"Sapienza" Università di Roma

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Corso di Laurea Specialistica in Fisica

APPRENDIMENTO DI MEMORIE CORRELATE CON SINAPSI BINARIE

Tesi di Laurea Specialistica

Relatore:
Prof. Enzo Marinari

Candidato:
Lorenzo Fontolan

Relatore esterno:
Prof. Stefano Fusi

Anno Accademico 2010/2011

In all truth I tell you,
unless a wheat grain falls into the earth and dies,
it remains only a single grain;
but if it dies it yields a rich harvest.

John, 12:24

Contents

Ringraziamenti

Questo lavoro di tesi giunge alla fine di un lungo, intenso, a volte entusiasmante e altre volte doloroso percorso universitario. Del resto non esiste metafora più evocativa della vita che non la vita stessa.

Non posso che cominciare i ringraziamenti dalla mia famiglia, che mi ha sempre accompagnato sopportando le mie stravaganze e supplendo amorevolmente alla mia proverbiale distrazione.

Ringrazio Alessandra che, pur lontana migliaia di chilometri, mi è stata vicina quotidianamente e che, suo malgrado, è stata costretta ad ascoltare innumerevoli sproloqui sugli argomenti di questa tesi, che ormai conosce meglio di chiunque altro.

La seconda persona che conosce questo lavoro in profondità è naturalmente il mio relatore esterno Stefano Fusi, che ringrazio per avermi ospitato per un anno intero nel suo gruppo di ricerca alla Columbia University di New York, e per avermi sempre guardato come un compagno di avventura e mai come un subalterno. Allo stesso modo gli altri membri del Center for Theoretical Neuroscience hanno davvero illuminato le mie giornate newyorkesi, a partire da Mattia e Kiyo che mi hanno aiutato con pazienza ammirevole e con i quali ho riso più di quanto abbia mai fatto negli ultimi dieci anni, Daniel che è stato un impareggiabile guida nelle neuroscience teoriche e

nel rock progessive, Anthony che è stato come un fratello, e poi Srdjan, Omri, Xaq, Evan e Dan.

Devo invece il fatto di non aver perso l'appetito, grazie alla cucina italiana di altissimo livello, ai miei compagni di appartamento Alberto e Valerio, con cui ho vissuto una vera e bella esperienza di comunione. Federica, Camilla e le altre riminesi, alternativamente a Charlotte e Claire, hanno invece movimentato (con risultati discutibili) le mie serate e sopportato i miei musi lunghi quando i risultati tardavano a venire.

Chi mi ha consentito di intraprendere questa avventura americana è stato il Prof. Enzo Marinari, la cui pazienza e fiducia verso di me sono state fondamentali per portare a termine questo (piccolo) lavoro di ricerca.

Ringrazio poi gli amici più cari di questi anni: Andrea, Giuliano, Giulia, Paola, Mariagiulia, Marta, Lorenzo, Luigi, Martino, Francesco, Sergio, Jonah, Aldo, gli altri membri del gruppo dei Tre Moschettieri, Marco, Alberto e Carlo, e tutti i fisici (in particolare Stanislao, Giulio, Arianna, Luciano, Carla, Sara, Michele e Francesca).

In ultimo, un pensiero a chi mi ha indirizzato verso le neuroscienze. Il responsabile maggiore del mio innamoramento per le neuroscienze teoriche è stato il Prof. Daniel Amit. Oltre al suo modo unico di guardare alla fisica, sono stati i suoi occhi pieni di curiosità e nostalgia che mi hanno spalancato lo sguardo. C'è un verso di T.S. Eliot che descrive quello che mi è accaduto seguendo il suo corso di Reti Neurali:

> "We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time."

Il dolore per la sua morte insieme alla gratitudine per averlo incontrato mi accompagneranno per tutta la vita.

Chapter 1

Introduction

1.1 Motivation of studies

In recent years, neuroscience has drawn the attention of scientists coming from many different backgrounds. Disciplines like medicine, psychology, biology, mathematics, philosophy, chemistry, bioengineering, cybernetics, and also physics, are now actively involved in what I personally perceive as the most interesting among the mysteries of human nature: the exploration of the human brain. Our perception of the world is mediated through the brain, at all levels: from tickling feet to the understanding of quantum electrodynamics, from the perception of beauty to the fits of anger, every external input must be processed by our brain. Since the age of the first scientific approaches, the number of questions and the complexity of the answers have been rapidly increasing, leading to an enormous corpus of data and evidences which, still, mostly lack an even basic understanding. This is, of course, due to several factors: first, the enormous, astonishing sophistication of the ner-

vous system; second, the fact that unharming experimental techniques for studying the brain have been developed only during the last seventy years; third, the insufficient use of mathematical and physical methods in favour of more psychological and heuristic attitudes. In recent years, however, amazing advancements have been made, boosted by the development of instruments capable of measuring the activity of single neurons, the "invasion" of a rising number of mathematicians, statistician and phycists into biological fields, and the availability of fast and powerful computers to test and simulate brain data.

With this dissertation we aim to contribute to the current theoretical effort in understanding the principles of learning and memory. In fact, neither the biological underpinnings, nor the physical dynamics of these processes have been fully understood yet. Theoretical models represent both a powerful source of ideas and a reliable ground for quantitative and predictive analysis (Abbott, 2008).

Our approach relies essentially on concepts and techniques borrowed from statistical physics and dynamical systems theory, along the research lines inaugurated by the first pioneers who brought physics into neuroscience. Since we are interested in the basic principles rather than subtle biological details, we will use suitably simplified models. However we will not make use of assumptions that are absurd from a biological point of view. In this way we hope to reveal the essential properties of the system, without blurring the analysis with irrelevant (to our level of description) biological details.

The fundamental motive of this thesis can be sketched as follows: building a plausible and robust model for the learning and storage of memories in a neural network with the aid of mathematical and physical tools, so that, ultimately (and not yet in this thesis), one could foresee the consequences of these formulations, and possibly formulate predictions which could be later verified or rejected by experiments.

1.2 Thesis overview

In this section we outline the main concepts presented in each chapter, hoping to offer the reader a concise overview of this dissertation.

Anatomical and biological structure of the brain

In Chapter 2 we will go over a wide-ranging but essential survey of the main biological features of the primate brain. At the macroscopic level, we can, at least, sketch the role of almost every single structure, but we are far from a deep understanding of specific functions and the way they are accomplished.
On the other side, at the microscopic scale ($\sim \mu m$), the brain is made up of a massive number of cells, the neurons, and an even larger number of connections, the synapses. Although there are many kinds of nervous cells, and several different synaptic mechanisms (triggered by disparate biochemical processes), our attention is more focused on the main universal properties of the brain basic components: the generation of an action potential in the neuron's soma, and the plasticity of the synaptic connections. Experimental evidences show that the ability to learn and retain an external stimulus primarily relies on those two properties, as we will point out. This raises the fundamental question of this thesis and, perhaps, of theoretical neuroscience: how does our nervous system manage and modulate the activity of billions of neurons to gather, elaborate and execute the variety of complex behaviours that we see in daily experience?

Neural Networks

Scientists have tried to answer this question for the last fifty years. Starting from very simplified and abstract models of binary computational units, McCulloch and Pitts (1943) have shown that such systems are capable of computing, i.e. perform-

ing any arbitrary sequence of logical operations. Moreover, the extensive studies of Rosenblatt (1962) prove that a simple network of these units (the *perceptron*) can learn, under some assumptions, to classify external inputs into different categories by dynamically adjusting the strength of the connections.

The studies on the perceptron laid the groundwork for the work of Little (1974), Amari (1977), Hopfield (1982), and Amit (1989), who developed the modern concept of *Attractor Neural Networks* (ANN), which, as we shall see in Chapter 3, is particularly important for memory and learning. An ANN is a recurrent network of interacting elementary units, whose time dynamics converge to a stable pattern of activity. The fixed points of the dynamics (the attractors) are representations of the memorized patterns: whenever we feed the network with a stimulus (that can be seen as the initial conditions of the dynamics), it will soon settle into the closest attractor and remain there until a new stimulus is presented. It is straightforward to notice how promising these network are for modelling associative memory, i.e. a memory that can be retrieved just by specifying a piece of its content[1].

Once established the adequate mathematical formalism, researchers embarked in a tour de force to determine the scaling properties of associative ANN with the tools of statistical mechanics. Amit et al. (1985) found that the maximum number of uncorrelated memories that can be stored in a ANN grows linearly with the size of the network. When tested under more realistic biological constraints, however, this fairly general result proved to be a very optimistic estimate of the real capacity. Only successively, thanks to the work of Amit and Fusi (1994), the introductions of a stochastic component in the learning process allowed to restore the previous estimation. In addition, it has been noticed by Tsodyks and Feigelman (1988) that the

[1] An example of associative memory is a system that would recall the word "shakespeare" when feeded with the phrase "to be or not to be, that is the question". The system should be robust to small errors: for example, we would like the network to retrieve the word "shakespeare" even when presented with the corrupted stimulus "to beat or not to beat, that is the question".

capacity might be enhanced if memories are sparsely coded, which means activating only a small number of neurons per pattern.

In the last section of Chapter 3 we illustrate a possible framework for generating a hierarchy of correlated patterns: ultrametric trees. Correlated stimuli are, in fact, much more interesting than uncorrelated ones, since they constitute a richer environment more similar to what we do actually perceive from the world around us.

Learning of correlated memories

The question of how to build a network capable of storing an ultrametric hierarchy of memories is the main topic of Chapter 4. In the first part of the chapter, in order to get acquainted with the model, we present an extended discussion of the case of uncorrelated sparse memories.

We consider a recurrent network of binary neurons whose synapses have only two stable states, whose input is a sequence of memories presented one at a time. Each memory elicits a peculiar pattern of neural activity, which in turn modifies a small group of synapses in a stochastic fashion. The network stores a certain amount of information from the currently presented stimulus by dynamically updating a certain number of synaptic strength through a stochastic Hebbian rule. Hence, the learning process can be visualized as a random walk between the stable synaptic states. Changes induced in the synaptic matrix represent the memory trace of the stimulus: as long as some of the synapses preserve their value, the original pattern of activity produced by the stimulus may be recalled.

In this way, older patterns are progressively erased from the network, since new patterns cause further synaptic modifications that may overwrite previous changes. In the limit of *slow learning*, in which the probability that a pattern leaves a trace

in the synaptic matrix is kept low, memories stay in the network for a long time without being lost. We provide an analytic result for the memory lifetime (which is perfectly equivalent to calculating the maximum storing capacity of the network) by performing a signal to noise analysis. In fact, we define the memory lifetime as the time after which the original trace left by the tracked memory becomes indistinguishable from the random fluctuations of the system. The maximum lifetime is seen to be inversely proportional to the sparseness of patterns.

In the second part of Chapter 4 we propose a model of a network whose input is a set of memories organized in classes. Each class can be seen as an ultrametric tree whose leaves represent a group of correlated stimuli, while at the branching node lies a prototype pattern that collects the average features of all descending leaves. The stimuli are random selected among the members of each class, and never among the prototypes. If the learning process is slow and the intra-class correlation is high, the network learns the prototype pattern more effectively than any of the original stimuli.

We exploit this feature by constructing a further network capable of storing only the differences between the prototype and the stimulus representations. The higher is intra-class correlation, the sparser is the representaton of the difference. The number of patterns that can be learned sensibly increases by a factor proportional to correlation.

We first studied an ultrametric hierarchy with two levels, and then with three levels. In the latter case we show that the network may also learn any of the subclasses parents from the intermediate level if the network parameters are appropriately tuned. The analysis has been carried out using the mean field approximation, and the results have been afterwards tested with extensive computer simulations.

Conclusions and future directions

In the last chapter we briefly summarize and discuss the work done in this dissertation and its possible consequences, suggesting some interesting advancements that could follow the present work in the near future.

Chapter 2

Biology of the Nervous System

The first scientific enquiries into the brain function trace back to the Neolithic period, attested by remains of tools used during brain operations. By the 3rd millennium B.C., the Egyptians were even able to reach a remarkable rate of success in brain surgery, as it is reported, for example, in the Edwin Smith Papyrus[1]. In this manuscript of inestimable value, the author recounts several cases of patients suffering from head wounds, who underwent neurosurgical treatments and could sometimes improve their previous condition. Every age had its surgeons: Celsus in ancient Rome, Galenus in Greece, many clerics and churchmen (even the Pope's personal confessors) in the Christian Middle Ages. Several philosophers from the past, from Aristotle to Cartesius, attempted, as well, to define the role of brain, in the effort to trace what makes Man exceptional in the realm of nature. Despite a constant curiosity for its mysteries, the history of scientific investigation of the human mind has been slow and complicated. Only during the second half of the

[1] http://www.touregypt.net/edwinsmithsurgical.htm

19th century, the work of Camillo Golgi and Santiago Ramon y Cajal inaugurated a new scientific approach to the study of the brain, granting them the Nobel prize and giving birth to modern neuroscience.

2.1 Basic functional anatomy

In this section we will shortly sketch the anatomical structure of the human brain (see Fig. 2.1). The bigger part of the brain, located immediatly below the skull, is represented by the two (left and right) cerebral hemispheres, consisting in a four millimeters thick, highly convoluted stratum of gray matter[2] called *cortex*, and an underlying core of white matter linking the cortex to the spinal cord. The cortex

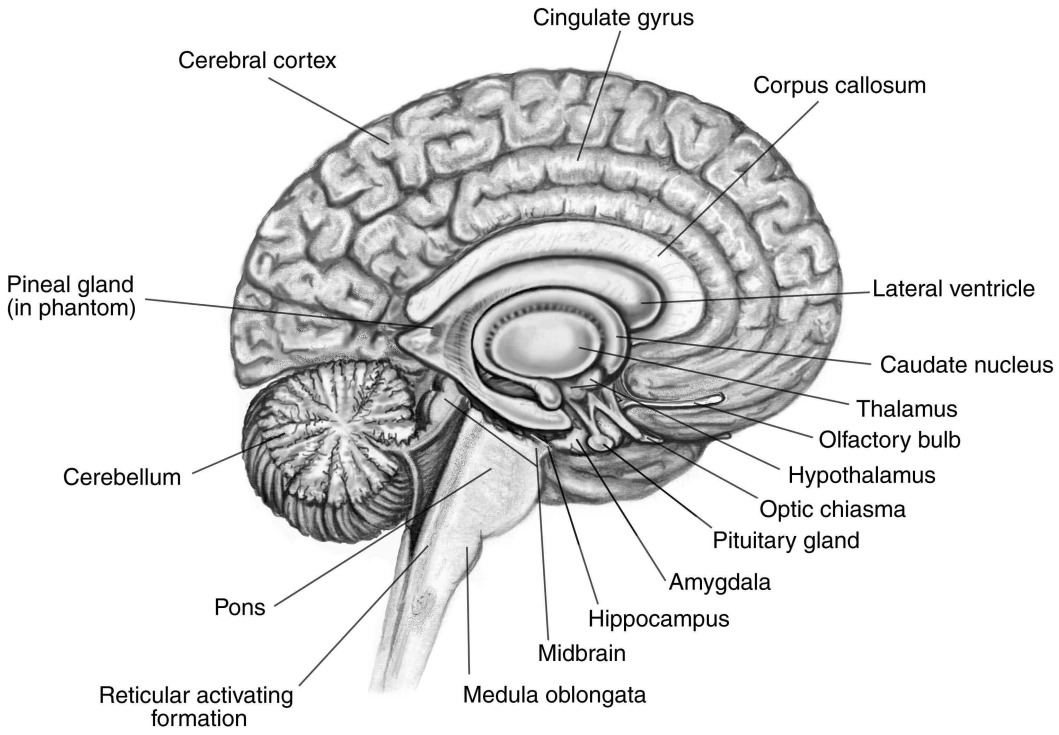


Figure 2.1: Cross section of the human brain. Regions of major interest are shown.

[2]Gray matter comprehends nerve cell bodies, dendrites, myelinated and unmyelinated axons and blood capillaries. White matter, on the contrary, mainly constitutes of myelinated axons, determining the white color.

directs the brain's higher cognitive and emotional functions. It takes part in a myriad of processes: primary sensory areas elaborate sensory information, motor areas select volontary movements of the body and produce the signal in order to execute them, associative areas are involved in building the perceptual representations of reality, such as three-dimensional space, but also in abstract thinking and in conscious experience. A hard packet of nerve fibers connects the two hemispheres: the *corpus callosum*.

Beneath the cortex resides the limbic system, consisting of the amigdala, which plays a role in motivated behaviours (like sexual desires) and some emotional states (fear, aggressiveness), the hippocampus, that mediates learning and memory processes, and the hypothalamus, an ensemble of nuclei devoted to the monitoring of body temperature, blood pressure, reproductive behaviour, hunger and thirst.

Basic life processes, including breathing, heart pulse, arousal, balance and sleep, are controlled by a group of structures placed below the lymbic system. This core compound is evolutionary old, since it can be found in all vertebrates, and includes some important brain architectures such as the thalamus and the cerebellum, that are, respectively, the first filter of sensory information and the coordinator of body movements, posture and equilibrium. Going further down one finds the medulla, a center for many autonomic functions but also the relay of nerve signals between the brain and spinal cord. Traveling inside a tubular bundle of nerve channels, neural signals are eventually transmitted to muscles and organs all over the body.

2.2 Neurons and synapses

To develop its computational paradigm, that we are still far from understanding, the brain makes use of a huge number of slow, unreliable[3] and densely interconnected

[3]Neuron's behaviour is nonlinear, showing significant probability of error per unit.

units: the *neurons*. The human brain contains between 10^{10} and 10^{11} nerve cells, which are mutually linked through chemical junctions called *synapses*. In the cortex, each neuron receives around 10000 synaptic inputs, for a total of $\sim 10^{13}$ synapses and a density of 10^8 synapses per cubic millimeters. Since they are the building blocks of the brain, we are about to give a brief account of the characteristics of neurons and synapses.

2.2.1 The neuron

The basic computing unit in the brain is the single nerve cell, the *neuron*. Neurons communicate through electrical impulses produced inside the cell body and then transmitted to other structures, such as muscle fibers or, in the majority of cases, other neurons. The neuron has a complex structure, which is illustrated in Fig. 2.2. The *soma*, which embodies the nucleus, is the place where the most important biochemical reactions take place and, therefore, where the communicating signal of the neuron, a fast and strong electrical impulse (the *action potential*), is first produced. The action potential is generated in correspondence with a shift in the electrochemical potential of the cell, caused by a depolarization (that would provoke an excitatory current spike) or a hyperpolarization (which induces an inhibitory effect: the cell is less likely to produce an action potential) of the cell membrane, the protective layer coating the cell. For reasons that we will explain later, the impulse is largely independent of the size and shape of the depolarization. Hence, it travels always with (about) the same amplitude and speed through the main output channel, the *axon*, a long cable surrounded by a myelin sheath to facilitate the propagation of the signal. The axon may split into several branches, each one containing several terminals connected to other neurons. At the end of his travel in the axon, the current spike reaches a pre-synaptic terminal, a small structure con-

taining various chemicals that are released upon the spike arrival, and transmit the impulse to the post-synaptic terminals of target neurons. These chemicals, called *neurotransmitters*, move across a small gap (about $20nm$ wide) that separates the pre- and the post-synaptic terminals, where they are captured by receptor molecules of the target cell, eventually delivering the signal. Post-synaptic terminals are disseminated in the dendrites, which form a tree-like structure, connected to the soma, where the signal ends its travel. The initial fast and all-or-none spike, generated inside the soma, is modified in amplitude, frequency and phase during both synaptic

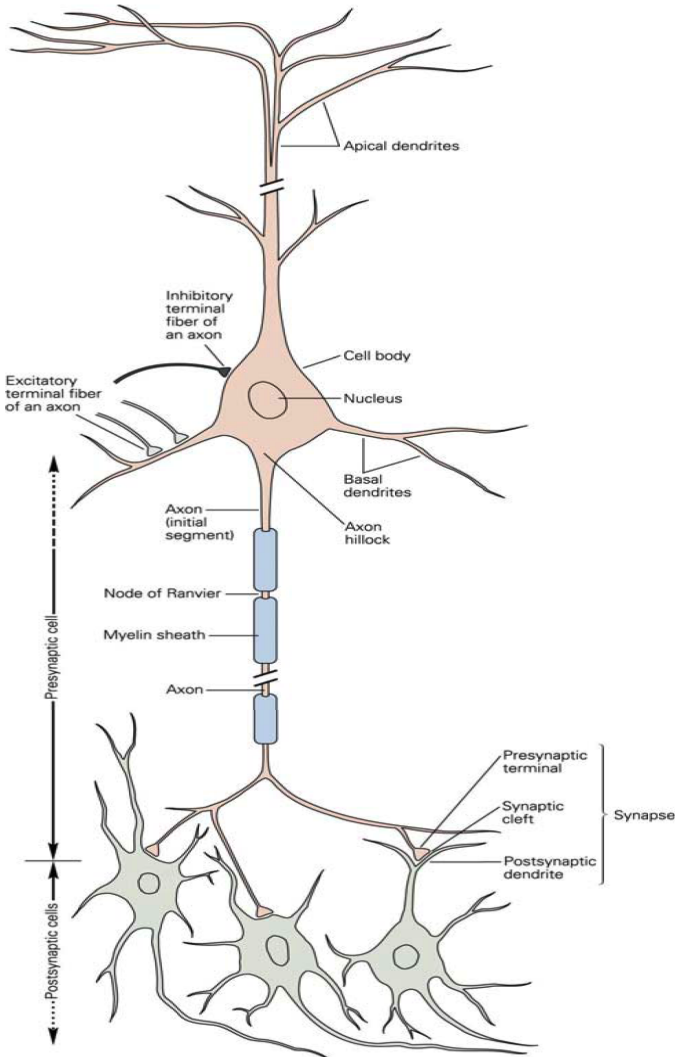


Figure 2.2: Simplified anatomical structure of a single nerve cell.

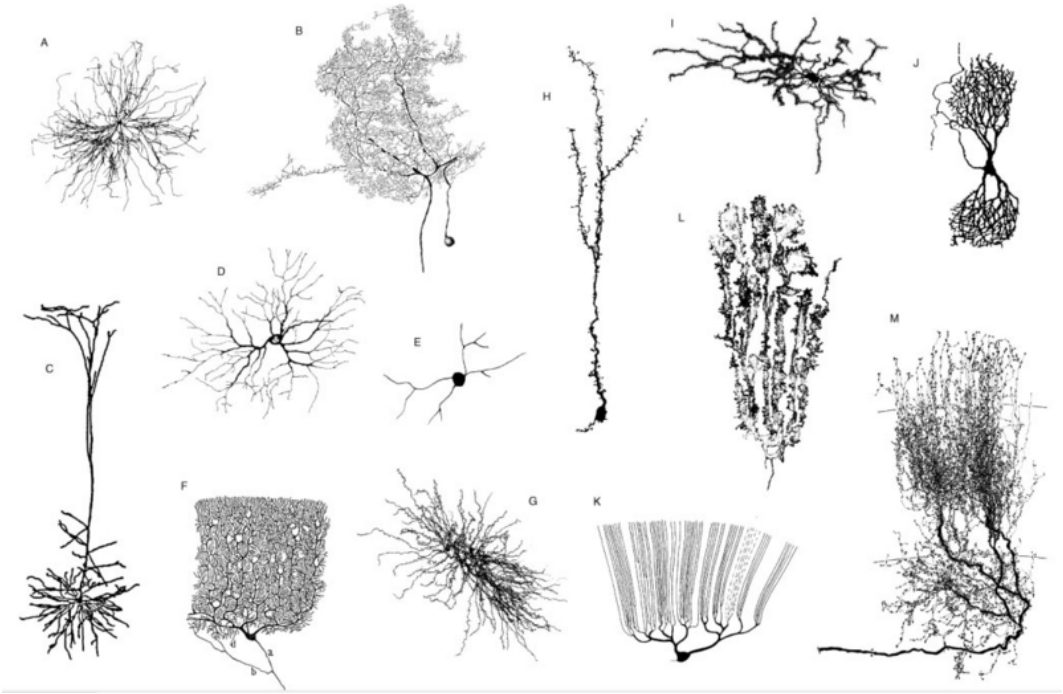


Figure 2.3: Some different single neurons, lengths given are approximate and correspond to direction of maximal extent. Adapted from Mel (1994). **A**. Alpha motorneuron in spinal cord of cat (2.6 mm) **B**. Spiking interneuron in mesothoracic ganglion of locust (540 **C**. Layer 5 neocortical pyramidal cell in rat (1030 μm) **D**. Retinal ganglion cell in postnatal cat (390 μ m) **E**. Amacrine cell in retina of larval tiger salamander (160 μ m) **F**. Cerebellar Purkinje cell in human **G**. Relay neuron in rat ventrobasal thalamus (350 μm) **H**. Granule cell from olfactory bulb of mouse (260 μm) **I**. Spiny projection neuron in rat striatum (370 μm) **J**. Nerve cell in the Nucleus of Burdach in human fetus **K**. Purkinje cell in mormyrid fish (420 μm) **L**. Golgi epithelial (glial) cell in cerebellum of normal–reeler mutant mouse chimera (150 μm) **M**. Axonal arborization of isthmotectal neurons in turtle (460 μm).

release and reception, according to the different characteristics of the many chemicals involved in the process.

Although they share many of the features we have just mentioned, there exist many kind of different neurons varying in shape, size and electrochemical properties, characteristics that determine their highly specialized functions in the nervous system (see Fig 2.3).

2.2.2 Spike generation mechanism

When the neuron is at rest, the cell membrane mantains a difference between the electric potential of the soma and of the outside of the cell. The cell body is negatively charged with respect to the external matter, whose potential is by convention set to zero; the value of this resting potential is about $-65mV$ (Squire and Kandel, 2009). The electrical conductivity of the membrane (a thin lipid bilayer) is, on average, remarkably small ($\sim 1\mu F$), but the presence of ion channels can increase local conductivity, so that charged ions and molecules may enter or exit the cell. In particular, biologically important ions, such as sodium (Na^+), potassium (K^+), calcium (Ca^{++}) and chloride (Cl^-), can cross the membrane through specialized channels (or gates) provided by proteins embedded into the phospholipidic layer. When the electrochemical equilibrium is violated, the membrane potential can either be as low as $-90mV$, which is called *hyperpolarization*, or increase to $-50mV$, that is called *depolarization*. These two mechanisms are activated in the post-synaptic neuron by the activity of the pre-synaptic cells. When a presynaptic neuron is excited, it releases a certain combination of neurotransmitters into the synaptic gap. The postsynaptic neuron has different kinds of receptors, that match with distinct neurotransmitters: depending on the combination of presynaptic transmitters and postsynaptic receptors, the resting potential of the postsynaptic membrane will be lowered or raised. As the potential becomes more negative, we say the synapse has an inhibitory effect, because it makes the generation of an action potential less likely. On the contrary, when the electric potential becomes less negative, the synapse is called excitatory, since it helps the postsynaptic cell reaching the threshold for the origination of the action potential. In the cortex , about 85% of synapses are excitatory, the majority of which connects pyramidal neurons, while only 15% are believed to be inhibitory. The majority of the neurotransmitters are either always

excitatory or always inhibitory, but there are some that can be both excitatory and inhibitory, depending on the receptors. The induced excitatory or inhibitory signal, regardless of its amplitude, is propagated through the dendrites and eventually reaches the cell body of the postsynaptic neuron, progressively vanishing over time. However, when a neuron is excited by more than one synapse in a short period of time, or when one synapse is repeatedly activated, the neuron's threshold potential (typically around $-50mV$) may be reached, and an action potential may then be generated. When the potential goes above threshold, all voltage activated sodium channels are opened simultaneously, and positively charged sodium ions rapidly flow into the neuron. Therefore, the potential of the neuron rises rapidly to a peak of about $25mV$. Then the sodium gates shut down completely for a few milliseconds, while the potassium gates gradually start opening, letting the K^+ move outbound.

