UNIVERSITÉ PARIS DESCARTES
Laboratoire de Physiologie Cérébrale

THÈSE DE DOCTORAT DE l'UNIVERSITÉ PARIS DESCARTES

présentée par

Alexis Dubreuil

pour obtenir le grade de Docteur de l'Université
Paris Descartes

École Doctorale ED3C Cerveau Cognition Comportement

Sujet de la thèse :

Mémoire et Connectivité Corticale
Soutenance le 1 Juillet 2014
Devant le jury composé de :

Nicolas Brunel	Directeur de thèse
Boris Gutkin	Examinateur
Peter Latham	Rapporteur
Alain Marty	Directeur de thèse
Jean-Pierre Nadal	Examinateur
Alessandro Treves	Rapporteur

Résumé

Le système nerveux central est capable de mémoriser des percepts sur de longues échelles de temps (mémoire à long terme), ainsi que de maintenir activement ces percepts en mémoire pour quelques secondes en vue d'effectuer des tâches comportementales (mémoire de travail). Ces deux phénomènes peuvent être étudiés conjointement dans le cadre de la théorie des réseaux de neurones à attracteurs. Dans ce cadre, un percept, représenté par un patron d'activité neuronale, est stocké en mémoire à long terme et peut être chargé en mémoire de travail à condition que le réseau soit capable de maintenir de manière stable et autonome ce patron d'activité. Une telle dynamique est rendue possible par la forme spécifique de la connectivité du réseau.

Ici on examine des modèles de connectivité corticale à différentes échelles, dans le but d'étudier quels circuits corticaux peuvent soutenir efficacement des dynamiques de type réseau à attracteurs. Ceci est fait en montrant comment les performances de modèles théoriques, quantifiées par la capacité de stockage des réseaux (nombre de percepts qu'il est possible de stocker, puis réutiliser), dépendent des caractéristiques de la connectivité.

Une première partie est dédiée à l'étude de réseaux complètement connectés où un neurone peut potentiellement être connecté à chacun des autre neurones du réseau. Cette situation modélise des colonnes corticales dont le rayon est de l'ordre de quelques centaines de microns. On s'intéresse d'abord à la capacité de stockage de réseaux où les synapses entre neurones sont décrites par des variables binaires, modifiées de manière stochastique lorsque des patrons d'activité sont imposés sur le réseau. On étend cette étude à des cas où les synapses peuvent être dans K états discrets, ce qui, par exemple, permet de modéliser le fait que les connections entre deux cellules pyramidales voisines du cortex sont connectées par l'intermédiaire de plusieurs contacts synaptiques.

Dans un second temps, on étudie des réseaux modulaires où chaque module est un réseau complètement connecté et où la connectivité entre modules est diluée. On montre comment la capacité de stockage dépend de la connectivité entre modules et de l'organisation des patrons d'activité à stocker. La comparaison avec les mesures expérimentales sur la connectivité à grande échelle du cortex permet de montrer que ces connections peuvent implémenter un réseau à attracteur à l'échelle de plusieurs aires cérébrales.

Enfin on étudie un réseau dont les unités sont connectées par des poids dont l'amplitude a un coût qui dépend de la distance entre unités. On utilise une approche à la Gardner pour calculer la distribution des poids qui optimise le stockage de patrons d'activité dans ce réseau. On interprète chaque unité de ce

réseau comme une aire cérébrale et on compare la distribution des poids obtenue théoriquement avec des mesures expérimentales de connectivité entre aires cérébrales.

Summary

The central nervous system is able to memorize percepts on long time scales (long-term memory), as well as actively maintain these percepts in memory for a few seconds in order to perform behavioral tasks (working memory). These two phenomena can be studied together in the framework of the attractor neural network theory. In this framework, a percept, represented by a pattern of neural activity, is stored as a long-term memory and can be loaded in working memory if the network is able to maintain, in a stable and autonomous manner, this pattern of activity. Such a dynamics is made possible by the specific form of the connectivity of the network.

Here we examine models of cortical connectivity at different scales, in order to study which cortical circuits can efficiently sustain attractor neural network dynamics. This is done by showing how the performance of theoretical models, quantified by the networks storage capacity (number of percepts it is possible to store), depends on the characteristics of the connectivity.

In the first part we study fully-connected networks, where potentially each neuron connects to all the other neurons in the network. This situation models cortical columns whose radius is of the order of a few hundred microns. We first compute the storage capacity of networks whose synapses are described by binary variables that are modified in a stochastic manner when patterns of activity are imposed on the network. We generalize this study to the case in which synapses can be in K discrete states, which, for instance, allows to model the fact that two neighboring pyramidal cells in cortex touches each others at multiple contact points.

In the second part, we study modular networks where each module is a fully-connected network and connections between modules are diluted. We show how the storage capacity depends on the connectivity between modules and on the organization of the patterns of activity to store. The comparison with experimental measurements of large-scale connectivity suggests that these connections can implement an attractor neural network at the scale of multiple cortical areas.

Finally, we study a network in which units are connected by weights whose amplitude has a cost that depends on the distance between the units. We use a Gardner's approach to compute the distribution of weights that optimizes storage in this network. We interpret each unit of this network as a cortical area and compare the obtained theoretical weights distribution with measures of connectivity between cortical areas.

Contents

1 Introduction **6**

1.1 Attractor neural network (ANN) and memory 7

1.1.1 Neural networks as dynamical system 7

1.1.2 Dynamical systems and attractors 8

1.1.3 Modeling brain memory processes with ANNs 9

1.1.4 Associative memory and fixed point attractor 10

1.1.5 A brief history of brain modeling with ANNs 13

1.2 Alternative models of short term memory 14

1.2.1 Bistability from single cells properties 14

1.2.2 Working memory with short term plasticity 14

1.2.3 Long-term memory and synfire chains 15

1.2.4 Short term memory and transient dynamics of neural networks 16

1.3 Experimental evidence of ANN dynamics 17

1.3.1 Persistent activity 17

1.3.2 Connectivity in local cortical circuits 18

1.4 The whole cortex as an attractor neural network ? 20

1.4.1 Data on long-range connectivity 20

1.4.2 Connectivity of the entire network of cortical areas 22

1.4.3 Theoretical studies on large-scale ANNs 24

2 Storage capacity of local networks with discrete synapses **26**

2.1 Results on the storage capacity of various models 26

2.1.1 Standard coding and continuous synapses 27

2.1.2 Sparse coding and continuous synapses 29

2.1.3 Discrete synapses 29

2.1.4 On-line learning 31

2.1.5 Beyond binary neurons 31

2.2 Memory capacity of networks with stochastic binary synapses . . . 33

2.2.1 Summary 33

2.2.2 Introduction . . . 33
2.2.3 Storage capacity in the $N \to \infty$ limit . . . 35
2.2.4 Willshaw model . . . 38
2.2.5 Amit-Fusi model . . . 39
2.2.6 Multiple presentations of patterns, slow learning regime . . . 42
2.2.7 Finite-size networks . . . 43
2.2.8 Storage capacity with errors . . . 47
2.2.9 Increase in capacity with inhibition . . . 48
2.2.10 Discussion . . . 49
2.2.11 Methods . . . 54
2.3 Synapses with multiple states . . . 61
2.3.1 Capacity calculation . . . 61
2.3.2 K-states synapses in the single-presentation (SP) learning scenario . . . 62
2.3.3 K-states synapses in the multiple-presentation (MP) learning scenario . . . 65
2.3.4 Discussion on multiple states connections . . . 67
2.3.5 Methods . . . 69

3 Modular networks **78**
3.1 Introduction . . . 78
3.2 Modular networks and storage capacity . . . 80
3.2.1 Connectivity: pre-existing architecture and activity dependent learning . . . 80
3.2.2 Pattern stability and storage capacity . . . 82
3.3 Maximal storage capacity . . . 84
3.3.1 Unstructured macroscopic patterns . . . 84
3.3.2 Patterns organized in categories . . . 88
3.4 Storage capacity with finite size basin of attractions . . . 89
3.4.1 Macroscopic pattern completion . . . 91
3.4.2 Disambiguation . . . 92
3.4.3 Effect of local inhibition on error correction . . . 93
3.4.4 Pattern completion and disambiguation in the same network 94
3.5 Discussion . . . 94
3.6 Methods . . . 99
3.6.1 Distribution of inputs for the unstructured model . . . 100
3.6.2 Distribution of inputs for the categorized model . . . 104
3.6.3 Storage capacity with error correction . . . 106

4 The whole cortex as an attractor network ? 107
4.1 Introduction 107
4.2 Model: Perceptron with a distance constraint 108
4.3 Calculation of storage capacity with the replica method 109
4.4 Distribution of weights 112
4.5 Comparison with experimental data 114
4.5.1 Networks at maximal capacity 114
4.5.2 Networks below maximal capacity 116
4.6 Discussion 118
4.7 Methods 118

5 Discussion 123

Chapter 1

Introduction

Systems neuroscience has been concerned in elucidating how neural circuits can give rise to high-level mental processes like perception, attention, decision-making, motor movements or memory. In this context, theoretical tools allow to formalize mechanistic models of brain circuitry, which can be used to make quantitative predictions about observable quantities. These predictions can be confronted to experimental data in order to assess the relevance of given mechanistic models, thus leading to a better understanding of brain processes.
Theoretical tools have been widely used to propose and test mechanisms of memory formation and memory retrieval. In particular, the proposals that long-term memories are maintained in changes in the synaptic connectivity of neural circuits and that short term memories are maintained by persistent neural activity have been formalized conjointly in the framework of the theory of attractor neural networks (ANNs). This theory has been corroborated by multiple experimental observations, like the recordings of neurons showing persistent activity in animals engaged in working memory tasks. In the present thesis, we will study the memory performance of several ANN models, focusing on various features of network connectivity. Comparison with experimental measures of cortical connectivity at different scales will allow us to discuss which cortical circuits are likely to support ANN dynamics.

In this introduction we will define more precisely ANNs and describe how they have been used to model memory processes in brain networks. We will also review experimental data describing connectivity at the scale of local networks of size of the order of hundred microns, the scale of a few centimeters of cortex, and at the scale of the whole cortex.
In the second chapter, we will focus on models of fully-connected neural networks that can be thought of as modeling the local cortical circuits. In particular we will

quantify the storage capacity of networks with binary synapses subject to different learning scenarios. This will then be extended to two specific models of synapses that can be in $K \geq 2$ discrete states.
In the third chapter, we will study modular networks where each module is a fully-connected network and where the connectivity between modules is diluted. We will show how the memory performance of these networks depends on the connectivity between modules and discuss the idea that cortical circuits involving multiple local circuits can sustain ANN dynamics.
In the last chapter, we will study a network whose units are connected by weights whose amplitude has a cost that depends on the distance between units. We will use a Gardner approach to compute the distribution of weights that optimizes the storage of memory. We will interpret this network as the network of cortical areas and compare features of the theoretical distribution of weights to experimental measures of connectivity between cortical areas.

1.1 Attractor neural network (ANN) and memory

1.1.1 Neural networks as dynamical system

A neural network can be regarded as a dynamical system whose state evolves in time driven by external inputs and constrained by its own dynamical properties, that are determined by the properties of the individual units of the network and the form of the interactions between these units. In order to describe such dynamics, one has to choose variables that characterize the state of each unit. Ideally these variables should relate to experimentally measurable quantities. The activity of neurons or groups of neurons can be characterized in different ways. Common experimental techniques allow to monitor the membrane potential of neurons (via intracellular recordings), and/or their spiking activity (extra-cellular recordings). Many studies of neural networks focus on measures of spiking activity as the standard theory states that neurons communicate with each other through spikes. Spikes are all-or-none events that correspond to a sharp rise and decay of the membrane potential during a time of approximately 2ms. Whether the precise time at which spikes occur matters for the collective behavior of groups of neurons has been much debated. In the work presented here, we assume that it does not for the phenomenon we are trying to model. Thus the variable we use to characterize the state of a neural unit is its firing rate, i.e. the averaged number of spikes emitted during a given time window. The simplest way to model the firing activity of a neuron, and that is used in this thesis, is to use binary $0-1$ variables,

where 0 stands for a neuron firing at spontaneous activity ($\simeq$ 2Hz) and 1 for a neuron firing at an elevated rate (e.g. 20Hz). Although it is a crude way to model the state of a neuron, it allows to gain insight into the properties of biological networks, assuming that their dynamics is mainly governed by the specificity of the connections between units rather than the precise input-output relationship of neurons.
In the above paragraph, we focused on description of neural networks at the level of neurons. In order to grasp brain properties that involve large neural networks (e.g. multiple brain areas), experimentalists have used techniques that allow to monitor the averaged activity of ensembles of neurons (e.g. fMRI, LFP recordings). These techniques do not give access to the activity of single neurons, thus it can be of interest to find relevant variables to describe neural networks at the level of the neural population. One example of such a model is the work by Wilson and Cowan (1972) who introduced what is now called neural field/neural mass models in which a continuous variable $r(\vec{x}, t)$ describes the average activity of a piece of neural tissue located at a position $\vec{x}$ at time t. Another example is a Potts model in which each unit can take several discrete values representing different possible states of an ensemble of neurons, that have been used to describe interactions between cortical areas, for instance see Kropff and Treves (2005).

1.1.2 Dynamical systems and attractors

A neural network can thus be described by an ensemble of N interconnected units (where each unit can model a single neuron, a local network, a brain area, etc.) whose state is characterized by a set of variables. We will consider models where the whole network is described by a vector of length N. The set of all the possible network states, a N-dimensional space, is called the state space of the network. For a network of N binary neurons, its state is described by a vector $\vec{\sigma} \in \{0, 1\}^N$, and the state space is constituted by all the binary vectors of length N, a set of size 2^N.
The evolution in time of the network state, the trajectory of the network, is governed by a dynamic rule that can take the form of a set of differential equations (one for each unit) if time is taken to be continuous or a map if time is discrete. The form of these equations depends on the connectivity between units and the specificity of the input-output relationship of each unit. A simple form of dynamics for a network of binary neurons is the following

$$\sigma_i(t+1) = \Theta\left(\sum_{j=1}^{N} W_{ij}\sigma_j(t) - \theta\right) \text{ for all } i \tag{1.1}$$

where Θ is the Heaviside function, θ an activation threshold and $\mathbf{W}$ is the connectivity matrix that specifies how neurons interact. In words, each neuron i receives an input that is the sum of the activity of the other neurons j's, weighted by the connection strenghts from j's to i. The input-output relationship of these units consists in comparing the input with the activation threshold θ and producing the output 1 if the input is above θ and 0 if it is below.

Given a network defined by its dynamical equation, the trajectory of the network tends to flow towards particular states that are called attractor states. An attractor is a subset of phase space in which the network remains if it enters in. Each attractor A is associated with a basin of attraction $B(A)$. $B(A)$ is defined as a subset of phase space such that if the network is in a state in B it will eventually flow towards states in A. Attractors can consist of a single state (fixed point attractors), a discrete set of states, a continuous set of states (if variables describing units are continuous) like a line or a plane, or more complicated sets. The trajectory of the system in an attractor can be of various types. For instance, if the network visits states of A in a periodic manner, we refer to A as a limit cycle. In the next section, we describe examples in which attractor networks have been used to model memory properties of the brain.

1.1.3 Modeling brain memory processes with ANNs

Fixed point attractors and memory of discrete items

These models rely on the assumption that each item μ is represented by a state of the network $\vec{\xi^{\mu}}$ (e.g. the value of the firing rate of each neuron in the network). A given item μ is said to be memorized if $\vec{\xi^{\mu}}$ is a fixed point attractor of the network (Amit, 1989). These models share properties with human memory like associativity. This will be discussed in more detail in the following.

Line attractor and the memory of eye position

It has been proposed that the ability to hold the eyes still necessitates to form a memory of eye position. There is experimental evidence that neural circuits in the brain stem and cerebellum can perform this task. Electrophysiologists have found neurons that receive transient inputs whose amplitude is proportional to the

amplitude of the saccadic eye movement made by the animal. These neurons respond with a transient modulation of their firing rate that eventually stabilizes at a value linearly related to the eyes position, i.e. they perform an integration of their inputs thus maintaining a signal coding for eye position. Seung (1996) proposed that these neurons form a network with a line attractor where each stable state (in the absence of external input) codes for one position of the eye. Seung explicited on which conditions on the connectivity between neurons the network can sustain such states. In this first model these conditions required an implausible fine tuning of network connectivity. A recent study of a line attractor in a network with excitatory and inhibitory populations has implemented a more robust version of a network with a line attractor (Lim and Goldman, 2013). Line attractors have been used to model other electrophysiological recordings while animals where required to hold in memory the amplitude of a stimulus (Machens et al., 2005).

Plane attractor and 2-D spatial memory

It has been shown that firing patterns of neural activity in the hippocampus of rats form an internal representation of the animal's current location in a given environment (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978; Wilson and McNaughton, 1993). This corresponds to a form of memory as it has been shown that such a representation is indeed internal, since it does not rely on external cues. As the animal is moved from one environment to another one with the same geometry (but with a different lighting for instance), equivalent positions are encoded by different patterns of activity. Circuits of the hippocampus giving rise to such representations have been modeled by plane attractor networks in which each 2-D location (parametrized by 2 numbers) in an environment is represented by a self-sustained pattern of activity (Samsonovich and McNaughton, 1997; Battaglia and Treves, 1998). The set of patterns of activity coding for each location forms a plane attractor. In the network models that have been proposed, multiple such attractors/environments can be embedded in the connectivity of the networks.

1.1.4 Associative memory and fixed point attractor

Properties of associative memories

As discussed above, neural networks with fixed point attractors can be used to model properties of human memory. Two main aspects of memory are captured by these models. Associative retrieval from long term memory, the fact that humans are able to recall previously experienced percepts from a cue signal that share some

resemblance with these percepts (see figure 1.1B). In the framework of ANNs, a previously experienced percept $\vec{\xi}^{\mu}$ corresponds to a network state that is one of the fixed points of the network dynamics (figure 1.1A). The partial cue is considered to set the network in a state that is within the basin of attraction $\vec{\xi}^{\mu}$, and the time required for memory retrieval corresponds to the flow of the network towards the attractor state.

Another related aspect of human memory that can be easily interpreted in this framework is working memory, the ability to maintain a memory of a percept for behavioral purposes. The maintenance of the memory is simply interpreted as the fact that the network remains in the fixed point state corresponding to this memory. Thus one requirement for an ANN to load a percept in working memory is that this percept is a fixed point of the dynamics.

The properties of associative memories have also been described from the point of view of computer science (Knoblauch et al., 2010). In ordinary computers, two distinct modules with different physical substrates are respectively in charge of computations and data storage while in an ANN, computations and data storage are realized by the same physical substrate. This leads to different properties, for instance to retrieve a memory in a ordinary computer, a precise address has to be provided, while an ANN can accept arbitrary queries and associate them with one of the stored memories. Concretely this means that ANNs can be used to categorize inputs. Other tasks that can be performed thanks to the associative property are pattern completion and denoising.

Figure 1.1: *Associative memory with recurrent networks. A- Schematic state space representation of network states. Each state is labeled by a coordinate (x, y), circles represent fixed point states that correspond to particular memories, and lines delimit the basin of attraction associated with each memory. B- Illustration of the associative property of a Hopfield network. The color of each pixel (black or white) is the activity of a neuron ($+1$ or -1). The connectivity matrix is chosen such that network states represented in the right column are fixed points of the network's dynamics. Initializing the network in states that partially overlap with the stored patterns (left column) leads to pattern completion, i.e. memory retrieval. C- Controlling fixed point dynamics via recurrent excitation. Blue connections create a positive feed-back loop in which activity is reverberated, making the represented pattern of activity a fixed point of the network's dynamics. Red connections, that are here to store a different pattern of activity, tend to destabilize the represented pattern. Computing storage capacity in this kind of network consists in quantifying how many stable patterns can be imprinted in the synaptic matrix.*

Implementing attractor dynamics with excitatory recurrent connectivity

A common way to implement attractor dynamics in a neural network, and that is at the root of all the models presented in this thesis, is to use recurrent excitatory connectivity to create feedback loops in which activity is able to self-sustain. This mechanism is illustrated on figure 1.1C where an item is represented by the state

of a network of binary neurons. This item can be stored in long term memory and used in working memory by potentiating the synapses between neurons that are active in this state.

The shaping of recurrent connectivity is supposed to be the result of synaptic plasticity that occurs during a learning phase in which the network is presented the patterns to be learned. The scheme described above can be implemented by Hebbian type learning rules that tend to connect neurons that are simultaneously active, and to disconnect neurons that have opposite activity. Such kind of plasticity rule is consistent with experimental data on plasticity (Bienenstock et al., 1982; Kirkwood and Bear, 1994; Sjöström et al., 2001).

Characterizing the performance of associative memories

The storage capacity can simply be quantified by the number of stable fixed point states P_{max} that can be imprinted in the synaptic matrix of the network. As patterns do not have the same information content, P_{max} is not always the most appropriate measure. To illustrate this point, let consider a network of binary neurons, in which one is willing to store a set of patterns in which each neuron is active independently with probability f (f is referred to as the coding level of the memories). The size of the set of all possible patterns of activity with coding level f increases when f increases from 0 to $\frac{1}{2}$, thus for a fixed coding level observing a given state is more informative at $f = \frac{1}{2}$ than at $f \to 0$. To take this into account, for the example of a fully connected network of N binary units storing independent patterns, the storage capacity can be quantified by the information capacity, measured in bits per synapse

$$i = \frac{P_{max} N(-f \ln_2 f - (1-f)\ln_2(1-f))}{N^2} \tag{1.2}$$

That is the the total number of stored patterns multiplied by the information carried by the observation of a specific pattern among the set of patterns of coding level f ; divided by the total number of synapses in the network, which are the physical substrate on which patterns are imprinted.

Another important quantity is the size of the basin of attraction associated to each pattern, that determines to which extent the memory is associative for these patterns.

The way patterns are learned from the environment is also of interest. Some

models of associative memory do not possess any forgetting mechanism. In this case, if too many patterns are presented to the network, the fixed points of the network eventually become unrelated to any of the previously stored memories (e.g. Amit et al. (1985)). The second chapter of this thesis is devoted to study the storage capacity of simple models with different learning scenarios that have been proposed to circumvent this issue.

1.1.5 A brief history of brain modeling with ANNs

In Hebb's book the *Organization of Behaviour* (Hebb, 1949) it is proposed that neurons could group together to form processing units, or *cell-assemblies*, under the influence of activity dependent synaptic modifications. ANNs are a direct realization of this idea. Willshaw et al. (1969) proposed a network of binary units that exhibits associative memory properties. They showed that their network has remarkable storage capacity, as it allows to store 0.69 bits per binary synapse, which is not too far from the 1 bit theoretical limit for such synapses. Later, Hopfield (1982) introduced a model of associative memory, of which he described the dynamics with numerical simulations, observing "stable limit points" corresponding to the patterns he was willing to store in the synaptic matrix. Importantly he proposed a parallel between his network and models of spin glasses that were studied by physicists at that time. This allowed Amit et al. (1985) to provide a very detailed description of the various stable states of the network, and to compute the storage capacity of the Hopfield network. They confirmed that the number of patterns that could be stored scaled with the number of neurons in the network. More recently Coolen and Sherrington (1993) presented a dynamical theory for the Hopfield model. Such a theory is important as it allows to study 'dynamical' quantities, like the basin of attractions associated to each memory.
Another line of work that makes use of spin glass theory tools was introduced by Gardner (1988). Her work, that will be described in more detail in the last chapter, allows to derive general bounds on the storage capacity of ANNs, independently of the choice of a specific learning rule, as is done in the Hopfield model for instance where the form of the connectivity matrix is given a priori. Such a bound was derived by Cover (1965) for perceptrons with continuous weights using a geometric proof. The interest of Gardner's method resides in its generality, many assumptions on the statistics of patterns to store or on the nature of synaptic weights can be explored (Gardner, 1988; Gardner and Derrida, 1988; Gutfreund and Stein, 1990). Results obtained with this method are presented in chapter 2.

Since these early studies, one line of research in computational neuroscience has consisted in including more biologically realistic features in these models in order to confront with experimental observations the hypothesis that memory and associative properties of brains are realized by ANNs.

1.2 Alternative models of short term memory

In the models we have discussed so far, a special form of the connectivity between neurons creates "cell-assemblies" in which activity can be self-sustained. In this framework neurons can be observed in two states, a spontaneous state of activity or an enhanced state of activity if they belong to the cell assembly in which activity is reverberated. This enhanced activity allows to interpret the observed persistent activity during memory tasks (see section 1.3.1). The most common proposal to account for the form of the synaptic matrix leading to such a behavior, is to assume that previous presentations of stimuli to be learned by the network lead to Hebbian long-term plastic changes (see Wang 2001 for a review). Here we present models based on different principles that have been proposed to model memory related phenomena.

1.2.1 Bistability from single cells properties

An alternative mechanism to explain persistent activity is to postulate that individual cells are bistable. Such bistability could be achieved by positive feed-back loops created by active ionic currents (Hodgkin and Huxley, 1952; Booth and Rinzel, 1995; Delord et al., 1997; Loewenstein and Sompolinsky, 2003). A hybrid mechanism allowing bistability thanks to a combination of after-depolarization currents (observed in pre-frontal pyramidal cells under the influence of neuro-modulators like acetylcholine or serotonin) and external oscillatory inputs has been proposed by Lisman and Idiart (1995). Such models have the advantage of being able to maintain in persistent activity arbitrary patterns of activity, even those that have not been previously presented to the network, however, without recurrent connections these models do not have associative properties.

1.2.2 Working memory with short term plasticity

Mongillo et al. (2008) have proposed that working memory could be achieved by synaptic facilitation (a form of short term synaptic plasticity), a phenomenon observed in pre-frontal areas. A pattern of activity is presented to the network, synapses between active neurons are facilitated (increase of their efficacy to excite the post-synaptic cell) for a time of the order of 1s after the initial presentation. Their network can operate in different regimes depending on the level of unspecific (i.e. with a uniform spatial distribution) background activity exciting the network after an object has been presented. If the level of background activity is low, the elevated spiking activity elicited at pattern presentation goes away but the pattern can still be read out during a time of the order of a second using an unspecific excitation signal that preferentially reactivates neurons connected via facilitated synapses (figure 1.2A). If there is an intermediate level of the unspecific background excitation after stimulus presentation, the network becomes bistable, i.e. the sub-population that was initially stimulated is periodically reactivated with a period controlled by the depression time constant of the short term plasticity model (figure 1.2B). If the level of maintained unspecific excitation is increased further, and if the recurrent connections are high enough, the network exhibit an asynchronous state of persistent activity (figure1.2C) which is similar to the classical states sustained by recurrent excitation alone. Note than in all of the three regimes they studied, there are recurrent 'long-term' connections between neurons coding for the same object. It is not clear to what extent this pre-training is necessary, but the existence of these three regimes is interesting as a flexible manner of memorizing objects for short periods of time. Indeed, the observed flexibility of working memory is not accounted for by the basic ANN framework, but note that recently Dipoppa and Gutkin (2013) proposed oscillatory mechanisms to allow for a flexible control of working memory in ANNs.

Figure 1.2: *Working memory with short term plasticity. Raster plot of spiking activity in a spiking network of excitatory and inhibitory neurons with short-term plasticity. Black dots are spikes emitted by neurons excited by the stimulus presented at $t = 0$, green dots are spikes emitted by neurons non-selective for the stimulus. The time course of the short term plasticity parameters of the stimulus-selective neurons are shown by the red and blue lines. A- Regime with low background excitation. Spiking activity decays after stimulus presentation, a population spike coding for the stimulus is elicited when a brief non-specific input (second grey shading) stimulates the network. B- Regime with intermediate background excitation, population spikes coding for the initial stimulus are emitted regularly at a frequency controlled by the short-term plasticity parameters. C- Regime with high background excitation and high level of recurrent connections between cells coding for the stimulus. In this case selective neurons fire asynchronously at an elevated firing rate after stimulus presentation.*

1.2.3 Long-term memory and synfire chains

Feedforward networks are also at the root of a theory of neural coding that is different from the coding scheme on which ANNs rely (Abeles, 1991). In the so called synfire chains, percepts are coded by sequential transient activations of specific pools of neurons. Here spike timing is fundamental as opposed to models with reverberating activity in which items are assumed to be coded by spatial patterns of firing rates. Whether neural networks can learn to generate reproducible sequences of activity has been much debated. Spike-timing dependent plasticity seems a good candidate to achieve such learning. Studies using STDP were able to produce only short and closed sequences even with fine tuning of parameters (Hertz and Prügel-Bennett, 1996; Diesman et al., 1999; Levy et al., 2001; Izhikevich et al., 2004). In a more recent study, Liu and Buonomano (2009) using both STDP and an homeostatic learning rule, were able to learn up to 5 sequences in a network of 500 neurons where each neuron is activated once in the sequence (figure 1.3). Each sequence lasts for $100 - 200$ms, and is initiated by the activation of a specific subset of neurons, which can be seen as a sort of memory retrieval in which the initial activation is a cue and the sequence of activity the retrieval of a specific memory. To what kind of memory it could correspond is unclear. It is tempting to say it would be a good mechanism to store percepts that involve precisely timed sequences like auditory sequences or motor sequences.

Figure 1.3: *Long-term memories retrieved by sequential activation of neurons in a recurrent network. A network of* 400 *excitatory and* 100 *inhibitory spiking neurons is stimulated with a pulse of activity targeting a specific subset of excitatory neurons. After* $\simeq 400$ *such stimulations, thanks to the combination of a spike-timing-dependent-plasticity and a homeostatic plasticity rule, the connectivity matrix organizes such that stimulations lead to the generation of sequences of activity in which each excitatory neuron fire* 1 *or* 2 *spikes at a precise time. Each sequence of activity last for a duration of* 100*ms. Only up to 5 such sequences can be stored. A- Raster plots where neurons are ordered according to the moment they are activated in the sequence of activity triggered when a subset A of the excitatory neurons are briefly stimulated. B- Same for an ordering determined by the stimulation of a subset B.*

1.2.4 Short term memory and transient dynamics of neural networks

To maintain information about sequences of stimuli, models based on a different principle have been developed in the last ten years. Instead of storing the memory of stimuli in the synaptic matrix of neural networks, the memory is stored in the activity of neural networks. In the framework described in Maass et al. (2002), a first neural network is seen as a medium excited by a stream of stimuli, the transient dynamical state of the network carrying information on the sequence of inputs. Other ensembles of readout neurons can then be trained to extract task

relevant information from this transient dynamics. Maass et al. (2002) have described under what general conditions on the medium and on the readout neurons this process is efficient to perform non trivial computations on sequences of inputs. The ability of the medium to embed information about the history of the signal has been quantified in particular models (White et al., 2004; Ganguli et al., 2008; Lim and Goldman, 2012), in which a network of N neurons has to build the memory of a time-varying scalar input presented to the network. It has been shown that feed-forward structure are optimal to perform such tasks (Lim and Goldman, 2012), this is because as past signals propagate through the network, they avoid interfering with signals injected more recently. An illustration of this principle is shown in figure 1.4. But recurrent networks with a hidden feed-forward architecture could also subserve such dynamics (Goldman, 2009). However, it is not clear what kind of recurrent connectivity matrix is well suited to implement this mechanism.

Figure 1.4: *Short term memory and transient dynamics of neural network. Example of a neural integrator. A- A feed-forward architecture whose activity is read-out by a single neuron is well suited to perform such a task. B- The first neuron in the feed-forward chain is excited by a brief pulse of activity, the activity is then propagated along the chain. If the weights to the read out neuron are well chosen, it can produce an output proportional to the amplitude of the pulse for a time proportional to the number of neurons in the chain.*

1.3 Experimental evidence of ANN dynamics

1.3.1 Persistent activity

A first neural correlate of memory maintenance predicted by fixed point ANNs is persistent activity (PA). It has been observed with single cell recordings in the prefrontal and temporal cortices of primates, and also in the basal ganglia, the superior colliculus and brainstem. It has also been observed in the thalamus, hippocampus and midbrain of rodents, and in the brainstem of non-mammalian vertebrates (Major and Tank, 2004).
A standard memory task during which PA has been observed is the delay match to sample (DMS) task. DMS tasks consist of three epochs (see figure 1.5A). First the monkey is presented an image on a screen. Then there is a delay period with a blank screen during which the monkey has to remember the shown image. In the third epoch the monkey is presented a new image and has to decide whether it is the same image as the one presented in the first epoch. An example of cells in prefrontal cortex exhibiting PA during a DMS task is shown in figure 1.5B.

An alternative interpretation of persistent activity has been proposed in the

Figure 1.5: *Persistent activity observed during a working memory task. A- The Delay-Match-to-Sample (DMS) task has been widely used to study the neural correlates of working memory in primates. The monkey fixates a screen where a first image (sample) is presented for a brief time. During a delay of the order of the second it has to remember this image. Then a test image is presented and the monkey has to signal with a saccade whether the test image matches the sample. B- Example of neural activity in a temporal cortex neuron showing persistent activity during a DMS task on a familiar stimulus (taken from Miyashita (1988)). This neuron is recorded during multiple repetition of the task as shown by the raster plot on the upper part of the panel, and the histogram shows the average activity of the cell during the different epochs of the DMS task. C- Same as B when the monkey is tested with a non familiar stimulus.*

framework of memory models based on transient dynamics. Goldman (2009) proposed that neurons exhibiting persistent activity are neurons reading out (with appropriate weights) the activity of a feedforward network in which the image presented in the first epoch is propagating (see figure 1.4). During DMS tasks the firing rate of single units can exhibit various trajectories. These can be accounted for by ANN models by assuming they are the result of phenomena like neural fatigue. However one experiment put forward by Goldman (Batuev et al., 1979) shows recordings in which some cells are transiently activated in the middle of the delay period, which is easier to interpret with his theory (such cell is a readout neuron receiving input only from neurons in the middle of the feed-forward circuit) than with the theory of ANNs. Also, Harvey et al. (2012) have shown that some cells in the parietal cortex of rats are sequentially and transiently active during the delay period of a memory task, with different sets of cells active for trials leading to different behavioral choices. A read-out with fine tuned connectivity from these cells could generate persistent activity.

Exactly which brain circuits can implement ANN dynamics is a matter of debate. For instance is it only present in the prefrontal areas of cortex ? It has recently been shown that even in the auditory cortex of mice, neural responses exhibit ANN like dynamics in a task involving sound discrimination (Bathellier et al., 2012). It has also been proposed that attractor like dynamics can result from choosing synaptic weights that allow an optimal representation of a distribution of input stimuli in the presence of high level of noise (Tkavcik et al., 2010).

1.3.2 Connectivity in local cortical circuits

The different memory mechanisms described above predict very different dynamical behavior for the networks involved. As dynamical properties are shaped by connectivity matrices, it can be relevant to give a close look at these data. We will present data on the connectivity of local cortical networks, and then discuss

theoretical studies that relate these measures with memory models.

Experimental data on local cortical connectivity

Current techniques to estimate the strength of the connectivity between two neurons consist in recording intra-cellularly potential post-synaptic neurons while triggering spikes from pre-synaptic neurons. The measure of the amplitude of excitatory/inhibitory post-synaptic potential/current allows to quantify the synaptic strength between the two neurons. One crude way to analyze these connectivity measure is to describe a synapse by a binary variable (absent or present), multiple pair recordings can then give the probability of connections between two neurons as well as the probability to find a pair bidirectionally connected. A less crude way is to consider the distribution of synaptic weights as defined by the amplitude of the measured voltage/currents elicited in the post-synaptic cell.

Connections from excitatory to excitatory cells

Different studies in different cortical areas and different layers have been carried, and they all find a low connection probability $< 15\%$ (see figure 1.6 for some example). The connection probability seems not to depend on the distance between the two recorded neurons for short distances ($< 100\mu m$), at least in layer 5 of visual cortex as shown in figure 1.7A, but decays for larger distance as shown in figure 1.7B. In figure 1.8 blue histograms show distributions of synaptic weights obtained in nine different studies.

