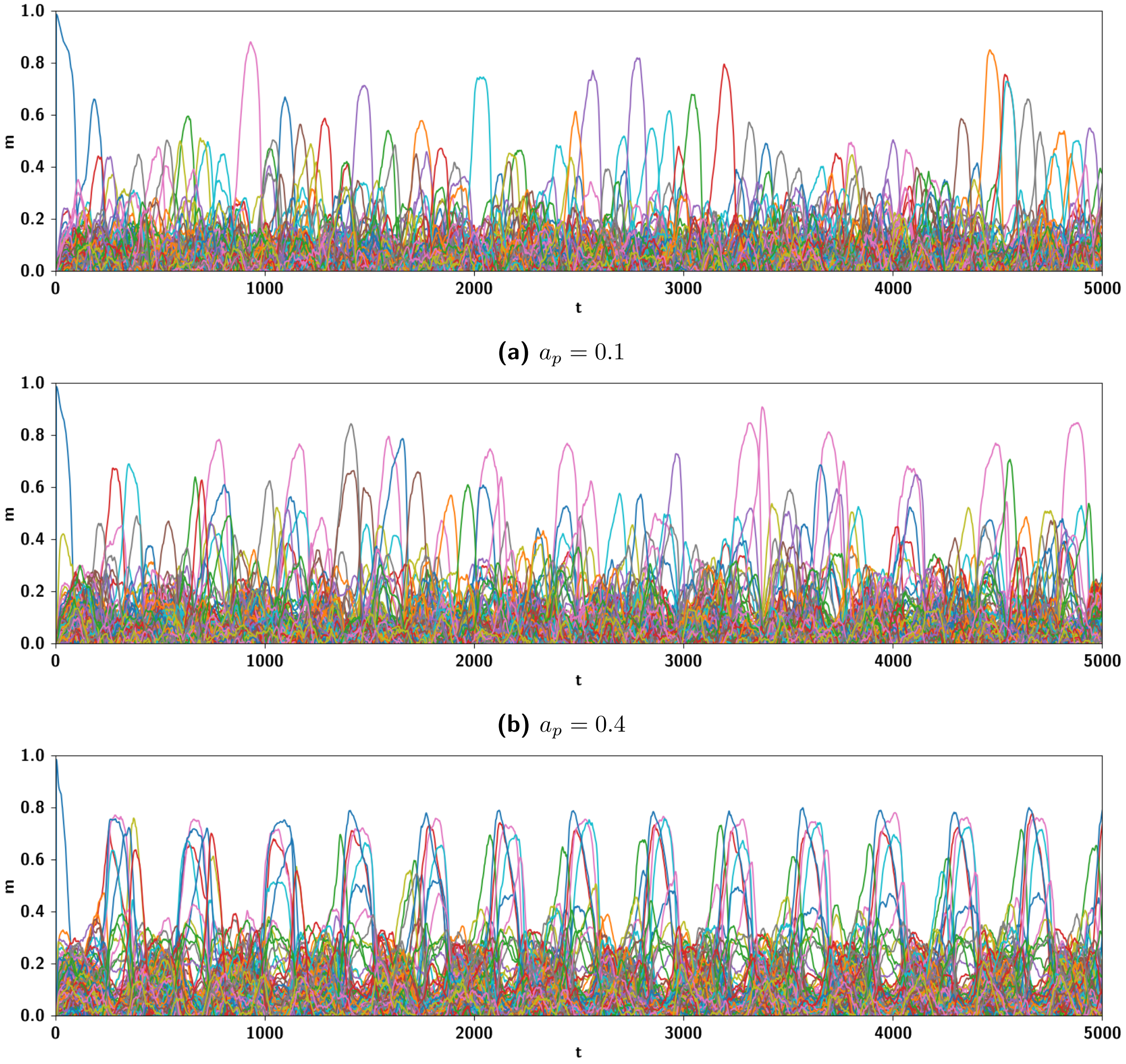


**(a)** $f = 0.01$

**(b)** $f = 0.03$

**(c)** $f = 0.05$

**Fig. 6.4: Latching chains. (a)** One example of a latching chain with $f = 0.01$ and $p = 150$. We observe parallel retrieval of patterns, or clusters. One such cluster that appears frequently is the (yellow-orange-brown). Another is (yellow-green). Yet another is (pink-brown). With this value of the $f$ parameter, many children share a single common parent, such that they form clusters. They are retrieved with high overlap. Transitions *between* clusters, or patterns tend to occur between those that do not share common parents, and are therefore randomly correlated. **(b)** Latching chain with $f = 0.03$. Patterns are not retrieved in parallel anymore, and when they are, are typically retrieved with lower overlap. **(c)** Latching chain with $f = 0.05$. Another decrease in the frequency of patterns retrieved with high overlap, respect to Fig. 6.4b. The oscillatory behavior of the uncondensed patterns starts to manifest itself. Network parameters are identical as in Fig. 6.1, correlation parameters are $a_p = 0.4$, $\zeta = 0$ and $\Pi = 150$.



**(a)** $a_p = 0.1$

**(b)** $a_p = 0.4$

**(c)** $a_p = 1.0$

**Fig. 6.5: Latching chains.** **(a)** Latching chain with $a_p = 0.1$ and $p = 200$. The network cannot retrieve individual patterns very well, due to the (relatively) high loading. **(b)** Latching chain with higher sparsity of input from parents $a_p = 0.4$. **(c)** $a_p = 1.0$. We observe oscillations of clusters of highly correlated patterns. Network parameters are identical as in Fig. 6.1, and correlation parameters are $f = 0.03$, $\zeta = 0$ and $\Pi = 150$.

retrieves clusters, but there are bouts in which it oscillates around a cluster of patterns. Increasing $\zeta$ further, the network goes into a regime in which it is not able to retrieve any pattern anymore, but oscillates indefinitely around the same cluster of patterns.