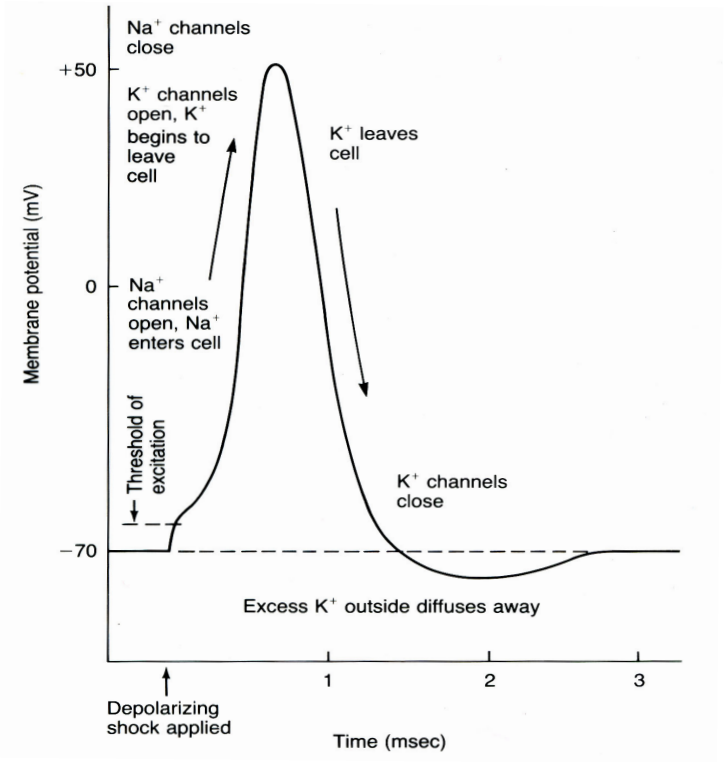


Figure 2.4: Sequence of biochemical processes involved in the generation of an action potential (Pinel, 2009).

After the potassium gates have been closed, the neuron's potential overshoots the resting potential, causing a hyperpolarization. During this interval, that is called *absolute refractory period* and lasts $\sim 1ms$, the cell is almost completely inhibited from producing further action potentials, and it reaches back the resting potential in a few milliseconds. The precise sequence of ion channels activation, which leads to the generation of an action potential, is reported in Fig. 2.4. The intensity of the produced electric signal is independent from the magnitude of the depolarization; it is a large, fast traveling spike that propagates along the axon always with the same amplitude. The signal does not suffer from any substantial attenuation, since the myelin sheath acts as a regenerator, helping the spike in mantaining its speed and amplitude.

2.2.3 Synaptic Plasticity

The second fundamental constituent that we wish to examine is the synapse. With the term "synapse" we will refer only to the chemical process described in the previous section, and not to the so called *electrical* synapses, which are much less commonly found in the vertebrate brain and possess different characteristics. Synapses mediate the transmission of the action potential from the pre-synaptic axon to the post-synaptic dendrites. Electrophysiological experiments have shown that the amplitude of the synaptic response varies over time, depending on the activity of the two connected neurons. This ability to change in strength is called *synaptic plasticity*. Theorists and experimentalists are convinced, on the basis of increasing evidences (Bliss and Collingridge, 1993; Bredt and Nicoll, 2003), that synaptic plasticity plays a key role in memory storage, learning and in the development of neuronal connectivity.

After a transition, the synapse may quickly return to its previous value (*short-term*

enhancement), or persist in its new state for a few minutes, as well as for several months. In this case it is called *Long Term Potentiation* (LTP), whenever an increase in the synaptic efficacy has occurred, or *Long Term Depression*(LTD) when the synaptic efficacy has decreased. Indeed, this is an example of adaptation in the nervous system, a property that, since the end of the 19th century, has been thought to be related to memory and learning processes (Ramon y Cajal, 1909; Tanzi, 1893). In theoretical applications, such as the neural network models outlined in Chapter 2, synaptic plasticy rules have been widely used long before experiments proved their validity, inspired principally by the so called *Hebb rule* (Hebb, 1949). Putting together behavioural evidences and neurophysiological data, canadian psychologist Donald Hebb formulated the following principle:

> Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability[...]. When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

The postulate suggests that activities of cells A and B during past times may trigger a synaptic modification that, in turn, will affect the future behaviour of the neurons. More generally, Hebb rule is interpreted as stating that synaptic efficacy variations are driven by temporal correlations in the activities of the pre-synaptic and post-synaptic cells.

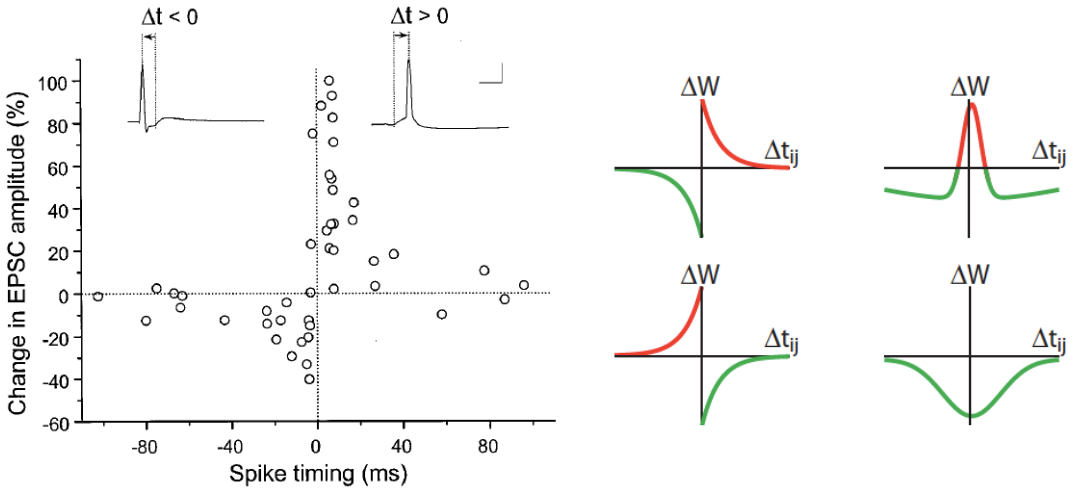
The first experimental clues on Long Term Potentiation were identified by Bliss and Lomo (1973) in the dentate gyrus of anaesthetized rabbit, and, in following years, many further evidences of synaptic plasticity have been noticed in various other areas of the mammalian brain, such as the hippocampus, the neocortex and the cerebellum. The most studied region, however, is the hippocampus, specifically the

synapses linking CA3 to CA1[4] pyramidal cells. Low-frequency stimulation of the pre-synaptic afferent does not result in any modification of the synaptic efficacy, while high frequency stimulation provokes LTP when both cells are firing. Since it requires both neurons to be active, this kind of LTP obeys the Hebb rule, which, in its original formulation, asserts that pre- and post-synaptic cells need to be active together in order to provoke the synaptic shift. In a real biological network, however, neurons are almost never temporally synchronized, and thus pre- and post-synaptic neurons may fire at different times. Indeed, what has emerged from experiments is that spike timing between neurons is extremely important to determine the effectiveness and the sign of the synaptic modifications (Bell et al., 1997; Bi and Poo, 1998; Markram et al., 1997). This property is commonly referred to as *spike timing dependent plasticity* (STDP).

The usual experimental protocol consists in measuring the magnitude of the change in synaptic weight, depending on the time delay elapsing between the evoked pre and post-synaptic action potentials. In Fig. 2.5 we have reported an example of experimental data from Bi and Poo (1998), on the left, and an exemplified representation of the STDP curves, on the right, drawn from various type of synapses found in different parts of the brain. While some synapses obey the Hebb rule, as can be noticed by looking at 2.5(a) and 2.5(b) left-top, one can also find anti-Hebbian plasticity (left-bottom) and even non-Hebbian rules (right-top and right-bottom).

After a number of early successes, however, researchers have struggled to obtain reliable and definitive findings about LTP. Moreover, given the many controversial outcomes collected from experiments, the biochemical processes underlying LTP are still greatly debated, and none of the many possible mechanisms that have been proposed is able to provide a full explanation of the phenomenon. Accordingly,

[4] *Cornu Ammonis*. The name comes from the anatomical resemblance between the hippocampus and the ram's horn, symbol of Jupiter.



(a) (b)

Figure 2.5: (a) Measurements of induced change in the Excitatory Post-Synaptic Potential as a function of time delay Δt between pre- and post-synaptic spikes. In vitro data from hippocampus (Bi and Poo, 1998). (b) Synaptic plasticity curves for different synapse types (from Shouval et al., 2010). The red portion indicates a positive change in the synaptic weight (LTP), while green portion is associated to a negative shift (LTD). Each kind of synapse can be associated with a specific learning rule (see the review of Abbott and Nelson, 2000).

although STDP has become a cornerstone for both experimentalists and theorists, the straightforward, linear approach based on temporal synchronization has shown its cracks, and the time has probably arrived to embody other features in the actual theoretical framework. Indeed, STDP role as a dominant mechanism for synaptic plasticity has been recently questioned by Shouval et al. (2010), who propose a broader theory based on cellular mechanisms, where STDP would be only one of the key parameters involved. In this thesis, however, we will take the Hebb rule as the paradigm of reference for synaptic plasticity, as we shall see in Section 3.2.1.

2.2.4 Binary synapses

So far, plasticity has been treated as a dynamical process that provokes a sudden shift in the value of the synaptic weights, depending on the activity of pre- and

post-synaptic neurons. Such a formulation leads, unavoidably, to an intrinsic instability. In fact, whenever two linked cells obeying the Hebb rule fire persistently closely in time, the synaptic coupling between them is strengthened. This higher synaptic efficacy facilitates, again, the activity of the two neurons, causing a further reinforcement of synaptic weight, and so on.

However, in a biological network, one would expect synapses to be bounded at some value, in order to avoid the infinite growth of activity that produces an unstable behaviour. This view is supported by in vitro experiments made in hippocampal CA3-CA1 regions by O'Connor et al. (2005) and Petersen et al. (1998). Results from collective data of synaptic populations had shown graded response to cell stimulation, endorsing the hypothesis that efficacies range over a broad spectrum of analog values. Instead, the aforementioned studies investigated the behaviour of single synapses, under stimulation of both pre- post-synaptic neurons. Their analysis shows that modifications of synaptic weight occurs swiftly (<1 min) and in an all-or-none fashion, with different thresholds for individual synapses. After a successful transition, the synapse saturates to one of the two stable states and gets stuck there for several minutes, even in presence of further stimulation.

According to these observations, stimuli are thus encoded in the network in a digital way, by setting the value of synaptic efficacies to one of the two stable states. Yet, two more problems are associated with this view: first, all synapses may saturate to one value causing the loss of stimulus selectivity, and second, if each new stimulus provoked a shift in all synaptic weights, older memories would be quickly erased and forgotten. Collective saturation can be fixed assuming competitiveness, that is, in correspondence of a stimulus, there would always be a certain number of synapses getting depressed while all others are getting potentiated and vice versa. At the same time, the problem of memory stability has been tackled by Amit and

Fusi (1994), who introduced the mechanism of stochastic learning (see Section 3.3.1) to reduce the average number of synaptic transition per stimulus. This way, stored information relating to old memories is preserved for a longer time. Furthermore, an appealing biological device, that could possibly account for stochastic transitions, is the fact that plasticity events are triggered with dissimilar thresholds in different synapses (O'Connor et al., 2005; Petersen et al., 1998).

Despite all these successes, however, experimental evidences for bistability are still scarce, further studies are hence needed to uncover the many unknown aspects of synaptic plasticity and put the work of theoreticians on the right path.

Chapter 3

Artificial Neural Networks

3.1 Early models

The previous chapter contains a brief overview of the biological bricks that lead to the computation and the storage of information in the brain. Quite obviously, the brain is immensely more complex than the description we have given. Apart from the extremely variegated and highly specialized functions expressed by different nerve cells, the brain is disseminated of many different substructures performing very different tasks, from receiving external sensory inputs to generating emotions and feelings. Almost inevitably, the first mathematical models of neural networks not only operate several drastic simplifications, but they also rely on very strong hypotheses that have not been necessariliy verified in experiments. In fact, one has to be aware that these pioneeristic models are overmuch elementary when considered from a neurophysiological point of view, and should not treat them as realistic, biologically exact descriptions. Nevertheless, they are a precious resource for un-

derstanding the basic strategies and principles of neural computation, both from a physiological and a bioengineering point of view.

In this thesis we are interested in the network properties, i.e. in the collective behaviour of a huge number of units (neurons) interacting with each other. For this reason, we make the (rather strong) hypothesis that we do not need the level of elaboration provided by detailed single neuron models, such as the Hodgkin-Huxley neuron. Instead, we will start from the simplest of the neurobiologically inspired computational units available in literature: the formal neuron of McCulloch and Pitts.

3.1.1 McCulloch-Pitts neuron

McCulloch and Pitts (1943) summarized the most salient characteristic of a biological neuron, already stated in Section 2.2.1, in a model that regards the neuron as a binary computing unit, in the context of Boolean logic. Of course, they had to make a number of radical preliminary simplifications:

- all neurons are identical
- subthreshold inputs do not trigger the release of any neurotransmitter, and suprathreshold inputs always leads to the generation of a single, identical spike
- the output of the unit depends only on the distribution of the upcoming inputs and on the set of synaptic efficacies

All these hypotheses are far from being fully justifiable from a neurophysiological perspective, but can represent a good approximation if we are interested primarily in the collective computational properties of the brain. In the McCulloch-Pitt's picture, the neuron is merely a discrete unit that elaborates a set of incoming inputs, which

in turn are generated from other binary elements, and returns a binary variable as the output.

The ideal scheme is displayed in Fig. 3.1. Binary inputs $x_1, ..., x_n$, coming from other neurons, are connected to the soma through the logical equivalent of the axons. Each channel, when activated, produces an input signal that is modulated by the synaptic efficacies (or *weights*) w_{ij}, where i is the pre-synaptic and j the post-synaptic neuron. The soma adds up these modified inputs

$$h_j(t+1) = \sum_i w_{ij} x_i(t)$$

and compares the result with the neuron's firing threshold θ_j. The unit emits a logical output at the next time step, according to whether the sum of the weighted inputs lies above threshold or not:

$$\sigma_j(t+1) = \varphi[h_j(t+1) > \theta_j]$$

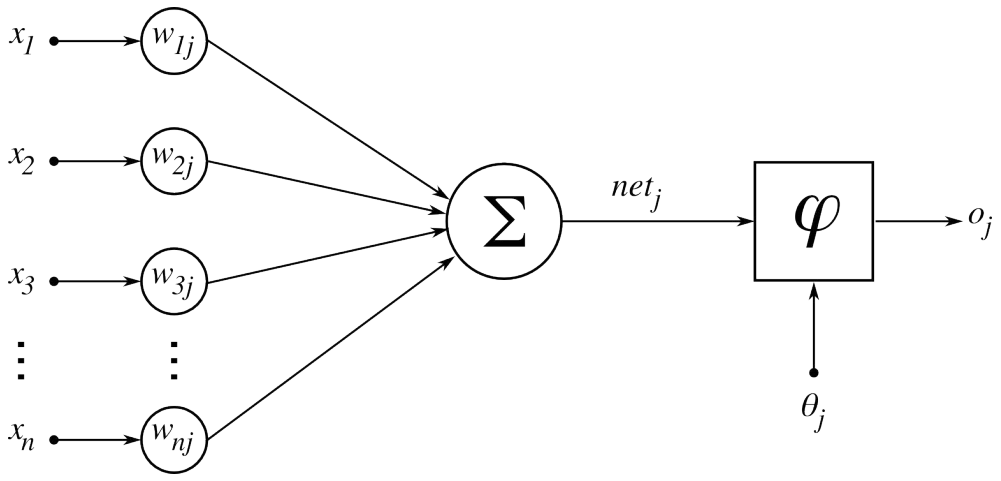


Figure 3.1: Scheme of a McCulloch-Pitt's formal neuron. Inputs x_i reach the soma after being multplied by synaptic weights w_{ij}. The soma releases an output if the total weighted sum exceeds threshold θ.

where the transfer function φ can be the Heavyside function or any nonlinear monotonic function.

McCulloch and Pitts had already noticed that a network of formal neurons is capable of universal computation[1], by choosing an appropriate set of weights $\{w_{ij}\}$. In later years other groups expanded the field, studying large systems of logical units and their computational properties, and nowadays McCulloch-Pitt's neurons are still used as the network basic elements when single neuron features can be put aside.

3.1.2 Rosenblatt's Perceptron

The most interesting among the many possible applications of the formal neuron is the *perceptron*. This system, that initially has been investigated by Rosenblatt (1962), consists in one or more layers of McCulloch-Pitts neurons with feed-forward pathways going from lower layers to upper ones. The lowest layer receives the external input, the highest computes the final output. Why is the perceptron interesting? First because it involves parallel computing, which is actually what the brain must be doing to produce complex responses in a few milliseconds, albeit the activation time of a single neuron pertains to the same order of magnitude. The second reason is that Rosenblatt proved that the perceptron, at least the most simple version with only one layer, is capable of *learning*. He found a convergence theorem ensuring that, following a specific learning rule, the weights can be updated iteratively to reach the desired output result, given a specific input. At the time, this finding provoked a wave of optimism in the scientific community, who believed that the perceptron could be a promising candidate to be the basis for artificial intelligence

[1]The power of of calculating every sequence of logical functions in a finite time. To do so, the device must be able to perform: boolean negation, one between conjunction and disjunction, and an associative relation defined in the space of logical operations (Russel and Whitehead, 1910).

devices. The enthusiasm partially dissolved when Minsky and Papert (1969) pointed out that non linearly separable[2] tasks, like for example the boolean exclusive OR (XOR), could not be worked out by a single-layer perceptron. Furthermore, they showed that performing some of these easy computations with the perceptron would take an excessive amount of time, a result that has later been extended to all linear threshold devices.

During the successive years, Rosenblatt and many others tried to bypass the problem using multi-layer perceptrons, that in principles possess the capabilities to overcome Minsky and Papert's limitations. On the other hand, the learning algorithms for multi-layer schemes are much more complicated, and there is no simple rule to find the desired set of weights. For many years then, the artificial neural networks field was abandoned by the majority of the researchers for more promising paradigms.

The interest in perceptron-like machines raised again in the community during the mid-eighties, when the *back-propagation* algorithm, conceived by P. Werbos in 1974, has been rediscovered and actively employed. To summarize its core in a few words,

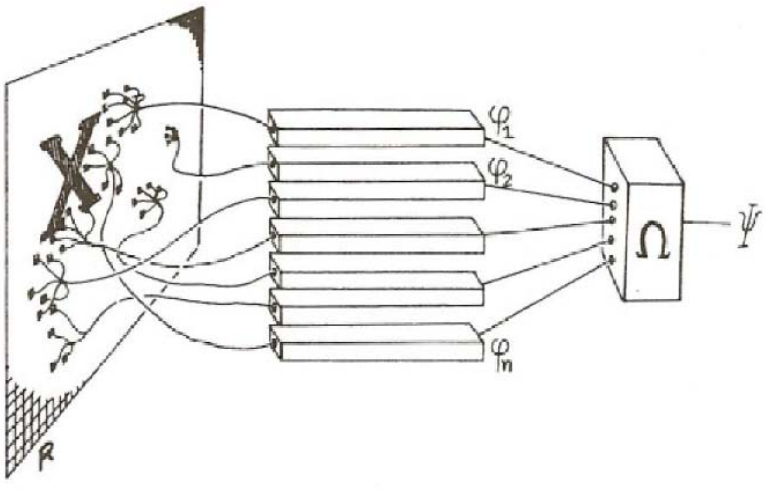


Figure 3.2: The perceptron as it is described in Minsky and Papert (1969).

[2]Two sets of points in n dimensions are said to be *linearly separable* if they can separated by a hyperplane in $n-1$ dimensions.

back-propagation allows multi-layered networks to find the correct weights associated with a specific set of input-output patterns. The key aspect is that errors (i.e. the difference between the desired and the actual outputs) are sent back from the last layer up to the first, where the weights are consequently updated for the next upcoming pattern. In this way, the perceptron can be trained to perform complex tasks and it is now utilized in data-mining, speech recognition tasks or financial forecasting. Unfortunately, the algorithm has still some disadvantages: first it is not guaranteed that the network reaches a global minimum (although the local minima problem has been solved using stochastic noise in the dynamics), and second the speed of convergence with the standard algorithm is extremely low.

3.2 Hopfield Networks

More interesting for us are, however, the developments that originated from the work on *associative content-addressable memory* by Willshaw et al. (1969), Marr (1969; 1971), Anderson (1970), Amari (1972; 1977), Little (1974; 1978), and culminated in the works of J.J. Hopfield (1982; 1984). Hopfield pointed out several subsantial connections and similarities between recurrent neural networks and physical systems with many degrees of freedom, such as magnetic lattices. Once established, this equivalence opened up the field to the application of many theoretical and mathematical tools that pertain to the field of statistical physics, leading to a much deeper comprehension of neural networks from innumerable points of view.

Let us start from the original problem, which may be stated in the following way: we would like to store p patterns ξ^μ ($\mu = 1, ..., p$) in a McCulloch-Pitt's neural network and obtain that, when the initial condition is represented by a pattern ζ^0, the network extracts the memory that is most "close" to ζ^0 among the stored ones. What do we mean by "close"? In this context, memories are represented through

binary words, whose size is determined by the number of units N populating the network. Closeness may then be measured using the Hamming distance, defined as the number of bits in which two patterns differ from each other. The network should be able to select the pattern with the smallest Hamming distance from ζ^0.
In this way, the network will retrieve the proper stored memory even when a partial match or a noisy, flawed version of the original pattern is presented. This kind of memory is *content-addressable*, meaning that one does not need to specify the physical location of the pattern in the network (unlike in ordinary calculators), but needs only to provide some clues about the content of the stored information. And, as we have said, it is *robust* to small errors in the input pattern, at least to some extent.

3.2.1 Networks dynamics: analogy with magnetic systems in physics

The kind of network suited for this task is a recurrent network of formal neurons, that is a perceptron where the output signal becomes the new input at the next temporal cycle. However, unlike Rosenblatt's formulation of the perceptron, in which all units belonging to the same layer operate synchronously, neurons are now updated in a random way (*asynchronously*). Synchronization has, in fact, the disadvantage, from the point of view of both biology and physics, that some kind of global information is needed: a universal time to which computational units have to adapt. Whereas in conventional computers synchronization of the digital components is achieved using a clock signal, there is no such global clock in biological systems. On the other side, it is quite difficult to build a realistic model of temporal interactions among nerve cells: real neurons fire when the membrane potential exceeds the threshold, which, in turn, happens when a sufficient number of action potentials have reached

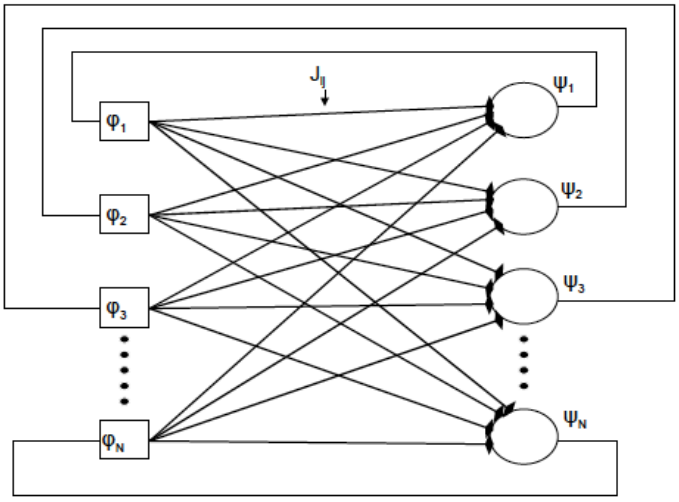


Figure 3.3: Example of feed-forward recurrent network of McCulloch-Pitt's units (Amit, 1989). The output of neurons ψ_i becomes the input φ_i at the next temporal step.

the pre-synaptic terminals. Thus, the activity of one neuron strongly depends on the activity of the others. The easiest way to deal with this complexity is to make the update process asynchronous and random. This, together with recurrency, gives rise to nonlinear effects that determine the dynamics of the system.

Before going further, let us make a linear change of variables to facilitate calculations in this section: instead of using 1 for firing and 0 for quiescent neurons (σ_i), we will use, respectively, 1 and -1 (S_i). Most results remain valid in both formulations, and one can always go back to the original choice with a linear transformation. In terms of the new variables, the equations of the dynamics become

$$S_i(t+1) = \mathrm{sgn}[h_i(t+1) + h_i^e > \theta_i]$$

$$h_i(t+1) = \frac{1}{2}\sum_{j=1}^{N} w_{ij} S_j(t) \qquad h_i^e = \frac{1}{2}\sum_{j=1}^{N} w_{ij}$$

where $\mathrm{sgn}(x)$ is the sign function, w_{ij} are the fixed synaptic efficacies, $h_i(t)$ is the local field acting on the i-th spin (neuron), induced by all surrounding spins (neurons), and h_i^e is analogous to a static external field, independent from the state of the other

neurons. This external contribution may exert a strong influence on the network behaviour, but here, for the sake of clarity, we choose the thresholds to be $\theta_i = h_i^e$ in order to balance its effect. Then, by redefining the synaptic weights as $J_{ij} = \frac{1}{2}w_{ij}$, we get

$$\begin{aligned} S_i(t+1) &= \text{sgn}[h_i(t+1) > 0] \\ h_i(t+1) &= \sum_{j=1}^{N} J_{ij} S_j(t) \end{aligned} \tag{3.1}$$

Keeping in mind these equations, we can illustrate the two crucial hypotheses made by Hopfield in his 1982 work:

1. synaptic weights are chosen according to the *generalized Hebb's rule* (see Section 2.2.3), which states that changes in the synaptic plasticity are determined by the correlations in the activities of pre- and -post-synaptic neurons during learning. This prescription, which is part of a much wider and deeper theory, has often been condensed into the catchy phrase *neurons that fire together, wire together*: two neurons that had repeatedly fired together will be more likely to fire in future. Applied to the present case, since we train the network to learn p patterns, the simplest rule is given by a superposition of these patterns:
$$J_{ij} = \frac{1}{N} \sum_{\mu}^{p} \xi_i^\mu \xi_j^\mu \tag{3.2}$$

2. the network is *fully connected* (every neuron is connected to all other neurons), and all connections are *symmetric*: $J_{ij} = J_{ji}$. While it is possible to find (almost) fully connected networks in the brain at a local scale, symmetric synapses are not biologically plausible for several reasons, including the fact that most neurons are either inhibitory or excitatory only[3]. However,

[3]This principle is known as *Dale's law*.

the postulate has proved to be immensely useful, for it is fundamental to the equivalence with disordered physical systems, which, in turns, made possible the employment of statistical mechanics for calculations. For the sake of simplicity, in this thesis we ignore autorecursive connections ($J_{ii} = 0$) that, by the way, do not influence the qualitative behaviour of the model.

These assumptions allow for the definition of a *Lyapunov function* for the system[4]

$$E = -\frac{1}{2}\sum_{i}\sum_{j\neq i} J_{ij}S_iS_j \tag{3.3}$$

whose minima are, by definition, attractors of the dynamics (Khalil, 2002). In fact, if we calculate the variation in the Lyapunov function caused by the inversion of one bit $S_i' = -S_i$:

$$\Delta E = E' - E = 2S_i\sum_{j\neq i} J_{ij}S_j$$

that is always negative because

$$S_i' = sgn\Big[\sum_{j} J_{ij}S_j\Big] \equiv -S_i$$

Hopfield noticed that this model is isomorphic to the well-known *Ising model* for magnetic lattices at zero temperature. Hence, the properties of a Hopfield network, also called *Attractor Neural Network* (ANN), can be investigated with the tools developed in the statistical mechanics of disordered systems and spin glass theory. One simply needs to replace the variable describing neuronal activity with a quantized variable, that schematically represents the orientation of the magnetic spin on each site of the lattice. At the same time, the interaction between post- and pre-synaptic neurons is supplanted by the spin-spin magnetic force, and the Lyapunov function

[4]Indeed, while the condition of complete connectivity may be relaxed (Derrida et al., 1987; Sompolinsky, 1986), the hypothesis of symmetric synapses is crucial for the definition of the Lyapunov function (Amit, 1989).