In all the studies an over representation of bidirectional motifs (pair of neurons reciprocally connected) was found compared to what it would be in a random graph with the same connection probability. These results are summarized in figure 1.6. In the study by Song et al. (2005), they mentioned that this over representation of bidirectional motifs could originate from non-uniform probability of connections but checked that connection probability was not sensitive to horizontal distance, vertical distance, and ruled out a few experimental artifacts. They also recorded triplets of cells simultaneously and were able to show that some 3-neurons motifs were over-represented compare to a network whose 2-neurons motifs are randomly drawn in numbers consistent with 2-neurons motifs observed experimentally.

Figure 1.6: *Connection probability and over-representation of bidirectional motifs in three studies reporting results of intracellular recordings in pairs of excitatory neurons. P_{E-E} is the probability to find a connection present from one neuron to another. The last column reports the fraction of pairs connected bi-directionally, the number in parenthesis is what would be observed in a network where connections are randomly active with probability P_{E-E}.*

Figure 1.7: *Intracellular recordings of pairs of neurons allow to describe some features of local cortical connectivity. A- Connection probability between excitatory pyramidal cells as a function of distance between the neurons in layer 5 of visual cortex. B Connection probability between excitatory pyramidal cells as a function of distance between the neurons in layers 2/3 of visual or somato-sensory cortex. Bars are connection counts and black line is the connection probability. C- Connection probability from an excitatory pyramidal cell to an inhibitory cell (black squares) and from an inhibitory to an excitatory cell (black circles) as a function of distance between the neurons in layer 2/3 of visual or somato-sensory cortex.*

Connections from inhibitory to excitatory and excitatory to inhibitory cells

In Holmgren et al. (2003) they also recorded pairs of excitatory/inhibitory neurons. Results are summarized in figure 1.7C. The probability of connections is higher than for excitatory-excitatory connections, and the decrease with distance is slower. The occurrence of bidirectional motifs is also reported in figure 1.7.

Interpretation of local cortical connectivity

Barbour et al. (2007); Chapeton et al. (2012) have reported that these connection probabilities of local cortical networks can be interpreted as a signature that they operate as ANNs maximizing the number of memories they store. As mentioned in a previous section, the method introduced by Gardner allows to compute the maximal storage capacity of ANNs, moreover this technique can be used to compute statistical properties of the synaptic matrix that achieves maximal capacity (Brunel et al., 2004). This method will be described in more details in the last chapter. Chapeton et al. (2012) considered networks of excitatory and inhibitory neurons and found that at maximal capacity the probability that two excitatory neurons are connected is $< 50\%$ while the connection from an inhibitory to an excitatory neuron is $> 50\%$, as observed in experimental data. Moreover experimental distributions of non-zero synaptic weights can be fitted by theoretical predictions as shown in figure 1.8.

Figure 1.8: *Blue histograms: experimental distribution of PSP amplitude measured in pair recordings in* 9 *different studies. Green lines: experimental distributions are well fitted by distributions of synaptic weights that optimize storage capacity in a network of excitatory and inhibitory neurons.*

1.4 The whole cortex as an attractor neural network ?

These interpretations of local cortical circuitry suggest that these circuits can, by themselves, sustain PA and behave as ANNs. Other proposals have suggested that ANNs could be implemented in circuits involving multiple brain areas like thalamo-

cortical loop, which is consistent with the observation of PA in thalamus during the delay period of memory tasks (Wang, 2001; Fuster and Alexander, 1971). Another proposal (Braitenberg and Schütz, 1991), complementary to the idea that local cortical areas can sustain PA by themselves, is that long range connections between different cortical columns/areas can be used to reverberate activity between these areas and imprint stable states in the dynamics of the whole cortex. We will present recent data on long-range cortical connectivity and discuss previous theoretical work on this idea.

1.4.1 Data on long-range connectivity

The previous section focused on describing the statistics of connectivity between pairs of neurons distant from less than $200\mu m$ horizontally. Here we present data describing connectivity on larger scales.

Short-range versus long-range connectivity

Stepanyants et al. (2009) reconstructed neurons from multiple sections of cat visual cortex and by counting synaptic terminals on axons and dendrites, they were able to estimate the fraction of connections whose pre-synaptic neurons are located inside a cylinder of radius R centered on the post-synaptic neuron (see figure 1.9A). For $R = 200\mu m$ (the size of an iso-orientation column) they found that only 18% of the synaptic contacts onto excitatory neurons originated in the cylinder (see figure 1.9) ; 34% for $R = 800\mu m$ (the size of an occular-dominance column) ; 36% for $R = 1,000\mu m$. The method they use do not allow to know where the non-local, longer-range connections originate.

Figure 1.9: *Estimation of the proportion of short-range versus long-range connections. A- Stepanyants et al. (2009) have estimated the number of synaptic terminals whose pre-synaptic cell is outside/inside a cylinder of radius R centered on the post-synaptic neuron. B- Fraction of local synapses whose pre-synaptic neuron is excitatory or inhibitory.*

Patchy connectivity in pre-frontal cortex

Another common technique to study brain connectivity is to inject tracers into a local patch of cortex. These tracers are viruses that propagate between neurons and label them. They can be retrograde, if they travel from the injected neurons to its pre-synaptic neurons or they can be anterograde if they travel from the injected neurons to their post-synaptic neurons. As it is difficult to inject into a single neuron, the measures of connectivity obtained from these experiments concern

the patches of neurons that are affected by the initial injection of virus (typically patches of radius $\simeq 0.2 - 1mm$). Pucak et al. (1996) performed such injections in the pre-frontal areas 6 and 46 of monkeys and described the connectivity. For both anterograde and retrograde tracers, they found that labeled neurons were clustered in groups of $\simeq 0.3 \times 1.5mm$ that often take the form of stripes. An example of injection and reconstructed patches of labeled neurons is shown in figure 1.10A. They were able to distinguish patches that are connected to the injection sites via grey matter connections ('intrinsic patches') from patches connected through white matter ('associative patches'). They spotted an average of $\simeq 15$ intrinsic patches per injection (12 on average for 5 anterograde labeling and 17 on average for 4 retrograde labeling) that were located between 0.8 and $6mm$ from the injection site. For the 5 anterograde injection, they spotted 47 associative patches, 42 of which were further than $8mm$ from the injection site and 39 patches at similar distances for the 4 retrograde injections. These patches can be located in different brain areas, and have sizes similar to the intrinsic patches. Some associational patches together with intrinsic patches are reconstructed in figure 1.10B. 4 of the injections lead to both anterograde and retrograde labeling which lead the authors to claim that these stripes are connected reciprocally for intrinsic patches. They have suggested that associational patches are also connected in a bi-directional manner. Similar experiments have also reported patchy connectivity in sensory cortices (DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Bosking et al., 1997).

Figure 1.10: *Patchy connectivity in pre-frontal cortex. A- Labeled neurons in area 46 of a macaque monkey following injection of retrograde tracers at the border between areas 46 and 9 (left part of the image). B- Reconstruction of the intrinsic patches (connections between the injection and source neurons are via grey matter connection) of labeled neurons following an injection of anterograde tracers in area 9 (white circle). C- Reconstruction of intrinsic patches (black) and associative patches (black-white stripes, connected to the injection site via white matter connections) following an injection of tracers in area 9 (white circle).*

1.4.2 Connectivity of the entire network of cortical areas

Description of the experiments. The study presented above reported some connections between neighboring brain areas, but until recently, experimental data characterizing brain connectivity at the level of brain areas have been rather scarce. Anatomical experiments reported only the presence or absence of connections between brain areas (Felleman and Van Essen, 1991), leading to a description of large-scale connectivity with binary connectivity matrices. The recent work of Markov et al. (2011, 2012) allows a quantitative description of brain connectivity, and in particular cortical connectivity. The principle of their experiments is sim-

ilar to the study mentioned above, after injecting retrograde tracers (viruses that travel from a neuron to its presynaptic neurons) in a given "target" cortical area, they can count the number of neurons that project to this areas (see figure 1.11). Based on anatomical considerations they delimit the cortex of macaque monkeys in 91 areas and inject retrograde tracers in 29 of these areas. After slicing the brain, they count the number of neurons labeled in each area for one every three slice.

Figure 1.11: *Study of large-scale connectivity via tracers injections in cortical areas of macaque monkey. A-B- Markov et al. (2011) performed injections of tracers in 29 out 91 cortical areas delimited based on anatomical considerations. C- For each injection, after slicing the brain they count the number of retrogradely labeled neurons in each brain areas allowing them to establish the connectivity profile of each injected area.*

Quantification of connectivity. After counting the number of source neurons projecting to each injected target area, they quantify the projection strength from a source area (S) to a target area (T) with the total fraction of labeled neurons $FLNt$, which is the ration between the number of labeled neurons in area S and the total number of labeled neurons in the whole brain ; and with $FLNe$, the ratio between the number of labeled neurons in area S and the total number of labeled neurons in cortex minus those labeled in area T. In the following we also refer to $FLNe$ as weights from (T) to (S).

Origin of long-range connections. For 5 injections in visual cortex and one in prefrontal cortex, they measured $FLNt$ for each source areas including subcortical areas. Results are shown in figure 1.12A, which shows that the majority, 79%, of neurons targeting at least one of the neuron initially infected by the virus originate from the area in which the injection is made ; 16% come from neurons in neighboring cortical areas, 5% from other cortical areas and only 1% from subcortical areas.
They also measured sizes of cortices around the injection site with given $FLNt$. Results are shown in figure 1.12B. From this figure, connectivity seems far more local than shown in the previous study by Stepanyants et al. (2009). Many reasons could explain such discrepancy, including methodological considerations.

Figure 1.12: *Origin of long-range connections. Report from* 5 *injections in visual cortex and one in prefrontal cortex. The FLNt measure the strength of the connection from a brain area to the injected area (see text for details). A- Percentage of labeled neurons that are intrinsic to the injected area, in an area that touches the injected area (short), in other cortical areas (long), in subcortical areas (SC). B- Fraction of labeled neurons that are at a given distance from the injection site for* 6 *different injections.*

Details of cortical projections. Injections in 29 of the 91 areas, allow to build a 29×91 weighted connectivity matrix describing the connections from the whole cortex to the 29 injected areas (figure 1.13A). It also allows to build a 29×29 matrix W which fully characterizes the connections of a subset of cortex. This matrix is shown in figure 1.13B. Note that in this matrix, the weight W_{ij} from area i to area j is obtained from one injection of tracer, while the weight W_{ji} from j to i is obtained from another injection. As the $FLNe$ depends on the total number of labeled neurons projecting to a given area, one has to be careful when comparing weights obtained from different injections.

Figure 1.13: *Connectivity profiles between cortical areas. A- Connectivity matrix obtained from injections in* 29 *out of* 91 *delimited cortical areas. B- Full connectivity matrix for the network composed of the* 29 *injected areas. C- Distribution of connection weights between cortical areas. Blue line is a log-normal fit.*

Distribution of cortical connection weights. For each injection the obtained weights range approximately between 10^{-1} and 10^{-5}. The distribution of weights is well fitted by a log-normal distribution (if X is log-normally distributed, then log(X) is normally distributed), as shown in figure 1.13C. Also the weights from areas neighbors of the target area are more strongly connected to it than more distant areas, as shown in figure 1.14B.

Binary features of the cortical connectivity matrix. The authors of this study also carried extensive analysis of the binary connectivity matrix W_b of the 29×29 matrix W, and its associated graph $\mathcal{G}_b$. They found a density of 64%, that is 64% of the entry of W_b are equal to one. The probability to find a link from one area to another decreases with distance, as shown by the red line in figure 1.14A. They characterize the degree of symmetry of $\mathcal{G}_b$, it can be done by looking at the occurrence of the possible connection motifs between two areas. The motif "bidirectional connectivity" is observed for 53% of the pairs of areas, for a binary random graph where each link is 'on' with probability 0.66 this motif would on average occur for 44% pairs. The "unidirectional connectivity" motif is observed for 27% of the pairs (45% for the random graph), and the motif with no connection for 20% of the pairs (11% for the random graph). They have also described the three neurons motifs.
In this study they have shown that the main properties (e.g. over representation of bidirectional motifs, of some 3-neurons motifs, graph spectra) of the binary connectivity matrix are well accounted for by a model in which the probability to find a connection between two neurons decreases exponentially with the distance

between them (Ercsey-Ravasz et al., 2013).

Figure 1.14: *Dependence of connectivity features with distance. A- Histogram shows the distribution of distances between brain areas as measured by the distance to travel from one area to another through the white-matter ; red line shows how the probability to find at least one connection from one area to another decay with distance. B- Each point corresponds to the value of the connection strength from one area to another separated by a distance d.*

1.4.3 Theoretical studies on large-scale ANNs

The possibility of using long-range connections to sustain activity between distant brain areas was investigated by O'Kane and Treves (1992) who studied a network made of modules with dense connectivity between neurons inside the same modules and diluted long-range connections between neurons belonging to different modules. The ratio between the number of short and long range connections a neuron receives was taken of order one according to experimental data provided by Braitenberg and Schütz (1991). In this first model, long-ranged connections onto a neuron could originate from any of the other modules ; each pattern of activity to be stored involved specific patterns of activity in all the modules. This model suffered from two main issues. First the storage performance of the network was not satisfactory in the sense that the number of stored patterns was not increasing with the number of modules in the network. Second the authors discovered the presence of 'memory glass' states that were stable states co-existing along with the patterns to be stored. These memory glass states are states in which local patterns of activity in each module are patterns of activity corresponding to different global patterns.

In a following paper Mari and Treves (1998) solved these two issues. First they stored patterns in which only a fraction F of modules are active. They showed that the number of stored patterns can scale with F. Thus, having F decreasing with the number of modules, the number of stored patterns can increase with network size. Second, instead of distributing long-range connections between all modules, they focused long-range connections only between a limited number of pairs, such that each module is connected to a finite number of modules. This modification allows to increase the relative drive of long range connections and to destabilize the memory glass states.

In the third chapter of this thesis similar modular networks are studied and are confronted with data on long-range cortical connectivity.

Chapter 2

Storage capacity of local networks with discrete synapses

In this chapter, we focus mainly on the storage property of fully connected networks. For such connectivity it is assumed that potentially every neuron can be connected to any other in the network, the actual connectivity being shaped solely by the patterns of activity to be learnt. Models of fully connected networks seem to be well suited to describe cortical networks with dimensions $\leq 150\mu m$, as supported by the study of Kalisman et al. (2005) showing that -in layer 5 of the rat's somato-sensory cortex- an axon originating in such a network touches all the neighboring dendrites (without necessarily forming a functional connection) without any bias. More roughly, fully connected networks can model networks of size $\leq 500\mu m$ for which the probability that two neurons touch each others varies smoothly from 0.8 to 0.1 with increasing distance, as shown by Hellwig (2000) -for neurons in layer 2/3 of the rats visual cortex. A cortical network of this size contains approximately $10,000$ neurons.
In the first section we summarize various previous results on the storage capacity of networks of binary neurons. In the second section, we present a study of the storage capacity of networks with binary synapses and on-line learning rules in which synapses are updated stochastically upon pattern presentation. This work is currently under-review for publication, a summary is provided at the beginning of the section. In the third section, we generalize this study to the case where synapses can take not only 2 but K discrete values.

2.1 Results on the storage capacity of various models

Since the introduction of the first ANNs, multiple models with increasing levels of biological realism have been studied in order to estimate the relevance of the ANN

idea to model brain functions. As discussed in the introduction, an important quantity to assess the viability of a given model is to compute its storage capacity. Two main approaches have been taken in this direction. One is to use Gardner's method (Gardner, 1988) in order to compute theoretical bounds on storage capacity of ANNs. The principle of this method is to explore the space of all possible network connectivities, and to measure the volume of the sub-space corresponding to connectivities that satisfy the learning problem (i.e. all P patterns being fixed points of the dynamics of the network). Intuitively, when more patterns are required to be stored, this volume should get smaller. The maximal storage capacity is then given by the number of patterns for which this volume becomes zero. We refer the reader to chapter 4 for a more explicit example. The power of this method is to be general enough to study a large variety of models. For instance in her original paper, Gardner computed the maximal storage capacity for networks of binary neurons, with continuous synapses storing patterns with arbitrary coding level. Note that this method does not provide an explicit connectivity matrix satisfying the learning problem.

A second approach, is to choose a specific rule ('learning rule') that determines the synaptic matrix for a given set of memories to store. The performance of a given learning rule can then be assessed by comparing the resulting storage capacity with the corresponding theoretical bound established with Gardner's approach. In the following we present results obtained in networks of binary neurons with various constraints on the nature of the synaptic connectivity and on the distribution of patterns to be stored. Note that often in these works, the problem of learning a set of P fixed points in an ANN is reduced to learning a set of P input-output associations in N perceptrons.

We then discuss more recent works using ANNs of spiking neurons.

2.1.1 Standard coding and continuous synapses

Here the focus is on networks of binary neurons (whose activity is described by a set of binary variables $\sigma_i \in \{-1; +1\}$, $i = 1...N$ where N is the size of the network) connected through real value synapses. The patterns of activity $\vec{\xi}^{\mu}$ that are stored in the synaptic matrix are drawn randomly and independently with a 'standard' coding level $f = \frac{1}{2}$. That is for each pattern μ the activity of each neuron i is given by

$$\xi_i^\mu = \begin{cases} 1 & \text{with probability } f = \frac{1}{2} \\ -1 & \text{with probability } 1 - f = \frac{1}{2} \end{cases} \tag{2.1}$$

With a network dynamics

$$\sigma_i(t+1) = sign(\sum_{i=1}^{N} W_{ij}\sigma_j(t)) \tag{2.2}$$

where $W_{ij} \in [-\infty, +\infty]$ is the connection from neuron j to neuron i. The problem of finding a connectivity matrix W such that P given patterns are fixed points of this dynamics can be reduced to the problem of learning P input-output associations $(\vec{\xi^\mu}, \xi_{i_0}^\mu)$ in N perceptrons. It is known since the work of Cover (1965) that in this framework a maximum of $P = 2N$ patterns (or input-output associations) can be learned in the limit $N \to +\infty$. The same result has been derived using Gardner's approach (Gardner, 1988).
Hopfield (1982) has proposed an explicit learning rule to store a set of P such patterns with real value synapses. The symmetric synaptic matrix is related to the patterns to store by

$$W_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{2.3}$$

This synaptic matrix reflect the idea of Hebb that neurons that have similar activity in a given pattern tend to be connected together and neurons with dissimilar activity tend to be disconnected. Because the synaptic matrix is symmetric, standard spin glass tools can be used to study its storage capacity. It has been shown by Amit et al. (1985) that for this explicit learning rule, the maximal number of patterns that can be stored is $P = 0.14N$. A study by Sompolinsky (1986) has shown that similar performance is obtained when perturbing this learning rule while keeping synapses symmetric, for instance when: introducing a symmetric noise η_{ij} at each synapse $W_{ij} = W_{ij} + \eta_{ij}$; clipping each weights $W_{ij} = sign(W_{ij})$; diluting the connectivity $W_{ij} = c_{ij}W_{ij}$ where c_{ij} is zero or one with probability $c = O(1)$. Remarkably he has found that the ratio of the number of patterns that can be stored to the number of available synapses is increasing with dilution. A similar observation has been made for extreme asymmetric dilution ($c_{ij} \neq c_{ji}$ and $c \ll \frac{\ln N}{N}$) for which the maximal number of patterns that can be stored is $P = 0.64C$ (Derrida et al., 1987), where $C \times N$ is the total number of synapses in the network after dilution. These studies on dilution were motivated by the fact that the Hopfield model had the unrealistic feature that every neuron is connected

to every other in the network. Many other features of the Hopfield network are unrealistic. For instance there is no distinction between excitatory and inhibitory neurons (in particular Dale's law is not verified), binary units is a crude way to model real neurons that emit spikes, synapses are assumed to have an infinity of stable states, and the hypothesis of standard coding is not corroborated by electrophysiological recordings which suggest that only a small fraction of the neurons are active when coding for a given memory. The next section focuses on storing patterns with biased coding levels $f < \frac{1}{2}$.

2.1.2 Sparse coding and continuous synapses

For patterns with a coding level f, Gardner computed the storage capacity and in particular found that for small f, the maximal number of stored patterns is

$$P = \frac{1}{2f|\ln f|}N \tag{2.4}$$

As mentioned in the introduction, patterns with a smaller coding level are less informative, which can be accounted for by quantifying performance using the information capacity

$$i = \frac{PI_{pattern}}{N^2} \text{ with } I_{pattern} = N\left(-f\log_2 f - (1-f)\log_2(1-f)\right) \tag{2.5}$$

for a fully connected network with N^2 modifiable synapses in which P patterns of coding level f are stored as fixed points of the dynamic. The information capacity of such network is found to decrease with the coding level from $i = 2$bits/synapse at $f = 0.5$ to $i = 0.72$bits/synapse at $f \ll 1$.
An adaptation of the rule used by Hopfield (2.2) to sparse patterns $f \ll 1$ has been studied by Tsodyks and Feigel'man (1988). Instead of using $\{-1; +1\}$ neurons he used $\{0, +1\}$ neurons governed by the dynamics

$$\sigma_i(t+1) = \Theta\left(\sum_{j=1}^{N} W_{ij}\sigma_j - \theta\right) \tag{2.6}$$

where θ is an activation threshold. This model has the advantage that the state $\vec{\sigma} = \vec{0}$ is always stable, which would correspond to the spontaneous state of the network. Remarkably they showed that for a well chosen threshold the maximum number of patterns that can be stored reaches the upper bound (2.4) derived by Gardner.

2.1.3 Discrete synapses

It has been argued that it would be difficult to build robust synapses that can take continuous values with known biophysical processes (see Brunel 2003). Also for practical applications, discrete states synapses seem easier to implement, and there is experimental evidence that synapses can take only discrete values, at least in hippocampus as shown by Petersen et al. (1998); O'Connor et al. (2005). Gutfreund and Stein (1990) applied Gardner's approach to the case of discrete couplings and described multiple cases. They first studied the standard coding case $f = \frac{1}{2}$. For binary synapses $W_{ij} \in \{-1, +1\}$ they found that up to $0.83N$ patterns (or 0.83bits/synapse) can be stored, as previously observed by Krauth and Mezard (1989). For $W_{ij} \in \{0, 1\}$ up to $0.59N$ patterns (or 0.59bits/synapse) can be stored. In this case they can also estimate the fraction of synapses g that are active at maximal capacity, $g = 0.32$. They have also studied the case where synapses can be in a finite number of synaptic states, and found that for synapses taking both positive and negative values the storage capacity goes toward the value for continuous synapses. For instance for $K = 6$ up to $1.53N$ patterns can be stored. Similarly for positive values, it goes towards $1N$ the theoretical value for positive continuous synapses (Amit, 1989). They also studied the case $f \ll 1$ for $W_{ij} \in \{0, 1\}$. They were able to derive an upper bound on the maximal storage capacity and found $i \leq 0.29$ bits/synapse. Note that these models are very sensitive to fluctuations in the number of active neurons per pattern, in this last case, if all patterns have the same number of active neurons, the capacity increases to 0.45 bits/synapse (Brunel, 1994).

The Hebbian learning rule for networks with $\{0, 1\}$ synapses was introduced by Willshaw et al. (1969). It takes the following form

$$W_{ij} = \Theta\left(\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu\right) \tag{2.7}$$

The storage capacity of this model has been studied under different definitions. In a recurrent network, for extremely sparse coding levels $\frac{\ln N}{N}$ and fixed number of active neurons in each pattern, the capacity is 0.23bits/synapse which is also almost optimal compared to the theoretical upper bound (Knoblauch et al., 2010). For larger coding levels, the capacity goes to zero in the large N limit.

It is also worth noting, that contrary to the case of continuous synapses, there exists no algorithm polynomial in time that guarantees to find a weight vector

that allows to reach maximal capacity for a given set of patterns (Amaldi, 1991; Garey and Johnson, 1990). Baldassi et al. (2007) have proposed a supervised algorithm that in practice can find the optimal weights in a reasonable time for $P \lesssim 0.65N$ for the case of $\{-1, 1\}$ synapses and standard coding (for which $i \simeq$ 0.83bits/synapse). This algorithm requires to characterize a given synapse with some 'hidden' states, that keep track of the patterns presented to the network without necessarily modifying the effective value of the binary synapses. It has been proposed that such 'meta-plasticity' would arise from the multi-stability of the protein interaction network at post-synaptic densities (Baldassi et al., 2007). Thus 'meta-plasticity' could be a good candidate to build learning rules achieving good storage capacity for coding levels of order 1 (see also Fusi et al. 2005, Fusi and Abbott 2007).

	Optimal analog synaptic matrix	Hebbian analog synaptic matrix	Optimal binary ({0,1}) synaptic matrix	Hebbian binary ({0,1}) synaptic matrix (Willshaw)
f=$\frac{1}{2}$ (nb of patterns)	2N	0.14 N	0.59 N	o(N)
f$\ll$1 (nb of patterns)	$\frac{N}{2f\lvert lnf\rvert}$	$\frac{N}{2f\lvert lnf\rvert}$	$\leq 0.2\frac{N}{f\lvert lnf\rvert}$ (f small but O(1))	$0.69\frac{N^2}{(lnN)^2}$ (f=1.5 $\frac{lnN}{N}$)
f=$\frac{1}{2}$ (information)	2 bits/synapse	0.14 bits/synapse	0.59 bits/synapse	0 bits/synapse
f$\ll$1 (information)	0.72 bits/synapse	0.72 bits/synapse	$\leq$0.29 bits/synapse (f small but O(1)) $\leq$0.45 bits/synapse (fixed nb of selective neurons)	? 0.23 bits/synapse (with f=1.5 $\frac{lnN}{N}$)

2.1.4 On-line learning

All the learning rules presented so far suffer from the so called black-out catastrophe. If too much patterns are presented to the network, all the presented patterns are eventually lost, see e.g. Amit et al. (1985) for the Hopfield model. To remedy this bothering situation, several authors introduced palimpsest learning rule avoiding this saturation. Nadal et al. (1986) modified the Hopfield learning rule by introducing a decay term that made synaptic changes resulting from the presentation of a given pattern to decay when new patterns are presented. With this modification the capacity drops from 0.14 to 0.05 bits/synapse. The Willshaw learning rule has also been modified to embed networks with binary synapses with palimpsest properties. The main modification is to introduce in the learning rule an activity dependent depression term that naturally allows to forget memories

presented far in the past. Evaluating the storage capacity of such rules is the subject of the work presented in section 2.2.

2.1.5 Beyond binary neurons

Following the work on binary neurons, models of ANN with more realistic neurons have been developed. First, authors introduced models with neurons showing a graded range of activity (Hopfield, 1984; Amit and Treves, 1989; Amit and Tsodyks, 1991). Treves (1990) has shown that such a model can have a storage capacity similar to models with binary neurons. Then, in order to carry more quantitative comparisons with experimental data, networks of spiking neurons were introduced (Gerstner and van Hemmen, 1992; Amit and Brunel, 1997b,a; Brunel, 2000b). The first models focused on studying the bi-stability properties of networks storing a finite number of patterns (see Renart et al. 2003 for a review of these models). A capacity analysis was carried by Roudi and Latham (2007) in a model where excitation and inhibition where balanced. They have shown that the number of stored patterns can scale with the number of connections per neurons at finite coding level. In their model, the coding level is required to be high enough, otherwise selective neurons fire at a too elevated rate and activity is not as irregular as seen in experiments. With such a high coding level, the number of patterns that can be stored would be relatively low (they exhibit a network storing 100 patterns for $10,000$ synapses onto each neuron). It has been suggested that the pattern capacity could be increased further by decreasing the coding level of memories as has been done in networks of binary neurons. However in spiking networks, the coding level can not be decreased too much because the activity of selective neurons has to be distinguishable from the fluctuations of the background activity. A network model with small coding level and spiking statistics consistent with experimental data is still to be found (see Renart et al. (2007) for an attempt)

2.2 Memory capacity of networks with stochastic binary synapses

2.2.1 Summary

In standard attractor neural network models, specific patterns of activity are stored in the synaptic matrix, so that they become fixed point attractors of the network dynamics. The storage capacity of such networks has been quantified in two ways: the maximal number of patterns that can be stored, and the stored information measured in bits per synapse. In this paper we compute both quantities in fully connected networks of N binary neurons with binary synapses, storing patterns with coding level f, in the large N and sparse coding limits ($N \to \infty$, $f \to 0$). We also derive finite-size corrections that accurately reproduce the results of simulations in networks of tens of thousands of neurons. These methods are applied to three different scenarios: (1) the classic Willshaw model, (2) networks with stochastic learning in which patterns are shown only once (one shot learning), (3) networks with stochastic learning in which patterns are shown multiple times. The storage capacities are optimized over network parameters, which allows us to compare the performance of the different models. We show that finite-size effects strongly reduce the capacity, even for networks of realistic sizes. We discuss the implications of these results for memory storage in hippocampus and cerebral cortex.

2.2.2 Introduction

Attractor neural networks have been proposed as long-term memory storage devices (Hopfield, 1982; Amit, 1989; Brunel, 2003). In such networks, a pattern of activity (the set of firing rates of all neurons in the network) is said to be memorized if it is one of the stable states of the network dynamics. Specific patterns of activity become stable states thanks to synaptic plasticity mechanisms, including both long term potentiation and depression of synapses, that create positive feed-back loops through the network connectivity. A long standing question in the field has been the question of the storage capacity of such networks. Much effort has been devoted to compute the number of attractor states that can be imprinted in the synaptic matrix, in networks of binary neurons (Amit et al., 1985; Sompolinsky, 1986; Gardner, 1988; Tsodyks and Feigel'man, 1988). Models storing patterns with a covariance rule (Sejnowski, 1977; Hopfield, 1982; Tsodyks and Feigel'man, 1988) were shown to be able to store a number of patterns that scale linearly with the number of synapses per neuron. In the sparse coding limit (in which the average

fraction of selective neurons per pattern f goes to zero in the large N limit), the capacity was shown to diverge as $1/(f|\log(f)|)$. These scalings lead to a network storing on the order of 1 bit per synapse, in the large N limit, for any value of the coding level. Elizabeth Gardner (Gardner, 1988) computed the maximal capacity, in the space of all possible coupling matrices, and demonstrated a similar scaling for capacity and information stored per synapse.

These initial studies, performed on the simplest possible networks (binary neurons, full connectivity, unrestricted synaptic weights) were followed by a second wave of studies that examined the effect of adding more neurobiological realism: random diluted connectivity (Sompolinsky, 1986), neurons characterized by analog firing rates (Amit and Tsodyks, 1991), learning rules in which new patterns progressively erase the old ones (Nadal et al., 1986; Parisi, 1986). The above mentioned modifications were shown not to affect the scaling laws described above. One particular modification however was shown to have a drastic effect on capacity. A network with binary synapses and stochastic on-line learning was shown to have a drastically impaired performance, compared to networks with continuous synapses (Tsodyks, 1990; Amit and Fusi, 1994). For finite coding levels, the storage capacity was shown to be on the order of $\sqrt{N}$, not N stored patterns, while the information stored per synapse goes to zero in the large N limit. In the sparse coding limit however ($f \sim \log(N)/N$), the capacity was shown to scale as $1/f^2$, and therefore a similar scaling as the Gardner bound, while the information stored per synapse remains finite in this limit. These scaling laws are similar to the Willshaw model (Willshaw et al., 1969), which can be seen as a particular case of the Amit-Fusi (Amit and Fusi, 1994) rule. The model was then subsequently studied in greater detail by Amit and Huang (2010); Huang and Amit (2011) who computed the storage capacity for finite values of N, using numerical simulations and several approximations for the distributions of the 'local fields' of the neurons. However, computing the precise storage capacity of this model in the large N limit remains an open problem.

In this article we focus on a model of binary neurons where binary synapses are potentiated or depressed stochastically depending on the states of pre and post synaptic neurons (Amit and Fusi, 1994). We first introduce analytical methods that allow us to compute the storage capacity in the large N limit, based on a binomial approximation for the synaptic inputs to the neurons. We first illustrate it on the Willshaw model and to recover the well-known result on the capacity of this model (Willshaw et al., 1969; Nadal, 1991; Knoblauch et al., 2010). We then move to a stochastic learning rule, in which we study two different scenarios: (i) in which patterns are presented only once - we will refer to this model as the SP

(Single Presentation) model (Amit and Fusi, 1994); (ii) in which noisy versions of the patterns are presented multiple-times - the MP (Multiple presentations) model (Brunel et al., 1998). For both models we compute the storage capacity and the information stored per synapse in the large N limit, and investigate how they depend on the various parameters of the model. We then study finite size effects, and show that they have a huge effect even in networks of tens of thousands of neurons. Finally we show how capacity in finite size networks can be enhanced by introducing inhibition, as proposed in Amit and Huang (2010); Huang and Amit (2011). In the discussion we summarize our results and discuss the relevance of the SP and MP networks to memory maintenance in the hippocampus and cortex.

2.2.3 Storage capacity in the $N \to \infty$ limit

The network

We consider a network of N binary (0,1) neurons, fully connected through a binary (0,1) synaptic connectivity matrix. The activity of neuron i ($i = 1...N$) is described by a binary variable, $\sigma_i = 0, 1$. Each neuron can potentially be connected to every other neurons, through a binary connectivity matrix $\mathbf{W}$. This connectivity matrix depends on P random uncorrelated patterns ('memories') $\vec{\xi}^{\mu}, \mu = 1, \ldots, P$ that are presented during the learning phase. The state of neuron $i = 1, \ldots, N$ in pattern $\mu = 1, \ldots, P$ is

$$\xi_i^{\mu} = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } 1 - f \end{cases} \tag{2.8}$$

where f is the coding level of the memories. We study this model in the limit of low coding level, $f \to 0$ when $N \to \infty$. In all the models considered here, P scales as $1/f^2$ in the sparse coding limit. Thus, we introduce a parameter $\alpha = Pf^2$ which stays of order 1 in the sparse coding limit.

After the learning phase, we choose one of the P presented patterns $\vec{\xi}^{\mu_0}$, and check whether it is a fixed point of the dynamics:

$$\sigma_i(t+1) = \Theta[h_i(t) - fN\theta], \tag{2.9}$$

where

$$h_i(t) = \sum_{j=1}^{N} W_{ij}\sigma_j(t) \tag{2.10}$$

is the total synaptic input ("field") of neuron i, $\theta = O(1)$ is a scaled activation threshold, and Θ is the Heaviside function.

Field averages

When testing the stability of pattern $\vec{\xi}^{\mu_0}$ after learning P patterns, we need to compute the distribution of the fields on selective neurons (sites i such that $\xi_i^{\mu_0} = 1$), and of the fields on non-selective neurons (sites i such that $\xi_i^{\mu_0} = 0$). The averages of those fields are fNg_+ and fNg respectively, where

$$g_+ = \mathbb{P}(W_{ij} = 1 | \xi_i^{\mu_0} = \xi_j^{\mu_0} = 1) \tag{2.11}$$

and

$$g = \mathbb{P}(W_{ij} = 1 | \, (\xi_i^{\mu_0}, \xi_j^{\mu_0}) \neq (1,1) \,). \tag{2.12}$$

Pattern $\vec{\xi}^{\mu_0}$ is perfectly imprinted in the synaptic matrix if $g_+ = 1$ and $g = 0$. However, because of the storage of other patterns, g_+ and g take intermediate values between 0 and 1. Note that here we implicitly assume that the probability of finding an potentiated synapse between two neurons i, j such that $\xi_i^{\mu_0} = \xi_j^{\mu_0} = 0$ or $\xi_i^{\mu_0} \neq \xi_j^{\mu_0}$ is the same. This is true for the models we consider below. g_+ and g are function of α, f, and other parameters characterizing learning.