### 6.2.4 The emergence of a new, oscillatory phase?

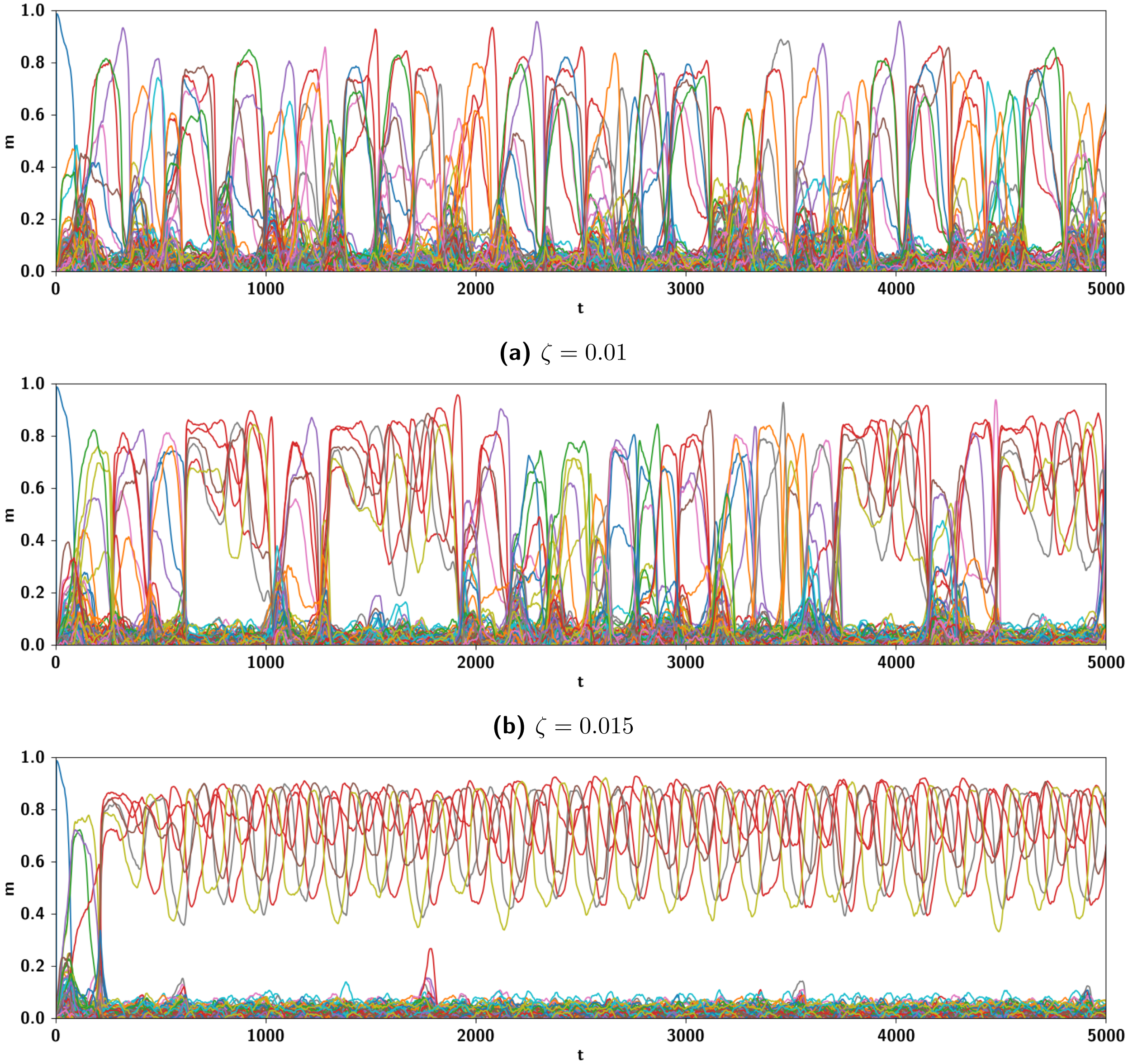
One observation, proper to sequences of latching with correlated patterns, is the appearance of oscillatory behavior of the non-condensed overlaps, appearing notably at higher loading. We have computed two quantities:

$$\overline{\sigma}_0(t) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^0(t)\,, \tag{6.15}$$

$$\overline{q}(t) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{S} \frac{(\sigma_i^k(t))^2}{(1-\sigma_i^0(t))^2} \tag{6.16}$$

that measure respectively, the mean activity in the null state and the mean spread of the activity in the different Potts states. High values of $\overline{q}$ means that activity is concentrated in a few Potts states, while a low value means that activity is spread in different Potts states. We have plotted these two values for the loading $p = 200$ for the correlation parameters $a_p = 0.4$ and $f = 0.05$, and as a control, for uncorrelated patterns. Figs. 6.7a and 6.7b suggest that this behavior is proper to correlated patterns. The appearance of such oscillations is probably linked to the additional noise: as the network attempts to retrieve a pattern, correlations increase the noisy activity, and it takes a while for the adaptive thresholds to lower this noisy activity. From the resulting "attractor ruins" [Tsuda, 2001], the network then retrieves a new pattern, one that has been subjected to a less harsh treatment from the adaptive thresholds.

The simulations presented in this chapter, while purely descriptive, point therefore to a rich latching behavior of even the simplest version of the Potts network, once it operates with correlated patterns. It is possible that new dynamical phases be identified, and that a quantitative analysis along the lines just sketched here may be developed and applied to characterize a novel phase transition, potentially relevant to understand real cortical dynamics.