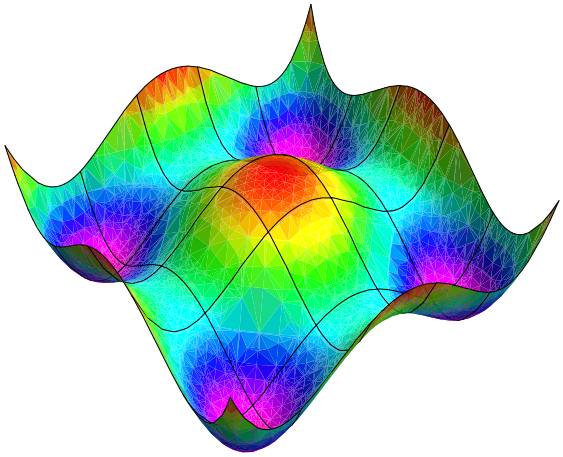


Figure 3.4: Example of energy landscape in the configurations space. Purple-labeled energy minima are attractors of the dynamics.

is identified with the energy of the system. If we map each state of the network, corresponding to a particular configuration of neural activities, to a 2^N hypercube, we can imagine a "landscape" formed by the energy function in the states space where lower points and valleys correspond to fixed points[5]. In absence of noise, when the dynamics is fully deterministic, the configuration of the network will always evolve towards one of this attractors, and, once inside the *basin of attraction*, will remain there forever. These attractors are, indeed, the recalled patterns: the retrieval depends on the initial stimulus (i.e. a corrupted version of the original pattern) and it is stable, since the network remains stuck there until a new stimulus is presented.

3.2.2 Stability of retrieved memories

Attractors of the network dynamics are determined by the synaptic weights[6]: in the previous noiseless case it is easy to see that the stored attractors are stable

[5]Be aware that the landscape picture makes sense only if it is possible to find a Lyapunov function for the system. If, for example, connections are not symmetric, it would not be possible to define such function, since the distance between two points in the configurations space would depend on the direction of moving.

[6]In the context of spin glass theory, the role of the synaptic weights is played by the magnetic interaction.

fixed points of the dynamics, when eq. (3.2) holds (Amit, 1989; Hertz et al., 1991). Let us sketch the key points of the demonstration. In general, it is preferred to draw input memories from some probability distributions, instead of specifying the value of every single bit. A common choice that gives uncorrelated patterns is the binomial distribution on $\{1, -1\}$. Using the Hebb generalized, one can train a fully connected network with asynchronous dynamics to memorize p random and uncorrelated (almost orthogonal) patterns. The condition for the stability of the attractor corresponding to memory ξ^μ is (from eqs. 3.1):

$$\xi_i^\mu = sgn\left[h_i^\mu\right] \qquad \forall\ i$$

that is equivalent to requiring

$$\xi_i^\mu h_i^\mu > 0 \qquad \forall\ i \tag{3.4}$$

The local field perceived by the i-th neuron is

$$h_i^\mu = \sum_{\substack{j=1 \\ j\neq i}}^{N} J_{ij}\xi_j^\mu = \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\sum_{\nu=1}^{p} \xi_i^\nu \xi_j^\nu \xi_j^\mu$$

using Hebb rule. Hence

$$\xi_i^\mu h_i^\mu = \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\sum_{\nu=1}^{p} \xi_i^\mu \xi_i^\nu \xi_j^\nu \xi_j^\mu$$

separating from the sum the term that concerns ξ^μ, one gets:

$$\begin{aligned} \xi_i^\mu h_i^\mu = \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\left(\xi_i^\mu\right)^2\left(\xi_j^\mu\right)^2 + \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\sum_{\substack{\nu=1 \\ \nu\neq\mu}}^{p} \xi_i^\mu \xi_i^\nu \xi_j^\nu \xi_j^\mu = \\ = \frac{N-1}{N} + \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\sum_{\substack{\nu=1 \\ \nu\neq\mu}}^{p} \xi_i^\mu \xi_i^\nu \xi_j^\nu \xi_j^\mu > 0 \end{aligned} \tag{3.5}$$

the first factor is the *signal*, since it expresses the contribution of the tracked memory, the second term gathers the interfering effect of the other patterns, and is therefore called *noise*[7]. As far as the noise factor is smaller than the signal, the attractors are stable and the memory is successfully recalled. Noise is of course zero when all patterns are all perfectly orthogonal. Instead, for uncorrelated, unbiased patterns, the sums inside the noise term are equivalent to a random walk of $\sim (N-1)(p-1)$ steps of magnitude ± 1. The mean of a random walk, whose probability distribution is simply a binomial, is zero and the variance, i.e. the average fluctuation around the mean, is of order $(N-1)(p-1)$. Hence, in the limit of large networks ($N \to \infty$), eq. (3.5) may be rewritten as

$$\xi_i^\mu h_i^\mu \approx 1 + Noise > 0 \qquad\qquad |Noise| = \sqrt{\frac{p}{N}}$$

from which we notice that, if the number of stored patterns is much smaller than the number of units ($p \ll N$), the dynamics would not jump out of attractors once has reached one of them, i.e. memories are stable. Moreover, if the network is shown a corrupted stimulus (obtained by flipping a fraction d of bits) instead of correct pattern ξ^μ, the dynamics will anyway converge and stay into the desired stable state, provided that d is not too large. In fact the local field acting on the i-th neuron would read

$$h_i = 1 - 2d + Noise \approx 1 - 2d + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$$

meaning that, as long as $1-2d \gg N^{-1/2}$, the network is still in the basin of attraction of the desired memory ξ^μ.

Unfortunately, in addition to the p requested pattern, some unwanted, *spurious*

[7]Sometimes, like in Sections 3.2.1 and 3.2.2, the word *noise* refers to the presence of a stochastic term in the dynamics which can destabilize spourious attractors, that may be introduced into the equations as a fictious temperature. Here, in Section 3.3.2 and in Chapter 4, it refers to the fluctuations induced in the signal by non-orthogonal patterns.

states also obey condition (3.4) and are, therefore, local energy minima, causing errors in the retrieval. Such states are, for example, linear combinations of an odd number of stored patterns, or simply the reversed version of the originals $\left(-\xi^{\mu}\right)$. To avoid the possibility that the system selects one of these states, one must introduce some sort of *stochastic noise* in the dynamics[8], so that the system can jump out of the basin of attraction of spurious states (which by the way is smaller than for the originally stored patterns). In the equations, noise is introduced as a pseudo-temperature, since it plays the role of temperature in spin glasses, representing the stochastic disorder always present in real networks (Amit, 1989). A disorder that may be due to fluctuations in neurons' firing rate and in the synaptic efficacies, or to any kind of external noise acting on the network.

In the presence of noise, of course, the network is not deterministic anymore, and each time we run eqs. (3.1) we would get a different realization. Thus, to calculate the relevant quantities, it becomes necessary to average over all possible realizations, by making use of the ordinary methods of statistical mechanics[9].

3.2.3 Network storage capacity

One of the most powerful results in the field has been obtained by D. Amit, H. Gutfreund and H. Sompolinsky (1985; 1987), who applied mean field theory and the replica method to calculate the memory capacity of a Hopfield network. The patterns to be stored were generated at random $\left(Pr(\xi_i^{\mu}=1)=Pr(\xi_i^{\mu}=-1)=\frac{1}{2}\right)$ and the network was tested with and without noise in the dynamics.

The final result is exhibited in Fig. 3.5 in the form of a phase diagram whose axis are the noise level, i.e. the pseudo-temperature T, and the load parameter $\alpha=\frac{p}{N}$,

[8]This kind of noise is often referred to as *fast* noise, to distinguish it from *slow* noise, which is caused by interferences among the stored patterns.

[9]These methods are valid if the system has reached an *equilibrium state*. For systems that are out of equilibrium, things are much more complicated since Gibbs' formulation loses its validity.

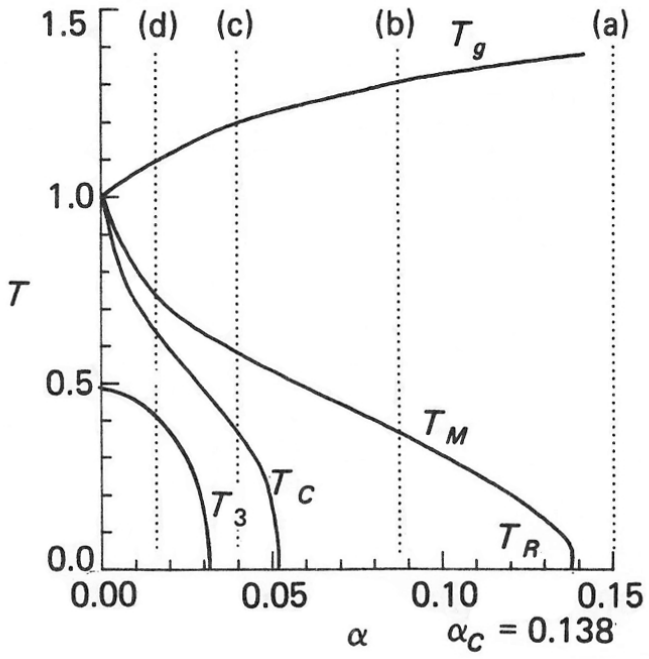


Figure 3.5: Phase Diagram of the Hopfield model (Amit, 1989).

that is the number of desired patterns divided by the number of neurons composing the network. For very high T the system is a paramagnetic-like phase, where the dynamics is ergodic and every state with $\langle S_i \rangle = 0$ is visited with equal probability. The energy function[10] has no minima and no retrieval is possible at all. Moving down along one of the lines indicated in the figure, the system enters in the so called *spin glass* phase below T_g: there are an infinite number of energy minima but none of them has a significant overlap with the desired state. If the load parameter is lower than $\alpha_c \approx 0.138$ the system goes into a ferromagnetic phase and retrieval of desired patterns becomes possible. Inside this area, three distinct subphases may be put on evidence: one between T_M and T_C, in which target memory states are local minima and can be recalled even though the probability or errors is considerable. Then, below T_C, the desired memories are global minima of the free energy and retrieval

[10]When $T \neq 0$ the appropriate Lyapunov function is the free energy $E - TS$, since the entropy contribution must included.

is optimal, but, if α is too small, spurious stases, such as mixed combinations of the stored memories, evolve into local minima and the system may fall into one of them. Along the $T = 0$ line the replica symmetry is broken and an abrupt transition is encountered when crossing the critical value of the load parameter α_C: the average overlap with desired states drops from about 0.97 to 0. The failure is due to the fact that the number of spurious attractors grows exponentially with the number of desired memories.

In the simplest case (T=0), then, the maximum number of retrievable patterns is $p_{max} \approx \alpha_C N$, allowing for small errors ($< 1.5\%$). This result is compatible with the work of Weisbuch and Fogelman-Soulié (1985), in which they performed a signal to noise[11] analysis, similar to that of the previous section, for very large N. In that limit the binomial distribution, typical of discrete random walks, can be approximated by a continuous gaussian distribution. Using this fact and requiring that the probability of error in any bit of all desired patterns goes to zero in the thermodynamic limit, one gets an expression for the maximum capacity in the zero temperature limit:

$$p_{max} \simeq \frac{N}{4\log(N)} \tag{3.6}$$

but, of course, it is still true that, as soon as p is higher than $0.138N$, all memories are suddenly lost, since their overlaps drop to zero.

3.3 Realistic constraints on learning

Many studies highlighted the robustness of the Hopfield model, even to rather drastic alterations of its original hypotheses, like a less strict hebbian prescription, asymmetrical weights or diluited connections. However we have seen that, as soon as

[11] Here noise simply denotes the square root of the fluctuations in the network activity caused by the corrupted inputs, and it is not related to the pseudo-temperature T.

we exceed the maximum capacity by adding more patterns to be stored, the entire memory of all patterns is suddenly lost[12]. This makes the model inconvenient for practical purposes, such as artificial implementations, but is also against common human experience: we do not lose memories all together when a certain threshold is crossed, instead we tend to progressively forget about past memories and to recall recent facts more vividly.

3.3.1 Bounded synapses and stochastic learning

The first effort to overcome the disastrous blackout effect has been attempted by Nadal et al. (1986), Mezard et al. (1986) and Parisi (1986), who proposed to modify the Hebbian rule in such a way that the synaptic efficacies cannot become arbitrarily large but, rather, run over a finite set of discrete values. This assumption has also full biological relevance: as we have seen in Section 2.2.4, discrete synapses are compatible with experimental data from various areas of the brain. The second central feature of the model, if the synaptic values have been properly set and bounded, is that the network exhibits the *palimpsest* property: older memories are progressively forgotten and only most recent patterns can be successfully retrieved. A comprehensive analysis was carried out by Sompolinsky (1986) for two- and three-states clipped synapses. If synaptic efficacies are discrete and bounded, the network does not fall into the total blackout state and its capacity is only slightly lower than the result obtained by Amit and collaborators. But there was a key issue not considered by these authors: those results were obtained by clipping the synaptic strength *after* constructing the synaptic matrix, thereby making the implicit assumption that the desired patterns had already been learned somehow. What happens if, instead, synapses are limited from the very beginning, before learning takes place?

[12]This is strictly true only for noiseless dynamics, but the effect is very similar, although more gradual, also in the presence of noise.

The consequences, exposed in Amit and Fusi (1992), lead to a drastic reduction of the storage capacity from $\backsim \mathcal{O}(N)$ to $\backsim \mathcal{O}(\log N)$, which makes this kind of memory practically useless. The reason is that, in this scenario, the dynamics of the learning process becomes important for studying the properties of the model. In fact, the rate of pattern presentation, in respect to the synaptic refreshing time constant, affects the properties of the network and can lead to very different results for capacity. The simplest case happens when only one pattern is presented between two synaptic clipping events. Each memory is presented only once and never again (*one-shot learning*). Suppose we present a first memory pattern ξ^1 to a network with two-states synaptic strength: neuronal activity will be updated according to eq. (3.1) with synaptic efficacies

$$J_{ij} = \xi_i^1 \xi_j^1$$

The synaptic dynamics is usually much slower than neuronal dynamics, thefore the state of the system would rapidly converge onto the attractor corresponding to ξ^1, before the occurrence of a new synaptic update, and the pattern is successfully learned. However, when the next pattern, ξ^2, is imposed, some synapses will be refreshed again, according to $J_{ij} = \xi_i^2 \xi_j^2$, and the memory of previous pattern previously contained in those synapses is completely erased. After p-patterns, the fraction of synapses still preserving the initial memory are only $\backsim 2^{-p}$. Thus, after requiring at least one synapse to be in the untouched group, one gets $p_{max} \backsim \log N$. This learning dynamics quickly destroy any trace of older patterns, but, on the other side, new patterns are perfectly learned.

In order to improve the network capacity, the learning process should be somehow modified to slow down the forgetting process, at the cost of reducing the precision of one-shot learning. This can be done by limiting the number of synapses that update

their values at each refreshing cycle. Ideally, the network would need a mechanism to select only some of the synaptic strength to be updated, upon the arrival of a certain input. A suitable solution for unsupervised[13] learning is *stochastic* synaptic plasticity: each synaptic weight undergoes a transition only with some probability. When a new stimulus approaches, only some fraction of all synapses is actually refreshed. As a consequence, the activity of the network detected right after the pattern presentation will show a smaller overlap with the original stimulus, but, on the other side, synapses bearing the memory traces of older patterns will be less likely changed.

In this thesis we will use the stochastic learning framework to memorize patterns. For this reason a more detailed account of the model is given in Section 4.1.1.

3.3.2 Sparse coding

So far, input patterns were generated completely at random: the average activity triggered by one pattern was $\langle \xi_i^\mu \rangle = 0$ for $\{+1, -1\}$ units, or equivalently $\langle \xi_i^\mu \rangle = \frac{1}{2}$ for $\{1, 0\}$ neurons. We will now discuss the results for unbalanced (or *biased*) memories with neurons taking $0, 1$ values, in particular when patterns are *sparse*, i.e. trigger low average activity levels in the network. Patterns are drawn from a binomial distribution with probabilities

$$Pr(\xi_i^\mu = 1) = f \qquad Pr(\xi_i^\mu = 0) = 1 - f$$

where of course $f = \frac{1}{2}$ in the unbiased case, and $f \ll \frac{1}{2}$ in the sparse limit. The average number of active neuron per pattern is not zero anymore since

$$\langle \xi^\mu \rangle = \frac{1}{N} \sum_i^N \xi_i^\mu = f$$

[13] Unsupervised learning means that the network does not make use of an external guidance for modifying the learning parameters, such as synaptic weights.

There are at least two reasons for considering biased memories. The first comes from the results exposed in the last section: the overlapping portion between two patterns acts as a noise for the dynamics, thereby reducing the effective size of the basin of attraction of each memory. A lower average activity means less interferences, and thus less noise and more stability, suggesting some advantages in terms of capacity. The second reason regards experimental evidences of neural activity in the cortex: leaving out some nontrivial difficulties in measuring the sparseness in real brain, experiments suggest a high degree of sparseness in many areas, such as the hippocampus (Barnes et al., 1990; Jung and McNaughton, 1993), medial temporal lobe (Quiroga et al., 2005), visual memory experiments in Inferotemporal cortex (Brunel, 1994; Miyashita, 1988; Rolls and Tovee, 1995; Sato et al., 2007).

The general problem for biased patterns (with unbounded synapses) has been treated by Tsodyks and Feigelman (1988), Buhmann et al. (1989), and Gardner (1988), who found some very interesting results at the price of little modifications in the construction of the synaptic matrix.

They took the local hebbian rule of eq. (3.2) adding a global factor proportional to the average activity of the network:

$$J_{ij} = \frac{1}{f(1-f)N} \sum_{\mu}^{p} (\xi_i^\mu - f)(\xi_j^\mu - f) \tag{3.7}$$

Furthermore, they did not set the threshold value as to balance the external field ($\theta_i = h_i^e$), but instead used a uniform threshold U that, they noticed, optimizes the network memory capacity. The local field acting on the i-th bit, when the network

lies in the attractor corresponding to stimulus ξ^1, reads

$$h_i^1 = \sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}\xi_j^1 - U = \frac{1}{f(1-f)N}\sum_{\substack{j=1\\ j\neq i}}^{N}(\xi_i^1 - f)(\xi_j^1 - f)\xi_j^1 + \\ + \frac{1}{f(1-f)N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\mu=2}^{p}(\xi_i^\mu - f)(\xi_j^\mu - f)\xi_j^1 - U$$

where we have divided the signal and the noise term. The signal is simply given by

$$\mathcal{S} = \xi_i^1 - f$$

and thus depends on the the value of ξ_i^1, being equal to either $1-f$ when $\xi_i^1 = 1$ or to $-f$ when $\xi_i^1 = 0$. The threshold U is conveniently chosen to optimize the contributions coming from properly oriented spins and, at the same time, minimize the noise provoked by the wrongly flipped ones:

$$U = \frac{1}{2} - f$$

The global factor inside eq. (3.7) cause the noise term to be zero on average, but, similarly to Section 3.2.2, it fluctuates with variance given by:

$$\mathcal{N}^2 = \frac{1}{\big[f(1-f)N\big]^2}\sum_{j\neq i}^{N}\sum_{k\neq i}^{N}\sum_{\mu=2}^{p}\sum_{\nu=2}^{p}(\xi_i^\mu - f)(\xi_j^\mu - f)\xi_j^1(\xi_i^\nu - f)(\xi_k^\nu - f)\xi_k^1$$

where the only surviving terms are those such that $\mu = \nu$ and $j = k$. Hence the variance is equal to

$$\mathcal{N}^2 = \frac{1}{\big[f(1-f)N\big]^2}\sum_{j\neq i}^{N}\sum_{\mu=2}^{p}(\xi_i^\mu - f)^2(\xi_j^\mu - f)^2\big(\xi_j^1\big)^2 = \\ = \frac{p}{\big[f(1-f)\big]^2 N}\langle(\xi_i^\mu - f)^2\rangle\langle(\xi_j^\mu - f)^2\rangle\langle\xi_j^1\rangle = f\frac{p}{N} \equiv f\alpha$$

The signal to noise ratio is

$$\frac{\mathcal{S}}{\mathcal{N}} = \frac{1}{2\sqrt{f\alpha}}$$

A rough estimation of the capacity is obtained requiring the signal-to-noise ratio to be at least of order one, then

$$\alpha_c(f) \approx \frac{1}{4f}$$

The more rigorous calculation performed by (Buhmann et al., 1989; Tsodyks and Feigelman, 1988), who employ the methods of mean field theory, returns for the capacity

$$\alpha_c(f) = -\frac{1}{2f \log f}$$

This expression diverges for very small f, confirming the previous intuition about interfering patterns. Reducing the coding level allows for a much higher storage capacity, but, of course, it diminishes the average amount of information carried by each pattern. To mantain a finite amount of information per memory, in the limit of large networks ($N \to \infty$), one could set the coding level as low as $f = \frac{\log N}{N}$, and obtain an optimal storage capacity of

$$p_{max} \approx \left(\frac{N}{\log N}\right)^2$$

At this point the advantage of using sparsely coded patterns should be sufficiently evident: at the price of losing some of the information content carried by each pattern, the capacity is enormously enhanced compared to the unbiased patterns case. Furthermore, this feature arises for $\{1, 0\}$ units only, which is the most natural choice for modelling a firing and a quiescent neuron. Instead, some caution is needed when mapping to the $\{+1, -1\}$ picture, since correlations give rise to different results.

3.4 A hierarchy of memories

Starting from sparse patterns, some authors studied the network behaviour when the desired memories exhibit a certain degree of correlation with each other. Initially, the effort was inspired by the fact that the energy minima in the Sherrington-Kirkpatrick (SK) spin glass model (Sherrington and Kirkpatrick, 1975) with long range interactions form an *ultrametric tree* (Mezard et al., 1987). The well known analogy between spin glass models and Hopfield networks elicited a wave of interests towards the utilization of ultrametric hierarchies in ANN (see next section).

But a second important reason for studying correlated patterns is that the encoding of uncorrelated quasi-orthogonal patterns does not seem to be a suitable strategy, if one aims to model the human memory. It appears indeed evident that our memories are deeply connected, tied to each other to the point that recalling a single event may evoke a cascade of closely related experiences (Klatzky, 1980). A natural way to build these relationships among experiences would be to organize them into categories: similar stimuli cluster together to form different groups and subgroups. The formation of a structured hierarchy helps both memory storage and retrieval. For example if we need to store a sequence of animals, it would be more convenient to memorize a stylized patterns that embrace some common features, such as the number of legs or the presence of a tail, and then distinguish every single animal for its peculiar characteristic (e.g. giraffe's long neck or rhinoceros' horn). This way, the network does not have to memorize all the informations about each pattern, but can, instead, store a smaller portion related to the specific differences between the pattern and the category representation. The more similar is the pattern to the cluster prototype, the lesser would be the stored information, and hence, the higher the network capacity.

Evidences for memory categorization emerge in the inferior-temporal cortex (Kriegesko-

rte et al., 2008; Sato et al., 2007; Tsunoda et al., 2001; Wang et al., 1996), in the parietal cortex (Freedman and Assad, 2006), as well as in the hippocampus (Hampson et al., 2004) and in the medial-temporal lobe (Kreiman et al., 2000). The results of Fig. 3.6, taken from the work of Kriegeskorte et al., show that the neural representations of stimuli tend to form clusters based on their mutual similarity.

3.4.1 Ultrametric trees

We have said in the previous section that a candidate model for representing hierarchical structures is given by ultrametric trees, given the analogy with energy minima in the SK model. The mathematical concept of ultrametricity is often employed in disciplines like semantics, taxonomy and data analysis, since it allows the study of hierarchial categorizations with an accessible mathematical formalism.

Ultrametric spaces are characterized by a different definition of the metric distance compared to usual euclidean spaces. The ultrametric version of the usual triangular inequality is

$$d(A,C) \leq Max\Big(d(A,B), d(B,C)\Big) \tag{3.8}$$

which implies that three points in an ultrametric space may only form equilateral triangles or isosceles with one edge shorter than other two[14].

An explicit graphical representation of an ultrametric space is an indexed hierarchical tree (see Fig. 3.7), where the distance between points, lying at the bottom of the tree, is a function of the number of steps one has to climb in order to find the first common ancestor. It has been demonstrated by Rammal et al. (1986) that any unbiased evolution process, acting on a system with many degrees of freedom (high number of neurons N in our case), gives rise to an ultrametric set of vectors

[14]This is not the only topological oddity of these spaces, in fact each point lying inside a ball of radius r is itself at the center of the ball, and the radius of the ball coincides with its diameter.

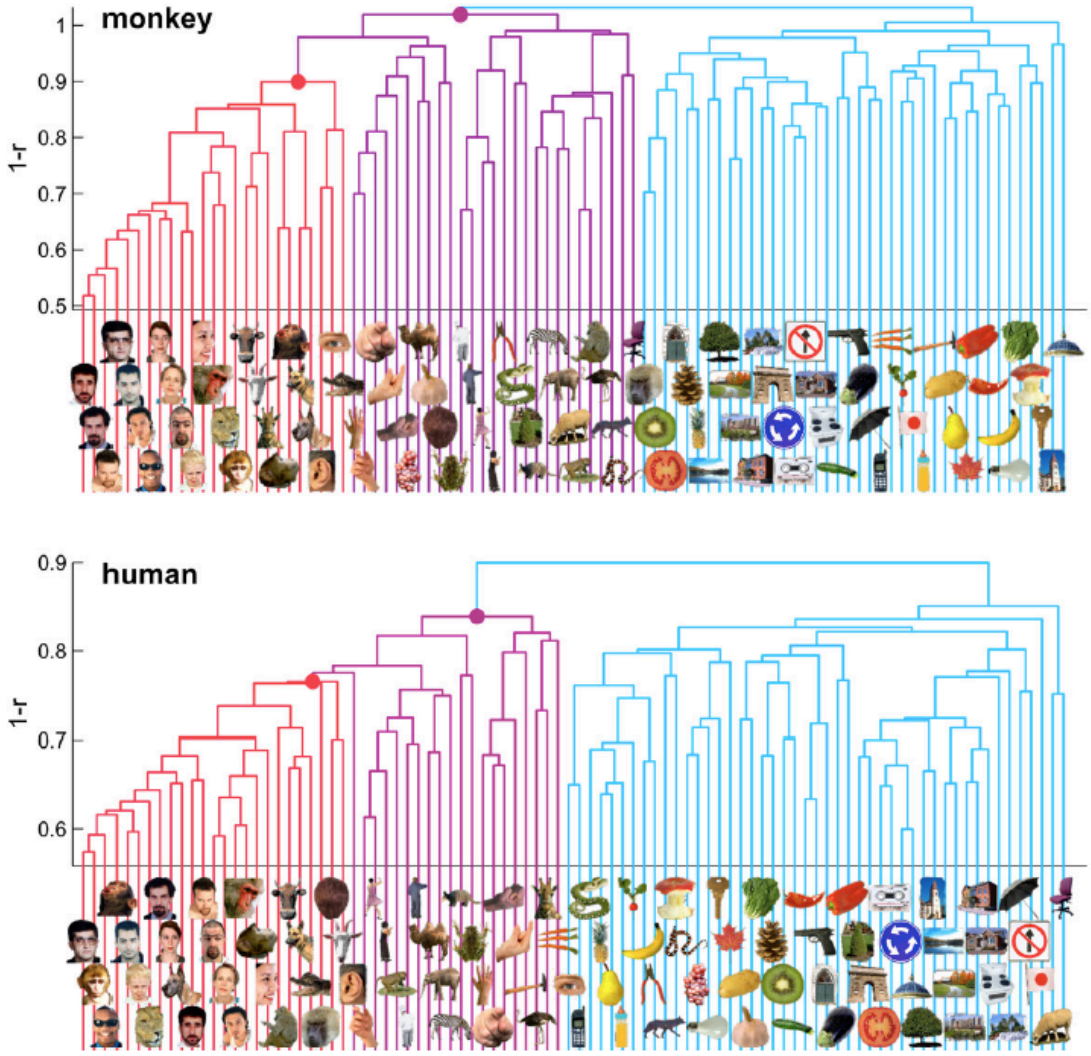


Figure 3.6: Hierarchical clustering of patterns in human (top) and monkey (bottom) IT cortex, from Kriegeskorte et al. (2008). Experimenters measured the degree of dissimilarity (**1-r**) among neural patterns, in response to the presentations of objects in the visual field. Human data are collected from fMRI study while monkey data come from single-cell recordings. The intraspecies analysis proceeds from single-image clusters (bottom of each panel) and successively combines the two clusters closest to each other in terms of the average response-pattern dissimilarity, so as to form a hierarchy of clusters (tree structure in each panel). On the vertical axis is shown the average response-pattern dissimilarity between the stimuli of the two linked subclusters. Quite amazingly, hierarchical trees for monkey and human present a similar categorization tendency, although they are the result of completely independent experiments and analysis techniques. For example, all human faces trigger a similar neural response-pattern in both experiments. Subcluster trees are coloured for easier comparison (faces, red; bodies, magenta; inanimate objects, light blue).