Information stored per synapse

One measure of the storage capability of the network is the information stored per synapse :

$$\begin{aligned} i &= \frac{P_{max} N(-f \ln_2 f - (1-f)\ln_2(1-f))}{N^2} \\ &\underset{f \to 0}{\simeq} \alpha \frac{|\ln_2 f|}{fN} \end{aligned} \tag{2.13}$$

where P_{max} is the size of a set of patterns in which each pattern is a fixed point of the dynamics with probability one. With $\alpha = O(1)$, for the information per synapse to be of order one in the large N limit, we need to take f as

$$f = \beta \frac{\ln N}{N}. \tag{2.14}$$

In this case the information stored per synapse has the simple expression:

$$i = \frac{\alpha}{\beta \ln 2} \tag{2.15}$$

Computing the storage capacity

Our goal here is to compute the size $P_{max} = \alpha / f^2$ of the largest set of patterns that can be stored in the connectivity matrix. The criterion for storage that we adopt is that if one picks a pattern in this set, then this pattern is a fixed point of

the dynamics with probability 1. We thus need to compute the probability $\mathbb{P}_{ne}$ of no error in retrieving a particular pattern μ_0. To compute this probability, we first need to estimate the probabilities that a single selective/non-selective neuron is in its right state when the network is initialized in a state corresponding to pattern μ_0. For a pattern with M selective neurons, and neglecting correlations between neurons (which is legitimate if $f \ll 1/\sqrt{N}$ (Amit and Fusi, 1994)), we have

$$\mathbb{P}_{ne} = (1 - \mathbb{P}(h_i \leq fN\theta | \xi_i^{\mu_0} = 1))^M (1 - \mathbb{P}(h_i \geq fN\theta | \xi_i^{\mu_0} = 0))^{N-M} \tag{2.16}$$

Clearly, for $\mathbb{P}_{ne}$ to go to 1 in the large N limit, the probabilities for the fields of single neurons to be on the wrong side of the threshold have to vanish in that limit. A first condition for this to happen is $g_+ > \theta > g$ - if these inequalities are satisfied, then the average fields of both selective and non-selective neurons are on the right side of the threshold. When g_+ and g are sufficiently far from θ, the tail probabilities of the distribution of the fields are

$$\begin{aligned} \mathbb{P}\left(h_i \leq fN\theta | \xi_i^{\mu_0} = 1\right) &= \exp\left(-M\Phi(g_+, \theta) + o(M)\right) & (2.17) \\ \mathbb{P}\left(h_i \geq fN\theta | \xi_i^{\mu_0} = 0\right) &= \exp\left(-M\Phi(g, \theta) + o(M)\right) & (2.18) \end{aligned}$$

where $\Phi(g_+, \theta)$, $\Phi(g, \theta)$ are the rate functions associated with the distributions of the fields (see Methods A). Neglecting again correlations between inputs, the distributions of the fields are binomial distributions, and the rate functions are

$$\Phi(x, \theta) = \theta \ln \frac{\theta}{x} + (1 - \theta) \ln \frac{1 - \theta}{1 - x} \tag{2.19}$$

Inserting Eqs. (2.17,2.18,2.19,2.14) in Eq. (2.16), we find that

$$\mathbb{P}_{ne} = \exp\left[-\exp\left(X_s\right) - \exp\left(X_n\right)\right] \tag{2.20}$$

where

$$\begin{aligned} X_s &= -\beta\Phi(g_+, \theta) \ln N + \ln(\ln(N)) + o(\ln(\ln(N))) \\ X_n &= -\beta\Phi(g, \theta) \ln N + \ln(N) + o(\ln(N)). \end{aligned} \tag{2.21}$$

For $\mathbb{P}_{ne}$ to go to 1 in the large N limit, we need both X_s and X_n to go to $-\infty$ in that limit. This will be satisfied provided

$$\begin{aligned} \Phi(g_+, \theta) &> \frac{\ln(\ln N)}{\beta \ln N} & (2.22) \\ \Phi(g, \theta) &> \frac{1}{\beta} & (2.23) \end{aligned}$$

These inequalities are equivalent in the large N limit to the inequalities

$$g_+ > \theta > g + \zeta \tag{2.24}$$

where ζ is given by the equation $\Phi(g+\zeta,\theta) = 1/\beta$.

The maximal information per synapse is obtained by saturating inequalities (2.22) and (2.23), and optimizing over the various parameters of the model. In practice, for given values of α, and parameters of the learning process, we compute g and g_+; we can then obtain the optimal values of the threshold θ and the rescaled coding level β as

$$\theta = g_+ \tag{2.25}$$

$$\beta = \frac{1}{\Phi(g,\theta)}, \tag{2.26}$$

and compute the information per synapse using Eq. (2.15). We can then find the optimum of i in the space of all parameters.

Before applying these methods to various models, we would like to emphasize two important features of these calculations:

- In Eq. (2.22), note that the r.h.s. goes to zero extrememely slowly as N goes to ∞ (as $\ln(\ln N)/\ln(N)$) - thus, we expect huge finite size effects. This will be confirmed in Section 5 where these finite size effects are studied in detail.
- In the sparse coding limit, a Gaussian approximation of the fields gives a poor approximation of the storage capacity, since the calculation probes the tail of the distribution.

2.2.4 Willshaw model

The capacity of the Willshaw model has already been studied by a number of authors (Willshaw et al., 1969; Nadal, 1991; Knoblauch et al., 2010). Here, we present the application of the analysis described in Section 1 to the Willshaw model, for completeness and comparison with the models described in the next Section. In this model, after presenting P patterns to the network, the synaptic matrix is described as follows: $W_{ij} = 1$ if at least one of the P presented patterns had neuron i and j co-activated, $W_{ij} = 0$ otherwise. Thus, after the learning phase, we have,

$$\begin{aligned} g_+ &= 1 \\ g &= 1-(1-f^2)^P \simeq 1-\exp(-\alpha) \text{ for small } f \end{aligned} \tag{2.27}$$

Saturating the inequalities (2.25),(2.26) with g fixed, one obtains the information stored per synapse,

$$i_{opt} = \ln(1-g)\ln g \frac{1}{\ln 2} \tag{2.28}$$

The information stored per synapse is shown as a function of g in figure 2.1a. It reaches a maximum is reached for $g = 0.5$ at $i_W = \ln 2 = 0.69$ bits/synapse, but goes to zero in both the $g \to 0$ and $g \to 1$ limits. The model has a storage capacity comparable to its maximal value, $i_{opt} > 0.5 i_W$ in a large range of values of g (between 0.1 and 0.9. We can also optimize capacity for a given value of β, as shown in figure 2.1b. It reaches its maximum at $\beta = 1.4$, and goes to zero in the small and large β limits. Again, the model has a large storage capacity for a broad range of β, $i_{opt} > 0.5 i_W$ for β between 0.4 and 10.

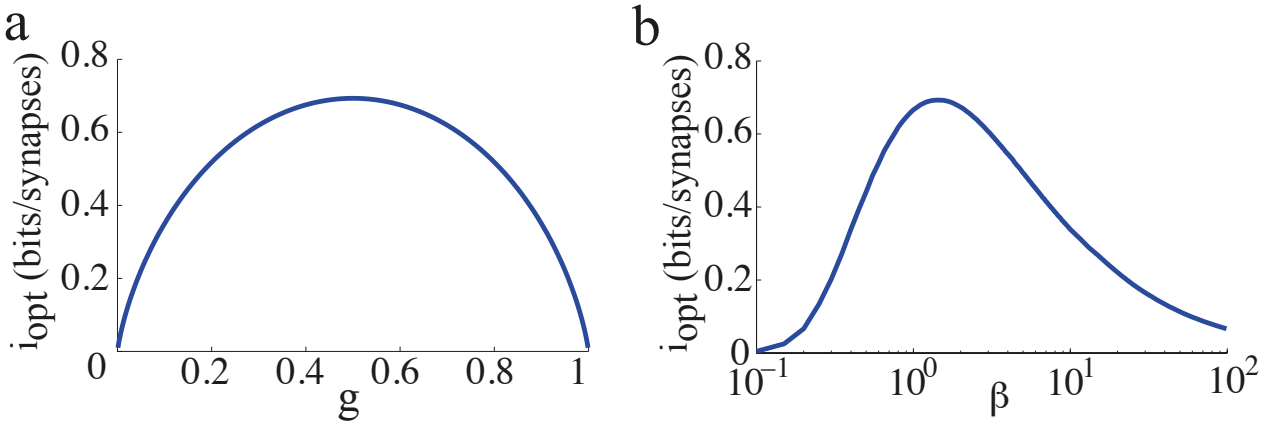

Figure 2.1: *Optimized information capacity of the Willshaw model in the limit $N \to +\infty$. Information is optimized by saturating (2.25) ($\theta = 1$) and (2.26): a. i_{opt} as a function of g, b. i_{opt} as a function of $\beta = fN/\ln N$.*

2.2.5 Amit-Fusi model

A drawback of the Willshaw learning rule is that it only allows for synaptic potentiation. Thus, if patterns are continuously presented to the network, all synapses will eventually be potentiated and no memories can be retrieved. In (Amit and Fusi, 1994), Amit and Fusi introduced a new learning rule that maintains the simplicity of the Willshaw model, but allows for continuous on-line learning. The proposed learning rule includes synaptic depression. At each learning time step μ, a new pattern $\vec{\xi}^{\mu}$ with coding level f is presented to the network, and synapses are updated stochastically:

- for synapses such that $\xi_i^\mu = \xi_j^\mu = 1$:
  if $W_{ij}(\mu - 1) = 0$, then $W_{ij}(\mu)$ is potentiated to 1 with probability q_+ ; and if $W_{ij}(\mu - 1) = 1$ it stays at 1.
- for synapses such that $\xi_i^\mu \neq \xi_j^\mu$:
  if $W_{ij}(\mu - 1) = 0$, then $W_{ij}(\mu)$ stays at 0 ; and if $W_{ij}(\mu - 1) = 1$ it is depressed

to 0 with probability q_-.

- for synapses such that $\xi_i^\mu = \xi_j^\mu = 0$, $W_{ij}(\mu) = W_{ij}(\mu - 1)$.

The evolution of a synapse W_{ij} during learning can be described by the following Markov process :

$$\begin{bmatrix} \mathbb{P}(W_{ij}^{\mu+1} = 0) \\ \mathbb{P}(W_{ij}^{\mu+1} = 1) \end{bmatrix} = \begin{bmatrix} 1-a & b \\ a & 1-b \end{bmatrix} \times \begin{bmatrix} \mathbb{P}(W_{ij}^{\mu} = 0) \\ \mathbb{P}(W_{ij}^{\mu} = 1) \end{bmatrix} \tag{2.29}$$

where $a = f^2 q_+$ is the probability that a silent synapse is potentiated upon the presentation of pattern μ and $b = 2f(1-f)q_-$ is the probability that a potentiated synapse is depressed. After a sufficient number of patterns has been presented the distribution of synaptic weights in the network reaches a stationary state. We study the network in this stationary regime.

For the information capacity to be of order 1, the coding level has to scale as $\frac{\ln N}{N}$, as in the Willshaw model, and the effects of potentiation and depression have to be of the same order (Amit and Fusi, 1994). Thus we define the *depression-potentiation ratio* δ as,

$$\delta = \frac{2f(1-f)q_-}{f^2 q_+} \tag{2.30}$$

We can again use equation (2.15) and the saturated inequalities (2.25,2.26) to compute the maximal information capacity in the limit $N \to \infty$. This requires computing g and g_+, defined in the previous section, as a function of the different parameters characterizing the network. We track a pattern $\xi^{\vec{\mu_0}}$ that has been presented P time steps in the past. In the following we refer to P as the age of the pattern. In the sparse coding limit, g corresponds to the probability that a synapse is potentiated. It is determined by the depression-potentiation ratio δ,

$$g = \frac{1}{1+\delta} \tag{2.31}$$

and

$$g_+ = g + q_+(1-g)(1-a-b)^P \simeq g + q_+(1-g)\exp(-\frac{q_+\alpha}{g}) \text{ for } f \ll 1 \tag{2.32}$$

where $\alpha = Pf^2$. Our goal is to determine the age P of the oldest pattern that is still a fixed point of the network dynamics, with probability one. Note that in this network, contrary to the Willshaw model in which all patterns are equivalent, here younger patterns, of age $P' < P$, are more strongly imprinted in the synaptic matrix, $g_+(P') > g_+(P)$, and thus also stored with probability one.

Choosing an activation threshold and a coding level that saturate inequalities (2.25) and (2.26), information capacity can be expressed as :

$$\begin{aligned} i_{opt} &= \frac{g}{q_+}\ln\left[q_+\frac{1-g}{g_+-g}\right]\left[g_+\ln_2\frac{g_+}{g}+(1-g_+)\ln_2\frac{1-g_+}{1-g}\right] \\ &= \frac{\alpha}{1+\delta}\Big[\left(1+\delta q_+e^{-\alpha(1+\delta)q_+}\right)\ln_2\left(1+\delta q_+e^{-\alpha(1+\delta)q_+}\right)+ \\ &\quad \delta\left(1-q_+e^{-\alpha(1+\delta)q_+}\right)\ln_2\left(1-q_+e^{-\alpha(1+\delta)q_+}\right)\Big] \end{aligned} \tag{2.33}$$

The optimal information $i_{SP} = 0.083$ bits/synapse is reached for $q_+ = 1\,,\ \theta = 0.72\,,\ \beta = 2.44\,,\ \alpha = 0.14\,,\ \delta = 2.57$ which gives $g = 0.28\,,\ g_+ = 0.72$.

The dependence of i_{opt} on the different parameters is shown in figure 2.2. Panel a shows the dependence on g the fraction of activated synapses in the asymptotic learning regime. Panels b, c and d show the dependence on δ, β and q_+. Note from panel c that there is a broad range of values of β that give information capacities similar to the optimal one. One can also observe that the optimal information capacity is about 9 times lower in the SP model than in the Willshaw model. This is the price one pays to have a network that is able to continuously learn new patterns. However, it should be noted that at maximal capacity, in the Willshaw model, every pattern has a vanishing basin of attraction while in the SP model, only the oldest stable patterns have vanishing basins of attraction. This feature is not captured by our measure of storage capacity.

2.2.6 Multiple presentations of patterns, slow learning regime

In the SP model, patterns are presented only once. Brunel et al (Brunel et al., 1998) studied the same network of binary neurons with stochastic binary synapses but in a different learning context, where patterns are presented multiple times. More precisely, at each learning time step t, a noisy version $\vec{\xi}^{\mu(t),t}$ of one of the P prototypes $\vec{\xi}^{\mu}$ is presented to the network,

$$\begin{cases} \mathbb{P}(\xi_i^{\mu(t),t}=1)=1-(1-f)x \ \text{ and } \ \mathbb{P}(\xi_i^{\mu(t),t}=0)=(1-f)x \ \text{ for } \ \xi_i^{\mu(t)}=1 \\ \mathbb{P}(\xi_i^{\mu(t),t}=1)=fx \ \text{ and } \ \mathbb{P}(\xi_i^{\mu(t),t}=0)=1-fx \ \text{ for } \ \xi_i^{\mu(t)}=0 \end{cases} \tag{2.34}$$

Here x is a noise level: if $x = 0$, presented patterns are identical to the prototypes, while if $x = 1$, the presented patterns are uncorrelated with the prototypes. As for the SP model this model achieves $i_{opt} = O(1)$ and has good storage properties if the depression-potentiation ratio δ is of order one and if the coding level scales

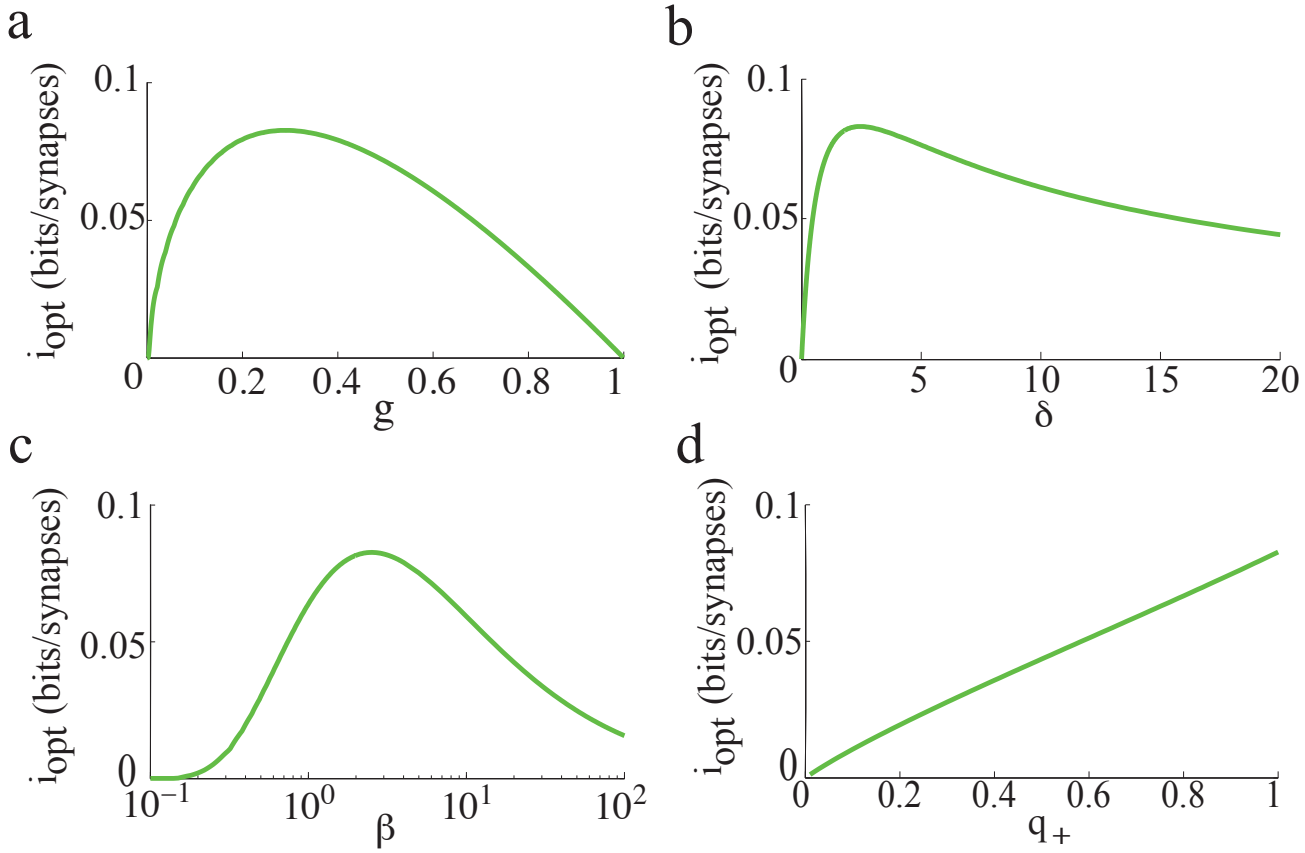

Figure 2.2: *Optimized information capacity for the SP model in the limit $N \to +\infty$. a. i_{opt} as a function of g, b. i_{opt} as a function of δ, the ratio between the number of depressing events and potentiating events at pattern presentation, c. i_{opt} as a function of $\beta = f\frac{N}{\ln N}$, d. i_{opt} as a function of the LTP transition probability q_+.*

with network size as $f \propto \frac{\ln N}{N}$. If learning is slow, $q_+, q_- \ll 1$, and the number of presentations of patterns of each class become large the probabilities g and g_+ are (Brunel et al., 1998):

$$g = \sum_{\Pi=0}^{+\infty} \frac{(1-x)^2\Pi + \alpha x(2-x)}{(1-x)^2\Pi + \alpha(\delta + x(2-x))} \frac{\alpha^\Pi \exp(-\alpha)}{\Pi\,!} \tag{2.35}$$

and

$$g_+ = \sum_{\Pi=0}^{+\infty} \frac{(1-x)^2(\Pi+1) + \alpha x(2-x)}{(1-x)^2(\Pi+1) + \alpha(\delta + x(2-x))} \frac{\alpha^\Pi \exp(-\alpha)}{\Pi\,!} \tag{2.36}$$

We inserted those expressions in Eqs. (2.25,2.26) to study the maximal information capacity of the network under this learning protocol. The optimal information $i_{MP} = 0.69$ bits/synapse is reached at $x = 0$ for $\theta \to 1$, $\beta \to 1.44$, $\delta \to 0$, $\alpha \to 0.69$ which gives $g \to \frac{1}{2}$, $g_+ \to 1$. In this limit, the network becomes equivalent to the Willshaw model.

The maximal capacity is about 9 times larger than for a network that has to learn in one shot. On figure 2.3a we plot the optimal capacity as a function of g.

The capacity of the slow learning network with multiple presentations is bounded by the capacity of the Willshaw model for all values of g, and it is reached when the depression-potentiation ratio $\delta \to 0$. For this value, no depression occurs during learning: the network loses palimpsest properties, i.e. the ability to erase older patterns to store new ones, and it is not able to learn if the presented patterns are noisy. The optimal capacity decreases with δ, for instance at $\delta = 1$ (as many potentiation events as depression events at each pattern presentation), $i_{opt} = 0.35$ bits/synapse. Figure 2.3c shows the dependence as a function of $\beta = f\frac{N}{\ln N}$. In figure 2.3d, we show the optimized capacity for different values of the noise x in the presented patterns. This quantifies the trade-off between the storage capacity and the generalization ability of the network (Brunel et al., 1998).

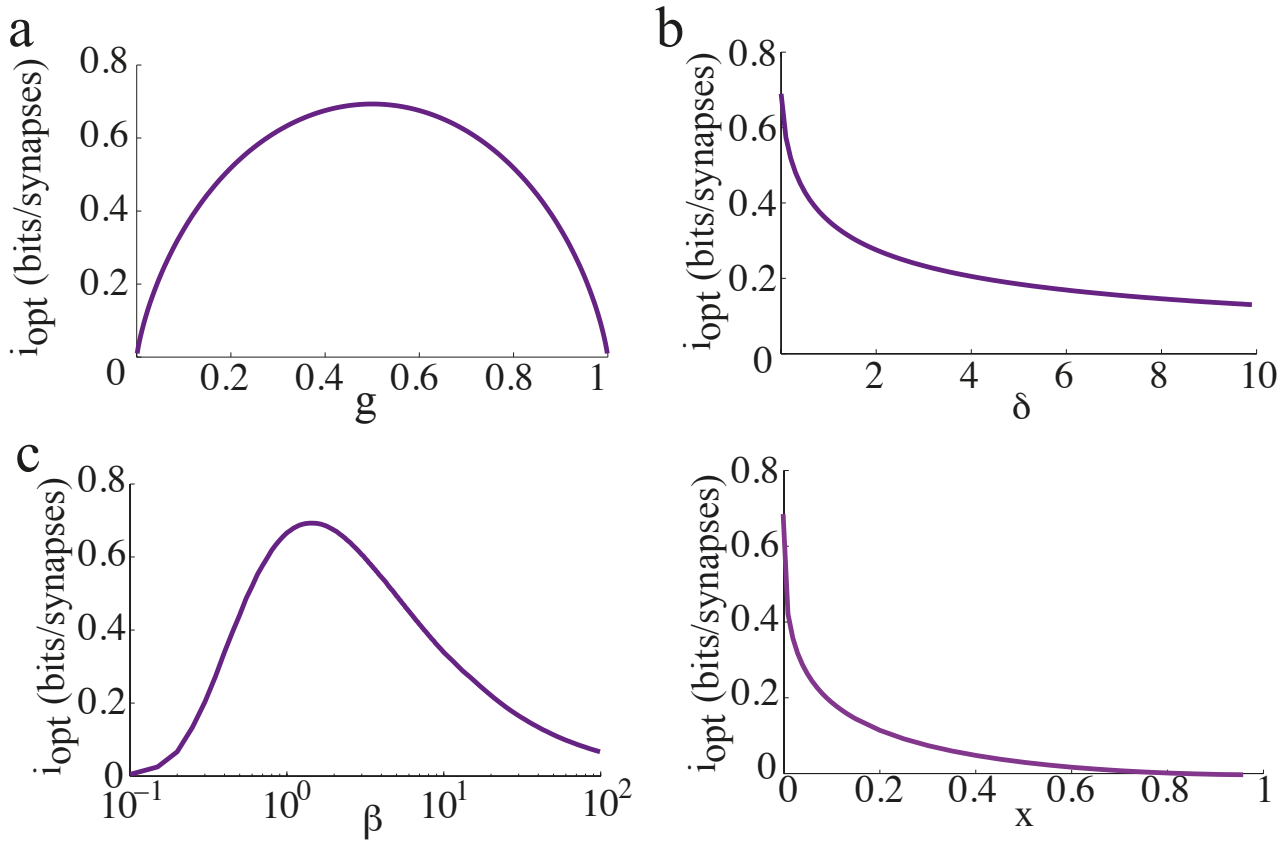

Figure 2.3: *Optimized information capacity for the MP model in the limit $N \to +\infty$. a. Optimal information capacity as a function of g, the average number of activated synapses after learning. Optimal capacity is reached in the limit $\delta \to 0$ and at $x = 0$ where the capacity is the same as for the Willshaw model. b. Dependence of information capacity on δ, the ratio between the number of depressing events and potentiating events at pattern presentation. c. Dependence on $\beta = f\frac{N}{\ln N}$. d. Dependence on the noise in the presented patterns, x. This illustrates the trade-off between the storage capacity and the generalization ability of the network.*

2.2.7 Finite-size networks

The results we have presented so far are valid for infinite size networks. Finite-size effects can be computed for the three models we have discussed so far (see Methods B). The main result of this section is that the capacity of networks of realistic sizes is very far from the large N limit. We compute capacities for finite networks in the SP and MP settings, and we validate our finite size calculations by presenting the results of simulation of large networks of sizes $N = 10,000$, $N = 50,000$.

We summarize the finite size calculations for the SP model (a more general and detailed analysis is given in Methods B). In the finite network setting, conditional on the tested pattern μ_0 having $M+1$ selective neurons, the probability of no error is $\mathbb{P}_{ne}$ is given by

$$\mathbb{P}_{ne} = \exp\left[-\exp\left(X_s\right) - \exp\left(X_n\right)\right]$$

with

$$\begin{aligned} X_s &= -\beta_M \Phi\left(g_+, \theta_M\right) \ln N + \frac{1}{2} \ln \ln N \\ &\quad -\frac{1}{2} \ln \left[\frac{\left(1 - \exp(\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M))\right)^2 2\pi\theta_M(1-\theta_M)}{\beta_M} \right] + o(1) \\ X_n &= (-\beta_M \Phi(g, \theta_M) + 1) \ln N - \frac{1}{2} \ln \ln N - \\ &\quad \frac{1}{2} \ln \left[\left(1 - \exp(-\frac{\partial \Phi}{\partial \theta}(g, \theta_M))\right)^2 2\pi\theta_M(1-\theta_M)\beta_M \right] + o(1) \end{aligned} \tag{2.37}$$

where $\beta_M = \frac{M}{\ln N}, \theta_M = \theta \frac{fN}{M}$ and Φ is given by Eq. (2.19). In the calculations for $N \to +\infty$ discussed in Sections 1-4 we kept only the dominant term in $\ln N$, which yields equations (2.25) and (2.26).

In the above equations, the first order corrections scale as $\frac{\ln \ln N}{\ln N}$, which has a dramatic effect on the storage capacity of finite networks. In figure 2.4a,b, we plot $\overline{\mathbb{P}}_{ne}$ (where the bar denotes an average over the distribution of M) as a function of the age of the pattern, and compare this with numerical simulations. It is plotted for $N = 10,000$ and $N = 50,000$ for learning and network parameters chosen to optimize the storage capacity of the infinite-size network (see Section 3). We show the result for two different approximations of the field distribution: a binomial distribution (magenta), as used in the previous calculations for infinite size networks ; and a gaussian (red) approximation (see Methods C for calculations) as used by previous authors Amit and Huang (2010); Huang and Amit (2011); Leibold and Kempter (2008). For these parameters the binomial approximation gives an accurate estimation of $\overline{\mathbb{P}}_{ne}$, while the gaussian calculation overestimates

it.

The curves we get are far from the step functions predicted for $N \to +\infty$ by Eq. (2.51). To understand why, compare equations (2.77), and (2.37): finite size effects can be neglected when $|(-\beta\Phi(g_+,\theta))| \gg \frac{\ln\ln N}{\ln N}$ and $|(-\beta\Phi(g,\theta)+1)| \gg \frac{\ln\ln N}{\ln N}$. Because the finite size effects are of order $\frac{\ln\ln N}{\ln N}$, it is only for huge values of N that the asymptotic capacity can be recovered. For instance if we choose an activation threshold θ slightly above the optimal threshold given in Section 3 ($\theta = \theta_{opt} + 0.01 = 0.73$), then $-\beta\Phi(g,\theta)+1 = -0.06$, and for $N = 10^{100}$ we only have $|-\beta\Phi(g,\theta)+1| \simeq 3\frac{\ln\ln N}{\ln N}$. In figure 2.4c we plot $\mathbb{P}_{ne}$ as a function of $\frac{\alpha}{\alpha_{opt}}$ where $\alpha_{opt} = 0.14$ is the value of α that optimizes capacity in the large N limit, $\theta = 0.73$ and the other parameters are the one that optimizes capacity. We see that we are still far from the large N limit for $N = 10^{100}$. Networks of sizes $10^4 - 10^6$ have capacities which are only between 20% and 40% of the predicted capacity in the large N limit. Neglecting fluctuations in the number of selective neurons, we can derive an expression for the number of stored patterns P that includes the leading finite size correction for the SP model,

$$P(N) = c_1 \frac{N^2}{(\ln N)^2}\left[1 - c_2\sqrt{\frac{\ln(\ln N)}{\ln N}} + o\left(\sqrt{\frac{\ln(\ln N)}{\ln N}}\right)\right] \tag{2.38}$$

where c_1 and c_2 are two constants (see Methods B).

If we take fluctuations in the number of selective neurons into account, it introduces other finite-size effects as can be seen from equations (2.49) and (2.50) in the Methods section. These fluctuations can be discarded if $|(-\beta\Phi(g_+,\theta))| \gg \frac{\sqrt{\beta}}{\sqrt{\ln N}}\frac{1-\theta}{1-g_+}$ and $|(1-\beta\Phi(g,\theta))| \gg \frac{\sqrt{\beta}}{\sqrt{\ln N}}\frac{1-\theta}{1-g}$. In figure 2.4d we plot $\overline{\mathbb{P}}_{ne}$ for different values of N. We see that finite size effects are even stronger in this case.

To plot the curves of figure 2.4, we chose parameters to be those that optimize storage capacity for infinite network sizes. When N is finite, those parameters are no longer optimal. To optimize parameters at finite N, since the probability of error as a function of age is no longer a step function, it is not possible to find the last pattern stored with probability one. Instead we define the capacity P_c as the pattern age for which $\overline{\mathbb{P}}_{ne} = \frac{1}{2}$. Using equations (2.37) and performing an average over the distribution of M, we find parameters optimizing pattern capacity for fixed values of β. Results are shown on figure 2.5a,b for $N = 10,000$ and $N = 50,000$. We show the results for the different approximations used to model the neural fields: the blue line is the binomial approximation, the cyan line the gaussian approximation and the magenta one is a gaussian approximation with a covariance term that takes into account correlations between synapses (see Methods C and Amit and Huang (2010); Huang and Amit (2011)). For $f < \frac{1}{\sqrt{N}}$ the storage

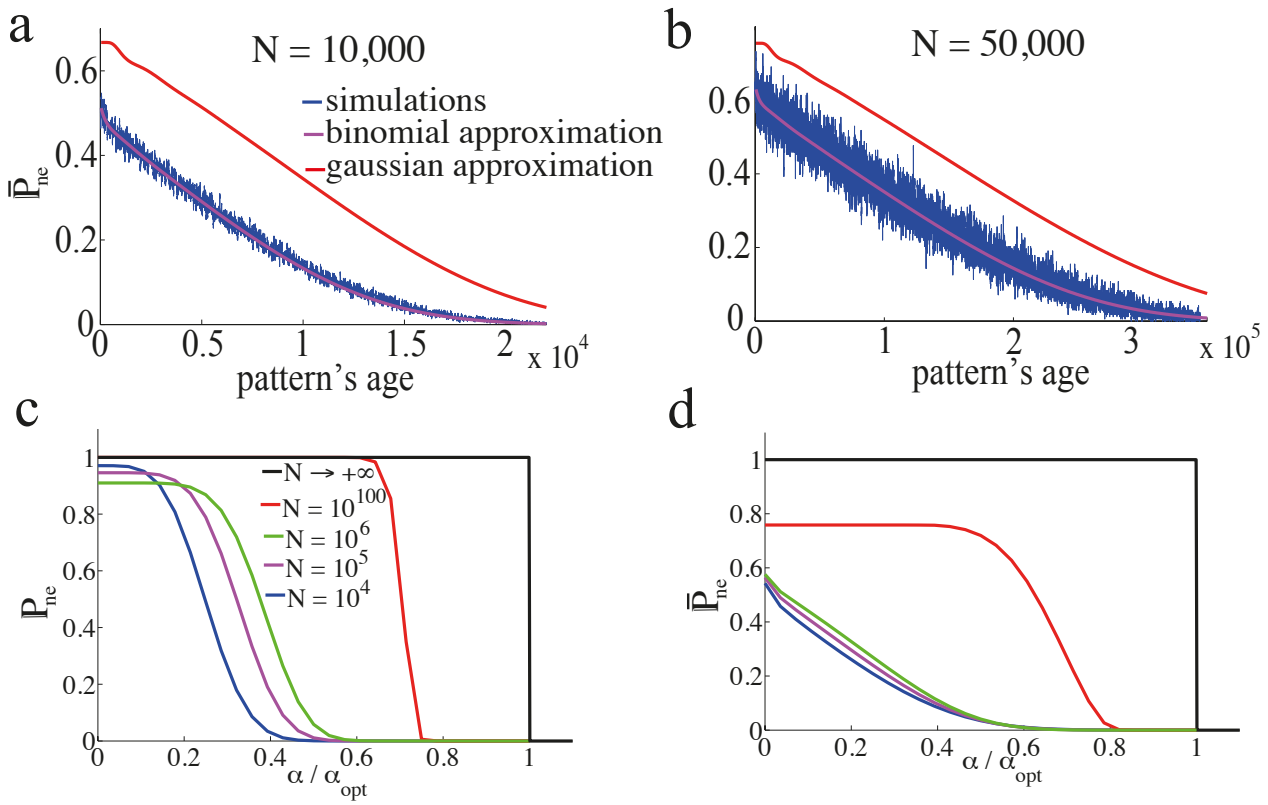

Figure 2.4: *Finite size effects. Shown is $\mathbb{P}_{ne}$, the probability that a tested pattern of a given age is stored without errors, for the SP model. a. $\mathbb{P}_{ne}$ as a function of the age of the tested pattern. Parameters are those optimizing capacity at $N \to +\infty$ (see Section 3), results are for simulations (blue line) and calculations with a binomial approximation of the fields distributions (magenta) and a gaussian approximation (red) ; $\mathbb{P}_{ne}$ is averaged over different value of M, the number of selective neurons in the tested pattern (magenta line). b Same for $N = 5.10^4$. c. $\mathbb{P}_{ne}$ as a function of a scaled version of pattern age (see text for details), fluctuations in M are discarded on this plot. d. Same as c with an average of $\mathbb{P}_{ne}$ over different M.*

capacity of simulated networks (black crosses) is well predicted by the binomial approximation while the gaussian approximations over-estimates capacity. For $f > \frac{1}{\sqrt{N}}$, the correlations between synapses can no longer be neglected (Amit and Fusi, 1994). The gaussian approximation with covariance captures the drop in capacity at large f.