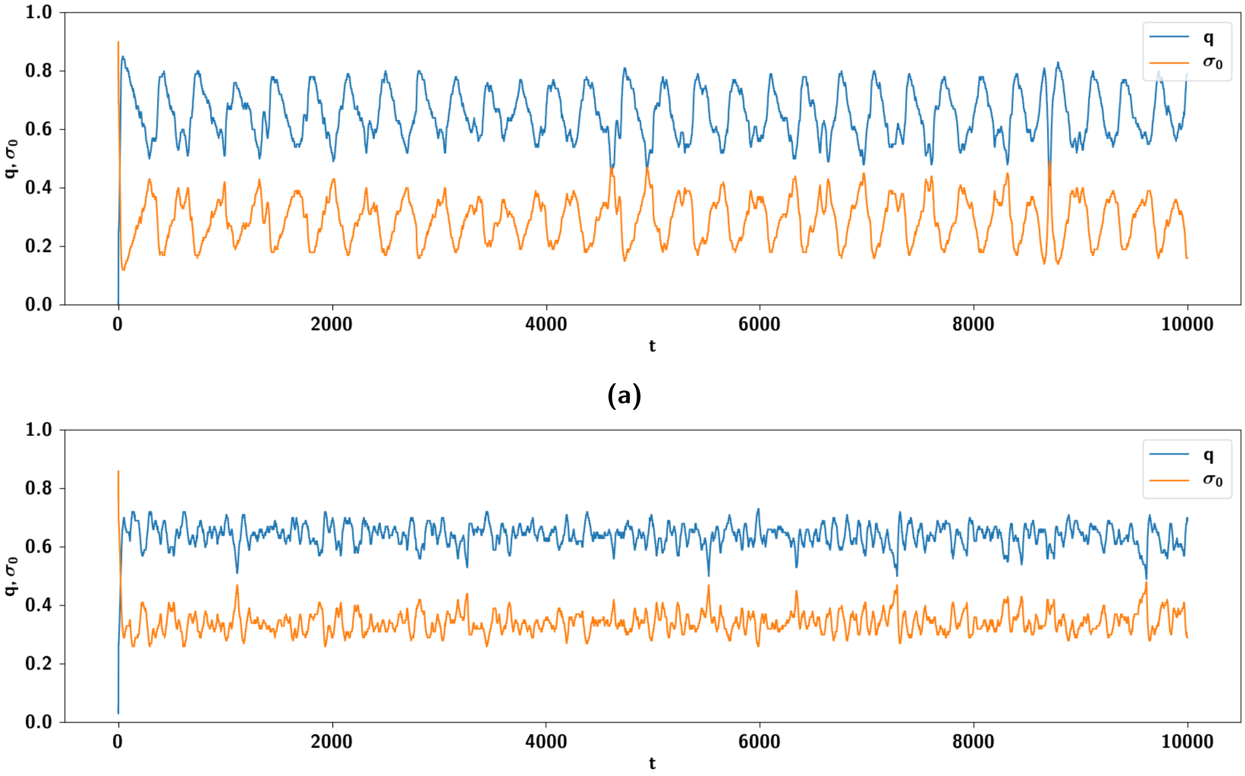


**(a)** $\zeta = 0.01$

**(b)** $\zeta = 0.015$

**(c)** $\zeta = 0.02$

**Fig. 6.6: Latching chains. (a)** One example of a latching chain with $\zeta = 0.001$ and $p = 150$. The network retrieves clusters of highly correlated patterns. **(b)** Latching chain with $\zeta = 0.015$: at the cusp of the transition of the residual information (see Fig. 5.6b), the network retrieves clusters, with bouts in which it oscillates around an attractor without being able to completely destabilize it. **(c)** The network is not able to completely destabilize an attractor it visits, leading to oscillations in the overlap. Network parameters are identical to Fig. 6.1, and correlation parameters are $f = 0.03$, $a_p = 0.4$ and $\Pi = 150$.



**(a)**

**(b)**

**Fig. 6.7: Oscillations? (a)** The activity of the network, as measured by $\overline{\sigma}_0$ and $\overline{q}$ (see text) for $f = 0.05$ and $p = 200$. Both quantities oscillate in time. **(b)** The activity of the network, as measured by $\overline{\sigma}_0$ and $\overline{q}$ (see text) for uncorrelated patterns.

# 7

# Entering a new phase with correlated memories

In recent years, the Potts network has been proposed as an effective model of a network organized with distinct local and global properties. Many physical aspects of the model have been studied in detail, under many simplifying assumptions, including that of full or highly diluted connectivity. One of the additional steps explored in this thesis are the properties of differing global connectivity on the storage capacity, namely the RD, SD and SDRD models. In [Kropff and Treves, 2005], the fully connected network (in which the connectivity models become equivalent) and the highly diluted (HD) limits were investigated. We have corroborated the outcome of previous analyses by showing that the intermediate connectivity indeed interpolates the fully connected and highly diluted limits.

The second and more crucial element in the direction of cortical plausibility has been that of *correlations*. Most of the studies of auto-associative networks have featured uncorrelated patterns, or else correlations that are hardly plausible, irrelevant for the study of semantic memory. The various feats of "mind-reading", achieved with fMRI studies (e.g. [Haxby et al., 2001, Norman et al., 2006, Mitchell et al., 2008]), are reminders that correlations between memories are not details with somewhat diminishing analytical returns, but may be reflective of the core capacity of the cortex as a machine for encoding structured information. The study the statistical and dynamical properties of the Potts network encoding correlated patterns then arises naturally. In order to make progress in this direction, however, it becomes crucial to devise a plausible way of generating correlations. The question of plausibility taps directly into the ability of the cortex to learn and perform generalization, and any algorithm aimed at producing memories that are semantically relevant must produce patterns of activity that are representative of such computation.

We have attempted this by designing the multi-parent pattern generation algorithm. In the algorithm, the patterns are generated through a series of factors, that can be considered as semantic category generators, or else features in a somewhat looser sense, that carry information on the statistical co-occurrence of events. Through a competition, those events that co-occur most frequently together (as reflected in the parameter $f$), their absolute (parametrized by $a_p$) and relative strengths (parametrized by $\zeta$) construct the statistical structure of the memories.

We further studied the storage capacity of the network as a function of both network parameters and correlation parameters through extensive simulations and find that with a Hebbian rule for the storage of patterns, we can store and retrieve *fewer* patterns, a result that is well-known in the literature. Other, Hebbian derived prescriptions for learning, enhancing the capacity may be explored and studied, and we leave such studies for future investigations. However, though all of the correlation parameters, at least for reasonably low values of the sparsity, decrease the storage capacity of the network, the effect of $\zeta$ is particularly interesting. $\zeta$ is the parameter that mediates the strength of input of different parents. $\zeta \sim 0$ corresponds to a situation in which all parents are at equal footing, and none of them has a higher relative influence, while the opposite limit corresponds to only a handful out of the total being relevant. For high enough values of $\zeta$, we observe *correlated retrieval*, in that with the decrease of the fraction retrieved, the fraction retrieved of another, correlated pattern, different from the one cued, increases. In terms of the mutual information between the cued pattern and the final configuration of the network, after retrieval dynamics, that is independent of the criterion set for retrieval, we observe that the information does not go to zero, but stabilizes at a constant value, after the capacity limit has been reached. We call this remaining information the *residual information.* The residual information displays a nontrivial dependence on $\zeta$. It is found that for $\zeta < \zeta_c = 0.02$, this residual information is approximately zero. At $\zeta_c$, there is a sharp increase in the residual information that stabilizes at large values of $\zeta$, when only a handful of parents become relevant. This sharp increase is reminiscent of a phase transition, in which the residual information is the order parameter and $\zeta$ is the control parameter. The residual information has an interesting interpretation: it can be thought of as the information pertaining to the *gross*, core semantic component of the memories, after the *fine* details have been compromised.