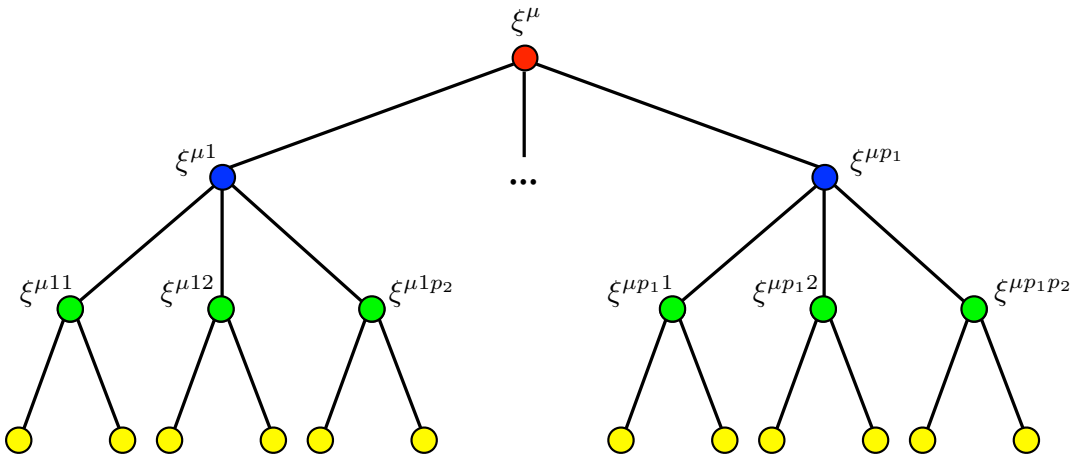


Figure 3.7: Ultrametric tree with four generations. The distance between any of the yellow states obeys the ultrametric inequality eq. (3.8). The branching ratio is not globally fixed but depends on the layer at which the branching occurs.

that can be organized within a hierarchical tree. Thus, by choosing the appropriate stochastic process responsible for the evolution, one could generate a correlated set of neural activity patterns placed at the bottom of an ultrametric tree.

As an example, let us examine the ultrametric tree shown in Fig. 3.7. The highest node is occupied by the prototype, a vector pattern ξ^{μ} whose N elements are chosen at random from a discrete, two-state (0,1) distribution. The first generation of descendants $\xi^{\mu 1}$, $\xi^{\mu 2}$, ..., $\xi^{\mu p_1}$ is generated by flipping each component of ξ^{μ} with probability distribution $P(\xi_i^{\mu\kappa}|\xi_i^{\mu})$, fixed for all $\xi^{\mu\kappa}$. Repeating this procedure with all the newborn patterns, one obtains a third layer of branches defined by probability distribution $P(\xi_i^{\mu\kappa\rho}|\xi_i^{\mu\kappa}, \xi_i^{\mu})$, and so forth for as many generations as wished.

For patterns belonging to the very last ramification, there exists a metric function that obeys eq. (3.8) when $N \to \infty$:

$$d_h\big(\xi^a, \xi^b\big) = \frac{1}{N} \sum_{i=1}^{N} \big[\xi_i^a(1-\xi_i^b) + (1-\xi_i^a)\xi_i^b\big] \tag{3.9}$$

that is the Hamming distance for binary ($\xi_i = 0, 1$) patterns. This is, by definition, a measure of pattern similarity. In fact, patterns whose common ancestor is found in a lower layer share more features than patterns whose first commutual forefather is located at the top of the tree. Moreover, it is important to notice that, because of the way we defined the stochastic evolution, ancestor vectors coincide with the average of all their offsprings. It is easy, then, to identify vectors that branch from the same node as elements of a cluster or subcluster, whose prominent common features are summarized in the correspondent ancestor.

Parga and Virasoro (1986), at the price of a slight alteration of the Hebbian learning rule, constructed a network that is able to memorize these ultrametric hierarchies in the static case and with unbounded synapses. In the next chapter, we will use ultrametric trees to model input patterns stored in a dynamic manner, to test if the network is somehow capable of generalization, that is to retain the features common to the whole cluster, and discard those peculiar to each single vector.

Chapter 4

Learning models and simulations

Summary of results

In this section we investigate the behaviour of a recurrent network of binary neurons ($\xi_i = 0, 1$) and bounded synaptic strength performing an associative memory task (Amit and Fusi, 1994). As we exposed in the previous chapter, this network displays the palimpsest property, i.e. it retains only the most recent memories, while older ones are gradually overwritten. Our aim is to estimate and optimize the average lifetime of a typical memory that, we will see, it is a good estimator to the storage capacity of the network. We postulate that the synaptic efficacy can take only two values ($J = 0, 1$), that the network is fully connected, and that synaptic updating occurs on a sufficiently large time scale, so that the underlying neural dynamics converges to the desired attractor before the arrival of a new stimulus.

First we carry out the complete analysis of the random and uncorrelated case in the sparse coding limit. We calculate the signal produced by a generic memory, defined

as the degree of correlation between the synaptic matrix and the given pattern, and the variance of the signal caused by the learning of other memories. This method is known in the physics literature as mean field approximation. The maximum memory capacity of the network is derived imposing a lower bound on the signal-to-noise ratio: if the memory signal is so low that fluctuations may dominate, the memory is considered to be lost. With random and uncorrelated patterns the optimal capacity is reached for very sparse patterns.

In the second part of the chapter, we turn our attention to ultrametric hierarchies consisting first of two and then of three levels. Each ultrametric tree consititutes a separate class that spreads from a prototypical pattern. This pattern is constructed as the average of all patterns belonging to that specific cluster. Similarly, when the hierarchy is composed of three stages, patterns occupying intermediate nodes are the averages of their descendants. This property yields the definition of subclass, a group of patterns that are more correlated with themselves than with other members of the class.

We derive the mean field equations that determine the expression of the overlaps between the synaptic matrix and the pattern of interest, and we test them with several simulations. Interestingly we found that the network is capable of categorization, i.e. of learning the class or the subclass prototype only by extracting the average features of the presented stimuli. Categorization occur when the following conditions are verified:

1. slow learning, meaning that only a small fraction of synapses is affected by the presentation of a new pattern. One the one hand, decreasing learning speed helps old patterns not to be erased, on the other hand, however, it makes more difficult for new memories to be learned.

2. sufficiently high correlations between class (or subclass) members and proto-

types. In fact, if intra-class patterns have only a small number of features in common, the overlap of the synaptic matrix with the prototype would be weak.

The categorization property is then exploited to build a sparser representation of the stimuli. In fact, since we can extract the class ancestor using a properly tuned network, we can imagine to have a second network that stores the uncorrelated fraction of bits between a prototype and a class member. The original pattern is then retrieved automatically, as the input recalls the prototype pattern stored in the first network, which in turn projects to the other network cointaining the uncorrelated portion. This way we significantly enhance the number of stored memories, at the price of adding a network to the system.

4.1 Random and uncorrelated memories

Each memory that we impose to the network produces a different pattern of neuronal activity, and consequently triggers the learning process through the modification of the synaptic efficacies. At this point two distinct sources of stochasticity may affect learning: the first is given by the sequence of presentation of the inputs and the second by the stochasticity in the synaptic potentiation and depression mechanism, that we have seen to be a condition for learning in Section 3.3.1. This stochasticity may play the part of a non-deterministic noise in the real brain, either due to intrinsic fluctuations in the system (in the neurons spiking activity or in synaptic thresholds) or to the fact that the brain can suffer unequal conditions in different contexts. In this section we will present past results regarding uncorrelated streams of inputs, mainly due to the work of Amit, Fusi and Brunel. The brain receives a massive amount of external stimuli from a reality that is often coherent: both natural and human crafted environments show structured (though sometimes

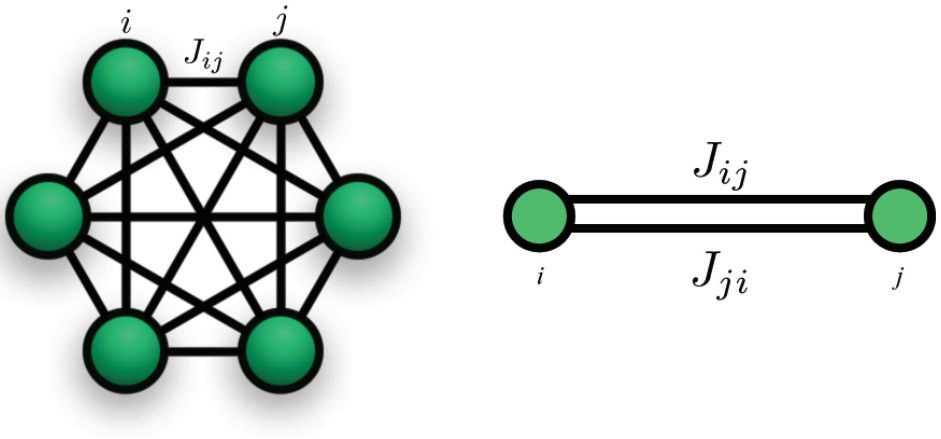


Figure 4.1: Topology of the fully connected network we study in this thesis. We consider connections J_{ij} and J_{ji} between neurons i and j as distinct synapses.

extremely complex) patterns. Sensory patterns approaching neural networks are therefore correlated, both spatially and temporally. It seems reasonable to argue that neural circuits have been shaped by evolution in order to be optimized for the world we live in, and correlations must be somehow exploited for this purpose. As a first primitive approximation it is anyway preferrable to start investigating the learning of a group of uncorrelated, randomly generated stimuli, before proceeding towards more complex scenarios.

A memory is represented by a string of N binary variables η_i, $(i = 1, ..., N)$:

$$\eta^{\mu}_{\{i\}} \equiv 1000101101000001001010100001100101 0$$

The variables may assume values 0 or 1, meaning that i-th neuron is respectively quiescent or active with probabilities

$$P(\eta^{\mu}_i = 1) = f \qquad P(\eta^{\mu}_i = 0) = 1 - f \tag{4.1}$$

where f is called *coding level*. Thus the average number of active neurons per memory is fN.

4.1.1 Synaptic dynamics and input statistics

As we have previously pointed out in section 3.2, the ability to store memories resides in the modification of the connections among pairs of neurons. When a certain pattern of activity is imposed to the network, it triggers a change in the synaptic strengths depending on the pre- and post-synaptic neurons status. Synaptic plasticity mechanisms are regulated by complex biochemical processes, for this reason synaptic modifications may occurr over many different time scales leading possibly to graded changes, but this aspect is still strongly debated.

Over longer time scales, however, the memory should be mantained even in the absence of the stimulus leading to a narrow number of stable synaptic states. In our model, to keep things simple, we have chosen to bound the synapses between two values, a depressed ($J = 0$) and a potentiated state ($J = 1$). Along the same lines of simplicity, we will make use of a convenient learning rule, which establishes that a synapse is potentiated with probability q_+ when both the pre- and the post-synaptic units are firing, is depressed with probability q_- when only one of them is activated, and stays unchanged when they are both quiescent:

$$\begin{aligned}
J_{ij}(t-1) = 0 \quad &\rightarrow \quad J_{ij}(t) = 1 \qquad \text{with probability} \quad q_+ \eta_i^t \eta_j^t \\
J_{ij}(t-1) = 1 \quad &\rightarrow \quad J_{ij}(t) = 0 \qquad \text{w. p.} \quad q_- [\eta_i^t (1-\eta_j^t) + (1-\eta_i^t)\eta_j^t] \\
&J_{ij}(t) = J_{ij}(t-1) \qquad\qquad \text{otherwise}
\end{aligned}$$

the first transition is identified as LTP, the second as LTD. Another interesting learning rule is the one proposed by Tsodyks and Feigelman (1988) for unbounded synapses and extended to bounded ones by Ben Dayan Rubin and Fusi (2007), which allows the suppression of the linear correlation term in the variance of the signal that we will shortly define. We preferred the former because the latter requires a finest

tuning when optimizing the network for better performance. Thus, according to our prescription, each synapse can shift stochastically between two discrete states, 0 and 1, as more and more patterns are shown to the network.

4.1.2 Signal to Noise analysis

To keep track of the changes provoked by each memory, we will look at the conditional probability distribution function for the synaptic strength $g_J^\mu(t)$, which conveys the probability that a synapse has value J at time t conditioned on the pattern of neuronal activity imposed by a generic memory μ. More formally, the *normalized signal* is defined as

$$\mathcal{S}_n(t) = g^\mu_{J=1}(t) = E\left[\frac{1}{f^2 N(N-1)} \sum_{i \neq j} J_{ij}(t)\eta_i^\mu \eta_j^\mu\right] \tag{4.2}$$

namely the overlap between the synaptic matrix and a given pattern μ at time t. The normalization keeps the signal between zero and one, detailed calculation are reported in the *Appendix A*.

Following the work of Amit and Fusi (1994), the procedure is summarized in the following steps:

1. we randomly generate a long sequence of uncorrelated memory patterns and impose it to the network. Each pattern is displayed only once in the sequence.

2. keeping constant the rate of stimuli presentation, we let the synapses reach the equilibrium distribution J_∞.

3. we choose one particular memory μ, which is not special in any sense, and show it to the network.

4. we track the memory trace by counting the number of synaptic modifications

originally produced by pattern μ that are still left in the system. In the context of binary neurons and bistable synapses, the signal defined in 4.2 is estimated by the conditional distribution of potentiated ($J = 1$) synapses:

$$g^{\mu}(t) = P(J_{ij}(t) = 1 | \eta_i^{\mu} = 1, \eta_j^{\mu} = 1)$$

This quantity measures how well the prototype pattern μ has been memorized by the network during the learning process. Distribution $g^{\mu}(t)$ critically depends on the state of the synaptic matrix, which fluctuates reflecting the statistics of a specific input sequence imposed to the network. As we do not consider the details of neural dynamics, looking at this observable does not exhaust in any sense the problem of memory retrieval. Instead, by restricting our attention to it, we can at least determine the necessary conditions for memory storage and retrieval, and, in the future, look at the sufficient ones including all underlying features.

In our protocol time is discrete, meaning that first pattern is presented at $t = 1$, second at $t = 2$, n-th pattern at $t = n$ and so on[1]. As a consequence the synaptic update process is a Markov chain in discrete time. Since, for any given synapse, it is always possible to leave one state with finite probability, i.e. there are not any absorbing states, the Markov chain is ergodic and irreducible (Cox and Miller, 1977), and the corresponding transition matrix M has one eigenvalue equal to one, leading to an equilibrium distribution after a sufficient number of presented memories, and a second eigenvalue related to the relaxation time[2]. The structure of matrix M is of the form

$$M(J(t+1)|J(t)) = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \quad \text{with eigenvalues} \begin{cases} \lambda_1 = 1 \\ \lambda_2 = 1-\alpha-\beta \end{cases} \tag{4.3}$$

[1]During the period separating two consecutive presentiations the dynamics is supposed to converge to the proper attractor.

[2]The amount of time taken by the system to reach the equilibrium state after a perturbation

In the case of random and uncorrelated patterns we have

$$M = \begin{pmatrix} 1-2f(1-f)q_- & 2f(1-f)q_- \\ f^2q_+ & 1-f^2q_+ \end{pmatrix} \quad \begin{cases} \lambda_1 = 1 \\ \lambda_2 = 1-2f(1-f)q_- - f^2q_+ \end{cases}$$

where $2f(1-f)q_-$ is the probability of finding a firing-quiescent pair of neurons times the synaptic depression probability, and f^2q_+ is the probability of finding a pair of active neurons times the synaptic potentiation probability. The dynamics of conditional distribution $g_J^\mu(t)$ may be written as

$$g_J^\mu(t) = \sum_K g_K^\mu(t-1)M_{KJ} = \sum_K g_K^\mu(0)(M^t)_{KJ} \tag{4.4}$$

where we can write the second equivalence only if the process is time homogeneous, i.e. M does not depend on time and on the particular sequence of presentation. This is always true when patterns are randomly chosen and uncorrelated. The spectral representation of M^t (if M has distinct eigenvalues) is

$$M^t = \mathsf{U}\ \Lambda^t\ \mathsf{V}$$

where Λ is the diagonal matrix of the eigenvalues λ_i $(i = 1, 2)$, columns u^i of matrix U are the right eigenvectors of M and rows v^i of V the left eigenvectors:

$$\mathsf{U} = \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \qquad \mathsf{V} = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \alpha \\ 1 & -1 \end{pmatrix}$$

Using this decomposition in (4.4) for potentiated synapses $(J = 1)$ we find

$$g_{J=1}^\mu(t) = \sum_{i=1}^{2} \lambda_i^t \ \sum_K g_K^\mu(0)\ \mathsf{u}_K^i \mathsf{v}_{J=1}^i = g_\infty^\mu\ (1-\lambda_2^t)\ +\ g^\mu(0)\ \lambda_2^t \tag{4.5}$$

where we have defined the steady state for the distribution of potentiated synaptic efficacies as:

$$g^{\mu}_{\infty} \equiv \frac{\beta}{\alpha + \beta} \tag{4.6}$$

It is clear, indeed, that any distribution converges to g^{μ}_{∞} in a finite time, leaving no trace of initial conditions $g^{\mu}(0)$. Eigenvalue λ_2 rallies the decay towards the asymptotic state g^{μ}_{∞}. If memory pattern μ is presented at time $t = 1$ and the network has reached the equilibrium distribution, we have that the initial value of $g^{\mu}_{J=1}$ is given by the conditional probability that a synapse was already potentiated, plus the distribution representing those that were previously depressed but now are potentiated with probability q_+:

$$g^{\mu}(0) = g^{\mu}_{\infty} + (1 - g^{\mu}_{\infty})\ q_+$$

After the presentation of t patterns the synaptic matrix has undergone many changes that must have weakened the original trace created at $t = 0$. In fact, synaptic modifications induced by the imposition of a new pattern may regard some of the synapses encoding our tracked memory, causing the progessive erasure of its trace. Indeed, time evolution of the conditional distribution g^{μ} follows equation (4.5):

$$g^{\mu}(t) = g^{\mu}_{\infty}\ +\ (1 - g^{\mu}_{\infty})\ q_+ \lambda_2^t \tag{4.7}$$

For $t \to \infty$ the breakdown is exponential, in fact

$$\lambda_2^t = (1 - \alpha - \beta)^t \xrightarrow[t\to\infty]{} e^{-(\alpha+\beta)t}$$

$$g^{\mu}_J(t) = g^{\mu}_{\infty}\ +\ (1 - g^{\mu}_{\infty})\ q_+ e^{-(\alpha+\beta)t}$$

from which we see that the network relaxes to the equilibrium state with a time constant $\tau = (\alpha+\beta)^{-1}$. How is the decay constant related to memory lifetime? How many presentations does it take for the memory trace to be definitely lost? Quite obviously, studying the signal is not sufficient if we want to answer these questions.

Randomness in the order of presentation and uncertainty in the synaptic plasticity rule cause the presence of stochastic fluctuations in the network. A particular state of the system, apparently suggesting that the μ-th memory trace is still imprinted in the synaptic efficacies, might instead be the result of a random deviation from the equilibrium state. This happens because patterns are not mutually orthogonal, so presenting a certain memory affects synapses that carry other memories' signal. We have already pointed out that our network does not take into account any underlying neuronal dynamics, nonetheless we should not forget that the Hopfield model treats memories as attractors of the neural dynamics. In this picture, two non orthogonal memories are represented by attractors whose basins of attraction are ovelapping: when the dynamics of the system falls in the shared portion, both basins are likely to be reached. The network might then retrieve the wrong pattern, although it has been showed the correct stimulus. This is due to the *noise*: random fluctuations of the dynamics, coming from inside the synaptic matrix (non orthogonal patterns) and from the variability in the number of active neurons per memory. We formalize this concept by interpreting the noise as the square root of the variance:

$$\mathcal{N}(t) = \sqrt{Var\left[\sum_{i \neq j} J_{ij}(t)\eta_i\eta_j\right]}$$

In fact each memory pattern has fN nonzero bits on average but, since we generate memories according to (4.1), the variance of the coding level about its mean is $f(1-f)N$. This variability in the coding level leads to dramatic consequences in the noise level, as we may see later.

Before proceeding any further, let us ask a question: which is the fraction of neurons that are activated, on average, by an upcoming stimulus in the real brain? This question is still open and debated at all levels, and no exhaustive answer has been

provided yet. Three possible strategies for coding are discussed in literature: *dense coding*, *local coding* and *sparse coding*. In a dense distributed code each pattern is encrypted by almost all neurons present in the network, which produces a massively redoundant representation that makes the newtork resistent to possible errors and single units faults. On the contrary, dense coding implies a high probability of interference amid different patterns which makes the decoding of the output particularly difficult, since interfering memories would be not linearly separable. The opposite limit corresponds to local coding, where each neuron, or small group of neurons, represents a whole stimulus. Although such strategy avoids any interference or correlation among different patterns, the capacity would be linearly proportional to the number of units (neurons or small aggregations of neurons) encoding each single memory, and it is therefore much smaller than in the dense coding representation. Moreover, any correlation or association present in the upcoming input sequence is disregarded even when it would be beneficial for generalization purposes. Sparse codes combine advantages of local and dense codes while averting most of their drawbacks. Increasing the number of active units introduces some overlap in the patterns representation, but the network can, nonetheless, mantain a quite high representational capacity. However, we have pointed out in Section 3.3.2 that the number of input-output pairs that can be stored in an associative memory is far greater for sparse than for dense patterns (Meunier and Nadal, 1995), because of the reduced amount of information contained in the representation of any stored pattern. As a much larger fraction of all input-output functions are linearly separable using sparse coding, decoding becomes easier and less complex. In addition, since generalisation takes place only between overlapping patterns, new associations will not interfere with previous associations to nonoverlapping patterns.

In *Appendix A* the noise has been calculated explicitly in the sparse coding limit,

i.e. $f \to 0$. Nevertheless, this assumption is supported by several experimental results in various areas of the brain (see Section 3.3.2). Thus, in the limit of high sparseness and large network ($f \to 0$, $N \to \infty$, $fN \to \infty$), we have from *Appendix A*:

$$\mathcal{N}(t) = 2(fN)^{3/2}\sqrt{\gamma^{\mu}(t) - fg^{\mu}(t)^2}$$

where $\gamma^{\mu}(t) \equiv P(J_{ij} = 1, J_{il} = 1|\eta_i^{\mu}\eta_j^{\mu}\eta_l^{\mu} = 1)$ is the conditional distribution probability for a pair of correlated synapses, i.e. sharing one pre- or post-synaptic neuron. The second term within brackets is just the square of the conditional distribution encountered in the signal times the coding level f. Both are probability distribution, so all the dependency on the network size is contained in the factor $N^{3/2}$.
For large t all time dependent terms vanish, thus we can rewrite the noise as

$$\mathcal{N}(t) = 2(fN)^{3/2}\sqrt{\gamma_{\infty}^{\mu} - f(g_{\infty}^{\mu})^2}$$

Now that we have obtained the expression for the noise, we are about to compare it to the signal in order to check the attractor stability ad robustness, as it is suggested in Amit and Fusi (1994). This will provide us with an upper bound on the storage capacity of the network. To do this, we slightly modify the signal as it has been introduced in (4.2): we remove the normalization factor and we subtract out its asymptotic value, as we assume that a memory is certainly forgotten when its signal is indistinguishable from the spontaneous activity of the network. Thus we have

$$\mathcal{S}(t) = E\left[\sum_{i \neq j} J_{ij}(t)\eta_i^{\mu}\eta_j^{\mu}\right] - S_{\infty}$$

If memory μ is imposed at $t = 0$, the initial signal, in analogy with (4.7), is given by

$$S(0) = f^2N(N-1)\ (1 - g_{\infty}^{\mu})\ q_{+}$$

and after the presentation of t uncorrelated memories the original signal has decreased to

$$S(t) = f^2 N(N-1)\ g^{\mu}(t) \ = \ f^2 N(N-1)\ (1 - g^{\mu}_{\infty})\ q_{+} e^{-(\alpha+\beta)t}$$

for a large network ($N \to \infty$ and the product $fN \to \infty$) we can ignore the linear term in N, finally getting:

$$S(t) = f^2 N^2 (1 - g^{\mu}_{\infty})\ q_{+} \exp\Big[- (\alpha + \beta)\ t \Big]$$

Hence, the *signal-to-noise ratio* (SNR) scales as:

$$\frac{\mathcal{S}(t)}{\mathcal{N}(t)} = \sqrt{fN}\ \ \frac{(1 - g^{\mu}_{\infty})\ q_{+}}{2\ \sqrt{\gamma^{\mu}_{\infty} - f(g^{\mu}_{\infty})^2}}\ \exp\Big[- (\alpha + \beta)\ t \Big] \approx \sqrt{fN} \exp(-t/\tau) \qquad (4.8)$$

If we establish a threshold θ such that when the SNR value lies below θ the memory is lost, we see that the maximum memory lifetime is limited by:

$$t_{max} < \Big(\frac{1}{2} \log(fN) - \log(\theta) \Big) \tau \qquad (4.9)$$

θ is chosen according to the desired error tolerance: the higher is the threshold in the SNR, the lesser the mistakes in pattern retrieval process. Refer to Weisbuch and Fogelman-Soulié (1985) for a deeper perspective and detailed calculations.

We consider memory lifetime and storage capacity of the network to be equivalent. In fact, we argue that if a memory can be retrieved up to a certain time t_{max} after its first appearence, then every memory that has been presented during this interval could be recalled as well. The maximum storage capacity p_{max} grows, therefore, linearly with decay factor τ, but is only proportional to the logarithm of N. In the present case of random and uncorrelated patterns, the decay constant is

$$\tau = (\alpha + \beta)^{-1} = (2f(1-f)q_- + f^2 q_+)^{-1} \approx \mathcal{O}(f^{-1})$$

Memory lifetime may therefore be optimized by adjusting q_+ and q_- to balance the depressing and potentiating synaptic transitions:

$$q_+ = q \qquad\qquad q_- = \frac{qf}{2(1-f)}$$

Theferore the equilibrium distribution and the decay factor are rewritten as

$$g_\infty^\mu = \frac{\beta}{\alpha+\beta} = \frac{q_+ f^2}{2q_- f(1-f) + q_+ f^2} = \frac{1}{2} \qquad\qquad \tau = (2f^2 q)^{-1} \approx \mathcal{O}(f^{-2}) \quad (4.10)$$

This result, first obtained in Amit and Fusi (1994), states that memory lifetime depends quadratically on the inverse of the coding size f and linearly on potentiation probability q_+. Hence, in principle we could achieve longer memory lifetimes in two ways:

1. *slow learning*: within our framework old memories are erased progressively when new memories are learned, a kind of memory that has been called *palimpsest* (Nadal et al., 1986; Parisi, 1986). The learning (and forgetting) speed is poportional to the probability that a synapse changes its state consequently to the impinging memory, q. Learning can be made slow by lowering q, thereby helping memory preservation by reducing the rate of changes in the synaptic matrix.