For $N = 10,000$, the SP model can store a maximum of $P_c = 7,800$ patterns at a coding level $f = 0.0015$ (see blue curve in figure 2.5c). As suggested in figures 2.4c,d, the capacity of finite networks is strongly reduced compare to the capacity predicted for infinite size networks. More precisely, if the network of size $N = 10,000$ had the same information capacity as the infinite size network (2.33), it would store up to $P = 70,000$ patterns at coding level $f = 0.0007$. Part of this decrease in capacity is avoided if we consider patterns that have a fixed number fN of selective neurons. This corresponds to the red curve in figure 2.4c. For

fixed sizes the pattern capacity is approximately twice as large. In figure 2.5d, we do the same analysis for the MP model with $N = 10,000$. Here we have also optimized all the parameters, except for the depression-potentiation ratio which is set to $\delta = 1$, ensuring that the network has the palimpsest property and the ability to deal with noisy patterns. For $N = 10,000$, the MP model with $\delta = 1$ can store up to $P_c = 70,000$ patterns, at $f = 0.001$ (versus $P_c = 7,800$ at $f = 0.0015$ for the SP model). One can also compute the optimized capacity for a given noise level. At $x = 0.1$, $P_c = 20,900$ for $f = 0.0012$ and $\delta = 4.3$ or at $x = 0.2$, $P_c = 8,900$ for $f = 0.0018$ and $\delta = 6.9$.

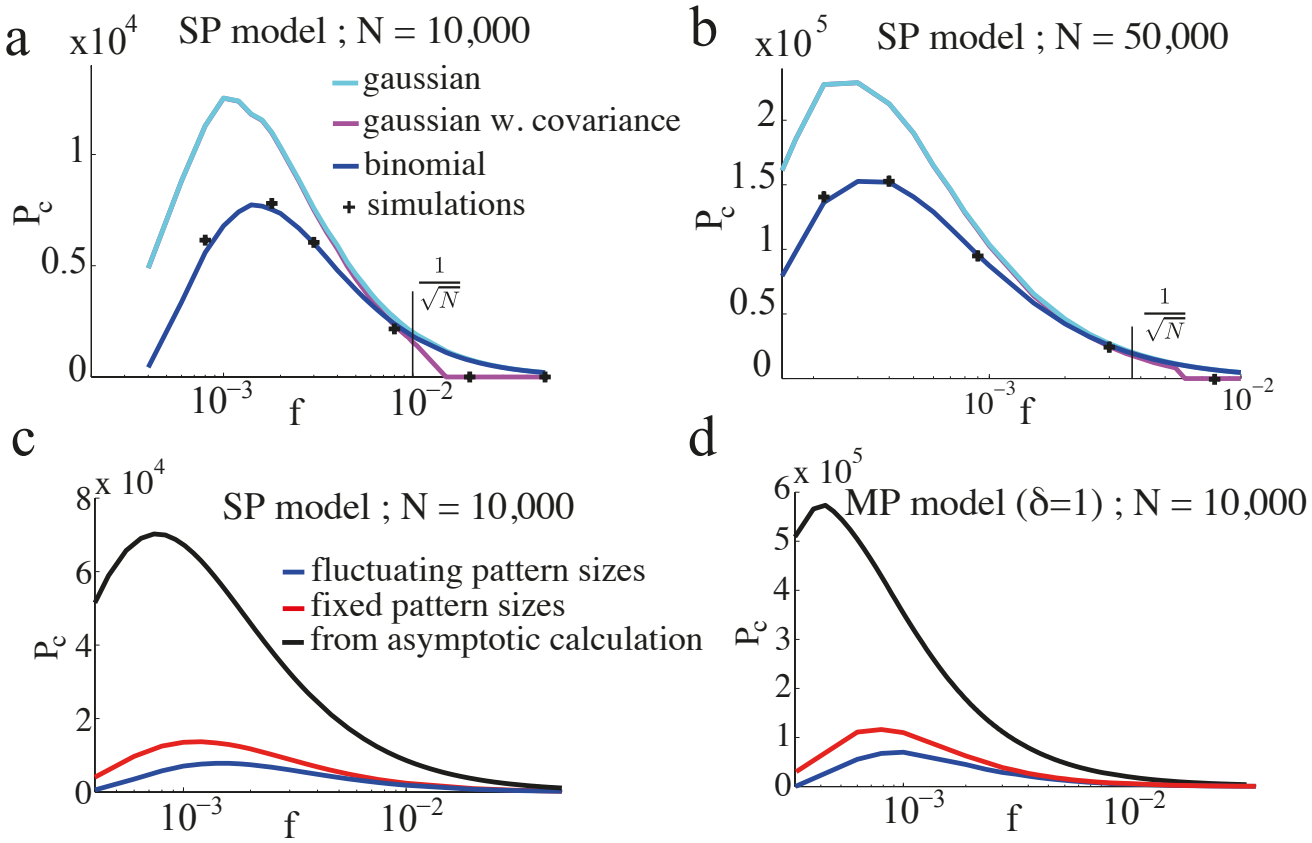

Figure 2.5: *Capacity at finite N. a,b.* P_c *as a function of* f *for the SP model and* $N = 10^4, 5.10^4$ *Parameters are chosen to optimize capacity under the binomial approximation. Shown are the result of the gaussian approximation without covariance (cyan) and with covariance (magenta) for these parameters. c. Optimized* P_c *as a function of* f *for the SP model at* $N = 10,000$*. The blue curve is for patterns with fluctuations in the number of selective neurons. The red curve is for the same number of selective neurons in all patterns. The black curve is the number of patterns that would be stored if the network were storing the same amount of information as in the case* $N \to +\infty$*. d. Same for the MP model, where parameters have been optimized, but the depression-potentiation ratio is fixed at* $\delta = 1$*.*

2.2.8 Storage capacity with errors

So far we have only considered patterns that are perfectly retrieved. It is difficult to estimate analytically the stability of patterns that are retrieved with errors as it requires analysis of the dynamics at multiple time steps, not only the first time step. However, in simulations we can verify whether a tested pattern is retrieved as a fixed point of the dynamics at a reasonable error level. To quantify the degree of error, we introduce the overlap $m\left(\vec{\sigma}^*, \vec{\xi}^{\mu_0}\right)$ between the network fixed point $\vec{\sigma}^*$ and the tested pattern $\vec{\xi}^{\mu_0}$ (with M selective neurons)

$$m\left(\vec{\sigma}^*, \vec{\xi}^{\mu_0}\right) = \frac{1}{M(1-f)} \sum_{i=1}^{N} \left(\xi_i^{\mu_0} - f\right) \sigma_i^* \tag{2.39}$$

In figure 2.6a we show $P_c(m)$, where $P_c(m=1)$ corresponds to the same definition of P_c used above, and $P_c(m<1)$ is defined similarly but instead of taking into consideration only fixed points that are exactly pattern $\vec{\xi}^{\mu_0}$, we consider all the patterns that lead to a fixed point with an overlap larger than m. In the figure we plot this for $m=1$, $m=0.99$ and $m=0.7$. Taking into account patterns with m smaller than 0.7 leads to values of the capacity similar to the case $m=0.7$ as only a negligible number of tested patterns lead to fixed points with m smaller than 0.7. Note that for sparse coding levels, the overlap m is much more sensitive to missed activation of selective neurons than false activation of non-selective neurons. For instance for $f=0.008$ and $N=10,000$, if a pattern for which $M=80$ is retrieved with 79 selective neurons active and 0 non-selective neurons active, the overlap is $m=0.987$. On the other hand, if it is retrieved with 80 selective neurons active and 50 non-selective neuron active, it leads to an overlap $m=0.995$.
Considering fixed points with errors leads to a substantial increase in capacity from $P_c(m=1)=7,800$ to $P_c(m=0.7)=10,400$ at $f=0.0018$ for instance.
In figure 2.6b, we quantify the storage capacity with $i = \frac{P_c(-f\log_2 f-(1-f)\log_2(1-f))}{N}$. This resembles the information capacity, but note that its meaning is not as clear as in the case $N \to +\infty$ since now P_c is the pattern age at which the probability of retrieval equals $\frac{1}{2}$, and some patterns are retrieved with errors. With this measure, the optimal coding level is larger, $f_{opt} \simeq 0.003$ for i against $f_{opt} \simeq 0.002$ for P_c.

2.2.9 Increase in capacity with inhibition

As we have seen above, the fluctuations in the number of selective neurons in each pattern lead to a reduction in storage capacity in networks of finite size (e.g. figure 2.5c,d). The detrimental effects of these fluctuations can be mitigated by adding

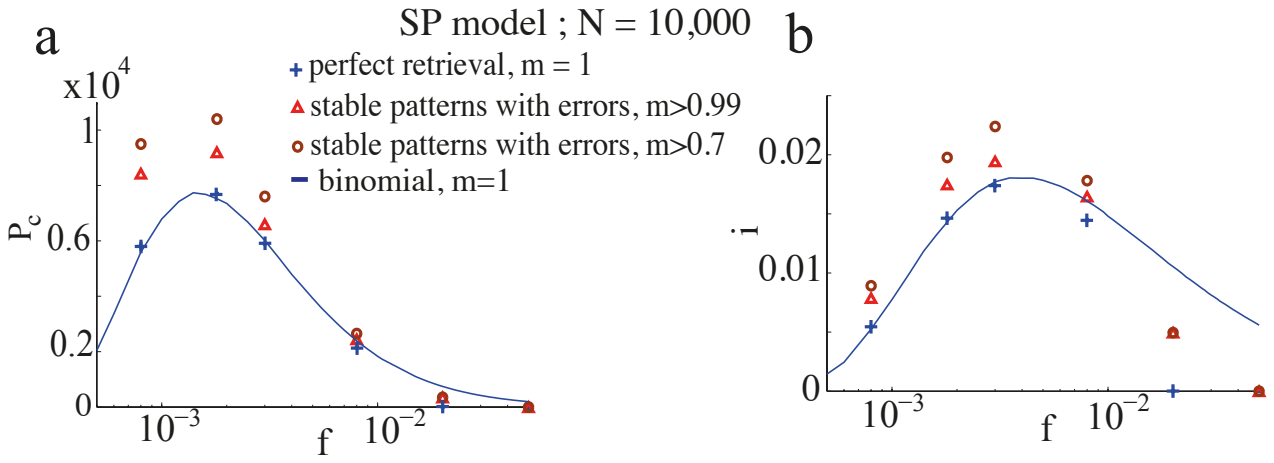

Figure 2.6: *Storage capacity with errors in the SP model. Instead of counting only patterns that are perfectly retrieved, patterns that lead to fixed points of the dynamic overlapping significantly (see text for the definition of the overlap) with the tested memory pattern are also counted. Simulations are done with the same parameters as in figure 2.5a. a. P_c as a function of f. Blue crosses correspond to fixed points that are exactly the stored patterns. Red triangles correspond to fixed points that have an overlap larger than 0.99, and brown circles an overlap larger than 0.7. b. Same as a. but instead of quantifying storage capacity with P_c, it is done with $i = \frac{P_c(-f\log_2 f-(1-f)\log_2(1-f))}{N}$.*

a uniform inhibition η to the network (Amit and Huang, 2010). Using a simple instantaneous and linear inhibitory feed-back, the local fields become

$$h_i = \sum_{k=1}^{N} W_{ik}\xi_k^{\mu_0} - \eta \sum_{k=1}^{N} \xi_k^{\mu_0} - fN\theta \tag{2.40}$$

For infinite size networks, adding inhibition does not improve storage capacity since fluctuations in the number of selective neurons vanish in the large N limit. However, for finite size networks, minimizing those fluctuations leads to substantial increase in storage capacity. When testing the stability of pattern $\vec{\xi^1}$, if the number of selective neurons is unknown, the variance of the field on non-selective neurons is $Nf(g - 2\eta g + \eta^2)$, and $Nf(g_+ - 2\eta g_+ + \eta^2)$ for selective neurons (for small f). The variance for non-selective neurons is minimized if $\eta = g$, yielding the variance obtained with fixed sized patterns. The same holds for selective neurons at $\eta = g_+$. Choosing a value of η between g and g_+ brings the network capacity towards that of fixed size patterns. On figure 2.7, pattern capacity is shown as a function of f in these three scenarios. Optimizing the inhibition η increases the maximal capacity by 54% (green curve) compared to a network with no inhibition (blue curve). Red curve is the capacity without fluctuations.

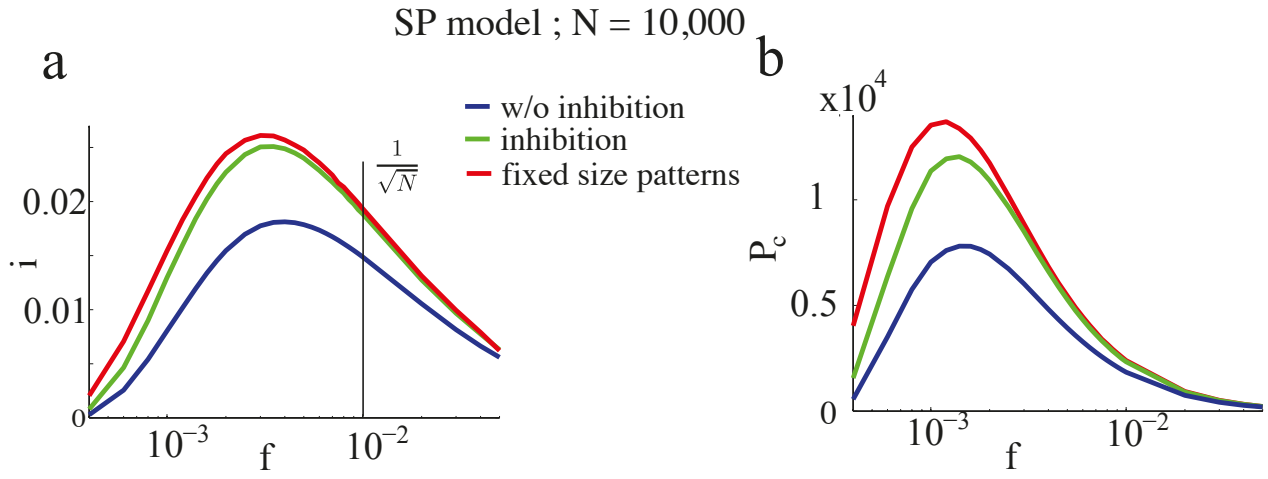

Figure 2.7: *Storage capacity optimized with inhibition in the SP model. Blue is for a fixed threshold and fluctuations in the number of selective neurons per pattern. Green, the fluctuations are minimized using inhibition. Red, without fluctuations in the number of selective neurons per pattern. a. Storage capacity as measured by* $i = \frac{P_c(-f\log_2 f-(1-f)\log_2(1-f))}{N}$. *b. Storage capacity measured by* P_c.

2.2.10 Discussion

We have presented an analytical method to compute the storage capacity of networks of binary neurons with binary synapses in the sparse coding limit. When applied to the classic Willshaw model, in the infinite limit, we find a maximal storage capacity of $\ln(2) = 0.69$ bits/synapse. In previous studies (Willshaw et al., 1969; Nadal, 1991) this value was obtained for a feed-forward network with a single output neuron, with no fluctuations in the number of selective neurons per pattern, and by requiring that the number of errors on silent outputs is of the same order as the number of selective outputs in the whole set of patterns. In the calculations presented here, we have used a different criteria, namely that a given pattern (not all) is exactly a fixed point of the dynamics of the network with a probability that goes to one in the large N limit. Another possible definition would be to require that **all** the P patterns are exact fixed point with probability one. In this case, the capacity, for patterns with a fixed numbers of selective neurons, the capacity drops by a factor 3, $\ln(2)/3 = 0.23$, as already computed by Knoblauch et al (Knoblauch et al., 2010).

We then used this method to study the storage capacity of a network with binary synapses and stochastic learning, in the single presentation (SP) scenario (Amit and Fusi, 1994). The main advantage of this model, compared to the Willshaw model, is its palimpsest property, that allows it to do on-line learning in an ever changing environment. Amit and Fusi showed that the optimal storage capacity was obtained in the sparse coding limit, $f \propto \frac{\ln N}{N}$ and with a balance between the effect of depression and potentiation. The storage capacity of this network has

been further studied for finite size networks in Amit and Huang (2010); Huang and Amit (2011). We have complemented this work by computing analytically the storage capacity in the large N limit. The optimal capacity of the SP model is 0.083 bits/synapse, which is about 9 times lower than the one of the Willshaw model. This decrease in storage capacity is similar to the decrease seen in palimpsest networks with continuous synapses - for example, in the Hopfield model the capacity is about 0.14 bits/synapse, while in a palimpsest version the capacity drops to about 0.05 bits/synapse. The reason for this decrease is that the most recently seen patterns have large basins of attraction, while older patterns have smaller ones. In the Willshaw model, all patterns are equivalent, and therefore they all have vanishing basins of attraction at the maximal capacity.

We have also studied the network in a multiple presentation (MP) scenario, with in which patterns presented to the network are noisy versions of a fixed set of prototypes, in the slow learning limit in which transition probabilities go to zero (Brunel et al., 1998). In the extreme case in which presented patterns are the prototypes, all synaptic weights are initially at zero, and if the synapses do not experience depression, this model is equivalent to the Willshaw model with a storage capacity of 0.69 bits/synapse, which is about 9 times larger than the capacity of the SP model. A more interesting scenario is when depression is present. In this case then the network has generalization properties (it can learn prototypes from noisy versions of them), as well as palimpsest properties (if patterns drawn from a new set of prototypes are presented it will eventually replace a previous set with the new one). We have quantified the trade-off between generalization and storage capacity (see figure 2.3d). For instance, if the noisy patterns have 80% of their selective neurons in common with the prototypes to be learned, the storage capacity is decreased from 0.69 to 0.12bits/synapse.

A key step in estimating storage capacity is deriving an accurate approximation for the distribution of the inputs neurons receive. These inputs are the sum of a large number of binary variables, so the distribution is a binomial if one can neglect the correlations between these variables, induced by the learning process. Amit and Fusi (1994) showed that these correlations can be neglected when $f \ll 1/\sqrt{N}$. Thus, we expect the results with the binomial approximation to be exact in the large N limit. We have shown that a Gaussian approximation of the binomial distribution gives inaccurate results in the sparse coding limit, because the capacity depends on the tail of the distribution, which is not well described by a Gaussian. For larger coding levels ($f \sim 1/\sqrt{N}$), the binomial approximation breaks down because it does not take into account correlations between inputs. Following Amit and Huang (2010) and Huang and Amit (2011), we use a Gaussian approxima-

tion that includes the covariance of the inputs, and show that this approximation captures well the simulation results in this coding level range.

We computed storage capacities for two different learning scenarios. Both are unsupervised, involve a Hebbian-type plasticity rule, and allow for online learning (providing patterns are presented multiple times for the MP model). It is of interest to compare the performance of these two particular scenarios with known upper bounds on storage capacity. For networks of infinite size with binary synapses such a bound has been derived using the Gardner approach (Gutfreund and Stein, 1990). In the sparse coding limit, this bound is $\simeq 0.29$ bits/synapse with random patterns (in which fluctuations in the number of selective neurons per pattern fluctuates), and $\simeq 0.45$ bits/synapse if patterns have a fixed number of selective neurons (Brunel, 1994). We found a capacity of $i_{SP} = 0.083$bits/synapse for the SP model and $i_{MP} = 0.69$ bits/synapse for the MP model, obtained both for patterns with fixed and variable number of selective neurons. The result for the MP model seems to violate the Gardner bound. However, as noticed by Nadal (1991), one should be cautious in comparing these results: in our calculations we have required that a given pattern is stored perfectly with probability one, while the Gardner calculation requires that **all** patterns are stored perfectly with probability one. As mentioned above, the capacity of the Willshaw and MP models drops to $i_{opt} = 0.23$bits/synapse in the case of fixed-size patterns, if one insists that **all** patterns should be stored perfectly, which is now consistent with the Gardner bound. This means that the MP model is able to reach a capacity which is roughly half the Gardner bound, a rather impressive feat given the simplicity of the rule. Note that supervised learning rules can get closer to these theoretical bounds (Baldassi et al., 2007).

We have also studied finite-size networks, in which we defined the capacity as the number of patterns for which the probability of exact retrieval is at least 50%. We found that networks of reasonable sizes have capacities that are far from the large N limit. For networks of sizes $10^4 - 10^6$ storage capacities are reduced by a factor 3 or more (see Fig. 2.4). These huge finite size effects can be understood by the fact that the leading order corrections in the large N limit are in $\frac{\ln(\ln N)}{\ln N}$ - and so can never be neglected unless N is an astronomical number (see Methods A). A large part of the decrease in capacity when considering finite-size networks is due to fluctuations in the number of selective neurons from pattern to pattern. In the last section, we have used inhibition to minimize the effect of these fluctuations. For instance, for a network of $N = 10,000$ neurons learning in one shot, inhibition allows to increase capacity from $P = 7,800$ to $P = 12,000$. For finite size networks, memory patterns that are not perfectly retrieved can still lead to fixed points

where the activity is significantly correlated with the memory patterns. We have investigated with simulations how allowing errors in the retrieved patterns modifies storage capacity, for instance for $N = 10,000$, capacity increases from $P = 7,800$ to $P = 10,400$.

Models with binary synapses are a simple alternative to models with continuous synapses that have infinite resolution. It has been argued that it would be difficult to built robust synapses that can take continuous values with known biophysical processes Brunel (2003). Also for practical applications, discrete states synapses seem easier to implement, and there is experimental evidence from minimal stimulations in hippocampal slices that some synapses can be well described by binary variables Petersen et al. (1998); O'Connor et al. (2005). In this study we have used binary neurons, which allowed us to track analytically the storage properties of the networks. It remains to be investigated how these results will generalize to networks of more realistic neurons. In strongly connected networks of spiking neurons operating in the balanced mode (van Vreeswijk and Sompolinsky, 1996; Amit and Brunel, 1997b; van Vreeswijk and Sompolinsky, 1998; Brunel, 2000a), the presence of ongoing activity presents strong constraints on the viability of sparsely coded selective attractor states. This is because 'non-selective' neurons are no longer silent, but are rather active at low background rates, and the noise due to this background activity can easily wipe out the selective signal (Amit and Brunel, 1997b; Roudi and Latham, 2007). In fact, simple scaling arguments in balanced networks suggest the optimal coding level would become $f \sim 1/\sqrt{N}$ (Brunel, 2003; van Vreeswijk and Sompolinsky, 2005). The learning rules we have considered in this paper lead to a vanishing information stored per synapse with this scaling. Finding an unsupervised learning rule that achieves a finite information capacity in the large N limit remains an open question. However, the results shown here show that for networks of realistic sizes, the information capacity at such coding levels is in fact not very far from the optimal one that is reached at lower coding levels (see vertical lines in Fig. 2.5-2.7). A priori, a main drawback of the models we have studied here is the requirement of ultra-sparse coding level, $f \propto \frac{\ln N}{N}$, for having good storage capacity. To compare this quantity with experiments, one first has to set a network size. Fully-connected networks we have studied here do not seem appropriate to study networks larger than $0.5mm^3$ for which connection probability can hardly be considered homogeneous (see e.g. Hellwig 2000 for cortical networks). A network of this dimension contains approximately $N = 10,000$ neurons, and for this value of N we have seen that for coding levels up to 10^{-2} one can get close to the optimal capacity. This is one order of magnitude larger than one would have guessed from the scaling $f \propto \frac{\ln N}{N}$ and falls in a reasonable range

of coding level.

The SP and MP models investigated in this paper can be thought of as minimal models for learning in hippocampus and neocortex. The SP model bears some resemblance to the function of hippocampus, which is supposed to keep a memory of recent episodes that are learned in one shot, thanks to highly plastic synapses. The MP model relates to the function of neocortex, where a longer-term memory can be stored, thanks to repeated presentations of a set of prototypes that occur repeatedly in the environment, and perhaps during sleep under the supervision of the hippocampus. The idea that hippocampal and cortical networks learn on different time scales has been exploited in several modeling studies (Alvarez and Squire, 1994; Káli and Dayan, 2004; Roxin and Fusi, 2013), in which the memories are first stored in the hippocampus and then gradually transferred to cortical networks. It would be interesting to extend the type of analysis presented here to coupled hippocampo-cortical networks with varying degrees of plasticity.

2.2.11 Methods

A - Capacity calculation for infinite size networks

We are interested at retrieving pattern $\vec{\xi^\mu}$ that has been presented during the learning phase. We set the network in this state $\vec{\sigma} = \vec{\xi^\mu}$ and ask whether the network remains in this state while the dynamics (2.9) is running. At the first iteration, each neuron i is receiving a field

$$h_i = \sum_{j=1}^{N} W_{ij}\xi_j^\mu = \sum_{k=1}^{M} X_k^i \tag{2.41}$$

Where M+1 is the number of selective neurons in pattern $\vec{\xi^\mu}$, with $M = O(\ln N)$ and $N \to +\infty$. We recall that $g_+ = \mathbb{P}(W_{ij} = 1|\xi_i^\mu = \xi_j^\mu = 1)$ and $g = \mathbb{P}(W_{ij} = 1|(\xi_i^\mu, \xi_j^\mu) \neq (1,1))$. Thus X_k^i is a binary random variable which is 1 with probability, either g_+ if i is a selective neuron (sites i such that $\xi_i^\mu = 1$), or g if i is a non-selective neuron (sites i such that $\xi_i^\mu = 0$). Neglecting correlations between W_{ij_1} and W_{ij_2} (it is legitimate in the sparse coding limit we are interested in, see (Amit and Fusi, 1994)), the X_k^i's are independent and the distribution of the field on selective neurons can be written as

$$\begin{aligned}\mathbb{P}(h_i^s = S) &= \binom{M}{S} g_+^S (1-g_+)^{M-S} \\ &= \exp\left[-M\Phi\left(g_+, \frac{S}{M}\right) - \frac{1}{2}\ln\left(S\left(1-\frac{S}{M}\right)\right) - \frac{1}{2}\ln(2\pi)\right] \end{aligned} \tag{2.42}$$

where we used Stirling formula for $M, S \gg 1$, with Φ defined in (2.19). For non-selective neurons

$$\begin{aligned}\mathbb{P}\left(h_i^n = S\right) &= \binom{M}{S} g^S (1-g)^{M-S} \\ &= \exp\left[-M\Phi\left(g, \frac{S}{M}\right) - \frac{1}{2}\ln\left(S\left(1-\frac{S}{M}\right)\right) - \frac{1}{2}\ln(2\pi)\right] \quad (2.43)\end{aligned}$$

Now write

$$\begin{aligned}\mathbb{P}(h_i^s \leq \theta f N) &= \mathbb{P}(h_i^s = \theta f N) \sum_{S \leq \theta f N} \frac{\mathbb{P}(h_i^s = S)}{\mathbb{P}(h_i^s = \theta f N)} \\ \mathbb{P}(h_i^n \geq \theta f N) &= \mathbb{P}(h_i^n = \theta f N) \sum_{S \geq \theta f N} \frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = \theta f N)} \quad (2.44)\end{aligned}$$

In the limit $N \to +\infty$ we are considering in this section, and if $Mg < fN\theta < Mg_+$, the sums corresponding to the probabilities $\mathbb{P}(h_i^s \leq fN\theta), \mathbb{P}(h_i^n \geq fN\theta)$ are dominated by their first term (corrections are made explicit in the following section). Keeping only higher order terms in M in equations (2.42) and (2.43), we have:

$$\mathbb{P}(h_i^s \leq fN\theta) \simeq \exp(-M\Phi(g_+, \theta_M)) \quad (2.45)$$

and

$$\mathbb{P}(h_i^n \geq fN\theta) \simeq \exp(-M\Phi(g, \theta_M)), \quad (2.46)$$

yielding equation (2.77) with $\theta_M = \theta\frac{fN}{M} = O(1)$. Note that with the coding levels we are considering here ($f \propto \frac{\ln N}{N}$), M is of order $\ln N$. When the number of selective neurons per pattern is fixed at fN, we choose $M\theta$ for the activation threshold and these equations become:

$$\begin{aligned}X_s &= -\ln N \beta \Phi(g_+, \theta) + O(\ln\ln N) \\ X_n &= \ln N(-\beta\Phi(g, \theta) + 1) + O(\ln\ln N) \quad (2.47)\end{aligned}$$

where $\beta = f\frac{N}{\ln N}$

For random numbers of selective neurons we need to compute the average over M: $\overline{\mathbb{P}}_{ne}(N) = \Sigma_{M=0}^{N} \mathbb{P}(M)\mathbb{P}_{ne}(M, N)$. Since M is distributed according to a binomial of average Nf and variance $Nf(1-f) \simeq Nf$, for sufficiently large Nf, this can be approximated as $M = fN + z\sqrt{fN}$ where z is normally distributed:

$$\overline{\mathbb{P}}_{ne}(N) = \int_{-\infty}^{+\infty} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \exp(-\exp(X_s(z, N)) - \exp(X_n(z, N))) \quad (2.48)$$

with

$$\begin{aligned} X_s(z,N) &= -M\Phi\left(g_+, \frac{\theta}{1+\frac{z}{\sqrt{fN}}}\right) + O(\ln\ln N) \\ &\simeq -\beta \ln N \left[\Phi(g_+,\theta) + \frac{z}{\sqrt{fN}}\left(\Phi(g_+,\theta) - \theta\frac{\partial\Phi}{\partial\theta}(g_+,\theta)\right)\right] + O(\ln\ln N) \\ &\simeq -\beta \ln N \left[\Phi(g_+,\theta) + \frac{z}{\sqrt{fN}}\ln\frac{1-\theta}{1-g_+}\right] + O(\ln\ln N) \end{aligned} \tag{2.49}$$

and

$$\begin{aligned} X_n(z,N) &= -M\Phi\left(g, \frac{\theta}{1+\frac{z}{\sqrt{fN}}}\right) + \ln N + O(\ln\ln N) \\ &\simeq \ln N \left[1 - \beta\left(\Phi(g,\theta) + \frac{z}{\sqrt{fN}}\ln\frac{1-\theta}{1-g}\right)\right] + O(\ln\ln N) \end{aligned} \tag{2.50}$$

When N goes to infinity, we bring the limit into the integral in equation (2.48) and obtain

$$\begin{aligned} \lim_{N\to+\infty} \overline{\mathbb{P}}_{ne}(N) &= \int_{-\infty}^{+\infty} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{\pi}} \lim_{N\to+\infty} \exp\left[-\exp(X_s(z,N)) - \exp(X_n(z,N))\right] \\ &= \Theta(\Phi(g_+,\theta))\Theta(-\beta\Phi(g,\theta)+1) \end{aligned} \tag{2.51}$$

where Θ is the Heaviside function. Thus in the limit of infinite size networks, the probability of no error is a step function. The first Heaviside function implies that the only requirement to avoid errors on selective neurons is to have a scaled activation threshold θ below g_+. The second Heaviside function implies that, depending on β, θ has to be chosen far enough from g. There is no constraint on the distance between θ and g_+ because in each pattern there are only $O(\ln N)$ selective neurons while there are $O(N)$ background neurons. The above equation allows to derive the inequalities (2.25) and (2.26).

B - Capacity calculation for finite-size networks.

We now turn to a derivation of finite-size corrections for the capacity. Here we show two different calculations. In the first calculation, we derive Eq. (2.38), taking into account the leading-order correction term in Eq. (2.49). This allows us to compute the leading-order correction to the number of patterns P that can be stored for a given set of parameters. However, it does not predict accurately the storage capacity of the large-size but finite networks that we simulated. In the second calculation presented, we focus on computing the probability of no error in a given pattern $\mathbb{P}_{ne}$, including a next-to-leading-order correction.

Equation (2.38) is derived for a fixed set of parameters, assuming that the set of active neurons have a fixed size, and that the activation threshold θ has been chosen large enough such that the probability to have non-selective neurons activated is small. From the Stirling expansion, adding the first finite-size correction term in Eq. (2.47), we get

$$X_s \simeq -\ln N \beta_M \Phi(g_+, \theta) + \frac{1}{2}\ln(\ln N) \tag{2.52}$$

with $\beta_M = M/\ln N$. For large N, the number of stored patterns P can be increased until $g_+(P) \gtrapprox \theta$. Setting $g_+ = \theta + \epsilon$, an expansion of Φ in ϵ allows to write

$$X_s \simeq -\ln N \beta_M \frac{\epsilon^2}{2\theta(1-\theta)} + \frac{1}{2}\ln(\ln N) \tag{2.53}$$

The P patterns are correctly stored as long as $X_s \ll -1$. This condition is satisfied for $\epsilon < \sqrt{\frac{\theta(1-\theta)}{\beta_M}\frac{\ln(\ln N)}{\ln N}}$. For the SP model, we can deduce which value of P yields this value of ϵ (see Eq. (2.32)). This allows to derive Eq. (2.38),

$$\begin{aligned} P &= \frac{g}{q_+\beta^2}\ln\left(\frac{q_+(1-g)}{\theta-g}\right)\frac{N^2}{(\ln N)^2} \times \\ &\quad \left[1 - \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\beta_M}(\theta-g)\ln\left(\frac{q_+(1-g)}{\theta-g}\right)}\sqrt{\frac{\ln(\ln N)}{\ln N}} + o\left(\frac{\ln(\ln N)}{\ln N}\right)\right] \end{aligned} \tag{2.54}$$

We now turn to a calculation of the probability of no error on a given pattern $\mathbb{P}_{ne}$, taking into account the next-to-leading order correction of order one, in addition to the term of order $\ln\ln N$ in Eq. (2.47). This is necessary to predict accurately the capacity of realistic size networks (for instance for $N = 10,000$, $\ln(\ln N) \simeq 2 = O(1)$). $\mathbb{P}_{ne}(M)$ is computed for a memory pattern with M selective neurons. The estimation of $\overline{\mathbb{P}}_{ne}$ used in the figures is obtained by averaging over different values of M, with M drawn from a binomial distribution of mean fN.