This result, however, has to be considered also in complement to the potential contribution of other memory structures to the retrieval of memories. In [Lauro-Grotto et al., 1997], for example, it was found that different access modes to information stored in long-term memory lead to different distributions of classification errors of different groups with memory disorders. An information derived measure, the metric content, quantifying the concentration of errors was computed: high levels of metric content are indicative of a strong dependence on perceived relations among the set of stimuli, and therefore of a relatively preferred semantic access mode, while low levels (and similar correct performance), suggest a preferential episodic access mode.

It was found that compared with normal controls, the metric content index was increased in patients with Alzheimer's disease, decreased in patients with herpes encephalitis, and unvaried in patients with damage to the prefrontal cortex. Moreover, a significant correlation between the metric content and measures quantifying episodic and semantic retrieval mode in the remember/know paradigm introduced by Tulving [Tulving, 2002] was found. If we think of the access modes, to a first approximation, as being the result of a differential reliance on a given memory structure, the distribution of errors may then be a window into understanding the relative contributions of each of them. Within this larger picture, the loss of the ability of the Potts network to perform perfect memory recall may be somewhat mitigated by a complementary episodic mode of access, supported by other structures.

In a final short descriptive chapter, we embark on the dynamical properties of the Potts network with correlated patterns, as a last step towards cortical plausibility. We find that the latching chains cannot be fully described in terms of their *length* [Russo and Treves, 2012] or their *quality* [Kang et al., 2017] even though the latter studies have already been quite exhaustive. The qualitative changes in the structure of the patterns with correlations give rise to some interesting behaviors that we have only begun to study. We find that for parameters in which there exist highly correlated patterns, obtained with values of $f$ for which the algorithm corresponds, statistically, to a hierarchical algorithm, the network retrieves clusters, or patterns in parallel. This gives additional motivation to the study of the multi-parent algorithm. Increasing this parameter, latching resembles more the uncorrelated type (sequential retrieval), though the network also half-retrieves other patterns simultaneously. Increasing the correlations even further, latching becomes noisy, but concerted oscillatory behavior of the aggregate uncondensed patterns starts to manifest itself. Such oscillatory behavior is also found by increasing $a_p$; we observe the periodic retrieval of clusters of highly correlated patterns. Perhaps the most interesting behavior is observed at the cusp of the transition of the residual information when increasing $\zeta$. We observe, below the transition, parallel retrieval of patterns; at the cusp of the transition, the network retrieves clusters of patterns, but there are bouts in which it oscillates around a memory pattern and in which the adaptive thresholds are not strong enough to reduce the activity of the network. After the cusp, the network goes into a state of indefinite oscillatory behavior around a cluster of patterns.

In [Russo and Treves, 2012], the possibility of such various types of dynamics, as a consequence from complexity, but not as an added ingredient, was evoked. Here we find some preliminary evidence of such behavior. The emergence of such abundant types of behavior, allowed by rethinking semantic memories, opens new, unexpected perspectives warranting further investigation.

# Appendices

## A. Self consistent signal to noise analysis

In this section we outline the important steps in the calculation of the mean field equations yielding the storage capacity through the method of the self-consistent signal to noise analysis. We start from Eq. (3.33). Since the l.h.s. includes $p-1 \gg 1$ terms, the ansatz is still valid also when singling out one of these many contributions, so that we can equivalently write it as

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

where $\gamma_i^k$ and $\rho_i^n$ are independent of $\mu$. 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 $v_{\xi_i^\mu,k} m_i^\mu$:

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

Reinserting the expansion into the r.h.s of Eq. (3.28) we recognize a relation of the form

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

where

$$K_{ij}^{\mu} \equiv \frac{1}{c_m a(1-\tilde{a})} \sum_{l,n} c_{ij} v_{\xi_j^\mu,l} v_{\xi_j^\mu,n} \frac{\partial G_j^l}{\partial y^n} \ , \tag{A.4}$$

$$L_i^{\mu} \equiv \frac{1}{c_m a(1-\tilde{a})} \sum_j \sum_l c_{ij} v_{\xi_j^\mu,l} G_j^l \ . \tag{A.5}$$

The overlap $m_i^\mu$ can be found by iterating Eq. (A.3),

$$m_i^\mu = L_i^\mu + \sum_{j_1} L_{j_1}^\mu \Big\{ K_{ij_1}^\mu + \sum_{j_2} K_{ij_2}^\mu K_{j_2 j_1}^\mu + \sum_{j_2} \sum_{j_3} K_{ij_2}^\mu K_{j_2 j_3}^\mu K_{j_3 j_1}^\mu + ... \Big\} . \tag{A.6}$$

Therefore, the noise term can be written explicitly as

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

In order to obtain the expression for $\gamma_i^k$, in Eq. (3.33) we consider only the terms with $j = i$ and $l = k$, and take the average over the connectivity and the patterns:

$$\begin{aligned} \gamma_i^k \ &= \ \frac{\alpha}{S} \lambda \Big\langle \frac{1}{S} \frac{1}{N} \sum_{j_1} \sum_{l_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} + ... \Big\rangle \\ &= \ \frac{\alpha}{S} \lambda \Big\{ \Omega/S + (\Omega/S)^2 + ... \Big\} \\ &= \ \frac{\alpha}{S} \lambda \frac{\Omega/S}{1-\Omega/S} \end{aligned} \tag{A.7}$$

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

$$\Omega = \Big\langle \frac{1}{N} \sum_{j_1} \sum_{l_1} \frac{\partial G_{j_1}^{l_1}}{\partial y^{l_1}} \Big\rangle . \tag{A.8}$$