2. *high sparseness*: only a fraction f of the neurons is active for each memory. Making memory sparse, i.e. adopting a small f, decreases the interference between different patterns, this is to say that the number of active neurons common to two or more patterns is scarce. Consequently the probability that a new memory would alter synapses that were potentiated by older memories is

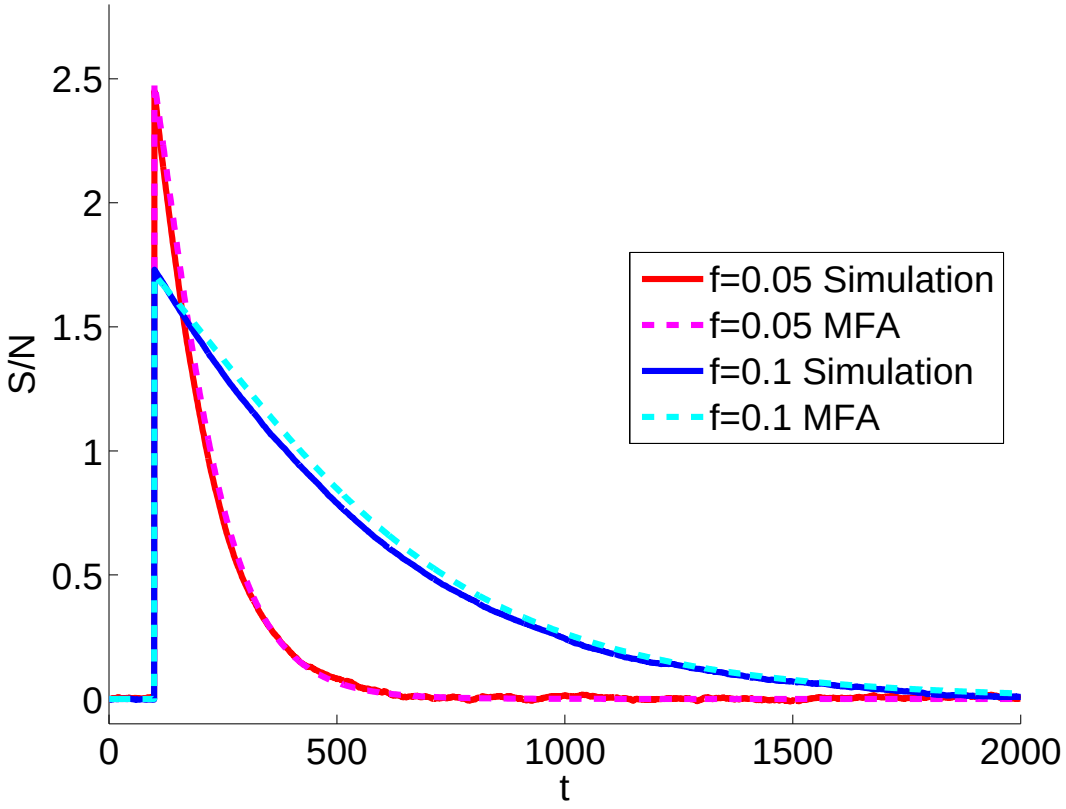


Figure 4.2: Signal-to-Noise ratio for random and uncorrelated memory patterns. Simulations are represented by the continous line, mean field approximation by the dotted line. In this example: $N = 1000$, $q = 0.5$, Trials$= 100$. Red line $f = 0.1$. Blue line $f = 0.05$.

low, and older memory are less likely to be erased when new ones are imposed.

Obviously, extending the storage capacity must have a cost in terms of amount of information initially stored in the network for each individual pattern. Slowing down learning would, in fact, provoke a linear loss in the initial SNR, and concomitantly a hyperbolic growth in the memory lifetime (for $\tau \approx q^{-1}$). To increase sparseness is more profitable: the linear contraction in the initial SNR is compensated by a quadratic gain in memory lifetime. Nevertheless sparseness cannot be increased ad infinitum, since the SNR needs to lay above threshold θ for the memory to be retrieved. Indeed, if the SNR is below threshold right after the presentation of the tracked memory, the pattern cannot be stored at all and its lifetime is zero regardless

of eq. (4.10). For this reason, Ben Dayan Rubin and Fusi (2007) impose a lower bound on f, limiting the sparseness but allowing memories to be hold back and then retrieved.

4.2 Correlated memories

In Section 3.4 we have drawn attention on the fact that human memories are always related to each other, giving rise to some complex hierarchical structure. To model these structures we make use of ultrametric trees as presented in Section 3.4.1, where patterns represent nodes, and the distance between two vectors is given by eq. (3.9).

4.2.1 Two generations hierarchy

The ultrametric architecture introduced in Section 3.4.1 is organized in *classes* that we will also label as *families* or *clusters*. Each class is originated from a prototypical pattern η^μ (where now index μ labels both the prototype and the class) as in the random and uncorrelated case. This ancestor memory collects the typical features of that particular class. The class members (*sons*) are then generated by corrupting the ancestor memory (or *father*) in a stochastic fashion (see Fig. 4.3). This way, the father shares with each son a conspicuous number of features, obviously bigger than in the uncorrelated protocol. Each son $\eta^{\mu\nu}$ is therefore a noisy version of ancestor μ. This is accomplished by reversing the activity of some neurons in the father representation as discussed in Section (4.1). The conditional probabilities, fixed for all classes and all sons, are

$$P(\eta_i^{\mu\nu} = 1 | \eta_i^\mu = 1) \equiv u \tag{4.11}$$

$$P(\eta_i^{\mu\nu} = 0 | \eta_i^\mu = 0) \equiv v \tag{4.12}$$

In principle the number of flipping bits per son may be arbitrary, leading to a variety of different coding levels across different sons and classes. For the sake of simplicity we preferred to keep f constant, meaning that each class member presents, on average, the same number of features of the prototype, some of which are common to both. The coding level of a son is

$$f^* = fu + (1-f)(1-v) \qquad \text{and must hold} \quad f^* \equiv f$$

Probabilities u and v will therefore depend on f and on a parameter expressing the grade of similarity between the sons and the father:

$$\begin{aligned} u &= 1-(1-f)(1-m) \\ v &= 1-f(1-m) \end{aligned}$$

$m = 0$ corresponds to the random patterns example, and hence no classes are formed, while the case $m = 1$ means that all the patterns that belong to one family are identical to their ancestor. By choosing a set of probabilities that preserves the average coding level f for all patterns, not only we simplify some of the following calculations, but we also avoid a generation-dependent coding that could lead to the trivial $f = 0$ or $f = 1$ cases as the number of generations grows.

Now we focus our attention on the dynamics of the synaptic matrix upon the presentation of a sequence of stimuli. It is very important, as a first step, to clarify the kind of sequence we aim to model: our effort is to describe a very simple scenario where inputs are correlated in a way which is quite far from the natural world, since, at this level of depth, we are only interested in the effect of correlation on storage capacity of these stochastic networks. It seems reasonable to us if we choose, at each time step, one father at random from one of the p uncorrelated classes, and

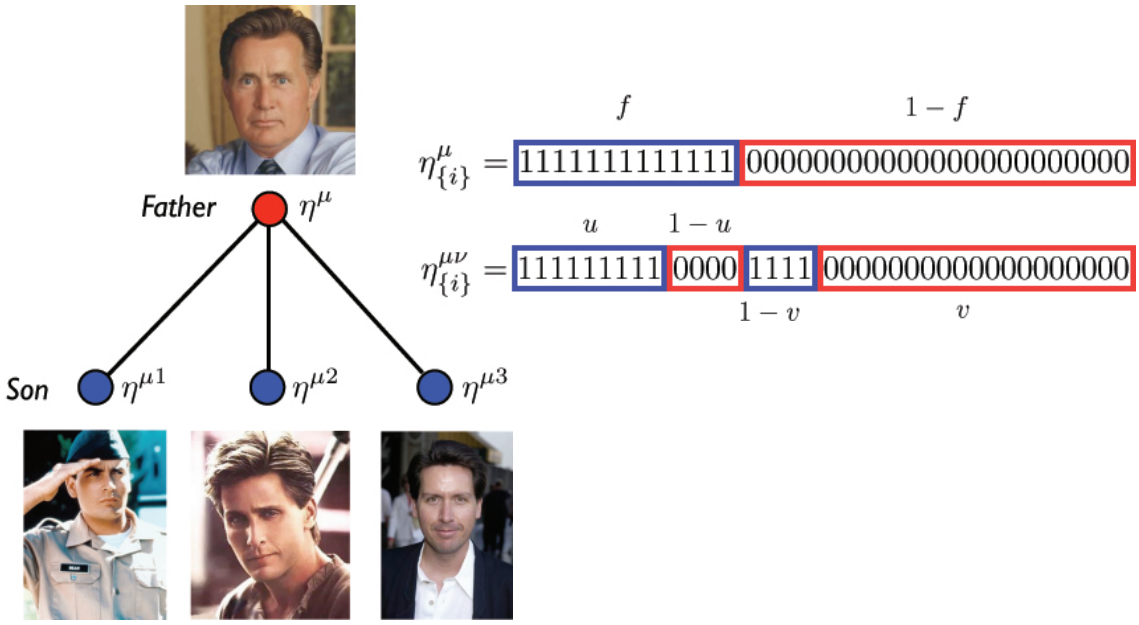


Figure 4.3: Two generations hierarchy of memories: starting from father $\eta^{\mu}_{\{i\}}$ a second layer of p sons is produced. When generating each son a fraction $(1-u)$ of father's active neurons are switched off, while a fraction $(1-v)$ of father's silent units are flipped.

then use it to generate a noisy pattern, i.e. a son, which is then imposed to the network. This way we construct a fixed number of categories p and a virtually infinite number of examples belonging to each category. The prototype never appears in the input sequence, nevertheless its memory trace can be found in the network when correlation with class members (the effective input) is high enough. A generic stream, as seen by the network, is then of the form:

$$\boxed{\eta^{\mu 1}, \eta^{\nu 1}, \eta^{\kappa 1}, ..., \eta^{p1}}\ ;\boxed{\eta^{\mu 2}, \eta^{\nu 2}, \eta^{\kappa 2}, ..., \eta^{p2}}\ ; ... \tag{4.13}$$

where the boxes delimit each sequence of inputs. To measure the correlation between the synaptic architecture and the prototype patterns during learning process we use the overlap introduced in Section 4.1.2:

$$g^{\mu}(t) = \frac{1}{f^2 N(N-1)} \sum_{i \neq j} J_{ij}(t) \eta_i^{\mu} \eta_j^{\mu} \tag{4.14}$$

The same observable may be extended at patterns standing at lower layers in the hierarchy. As we move down along the ultrametric tree we need to use more indexes to identify the memories and the classes or subclasses they belong to. In the two generations example, in order to check the correlation with the synaptic matrix of the ν-th pattern generated from μ-th prototype, we use:

$$g^{\mu\nu}(t) = \frac{1}{f^2 N(N-1)} \sum_{i \neq j} J_{ij}(t) \eta_i^{\mu\nu} \eta_j^{\mu\nu} \;=\; P(J_{ij}(t) = 1 | \eta_i^{\mu\nu} = 1, \eta_j^{\mu\nu} = 1) \quad (4.15)$$

and so on for any other memory. In general any synaptic configuration arises from the specific temporal arrangement of memories previously imposed to the network. If, for example, the network sees a pattern belonging to $\mu - th$ class, memory traces of class members and, of course, of the prototype will get stronger, while traces of other classes memories will not. Any attempt to approximate and model the dynamic behaviour of such system is therefore destined to suffer from some grade of inaccuracy.

But since we are interested in the mean properties and response of the system, to avoid this issue we adopt the point of view expressed in Brunel et al. (1998), who noted that, in the slow learning limit ($q_+ = 0$), averaging over all possible realizations of a sequence gives a good approximation of the synaptic matrix's behaviour when presented with a typical succession of patterns. Variability from sequence to sequence goes to zero when $q_+ \to 0$. Each sequence in (4.13) comprises p presentations of patterns chosen by picking out a prototype at random and then generating a noisy version of it. One class is represented, on average, only once per sequence. This convention is similar to the protocol used in visual memory experiments by Miyashita (1988). Still, as in the uncorrelated context previously described, any presentation causes a pool of stochastic transitions in the synaptic efficacies. The subsequent markovian stochastic process for a single synaptic efficacy J_{ij} is depicted

by a matrix similar to (4.3), where elements are the transition probabilities averaged over all possible sequences:

$$\begin{aligned}
\langle M_{ij}\rangle &= \begin{pmatrix} 1-\langle\alpha_{ij}\rangle & \langle\alpha_{ij}\rangle \\ \langle\beta_{ij}\rangle & 1-\langle\beta_{ij}\rangle \end{pmatrix} \\
\langle\alpha_{ij}\rangle &= \frac{q_-}{p}\sum_{\mu=1}^{p}[\eta_i^{\mu\nu}(1-\eta_j^{\mu\nu})+\eta_j^{\mu\nu}(1-\eta_i^{\mu\nu})] \\
\langle\beta_{ij}\rangle &= \frac{q_+}{p}\sum_{\mu=1}^{p}\eta_i^{\mu\nu}\eta_j^{\mu\nu}
\end{aligned}$$

Recalling eq. (4.6), we deduce that the expression for the steady state distribution is simply

$$g_\infty^\mu = \frac{1}{f^2N(N-1)}\sum_{i\neq j}\frac{\langle\beta_{ij}\rangle}{\langle\alpha_{ij}\rangle+\langle\beta_{ij}\rangle}$$

defining

$$\begin{aligned}
P_{ij} &\equiv \sum_{\mu=1}^{p}\eta_i^{\mu\nu}\eta_j^{\mu\nu} \\
D_{ij} &\equiv \sum_{\mu=1}^{p}[\eta_i^{\mu\nu}(1-\eta_j^{\mu\nu})+\eta_j^{\mu\nu}(1-\eta_i^{\mu\nu})]
\end{aligned}$$

we obtain for the asymptotic distribution

$$g_\infty^\mu = \frac{1}{f^2N(N-1)}\sum_{i\neq j}\frac{q_+P_{ij}}{q_+P_{ij}+q_-D_{ij}} \tag{4.16}$$

Now that we have averaged the efficacy of each single synapse over all possible sequences, we can carry out the *mean field approximation* (Parisi, 1988) by replacing the elements of the sum over all synapses by their average, denoted by double

brackets $\langle\langle\cdot\rangle\rangle$:

$$g_\infty^\mu = \left\langle\left\langle \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}} \right\rangle\right\rangle$$

A generic input sequence consists, on average, in one pattern belonging to the desired class (μ) and $(p-1)$ patterns residing in other families. In the former case, synapses will connect a pair of active neurons with probability u, and will therefore experience potentiation with probability $q_+ u$. Similarly depression will occurr with probability $q_- 2u(1-u)$, i.e. the chance of finding a firing neuron connected to a silent one times the synaptic depression probability.

Instead, since there is no correlation among classes, when we show a pattern that belongs to a different family we have that:

- if a synapse belongs to the particular subset in the $\{i,j\}$ space where $\eta_i^\mu \eta_j^\mu = 1$ (which happens with probability f^2), then it will be potentiated with probability $q_+ u^2$ or depressed with probability $q_- 2u(1-u)$;

- instead, if it belongs to the subset $\eta_i^\mu(1-\eta_j^\mu) + \eta_j^\mu(1-\eta_i^\mu) = 1$, that comprises, on average, a fraction $2f(1-f)$ of all the synapses, it will be potentiated with probability $q_+ u(1-v)$ and depressed with probability $q_-[uv + (1-u)(1-v)]$

- if a synapse belongs to the last subset, $\eta_i^\mu = 0, \eta_j^\mu = 0$ (with probability $(1-f)^2$), it ends up being potentiated with probability $q_+(1-v)^2$ or depressed with probability $q_- 2v(1-v)$;

Putting all different subsets together and adding the contribution of the intra-class

memory, we can rewrite (4.16) in the form:

$$
\begin{aligned}
g_\infty^\mu &= \frac{1}{f^2 N(N-1)} \sum_{i \neq j} \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}} \; \eta_i^\mu \eta_j^\mu = \\
&= \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p-1-\Pi} \psi_+(\Pi, \Delta) \; q_+ \Big(u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta) \Big) \times \\
&\times \Big(q_- [2u(1-u)(\Pi+1) + (uv + (1-u)(1-v))\Delta + 2v(1-v)(p-1-\Pi-\Delta)] \\
&+ q_+ [u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta)] \Big)^{-1} = \\
&= \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p-1-\Pi} \psi_+(\Pi, \Delta) \; \mathcal{A}_\infty^\mu(\Pi, \Delta)
\end{aligned}
$$

the sum is over $(p-1)$ patterns reflecting the fact that the network sees a member of the μ-th class with probability $1/p$, meaning that the pool of synapses matching the condition $\eta_i^\mu \eta_j^\mu = 1$ is potentiated at least once in the sequence with probability $q_+ u^2$ and depressed with probability $q_- 2u(1-u)$. The trinomial probability distribution

$$
\psi_+(\Pi, \Delta) = \frac{(p-1)!}{\Pi! \Delta! (p-1-\Pi-\Delta)!} [f^2]^\Pi \; [2f(1-f)]^\Delta \; [(1-f)^2]^{(p-1-\Pi-\Delta)}
$$

gives the joint probability $P(P_{ij} = \Pi, D_{ij} = \Delta)$.

If we include the full time dependency, the distribution may be rewritten as

$$
g^\mu(t) = \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p-1-\Pi} \psi_+(\Pi, \Delta) \; \Big(\mathcal{A}_\infty^\mu(\Pi, \Delta) \big(1 - \langle \lambda^\mu(\Pi, \Delta) \rangle^t \big) + c \langle \lambda^\mu(\Pi, \Delta) \rangle^t \Big)
$$

and the subdominant eigenvalue as

$$
\begin{aligned}
\langle \lambda^\mu(\Pi, \Delta) \rangle &= 1 - \frac{q_+}{p} \Big(u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta) \Big) + \\
&- \frac{q_-}{p} \Big(2u(1-u)(\Pi+1) + [uv + (1-u)(1-v)]\Delta + 2v(1-v)(p-1-\Pi-\Delta) \Big)
\end{aligned}
$$

where $\langle \lambda^{\mu}(\Pi, \Delta) \rangle$ is the sequence-averaged dominating eigenvalue of the Markov chain, and c is the initial distribution of potentiated synapses.

Things get more complicated when we look at the son, since nontrivial correlations between class members cause the formation of three different pools of synapses $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ that experience potentiation and depression with peculiar statistics:

$$g_{\infty}^{\mu\nu} = \frac{1}{f^2 N(N-1)} \sum_{i \neq j} \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}} \, \eta_i^{\mu\nu} \eta_j^{\mu\nu}$$
$$= \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p-1-\Pi} \psi_+(\Pi, \Delta) \Big(\mathcal{B}_{\infty}^{\mu\nu}(\Pi, \Delta) + \mathcal{C}_{\infty}^{\mu\nu}(\Pi, \Delta) + \mathcal{D}_{\infty}^{\mu\nu}(\Pi, \Delta) \Big)$$

where:

$$\mathcal{B}_{\infty}^{\mu\nu}(\Pi, \Delta) = u^2 \; q_+ \Big(u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta) \Big) \times$$
$$\times \Big(q_- [2u(1-u)(\Pi+1) + (uv + (1-u)(1-v))\Delta + 2v(1-v)(p-1-\Pi-\Delta)]$$
$$+ q_+ [u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta)] \Big)^{-1}$$

$$\mathcal{C}_{\infty}^{\mu\nu}(\Pi, \Delta) = \frac{2f(1-f)u(1-v)}{f^2} \, q_+ \Big(u^2\Pi + u(1-v)(\Delta+1) + (1-v)^2(p-1-\Pi-\Delta) \Big) \times$$
$$\times \Big(q_- [2u(1-u)\Pi + (uv + (1-u)(1-v))(\Delta+1) + 2v(1-v)(p-1-\Pi-\Delta)]$$
$$+ q_+ [u^2\Pi + u(1-v)(\Delta+1) + (1-v)^2(p-1-\Pi-\Delta)] \Big)^{-1}$$

$$\mathcal{D}_{\infty}^{\mu\nu}(\Pi, \Delta) = \frac{(1-f)^2(1-v)^2}{f^2} \, q_+ \Big(u^2\Pi + u(1-v)\Delta + (1-v)^2(p-\Pi-\Delta) \Big) \times$$
$$\times \Big(q_- [2u(1-u)\Pi + (uv + (1-u)(1-v))\Delta + 2v(1-v)(p-\Pi-\Delta)]$$
$$+ q_+ [u^2\Pi + u(1-v)\Delta + (1-v)^2(p-\Pi-\Delta)] \Big)^{-1}$$

As a function of time, starting from arbitrary initial conditions:

$$g^{\mu\nu}(t) = \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p-1-\Pi} \psi_+(\Pi,\Delta) \Big(\mathcal{B}^{\mu}_{\infty}(\Pi,\Delta)\big(1 - \langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{B}}\big) + \\ + \mathcal{C}^{\mu}_{\infty}(\Pi,\Delta)\big(1 - \langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{C}}\big) + \mathcal{D}^{\mu}_{\infty}(\Pi,\Delta)\big(1 - \langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{D}}\big) + \\ + \frac{c}{f^2}\big(f^2u^2\langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{B}} + 2f(1-f)u(1-v)\langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{C}} + \\ + (1-f)^2(1-v)^2\langle \lambda^{\mu}(\Pi,\Delta)\rangle^{t}_{\mathcal{D}}\big)\Big)$$

where each set of synapses is associated with a peculiar eigenvalue:

$$\langle \lambda^{\mu\nu}(\Pi,\Delta)\rangle_{\mathcal{B}} = 1 - \frac{q_+}{p}\Big(u^2(\Pi+1) + u(1-v)\Delta + (1-v)^2(p-1-\Pi-\Delta)\Big) + \\ - \frac{q_-}{p}\Big(2u(1-u)(\Pi+1) + [uv + (1-u)(1-v)]\Delta + 2v(1-v)(p-1-\Pi-\Delta)\Big)$$

$$\langle \lambda^{\mu\nu}(\Pi,\Delta)\rangle_{\mathcal{C}} = 1 - \frac{q_+}{p}\Big(u^2\Pi + u(1-v)(\Delta+1) + (1-v)^2(p-1-\Pi-\Delta)\Big) + \\ - \frac{q_-}{p}\Big(2u(1-u)\Pi + [uv + (1-u)(1-v)](\Delta+1) + 2v(1-v)(p-1-\Pi-\Delta)\Big)$$

$$\langle \lambda^{\mu\nu}(\Pi,\Delta)\rangle_{\mathcal{D}} = 1 - \frac{q_+}{p}\Big(u^2\Pi + u(1-v)\Delta + (1-v)^2(p-\Pi-\Delta)\Big) + \\ - \frac{q_-}{p}\Big(2u(1-u)\Pi + [uv + (1-u)(1-v)]\Delta + 2v(1-v)(p-\Pi-\Delta)\Big)$$

The formulas we have just derived are undoubtedly difficult to calculate, since the amount of terms inside each sum grows exponentially with the total number of patterns, and hard to read intuitively, for they are a non linear miscellany of several parameters. The exact solution can still be handled by a calculator if the number of classes is not high, namely less than 500. The parameters involved are:

- coding level f, that is the fraction of neurons activated on average by each

memory input.

- pattern similarity m, that controls correlations between layers inside the pattern hierarchy.

- synaptic transition probabiities. We use the same tuning as in the random case, to force the network towards an overall balance between number of potentiating and depressing events: $q_+ = q \qquad q_- = \dfrac{qf}{2(1-f)}$.

- total number of classes p, determining the richness of the external input received by the network.

Is there a set of parameters that allows the system to learn a prototype better than any of the class members? Note that we never show a prototype to the network, which means that learning must take place exclusively through correlations with class members.
To address this question we let the network reach the equilibrium distribution (i.e. when it has seen a sufficient number of patterns belonging to our hierarchy, according to the statistics previously introduced in this section), then, at time t^*, we impose a generic pattern belonging to class μ. Obviously, the son with the highest signal is the one that has been just presented, so we will look at the conditional distributions

$$g^{\mu}(t^*) = g^{\mu}_{\infty}\big(1 - q_- 2u(1-u)\big) + (1 - g^{\mu}_{\infty})q_+ u^2$$
$$g^{\mu\nu}(t^*) = g^{\mu\nu}_{\infty} + (1 - g^{\mu}_{\infty})q_+$$

to see if, for a given value of m and q

$$g^{\mu}(t^*) > g^{\mu\nu}(t^*)$$

In Fig. 4.4 we show the results of simulations in three configurations of parameters

m and q, while we keep fixed $f = 0.05$ and $p = 100$. Simulations are carried out in a network of $N = 1000$ binary neurons with no intrinsic dynamics, and are obtained averaging over 50 trials. In Fig. 4.4A is displayed what we have called the *poorly correlated sons regime*, where low correlations ($m = 0.1$), irrespective of learning speed, prevent the prototype from being effectually stored. The *strongly correlated*

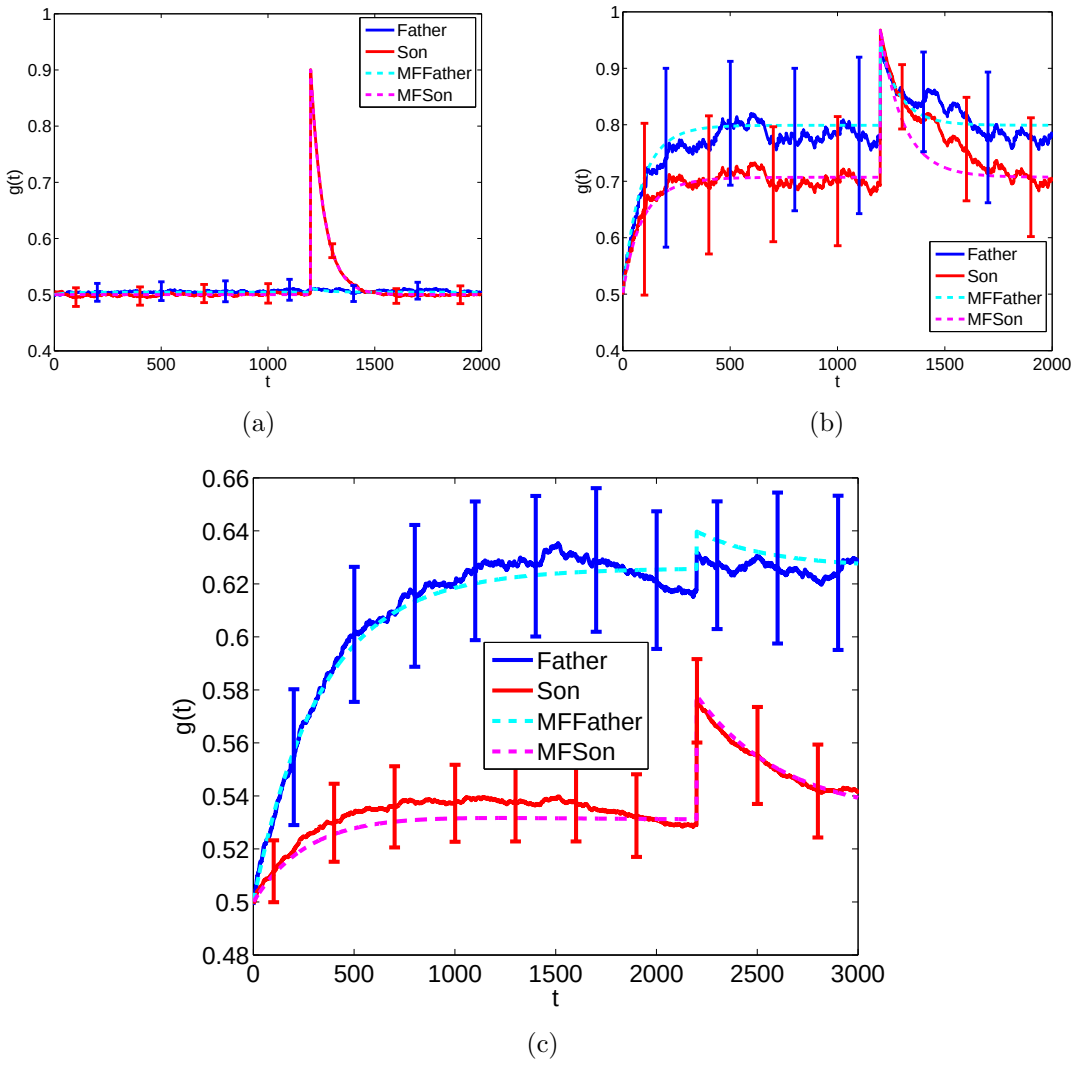
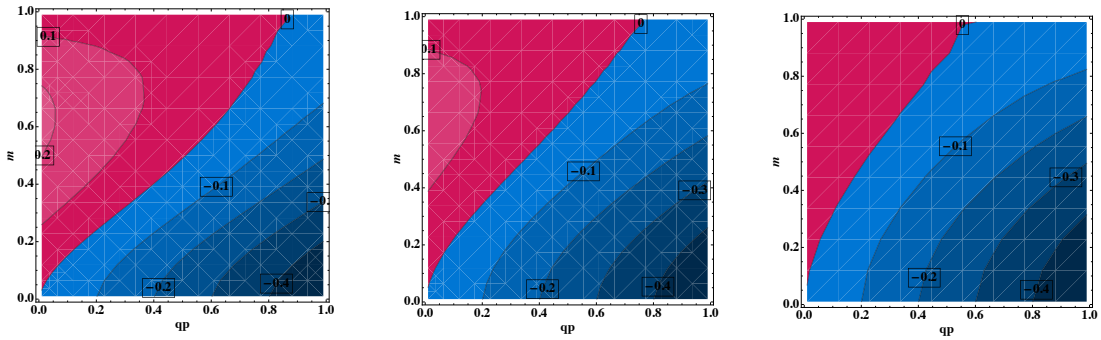


Figure 4.4: (a) Poorly Correlated Sons regime: correlations are too scarce to create a representation of the father; B. Strongly Correlated Sons regime: memory traces are strong and close together because of the high correlations, class members prevail at the time of their presentation; C. Dominating Father regime: the father trace is stronger than sons' signal when correlations and slow learning occur.