We first provide a more detailed expansion of the sums in equation (2.44). Setting $S = fN\theta + k$, with the Taylor expansions:

$$M\Phi\left(g, \theta_M + \frac{k}{M}\right) = M\Phi(g,\theta_M) + k\frac{\partial\Phi}{\partial\theta}(g,\theta_M) + \frac{k^2}{2M}\frac{\partial^2\Phi}{\partial\theta^2}(g,\theta_M) + O\left(\frac{1}{M^2}\right) \tag{2.55}$$

$$\ln\left(S\left(1-\frac{S}{M}\right)\right) = \ln(M\theta_M(1-\theta_M)) + \frac{k}{M}\Delta\theta_M^{-1} + O\left(\frac{1}{M^2}\right) \tag{2.56}$$

where $\theta_M = \theta \frac{fN}{M}$ and $\Delta\theta_M^{-1} = \frac{1}{\theta_M} - \frac{1}{1-\theta_M}$. Using (2.43) we can rewrite:

$$\sum_{S \geq fN\theta} \frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = fN\theta)} = \sum_{k=0}^{M-fN\theta} \exp\Big[-k\frac{\partial \Phi}{\partial \theta}(g, \theta_M) - \frac{1}{M}\left(\frac{k^2}{2}\frac{\partial^2 \Phi}{\partial \theta^2}(g, \theta_M) - k\Delta\theta_M^{-1}\right) + O\left(\frac{1}{M^2}\right)\Big] \tag{2.57}$$

In the cases we consider, we will always have $\frac{\partial \Phi}{\partial \theta}(g, \theta_M) \neq 0$ so that we can consider only the term of order 1 in M. The sum is now geometric, and we obtain

$$\sum_{S \geq fN\theta} \frac{\mathbb{P}(h_i^n = S)}{\mathbb{P}(h_i^n = fN\theta)} = \frac{1}{1 - \exp\left(-\frac{\partial \Phi}{\partial \theta}(g, \theta_M)\right)} + o(1) \tag{2.58}$$

The same kind of expansion can be applied for the selective neurons. Again if we are in a situation where $\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M) \neq 0$,

$$\sum_{S \leq fN\theta} \frac{\mathbb{P}(h_i^s = S)}{\mathbb{P}(h_i^s = fN\theta)} = \frac{1}{1 - \exp\left(\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M)\right)} + o(1) \tag{2.59}$$

When g_+ close to θ and thus $\frac{\partial \Phi}{\partial \theta}(g_+, \theta_M) \simeq 0$, we are then left with:

$$\sum_{k=0}^{\theta M} \exp\left[-\frac{1}{M}\left(\frac{k^2}{2}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M) - k\Delta\theta_M^{-1}\right)\right] \tag{2.60}$$

$$= \exp\left[\frac{1}{8M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)(\Delta\theta_M^{-1})^2\right] \sum_{k=0}^{+\infty} \exp\left[-\frac{(k - \Delta\theta_M^{-1})^2}{2M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)\right] + o(1)$$

$$= \int_0^{+\infty} dt e^{-\frac{(t-(\Delta\theta_M^{-1}))^2}{2M}\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)} + o(1)$$

$$= \sqrt{\frac{\pi}{2}\frac{M}{\frac{\partial^2 \Phi}{\partial \theta^2}(g_+, \theta_M)}} + o(1) \tag{2.61}$$

When g_+ is too close to θ, which is the case for the optimal parameters in the large N limit, we need to use (2.61). It only contributes a term of order $\ln \ln N$ in X_s and does not modify our results. In the figures of Sections 6 and 7, we use (2.59), which gives from (2.44) and (2.42), (2.43) and (2.59),(2.58):

$$\begin{aligned}
\mathbb{P}(h_i^s \leq fN\theta) &= \exp\Big[\ln N(-\beta_M \Phi(g_+,\theta_M)) - \frac{1}{2}\ln\ln N - \\
&\quad \frac{1}{2}\ln\left(2\pi\theta_M(1-\theta_M)[1-\exp(\frac{\partial\Phi}{\partial\theta}(g_+,\theta_M))]^2\right)\Big] \qquad (2.62)\\
\mathbb{P}(h_i^n \geq fN\theta) &= \exp\Big[\ln N(-\beta_M \Phi(g,\theta_M)) - \frac{1}{2}\ln\ln N - \\
&\quad \frac{1}{2}\ln\left(2\pi\theta_M(1-\theta_M)[1-\exp(-\frac{\partial\Phi}{\partial\theta}(g,\theta_M))]^2\right)\Big] \qquad (2.63)
\end{aligned}$$

The probability of no error is

$$\begin{aligned}
\mathbb{P}_{ne} &= (1-\mathbb{P}(h_i^s \leq fN\theta))^M (1-\mathbb{P}(h_i^n \geq fN\theta))^{N-M} \\
&= \exp(-\exp X_s - \exp X_n) \qquad (2.64)
\end{aligned}$$

which leads to equations (2.37)

$$\begin{aligned}
X_s &= -\beta_M \Phi\left(g_+,\theta_M\right)\ln N + \frac{1}{2}\ln\ln N - \\
&\quad \frac{1}{2}\ln\left[\frac{\left(1-\exp(\frac{\partial\Phi}{\partial\theta}(g_+,\theta_M))\right)^2 2\pi\theta_M(1-\theta_M)}{\beta_M}\right] + o(1)\\
X_n &= (-\beta_M\Phi(g,\theta_M)+1)\ln N - \frac{1}{2}\ln\ln N - \\
&\quad \frac{1}{2}\ln\left[\left(1-\exp(-\frac{\partial\Phi}{\partial\theta}(g,\theta_M))\right)^2 2\pi\theta_M(1-\theta_M)\beta_M\right] + o(1)
\end{aligned}$$

C - Gaussian approximation of the fields distribution.

For a fixed number $M+1$ of selective neurons in pattern $\vec{\xi}^1$, approximating the distribution of the fields on background neurons h_i^n and selective neurons h_i^s with a gaussian distribution gives:

$$\mathbb{P}^G(h_i^n = S) = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{(S-\mu_b)^2}{2\sigma_n^2}\right) \qquad (2.65)$$

where

$$\mu_b = Mg\,,\ \sigma_n^2 = Mg(1-g) \qquad (2.66)$$

and

$$\mathbb{P}^G(h_i^s = S) = \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left(-\frac{(S-\mu_f)^2}{2\sigma_s^2}\right) \qquad (2.67)$$

where

$$\mu_f = Mg_+ \, , \, \sigma_s^2 = Mg_+(1-g_+) \tag{2.68}$$

The probability that those fields are on the wrong side of the threshold are:

$$\mathbb{P}^G(h_i^n \geq fN\theta) = \int_{fN\theta}^{+\infty} P^G(h_i^n = z)dz \tag{2.69}$$

and

$$\mathbb{P}^G(h_i^s \leq fN\theta) = \int_{-\infty}^{fN\theta} P^G(h_i^s = z)dz \tag{2.70}$$

Following the same line of calculation than in Methods A, and keeping only terms that are relevant in the limit $N \to +\infty$, the probability that there is no error is given by:

$$\Theta(\Phi^G(g_+, \theta))\Theta(-\beta\Phi^G(g, \theta) + 1) \tag{2.71}$$

where the rate function Φ^G is

$$\Phi^G(x, \theta) = \frac{(\theta - x)^2}{2x(1-x)} \tag{2.72}$$

Calculations with the binomial versus the gaussian approximation differ only in the form of Φ. Finite size terms can be taken into account in the same way it is done in Methods B for the binomial approximation.

In all above calculations we assumed that fields are sums of independent random variables (2.41). For small f correlations are negligible (Amit and Fusi, 1994; Amit and Huang, 2010). It is possible to compute the covariances between the terms of the sum (see Eq. (3.9) in Amit and Huang (2010)), and take them into account in the gaussian approximation. This can be done using

$$\sigma_n^2 = Mg(1-g) + M(M-1)\gamma \tag{2.73}$$

$$\sigma_s^2 = Mg_+(1-g_+) + M(M-1)\gamma \tag{2.74}$$

in Eqs. (2.65),(2.67), where

$$\gamma = f\frac{\delta^2}{2(1+\delta)^3} \tag{2.75}$$

2.3 Synapses with multiple states

We now consider synaptic weights that take K discrete values, $W_{ij} \in \{0, ..., K-1\}$. Again we study the scenario in which patterns are presented only once to the network (SP model) and the scenario in which patterns are presented multiple times (MP model). Also, we study two different models of synaptic connection, a first one modeling a single synaptic contact that can be in K different states. In this case, upon pattern presentation, synaptic weights can be increased or decreased by at most one unit, as shown by the cartoon in figure 2.8A. And a second model for a connection made of $K-1$ binary contacts, that are independently updated upon pattern presentation, thus allowing transition of the total synaptic weight value (the sum of the values of the binary contacts) between non-neighboring values, this is shown by the cartoon in figure 2.8D. This last model of connection weights is inspired by the data of Kalisman et al. (2005) showing that in the somato-sensory cortex of rats, if a pre-synaptic neuron connects to a post-synaptic one, its axon forms an average of 5.6 contacts on the dendritic tree of the post-synaptic neuron. These models are studied in the limit $N \to +\infty$.

2.3.1 Capacity calculation

The method to compute the capacity is similar to what is described in the section 2.2.3. The only change is that the rate function ϕ is now a function of K variables instead of 2. More precisely, when testing the stability of a pattern $\vec{\xi}^{\mu_0}$ with M selective neurons ($M = O(\ln N)$), we need to compute the probability that it is perfectly stable (see Method section for derivation)

$$\begin{aligned} \mathbb{P}_{ne} &= \left(1 - \mathbb{P}(h_i \leq M\theta | \xi_i^{\mu_0} = 1)\right)^M \left(1 - \mathbb{P}(h_i \geq M\theta | \xi_i^{\mu_0} = 0)\right)^{N-M} \\ &= \exp(-\exp(X_s) - \exp(X_n)) \end{aligned} \tag{2.76}$$

with

$$\begin{aligned} X_s &= -M\Phi_\theta^K(\vec{P}^+) + o(M) \\ X_n &= -M\Phi_\theta^K(\vec{P}^-) + \ln N + o(M). \end{aligned} \tag{2.77}$$

Φ_θ^K is given in the method section and $\vec{P}^+ = \left(P_0^+, ..., P_{K-1}^+\right)$ describes the state of the synapses (i, j) that could have been potentiated when the tested pattern μ_0 was presented to the network; $\vec{P}^- = \left(P_0^-, ..., P_{K-1}^-\right)$ to describe the other synapses. With the notation

$$P_k^+ = P(W_{ij} = k|\xi_i^{\mu_0} = \xi_j^{\mu_0} = 1) \text{ and } P_k^- = P(W_{ij} = k|(\xi_i^{\mu_0}, \xi_j^{\mu_0}) \neq (1,1)) \quad (2.78)$$

where W describes the synaptic matrix at the end of the learning phase (i.e. once $P = \frac{\alpha}{f^2}$ patterns have been presented subsequently to the presentation of pattern μ_0 for the SP model, or presenting an infinity of sequences of the P patterns to store for the MP model).

Note that again, as we will work in the sparse coding limit and keep the depression-potentiation ratio of order 1, the depression probability will be small thus $P(W_{ij}(\mu) = k|(\xi_i^\mu, \xi_j^\mu) = (0,1)) \simeq P(W_{ij}(\mu) = k|(\xi_i^\mu, \xi_j^\mu) = (0,0))$.

In the large N limit, the condition for stability is expressed as before

$$\Phi_\theta^K(\vec{P}^+) > \frac{\ln(\ln N)}{\beta \ln N} \quad (2.79)$$

$$\Phi_\theta^K(\vec{P}^-) > \frac{1}{\beta} \quad (2.80)$$

and the optimal capacity is obtained by finding parameters (β, θ and all the parameters that determine $\vec{P}^\pm$) that saturates these inequalities and using the expression of the information capacity (2.15), $i = \frac{\alpha}{\beta \ln 2}$.

2.3.2 K-states synapses in the single-presentation (SP) learning scenario

Multi-stable synapses in the SP learning scenario

In this section, each pattern has to be learnt in one shot and only transitions between neighbor states are allowed (see figure 2.8A). The evolution of synapses is then described by

$$\begin{bmatrix} P(W_{ij}(\mu+1)=0) \\ P(W_{ij}(\mu+1)=1) \\ \vdots \\ \\ P(W_{ij}(\mu+1)=K-2) \\ P(W_{ij}(\mu+1)=K-1) \end{bmatrix} = \begin{bmatrix} 1-a & b & 0 & & \\ a & 1-a-b & b & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & \ddots & a & 1-a-b & b \\ & & 0 & a & 1-b \end{bmatrix} \times \begin{bmatrix} P(W_{ij}(\mu)=0) \\ P(W_{ij}(\mu)=1) \\ \vdots \\ \\ P(W_{ij}(\mu)=K-2) \\ P(W_{ij}(\mu)=K-1) \end{bmatrix} \tag{2.81}$$

with $a = f^2 q_+$ and $b = 2f(1-f)q_-$. Note that as in the case of 2 states synapses, we work in a regime where $f \propto \frac{\ln N}{N}$ and $\delta = \frac{2f(1-f)q_-}{f^2 q_+} = O(1)$. From this equation, $\vec{P}^{\pm}$ can be expressed by diagonalizing the transition matrix (see Methods), and the capacity can be computed. In figure 2.8B we plot the optimal capacity as a function of the number of synaptic states. It slightly increases from $K = 2$ to $K = 3$ and rapidly saturates.

The optimal capacities are always reached for $q_+ = 1$. Looking at the value of α that optimizes capacity gives a way to see how the network is dealing with patterns, as $\alpha = Pf^2$ is the average number of potentiating events a synapse experiences, between the time we probe the stability of the tested pattern and the time it has been presented. For $K = 2$, the network operates in a regime where $\alpha = 0.14$, while for $K = 10$, $\alpha = 4.8$ and for $K = 25$, $\alpha = 30.9$. For these values of K, the optimal δ are 2.57, 1.06 and 1.01.

In figure 2.9A we plot the values of P_k^-, which corresponds, in the sparse coding limit, to the probability that a randomly taken weight has a value k. This distribution can be written simply as it corresponds to the stationary distribution of synaptic states given by the eigenvector associated to the eigenvalue equals to 1 of the transition matrix in (2.81) (see derivation of $\vec{P}^{\infty}$ in Methods section).

$$P_k^- = \frac{\delta^{K-k}}{\sum_{n=0}^{K} \delta^n} \tag{2.82}$$

In the case $K = 2$, the optimal capacity is reached for parameters such that 72% of the synapses are silent. When increasing K, this fraction is decreasing to 30%

at $K = 5$ and to 4% at $K = 25$.

Poly-synaptic contacts in the SP learning scenario

Here each weight is the sum of K binary variables that are updated independently at each pattern presentation. This models the fact that an axon usually makes multiple contacts onto a single dendritic tree (see figure 2.8C) as shown by Kalisman et al. (2005) where they measure a 5.6 ± 3.57 synaptic contacts.
Here W_{ij} can take $K+1$ values. The evolution of each weight can then be described by the following Markov process

$$\begin{bmatrix} P(W_{ij}(t+1)=0) \\ P(W_{ij}(t+1)=1) \\ \vdots \\ \\ P(W_{ij}(t+1)=K-1) \\ P(W_{ij}(t+1)=K) \end{bmatrix} = \mathscr{M} \times \begin{bmatrix} P(W_{ij}(t)=0) \\ P(W_{ij}(t)=1) \\ \vdots \\ \\ P(W_{ij}(t)=K-1) \\ P(W_{ij}(t)=K) \end{bmatrix}$$

with

$$\mathscr{M}_{ij} = P(i \to j) = \begin{cases} f^2 \binom{K-j}{i-j} q_+^{i-j} (1-q_+)^{K-i} + 2f(1-f)\binom{i}{j} q_-^{j-i} (1-q_-)^i \text{ for } i \neq j \\ 1 - \sum_{k=0, k\neq i}^{K} P(i \to k) \text{ for } i = j \end{cases} \tag{2.83}$$

For instance for $K = 2$

$$\mathscr{M} = \begin{bmatrix} 1-a(2-q_+) & b & bq_- \\ 2a(1-q_+) & 1-a-b & 2b(1-q_-) \\ aq_+ & a & 1-b(2-q_-) \end{bmatrix} \simeq \begin{bmatrix} 1-a(2-q_+) & b & 0 \\ 2a(1-q_+) & 1-a-b & 2b \\ aq_+ & a & 1-2b \end{bmatrix}$$

We diagonalize such matrices for various values of K which allows us to compute the values of $\vec{P}^{\pm}$ and thus capacity. In figure 2.8D we show the maximal capacity for different values of K. In this case the number of bits stored per pair of neurons is significantly increased, almost a factor 2 from $K = 1$ to $K = 8$. The maximal capacities are still reached for $q_+ = 1$. The value of α optimizing are always smaller than 1 from $\alpha = 0.14$ for $K = 2$, to $\alpha = 0.21$ for $K = 8$. As for the case of a single multi-stable contact, the optimal value of δ decreases with increasing K, with $\delta = 1.09$ for $K = 8$.
We also plot the values of P_k^- describing the statistics of the synaptic matrix in figure 2.9C. In this case, we did not find analytical expressions for the eigenvalues

and eigenvectors of the transition matrix, that was diagonalized using Maple. Thus we don't have analytical expressions for the P_k^-. Similarly to the previous case, the fraction of silent synapses is decreasing with K, 30% for $K = 4$ and 13% for $K = 9$.

2.3.3 K-states synapses in the multiple-presentation (MP) learning scenario

Here we generalize the work of Brunel et al. (1998) to the case of synapses with K contacts. As before to compute capacity we first need to express $\vec{P}^{\pm}$ as a function of the different parameters characterizing the synapses, the statistics of patterns presented during learning and the total number of pattern in the sequence P, again defined through α with $P = \frac{\alpha}{f^2}$. We will sketch how to express these vectors (details can be found in the Method section) and then use these expressions to compute the optimal capacity for multi-state synapses and poly-synaptic contacts. The evolution of the state of a given synapse W_{ij} at time t (where here t corresponds to the number of patterns that have been presented to the network) is still described by a vector $(P(W_{ij}(t) = k))_{k=0...K-1}$, that evolves at each pattern presentation according to

$$\begin{bmatrix} P(W_{ij}(t+1)=0) \\ P(W_{ij}(t+1)=1) \\ \vdots \\ \\ P(W_{ij}(t+1)=K-1) \\ P(W_{ij}(t+1)=K) \end{bmatrix} = \mathscr{M}_{ij}(t) \times \begin{bmatrix} P(W_{ij}(t)=0) \\ P(W_{ij}(t)=1) \\ \vdots \\ \\ P(W_{ij}(t)=K-1) \\ P(W_{ij}(t)=K) \end{bmatrix}$$

where $\mathscr{M}_{ij}(t)$ is a matrix that depends on the activity ξ_i^t and ξ_j^t of the pattern presented at time t and on the parameters governing plasticity q_+ and q_-. In the SP model, this matrix was averaged over all the possible configurations of the activity of neurons i and j to give the matrix $\mathscr{M}$. Here we consider the average of this equation over all the possible sequences of presentation of the P patterns we are trying to store

$$\langle \vec{P}_{ij}(t+1) \rangle = \langle \mathscr{M}_{ij}(t) \rangle \langle \vec{P}_{ij}(t) \rangle \tag{2.84}$$

where $\langle \mathscr{M}_{ij}(t) \rangle$ depends on

$$C_{ij} = \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \quad \text{and} \quad D_{ij} = \sum_{\mu=1}^{P} \xi_i^\mu (1-\xi_j^\mu) + (1-\xi_i^\mu)\xi_j^\mu \tag{2.85}$$

the total number of patterns that tend to potentiate/depress synapse W_{ij}. We compute analytically $\left(\langle \vec{P}_{ij}(+\infty)\rangle\right)_k$ the probability that W_{ij} is k when an infinite number of sequences have been presented, by diagonalizing the averaged transition matrix. Then this allows to compute the expression of $\vec{P}^{\pm}$ (Brunel et al., 1998)

$$\begin{aligned} P_k^- &= \frac{1}{N(N-1)} \sum_{i \neq j} \langle P_{ij}^k(+\infty)\rangle (C_{ij}, D_{ij}) \\ &= \sum_{\Pi,\Delta=0}^{P} \Psi(\Pi,\Delta) \langle P_{ij}^k(+\infty)\rangle (\Pi,\Delta) \end{aligned} \tag{2.86}$$

where $\Psi(\Pi,\Delta) \underset{f\to 0}{\simeq} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \delta\left(\Delta - \frac{2\alpha}{f}\right)$ is the probability that a given synapse experiences Π potentiating activity events and Δ depressing activity events during a sequence of presentation of the P patterns. The distribution of fields can then be computed with equation (2.90). The distribution of synaptic values for synapses W_{ij} such that $\xi_i^{\mu_0} = \xi_j^{\mu_0} = 1$ is given by

$$\begin{aligned} P_k^+ &= \frac{1}{Nf(Nf-1)} \sum_{i \neq j} \langle P_{ij}^k(+\infty)\rangle (C_{ij}, D_{ij}) \xi_i^{\mu_0} \xi_j^{\mu_0} \\ &= \sum_{\Pi,\Delta=0}^{P-1} \Psi_+(\Pi,\Delta) \langle P_{ij}^k(+\infty)\rangle (\Pi+1, \Delta+1) \end{aligned} \tag{2.87}$$

where $\Psi_+(\Pi,\Delta) \underset{f\to 0}{\simeq} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \delta\left(\Delta - \frac{2\alpha}{f}\right)$ is the probability that a given synapse experiences Π potentiating activity events and Δ depressing activity events during a sequence of presentation of the $P-1$ patterns (where the tested pattern $\vec{\xi}^{\mu_0}$ has been already counted as a potentiating event).

Multi-stable synapses in the MP learning scenario

The expressions of $\vec{P}^{\pm}$ are given by equations (2.109),(2.108) in the Method section as a function of the depression-potentiation ratio δ, the amount of noise in the presented patterns x (see (2.34) for the definition of x), and the number of stored patterns via α. We applied the general method presented above to estimate the maximal capacity. Figure 2.8C shows how the optimal capacity behaves as a

function of K for different amounts of noise in the presented patterns. When the presented patterns are exactly the P prototypes to be learned (see (2.34)), increasing the number of synaptic states does not allow to increase the capacity beyond the 0.69bits/synapse value of the Willshaw model ($x = 0$, black line). The network reaches this capacity with $\alpha = 0.69$, that is on average a synapse is activated less than one time. The optimal depression-potentiation ratio being 0 as in the $K = 2$ case. For this noise-less case, the distributions of weights at the end of learning is extremely bimodal, with half of the connections at 0 and half at K. It is not clear why there is no parameters such that the intermediary synaptic states are used to lead higher capacity.
When the presented patterns are noisy versions of the prototypes (blue and magenta lines), having a synapse with multiple states allows to increase capacity, and eventually to reach the capacity obtained in the noiseless case. Note also that the optimal capacity is reached for $\delta \neq 0$, e.g. $\delta = 0.8$ for $K = 5$ and $\delta = 0.55$ for $K = 30$ at $x = 0.2$.
In figure 2.9B, we show the distribution of synaptic states after learning for $x = 0.2$. As shown by the panel with $K = 30$, the distribution is bimodal around the two extreme values 0 and K. Again the fraction of silent synapses is decreasing from 50% for $K = 2$ to 40% for $K = 5$ and 18% for $K = 30$.

Poly-synaptic contacts in the MP learning scenario

Again, the expression of $\vec{P}^{\pm}$ are given in the appendix and allows to compute the optimal capacity. Results are reported as a function of the number of synaptic contacts for different values of the noise x in figure 2.8F. The behavior is similar to K states synapses, the larger K the closer the capacity for noisy patterns is from the 0.69bits/synapse noiseless limit. Figure 2.9D shows the distribution of synaptic values, and as in the previous paragraph, this distribution is bi-modal, but this time, the peaks of probability are not on the extreme values 0 and K, but more at intermediate values. For instance, this makes the fraction of silent synapses to be 0 for $K = 30$.

2.3.4 Discussion on multiple states connections

We have computed the storage capacity of networks with two models of the connection between neurons and in the two learning scenarios SP and MP (see figure

2.8). In some cases having multiple states allow a significant increase in capacity, for instance from $i = 0.08$ bits/synapse for $K = 2$ synaptic states to $i = 0.16$ bits/synapse for $K = 6$ synaptic states, in the SP learning scenario where patterns are presented only once. We also found in the MP scenario where patterns are presented multiple times, that having multiple states was beneficial to reach a capacity of 0.69bits/synapse for noisy patterns. However, for non noisy patterns we have not found any increase in capacity with K, while Gardner's calculations performed on networks of synapses with multiple discrete positive states (Gutfreund and Stein, 1990) have shown that the optimal capacity is increasing with K, to reach the capacity for continuous synapses in the case of large K. Note that in this paper the calculations are done for the un-bias case (i.e. $f = \frac{1}{2}$), and that also our values of capacity still can not be directly compared to the Gardner bound as we are testing the stability of only one pattern while the Gardner calculation test the stability of all the patterns. We could only get fair comparison if we had the Gardner bounds for non-fluctuating numbers of selective neurons per patterns (for which we can test the stability of all patterns in our models) and sparse coding levels.

As in the case $K = 2$, we expect that finite size effects are going to be important and decrease capacity significantly. It would be interesting to compute these finite-size effects to quantify how capacity is reduced. It would also be interesting to see how the storage capacity depends on K at fixed coding levels, for finite-size networks and for the different models, as there is hope that multi-states synapses could be used to maintain good storage performance even for large coding levels. This question has been studied in the SP model for finite-size networks and gaussian distributions of the fields by Huang and Amit (2011). For instance for $f = 0.003$ and $N = 10,000$, they have found that the number of patterns that can be stored is almost the same for $K = 2$ and $K = 15$, but that for $f = 0.01$, the capacity is almost doubled between these two values of K, suggesting that the maximal information capacity can remain large on a larger range of coding level when K is larger.

The two models of connections we have studied are defined by a specific form of the transition matrix describing the evolution of synapses upon pattern presentation. For instance in the case of a single synaptic contact that can take various values, we have only allowed transitions between neighboring states (see figure 2.8A). It would be interesting to look for an optimal transition matrix (by changing independently each of the entry) in the SP and MP cases, to see whether further improvements can be achieved. This idea has been explored by Barrett and van Rossum (2008) for a perceptron performing a recognition task on patterns that are

presented only once during a learning phase (reminiscent of our SP model). They quantified performance using a different measure of information than the one we have used, and found that optimal learning rules were corresponding to transition matrices $\mathscr{M}$ that are band diagonal. In our study of the SP model, we have found for $K > 2$ that capacity is larger for multiple binary contacts independently updated at pattern presentation (non-band diagonal transition matrices) than for a single contact with multi-state with only transitions between neighboring states allowed (band-diagonal transition matrices). Although it is difficult to compare these two studies that use different models with different measures of information capacity, it would suggest, in order to increase capacity, to try to introduce heterogeneities in the transition probability between the different states (i.e. introduce heterogeneities in the a and b entries of the transition matrix (2.81)).

2.3.5 Methods

Computing the distribution of input fields

Here we assume that the $\vec{P}^{\pm}$ are known (see next paragraph for how to derive it). The goal is to compute the field distribution on neurons selective to pattern $\vec{\xi}^{\mu}$ (neuron i such that ξ_i^{μ}=1) and on neurons non-selective (neuron i such that ξ_i^{μ}=0). For a pattern with M selective neurons, the field $h_i^n = \sum_{j=1}^{N} W_{ij}\xi_j^{\mu}$ on a non-selective neurons is distributed according to (see Method for calculation)

$$\begin{aligned} P(h_i^n = S) &= \sum_{(S_1,...,S_{K-1})/\sum_{k=1}^{K-1} kS_k} (P_1^-)^{S_1}(P_2^-)^{S_2}...(P_{K-1}^-)^{S_{K-1}} \\ &\quad \times (1 - \sum_{k=1}^{K-1} P_k^-)^{M-\sum_{k=1}^{K-1} S_k} \binom{M}{S_1}\binom{M-S_1}{S_2}...\binom{M-\sum_{k=1}^{K-2} S_k}{S_{K-1}} \\ &\propto \sum_{(S_1,...,S_{K-1})/\sum_{k=1}^{K-1} kS_k} \exp(-M\Phi^K(\vec{P}^-, \frac{S_1}{M}, ..., \frac{S_{K-1}}{M}) + o(M)) \end{aligned} \tag{2.88}$$

Equation (2.88) is the generalization of the binomial distribution for K states (thus we treat W_{ij} as independent random variables, which we assume is legitimate in the sparse coding limit). Φ^K can be computed using Stirling formula

$$\Phi^K(\vec{P}^-, \frac{S_1}{M}, ..., \frac{S_{K-1}}{M}) = \sum_{k=1}^{K-1} \frac{S_k}{M} \ln \frac{S_k/M}{P_k^-} + (1 - \sum_{k=1}^{K-1} \frac{S_k}{M}) \ln \frac{1 - \sum_{k=1}^{K-1} \frac{S_k}{M}}{1 - \sum_{k=1}^{K-1} P_k^-} \tag{2.89}$$

In the limit of large size networks, the sum in equation 2.88 is dominated by its largest term

$$P(h_i^n = S) \propto \exp(-M\Phi^K(\vec{P}^-, \frac{S_1^*}{M}, ..., \frac{S_{K-1}^*}{M}) + o(M)) \tag{2.90}$$

where $(S_1^*, ..., S_{K-1}^*) = arg[min(\Phi^K(\vec{P}^-, \frac{S_1}{M}, ..., \frac{S_{K-1}}{M}))]$, which can be found using Lagrange multipliers method, we introduce:

$$L(\frac{S_1}{M}, ..., \frac{S_{K-1}}{M}) = \Phi^K(\vec{P}^-, \frac{S_1}{M}, ..., \frac{S_{K-1}}{M}) - \lambda(\sum_{k=1}^{K-1} k\frac{S_k}{M} - \frac{S}{M}) \tag{2.91}$$

The extremum of L is reached at a point $(\frac{S_1^*}{M}, ..., \frac{S_{K-1}^*}{M})$ such that $\frac{\partial L}{\partial S_k}(\frac{S_1^*}{M}, ..., \frac{S_{K-1}^*}{M}) = 0 \Leftrightarrow \frac{S_k^*}{M} = P_k^- \frac{1-X}{1-\sum_{k=1}^{K-1} P_k^-} e^{\lambda k}$, with $X = \frac{\sum_{k=1}^{K-1} P_k^- e^{\lambda k}}{1+\sum_{k=1}^{K-1} P_k^-(e^{\lambda k}-1)}$ and the Lagrange multiplier can be found by extracting roots of the following polynomial in e^λ :

$$\sum_{k=1}^{K-1} kP_k^- e^{\lambda k} - \frac{1 - \sum_{k=1}^{K-1} P_k^-}{1 - X} \frac{S}{M} = 0 \tag{2.92}$$

The root of the polynomial minimizing Φ^K is computed numerically. This allows us to obtain the distribution of h_i^n.
The distribution of the field on foreground neurons can be expressed similarly

$$\begin{aligned} P(h_i^s = S) &= \sum_{(S_1,...,S_{K-1})/\sum_{k=1}^{K-1} kS_k} (P_1^+)^{S_1}(P_2^+)^{S_2}...(P_{K-1}^+)^{S_{K-1}} \\ &\quad \times (1 - \sum_{k=1}^{K-1} P_k^+)^{M-\sum_{k=1}^{K-1} S_k} \binom{N}{S_1} \binom{M - S_1}{S_2} ... \binom{M - \sum_{k=1}^{K-2} S_k}{S_{K-1}} \\ &\propto \exp(-M\Phi^K(\vec{P}^+, \frac{S_1^*}{M}, ..., \frac{S_{K-1}^*}{M}) + o(M)) \end{aligned} \tag{2.93}$$

In the large N limit, the requirement for pattern stability is that $\Phi_\theta^K > 0$. This is satisfied as long as θ is chosen such that $\theta > \sum_{k=1}^{K-1} kP_k^+$, since $\Phi^K(\sum kP_k^+) = 0$.

Computing the distribution of synaptic states in the single presentation (SP) learning scenario

At each pattern presentation, synaptic state update is described by

$$\begin{bmatrix} P(W_{ij}(t+1)=0) \\ P(W_{ij}(t+1)=1) \\ \vdots \\ \\ P(W_{ij}(t+1)=K-1) \\ P(W_{ij}(t+1)=K) \end{bmatrix} = \mathscr{M} \times \begin{bmatrix} P(W_{ij}(t)=0) \\ P(W_{ij}(t)=1) \\ \vdots \\ \\ P(W_{ij}(t)=K-1) \\ P(W_{ij}(t)=K) \end{bmatrix}$$

It is possible to describe the state of a synapse after the presentation of μ patterns if the matrix M can be diagonalized:

$$\overrightarrow{P(W_{ij}(\mu)=\mu)} = \mathscr{P}(\mathscr{P}^{-1}\mathscr{M}\mathscr{P})^{\mu}\mathscr{P}^{-1}\overrightarrow{P(W_{ij}(\mu)=0)} \tag{2.94}$$

where $\mathscr{P} = (\vec{u}^0, ..., \vec{u}^{K-1})$ is a matrix to change basis such that M becomes a diagonal matrix $\mathscr{P}^{-1}M\mathscr{P}$. In other words, $\vec{u}^k$ is the eigenvector associated with λ_k, the kth eigenvalue of M. The expression of $\overrightarrow{P(W_{ij}(\mu)=0)}$ depends on whether we want to compute $\vec{P}^-$ or $\vec{P}^+$.

The transition matrix for the case of a single synaptic contact with K states in equation (2.81) can be diagonalized (Kouachi, 2006; da Fonseca, 2007). For $K = 2n$, the eigenvalues of M are,

- $\lambda_k = 1 - a - b + \sqrt{2ab(1+\cos\theta_k)} = 1 - f^2 q_+(1 + \delta - \sqrt{\delta}\sqrt{2(1+\cos\theta_k)})$ with $\theta_k = \frac{2k\pi}{K}$ for $k = 1, ..., n-1$
- $\lambda_k = 1 - a - b - \sqrt{2ab(1+\cos\theta_k)} = 1 - f^2 q_+(1 + \delta + \sqrt{\delta}\sqrt{2(1+\cos\theta_k)})$ with $\theta_k = \frac{2(k-n+1)\pi}{K}$ for $k = n, ..., K-2$
- $\lambda_{K-1} = 1 - a - b$
- $\lambda_K = 1$

The eigenvectors of M $\vec{u}^{(k)} = (u_1^{(k)}, ..., u_K^{(k)})$ are,

- for $k = 1, ..., n-1$

$$u_j^{(k)} = (-\sqrt{\delta})^{K-j} \begin{cases} (1 - \sqrt{\delta}\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j+1}{2}\theta_k) + \sin(\frac{K-j-1}{2}\theta_k) & \text{for j odd} \\ \sqrt{\delta}\sin((\frac{K-j}{2}+1)\theta_k) + (\sqrt{\delta} - \sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j}{2}\theta_k) & \text{for j eve} \end{cases} \tag{2.95}$$

- for $k = n, ..., 2n-2$

$$u_j^{(k)} = (-\sqrt{\delta})^{K-j} \begin{cases} (1+\sqrt{\delta}\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j+1}{2}\theta_k) + \sin(\frac{K-j-1}{2}\theta_k) & \text{for j odd} \\ \sqrt{\delta}\sin((\frac{K-j}{2}+1)\theta_k) + (\sqrt{\delta}+\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j}{2}\theta_k) & \text{for j eve} \end{cases} \quad (2.96)$$

- $u_{K-j+1}^{(K-1)} = (-1)^{[j/2]}\delta^{[(j-1)/2]}$
- $u_{K-j+1}^{(K)} = \delta^{j-1}$

For $K = 2n+1$ the eigenvalues of M are,

- $\lambda_k = 1-a-b+\sqrt{2ab(1+\cos\theta_k)} = 1-f^2q_+(1+\delta-\sqrt{\delta}\sqrt{2(1+\cos\theta_k)})$ with $\theta_k = \frac{2k\pi}{K}$ for $k = 1, ..., n$
- $\lambda_k = 1-a-b-\sqrt{2ab(1+\cos\theta_k)} = 1-f^2q_+(1+\delta+\sqrt{\delta}\sqrt{2(1+\cos\theta_k)})$ with $\theta_k = \frac{2(k-n)\pi}{K}$ for $k = n+1, ..., K-1$
- $\lambda_K = 1$

The eigenvectors of M $\vec{u}^{(k)} = (u_1^{(k)}, ..., u_K^{(k)})$ are,

- for $k = 1, ..., n$

$$u_j^{(k)} = (-\sqrt{\delta})^{K-j} \begin{cases} (1-\sqrt{\delta}\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j+1}{2}\theta_k) + \sin(\frac{K-j-1}{2}\theta_k) & \text{for j odd} \\ \sqrt{\delta}\sin((\frac{K-j}{2}+1)\theta_k) + (\sqrt{\delta}-\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j}{2}\theta_k) & \text{for j eve} \end{cases} \quad (2.97)$$

- for $k = n+1, ..., 2n-1$

$$u_j^{(k)} = (-\sqrt{\delta})^{K-j} \begin{cases} (1+\sqrt{\delta}\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j+1}{2}\theta_k) + \sin(\frac{K-j-1}{2}\theta_k) & \text{for j odd} \\ \sqrt{\delta}\sin((\frac{K-j}{2}+1)\theta_k) + (\sqrt{\delta}+\sqrt{2(1+\cos\theta_k)})\sin(\frac{K-j}{2}\theta_k) & \text{for j eve} \end{cases} \quad (2.98)$$

- $u_{K-j+1}^{(K)} = \delta^{j-1}$

For the case with poly-synaptic contacts, we diagonalize $\mathscr{P}$ with Maple.