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

$$(\rho_i^n)^2 = \frac{\alpha P_n}{S(1-\tilde{a})} q \Big\{ 1 + 2\lambda\Psi + \lambda\Psi^2 \Big\} , \tag{A.9}$$

where

$$q = \left\langle \frac{1}{Na} \sum_{j,l} (G_j^l)^2 \right\rangle \tag{A.10}$$

and

$$\Psi = \frac{\Omega/S}{1 - \Omega/S} \,. \tag{A.11}$$

The mean field received by a unit is then

$$\mathcal{H}_k^\xi = v_{\xi,k} m + \frac{\alpha}{S} \lambda \Psi (1 - \delta_{k,0}) + \sum_n v_{n,k} z^n \sqrt{\frac{\alpha P_n}{S(1-\tilde{a})} q \Big\{ 1 + 2\lambda\Psi + \lambda\Psi^2 \Big\}} - U(1 - \delta_{k,0}) \,. \tag{A.12}$$

# B. Calculation of the probability distribution of the field $S = 1$

In this section we outline the important steps in deriving Eq. (4.22). Given that the derivation is rather cumbersome and do not add to the comprehensibility of the main ideas presented in Chap. 4, we report them here. The distribution for $h_\mu$ can be computed by making use of the probability generating function

$$G(s) \equiv \mathcal{L}\{P(x = \xi_i^\pi)\} = \int\limits_0^\infty dx \, P(x) \, e^{-sx} = \frac{a_p}{s}(1 - e^{-s}) + (1 - a_p) \,. \tag{B.1}$$

$\xi_i^\pi$ being identically and independently distributed for all $\pi$, we can use the following property

$$P(h_\mu \,|\, n_p) = \mathcal{L}^{-1}\{G(s)^{n_p}\} \,. \tag{B.2}$$

The number of parents as well as the total field received by a pattern is i.i.d, so we drop the index $\mu$ for $h_\mu$.

$$P(h \,|\, n_p) = \lim_{\gamma \to \infty} \frac{1}{2\pi i} \int\limits_{c - i\gamma}^{c + i\gamma} ds \{1 - a_p + \frac{a_p}{s}(1 - e^{-s})\}^{n_p} \, e^{sh} \tag{B.3}$$

Using the binomial theorem

$$\{1 - a_p + \frac{a_p}{s}(1 - e^{-s})\}^{n_p} = \sum_{k=0}^{n_p} \binom{n_p}{k} (1 - a_p)^{n_p - k} \frac{a_p^k}{s^k} \sum_{j=0}^{k} (-1)^j e^{-sj} \,. \tag{B.4}$$

$$P(h\,|\,n_p)(h) = \lim_{\gamma\to\infty}\frac{1}{2\pi i}\int\limits_{c-i\gamma}^{c+i\gamma}\sum_{k=0}^{n_p}\sum_{j=0}^{k}(-1)^j\binom{n_p}{k}\binom{k}{j}(1-a_p)^{n_p-k}\,\frac{a_p^k}{s^k}\,e^{s(h-j)}ds \tag{B.5}$$

We can carry out the integral to find

$$I(k=0) = \delta(h)\,, \tag{B.6}$$

$$I(k\geq 1) = \frac{(h-j)^{k-1}}{(k-1)!}\,. \tag{B.7}$$

The distribution of the field $h$ for a given number of parents $n_p$ is then

$$P(h\,|\,n_p) = (1-a_p)^{n_p}\delta(h)\; + \sum_{k=1}^{n_p}\sum_{j=0}^{k}\frac{(-1)^j\,n_p!\;a_p^k\,(1-a_p)^{n_p-k}}{(n_p-k)!\,(k-j)!\,j!\,(k-1)!}\,(h-j)^{k-1}\,\Theta(h-j)\,. \tag{B.8}$$

The first term in this equation expresses the fact that the only way to get zero field is if all $n_p$ parents contribute zero field and this occurs with probability $(1-a_p)^{n_p}$. For a given pattern $\mu$, with $n_p$ parents, the field of each unit is distributed according to Fig. 4.4.

The cumulative distribution function writes

$$P(h' < h\,|\,n_p) = \int\limits_{-\infty}^{h} dh'\,P(h'\,|\,n_p) \tag{B.9}$$

$$= (1-a_p)^{n_p}\Theta(h) + \sum_{k=1}^{n_p}\sum_{j=0}^{k}\frac{(-1)^j\,n_p!\;a_p^k\,(1-a_p)^{n_p-k}}{(n_p-k)!\,(k-j)!\,j!\,k!}\,(h-j)^k\,\Theta(h-j)\,. \tag{B.10}$$

$h_m$ is the value for which the cumulative probability is equal to $1-a$.

$$P(h' < h_m\,|\,n_p) = 1-a\,. \tag{B.11}$$

# C. Calculation of the probability distribution of the field $S = 2$

In this section we outline the important steps in deriving Eq. (4.26). We start with the joint distribution of number of parents by state

$$P(\hat{n}^1 = n^1, ..., \hat{n}^S = n^S) = \frac{n_p!}{S^{n_p} \prod\limits_{k=1}^{S} n^k!} \,. \tag{C.1}$$

Note that we define the field to be identically distributed across states. The probability that the fields of all states are below that of the first is given by

$$P(h = h^1) = \int\limits_0^h P(h^1, ..., h^S) \prod_{k=2}^{S} dh^k \,. \tag{C.2}$$

The probability distribution of the maximal field is given by $S$ times the one above

$$P(h_{max}) = S \int\limits_0^{h^1} P(h^1, ..., h^S) \prod_{k=2}^{S} dh^k \,. \tag{C.3}$$

The joint distribution of the fields across states writes

$$P(h^1, ..., h^S | n_p) = \frac{n_p!}{S^{n_p}} \prod_{k=1}^{S} \sum_{n^k=1}^{n_p} \frac{P(h^k | n^k)}{n^k!} \, \delta_{n_p, \sum\limits_{k=0}^{S} n^k} \,, \tag{C.4}$$

where the constraint $n_p = \sum\limits_{k=0}^{S} n^k$ has been included in the last line. $P(h^k|n^k)$ is given by Eq. (B.8), replacing $n_p$ with $n^k$.