(a) p=50 (b) p=100 (c) p=200

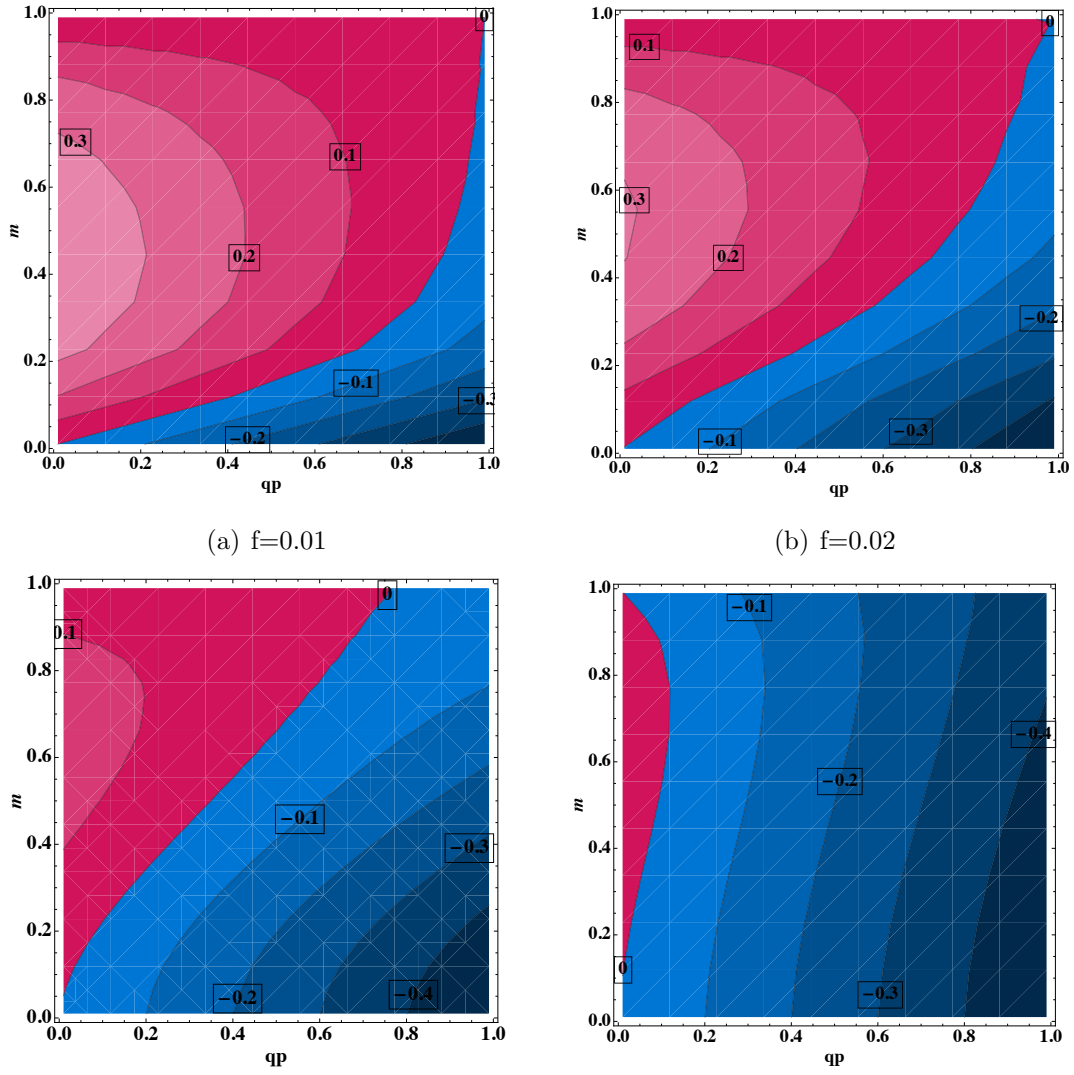
Figure 4.5: $g^{\mu}(t^*) - g^{\mu\nu}(t^*)$ as a function of two parameters, similarity m and synaptic transition probability q, for different number of classes: (a) $p = 50$; (b) $p = 100$; (c) $p = 200$. As p grows the red area, corresponding to the dominating father regime, becomes smaller, reflecting the higher probability of overlap beetween patterns belonging to different classes. In all figures $f = 0.05$.

sons regime (see Fig. 4.4B) represents a situation where similarity amid patterns is high enough and the learning rate is extremely fast ($m = 0.7$, $q = 0.9$); the son's memory trace stays much above his father's trace at the very moment of his presentation, even though it is asymptotically overcome by the prototype signal. The last and most interesting regime is shown in Fig. 4.4C: here patterns are quite correlated but learning speed is low ($m = 0.6$, $q = 0.2$), so that any imposed pattern is learned less effectively at the moment of the presentation, while the features of the prototypical pattern can still be stored. The agreement between simulations and mean field approximation is excellent, considering the relatively small size of our network. Extensive simulations have been performed with parameters f, m, and q ranging from 0 to 1, while the number of families varied between 10 and 500. In Fig. 4.5 and Fig. 4.6 we the two dimensional diagrams reporting the value of the difference $g^{\mu}(t^*) - g^{\mu\nu}(t^*)$ as a function of similarity and transition probability. For a given statistics in the input (i.e. a certain value of m), there could exist different networks with distinct synaptic transition probabilities that would allow the storage of the father or the sons respectively. The dominating father regime progressively

shrinks, eventually collapsing towards the $q = 0$ line, when p and f grow. In fact, a higher coding level translates in a higher overlap between uncorrelated patterns (i.e. belonging to different classes). The positive effect due to correlations is then counterbalanced by the higher average noise provoked by the learning of memories from other families. Similarly, breeding a higher number of families not only causes an increase in overlap probability but also extends the average time separating the presentations of two patterns from the same family.

4.2.2 Storing the difference

In the last section we have found a subset in the parameters' space for which the memory trace of a prototype prevails against every class member. A first implication of this result is that the network is able to generalize the representation of any group of correlated inputs when properly tuned. At the end of the learning process, any "noisy" pattern shown to the network elicits a stronger signal in the prototype representation than in its own. A second very important implication would be that, if we have a network capable of storing the prototype, we can imagine a second network, differently tuned, capable of storing the *differences* between the incoming queue of inputs (i.e., the class members) and the prototype. In other words we build a first network that keeps track of the features common to some groups of patterns, and then some connecting scheme that receives the actual input and matches it with the activity of the first network, keeping only the unmatching portion of each input pattern (i.e. the "noisy" part). A second network receives these differences in input and stores them as memories (see Fig. 4.7). Why storing differences instead of the whole original memories? We have seen in the random and uncorrelated case that the total storage capacity is proportional to f^{-2}. Reducing the coding size is therefore crucial for extending memory lifetime, and this is what happens if we store



(a) f=0.01 (b) f=0.02

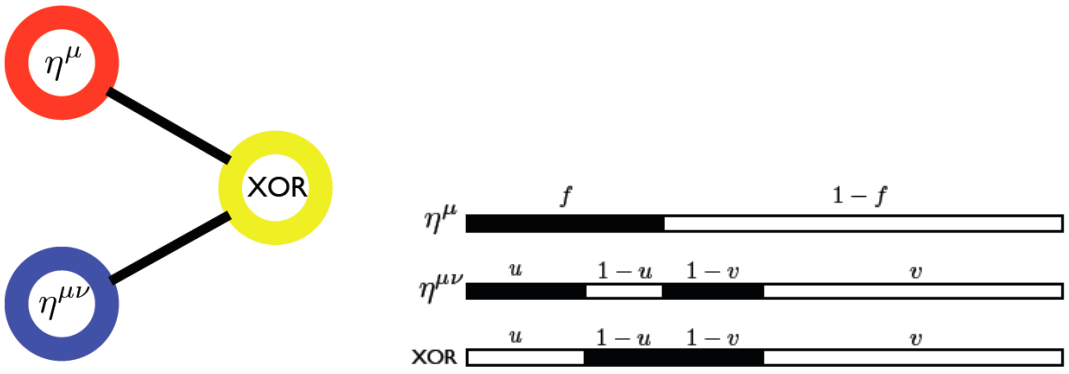
(c) f=0.05 (d) f=0.1

Figure 4.6: $g^{\mu}(t^*) - g^{\mu\nu}(t^*)$ as a function of m and q, for different coding sizes. When f grows the red area, corresponding to the dominating father regime, becomes smaller, reflecting the higher probability of overlap beetween patterns belonging to different classes. $p = 100$.

differencies, in fact the coding level of a difference pattern is on average

$$f_d = 2f(1-f)(1-m)$$

where f is the coding level of the original memories. For any $m > \dfrac{1-2f}{2(1-f)}$ we have that $f_d < f$. Since we assume a small f, this condition is easily satisfied. Moreover,



(a) (b)

Figure 4.7: (a) Schematic representation of two interacting networks: the first, in red, stores the father η^μ with the procedure described in Section 4.2.1; a second network, in yellow, receives both the instantaneous input stream (in blue) and the activity of the first network. (b) Taking the simple difference in this binary framework is equivalent to the boolean *exclusive or*, that we assume is performed somewhere in the connecting unit. Black (white)segments represent active (quiescent) neurons. In this work we do not investigate further any detailed underlying biological mechanism.

the correlations beetween differences of patterns belonging to the same class go to zero as $f \to 0$ or $m \to 1$:

$$E[\xi_i^1 \xi_i^2] - E[\xi_i^1]E[\xi_i^2] = f(1-u)^2 + (1-f)(1-v)^2 - f_d^2 = \\ = f(1-f)(1-m)^2\big(1 - 4f(1-f)\big) \xrightarrow[f\to 0,\ m\to 1]{} 0$$

In this limit all the calculations made in Section 4.1 for uncorrelated random patterns are valid, we simply need to replace f with f_d, the new coding level, and tune transition probabilities in the proper way to obtain the correct ratio of potentiating and depressing transitions:

$$q_+ = q \quad q_- = \frac{q f_d}{2(1-f_d)}$$

Chosen the network parameters, memory lifetime of a stored difference is, recalling eq. (4.10):

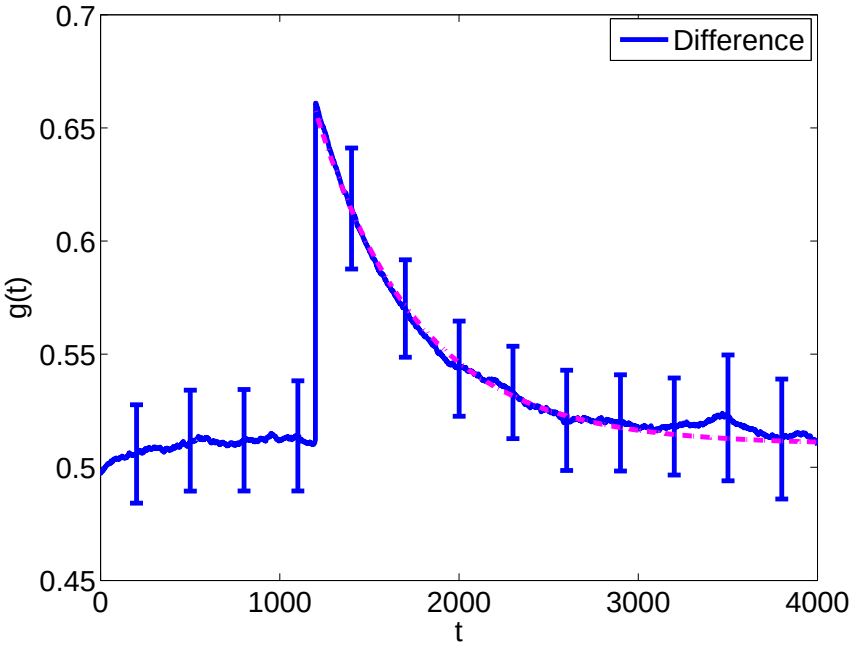


Figure 4.8: Memory trace of a pattern build taking the difference between a class member and the class prototype. Parameters are: $N = 1000$, $p = 50$, $f = 0.1$, $m = 0.7$, $q = 0.3$, simulations have been launched for 50 trials with random initialization of the synaptic matrix $(P(J_{ij}(0) = 1) = 0.5)$. When the son is presented to the original network $(t^* = 1200)$, the trace of its difference, shown to the XOR network, is also strengthened. Then the conditional distribution progressively decays to a value close to 0.5, as in the totally uncorrelated case. The dotted line represents the exponential fit from which we obtain $\tau_d^{sim} = 578\ unit\ time$, compared to a theoretical prediction of $\tau_d^{th} = 571.5\ u.t.$.

$$\tau_d = (2f_d^2 q)^{-1} = \left(8f^2(1-f)^2(1-m)^2 q\right)^{-1} \tag{4.17}$$

from which we see that bringing more correlated inputs (high m) boosts the capacity by reducing coding level f_d. To check the goodness of this prediction we have performed several simulations of this protocol. In particular, an example of the memory trace stored by this network is given in Fig 4.8 for $p = 50$, $f = 0.1$ and $m = 0.7$. In this case correlations between stored patterns are low, since the equilibrium distribution g_∞^d is slightly higher than 0.5, i.e. the value we obtained in the uncorrelated frame. The lifetime τ_d is obtained from simulation through an exponential fit (the dotted line in Fig. 4.8) of the curve with the time evolution of conditional probability:

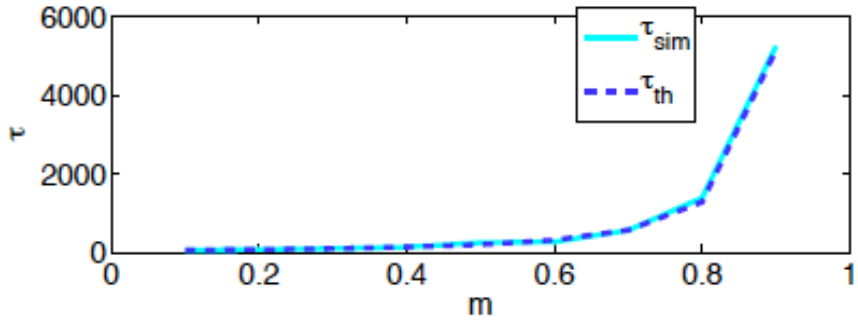


Figure 4.9: Memory lifetime τ_d vs. similarity parameter m. Continous light blue line represents the result obtained from the exponential fit (see eq. 4.18). Dotted blue line shows theoretical prediction.

$$g^d(t) = g^d_\infty + (1 - g^d_\infty)q \ \exp(-t/\tau_d) \tag{4.18}$$

In Fig. 4.9 we compare the theoretical prediction of eq. (4.17) with the value produced by the fit: the agreement is excellent, confirming that in the limit of small f correlations between difference patterns go to zero.

The advantage in terms of capacity deriving from the storage of the differences becomes substantial when the value of m approaches one, but, of course, there must be a drawback in terms of initial signal to noise ratio. In fact, as explained in Section 4.1, a decrease in the coding size would extend memory lifetime but would diminish the initial SNR, since

$$\frac{\mathcal{S}_0}{\mathcal{N}_0} \approx \sqrt{fN}$$

Thus, if we want to compare the capacity of a network storing a stream of uncorrelated single memories with one that receives a structured input, and memorizes the differences with the prototypes, we want both to start from the same initial SNR. Coding level is fixed, since it is a peculiar feature of the set of memories. Thus, the parameter we need to adjust is the synaptic transition probability q, the only free parameter left. Taking the usual limit ($f \to 0$, $N \to \infty$, $fN \to \infty$) the expressions

of the initial SNRs are

$$\frac{\mathcal{S}_0}{\mathcal{N}_0} = \sqrt{fN}\ \frac{\frac{1}{2}\ q}{2\sqrt{\frac{1}{4}(1+q)^2 + \frac{qf}{64}(1-q)^2 - f\big(\frac{1}{2}(1+q)\big)^2}}$$
$$\frac{\mathcal{S}_0^d}{\mathcal{N}_0^d} = \sqrt{f_d N}\ \frac{\frac{1}{2}\ q_d}{2\sqrt{\frac{1}{4}(1+q_d)^2 + \frac{q_d f_d}{64}(1-q_d)^2 - f_d\big(\frac{1}{2}(1+q_d)\big)^2}} \tag{4.19}$$

where we have approximated the conditional probabilities to the uncorrelated case:

$$g_\infty^d = \frac{1}{2} \qquad \gamma_0^d = \frac{1}{4}(1+q_d)^2 + \frac{q_d f_d}{64}(1+q_d)^2 + \mathcal{O}(f_d^2)$$

Neglecting terms of order f or f_d inside the denominators of eq. (4.19) and equating the two expressions, after some algebra we obtain:

$$q_d = \frac{\sqrt{f}\ q}{\sqrt{f_d}\ (1+q) - \sqrt{f}\ q} \tag{4.20}$$

This way we have balanced the learning speed of the two networks to get the same initial SNR. In eq. (4.9) we have seen that the decay time τ is a measure of the network storage capacity, especially on equal starting SNR. Hence, we can measure the capacity gain by looking at the ratio of the decay times, that grows as

$$\frac{\tau_d}{\tau} = \frac{\sqrt{2(1-f)(1-m)}(1+q) - q}{4(1-f)^2(1-m)^2} \tag{4.21}$$

The ratio reaches its maximum for

$$m_{max} = 1 - \frac{8q^2}{9(1-f)(1+q)^2} \approx 1 - \mathcal{O}(q^2)$$

Therefore, for m not too close to one and small q the advantage in terms of network

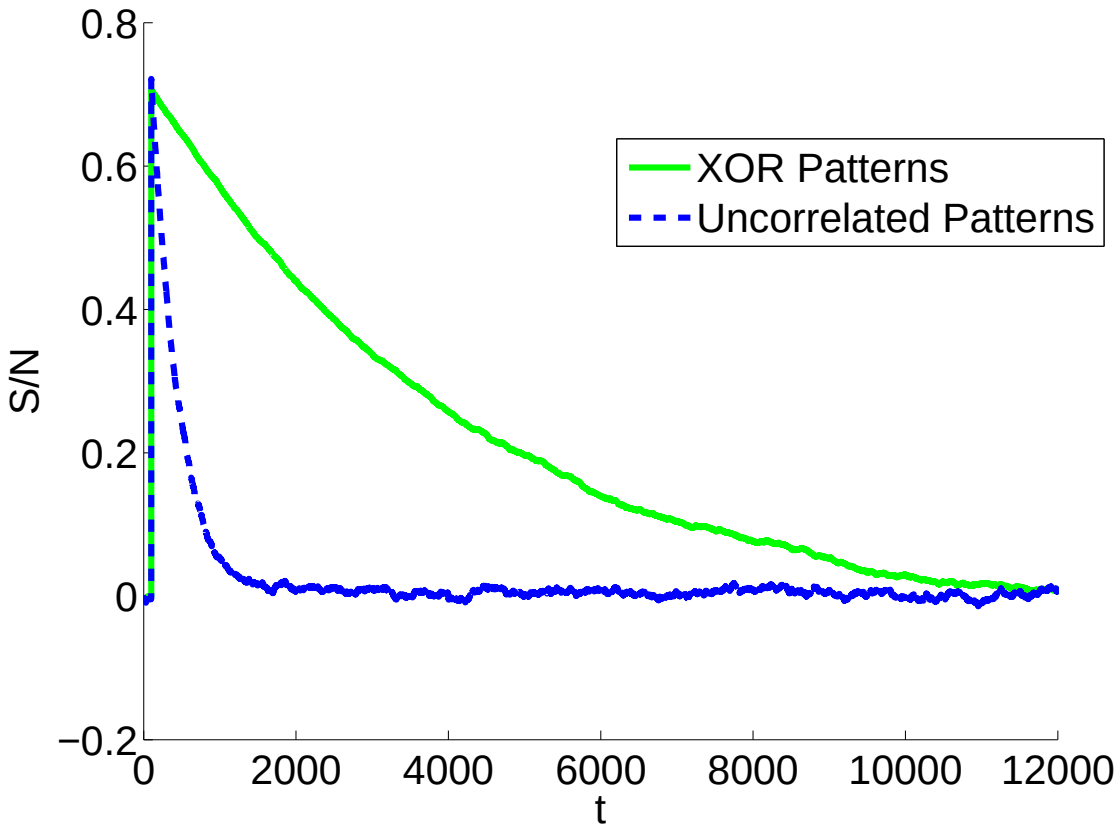


Figure 4.10: Signal-to-noise ratios of memories learned in two networks receiving different sets of memories. Blue line: uncorrelated patterns. Green line: differences between class members and prototypes imposed to a properly tuned network. In this simulation: $N = 1000$, $f = 0.1$, $q = 0.16$, $m = 0.9$, $f_d = 0.1$, $q_d = 0.5$, Trials$= 100$.

capacity scales like

$$\frac{\tau_d}{\tau} \approx (1-m)^{-3/2}$$

In Fig. 4.10 we exhibit the simulations of the two network settings. The blue line reproduces the SNR of a generic memory, when the input consists of a set of uncorrelated, randomly selected patterns. The tracked pattern is shown to the network at $t = 0$ and then erased upon the learning of new memories. Coding level is $f = 0.1$ and the learning rate is low: $q = 0.16$, slowing down the progressive decay of the SNR. The green trace, instead, corresponds to the SNR of a difference pattern, in the context of a two generations hierarchical structure of the input. The

coding size of the original patterns is equal to the uncorrelated situation ($f = 0.1$), but the similarity parameter is set to $m = 0.9$. This determines the average coding level of the difference patterns $f_d \equiv 2f(1-f)(1-m) = 0.018$, and, in consequence, the learning rate of the XOR network: $q_d \approx 0.5$. The resulting difference in memory lifetime is clear by looking at Fig. 4.10, however we have fitted the curves with two exponential functions of the form of eq. (4.8), obtaining[3]

$$\tau = 310 \ u.t. \quad \tau_d = 3142 \ u.t.$$

The storage capacity gain is then

$$\frac{\tau}{\tau_d} = 10.1$$

that is slightly less than the value calculated from eq. (4.21):

$$\frac{\tau_d^{th}}{\tau^{th}} \approx 10.2$$

4.2.3 Three generations

The previous scheme can be extended to complex nested hierarchies, containing three or more generations. So, for example, one could imagine an ultrametric tree representing a subsumptive hierarchy of objects, i.e. where the first branching nodes correspond to general categories and subsequent branches to more specific ones, as in the example of Fig. 4.11.

A generic three generations input structure is reported in Fig. 4.12, organized in the usual ultrametric tree scheme. The pattern set is built by generatinf p_1 uncorrelated *grandfathers* from distribution (4.1), and then p_2 *fathers* employing the same scheme

[3]Indeed the values extracted from the simulations are remarkably close to the theoretical results, predicted respectively by eq. (4.10) and eq. (4.17): $\tau^{th} = 312 \ u.t. \quad \tau_d^{th} = 3086 \ u.t.$

described in the two generations section. We will refer to the similarity parameter between grandfathers and fathers as m_1. From each of the second layer patterns we generate new sons according to equations

$$P(\eta_i^{\mu\nu\sigma} = 1|\eta_i^{\mu\nu} = 1) \equiv r = 1 - (1 - f)(1 - m_2)$$
$$P(\eta_i^{\mu\nu\sigma} = 0|\eta_i^{\mu\nu} = 0) \equiv s = 1 - f(1 - m_2)$$

that are simply an extension of eqs. (4.12).

Our goal is to find a first set of parameters that would permit the retrieval of the grandfather, a second one that would lead to the recall of the father, and a third that would favour the son attractor. In other words, we wish to build a network that is able to learn and memorize patterns belonging to any level of the hierarchy, just by varying the synaptic transition probability q, given that f and m are peculiar properties of the stimuli and are therefore fixed.

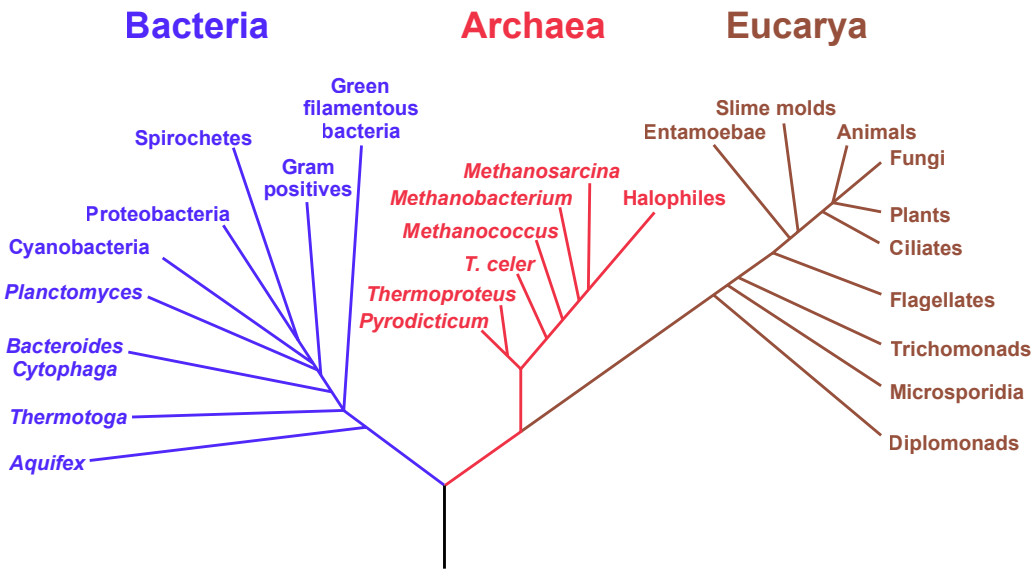


Figure 4.11: Simplified phylogenetic tree of living organisms, showing the evolutionary relationship between taxonomic groups. Each node represents a common ancestor for the descendants, located at the end of each branch.

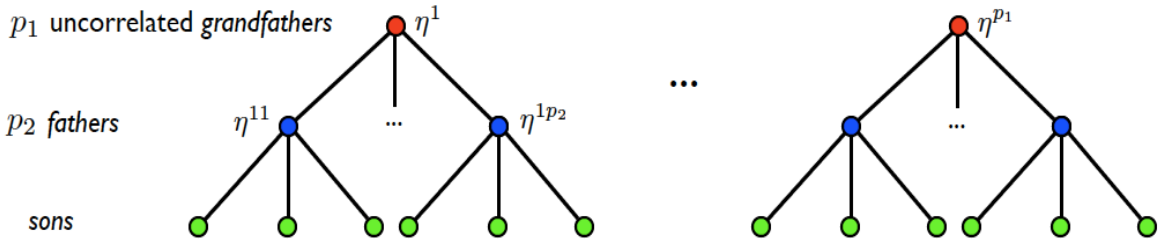


Figure 4.12: Three generations hierarchy of memories.