Computing the distribution of synaptic states in the multi-stable synapses, multi-presentation (MP) learning scenario

To establish the expressions of $\vec{P}^{\pm}$, we need to diagonalize the transition matrix $\mathscr{M}_{ij}(t)$ averaged over the possible sequences of presentation of the P patterns to

store. The expression of $\langle \mathscr{M}_{ij}(t) \rangle$ is similar to the transition matrix $\mathscr{M}$ defined in equation (2.81), with a and b replaced by $a_{ij} = q_+ \frac{C_{ij}}{P}$ and $b_{ij} = q_- \frac{D_{ij}}{P}$ where C_{ij} (resp. D_{ij}) is the number of potentiating events, i.e. the number of patterns μ such that $\xi_i^\mu = \xi_j^\mu = 1$ (resp. depressing events, the number of patterns μ such that $\xi_i^\mu \neq \xi_j^\mu$). We recall that $\vec{P}^{\pm}$ are obtained from the expressions of $\left(\langle \vec{P}_{ij}(+\infty) \rangle \right)_k$ that are obtained by diagonalizing $\langle \mathscr{M}_{ij}(t) \rangle$. More precisely

$$\langle \vec{P}_{ij}(t+1) \rangle = \mathscr{P}_{ij} \mathscr{D}_{ij} \mathscr{P}_{ij}^{-1} \langle \vec{P}_{ij}(t) \rangle \tag{2.99}$$

with $\mathscr{D}_{ij}$ the diagonalized version of $\mathscr{M}_{ij}$ and $\mathscr{P}_{ij} = (\vec{u}^0, ..., \vec{u}^{K-1})$ is a matrix to switch $\mathscr{M}_{ij}$ to $\mathscr{D}_{ij}$ and $\mathscr{P}_{ij}^{-1} = \left(\left(\vec{v}^0 \right)^T ; ... ; \left(\vec{v}^{K-1} \right) \right)^T$ its inverse. As we are interested only at the synaptic state after the presentation of an infinite number of sequences, $\langle \vec{P}_{ij}(+\infty) \rangle$, we only need to know the vectors $\vec{u}^0$ and $\vec{v}^0$ associated with the eigenvalue $\lambda_0 = 1$

$$\langle \vec{P}_{ij}(+\infty) \rangle = \left[\vec{v}^0 . \langle \vec{P}_{ij}(0) \rangle \right] \vec{u}^0 \tag{2.100}$$

In the multi-stable synapses case, we found

$$\left(\vec{u}^0 \right)^T = \left(\delta_{ij}^K, \delta_{ij}^{K-1}, ..., \delta, 1 \right) \tag{2.101}$$

and

$$\left(\vec{v}^0 \right)^T = \left(\frac{1}{1 + \delta + ... + \delta^K}, ..., \frac{1}{1 + \delta + ... + \delta^K} \right) \tag{2.102}$$

with $\delta_{ij} = \frac{q_- D_{ij}}{q_+ C_{ij}}$. It allows to express (by noting that $|| \langle \vec{P}_{ij}(0) \rangle ||^2 = 1$)

$$\left(\langle \vec{P}_{ij}(+\infty) \rangle \right)_k = \frac{\delta_{ij}^{K-k}}{\sum_{n=0}^K \delta_{ij}^n} \tag{2.103}$$

Finally, with the relationships (2.86) and (2.87) we can write the expressions of $\vec{P}^{\pm}$

$$P_k^- = \sum_{\Pi=0}^{P} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{(\alpha\delta)^{K-k} \Pi^k}{\sum_{n=0}^K (\alpha\delta)^n \Pi^{K-n}} \tag{2.104}$$

and

$$P_k^+ = \sum_{\Pi=0}^{P-1} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{(\alpha\delta)^{K-k} (1+\Pi)^k}{\sum_{n=0}^K (\alpha\delta)^n (1+\Pi)^{K-n}} \tag{2.105}$$

This expressions are valid when the presented patterns are exactly the prototypes we are trying to store ($x = 0$). When noisy versions of the prototypes are presented ($x > 0$), the same reasoning can be done by replacing C_{ij} by (see Brunel et al. (1998))

$$\begin{aligned} \tilde{C_{ij}} &= \left[1 - x(1-f)\right]^2 C_{ij} + fx(1 - x(1-f))D_{ij} + (fx)^2 \left[P - C_{ij} - D_{ij}\right] \\ &\underset{f\to 0}{\simeq} (1-x)^2 C_{ij} + \alpha x(2-x) \end{aligned} \tag{2.106}$$

and replacing D_{ij} by

$$\begin{aligned} \tilde{D_{ij}} &= 2(1-f)x\left[1 - x(1-f)\right] C_{ij} + \left[1 - x + 2f(1-f)x^2\right] D_{ij} + \\ &\quad 2fx(1-fx)\left[P - C_{ij} - D_{ij}\right] \\ &\underset{f\to 0}{\simeq} \frac{2\alpha}{f} \end{aligned} \tag{2.107}$$

which leads to

$$P_k^- = \sum_{\Pi=0}^{P} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{(\alpha\delta)^{K-k} \left[(1-x)^2\Pi + \alpha x(2-x)\right]^k}{\sum_{n=0}^{K} (\alpha\delta)^n \left[(1-x)^2\Pi + \alpha x(2-x)\right]^{K-n}} \tag{2.108}$$

and

$$P_k^+ = \sum_{\Pi=0}^{P-1} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{(\alpha\delta)^{K-k} \left[(1-x)^2(1+\Pi) + \alpha x(2-x)\right]^k}{\sum_{n=0}^{K} (\alpha\delta)^n \left[(1-x)^2(1+\Pi) + \alpha x(2-x)\right]^{K-n}} \tag{2.109}$$

Computing the distribution of synaptic states in the poly-synaptic binary contacts, multiple presentation (MP) learning scenario

We have to repeat the reasoning presented above with a different matrix $\mathscr{M}_{ij}$. Now this matrix has the same form as (2.83) again with a and b replaced by $a_{ij} = q_+ \frac{C_{ij}}{P}$ and $b_{ij} = q_- \frac{D_{ij}}{P}$. Note that we also assume that $q_+ \ll 1$. After diagonalizing several such matrices with different K, we were able to infer analytical expressions for $\vec{u}_0$ and $\vec{v}_0$

$$\left(u^0\right)_k = \binom{K}{k} \delta_{ij}^{K-k} \tag{2.110}$$

and

$$\left(v^0\right)_k = \frac{1}{(1+\delta_{ij})^K} \tag{2.111}$$

which leads to

$$P_k^- = \sum_{\Pi=0}^{P} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{\binom{K}{k} (\alpha\delta)^{K-k} \Pi^k}{(\Pi + \alpha\delta)^K} \tag{2.112}$$

and

$$P_k^+ = \sum_{\Pi=0}^{P-1} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{\binom{K}{k} (\alpha\delta)^{K-k} (1+\Pi)^k}{(1 + \Pi + \alpha\delta)^K} \tag{2.113}$$

When noisy patterns are presented

$$P_k^- = \sum_{\Pi=0}^{P} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{\binom{K}{k} (\alpha\delta)^{K-k} \left[(1-x)^2\Pi + \alpha x(2-x)\right]^k}{\left[(1-x)^2\Pi + \alpha\left(\delta + x(2-x)\right)\right]^K} \tag{2.114}$$

$$P_k^+ = \sum_{\Pi=0}^{P-1} e^{-\alpha} \frac{\alpha^\Pi}{\Pi!} \frac{\binom{K}{k} (\alpha\delta)^{K-k} \left[(1-x)^2(1+\Pi) + \alpha x(2-x)\right]^k}{\left[(1-x)^2(1+\Pi) + \alpha\left(\delta + x(2-x)\right)\right]^K} \tag{2.115}$$

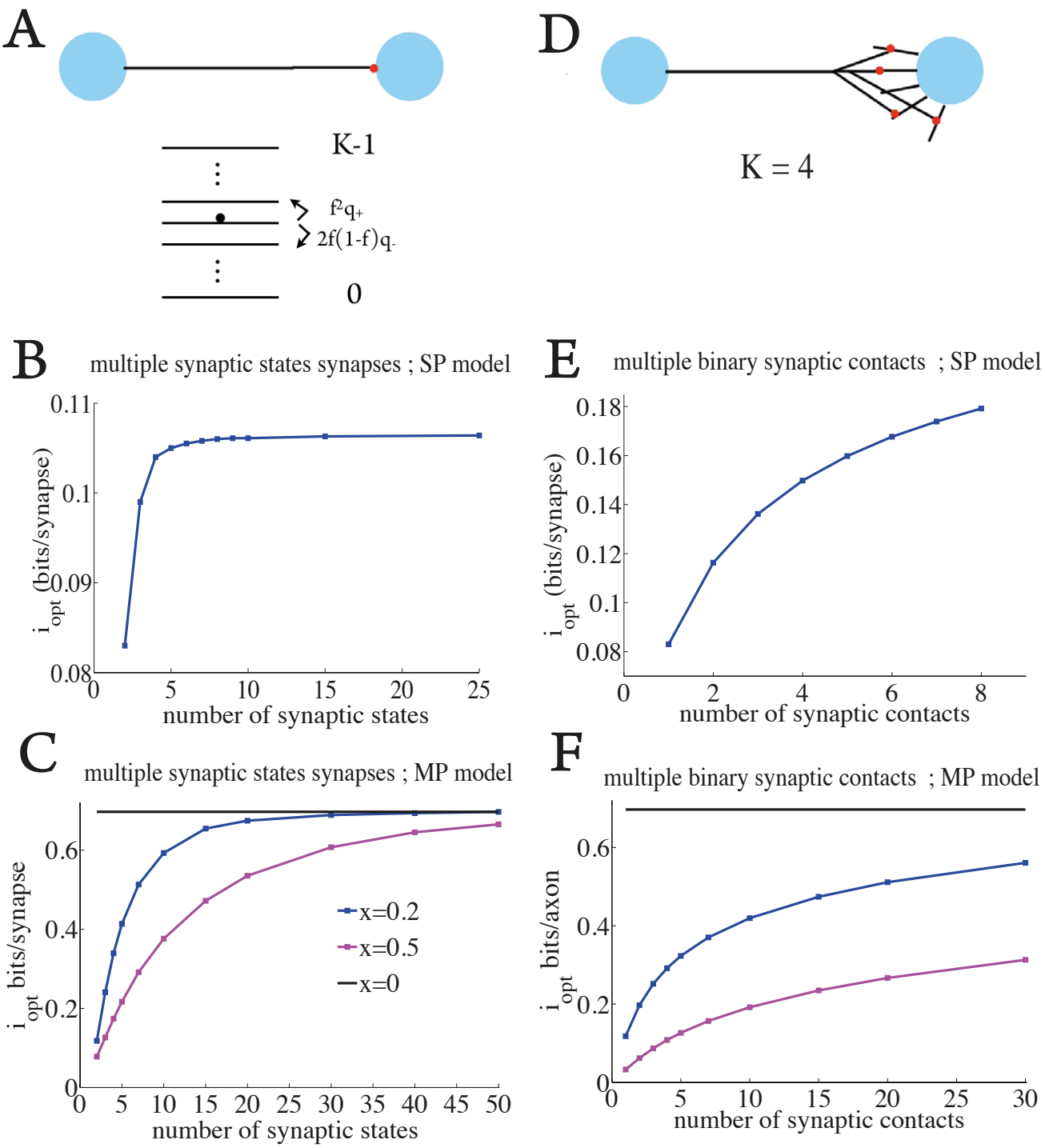

Figure 2.8: *Storage capacity in networks with K-states connections. A- Single synaptic contact that can take values* $0, ..., K-1$. *Upon pattern presentation only transitions between neighboring states are allowed. B- Storage capacity for the single contact with K states model in the SP scenario. We show the optimized capacity for different values of* K. *C- Storage capacity for the single contact with K states model in the MP scenario, for different levels of noise in the patterns presented during the learning phase (see* (2.34) *for a definition of the noise). D- As observed experimentally, axons have multiple contacts with a single dendritic tree (5 on average in cortex). A connection weight is thus modeled as a sum of K binary variables, each of them being updated independently at pattern presentation. E- Storage capacity for the poly-synaptic contacts model in the SP scenario. F- Storage capacity in the poly-synaptic contacts model in the MP scenario, for different levels of noise in the patterns presented during the learning phase.*

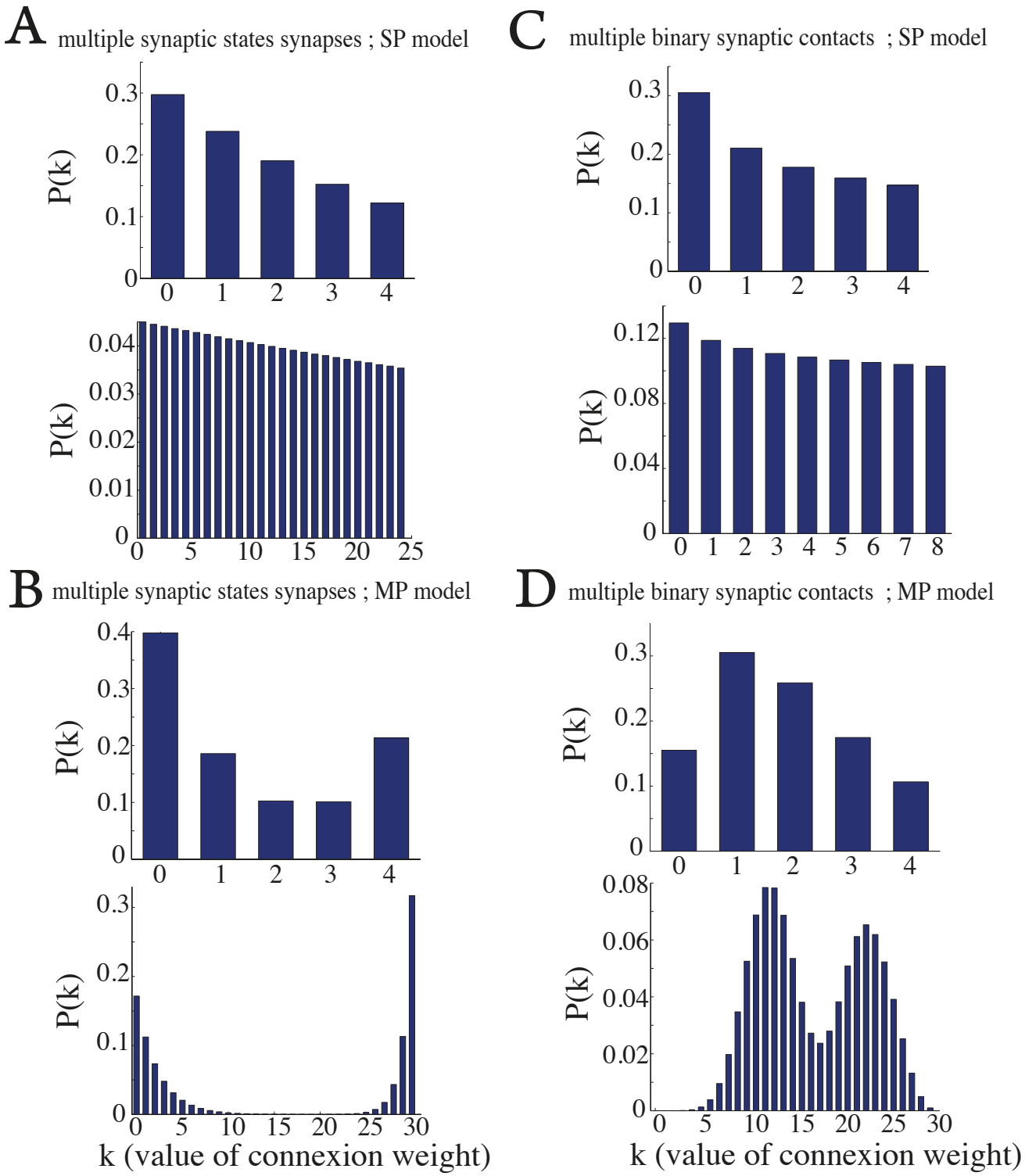

Figure 2.9: *Distribution of the strength of the connexion between two neurons at optimal capacity in the different models (multi-stable synapses and poly-synaptic contacts) and different learning scenarios, as given by the vector* $\vec{P}^{-}$. *A- Multi-stable synapses and single presentation (SP) learning scenario. B- Poly-synaptic binary contacts, SP scenario. C- Multi-stable synapses, MP model* ($x = 0.2$) *D- Poly-synaptic binary contacts, MP model* ($x = 0.2$)

Chapter 3

Modular networks

3.1 Introduction

In the previous chapter, the networks we have studied were fully-connected networks, i.e. every connection between two neurons can exist and be modified through activity dependent plasticity. Such models are well suited to describe local cortical networks with a size smaller than a few hundred microns, as supported by the study of Kalisman et al. (2005) which shows that potentially every pair of cells distant from $\leq 150\mu m$ can form a functional connection. When looking at cortical circuits at larger scales, connection probability is no more homogeneous with respect to the distance between neurons and drops with distance (Holmgren et al., 2003; Markov et al., 2011). Also, connectivity at a larger scale is not randomly distributed with a distance-dependent parameter, but rather shows some structure. For instance, as reported in the introduction, Pucak et al. (1996) have shown that connectivity from/to patches of pre-frontal cortex of monkeys of size $\simeq 0.3mm$ send/receive connections from other discrete patches of cortex that have the shape of stripes and sizes of $\simeq 0.25 \times 1.7mm$. One patch being connected to about $15-20$ other patches in the same or neighboring areas via grey matter connections and at least $15-20$ other patches connected via white matter connections. Other experimental studies have found such a patchy connectivity in sensory areas (DeFelipe et al., 1986; Gilbert and Wiesel, 1989; Bosking et al., 1997).

The possibility that local networks ($\leq 100\mu m$) sustain attractor dynamics has been supported by multiple experimental and theoretical studies. For instance, as mentioned in the introduction, the local connectivity between excitatory cells and from inhibitory cells to excitatory cells can be interpreted as the one of fully-connected networks optimizing storage capacity (Chapeton et al., 2012). Whether larger-scale cortical circuits (as the one described by Pucak et al. 1996 for instance) can sustain attractor dynamics remains far less studied (but see O'Kane and Treves 1992; Mari and Treves 1998; Kropff and Treves 2005; Johansson and Lansner 2007).

In this chapter, we study modular networks whose modules are fully connected networks ('short-range' connections) connected through 'long-range' diluted connections. In order to match the apparent large-scale structure of the cortex, we dilute connections between modules such that only a fraction of the pairs of modules can have connection between them. Again in order to match available experimental data, we impose that the numbers of 'short-range' and 'long-range' pre-synaptic connections onto a neuron are of the same order (Braitenberg and Schütz, 1991; Stepanyants et al., 2009).

We model these networks with binary neurons and binary synapses, for which the storage properties of fully-connected networks are well known. Moreover, for such models, the distribution of the synaptic currents can be expressed analytically, which allows to study the associative properties of the networks. The storage of patterns of activity is implemented using a Willshaw learning rule, that potentiates, at each pattern presentation, a fraction of the synapses that are allowed to exist by the network architecture. As we have seen in chapter 2, implementing this learning rule is equivalent to implementing an on-line learning rule in which patterns of activity are repeatedly imposed to the network. The patterns of activity, or memories, we are trying to store reflect the modular architecture, in the sense that each of the memory consists in the activation of only a subset of the modules. We study two different models, a first one where the identity of the modules that are active in each pattern is randomly chosen from one memory to another. And a second model where patterns are split in categories. Two patterns in the same category have the same modules activated. This is a more realistic situation since objects that are semantically close to each others are represented in very similar way on the cortical surface, as can be seen from fMRI recordings (with a spatial resolution of about $2mm$) of humans exposed to numerous visual stimuli (Huth et al., 2012).

The properties of the models are characterized in two steps. First we evaluate how the storage capacity, regardless of associative properties, behaves as a function of the different parameters defining the networks. Then we look at how the storage capacity of these networks is affected when we require the networks to have associative properties. In particular we will see that for having associative properties, the ratio of the number of long-range connections and the number of short-range connections plays a critical role.

In this work, we study modular networks for the two different kinds of patterns of activity described above. We first give a general description of these network models and explain how to compute their storage capacity. Then we explicit their

specificity for the two kinds of patterns.

3.2 Modular networks and storage capacity

We consider networks of M modules of N binary neurons connected through a binary connectivity matrix. The activity of each neuron (i, m) $(i = 1...N; m = 1...M)$ is described by a binary variable $\sigma_i^m = 0, 1$, evolving in time according to

$$\sigma_{i,m}(t+1) = \Theta\left[h_{i,m}(t) - \theta f N\right] \tag{3.1}$$

where

$$h_{i,m}(t) = h^l_{i,m}(t) + h^e_{i,m}(t) = \sum_{j=1}^{N} W_{ij}^{m,m}\sigma_{j,m} + \sum_{n\neq m}\sum_{j=1}^{N} W_{ij}^{m,n}\sigma_{j,n} \tag{3.2}$$

is the total synaptic input ('field') on neuron (i, m) that is composed of a local field resulting from the activity of neurons belonging to the same module and an external field resulting from the activity of neurons belonging to other modules. $\theta = O(1)$ is an activation threshold, Θ is the Heaviside function, and W describes the connectivity between neurons.

3.2.1 Connectivity: pre-existing architecture and activity dependent learning

The connection $W_{ij}^{m,n}$ from neuron (i, m) to neuron (j, n) is described by a binary variable $\in \{0, 1\}$. The connectivity between these two neurons is determined by two factors, an architectural constraint that can be thought of as accounting for whether or not the dendritic arbor of neuron (j, n) is within reach of the axon of neuron (i, m) ; and an activity dependent factor that accounts for the formation of a functional connection between the two neurons.

The architectural constraint is such that neurons belonging to the same module are fully connected (every neuron can potentially be connected to every other neurons in the network) and connections between neurons belonging to different modules are diluted. In the following intra-module connections will be qualified as 'short-range' and inter-modules connections as 'long-range'. Dilution is made at two scales. At a macroscopic scale, only some pairs of modules can have connections between their neurons (see dashed lines in figures 3.1A-B), which mimics the patchy connectivity observed in cortex (DeFelipe et al., 1986; Gilbert and Wiesel, 1989;

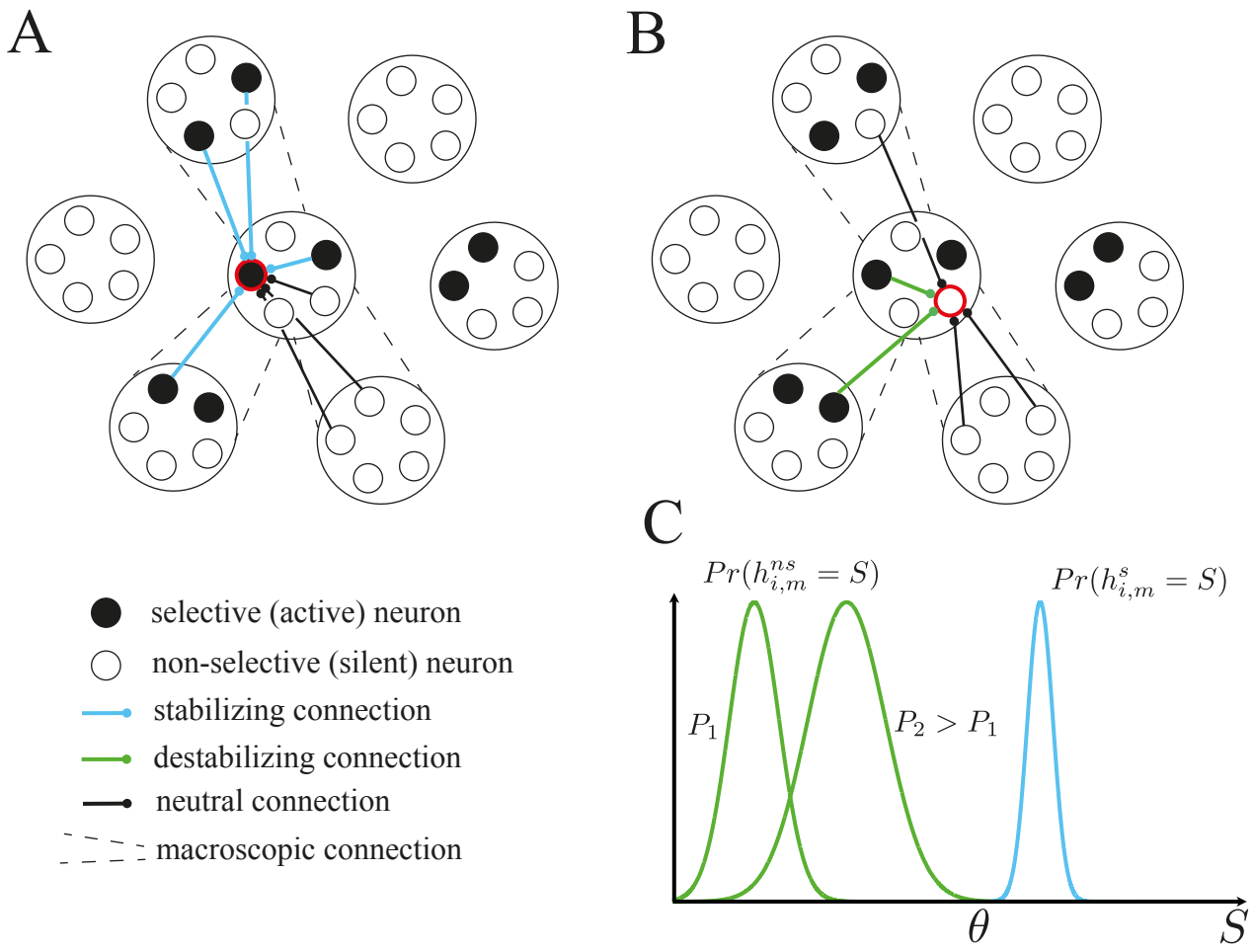

Figure 3.1: *Computing the number of patterns that can be stored in modular networks. After the learning phase (Willshaw learning rule), the network is set in one of the pattern $\vec{\Xi}^{\mu_0}$, which is said to be stored if it is a stable fixed point of the network dynamics. A- Focus on the connectivity onto a neuron selective for the memory μ_0 (red circle). Connections shown in blue create feed-back loops between selective neurons in which activity is reverberated, which stabilizes pattern $\vec{\Xi}^{\mu_0}$. Black connections from silent neurons do not influence the stability of this network state, they have been potentiated during the presentation of a pattern $\mu \neq \mu_0$. B- Focus on the connectivity onto a non-selective neuron (red circle). Connections onto this neuron result from the presentation of other patterns $\mu \neq \mu_0$ in which this neuron is active. The green connections from selective neurons provide excitation to this neuron, which tends to destabilize $\vec{\Xi}^{\mu_0}$. C- The typical stability of a pattern can be assessed by evaluating the probability distributions of the inputs on non-selective neurons (green) and on selective neurons (blue). A tested pattern is stable if the probability that the input on all non-selective neurons to be above the activation threshold θ, as well the probability that the input on all selective neurons to be below the activation threshold, are vanishingly small. When more patterns are stored in the synaptic matrix (from P_1 to $P_2 > P_1$), the number of green connections increases and the distribution of the input on non-selective neurons shifts its mean towards θ and gets wider. Evaluating storage capacity consists in computing the largest P for which a tested pattern is stable.*

Pucak et al., 1996; Bosking et al., 1997). At a microscopic scale, for a pair of connected modules only a fraction of the connections between neurons belonging to different modules can exist (see the specific models for details). In order to match available experimental data (Braitenberg and Schütz, 1991; Stepanyants et al., 2009) the amount of dilution is chosen such that each post-synaptic neuron receives

a number of long-range connections of the same order of short-range connections, thus we introduce

$$\gamma = \frac{\text{nb long-range connections}}{\text{nb short-range connections}} = O(1) \tag{3.3}$$

Besides the architectural constraint, the connectivity matrix is specified by P patterns of activity ('memories') that are stored in the network during a learning phase. For each pattern $\vec{\Xi}^{\mu}$ ($\mu = 1...P$), the activity of a neuron is also described at two scales by the product of two binary variables

$$\Xi^{\mu}_{i,m} = \Xi^{\mu}_{m} \xi^{\mu}_{i,m} \in \{0,1\} \times \{0,1\} \tag{3.4}$$

At the macroscopic scale, a fraction F (macroscopic coding level) of the modules are active

$$\sum_{m=1}^{M} \Xi^{\mu}_{m} = FM \tag{3.5}$$

At the microscopic scale, a fraction f (microscopic coding level) of the neurons are active in any active module m

$$\sum_{i=1}^{N} \xi^{\mu}_{i,m} = fN \tag{3.6}$$

We carry a study in a sparse coding limit where the microscopic coding level scales with the number of neurons in each module as

$$f = \beta \frac{\ln N}{N} \text{ with } \beta = O(1) \tag{3.7}$$

This allows the network to optimize its storage capacity (Willshaw et al., 1969). The scaling of the macroscopic coding level will be defined for each particular model. These patterns are stored using a Willshaw type learning rule. If the architecture of the network is such that there can be a connection between neurons (i,m) and (j,n), this becomes functional, $W_{ij}^{m,n} = 1$, if there exists a pattern in which these two neurons are co-activated.

3.2.2 Pattern stability and storage capacity

After the learning phase, we choose one of the P presented patterns $\vec{\Xi}^{\mu_0}$, set the network in this state $\vec{\sigma} = \vec{\Xi}^{\mu_0}$ and test whether it is a fixed point of the dynamic (3.1).

The stability of pattern $\vec{\Xi}^{\mu_0}$ is assessed by computing the probability $\mathbb{P}_{ne}$ that the fields on neurons are on the right side of the activation threshold $\theta f N$ (see figure 3.1C for an illustration). Doing so, we have to distinguish selective neurons (neurons (i,m) such that $\Xi^{\mu_0}_{i,m}=1$, figure 3.1A), neurons that are non-selective but that belong to an active module ((i,m) such that $\Xi^{\mu_0}_{i,m}=0$ but $\Xi^{\mu_0}_{m}=1$, figure 3.1B) and neurons that are non-selective and belong to an inactive module ((i,m) such that $\Xi^{\mu_0}_{i,m}=0$ and $\Xi^{\mu_0}_{m}=0$). The probability of $\vec{\Xi}^{\mu_0}$ being a fixed point of (3.1) can be written

$$\begin{aligned}\mathbb{P}_{ne} &= \big(1-\mathbb{P}(h_{i,m}\leq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=1)\big)^{FMfN} \\ &\times \big(1-\mathbb{P}(h_{i,m}\geq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=0,\, \Xi^{\mu_0}_{m}=1)\big)^{FM(1-f)N} \\ &\times \big(1-\mathbb{P}(h_{i,m}\geq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=0,\, \Xi^{\mu_0}_{m}=0)\big)^{(1-F)MN} \end{aligned} \tag{3.8}$$

In the limit of large networks and for a sparse microscopic coding level, these probabilities can be expressed for the two specific models we will consider (see Methods section), they take the form

$$\begin{aligned}\mathbb{P}(h_{i,m}\leq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=1) &= \exp\left[-fN\Phi^{s}+o(fN)\right] \\ \mathbb{P}(h_{i,m}\geq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=0,\, \Xi^{\mu_0}_{m}=1) &= \exp\left[-fN\Phi^{ns}+o(fN)\right] \\ \mathbb{P}(h_{i,m}\geq fN\theta \,|\, \Xi^{\mu_0}_{i,m}=0,\, \Xi^{\mu_0}_{m}=0) &= \exp\left[-fN\Phi^{ns'}+o(fN))\right] \end{aligned} \tag{3.9}$$

Where the Φ's are rate functions that depend on network parameters, patterns parameters and the number of stored patterns P. This allows to rewrite

$$\mathbb{P}_{ne}=\exp\left[-\exp(X_s)-\exp(X_{ns})-\exp(X_{ns'})\right] \tag{3.10}$$

with

$$\begin{aligned}X_s &= -\beta\Phi^{s}\ln N+o(\ln N) \\ X_{ns} &= -\beta\Phi^{ns}\ln N+\ln N+o(\ln N) \\ X_{ns'} &= -\beta\Phi^{ns'}\ln N+\ln N+o(\ln N) \end{aligned} \tag{3.11}$$

For $\mathbb{P}_{ne}$ to go to 1 in the large N limit, we need the X's to go to $-\infty$ in that limit. This will be satisfied provided

$$\Phi^{s}>0 \tag{3.12}$$

$$\Phi^{ns} > \frac{1}{\beta} \tag{3.13}$$

There is no inequality involving the rate function related to errors in modules that are silent because the probability to activate a neuron in these modules is much lower than the one to activate a neuron in active modules that receives local noise on top of external noise.
For a given set of parameters, one can thus find the maximal number of patterns P_{max} that can be imprinted in the synaptic matrix while keeping pattern $\vec{\Xi}^{\mu_0}$ a fixed point of the dynamics. This allows to compute the information capacity i of the network, defined as

$$i = \frac{P_{max} I_{pattern}}{\text{number of synapses}} \tag{3.14}$$

where $I_{pattern}$ is the information carried by each pattern. It is divided by the total number of synapses that can be potentiated, i.e. the physical substrate on which patterns are stored.

3.3 Maximal storage capacity

We now introduce the specific models for the two kinds of patterns and explicit the expressions for the information capacity as a function of the various parameters, which allows us to characterize their storage capacity.

3.3.1 Unstructured macroscopic patterns

In this model, each pattern of activity consists of FM active modules that are chosen randomly and **independently** from the other patterns. The architecture of the network on which these patterns are stored consists of the M fully connected modules that are connected to each other by connections randomly diluted at the macroscopic **and** microscopic level. The macroscopic dilution is such that connections from one module m to another module n can exist with probability $\frac{D}{M}$. The microscopic dilution is such that, if the macroscopic architecture allows connections from m to n, two neurons belonging to m and n can be connected with probability $\frac{d}{N}$. A given neuron in m is thus, on average, potentially connected to Dd external neurons. The constraint on the ratio of the numbers of long-range connections and short-range connections (3.3) is then written

$$Dd = \gamma N \tag{3.15}$$

In order to describe the noise due to the stored patterns on this architecture (see figure 3.1B), we introduce

$$g = 1 - (1 - f^2)^{PF} \underset{f \to 0}{\simeq} 1 - \exp(-\alpha F) \tag{3.16}$$

with

$$\alpha = Pf^2 \tag{3.17}$$

g is the probability to find a functional connection $W_{ij}^{m,m} = 1$ between two neurons belonging to the same module. We study the network in the limit where the number of stored patterns P scales as $\frac{1}{f^2}$, i.e. $\alpha = O(1)$. We also introduce

$$G = 1 - (1 - f^2)^{PF^2} \underset{f \to 0}{\simeq} 1 - \exp(-\alpha F^2) \tag{3.18}$$

the probability to find a functional connection between two neurons belonging to different modules, given the pre-existing architecture allows a connection to be formed.
As in the previous section, in order to estimate the storage capacity of such a network, we want to test the stability of a given pattern $\vec{\Xi}^{\mu_0}$. To do so we need to evaluate the distribution of inputs on neurons while the network is in a state $\vec{\sigma} = \vec{\Xi}^{\mu_0}$ (see figure 3.1C for an illustration). We focus on neurons that are selective for the pattern ((i,m) such that $\Xi_{i,m}^{\mu_0} = 1$), that should receive an input $h_{i,m}^s$ above the activation threshold for the pattern to be stable, and neurons that are non-selective but belong to an active module ((i,m) such that $\xi_{i,m}^{\mu_0} = 0$ but $\Xi_m^{\mu_0} = 1$) that should receive an input $h_{i,m}^n$ below the activation threshold. The other non-selective neurons that belong to inactive modules have a negligible chance to receive an input above threshold as they only receive external inputs. Before describing the distribution of these inputs, it is useful to write their average.

$$\begin{aligned} \langle h_{i,m}^s \rangle &= fN + FDfd = fN(1 + \gamma F) \\ \langle h_{i,m}^n \rangle &= fNg + FDfdG = fN(g + \gamma FG) \end{aligned} \tag{3.19}$$

The rate functions (3.9) describing the distribution of the inputs are given in the Methods section (equations (3.58),(3.59) for $FD \gg \ln N$ and equations (3.66),(3.67) if $FD = O(\ln N)$). They allow to derive the condition (3.12) for no error on selective neurons and the condition (3.13) for no error on non-selective

neurons. Condition (3.12) is met if the activation threshold is below the average of the input on selective neurons

$$\theta \underset{N\to+\infty}{\to} 1+\gamma F \tag{3.20}$$

We can choose θ to be the average of the field on selective neurons because in the large N limit the fluctuations (due to the random dilution) around the mean are small, and there is a relatively small total number of selective neurons ($FMfN$) compared to the number of non-selective neurons ($FM(1-f)N$). In condition (3.13), the Φ's in the right hand side of the inequality depend only on γ, F, and α through g and G. In order to find the optimal storage capacity of the network we also saturate this inequality, which can be done by increasing the number of stored patterns (controlled by the parameter α) at fixed γ, F and β.