$$P(h^1, ..., h^S | n_p) = \frac{n_p!}{S^{n_p}} \prod_{k=1}^{S} \sum_{n^k=1}^{n_p} \left\{ \frac{(1-a_p)^{n^k}}{n^k!} \, \delta(h^k) + \right. \tag{C.5}$$

$$\left. + \sum_{i=1}^{n^k} \sum_{j=0}^{i} (-1)^j \frac{(a_p)^i \, (1-a_p)^{n^k - i}}{(n^k - i)! \, (i-j)! \, j!} \, \frac{(h^k - j)^{i-1}}{(i-1)!} \Theta(h^k - j) \right\} \delta_{n_p, \sum\limits_{k=0}^{S} n^k} \tag{C.6}$$

For $S = 1$ all contributions go to a single state, so we automatically have $n^1 = n_p$, then the first sum disappears and we fall back onto Eq. (B.8). For $S = 2$ we have, denoting the state receiving the maximal field by $H$.

$$
\begin{aligned}
P(H \,|\, n_p) = \frac{n_p!}{2^{n_p-1}} \sum_{n^1=1}^{n_p} & \left\{ \frac{(1-a_p)^{n^1}}{n^1!} \delta(H) + \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1-i}}{(n^1-i)! \, (i-j)! \, j! \, (i-1)!} \, (H-j)^{i-1} \, \Theta(H-j) \right\} \\
& \left\{ \frac{(1-a_p)^{n_p-n^1}}{(n_p-n^1)!} \Theta(H) + \sum_{i'=1}^{n_p-n^1} \sum_{j'=0}^{i'} \frac{(-1)^{j'} (a_p)^{i'} \, (1-a_p)^{n_p-n^1-i'}}{(n_p-n^1-i')! \, (i'-j')! \, j'! \, i'!} \, (H-j')^{i'} \, \Theta(H-j') \right\}
\end{aligned}
\tag{C.7}
$$

where we drop the indices denoting the units (they are drawn from the same distribution). Note that the state does not appear in this expression because it is the distribution of the state receiving maximal input, regardless of which one it is. The $\mu$ dependence is through $n_p = n_p(\mu)$.

$$
P(H' < H_m \,|\, n_p) = \int_{-\infty}^{H_m} P(H' \,|\, n_p) \, dH' = 1 - a \tag{C.8}
$$

Finally

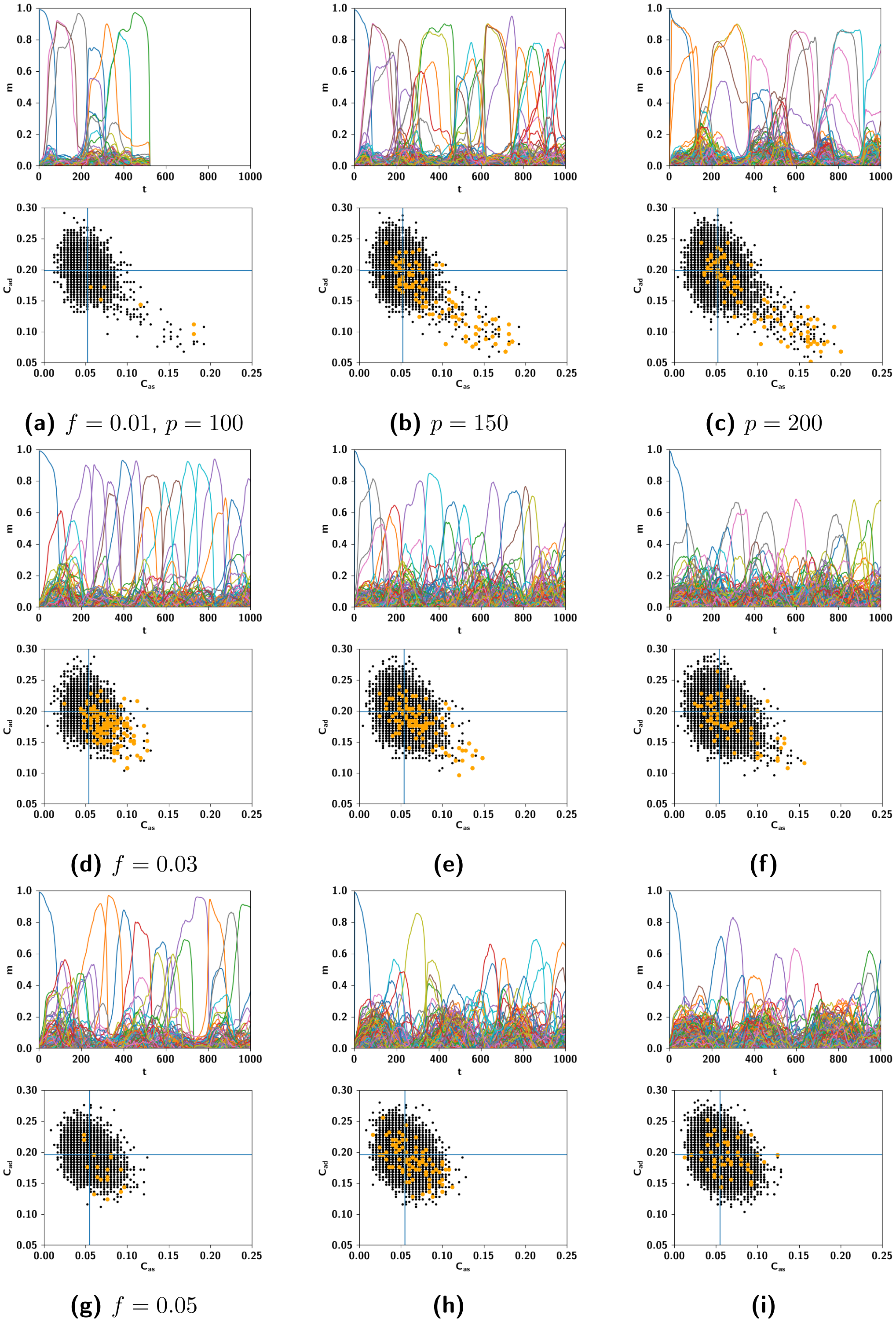
$$
\begin{aligned}
P(H' < H \,|\, n_p) = \; & \frac{n_p!}{2^{n_p-1}} \sum_{n^1=1}^{n_p} \left\{ \frac{(1-a_p)^{n_p}}{n^1! \, (n_p-n^1)! \, 2} \right. \\
& + \frac{(1-a_p)^{n_p-n^1}}{(n_p-n^1)!} \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1-i}}{(n^1-i)! \, (i-j)! \, j! \, i!} (H-j)^i \, \Theta(H-j) \\
& \left. + \sum_{i'=1}^{n_p-n^1} \sum_{j'=0}^{i'} \frac{(-1)^{j'} (a_p)^{i'} \, (1-a_p)^{n_p-n^1-i'}}{(n_p-n^1-i')! \, (i'-j')! \, j'! \, i'!} \sum_{i=1}^{n^1} \sum_{j=0}^{i} \frac{(-1)^j (a_p)^i \, (1-a_p)^{n^1-i}}{(n^1-i)! \, (i-j)! \, j! \, (i-1)!} \, I(H,i,i',j,j') \right\} .
\end{aligned}
\tag{C.9}
$$