In order to achieve our target, we proceed by following the procedure used for the two generations case. We shuffle the presentation order of stimuli by choosing a particular grandfather with probability $1/p_1$, then we pick one of the fathers descending from the chosen grandfather with probability $1/p_2$, and, at last, we generate a new son from this very father. In this way we never present the same pattern twice and we can average over the sequences as we did before. The typical sequence consists of $p1$ patterns constructed by corrupting one of the $p_1 \cdot p_2$ fathers chosen at random. Due to the mixed correlations between subclasses, the mean fields equations are more complex than the two generations protocol, and give rise to very long and time-consuming expressions. The expression for the grandfather would be

$$P\big(J_{ij}(\infty)=1\big) \equiv g_\infty = \sum_{\Pi=0}^{p_1} \sum_{\Delta=0}^{p_1-\Pi} \psi(\Pi,\Delta) \sum_{\pi_1=0}^{\Pi p_2} \sum_{\delta_1=0}^{\Pi p_2-\pi_1} \phi_1(\pi_1,\delta_1) \sum_{\pi_2=0}^{\Delta p_2} \sum_{\delta_2=0}^{\Delta p_2-\pi_2} \phi_2(\pi_2,\delta_2) \times$$
$$\times \sum_{\pi_3=0}^{(p_1-\Pi-\Delta)p_2} \sum_{\delta_3=0}^{(p_1-\Pi-\Delta)p_2-\pi_3} \phi_3(\pi_3,\delta_3) \left(\frac{\beta(\pi_1,\pi_2,\pi_3,\delta_1,\delta_2,\delta_3)}{\beta(\pi_1,\pi_2,\pi_3,\delta_1,\delta_2,\delta_3)+\alpha(\pi_1,\pi_2,\pi_3,\delta_1,\delta_2,\delta_3)} \right) \tag{4.22}$$

$$\beta(\pi_1,\pi_2,\pi_3,\delta_1,\delta_2,\delta_3) = q_+\Big(r^2(\pi_1+\pi_2+\pi_3) + r(1-s)(\delta_1+\delta_2+\delta_3) + \\ + (1-s)^2(3p_2-\pi_1-\pi_2-\pi_3-\delta_1-\delta_2-\delta_3)\Big)$$

$$\alpha(\pi_1,\pi_2,\pi_3,\delta_1,\delta_2,\delta_3) = q_-\Big(2r(1-r)(\pi_1+\pi_2+\pi_3) + (rs+(1-s)(1-r))\times \\ \times(\delta_1+\delta_2+\delta_3) + 2s(1-s)(3p_2-\pi_1-\pi_2-\pi_3-\delta_1-\delta_2-\delta_3)\Big)$$

where the following are the probability distributions:

$$\psi(\Pi,\Delta) = \frac{p_1!}{\Pi!\Delta!(p_1-\Pi-\Delta)!}[f^2]^{\Pi}\ [2f(1-f)]^{\Delta}\ [(1-f)^2]^{(p_1-\Pi-\Delta)}$$

$$\phi_1(\pi_1,\delta_1) = \frac{\Pi p_2!}{\pi_1!\delta_1!(\Pi p_2-\pi_1-\delta_1)!}[u^2]^{\pi_1}\ [2u(1-u)]^{\delta_1}\ [(1-u)^2]^{(\Pi p_2-\pi_1-\delta_1)}$$

$$\phi_2(\pi_2,\delta_2) = \frac{\Delta p_2!}{\pi_2!\delta_2!(\Delta p_2-\pi_2-\delta_2)!}[u(1-v)]^{\pi_2}\ [uv+(1-u)(1-v)]^{\delta_2}\times \\ \times[(1-v)^2]^{(\Delta p_2-\pi_2-\delta_2)}$$

$$\phi_3(\pi_3,\delta_3) = \frac{((p_1-\Pi-\Delta)p_2)!}{\pi_3!\delta_3!((p_1-\Pi-\Delta)p_2-\pi_3-\delta_3)!}[(1-v)^2]^{\pi_3}\ [2v(1-v)]^{\delta_3}\times \\ \times[v^2]^{\left((p_1-\Pi-\Delta)p_2-\pi_3-\delta_3\right)}$$

The sums count the number of distinct submatrices, each experiencing a different sequence of potentiating or depressing events inside the synaptic matrix. This number grows exponentially with the total number of branches ($p_{tot} = p_1 \cdot p_2$), leading to very long calculation times. In order to reduce the number of factors inside the sums and, consequently, reduce the time for simulations, we need to find an approximation for expression (4.22).

When p_2 is large and bigger than p_1, the average number of potentiating, depressing or neutral events is given by the mean of distributions ϕ_1, ϕ_2, ϕ_3. For example, the

vector of the averages for ϕ_1 is

$$\vec{\mu}_1 = \Pi\ p_2 \begin{bmatrix} u^2 \\ 2u(1-u) \\ (1-u)^2 \end{bmatrix}$$

Fluctuations around these means, i.e. the square root of the variances, are proportional to $\sqrt{p_2}$, therefore in the limit $p_2 \gg 0$ we can replace ϕ_1, ϕ_2, ϕ_3 with their means. Conceptually, this corresponds to generating always new stimuli directly from the grandfather with conditional probabilities

$$\begin{aligned} P(\eta_i^{\mu\nu\sigma} = 1 | \eta_i^{\mu} = 1) \ &\equiv U = ur + (1-u)(1-s) = 1 - (1-f)(1-m_1 m_2) \\ P(\eta_i^{\mu\nu\sigma} = 0 | \eta_i^{\mu} = 0) \ &\equiv V = (1-v)(1-r) + vs = 1 - f(1-m_1 m_2) \end{aligned}$$

Eq. (4.22) may be rewritten as

$$g_\infty = \sum_{\Pi=0}^{p_1} \sum_{\Delta=0}^{p_1-\Pi} \psi(\Pi, \Delta) \frac{\mathsf{B}(\Pi, \Delta)}{\mathsf{B}(\Pi, \Delta) + \mathsf{A}(\Pi, \Delta)} \tag{4.23}$$

$$\begin{aligned} \mathsf{B}(\Pi, \Delta) &= q_+ \big(U^2 \Pi + U(1-V)\Delta + (1-V)^2 (p_1 - \Pi - \Delta) \big) \\ \mathsf{A}(\Pi, \Delta) &= q_- \big(2U(1-U)\Pi + [UV + (1-U)(1-V)]\Delta + 2(1-V)V(p_1 - \Pi - \Delta) \big) \end{aligned}$$

that clearly leads to a significant reduction in the number of factors inside the sums. Following the same procedure, we can obtain approximated expression for g^μ, $g^{\mu\nu}$ and $g^{\mu\nu\sigma}$. The overlap value of the generic grandfather is primarily influenced by the fact that, with probability $1/p_1$, the network sees a member of μ class correlated with the grandfather. At the same time, though, there are other correlations acting on the network that also shape the synaptic matrix. Those are intra-class correlations

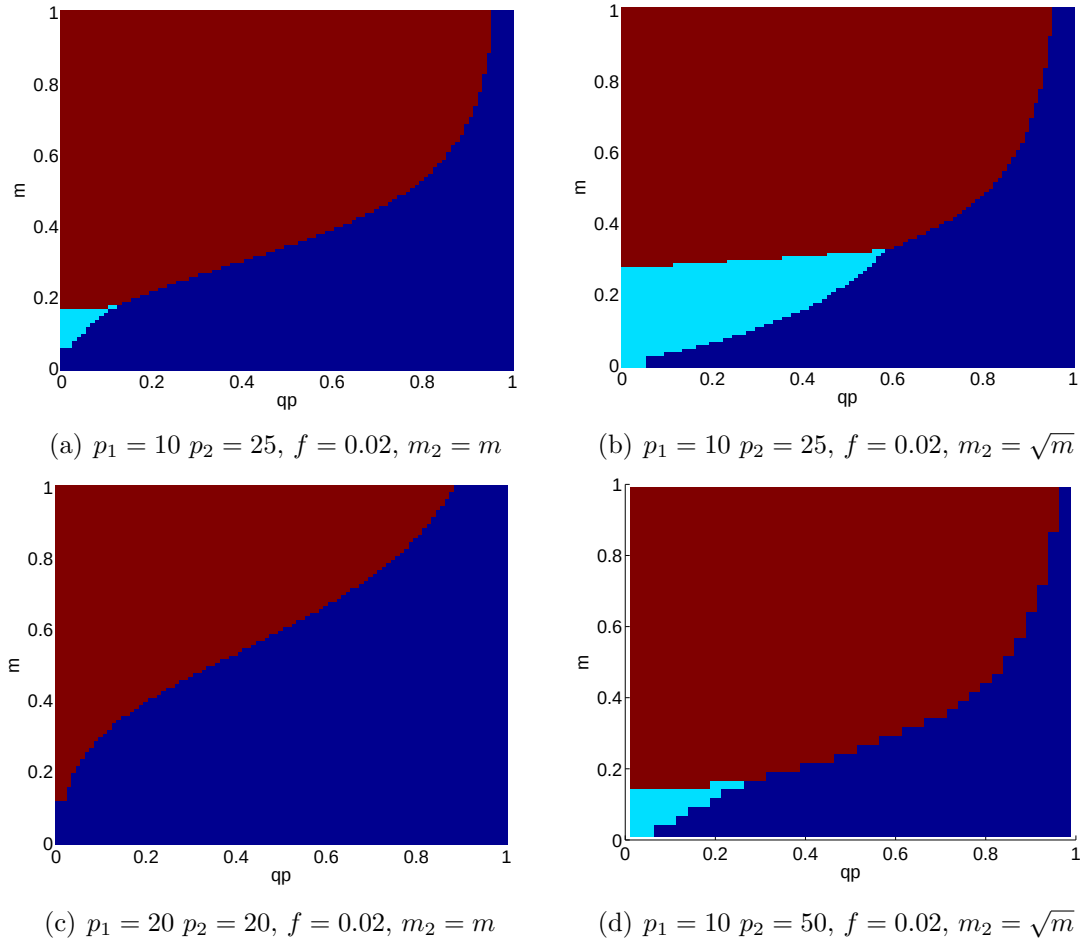
between members of other families (recall that there is no inter-family correlation). The average time separating the presentations of two patterns belonging to the same class is of order $\Delta t_1 \sim p_1$, while it is of order $\Delta t_2 \sim p_{tot}$ for intra-subclass memories. For $p_2 > p_1 \gg 0$ we can apply the approximation as in (4.23), keeping aside the contributions of all patterns belonging to the cluster of interest (μ). The grandfather signal is

$$g_\infty^\mu = \sum_{\Pi=0}^{p_1-1} \sum_{\Delta=0}^{p_1-1-\Pi} \psi_+(\Pi,\Delta) \sum_{\pi=0}^{p_2} \sum_{\delta=0}^{p_2-\pi} \phi_1(\pi,\delta) \left(\frac{\beta(\pi,\delta) + \mathsf{B}(\Pi,\Delta)}{\beta(\pi,\delta) + \mathsf{B}(\Pi,\Delta) + \alpha(\pi,\delta) + \mathsf{A}(\Pi,\Delta)} \right)$$

$$\begin{aligned}
\beta(\pi,\delta) &= q_+ \left(\frac{r^2\pi + r(1-s)\delta + (1-s)^2(p_2-\pi-\delta)}{p_2} \right) \\
\alpha(\pi,\delta) &= q_- \left(\frac{2r(1-r)\pi + (rs + (1-s)(1-r))\delta + 2s(1-s)(p_2-\pi-\delta)}{p_2} \right) \\
\mathsf{B}(\Pi,\Delta) &= q_+ \Big(U^2\Pi + U(1-V)\Delta + (1-V)^2(p_1-1-\Pi-\Delta) \Big) \\
\mathsf{A}(\Pi,\Delta) &= q_- \Big(2U(1-U)\Pi + (UV + (1-V)(1-U))\Delta + 2V(1-V)(p_1-1-\Pi-\Delta) \Big)
\end{aligned}$$

$$\begin{aligned}
\psi_+(\Pi,\Delta) &= \frac{(p_1-1)!}{\Pi!\Delta!(p_1-1-\Pi-\Delta)!} [f^2]^\Pi \ [2f(1-f)]^\Delta \ [(1-f)^2]^{(p_1-1-\Pi-\Delta)} \\
\phi_1(\pi,\delta) &= \frac{p_2!}{\pi!\delta!(p_2-\pi-\delta)!} [u^2]^\pi \ [2u(1-u)]^\delta \ [(1-u)^2]^{(p_2-\pi-\delta)}
\end{aligned}$$

Factors α and β express the contributions of family μ, while A and B are the approximations for the synaptic transitions provoked by all other families (uncorrelated to the μ-th one). The complete result for the second and the third stages of the hierarchy are exposed in Appendix B. For low values of p_2, the approximated expressions do not exactly reproduce the result of simulations, but are nevertheless qualitatively correct, i.e. the dominant memory trce is correctly predicted. This allows us to use this approximation for exploring the network parameters space. We wanted to know



(a) $p_1 = 10$ $p_2 = 25$, $f = 0.02$, $m_2 = m$

(b) $p_1 = 10$ $p_2 = 25$, $f = 0.02$, $m_2 = \sqrt{m}$

(c) $p_1 = 20$ $p_2 = 20$, $f = 0.02$, $m_2 = m$

(d) $p_1 = 10$ $p_2 = 50$, $f = 0.02$, $m_2 = \sqrt{m}$

Figure 4.13: Regions of the space spanned by parameters m and q_+ are colored according to the prevailing pattern: red for grandfather, light blue for the father, dark blue for the son. (a): the similarity parameter is the same for all layers; (b): $m_2 > m_1$ and the area corresponding to the father layer is larger; (c): when $p_2 \approx p_1$ the father region disappears completely (obtained from pure simulations); (d): increasing p_2 reduces the dominating father area like in (a).

which pattern is predominant (i.e. has the strongest memory trace in the network) within a certain set of parameters. Examples of the results are displayed in Fig. 4.13. Like the two generations case, there is a red region where the prototype (the grandfather) dominates, due to the high correlations with the presented stimuli, and a dark blue region where the stimulus presented last (the son) prevails. However, if the coding level is small (low probability of interference between patterns) it appears a light blue spot, which indicates that the father overcomes the other two.

This happens when m is not too high, so that the grandfather is not very much correlated with the sons, and q_+ is small, that is current stimulus does not predominate. Hence, as expected, when $m_2 > m_1$ the light blue spot expands, since the father is on average more correlated with the stimuli than the grandfather. Instead, the grandfather is more likely to dominate again when p_2 is very large, since the probability that the network sees a stimulus belonging to the father's subcluster is only $1/p_1p_2$, while the grandfather trace is strengthened with probability $1/p_1$.

Although we still need to perform a complete analysis of the results, it is indeed clear that the parameter space consists of three areas, one for each level of the ultrametric tree. We can therefore build three distinct networks, that memorize patterns at different stages of the hierarchy starting from exactly the same ensemble of input patterns.

As we did with a two layers hierarchy, we tested the possibility of learning the differences between the stimulus that is presently being shown to the network and the cluster or subcluster prototype. The capacity of a network K that learns the father-son differences is proportional to $\approx \left(8\ q\ f^2(1-f)^2(1-m_2)^2\right)^{-1}$. The gain is progressively less effective climbing up the hierarchy: if, for example, we introduce a network L capable of storing the grandfather-son difference, we gain a factor $\approx (1-m_1m_2)^{-2}$ in the memory storage capacity. Since m_1 is always smaller than one

$$(1-m_1m_2)^{-2} < (1-m)^{-2} \tag{4.24}$$

and thus

$$\tau_K > \tau_L \tag{4.25}$$

Chapter 5

Conclusions and future directions

In this dissertation we have explored the dynamic behaviour of associative recurrent networks with binary neurons and bistable synapses. The model is simple enough to be handled with the mathematical formalism of stochastic processes, but it indeed preserves many important features of real neuronal circuits, as described in Chapters 2 and 3, and it is therefore suitable even for future, more biologically oriented implementations. We measured the performance of the network in terms of memory capacity, when it is presented with either an uncorrelated random stream of input (for which we have also calculated the signal-to-noise ratio), or a structured, ultrametrically correlated hierarchy of stimuli. In the latter case, intrinsic network parameter q, that regulates the learning speed, may be slowed down so that the attractor of the network dynamics is a pattern representing the average of an ensemble of correlated patterns (categories), among which lies the currently presented stimulus.

Afterwards, we have verified that the maximum number of stored patterns increases

considerably when, instead of storing the whole patterns, the network memorizes only the bits that are uncorrelated with the average features of the class they belong to. Indeed, it is important to keep the prototype pattern stored in a separate network, for later retrieval of the original stimulus.

By extending this results to larger and more complex hierarchies, one could store memories representing the difference between patterns laying at higher levels of the hierarchy and the stimulus that is being shown to the network. For example, if we imagine to store the difference between current stimulus and one of its subclass prototypes, distant k steps in the ultrametric tree, we gain a factor $\approx (1 - m^k)^{-2}$ in terms of memory capacity (supposing that the similarity parameter m is the same at all stages). However, in order to memorize the whole hierarchy, we would need a network for each stage, which is not very convenient from a biological point of view. It could be the case, instead, that only the very first parent is stored, and the remaining subclasses are somehow reconstructed using that prototypical pattern as a reference. This way we can avoid the proliferation of subnetworks, which, indeed, does not seem to be biologically and evolutionary reasonable.

Evidences of categorization have been found all over the brain. Our model could be useful for describing the consolidation of short-term memories (the input patterns) to long-term memories (the class or subclass prototypes) in the hippocampus (Marr, 1971; Rolls, 1990). Before making any experimental prediction, however, we need to further analyse the material presented in this thesis. The complete model will be presented in a publication article that is currently being edited.

Still, there are many interesting questions and ideas regarding this work. First, our model is still far from real experimental protocols, an aspect that does not allow us to make strong predictions. Hence, in future studies, we hope to introduce more biological realism, for example embodying single neuron dynamics, in theoretical

modelling. Second, we still need to find a plausible mechanism that actually performs the difference between prototypes and class members. Given the huge number of connections in the brain, we suggest the implementation of *random projections* with a sparse matrix. This technique consists in projecting any set of points or vectors, lying in the N-dimensional space, to a randomly chosen M-dimensional space. M is independent on N and must be at least logarithmic in the size of the vector set.

In our case vectors represent the patterns of activity of the ANN which stores a cluster or subcluster prototype, and the flow of external stimuli. These vectors are multiplied by a $N \times M$ random matrix (where each entry is drawn independently from, for example, a gaussian distribution with zero mean and unitary variance) to obtain a projection in the M-dimensional space. A rather famous lemma by Johnson and Lindenstrauss (1984) guarantees that all pairwise distances among vectors (and hence similarities and differences between activity patterns) are maintained within an arbitrarily small factor. In particular, input patterns that are similar enough to let the recurrent network relax into the same attractor, will most likely have the same behavior in the projected network. The only difference is that all distances between patterns of neural activities would be stretched on average by some common factor.

A similar and possibly more appealing way of projecting the signal from the first to the second network should involve Compressed Sensing theory, a recent signal reconstruction theory developed by E. Candes, D. Donoho, J. Romberg, T. Tao and M. Wakin. Compressed Sensing (CS) exploits the fact that many interesting signals possess a structure which renders them very sparse when represented in the proper basis. While classical information theory would encode these signals by sampling at a rate above the Nyquist-Shannon criterion to subsequently compress the signal, CS

proceeds by directly sampling the signal at a low rate by projecting it onto a basis which is maximally incoherent with respect to the natural sparse basis of the signal, in order to obtain a representation which is already compressed. This allows high-resolution acquisition with low-resolution sensors. CS works because, if there exists a basis in which the signal is sparse, then the product of an incoherent sensing matrix with the matrix of basis vectors is roughly orthogonal when restricted to the space of sparse signals (Candes and Tao, 2005). Since orthogonal means invertible, the signal can be (almost) exactly reconstructed. The reconstruction technique doesn't involve inversion, though, but rather a minimization extremizing the sparseness of the signal. Since it does not imply any intelligent strategy, it would be interesting to look whether CS (or random projections theory) may be useful for our purposes. Third, and last, our model presents some rather impressive analogies with bayesian models of category learning (Sanborn et al., 2006) that are widely used in Machine Learning and Artificial Intelligence fields. We look forward to integrating some concepts and ideas borrowed from Bayesian statistics into our model, in the hope that we could contribute again to the understanding of memory and learning processes.

Appendix A

Floating Coding Level

A.1 Signal

Inside Section 4.1.1 we defined normalized signal as

$$\mathcal{S}_n(t) = E\left[\frac{1}{f^2 N(N-1)} \sum_{i \neq j} J_{ij}(t)\eta_i^\mu \eta_j^\mu\right]$$

The normalization coefficient comes from the statistics of the input pattern, as we will show in the following. Let us start with the unnormalized signal, i.e. the expectation value of the overlap between a memory and the synaptic matrix:

$$\mathcal{S}(t) = E\left[\sum_{i \neq j} J_{ij}(t)\eta_i^\mu \eta_j^\mu\right]$$

When taking the expectation value we have to be careful that the number of active neurons per pattern is equal to fN only on average, since it fluctuates with standard deviation $f(1-f)N$. To take into account this feature we notice that, if stochastic

variables X and Y live in the same probability space and , we can write

$$E[X] = E\big(E[X|Y]\big)$$

if X and Y run respectively over the sets of values $\{x\}$ and $\{y\}$:

$$E\big(E[X|Y]\big) = \sum_y E[X|y]\ P(Y=y) = \sum_y \Big(\sum_x x\ P(X=x|y)\Big) P(Y=y)$$

thus, conditioning our observable $\sum_{i\neq j} J_{ij}(t)\eta_i^\mu \eta_j^\mu \equiv X$ on the number of active neurons in tracked memory, which is itself a stochastic variable $\sum_a \eta_a^\mu \equiv Y$ with average fN, and variance $f(1-f)N$, we get

$$E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big] = E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \Big| \sum_a \eta_a^\mu\Big]\Big) =$$

$$= P(J_{ij}(t)=1|\eta_i^\mu=1,\eta_j^\mu=1)\sum_{y=0}^{N}(y^2-y)P\big(\sum_a \eta_a^\mu = y\big)$$

where the term inside the sum is simply the average number of synapses having value $J=1$ at time t, given that the effective coding level of pattern μ is $\frac{y}{N}$. In our simulations memories consist of N bits, each of which is taken to be one with probability f or zero with probability $(1-f)$. Consequently, the number of active units in the tracked memory $\sum_a \eta_a^\mu$ follows the binomial distribution

$$P\big(\sum_{a=1}^{N}\eta_a^\mu = y\big) = \frac{N!}{y!(N-y)!} f^y (1-f)^{N-y}$$

The first two moments around zero of this probability distribution are

$$\begin{aligned}\mathcal{M}^{(1)} &= \sum_{y=0}^{N} yP\big(\sum_{a=1}^{N}\eta_a^\mu = y\big) = fN \\ \mathcal{M}^{(2)} &= \sum_{y=0}^{N} y^2P\big(\sum_{a=1}^{N}\eta_a^\mu = y\big) = f^2N^2 + f(1-f)N\end{aligned}$$

the unnormalized signal is then a combination of these two moments times the conditional probability $g^\mu(t)$ introduced in Section 4.1.1:

$$E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big] = g^\mu(t)\big(\mathcal{M}^{(2)} - \mathcal{M}^{(1)}\big) = g^\mu(t)\big(f^2N(N-1)\big) \tag{A.1}$$

from which we derive the normalized signal

$$\mathcal{S}_n(t) = g^\mu(t) = E\Big[\frac{1}{f^2N(N-1)}\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big]$$

with the proper normalizing constant.

A.2 Noise

Now let us turn attention to *noise*. Recalling the definition from Section (4.1.2):

$$\mathcal{N}(t) = \sqrt{Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i\eta_j\Big]}$$

The *law of total variance* states that, if random variables X and Y belong to the same probability space and the variance of X stays finite, then:

$$Var[X] = E\Big[Var[X|Y]\Big] + Var\Big[E[X|Y]\Big]$$

So, like we just did with the signal, we condition our observable on the number of active neurons per memory, :

$$Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big] = E\Big(Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]\Big) + Var\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]\Big) \quad \text{(A.2)}$$

where the first is the mean of the conditional variance:

$$E\Big(Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]\Big) = \\ = E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu\Big|\sum_a \eta_a^\mu\Big]\Big) - E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]^2\Big) \quad \text{(A.3)}$$

and the second is the variance of the conditional expectation value:

$$Var\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]\Big) = \\ = E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]^2\Big) - E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big|\sum_a \eta_a^\mu\Big]\Big)^2 \quad \text{(A.4)}$$

Let us start with the expression (A.3). The first term is the mean of the conditional expectation value:

$$E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu\Big]\Big) = \\ = \sum_{y=0}^{N} E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu = y\Big] P\big(\sum_a \eta_a^\mu = y\big) \quad \text{(A.5)}$$

The expected value inside the sum consists of four terms describing all possible relationships between synapses:

$$E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu = y\Big] = \\ = P(J_{ij}(t)=1, J_{kl}(t)=1|\eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu = 1)E\big[\sum_{i\neq j\neq k\neq l} \eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu \big| \sum \eta_a^\mu = y\big] + \\ + P(J_{ij}(t)=1, J_{il}(t)=1|(\eta_i^\mu)^2\eta_j^\mu\eta_l^\mu = 1)E\big[\sum_{i=k\neq j\neq l} (\eta_i^\mu)^2\eta_j^\mu\eta_l^\mu \big| \sum \eta_a^\mu = y\big] + \\ + P(J_{ij}(t)=1, J_{ji}(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2 = 1)E\big[\sum_{i=k\neq j=l} (\eta_i^\mu)^2(\eta_j^\mu)^2 \big| \sum \eta_a^\mu = y\big] + \\ + P(J_{ij}^2(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2 = 1)E\big[\sum_{i=k\neq j=l} (\eta_i^\mu)^2(\eta_j^\mu)^2 \big| \sum \eta_a^\mu = y\big]$$

explicitating the expectation value in each term:

$$= P(J_{ij}(t)=1, J_{kl}(t)=1|\eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu = 1)\ y(y-1)(y-2)(y-3) + \\ + P(J_{ij}(t)=1, J_{il}(t)=1|(\eta_i^\mu)^2\eta_j^\mu\eta_l^\mu = 1)\ 4y(y-1)(y-2) + \\ + P(J_{ij}(t)=1, J_{ji}(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2 = 1)\ y(y-1) + \\ + P(J_{ij}^2(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2 = 1)\ y(y-1)$$

where the second contains the correlations given by synapses sharing one neuron, the third the correlations of synapses sharing both neurons, while the first and the

fourth are the uncorrelated contributions to the noise.