The information per synapse can be written as a function of network parameters. The information per pattern is

$$\begin{aligned} I_{pattern} &= M\Big[-F\ln_2(F)-(1-F)\ln_2(1-F)+ \\ &\quad FN(-f\ln_2 f-(1-f)\ln_2(1-f))\Big] \\ &\underset{fN\to+\infty}{\simeq} MFN(-f\ln_2 f) \end{aligned} \tag{3.21}$$

The number of synapses is equal to $MN(N-1)+M(M-1)\frac{D}{M}N(N-1)\frac{d}{N}\simeq MN^2(1+\gamma)$. Assuming α corresponds to the maximal number of patterns that can be stored without errors (i.e. α saturates (3.13)), the maximal information i_{max} that can be stored for parameters γ, F, β is

$$i_{max}=\frac{1}{\ln 2}\frac{\alpha F}{\beta(1+\gamma)} \tag{3.22}$$

In figure 3.2A, we plot the information optimized over the choice of the microscopic coding level (parameter β) as a function of the amount of long-range connections quantified by γ in the case $FD \gg \ln N$. For low γ the inputs to neurons are predominantly local, and the information capacity equals 0.69 bits/synapse the value of the storage capacity of the classical fully-connected Willshaw model (when we require the stability of only the tested pattern). For large γ, inputs are mainly controlled by external fields, and the storage capacity equals 0.26 bits/synapse, the capacity of the Willshaw model with diluted connections. For the values of interest $\gamma \simeq 1$, the storage capacity interpolates between these two values. The

different curves show the optimal capacity for different values of F. For $\gamma \simeq 1$, larger capacities are reached for smaller F since using smaller F minimizes the relative strength of external inputs (see (3.19)), and thus pushes capacity towards the larger capacity of non-diluted networks.

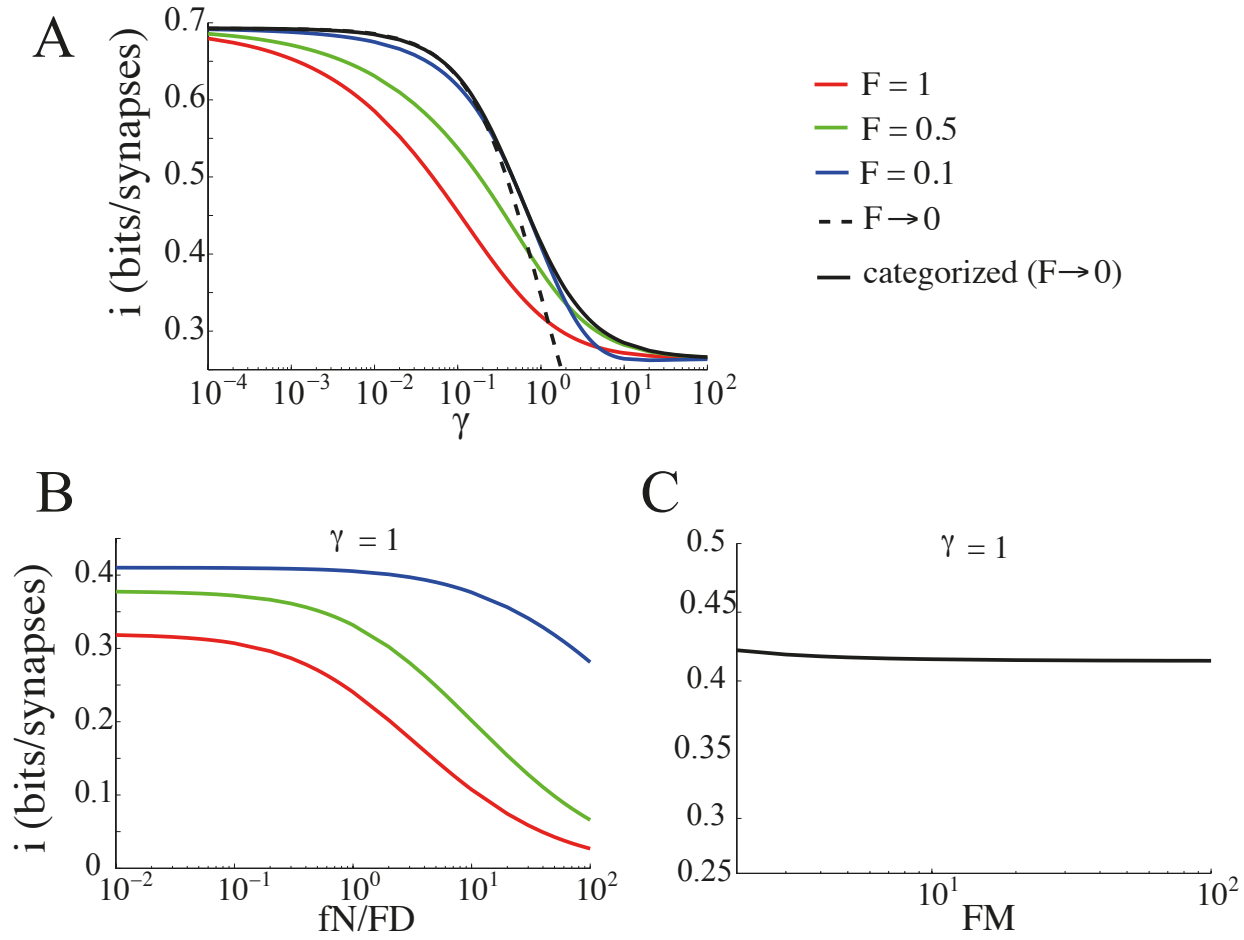

Figure 3.2: *Optimized storage capacity of the models. A- Storage capacity as a function of γ the ratio of long-range and short-range connections. Different curves correspond to different values of the macroscopic coding level. The full black curve is for the model in which patterns are organized in categories. The storage capacity interpolates between the storage capacity of the fully-connected Willshaw model at small γ and the storage capacity of a diluted Willshaw model for large γ. B- Storage capacity for the unstructured model when fluctuations (due to macroscopic dilution) in the number of modules sending inputs to a given module becomes large enough ($FD = O(fN)$) to decrease storage capacity. C- Storage capacity in the case of patterns organized in categories for a finite value of FM the number of active modules sending inputs to a given module.*

When the average number of modules sending inputs to a given module is small and of the order $FD = O(fN) = O(\ln N)$, the fluctuations of this number increases the width of the field distributions and capacity is reduced. Figure 3.2B shows how capacity is decreased for different values of F as a function of the ratio $\frac{fN}{FD}$. If one assumes $N \simeq 10,000$, and $FD \simeq 100$, the ratio is $\frac{fN}{FD} \simeq 0.1$ for which the capacity is only slightly impacted compare to the case $FD \gg fN$. It would be interesting to see in a finite network how these fluctuations due to macroscopic

dilution influence storage capacity. In the infinite size limit, if $FD = O(1)$ the capacity goes to zero.

3.3.2 Patterns organized in categories

Here each pattern of activity still consists of FM active modules chosen randomly, but the P patterns of activity are split into $\mathscr{P}$ categories. Patterns in the same category have the same active modules. Moreover, the patterns are chosen such that each module is activated in the same number of categories δ with

$$\delta = \mathscr{P}F \in \{1, ..., M\} \tag{3.23}$$

We study this model in the limits $\mathscr{P} \to +\infty$ and $F \to 0$. Here, there is no macroscopic dilution per se, but only pairs of modules that are co-activated in at least one category are connected. We introduce R, the probability that a given pair of modules is co-activated in at least one pattern

$$R = 1 - (1 - F^2)^{\mathscr{P}} \underset{F \to 0}{\simeq} \delta F \tag{3.24}$$

The microscopic dilution is described as before by the probability $\frac{d}{N}$ that two neurons belonging to a connected pair can be connected. The constraint (3.3) is now

$$RMd = \gamma N \tag{3.25}$$

Again we define

$$g = 1 - (1 - f^2)^{p\mathscr{P}F} \underset{f \to 0}{\simeq} 1 - \exp(-\alpha\delta) \tag{3.26}$$

the probability to find a functional connection $W_{ij}^{m,m} = 1$ between two neurons belonging to the same module. In equation (3.26) p is the number of patterns in each category, and

$$\alpha = \frac{p}{f^2} = O(1) \tag{3.27}$$

We also define

$$G = 1 - (1 - f^2)^{p(1+\mathscr{P}F^2)} \underset{f,F \to 0}{\simeq} 1 - \exp(-\alpha) \tag{3.28}$$

the probability to find a functional connection between two neurons belonging to different modules, given the pre-existing architecture allows a connection to be formed and given that the considered pair of modules is co-activated in at least one category.
In order to compute the storage capacity of this model, we proceed as above. First we establish the distributions of the inputs $h^s_{i,m}$ to selective neurons and $h^n_{i,m}$ to non-selective neurons, when the network state is set in one of the stored pattern. These fields now have averages

$$\begin{aligned} \langle h^s_{i,m} \rangle &= fN + FDfd = fN(1 + \frac{\gamma}{\delta}) \\ \langle h^n_{i,m} \rangle &= fNg + FDfdG = fN(g + \frac{\gamma}{\delta}G) \end{aligned} \tag{3.29}$$

and the rate functions describing their distributions is given in the Methods section by equations (3.69),(3.70) when $FM \gg 1$ and by equations (3.71),(3.72) when $FM = O(1)$. The maximal information per synapse for a given set of network parameters is given by

$$i_{max} = \frac{1}{\ln 2} \frac{\alpha\delta}{\beta(1+\gamma)} \tag{3.30}$$

where α is chosen to saturate (3.13). Note that in the expression of the field, $\frac{1}{\delta}$ plays the same role as F, but that in the expression of the information capacity δ plays the same role as F, thus the two models are not mathematically equivalent. On figure 3.2A we plot the storage capacity optimized over the choices of δ and β (black line) in the case where the number of active modules sending input to a given module is large $FM \gg 1$. As in the previous model, capacity interpolates between the two limits of the fully connected network and the diluted network. For $FM = O(1)$, the expressions of the rate functions change slightly and the value of the storage capacity is slightly modified as can be seen on figure 3.2C. Taking this limit means that the macroscopic coding level F can be taken to scale as $\frac{1}{M}$, thus the number of categories $\mathscr{P}$ and then the total number of stored patterns can scale with the number of modules in the network, as observed in the model studied by Mari and Treves (1998).

3.4 Storage capacity with finite size basin of attractions

In the previous section, the activation threshold is chosen very close to the average field on selective neurons (e.g. $\theta \to 1 + \gamma F$). With such a high threshold, the net-

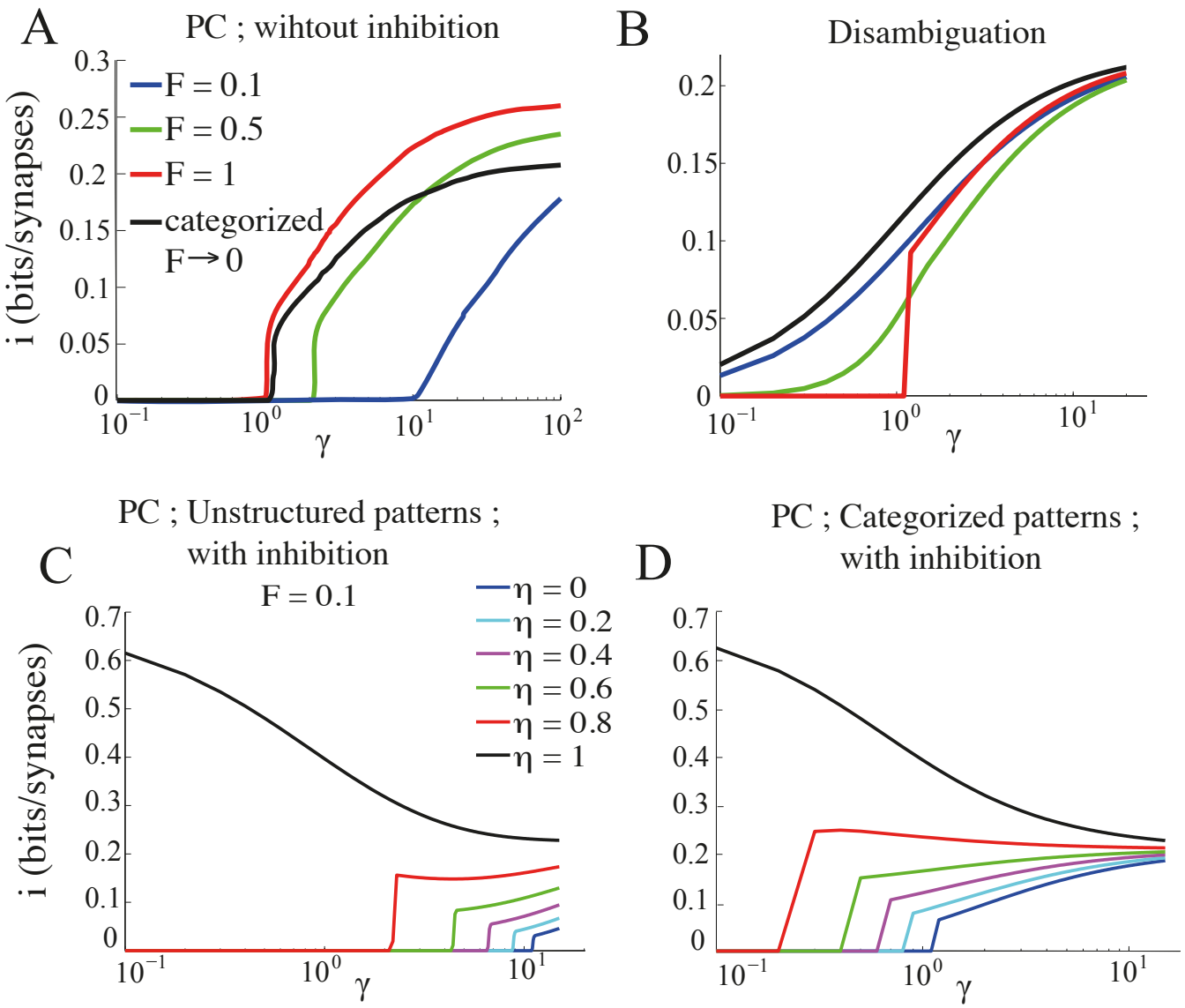

Figure 3.3: *Storage capacity with associative properties ($E = 0.1$). A- Storage capacity as a function of γ in the unstructured and categorized models when the network is required to perform macroscopic pattern completion. B- Same as A when the network is required to be able to disambiguate local patterns of activity. C- Storage capacity for $F = 0.1$ in the unstructured case with inhibition when the network is required to perform macroscopic pattern completion with the help of local inhibition. D- Same as C with categorized patterns and $F \to 0$.*

work does not have associative memory properties. We would like that a network initialized in a state $\vec{S}(t = 0)$ that resembles a pattern $\vec{\Xi}^{\mu_0}$, but with errors on the activation of a fraction of the modules, eventually reaches the state $\vec{\Xi}^{\mu_0}$. This can be done by using a lower activation threshold. As a consequence of the lower threshold, the network will store less memories. We study two kind of errors, a first one where the macroscopic pattern is different from the macroscopic pattern of $\vec{\Xi}^{\mu_0}$. When the network is able to recall the memory from such a pattern of activity we say the network performs 'pattern completion' (PC). Another kind of errors is

when the macroscopic pattern of activation is correct, but some of the modules are not in the correct attractor state. We say the network performs 'disambiguation' when it can recall the memory from such a pattern. We first show how the storage capacity drops when we require the network to perform pattern completion or disambiguation, then we show that adding local inhibition in the network helps for pattern completion. Finally we exhibit 'optimal' network parameters for which the networks have satisfactory storage capacity in both conditions.

3.4.1 Macroscopic pattern completion

For the pattern completion case, we set the network in an initial state such that

$$S_{i,m} = S_m s_{i,m} \text{ with } S_m = (1 - X_m^s)\Xi_m^{\mu_0} + X_m^{ns}(1 - \Xi_m^{\mu_0}) \tag{3.31}$$

where

$$X_m^s = \begin{cases} 1 & \text{with probability } (1-F)E \\ 0 & \text{with probability } 1-(1-F)E \end{cases} \tag{3.32}$$

$$X_m^{ns} = \begin{cases} 1 & \text{with probability } FE \\ 0 & \text{with probability } 1-FE \end{cases} \tag{3.33}$$

The microscopic activity $s_{i,m}$ in each correctly active module is $\xi_{i,m}^{\mu_0}$, and the activity in modules that are active by error is chosen as one of the local attractor of the module. In order to correct network activity in a single time step of the dynamics (3.1), we need an activation threshold that is above the input received by neurons to silence, neurons that are active in a module such that $\Xi_m^{\mu_0} = 0$, and below the input received by a neuron to activate ($\Xi_{i,m}^{\mu_0} = 1$ and $S_m = 0$). For the case of unstructured macroscopic patterns, the activation threshold is then required to satisfy

$$1 + \gamma FG < \theta < \gamma F(1 - E(1-F)(1-G)) \tag{3.34}$$

For the case of categorized patterns we need

$$1 < \theta < \frac{\gamma}{\delta}(1-E) \tag{3.35}$$

For the categorized case, there is no dependence on G since two randomly chosen modules have a probability to be co-activated in a pattern that is vanishingly small (see (3.24)). We study how these constraints on the activation threshold

reduces storage capacity, by optimizing capacity with a threshold satisfying the above constraints. The curves on figure 3.3A represent this optimized information capacity for the model with unstructured patterns and different values of F, for an amount of error $E = 0.1$. As can be noted from inequality (3.36), having a network performing error correction necessitates the relative amount of long-range connections quantified by γ to be large enough. The smaller the macroscopic coding level, the higher the amount of long-range connections has to be. This can be understood by examining (3.36) in the case where no patterns have been stored ($g = G = 0$). γ has to be high enough such that local inputs impinging on neurons to de-activate are smaller than external inputs impinging on neurons to activate. The optimized capacity for $E = 0.1$ and for the model with categorized patterns is shown by the black line in figure 3.3A. Again, γ has to be above 1 to allow for error correction, but note that here the amount of long range connections required does not depend on the macroscopic coding level. Even if $F \to 0$ a reasonable amount of long-range connections is sufficient to provide the network with associative properties.

3.4.2 Disambiguation

Now the network is initialized in a state $\vec{S}$ such that $S_m = \Xi_m^{\mu_0}$, but a fraction E of the active modules are not in the correct attractor: $S_{i,m} = \Xi_{i,m}^{\mu \neq \mu_0}$. To correct these errors, the activation threshold needs to be sufficiently high such that neurons whose activation is supported only by local connectivity are silenced, and sufficiently low such that long-range connections can activate neurons receiving only a noisy local input. Which gives for the model with unstructured macroscopic patterns

$$1 + \gamma FG < \theta < g + \gamma F(1 - E(1 - F)(1 - G)) \tag{3.36}$$

For the case of categorized patterns

$$1 + \frac{\gamma}{\delta} G < \theta < g + \frac{\gamma}{\delta}(1 - E(1 - G)) \tag{3.37}$$

Note that at low loading ($g = G = 0$) these inequalities are the same than for the case with unstructured patterns, the difference is in the way the storage of other patterns modify the right hand side and left hand side of the inequalities. We plot the capacity as a function of γ for $E = 0.1$ in figure 3.3B. Now, even for $\gamma < 1$ error correction is possible. However note that when $\gamma < 1$, the range of α

on which disambiguation is possible does not start at $\alpha = 0$. Such phenomenon is more clearly illustrated by the red bars in figures 3.4 and 3.5.

3.4.3 Effect of local inhibition on error correction

The need for having more numerous long-range connections than short-range connections in order to perform pattern completion, comes from the presence of the local term on the left hand side in inequalities (3.36),(3.37). Because of this term the activation threshold has to be high enough to prevent the activation of neurons that are only excited by local inputs. To relax this constraint on the activation threshold, the effective strength of local inputs has to be diminished via another mechanism. This can be done by introducing a local inhibitory term η proportional to the average activity in the local network. With this additional mechanism, the dynamics of the networks become

$$\sigma_{i,m}(t+1) = \Theta\left[\sum_{j=1}^{N} W_{ij}^{m,m}\sigma_{j,m} - \eta\sum_{j=1}^{N}\sigma_{j,m} + \sum_{n\neq m}\sum_{j=1}^{N} W_{ij}^{m,n}\sigma_{j,n} - \theta f N\right] \tag{3.38}$$

In this case, if no error correction is required it is optimal (in terms of storage capacity) to choose an activation threshold $\theta = 1 - \eta + \gamma F$ (resp. $\theta = 1 - \eta + \frac{\gamma}{\delta}$) for the unstructured model (resp. categorized model). The unconstrained storage capacity does not depend on the value of η. We first study the effect of inhibition in the pattern completion task. Here the constraints on the activation threshold become

$$1 - \eta + \gamma FG < \theta < \gamma F(1 - E(1-F)(1-G)) \tag{3.39}$$

for unstructured macroscopic patterns, and

$$1 - \eta < \theta < \frac{\gamma}{\delta}(1 - E) \tag{3.40}$$

for categorized patterns. In order to minimize the amount of long-range connections required to correct $\vec{S}$ into $\vec{\Xi}^{\mu_0}$, one should take $\eta = 1$. Once the value of the local inhibition is chosen, one can proceed as previously in order to estimate the maximal storage capacity that can be reached for a given amount of error correction (measured by E), i.e. optimize the choice of network parameters with an activation threshold satisfying the above constraints. In figure 3.3C, we plot the storage capacity as a function of γ for different values of the inhibition in the unstructured model and for $F = 0.1$. Figure 3.3D shows the same quantity in the

categorized model. As expected, the minimal γ required for having error correction decreases with η.
In the case of disambiguation, the constraints on θ in the presence of inhibition become

$$1 - \eta + \gamma FG < \theta < g - \eta + \gamma F(1 - E(1 - F)(1 - G)) \tag{3.41}$$

for unstructured macroscopic patterns, and

$$1 - \eta + \frac{\gamma}{\delta}G < \theta < g - \eta + \frac{\gamma}{\delta}(1 - E(1 - G)) \tag{3.42}$$

for patterns organized in categories. In this case, as can be noted from η appearing in both sides of the inequalities, adding inhibition is useless because the neurons we are trying to activate or silence are in the same module.

3.4.4 Pattern completion and disambiguation in the same network

So far we have optimized separately the capacity under the constraint of pattern completion or disambiguation. The optimal capacities exhibited in the different panels of figure 3.3 are reached for networks with different parameters (β and δ). We searched for parameters for which the network shows both the pattern completion and the disambiguation properties. In figure 3.4 each panel shows the performance of a network with unstructured macroscopic patterns in the three conditions. The brown, blue and red bars are respectively the capacity the network can reach if no associative property is required (0BA), pattern completion is required (PC), disambiguation is required (Dis.) for $E = 0.1$ and a fixed inhibition $\eta = 0.8$. The network performs better for large F and large γ, for which the external input is large and embeds the network with associative property at the macroscopic scale.

Similarly figure 3.5 shows the performance of networks in the case of categorized patterns for an error rate $E = 0.1$ and with fixed inhibition $\eta = 0.8$. Again the larger the external input (here measured by $\frac{\gamma}{\delta}$ instead of γF) the better are the macroscopic associative properties. One can note that at equal strength of external inputs ($\gamma F = \frac{\gamma}{\delta}$), the unstructured model has a larger capacity than the categorized model, because the information stored in the network is proportional to F in one case and inversely proportional to $\frac{1}{\delta}$ in the other case.

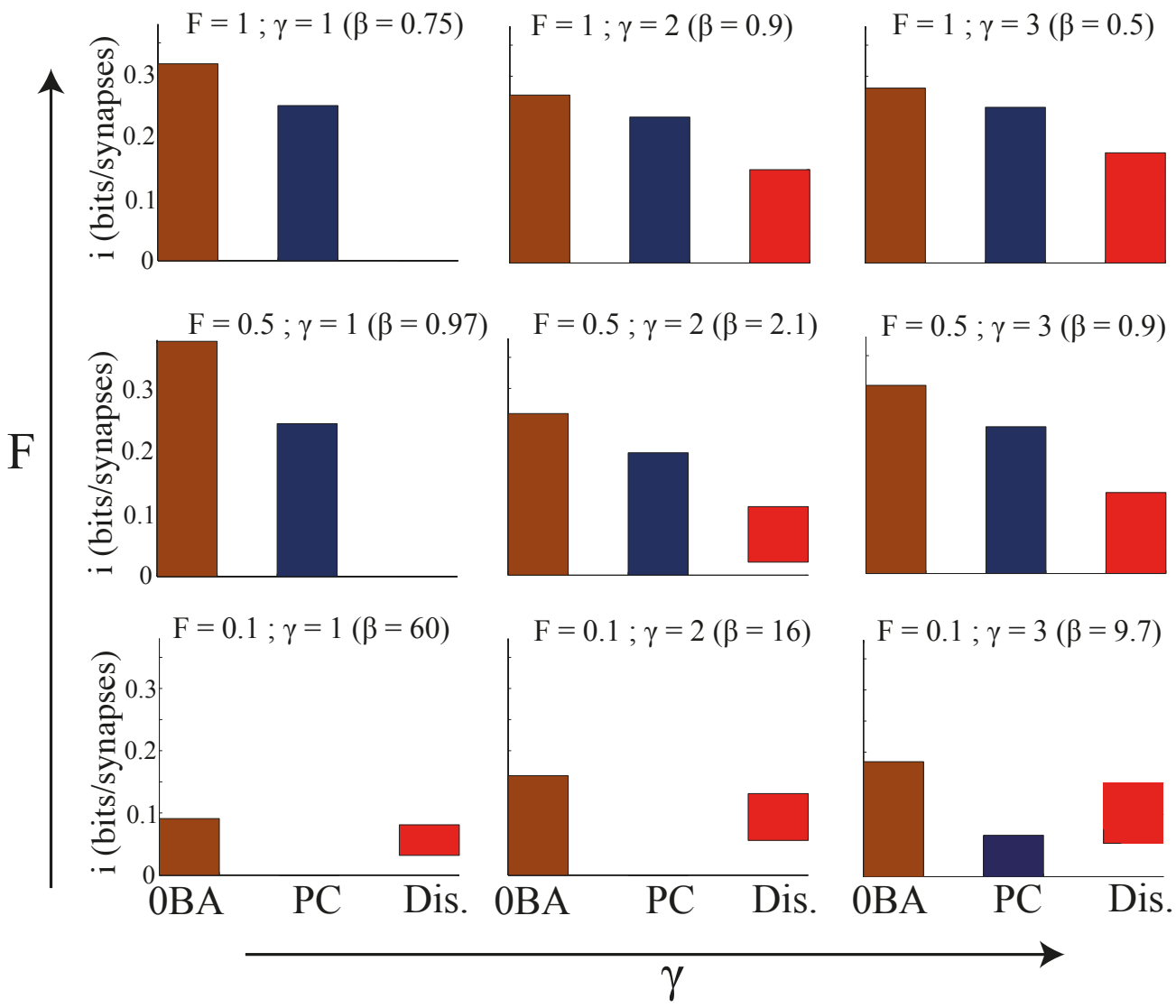

Figure 3.4: *Storage capacity of networks performing error-corrections in the case of unstructured patterns. From left to right, networks with increasing γ the ratio of the number of long-range connections to the number of short range connections. From bottom to up, increasing macroscopic coding level F. Different bars show the storage capacity that can be reached with an activation threshold θ that allows to perform disambiguation (red bars), pattern completion (blue bars) or that optimizes storage capacity without any error-corrections requirement (brown bars). The parameter $\beta = f\frac{\ln N}{N}$ has been chosen such that storage capacity are as high as possible in the different conditions, and would be chosen differently if it were optimized for each condition separately.*

3.5 Discussion

In this chapter, we have investigated modular ANNs where connectivity is potentially full inside modules and diluted between modules. This has been done using binary neurons and binary synapses on which patterns of activity are stored using a simple Hebbian learning rule as proposed by Willshaw et al. (1969). We have found that the storage capacities that can be reached on such a connectivity interpolate between the storage capacity of a fully connected Willshaw network

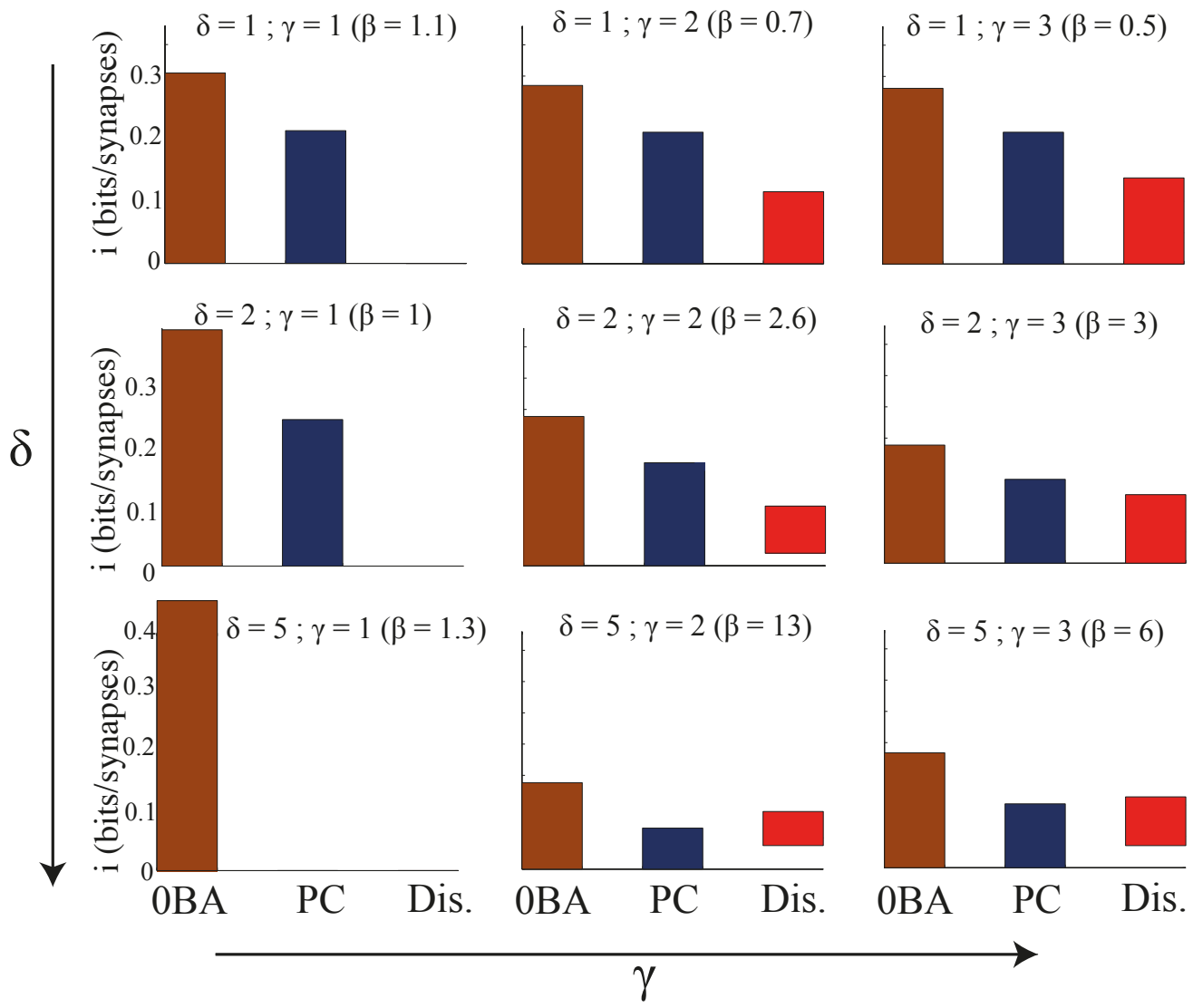

Figure 3.5: *Storage capacity of networks performing error-corrections in the case of patterns organized in categories. It reads as the previous figure, but from up to bottom, where δ, the number of categories in which a module is activated, is increasing.*

($i = 0.69$bits/synapse) and of a diluted Willshaw network ($i = 0.26$bits/synapse). These limits being respectively reached when the number of long-range connections is small or large compare to the number of local connections, i.e. $\gamma \ll 1$ or $\gamma \gg 1$. It seems therefore that the network which is best suited for memory storage is a network with disconnected modules ($\gamma = 0$). However, for disconnected modules, it is clear that the network does not have large-scale associative properties. That is, if the network is cued by a pattern of activity that resembles one of the memory patterns, but with some modules in a state too far from the memory pattern, the network's state does not flow towards the memory pattern.

We quantified how many long-range projections are needed to allow for such error-

corrections. This has been done for the models with unstructured macroscopic or categorized patterns, and two kinds of errors have been considered. A first kind ('pattern completion' task) for which the network is initialized with an erroneous macroscopic pattern, i.e. the pattern is one of the stored patterns but some modules that should be active are silent and some modules that should be silent are put in random attractors. We named the other kind of error-correction a 'disambiguation' task. In this case, the modules are correctly active or silent but some of the active modules are in a wrong attractor state. These errors can be corrected if the activation threshold is well adjusted such that wrongly active neurons are silenced and wrongly silent neurons are activated. This is possible if the synaptic drive carried by long-range range connections is high enough. More specifically, for the model with unstructured patterns, we have found that for the pattern completion task to be completed, the amount of long-range connections should satisfy $\gamma > \frac{1}{F}$; for the model with categorized patterns the constraint $\gamma > 1$ needs to be satisfied (see figure 3.3A). Note that in the case of categorized patterns, capacity is optimized for $\delta = 1$, i.e. each module is activated in only one category. If we want each module to be active in δ categories, the constraint becomes $\gamma > \delta$ (see equation (3.37)). Finally these constraints on γ simply translate the fact that in order to have pattern completion, the synaptic drive that allows a pattern to be stable should be dominated by its long-range component. This constraint on γ can be relaxed if a local inhibition, proportional to the activity in each module, is introduced (see figure 3.3C-D). This is because it becomes easier to activate silent modules in this case. For the disambiguation task, it seems there is no such hard constraint on the number of long-range connections, as there is always some parameters for which this task can be completed. However, for low values of γ ($\gamma < \frac{1}{F}$ in the unstructured macroscopic pattern case, and $\gamma < 1$ in the categorized case), disambiguation can be performed only if the network is loaded with enough memories (α high enough). Note that for the disambiguation task, adding inhibition does not relax the constraints on γ.

If the network are required to be able to perform both the pattern completion and the disambiguation tasks, storage capacity increases when the product γF increases (for unstructured patterns) with $\gamma > 1$ (see figure 3.4), or when the product $\frac{\gamma}{\delta}$ increases with $\gamma > 1$ (for categorized patterns, see figure 3.5). Again this is because the drives from long-range connections are proportional to γF or $\frac{\gamma}{\delta}$, and that this drive is crucial to perform macroscopic error correction. From this study, it appears that the networks that have the best memory performance are the one with $F = 1$ or $\delta = 1$ as they allow good storage capacity with a minimal number of long-range connections, which are supposed to be costly. In a sense these two

models are equivalent: the model with categorized patterns and in which modules are active in only one category is similar to a collection of multiple unconnected networks with $F = 1$. Thus it seems that having modules specialized for a single category of item is optimal in terms of memory storage.