where $max\{j, j'\} = j^*$

$$
\begin{aligned}
& I(H,i,i',j=j') = \frac{(H-j)^{i+i'}}{i+i'} \Theta(H-j) \\
& I(H,i,i',j \neq j') = \\
& = \begin{cases}
i'! \, (i-1)! \left[ \sum_{q=0}^{i'} (-1)^q \frac{(H-j)^{i+q} (H-j')^{i'-q}}{(i+q)! \, (i'-q)!} \, \Theta(H-j^*\}) + (-1)^{i'+1} \frac{(j^*-j)^{i+i'}}{(i+i')!} \right] & i-1 \geq i' \\
i'! \, (i-1)! \left[ \sum_{q=0}^{i-1} (-1)^q \frac{(H-j)^{i-q-1} (H-j')^{i'+q+1}}{(i-q-1)! \, (i'+q+1)!} \, \Theta(H-j^*) + (-1)^{i} \frac{(j^*-j)^{i+i'}}{(i+i')!} \right] & i-1 < i'
\end{cases}
\end{aligned}
\tag{C.10}
$$

## D. Latching dynamics with correlated patterns

In this section, we report several figures of latching chains with correlated patterns, accompanied by scatterplots in which black points denote correlation between all possible pairs of patterns learned by the network and yellow dots represent only those pairs between which a latching transition has occurred.



**(a)** $f = 0.01$, $p = 100$

**(b)** $p = 150$

**(c)** $p = 200$

**(d)** $f = 0.03$

**(e)**

**(f)**

**(g)** $f = 0.05$

**(h)**

**(i)**

# Bibliography

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

[Amit et al., 1985] Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review A*, 32(2):1007.

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

[Anderson, 1997] Anderson, O. R. (1997). A neurocognitive perspective on current learning theory and science instructional strategies. *Science Education*, 81(1):67–89.

[Bartol Jr et al., 2015] Bartol Jr, T. M., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., and Sejnowski, T. J. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife*, 4:e10778.

[Bollé et al., 1992] Bollé, D., Dupont, P., and Huyghebaert, J. (1992). Thermodynamic properties of the q-state Potts-glass neural network. *Physical Review A*, 45(6):4194.

[Bollé et al., 1991] Bollé, D., Dupont, P., and van Mourik, J. (1991). Stability properties of Potts neural networks with biased patterns and low loading. *Journal of Physics A: Mathematical and General*, 24(5):1065.

[Bollé et al., 1993] Bollé, D., Vinck, B., and Zagrebnov, V. (1993). On the parallel dynamics of the q-state Potts and q-Ising neural networks. *Journal of statistical physics*, 70(5):1099–1119.

[Borgo and Shallice, 2003] Borgo, F. and Shallice, T. (2003). Category specificity and feature knowledge: Evidence from new sensory-quality categories. *Cognitive Neuropsychology*, 20(3-6):327–353.

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

[Braitenberg, 1978b] Braitenberg, V. (1978b). Cortical architectonics: general and areal. brazier and h. petsche (eds).

[Braitenberg and Schüz, 1991] Braitenberg, V. and Schüz, A. (1991). *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media.

[Buhmann et al., 1989] Buhmann, J., Divko, R., and Schulten, K. (1989). Associative memory with high information content. *Physical Review A*, 39(5):2689.

[Capitani et al., 2003] Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). What are the facts of semantic category-specific deficits? a critical review of the clinical evidence. *Cognitive Neuropsychology*, 20(3-6):213–261.

[Caramazza et al., 1990] Caramazza, A., Hillis, A. E., Rapp, B. C., and Romani, C. (1990). The multiple semantics hypothesis: multiple confusions? *Cognitive neuropsychology*, 7(3):161–189.

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

[Cree and McRae, 2003] Cree, G. S. and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163.

[Crick, 1989] Crick, F. (1989). The recent excitement about neural networks. *Nature*, 337(6203):129–132.

[Derrida et al., 1987] Derrida, B., Gardner, E., and Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4(2):167.

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

[EilingYee and Thompson-Schill, 2013] EilingYee, E. G. C. and Thompson-Schill, S. L. (2013). Semantic memory. *The Oxford Handbook of Cognitive Neuroscience, Volume 1: Core Topics*, 1:353.

[Elston, 2000] Elston, G. N. (2000). Pyramidal cells of the frontal lobe: all the more spinous to think with. *Journal of Neuroscience*, 20:1–4.

[Engel, 1990] Engel, A. (1990). Storage capacity for hierarchically correlated patterns. *Journal of Physics A: Mathematical and General*, 23(6):L285.

[Engel et al., 2004] Engel, A., Monasson, R., and Hartmann, A. K. (2004). On large deviation properties of Erdös-Rényi random graphs. *Journal of Statistical Physics*, 117(3-4):387–426.

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

[Farah and McClelland, 1991] Farah, M. J. and McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of experimental psychology: General*, 120(4):339.