We may then rewrite (A.5) as

$$
\begin{aligned}
&E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu\Big]\Big) = \\
&= P(J_{ij}(t)=1, J_{kl}(t)=1|\eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu=1)\sum_{y=0}^{N} y(y-1)(y-2)(y-3)P\big(\sum_a \eta_a^\mu = y\big)+ \\
&+ P(J_{ij}(t)=1, J_{il}(t)=1|(\eta_i^\mu)^2\eta_j^\mu\eta_l^\mu=1)\sum_{y=0}^{N} 4y(y-1)(y-2)P\big(\sum_a \eta_a^\mu = y\big)+ \\
&+ P(J_{ij}(t)=1, J_{ji}(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2=1)\sum_{y=0}^{N} y(y-1)P\big(\sum_a \eta_a^\mu = y\big)+ \\
&+ P(J_{ij}^2(t)=1, |(\eta_i^\mu)^2(\eta_j^\mu)^2=1)\sum_{y=0}^{N} y(y-1)P\big(\sum_a \eta_a^\mu = y\big) \qquad \text{(A.6)}
\end{aligned}
$$

and again the sums inside equation (A.6) are a miscellany of moments around zero of the binomial probability distribution. The two more moments we need are:

$$
\begin{aligned}
\mathcal{M}^{(3)} &= \sum_{y=0}^{N} y^3 P\big(\sum_{a=1}^{N} \eta_a^\mu = y\big) = fN(1-3f+3fN+2f^2-3f^2N+f^2N^2) \\
\mathcal{M}^{(4)} &= \sum_{y=0}^{N} y^4 P\big(\sum_{a=1}^{N} \eta_a^\mu = y\big) = fN(1-7f+7fN+12f^2-18f^2N+6f^2N^2+ \\
&\qquad\qquad - 6f^3+11f^3N-6f^3N^2+f^3N^3)
\end{aligned}
$$

Inserting them back into (A.6) we get:

$$
E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu\Big]\Big) =
$$

$$
\begin{aligned}
&= P(J_{ij}(t)=1, J_{kl}(t)=1|\eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu=1)(\mathcal{M}^{(4)}-6\mathcal{M}^{(3)}+11\mathcal{M}^{(2)}-6\mathcal{M}^{(1)})+\\
&+ P(J_{ij}(t)=1, J_{il}(t)=1|(\eta_i^\mu)^2\eta_j^\mu\eta_l^\mu=1)(4\mathcal{M}^{(3)}-12\mathcal{M}^{(2)}+8\mathcal{M}^{(1)})+\\
&+ P(J_{ij}(t)=1, J_{ji}(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2=1)(\mathcal{M}^{(2)}-\mathcal{M}^{(1)})+\\
&+ P(J_{ij}^2(t)=1|(\eta_i^\mu)^2=1,(\eta_j^\mu)^2=1)(\mathcal{M}^{(2)}-\mathcal{M}^{(1)}) \qquad \text{(A.7)}
\end{aligned}
$$

Now we need to find the conditional probabilities of finding a pair of synapses in the potentiated state at time t given the presentation of memory pattern μ at $t=0$. Starting from the uncorrelated terms in the previous equation, we immediately see that the probability in the first term is equivalent to the square of the conditional probability of finding one enhanced synapse, since the activities of synapses connecting two separate pairs of neurons are uncorrelated:

$$P(J_{ij}(t)=1, J_{kl}(t)=1|\eta_i^\mu\eta_j^\mu\eta_k^\mu\eta_l^\mu=1) = \big(P(J_{ij}(t)=1|\eta_i^\mu\eta_j^\mu=1)\big)^2 = (g^\mu(t))^2$$

To derive the distribution in the fourth term we only need to note that in our framework $(\eta_i^\mu)^2=\eta_i^\mu$ and $J_{ij}^2=J_{ij}$, hence:

$$P(J_{ij}^2(t)=1|(\eta_i^\mu)^2(\eta_j^\mu)^2=1) = P(J_{ij}(t)=1|\eta_i^\mu\eta_j^\mu=1) = g^\mu(t)$$

The remaining terms account for synapses sharing one or two input neurons. The corresponding markov process has four states determined by the four possible combination of synaptic states 0 and 1, giving rise to a 4x4 transition probabilities matrix. We have obtained the two matrices, one for synapses receiving input from one common neuron, and the other for synapses sharing both neurons although with

opposite input-output roles. They have the form:

$$M(J_{ij}(t+1), J_{il}(t+1)|J_{ij}(t), J_{il}(t), \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) =$$

$$= \begin{pmatrix} 1-2m_{11}^{10}-m_{11}^{00} & m_{11}^{10} & m_{11}^{01} & m_{11}^{00} \\ m_{10}^{11} & 1-m_{10}^{11}-m_{10}^{01}-m_{10}^{00} & m_{10}^{01} & m_{10}^{00} \\ m_{01}^{11} & m_{01}^{10} & 1-m_{01}^{11}-m_{01}^{10}-m_{01}^{00} & m_{01}^{00} \\ m_{00}^{11} & m_{00}^{10} & m_{00}^{01} & 1-2m_{00}^{10}-m_{00}^{11} \end{pmatrix}$$

From now on we omit from notation that the matrices regard only synapses connecting neurons that were active upon the presentation of the tracked memory μ. When input is a sequence of random and uncorrelated memories, transition probabilities $m_{J^t J^t}^{J^{t+1} J^{t+1}}$ for synapses having one common neuron are:

$$m_{11}^{10} = m_{11}^{01} = P(J_{ij}(t+1) = 1, J_{il}(t+1) = 0|J_{ij}(t) = 1, J_{il}(t) = 1, \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) =$$

$$= q_- f(1-f)(2-q_-)$$

$$m_{11}^{00} = q_-^2 f(1-f)$$

$$m_{10}^{11} = m_{01}^{11} = q_+ f^2(1-q_-(1-f))$$

$$m_{10}^{01} = m_{01}^{10} = q_+ q_- f^2(1-f)$$

$$m_{10}^{00} = m_{01}^{00} = q_- f(1-f)(2-q_+ f)$$

$$m_{00}^{11} = q_+^2 f^3$$

$$m_{00}^{10} = m_{00}^{01} = q_+ f^2(1-q_+ f)$$

A 4x4 matrix has four eigenvalues λ_α ($\alpha = 1, ..., 4$). If the process is ergodic and irreducible, as in our case, the eigenvalues can be arranged according to magnitude: $\lambda_1 = 1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \lambda_4$. The leading eigenvalue λ_1 is associated to the asymptotic

distribution of synaptic efficacies, in fact it must hold:

$$\pi_1 M = \lambda_1 \pi_1$$

where, since $\lambda_1 = 1$, the left eigenvector π_1 is the equilibrium distribution of synaptic efficacies, undergoing the stochastic process defined by transition probability matrix M. The probability of finding two potentiated synapses sharing one neuron when $t \to \infty$ and conditioned on μ-th memory is then:

$$\gamma_\infty^\mu \equiv P(J_{ij} = 1, J_{il} = 1 | \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) =$$
$$= \frac{q_+^2 f^2 (2 - q_+ f)}{(2q_-(1-f) + q_+ f)(q_+ f(2 - q_+ f) - q_-^2(1-f) + 2q_-(1-f)(2 - q_+ f))}$$

Repeating the same argument for synapses sharing two neurons we find:

$$m_{11}^{10} = m_{11}^{01} = 2f(1-f)q_-(1-q_-)$$
$$m_{11}^{00} = 2q_-^2 f(1-f)$$
$$m_{10}^{11} = m_{01}^{11} = q_+ f^2$$
$$m_{10}^{01} = m_{01}^{10} = 0$$
$$m_{10}^{00} = m_{01}^{00} = 2q_- f(1-f);$$
$$m_{00}^{11} = q_+^2 f^2$$
$$m_{00}^{10} = m_{00}^{01} = q_+ f^2 (1 - q_+)$$

$$\sigma_\infty^\mu \equiv P(J_{ij} = 1, J_{ji} = 1 | \eta_i^\mu \eta_j^\mu = 1) =$$
$$= \frac{f q_+^2 ((1-f)2q_- + f(2 - q_-))}{((2(1-f)q_- + f q_+)(2(1-f)(2-q_-)q_- + 2f q_+ - f q_+^2)}$$

Inserting these results back into (A.7):

$$E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \sum_{k\neq l} J_{kl}(t)\eta_k^\mu\eta_l^\mu \Big| \sum_a \eta_a^\mu\Big]\Big) =$$

$$= (g^\mu(t))^2(\mathcal{M}^{(4)} - 6\mathcal{M}^{(3)} + 11\mathcal{M}^{(2)} - 6\mathcal{M}^{(1)}) + \gamma^\mu(t)(4\mathcal{M}^{(3)} - 12\mathcal{M}^{(2)} + 8\mathcal{M}^{(1)}) +$$
$$+ \sigma^\mu(t)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)}) + g^\mu(t)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)})$$

All correlations are embodied in the latter expression, the remaining quantities in eq. (A.2) are therefore easier to be calculated. Skipping superfluous details, the final results are:

$$E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \Big| \sum_a \eta_a^\mu\Big]^2\Big) = (g^\mu(t)^2)(\mathcal{M}^{(4)} + \mathcal{M}^{(2)} - 2\mathcal{M}^{(3)})$$

$$E\Big(E\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu \Big| \sum_a \eta_a^\mu\Big]\Big)^2 = (g^\mu(t)^2)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)})^2$$

We can now write the total variance as:

$$Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big] = (g^\mu(t))^2(\mathcal{M}^{(4)} - 6\mathcal{M}^{(3)} + 11\mathcal{M}^{(2)} - 6\mathcal{M}^{(1)}) +$$
$$+ \gamma^\mu(t)(4\mathcal{M}^{(3)} - 12\mathcal{M}^{(2)} + 8\mathcal{M}^{(1)}) + \sigma^\mu(t)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)}) +$$
$$+ g^\mu(t)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)}) - (g^\mu(t)^2)(\mathcal{M}^{(2)} - \mathcal{M}^{(1)})^2$$

keeping only the leading coefficients in the limit of high sparseness and large network ($f \to 0$, $N \to \infty$, $fN \to \infty$) the expression for the noise becomes:

$$\mathcal{N}(t) = \sqrt{Var\Big[\sum_{i\neq j} J_{ij}(t)\eta_i^\mu\eta_j^\mu\Big]} = 2(fN)^{3/2}\big(\gamma^\mu(t) - fg^\mu(t)^2\big)^{1/2}$$

it can be easily seen from (A.1) that signal in that limit has the form:

$$\mathcal{S}(t) = (fN)^2 g^\mu(t)$$

the signal to noise ratio grows with the square root of the network size and of the coding level:

$$\frac{\mathcal{S}}{\mathcal{N}} = \sqrt{fN}\frac{g^\mu(t)}{2\sqrt{\gamma^\mu(t) - f g^\mu(t)^2}}$$

it is worthwhile to remark that the leading term in the noise contains correlations through the probability distribution γ^μ, that therefore influence the capacity of retrieval, but for networks large enough the signal to noise ratio is guaranteed to be high enough, because all the dependance on N is contained in the coefficent $\sqrt{fN}$.

A.3 SNR expansion in a balanced network

Right after the presentation (at $t = 0$) of the tracked memory the conditional distribution $\gamma^\mu(t)$ is given by the probability that a pair of depressed correlated synapses underwent potentiation plus the probability that two correlated synapses with opposite efficacies end up being both potentiated, and obviously the fraction of correlated synapses that were already potentiated before the presentation of memory μ:

$$\begin{aligned}\gamma^\mu(0) = P(J_{ij} = 1, J_{il} = 1 | \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) \; + \; 2P(J_{ij} = 0, J_{il} = 1 | \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) \; q \; + \\ + \; P(J_{ij} = 0, J_{il} = 0 | \eta_i^\mu \eta_j^\mu \eta_l^\mu = 1) \; q^2 \; = \; \gamma_{11}^\mu \; + \; 2 \; \gamma_{10}^\mu \; q \; + \; \gamma_{00}^\mu \; q^2 \end{aligned} \quad \text{(A.8)}$$

where γ_{11}^μ, γ_{10}^μ and γ_{00}^μ are found by solving the eigenvalue problem for the 4x4 transition matrix $M(J_{ij}(t+1), J_{il}(t+1) | J_{ij}(t), J_{il}(t), \eta_i^\mu = 1, \eta_j^\mu = 1, \eta_l^\mu = 1)$. We have calculated these probabilities when memory patterns are uncorrelated. Skipping

unimportant details, we give the result:

$$\gamma_{11}^{\mu} = \frac{q_+^2 f^2 (2 - q_+ f)}{\big(2q_-(1-f) + q_+ f\big)\big(q_+ f(2-q_+ f) - q_-^2(1-f) + 2q_-(1-f)(2-q_+ f)\big)}$$

$$\gamma_{10}^{\mu} = \frac{q_+ q_- f(1-f)(4 - 2f q_+ - q_-)}{\big(2q_-(1-f) + q_+ f\big)\big(q_+ f(2-q_+ f) - q_-^2(1-f) + 2q_-(1-f)(2-q_+ f)\big)}$$

$$\gamma_{00}^{\mu} = \frac{(1-f) q_-^2 (2(1-f)(4-q_-) + f(4f-3)q_+)}{\big(2q_-(1-f) + q_+ f\big)\big(q_+ f(2-q_+ f) - q_-^2(1-f) + 2q_-(1-f)(2-q_+ f)\big)}$$

We now adopt the prescription

$$q_+ = q \qquad\qquad q_- = \frac{qf}{2(1-f)} \tag{A.9}$$

in order to balance the average number of potentiating and depressing events in the synaptic matrix. Achieved the correct simmetry, we expand the conditional probabilities γ_{11}^{μ}, γ_{10}^{μ} and γ_{00}^{μ} to the first order in f:

$$\gamma_{11}^{\mu} = \gamma_{00}^{\mu} = \frac{1}{4} + \frac{qf}{64} \qquad\qquad \gamma_{10}^{\mu} = \frac{1}{4} - \frac{qf}{64}$$

Using this and the fact that $g_{\infty}^{\mu} = \dfrac{1}{2}$ we can rewrite the initial SNR as

$$\frac{\mathcal{S}_0}{\mathcal{N}_0} = \sqrt{fN}\, \frac{\dfrac{1}{2}\ q}{2\sqrt{\dfrac{1}{4}(1+q)^2 + \dfrac{qf}{64}(1-q)^2 - f\big(\dfrac{1}{2}(1+q)\big)^2}}$$

Appendix B

Many subclasses limit

B.1 Approximated expressions for the overlaps

In Section 4.2.3 we have delineated the many patterns approximation in the three generations hierarchy. The effort has been motivated by the need for decreasing the number of different submatrices making up the synaptic matrix. This reduces the amount of time needed to calculate the expressions. The grandfather overlap is, we now give the expressions for the other generations.

The father's sequence averaged overlap is

$$g_{\infty}^{\mu\nu} = \sum_{\Pi=0}^{p_1-1} \sum_{\Delta=0}^{p_1-1-\Pi} \psi_{+}(\Pi,\Delta) \left[\sum_{\pi_1=0}^{p_2-1} \sum_{\delta_1=0}^{p_2-1-\pi} \phi_{1+}(\pi_1,\delta_1) \times \right.$$
$$\left. \times \left(\frac{\beta_{1+}(\pi_1,\delta_1) + \mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_1,\delta_1) + \mathsf{B}(\Pi,\Delta) + \alpha_{1+}(\pi_1,\delta_1) + \mathsf{A}(\Pi,\Delta)} \right) + \right.$$

$$+\sum_{\pi_2=0}^{p_2-1}\sum_{\delta_2=0}^{p_2-1-\pi}\phi_{2+}(\pi_2,\delta_2)\left(\frac{\beta_{1+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)+\alpha_{1+}(\pi_2,\delta_2)+\mathsf{A}(\Pi,\Delta)}\right)\Bigg]+$$
$$+\sum_{\pi_3=0}^{p_2-1}\sum_{\delta_3=0}^{p_2-1-\pi}\phi_{3+}(\pi_3,\delta_3)\left(\frac{\beta_{1+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)+\alpha_{1+}(\pi_3,\delta_3)+\mathsf{A}(\Pi,\Delta)}\right)\Bigg] \quad (11)$$

The memory trace of a generic son $\eta^{\mu\nu\sigma}$ is

$$\begin{aligned}
g_\infty^{\mu\nu\sigma} = &\sum_{\Pi=0}^{p_1-1}\sum_{\Delta=0}^{p_1-1-\Pi}\psi_+(\Pi,\Delta)\times\\
&\times\Bigg[U^2\sum_{\pi_1=0}^{p_2-1}\sum_{\delta_1=0}^{p_2-1-\pi}\phi_1(\pi_1,\delta_1)\Bigg(r^2\frac{\beta_{1+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)+\alpha_{1+}(\pi_1,\delta_1)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{2f(1-f)r(1-s)}{f^2}\frac{\beta_{2+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)}{\beta_{2+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)+\alpha_{2+}(\pi_1,\delta_1)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{(1-f)^2(1-s)^2}{f^2}\frac{\beta_{3+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)}{\beta_{3+}(\pi_1,\delta_1)+\mathsf{B}(\Pi,\Delta)+\alpha_{3+}(\pi_1,\delta_1)+\mathsf{A}(\Pi,\Delta)}\Bigg)+\\
&+\frac{2f(1-f)U(1-V)}{f^2}\sum_{\pi_2=0}^{p_2-1}\sum_{\delta_2=0}^{p_2-1-\pi}\phi_2(\pi_2,\delta_2)\times\\
&\times\Bigg(r^2\frac{\beta_{1+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)+\alpha_{1+}(\pi_2,\delta_2)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{2f(1-f)r(1-s)}{f^2}\frac{\beta_{2+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)}{\beta_{2+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)+\alpha_{2+}(\pi_2,\delta_2)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{(1-f)^2(1-s)^2}{f^2}\frac{\beta_{3+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)}{\beta_{3+}(\pi_2,\delta_2)+\mathsf{B}(\Pi,\Delta)+\alpha_{3+}(\pi_2,\delta_2)+\mathsf{A}(\Pi,\Delta)}\Bigg)+\\
&+\frac{(1-f)^2(1-V)^2}{f^2}\sum_{\pi_3=0}^{p_2-1}\sum_{\delta_3=0}^{p_2-1-\pi}\phi_3(\pi_3,\delta_3)\times\\
&\times\Bigg(r^2\frac{\beta_{1+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)}{\beta_{1+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)+\alpha_{1+}(\pi_3,\delta_3)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{2f(1-f)r(1-s)}{f^2}\frac{\beta_{2+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)}{\beta_{2+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)+\alpha_{2+}(\pi_3,\delta_3)+\mathsf{A}(\Pi,\Delta)}+\\
&+\frac{(1-f)^2(1-s)^2}{f^2}\frac{\beta_{3+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)}{\beta_{3+}(\pi_3,\delta_3)+\mathsf{B}(\Pi,\Delta)+\alpha_{3+}(\pi_3,\delta_3)+\mathsf{A}(\Pi,\Delta)}\Bigg)\Bigg]
\end{aligned}$$

where:

$$\beta_{1+}(\pi,\delta) = q_+\left(\frac{r^2(\pi+1)+r(1-s)\delta+(1-s)^2(p_2-1-\pi-\delta)}{p_2}\right)$$
$$\alpha_{1+}(\pi,\delta) = q_-\left(\frac{2r(1-r)(\pi+1)+(rs+(1-s)(1-r))\delta+2s(1-s)(p_2-1-\pi-\delta)}{p_2}\right)$$
$$\beta_{2+}(\pi,\delta) = q_+\left(\frac{r^2\pi+r(1-s)(\delta+1)+(1-s)^2(p_2-1-\pi-\delta)}{p_2}\right)$$
$$\alpha_{2+}(\pi,\delta) = q_-\left(\frac{2r(1-r)\pi+(rs+(1-s)(1-r))(\delta+1)+2s(1-s)(p_2-1-\pi-\delta)}{p_2}\right)$$
$$\beta_{3+}(\pi,\delta) = q_+\left(\frac{r^2\pi+r(1-s)\delta+(1-s)^2(p_2-\pi-\delta)}{p_2}\right)$$
$$\alpha_{3+}(\pi,\delta) = q_-\left(\frac{2r(1-r)\pi+(rs+(1-s)(1-r))\delta+2s(1-s)(p_2-\pi-\delta)}{p_2}\right)$$

are the sequence averaged potentiation and depression levels and

$$\phi_{1+}(\pi,\delta) = \frac{(p_2-1)!}{\pi!\delta!(p_2-1-\pi-\delta)!}[u^2]^\pi\ [2u(1-u)]^\delta\ [(1-u)^2]^{(p_2-1-\pi-\delta)}$$
$$\phi_{2+}(\pi,\delta) = \frac{(p_2-1)!}{\pi!\delta!(p_2-1-\pi-\delta)!}[u(1-v)]^\pi\ [uv+(1-u)(1-v)]^\delta\ [(1-v)^2]^{(p_2-1-\pi-\delta)}$$
$$\phi_{3+}(\pi,\delta) = \frac{(p_2-1)!}{\pi!\delta!(p_2-1-\pi-\delta)!}[(1-v)^2]^\pi\ [2v(1-v)]^\delta\ [(v^2]^{(p_2-1-\pi-\delta)}$$

the probability distributions.

Bibliography

Abbott, L. F. (2008). Theoretical neuroscience rising. *Neuron*, 60(3):489–495.

Abbott, L. F. & Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3 (suppl.):1178–1183.

Amari, S. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Trans. Comput.*, 21:1197–1206.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Amit, D. J. (1989). *Modelling Brain Functions*. Cambridge University Press, London.

Amit, D. J. & Fusi, S. (1992). Constraints on learning in dynamics synapses. *Network: Computation in Neural Systems*, 3:443–464.

Amit, D. J. & Fusi, S. (1994). Learning in neural networks with material synapses. *Neural Computation*, 6:957–982.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530–1533.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Information storage in neural networks with low levels of activity. *Phys. Rev. A*, 35:2293–2303.

Anderson, J. A. (1970). Two models for memory organization. *Mathematical Biosciences*, 8:137–160.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J., Leonard, B. W., & Lin, L. H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in seuqential stages of hippocampal processing. *Prog. Brain. Res.*, 83:287–300.

Bell, C. C., Han, V. Z., Sugawara, Y., & Grant, K. (1997). Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature*, 387:278–281.

Ben Dayan Rubin, D. & Fusi, S. (2007). Long memory lifetimes require complex synapses and limited sparseness. *Frontiers in Computational Neuroscience*, 1:7.

Bi, G. Q. & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strenght, and post-synaptic cell type. *J. Neurosci.*, 18:10464–10472.

Bliss, T. V. P. & Collingridge, G. L. (1993). A synaptic model of memory: long term potentiation in the hippocampus. *Nature*, 361:31–39.

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

Bredt, D. & Nicoll, R. A. (2003). Ampa receptor trafficking at excitatory synapses. *Neuron*, 40:361–379.

Brunel, N. (1994). Dynamics of an attractor neural network converting temporal into spatial correlations. *Network: Computation in Neural Systems*, 5:449–470.

Brunel, N., Carusi, F., & Fusi, S. (1998). Slow stochastic hebbian learning of classes of stimuli in a recurrent neural network. *Network: Computation in Neural Systems*, 9:123–152.

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

Candes, E. J. & Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51:4203–4215.

Cox, D. R. & Miller, H. D. (1977). *The Theory of Stochastic Processes*. Chapman and Hall, London.

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

Freedman, D. J. & Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–88.

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

Hampson, R. E., Pons, T. P., Stanford, T. R., & Deadwyler, S. A. (2004). Categorization in the monkey hippocampus: A possible mechanism for encoding information into memory. *Proc. Nat. Acad. Sci. USA*, 101(9):3184–3189.

Hebb, D. (1949). *The Organization of Behaviour*. Wiley, New York.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Addison-Wesley, Ma USA.

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

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci. USA*, 81:3088–3092.

Johnson, W. & Lindenstrauss, J. (1984). Extensions of lipshitz maps into a hilbert space. *Contemporary Math.*, 26:189–206.

Jung, M. W. & McNaughton, B. L. (1993). Spatial selectivity of unit activity in the hippocampal granular layer. *Hippocampus*, 3:165–182.

Khalil, H. K. (2002). *Nonlinear Systems*. Prentice Hall, NJ USA.

Klatzky, R. (1980). *Human Memory: structures and processes*. W. H. Freeman & Company.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3:946–953.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Hossein Esteky Keiji Tanaka, ., & Peter A. Bandettini1, . (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60:1126–1141.

Little, W. A. (1974). The existence of persistence states in the brain. *Math. Biosci.*, 19:101–120.

Little, W. A. & Shaw, G. L. (1978). Analytic study of the memory storage capacity of a neural network. *Math. Biosci.*, 39:281–290.

Markram, H., Lubke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275:213–215.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology (London)*, 202:437–470.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London*, 262:23–81.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of ideas immanents in neural activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Mel, B. W. (1994). Information processing in dendritic trees. *Neural Computation*, 6(6):1031–1085. Adapted from Fig. 1: A. Cullheim et al. (1987); Three–dimensional architecture of dendritic trees in type–identified alpha motoneurones; *J. Comp. Neurol.* (*JCN*), 255:85–96. B. courtesy G. Laurent. C. Amitai et al. (1993); Regenerative electrical activity in apical dendrites of pyramidal cells in neocortex; *Cerebral Cortex*, 3:26–38. D. Maslim et al. (1986); Stages in the structural differentiation of retinal ganglion cells; *JCN*, 254:382–402. E. Yang and Yazulla (1986); Neuropeptide–like immunoreactive cells in the retina of the larval tiger salamander: Attention to the symmetry of dendritic projections; *JCN*, 248:105–118. F. Ramon y Cajal (1909); *Histologie du système nerveux de l'homme et des vertébrés*, L. Azoulay, trans. Malaine, Paris; v.1, p. 61. G. Harris (1986); Morphology of physiologically identified thalamocortical relay neurons in the rat ventrobasal thalamus; *JCN*, 254:382–402. H. Greer (1987); Golgi analyses of dendritic organization among denervated olfactory bulb granule cells; *JCN*, 257:442–452. I. Penny et al. (1988); Relationship of the axonal and dendritic geometry of spiny projection neurons to the compartmental organization of the neostriatum; *JCN*, 269:275–289. J. Ramon y Cajal (1909); *Histologie du système nerveux de l'homme et des vertébrés*, L. Azoulay, trans. Malaine, Paris; v.2, p. 902. K. Meek and Nieuwenhuys (1991); Palisade pattern of mormyrid Purkinje cells–a correlated light and electron–microscopic study; *JCN*, 306:156–192. L. Terashima et al. (1986); Observations on the cerebellum of normal–reeler mutant mouse chimera; *JCN*, 252:264–278. M. Sereno and Ulinski (1987); Caudal topographic nucleus isthmi and the rostral nontopographic nucleus isthmi in the turtle; *Pseudemys scripta. JCN*, 261:319–346.

Meunier, C. & Nadal, J. P. (1995). Sparsely coded neural networks. In: *Handbook of Brain Theory and Neural Networks*, page 899. MIT Press.

Mezard, M., Nadal, J. P., & Toulouse, G. (1986). Solvable models of working memories. *J. Physique*, 47:1457–1462.

Mezard, M., Parisi, G., & Virasoro, M. A. (1987). *Spin glass theory and beyond.* World Scientific, Singapore.

Minsky, M. L. & Papert, S. A. (1969). *Perceptrons.* MIT Press, Boston.

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

Nadal, J. P., Toulouse, G., Changeux, J. P., & Dehaene, S. (1986). Networks of formal neurons and memory palimpsests. *Europhys. Lett.*, 1:535–542.

O'Connor, D. H., Wittenberg, G. M., & Wang, S. S. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proc. Nat. Acad. Sci. USA*, 102:9679–9684.

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

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

Parisi, G. (1988). *Statistical field theory*. Addison-Wesley, Ma USA.

Petersen, C. C., Malenka, R. C., Nicoll, R. A., & Hopfield, J. J. (1998). All-or-none potentiation at ca3-ca1 synapses. *Proc. Nat. Acad. Sci. USA*, 95:4732–4737.

Pinel, J. P. J. (2009). *Biopsychology 7th Edition*. Allyn and Bacon.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Rammal, R., Toulouse, G., & Virasoro, M. A. (1986). Ultrametricity for physicists. *Reviews of Modern Physics*, 58(3):765–788.

Ramon y Cajal, S. (1909). *Histologie du systéme nerveux de l'homme et des vertébrés*. Azouly, L. transl. Malaine, Paris.

Rolls, E. (1990). Principles underlying the representation and storage of information in neuronal networks in the primate hippocampus and cerebral cortex. In: *An Introduction to Neural and Electronic Networks*, S. F.Zornetzer, J. L. Davis, & C. Lau, ed., volume Ch. 4, pages 73–90. Academic Press, San Diego.

Rolls, E. T. & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.*, 73:713–726.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York.

Russel, B. & Whitehead, A. N. (1910). *Principia Mathematica*. Cambridge University Press, London.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. (2006). A more rational model of categorization. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, R. Sun & N. Miyake, ed.

Sato, T., Uchida, G., & Tanifuji, M. (2007). The nature of neuronal clustering in inferotemporal cortex of macaque monkey revealed by optical imaging and extracellular recording. In: *34th Ann. Meet. of Soc. for Neuroscience*. San Diego, USA.

Sherrington, D. & Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Physics Review Letters*, 35(26):1792–1796.

Shouval, H. Z., Wang, S. S.-H., & Wittenberg, G. M. (2010). Spike timing dependent plasticity: a consequence of more fundamental learning rules. *Frontiers in Computational Neuroscience*, 4:19.

Sompolinsky, H. (1986). Neural networks with nonlinear synapses and a static noise. *Phys. Rev. A*, 34(3):2572–2574.

Squire, L. R. & Kandel, E. R. (2009). *Memory: from mind to molecules 2nd ed.* Roberts and Company.

Tanzi, E. (1893). I fatti e la induzione nell'odierna istologia del sistema nervoso. *Riv. Sper. Freniatr. Med. Leg. Alienazioni Ment. Soc. Ital. Psichiatria*, 19:419–472.

Tsodyks, M. & Feigelman, M. (1988). Enhanced storage capacity in neural networks with low level of activity. *Europhysics Letters*, 6(2):101–105.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortx by the combination of feature columns. *Nature Neuroscience*, 4:832–838.

Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272:1665–1668.

Weisbuch, G. & Fogelman-Soulié, F. (1985). Scaling laws for the attractors of hopfield networks. *Journal de Physique Lettres*, 46(14):623–630.

Werbos, P. (1974). *Beyond regression: new tools for prediction and analysys in the behavioral sciences*. PhD thesis, Harvard University.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222:960–962.