[Frolov et al., 1997] Frolov, A. A., Husek, D., and Muraviev, I. P. (1997). Informational capacity and recall quality in sparsely encoded hopfield-like neural network: Analytical approaches and computer simulation. *Neural Networks*, 10(5):845–855.

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

[Haxby et al., 2001] Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

[Hebb, 2005] Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory.* Psychology Press.

[Hellwig, 2000] Hellwig, B. (2000). 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.

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

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

[Horn and Usher, 1989] Horn, D. and Usher, M. (1989). Neural networks with dynamical thresholds. *Physical Review A*, 40(2):1036.

[Humphreys and Forde, 2001] Humphreys, G. W. and Forde, E. M. (2001). Hierarchies, similarity, and interactivity in object recognition: "category-specific" neuropsychological deficits. *Behavioral and Brain Sciences*, 24(3):453–476.

[Kang et al., 2017] Kang, C. J., Naim, M., Boboeva, V., and Treves, A. (2017). Life on the edge: Latching dynamics in a Potts neural network. *Entropy*, 19(9):468.

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

[Kropff, 2009] Kropff, E. (2009). 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.

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

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

[Lambon Ralph et al., 2007] Lambon Ralph, M. A., Lowe, C., and Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, 130(4):1127–1137.

[Lauro-Grotto et al., 1997] Lauro-Grotto, R., Piccini, C., Borgo, F., and Treves, A. (1997). What remains of memories lost in alzheimer and herpetic encephalitis. In *Society for Neuroscience Abstracts*, volume 23, page 1889.

[Leutgeb et al., 2007] Leutgeb, J. K., Leutgeb, S., Moser, M.-B., and Moser, E. I. (2007). Pattern separation in the dentate gyrus and ca3 of the hippocampus. *science*, 315(5814):961–966.

[Leutgeb et al., 2004] Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B., and Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas ca3 and ca1. *Science*, 305(5688):1295–1298.

[Löwe, 1998] Löwe, M. (1998). On the storage capacity of Hopfield models with correlated patterns. *The Annals of Applied Probability*, 8(4):1216–1250.

[Mari and Treves, 1998] Mari, C. F. and Treves, A. (1998). Modeling neocortical areas with a modular neural network. *Biosystems*, 48(1):47–55.

[Markram et al., 2012] Markram, H., Gerstner, W., and Sjöström, P. J. (2012). Spike-timing-dependent plasticity: a comprehensive overview. *Frontiers in synaptic neuroscience*, 4.

[Marr et al., 1991] Marr, D., Willshaw, D., and McNaughton, B. (1991). Simple memory: a theory for archicortex. In *From the Retina to the Neocortex*, pages 59–128. Springer.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

[McNaughton and Morris, 1987] McNaughton, B. L. and Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in neurosciences*, 10(10):408–415.

[McRae et al., 2005] McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

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

[Mézard et al., 1987] Mézard, M., Parisi, G., and Virasoro, M. (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Co Inc.

[Mitchell et al., 2008] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

[Naim et al., 2017] Naim, M., Boboeva, V., Kang, C. J., and Treves, A. (2017). Reducing a cortical network to a Potts model yields storage capacity estimates. *arXiv preprint arXiv:1710.04897.*

[Norman et al., 2006] Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430.

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

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

[O'Reilly and Rudy, 2001] O'Reilly, R. C. and Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review*, 108(2):311.

[Osgood, 1964] Osgood, C. E. (1964). Semantic differential technique in the comparative study of cultures. *American Anthropologist*, 66(3):171–200.

[Parga and Virasoro, 1987] Parga, N. and Virasoro, M. (1987). The ultrametric organization of memories in a neural network. In *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, pages 436–443. World Scientific.

[Patterson and Plaut, 2009] Patterson, K. and Plaut, D. C. (2009). “shallow draughts intoxicate the brain”: Lessons from cognitive science for cognitive neuropsychology. *Topics in Cognitive Science*, 1(1):39–58.

[Plaut, 1995] Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th annual conference of the cognitive science society*, volume 17, pages 37–42.

[Preston and Eichenbaum, 2013] Preston, A. R. and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773.

[Pucak et al., 1996] Pucak, M. L., Levitt, J. B., Lund, J. S., and Lewis, D. A. (1996). Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *Journal of Comparative Neurology*, 376(4):614–630.

[Rogers et al., 2004] Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, 111(1):205.

[Rolls, 1989] Rolls, E. (1989). The representation and storage of information in neural networks in the primate cerebral cortex and hippocampus. In *The computing neuron*, pages 125–159. Addison-Wesley Longman Publishing Co., Inc.

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

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1:45–76.

[Russo et al., 2008] Russo, E., Namboodiri, V. M., Treves, A., and Kropff, E. (2008). Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, 10(1):015008.

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

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

[Shallice and Cooper, 2011] Shallice, T. and Cooper, R. (2011). *The organisation of mind.* Oxford University Press.

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

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

[Silveri and Gainotti, 1988] Silveri, M. C. and Gainotti, G. (1988). Interaction between vision and language in category-specific semantic impairment. *Cognitive Neuropsychology*, 5(6):677–709.

[Sompolinsky, 1986] Sompolinsky, H. (1986). Neural networks with nonlinear synapses and a static noise. *Physical Review A*, 34(3):2571.

[Tamarit and Curado, 1991] Tamarit, F. A. and Curado, E. M. (1991). Pair-correlated patterns in hopfield model of neural networks. *Journal of statistical physics*, 62(1):473–480.

[Toulouse et al., 1987] Toulouse, G., Dehaene, S., and Changeux, J.-P. (1987). Spin glass model of learning by selection. In *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, pages 432–435. World Scientific.

[Treves, 1997] Treves, A. (1997). On the perceptual structure of face space. *BioSystems*, 40(1-2):189–196.

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

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

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

[Tsuda, 2001] Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences*, 24(5):793–810.

[Tulving, 2002] Tulving, E. (2002). Episodic memory: from mind to brain. *Annual review of psychology*, 53(1):1–25.

[Tyler and Moss, 2001] Tyler, L. K. and Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in cognitive sciences*, 5(6):244–252.

[Vinson and Vigliocco, 2008] Vinson, D. P. and Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

[Warrington and Shallice, 1984] Warrington, E. K. and Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107(3):829–853.

[Zipser, 1988] Zipser, D. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:25.