In this study the number of neurons in each module as well as the number of modules go to infinity. How different the results would be if these numbers were to be finite ? As we have seen in the previous chapter, the storage capacity should be significantly reduced, one can bet on a factor around 5 given the previous results. In this previous chapter, by comparing the storage of patterns that all have the same number of active neurons, or a fluctuating numbers, we have realized that fluctuations in the input to selective neurons is detrimental to storage capacity. In the models studied here, we have removed this fluctuations by considering patterns of activity with a fixed number of active modules and fixed numbers of active neurons in each active module. However, in the model with unstructured macroscopic patterns another source of fluctuations is present, due to the random macroscopic dilution. Thus, in finite networks, we expect the storage capacity of this last model to be decreased more than the one of the model with categorized patterns.
We expect the constraints on γ we have established to be unaffected when going to finite size networks, as these constraints come from the need to have a sufficiently strong amount of non-local inputs to perform macroscopic error-corrections, which do not depend on network size.

ANNs with modules have been studied by a few other authors (O'Kane and Treves, 1992; Mari and Treves, 1998; Mari, 2004; Kropff and Treves, 2005; Johansson and Lansner, 2007). One comparable to ours is the one studied by Mari and Treves (1998). In their model, also a fraction F of the modules are activated in each pattern, and they show using a signal-to-noise analysis that the storage capacity of the network scales as $\frac{1}{F}$. By taking $F \propto \frac{1}{M}$, the number of patterns stored in the network can be proportional to the size of the whole network. In our models, we have found that only in the case of categorized patterns we could take the coding level to scale as low as $\frac{1}{M}$ without having a vanishingly small storage capacity. They got this feature for the two models they study, one with unstructured macroscopic patterns and one reminiscent of the categorized patterns we have studied, where macroscopic activity is not totally random but biased by the presence of connections between modules. One interesting feature of their second model, is that it allows to get rid of 'memory glass states', states that correspond to stable patterns where some modules are in local attractors that are not the one of the

memory. In our models, we have not encountered such states. One possibility for such difference is that we have used binary neurons while they used analog neurons. For binary neurons, in the large N limit, we take an activation threshold such that a neuron can stay active if it receives an input with both a strong local and external component (e.g. $\theta \to 1 + \gamma F$). With such a high threshold, a module that is in an attractor not supported by the macroscopic pattern will receive an input below threshold and be destabilized.

3.6 Methods

Estimating the storage capacity is done following the reasoning presented in section 3.2.2. This required to compute the distributions of the inputs on selective and non selective neurons (3.2). Because neural activity and synapses are binary, these inputs can be treated as sums of binary random variables. We first present some general results about such sums, results that are useful to explicit the distributions of inputs in the networks.

We consider a random variable h

$$h = \sum_{k=1}^{K} X_k \tag{3.43}$$

where the X_k are independent binary random variables described by a parameter q:

$$X_k = \begin{cases} 1 & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases} \tag{3.44}$$

The sum h is then distributed according to a binomial distribution

$$P(h = S) = \binom{K}{S} q^S (1-q)^{K-S} \tag{3.45}$$

Note that to get this binomial distribution, we have to assume the X_k's are independent. In our case, this mean that two synapses on the same neuron $W_{ij_1}^{m,n}$ and $W_{ij_2}^{m,n}$ are treated as independent variables. This is a reasonable assumption to make as the covariance between such random variables has been shown by Amit and Fusi (1994) to be negligible in the sparse coding limit $f \propto \frac{\ln N}{N}$ we are interested in. Moreover, in the second chapter, we have checked with numerical simulations that modeling the inputs with a binomial distribution is a very good approximation for such coding levels.

We will consider cases in which K and S are large, we can then use Stirling formula to express the binomial coefficients and write

$$P(h=S) = \exp\left(-K\Phi^{fc}(\frac{S}{K},q) + o(K)\right) \tag{3.46}$$

with

$$\Phi^{fc}\left(\frac{S}{K},q\right) = \frac{S}{K}\ln\left(\frac{S/K}{q}\right) + \left(1-\frac{S}{K}\right)\ln\left(\frac{1-S/K}{1-q}\right) \tag{3.47}$$

We use the superscript fc as this expression will be mainly used to describe fully connected sub-networks. For diluted enough networks, we will have $q, \frac{S}{K} \ll 1$, it is then useful to introduce Φ^{dc}

$$\Phi^{dc}\left(\frac{S}{K},q\right) = \frac{S}{K}\ln\left(\frac{S/K}{q}\right) - \frac{S}{K} + q \tag{3.48}$$

In our networks, when testing the stability of a given pattern $\vec{\Xi}^{\mu_0}$ it is useful to look at the total input as split into a local part and an external part. The local part is described by a couple (K,q), where $K = fN$ is the number of neurons active in a given local network, and $q = 1$ or g depending on whether we are considering the input onto a selective or a non-selective neuron. In most cases the external part can also be described by a couple (K,q) with $K = YfN$ (where Y is the number of modules that send input to the considered module) and $q = \frac{d}{N}$ or $\frac{d}{N}G$ for the selective or non selective neurons.

The distribution of the total input on a neuron can be written

$$\mathbb{P}(h_{i,m} = S) = \sum_{S_l,S_e/S_l+S_e=S} \mathbb{P}_l(S_l)\mathbb{P}_e(S_e) \tag{3.49}$$

To compute it, we first need to express the distribution of the inputs generated by the local module and the distribution of the inputs generated by the other modules. In the asymptotic limits we consider, this sum will be dominated by the most probable term of the sum, we will thus need to find the couple (S_l, S_e) that maximizes $\mathbb{P}_l(S_l)\mathbb{P}_e(S_e)$.

3.6.1 Distribution of inputs for the unstructured model

We apply the method sketched above to compute the distributions of inputs (3.9) (when the network is set in a memory state $\vec{\Xi}^{\mu_0}$) on selective ($(i,m)/\Xi_{i,m}^{\mu_0} = 1$) and

non selective neurons in the model with unstructured macroscopic patterns. We first derive these expressions in the case where the fluctuations on the number of active modules sending inputs to a given module are negligible. This is correct to assume when the average of this number is $FD \gg fN$ (or $FD \gg \ln N$). We then explain how to take these fluctuations into account as is done to get the curves of figure 3.2B.

Case $FD \gg \ln N$

Selective neurons - As the number of neurons active in each module is exaclty fN, and because the network is fully connected, the Willshaw learning rule ensures that the local part of the input is

$$\mathbb{P}_l^s(S_l) = \delta\left(S_l - fN\right) \tag{3.50}$$

If a module receives input from exactly the same number of modules FD, the external part of the field is simply written

$$\begin{aligned} \mathbb{P}_e^s(S_e) &= \binom{FDfN}{S_e} \left(\frac{d}{N}\right)^{S_e} \left(1 - \frac{d}{N}\right)^{FDfN - S_e} \\ &= \exp\left[-fN\Phi^{dc}\left(\frac{S_e}{fN}, \gamma F\right) + o(fN)\right] \end{aligned} \tag{3.51}$$

The total input on selective neurons is then

$$\mathbb{P}(h_{i,m} = S = fN + S_e | \Xi_{i,m}^{\mu_0} = 1) = \exp\left[-fN\Phi^{dc}\left(\frac{S_e}{fN}, \gamma F\right) + o(fN)\right] \tag{3.52}$$

Non-selective neurons - The local input now fluctuates because of inputs mediated by synapses that have been potentiated during the presentation of patterns $\vec{\Xi}^{\mu \neq \mu_0}$. It is distributed according to

$$\begin{aligned} P_l(S_l) &= \binom{fN}{S_l} g^{S_l} (1-g)^{fN - S_l} \\ &= \exp\left[-fN\Phi^{fc}\left(\frac{S_l}{fN}, g\right) + o(fN)\right] \end{aligned} \tag{3.53}$$

where g is defined in eq. (3.16). Similarly, the external part of the input can be modeled by

$$
\begin{aligned}
P_e(S_e) &= \binom{FDfN}{S_e}\left(\frac{d}{N}G\right)^{S_l}\left(1-\frac{d}{N}G\right)^{FDfN-S_e} \\
&= \exp\left[-fN\Phi^{dc}\left(\frac{S_e}{fN},\gamma FG\right)+o(fN)\right] \qquad (3.54)
\end{aligned}
$$

The distribution of the total input is now written

$$
\begin{aligned}
&\mathbb{P}(h_{i,m}=S|\Xi_{i,m}^{\mu_0}=0,\Xi_m^{\mu_0}=1)= \\
&\sum_{S_l,S_e/S_l+S_e=S}\exp\left[-fN\left(\Phi^{fc}\left(\frac{S_l}{fN},g\right)+\Phi^{dc}\left(\frac{S_e}{fN},\gamma FG\right)\right)+o(fN)\right] \\
&= \exp\left[-fN\left(\Phi^{fc}\left(s_l^*,g\right)+\Phi^{dc}\left(s_e^*,\gamma FG\right)\right)+o(fN)\right] \qquad (3.55)
\end{aligned}
$$

with $s_l^*=\frac{S_l^*}{fN}$ and $s_e^*=\frac{S_e^*}{fN}=s-s_l^*$ (where $s=\frac{S}{fN}$) such that $\frac{\partial\left(\Phi^{fc}(s1,g)+\Phi^{dc}(s-s_l,\gamma FG)\right)}{\partial s_l}(s_l^*)$ 0. $s_l^*(s)$ and $s_e^*(s)$ can be computed

$$
\begin{aligned}
s_l^* &= \frac{1}{2}\left(1+s+\frac{(1-g)\gamma FG}{g}\right)-\frac{1}{2}\sqrt{\left(1+s+\frac{(1-g)\gamma FG}{g}\right)^2-4s} \\
s_e^* &= -\frac{1}{2}\left(1-s+\frac{(1-g)\gamma FG}{g}\right)+\sqrt{\left(1-s+\frac{(1-g)\gamma FG}{g}\right)^2+4\frac{(1-g)\gamma FG}{g}s} \qquad (3.56)
\end{aligned}
$$

Probability of no errors - We have derived the expressions for the distribution of inputs to both selective and non-selective neurons. In order to compute the probability that there is no error $\mathbb{P}_{ne}$ in the retrieval of pattern $\vec{\Xi}^{\mu_0}$, we have to estimate the probability that the inputs are above or below threshold, as written in the main text in equations (3.9). To do so we first note that

$$
\mathbb{P}(h_{i,m}\geq\theta fN|\Xi_{i,m}^{\mu_0})=\mathbb{P}(h_{i,m}=\theta fN|\Xi_{i,m}^{\mu_0})\sum_{s\geq\theta}\frac{\mathbb{P}(h_{i,m}=sfN|\Xi_{i,m}^{\mu_0})}{\mathbb{P}(h_{i,m}=\theta fN|\Xi_{i,m}^{\mu_0})} \qquad (3.57)
$$

where the '$\sum$' term will not contribute to the final expression of $\mathbb{P}_{ne}$ in the large N limit, as shown in chapter 2. In practice we thus replace the probability to be above threshold by the probability to be at threshold. We can apply the same reasoning for the probability to be above the activation threshold for non-selective

neurons. We now have all the elements to express Φ^s and Φ^{ns} in formulas (3.9) for the case of unstructured macroscopic patterns

$$\Phi^s = \Phi^{dc}\left(\theta - 1, \gamma F\right) \tag{3.58}$$

and

$$\Phi^{ns} = \Phi^{fc}\left(s_l^*(\theta), g\right) + \Phi^{dc}\left(s_e^*(\theta), \gamma F G\right) \tag{3.59}$$

Case $FD = O(\ln N)$

The fluctuations in the number of active modules connected to a given module has now to be taken into account. It changes the distribution of the external input, for selective neurons

$$\begin{aligned}
\mathbb{P}_e^s(S_e) &= \sum_{Y=0}^{FM} \exp\left[-FD\Phi_0 - fN\Phi^e(\frac{S_e}{fN}, y\gamma F) + o(fN) + o(FD)\right] \\
&= \sum_{Y=0}^{FM} \exp\left[-FD\Phi_0 - fN\Phi^{dc}(\frac{S_e}{fN}, y\gamma F) + o(fN) + o(FD)\right] \\
&= \exp\left[-FD\Phi_0(y^*, \frac{D}{M}) - fN\Phi^{dc}(\frac{S_e}{fN}, y^*\gamma F) + o(fN) + o(FD)\right]
\end{aligned} \tag{3.60}$$

with

$$\Phi_0(y, \frac{D}{M}) = y\ln y + (\frac{M}{D} - y)\ln\left(\frac{1 - y\frac{D}{M}}{1 - \frac{D}{M}}\right) \tag{3.61}$$

and $Y = yFD$. With the same reasoning as before, the sum over different Y is dominated by the most probable term y^* that can be determined numerically by

$$\begin{aligned}
&\frac{\partial\left(FD\Phi_0(y, \frac{D}{M}) + fN\Phi^{dc}(\frac{S_e}{fN}, y\gamma F)\right)}{\partial y}(y^*) = 0 \\
\Leftrightarrow \quad & FD\ln\left(y^*\frac{1 - \frac{D}{M}}{1 - y^*\frac{D}{M}}\right) + fN\left(-\frac{S_e}{fNy^*} + \gamma F\right) = 0
\end{aligned} \tag{3.62}$$

The distribution of the field on selective neurons is then

$$\begin{aligned}
\mathbb{P}(h_{i,m} = S = fN + S_e | \Xi_{i,m}^{\mu_0} = 1) &= \exp\Big[-FD\Phi_0^s(y^*, \frac{D}{M}) - fN\Phi^{dc}(\frac{S_e}{fN}, y^*\gamma F) \\
&\quad + o(fN) + o(FD)\Big]
\end{aligned} \tag{3.63}$$

For non-selective neurons, the distribution of external input is also modulated by the fluctuations on Y and the distribution of the total input is

$$\begin{aligned}
&\mathbb{P}(h_{i,m} = S | \Xi_{i,m}^{\mu_0} = 0) \\
&= \sum_y \exp\left[-FD\Phi_0(y, \frac{D}{M})\right] \sum_{S_l, S_e / S_l + S_e = S} \exp\left[-fN\left(\Phi^{fc}(\frac{S_l}{fN}, g) + \Phi^{dc}(\frac{S_e}{fN}, y\gamma FG)\right)\right] \\
= &\sum_y \exp\left[-FD\Phi_0(y, \frac{D}{M})\right] \exp\left[-fN\left(\Phi^{fc}(s_l^*(y), g) + \Phi^{dc}(s_e^*(y), y\gamma FG)\right)\right] \\
= &\exp\left[-FD\Phi_0(y^*, \frac{D}{M}) - fN\left(\Phi^{fc}(s_l^*(y^*), g) + \Phi^{dc}(s_e^*(y^*), y^*\gamma FG)\right)\right] \qquad (3.64)
\end{aligned}$$

where the expressions of s_l^* and s_e^* are given by equations (3.56) with γF replaced by $y^*\gamma F$. Again y^* can be determined numerically from its definition

$$\begin{aligned}
&\frac{\partial\left(FD\Phi_0(y, \frac{D}{M}) + fN\left(\Phi^{fc}(s_l^*(y), g) + \Phi^{dc}(s_e^*(y), y\gamma FG)\right)\right)}{\partial y}(y^*) = 0 \\
\Leftrightarrow \quad &FD\ln\left(y^*\frac{1-\frac{D}{M}}{1-y^*\frac{D}{M}}\right) + fN\left(\gamma FG - \frac{s - s_1^*(y^*)}{y^*}\right) = 0 \qquad (3.65)
\end{aligned}$$

The probability that the field is below/above threshold for selective/non-selective neurons (3.9) is now given by

$$\Phi^s = \Phi^{dc}\left(\theta - 1, y^*\gamma F\right) + \frac{FD}{fN}\Phi_0(y^*, \frac{D}{M}) \qquad (3.66)$$

and

$$\Phi^{ns} = \Phi^{fc}\left(s_l^*(\theta), g\right) + \Phi^{dc}\left(s_e^*(\theta), y^*\gamma FG\right) + \frac{FD}{fN}\Phi_0(y^*, \frac{D}{M}) \qquad (3.67)$$

3.6.2 Distribution of inputs for the categorized model

In this case, there is no macroscopic dilution and each active module receives inputs from exactly $F(M-1)$. This removes the fluctuations on the number of active modules sending inputs to a given module, and allows to work with $FM = O(1)$. However, the expression of the rate functions will slightly differ when considering $FM \to +\infty$ (and $\frac{d}{N} \to 0$, see (3.25)) or $FM = O(1)$ (and $\frac{d}{N} = O(1)$). Thus we present results for the two cases separately.

Case $FM \to +\infty$

The strategy to compute the distribution of inputs is similar to the unstructured model. We first focus on selective neurons and split the input into a local component and an external component (as in eq. (3.49)).

Selective neurons - The local component do not fluctuate and is given by eq. (3.50). The external component is given by

$$\begin{aligned}\mathbb{P}_e(S_e) &= \binom{FMfN}{S_e}\left(\frac{d}{N}\right)^{S_e}\left(1-\frac{d}{N}\right)^{FMfN-S_e} \\ &= \exp\left[-fN\Phi^{dc}\left(s_e,\frac{\gamma}{\delta}\right)\right] \end{aligned} \tag{3.68}$$

with $s_e = \frac{S_e}{fN}$.

Non-selective neurons - The distribution of the local field is the same as (3.54) with γF replaced by $\frac{\gamma}{\delta}$. Also note that g and G are defined by (3.26) and (3.28). Doing so we have assumed that the amount of noise coming from the storage of patterns different than $\vec{\Xi}^{\mu_0}$ is the same for every pair of modules. This is the case if each pair is co-activated in the same number of categories. In the case we are considering, a given pair is co-activated in at least one category with a probability $R \underset{F\to 0}{\simeq} \delta F \to 0$. Thus in practice, a pair of module activated in pattern $\vec{\Xi}^{\mu_0}$ is activated only in this category.

Probability of no errors - The distribution of the total input can be expressed by taking values s_l^* and s_e^* defined by equations (3.56) with γF replaced by $\frac{\gamma}{\delta}$. The rate functions for the categorized model in the case $FM \to +\infty$ are

$$\Phi^s = \Phi^{dc}\left(\theta - 1, \frac{\gamma}{\delta}\right) \tag{3.69}$$

and

$$\Phi^{ns} = \Phi^{fc}\left(s_l^*(\theta), g\right) + \Phi^{dc}\left(s_e^*(\theta), \frac{\gamma}{\delta}\right) \tag{3.70}$$

Case $FM = O(1)$

Now the microscopic dilution term $\frac{d}{N}$ is finite and we have to use Φ^{fc} instead of Φ^{dc} to describe the external inputs. At the end the rate functions are

$$\Phi^s = (\theta-1)\ln\left(\frac{\theta-1}{FM\frac{d}{N}}\right) + (FM-(\theta-1))\ln\left(\frac{1-(\theta-1)/FM}{1-d/N}\right) \tag{3.71}$$

and

$$\Phi^{ns} = \Phi^{fc}\left(s_l^*(\theta), g\right) + s_e^*(\theta)\ln\left(\frac{s_e^*(\theta)}{FM\frac{d}{N}G}\right) + (FM - s_e^*(\theta))\ln\left(\frac{1 - s_e^*(\theta)/FM}{1 - d/N}\right) \tag{3.72}$$

with s_l^* and s_e^* given by

$$\begin{aligned} s_l^* &= \frac{\lambda(FM - s) + 1 + s}{2(1-\lambda)} - \frac{1}{2}\sqrt{\left(\frac{\lambda(FM - s) + 1 + s}{1 - \lambda}\right)^2 - 4\frac{s}{1-\lambda}} \\ s_e^* &= \frac{-\lambda(FM + s) + s - 1}{2(1-\lambda)} + \frac{1}{2}\sqrt{\left(\frac{-\lambda(FM + s) + s - 1}{(1 - \lambda)}\right)^2 + 4\frac{\lambda FMs}{1-\lambda}} \end{aligned} \tag{3.73}$$

3.6.3 Storage capacity with error correction

In the section 3.4. the network is able to correct for errors because we choose an activation threshold smaller than the average activity received by selective neurons when the network is in the pattern $\vec{\Xi}^{\mu_0}$. The expression of the rate functions we have established above are valid for arbitrary θ, thus they can be used to estimate storage capacity also for low activation threshold.

Chapter 4

The whole cortex as an attractor network ?

4.1 Introduction

In this chapter, we study the storage capacity of a network with a different approach than previously used. In the work presented so far, network connectivity was explicitly determined by the patterns of activity we were trying to imprint as fixed points of the network dynamics. For instance, in chapter 2 we found that networks with binary synapses updated stochastically at each pattern presentation with a simple Hebbian-learning rule were able to store 0.23bits/synapse if multiple presentations of the patterns are allowed, while it has been shown using Gardner's approach that networks with binary synapses are able to store up to 0.59bits/synapse (Gutfreund and Stein, 1990) . The principle of Gardner's approach is to explore the space of network connectivities, and to examine the sub-space in which a set of P patterns are fixed points of the dynamics of the network, independently of a learning rule explicitly specifying connectivity as a function of the identity of the patterns to store. Notably, this allows to compute storage capacities by finding the value P_c such that this sub-space has a size that vanishes. Moreover, this technique allows to give a statistical description of the connectivity storing a set of P patterns (Brunel et al., 2004). This has been applied to interpret experimental measures of network connectivity, for instance for synaptic connectivity from granule cells to Purkinje cells in the cerebellum. The network constituted by a Purkinje cell and its pre-synaptic granule cells can be seen as a perceptron storing input-output associations, and as presented in the introduction chapter, pair recordings of granule-Purkinje cells allow to quantify the connectivity of this network. Brunel et al. (2004) have studied a perceptron with positive weights (modeling excitatory synapses from granule to Purkinje cells) and derived an expression for the distribution of synaptic weights when the perceptron is at

maximal capacity. They found parameters such that the experimental distribution is well fitted by the theoretical distribution (notably the unexpected fact that more than 50% of the synapses are silent), allowing to interpret the experimental connectivity as maximizing the number of input-output associations the network can store. As mentioned in the introduction, Chapeton et al. (2012) have used a similar reasoning to interpret features of the connectivity of local cortical networks.

Here we study a network in which two units are connected via an excitatory connection whose amplitude is assigned a specific cost. This cost is used to take into account the physical distance separating two units in cortex. We interpret this network as the network of cortical areas and compare the theoretical distributions of weights with the quantitative experimental measures of weights between cortical areas presented in the introduction. Note that in practice the problem of finding weights that stabilize a given pattern of activity in a recurrent network with N neurons is equivalent to finding weights to store input-output associations for N perceptrons with N neurons in the input layer.

4.2 Model: Perceptron with a distance constraint

We study a perceptron with N synapses w_i that have to classify P input-output associations, $\{\vec{\xi}^\mu, \xi_0^\mu\}$ where $\vec{\xi}^\mu$ are vectors of length N and $\mu = 1, ..., P$. The input and output patterns are random independent binary variables with coding level f, that is

$$P(\xi_i^\mu = 1) = f \text{ and } P(\xi_i^\mu = 0) = 1 - f \tag{4.1}$$

for all $i = 0, ..., N$ and all μ. We define the stability of pattern μ

$$\Delta^\mu = \frac{\zeta^\mu}{\sqrt{N}} \left(\sum_{i=1}^{N} w_i \xi_i^\mu - N\theta \right) \tag{4.2}$$

where θ is an activation threshold and $\zeta^\mu = 2\xi_0^\mu - 1$. We say the P patterns are learned (or the 'storage problem' is satisfied) if the weights are such that

$$\Delta^\mu \geq \kappa \text{ for all } \mu \tag{4.3}$$

where κ is a robustness parameter. Note that when learning is successful, we expect that in the large N limit, the average weight $\overline{W} = \sum_i w_i$ is such that the

average field on the output unit is of the order of the threshold, which translates to $\overline{W} \simeq \frac{\theta}{f}$.
In order to model the fact that connections between two units (i, j) have a cost that depends on the distance between them, we introduce the *distance constraint*,

$$\sum_{i=1}^{N} w_i \rho_i = \lambda N \tag{4.4}$$

where ρ_i represents a cost that increases with the distance between the input unit indexed by i and the output unit. Calculations can be done for arbitrary $\vec{\rho}$, but we will focus on constraints such that $\sum_i \rho_i \simeq 1$. To better understand the meaning of equation (4.4), first consider a weight vector that has learned the P patterns without the distance constraint. For this we have $\sum_i w_i \rho_i = \overline{W} N + O(\sqrt{N})$, such that in the large N limit, the unconstrained learning leads to (4.4) with $\lambda = \overline{W}$. If we want (4.4) to be satisfied with $\lambda < \overline{W}$, we encourage weights i such that $\rho_i > 1$ to be smaller than $\overline{W}$ and weights such that $\rho_i < 1$ to be larger than $\overline{W}$, which corresponds to the intuitive idea that weights between distant units are penalized.

4.3 Calculation of storage capacity with the replica method

In order to evaluate how the distance constraint reduces storage capacity, we use the replica formalism to evaluate for which value of $\alpha = \frac{P}{N}$ the volume of weights satisfying the constraints (4.3) and (4.4) shrinks to zero. In the space of couplings the volume of weights satisfying the learning problem is

$$V = \frac{\int \left(\prod_{i=1}^{N} dw_i\right) \delta\left(\sum_i w_i \rho_i - \lambda N\right) \prod_{\mu=1}^{P} \Theta(\Delta^\mu - \kappa)}{\int \left(\prod_{i=1}^{N} dw_i\right) \delta\left(\sum_i w_i \rho_i - \lambda N\right)} \tag{4.5}$$

We are interested in the typical value of V. This typical value can be obtained for the logarithm of V by averaging over the distribution of random patterns ($\langle . \rangle$). We introduce n replicas of the system in order to compute $\langle \ln V \rangle$ using the replica trick (Mézard et al., 1987)

$$\langle \ln V \rangle = \lim_{n \to 0} \frac{\langle V^n \rangle - 1}{n} \tag{4.6}$$

The quantity $\langle V^n \rangle$ is the volume of the sub-space of weights that satisfy the learning problem in n ('replicated') neural networks. We start by computing the average over the distribution of patterns of the volume of the replicated system:

$$\langle V^n \rangle = \frac{\int \left(\prod_{i=1}^{N} \prod_{\alpha=1}^{n} dw_i^\alpha\right) \left(\prod_{\mu=1}^{P} \prod_{\alpha=1}^{n} \delta\left(\sum_i w_i^\alpha \rho_i - \lambda N\right)\right) \left\langle \left(\prod_{\mu=1}^{P} \prod_{\alpha=1}^{n} \Theta\left(\Delta^{\mu,\alpha} - \kappa\right)\right)\right\rangle}{\int \left(\prod_{i=1}^{N} \prod_{\alpha=1}^{n} dw_i^\alpha\right) \left(\prod_{\mu=1}^{P} \prod_{\alpha=1}^{n} \delta\left(\sum_i w_i^\alpha \rho_i - \lambda N\right)\right)} \tag{4.7}$$

In order to compute these integrals, a first step is to introduce integral representations of the δ and Θ functions. This allows to express this quantity only with integrals of regular functions, which makes it easier to perform the average over the distribution of patterns to store. $\langle V^n \rangle$ can then be estimated using a steepest descent method. To do so the number of variables on which we integrate has to be finite. Thus we introduce a finite number of variables describing macroscopic features of the synaptic connectivity :

$$q^{\alpha\beta} = \frac{1}{N} \sum_{j=1}^{N} w_j^\alpha w_j^\beta \tag{4.8}$$

$$Q^\alpha = \frac{1}{N} \sum_{j=1}^{N} \left(w_j^\alpha\right)^2 \tag{4.9}$$

$$\frac{M^\alpha}{\sqrt{N}} = \frac{1}{N} \sum_{j=1}^{N} w_j^\alpha - \frac{\theta}{f} \tag{4.10}$$

which respectively describes the overlap of the synaptic weight vectors of two different replicated networks indexed by α and β ($\alpha, \beta = 1, ..., n$), the norm of the synaptic weight vector of network α, the distance between the average weight of network α and activation threshold divided by the fraction of active units in each pattern of activity, which we expect to scale as $\frac{1}{\sqrt{N}}$ if an extensive number of patterns are stored (see equation (4.2)). In order to introduce these parameters in the expression of $\langle V^n \rangle$, we also need to introduce the conjugate parameters $\hat{q}^{\alpha\beta}$, $\hat{Q}^\alpha$, $\hat{M}^\alpha$ such that

$$\delta\left(N q^{\alpha\beta} - \sum_j w_j^\alpha w_j^\beta\right) = \int \frac{d\hat{q}^{\alpha\beta}}{2\pi i} \exp\left(-N q^{\alpha\beta} \hat{q}^{\alpha\beta} + \hat{q}^{\alpha\beta} \sum_j w_j^\alpha w_j^\beta\right) \tag{4.11}$$

$$\delta\left(N Q^\alpha - \sum_j \left(w_j^\alpha\right)^2\right) = \int \frac{d\hat{Q}^\alpha}{2\pi i} \exp\left(-N Q^\alpha \hat{Q}^\alpha + \hat{Q}^\alpha \sum_j \left(w_j^\alpha\right)^2\right) \tag{4.12}$$

$$\delta\left(\sum_j w_j^\alpha - N\frac{\theta}{f} - M^\alpha \sqrt{N}\right) = \int \frac{d\hat{M}^\alpha}{2\pi i} \exp\left(\sqrt{N} M^\alpha \hat{M}^\alpha + N\frac{\theta}{f} \hat{M}^\alpha - \hat{M}^\alpha \sum_j w_j^\alpha\right) \tag{4.13}$$

We also introduce a parameter E^α to enforce the distance constraint (4.4)

$$\delta\left(\sum_j \rho_j w_j^\alpha - \lambda N\right) = \int \frac{dE^\alpha}{2\pi i} \exp\left(\lambda N E^\alpha - \sum_j \rho_j w_j^\alpha\right) \tag{4.14}$$

All these elements allow to prepare $\langle V^n \rangle$ for the steepest descent method (see Methods section for details)

$$\langle V^n \rangle = \frac{\int dM d\hat{M} dQ d\hat{Q} dq d\hat{q} dE \exp(NnF)}{\int dE \exp(NnH)} \tag{4.15}$$

with

$$F = \alpha Z_1 + Z_2 + Z_3 + O(\frac{1}{\sqrt{N}}) \tag{4.16}$$

where

$$Z_1 = \int Dt \sum_{\zeta=\pm 1} p_\zeta \ln\left[H\left(\frac{\kappa - f\zeta M + t\sqrt{f(1-f)q}}{\sqrt{f(1-f)(Q-q)}}\right)\right] \tag{4.17}$$

$$Z_2 = \int d\rho P(\rho) \int Dt \ln\left[\int_0^{+\infty} dw \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)\right] \tag{4.18}$$

$$Z_3 = -Q\hat{Q} + \frac{1}{2} q\hat{q} + \frac{\theta}{f}\hat{M} + \lambda E \tag{4.19}$$

With $P(\rho)$ the density probability function from which the elements of $\vec{\rho}$ are drawn and $p_{\zeta=+1} = f$, $p_{\zeta=-1} = 1 - f$. The order parameters M, Q, q and the conjugate parameters $\hat{M}, \hat{Q}, \hat{q}$ do not depend on the replica indeces anymore, as we have assumed that for each replicated network these macroscopic quantities are the same (replica symmetric solution). For n finite and $N \to +\infty$ we apply a steepest descent method and found the set of $q, Q, M, \hat{q}, \hat{Q}, \hat{M}, E$ maximizing F. The saddle point equations defining these parameters are given in the Methods section (4.39) to (4.47). Note that they depend on the memory load α.

The storage capacity corresponds to the maximum number of patterns that can be stored. As mentioned above, at maximal capacity we expect that the volume of the sub-space of weights satisfying (4.3) shrinks to zero, i.e. only one synaptic weight vector satisfies the storage problem. In this case, the weights are the same in each replicated system, and we expect

$$q \to Q \tag{4.20}$$

Solving the saddle point equations (4.49) to (4.55) in this limit allows to compute the storage capacity as a function of the various parameters. In figure 4.1A we show how the storage capacity depends on the distance constraint via the parameter λ, for $\theta = 0.5$, $f = 0.5$ and $\kappa = 0$. In this figure and in the following, the value of the elements of the cost vector ρ_i's are distributed normally with mean 1 and standard deviation $\sigma = 0.38$ (see section 'Comparison with experimental data' for an explanation of this choice). For $\lambda = \frac{\theta}{f} = 1$, we recover the known result that $\alpha_c = 1$ when no distance constraint is imposed (when no constraint is imposed, the average weight equals $\frac{\theta}{f} + O\left(\frac{1}{\sqrt{N}}\right)$). As the constraint gets stronger $\lambda < \frac{\theta}{f}$, capacity is reduced. The solid line shows the theoretical prediction and the squares with error bars are obtained by simulating learning in 100 perceptrons with $N = 100$ and the above mentioned parameters.

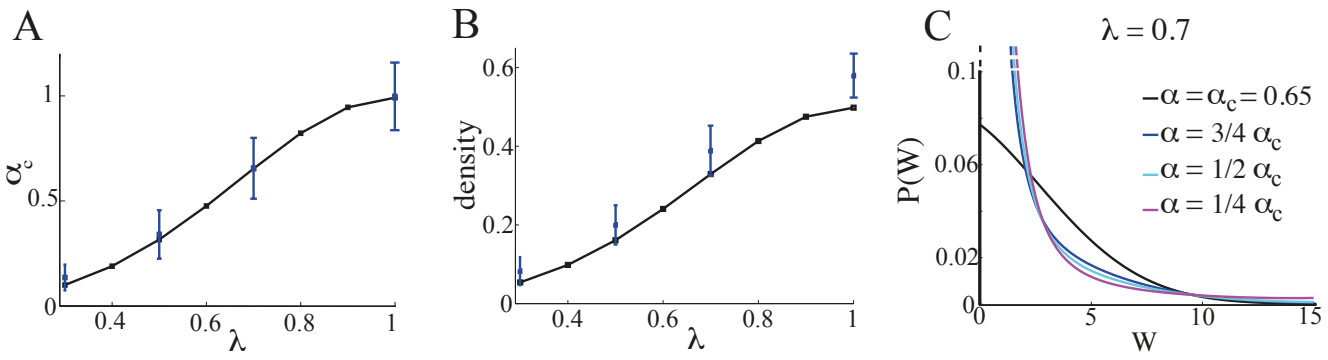

Figure 4.1: *Properties of the perceptron with a distance constraint. In these figures the costs ρ_i's are distributed normally with a mean* 1 *and standard deviation $\sigma = 0.38$ The other parameters are $f = 0.5$, $\theta = 0.5$ and $\kappa = 0$. A- Storage capacity as a function of λ (see equation* (4.4) *for definition). Black line is the value of α_c obtained by solving the saddle point equations* (4.49)-(4.55) *in the limit $q \to Q$. Blue line is the average of $\alpha_c = \frac{P}{N}$ corresponding to the number of patterns that can be learned by a simulated perceptron with $N = 100$ input units. Error bars are the standard deviations obtained from* 100 *such simulations. B- Density (fraction of non-zero synapses) of the weights vector at maximal capacity. C- Theoretical distributions of weights at different loading for $\lambda = 0.7$.*

4.4 Distribution of weights

We want to compute the typical probability $\langle P(W_{i_0}) \rangle$ that a weight with cost ρ_{i_0} takes the value W_{i_0}. It writes,

$$\langle P(W_{i_0}) \rangle = \langle \frac{1}{V} \frac{\int \left(\prod_{i=1}^{N} dw_i\right) \delta(W_{i_0} - w_{i=i_0}) \delta\left(\sum_i w_i \rho_i - \lambda N\right) \prod_{\mu=1}^{P} \Theta(\Delta^\mu - \kappa)}{\int \left(\prod_{i=1}^{N} dw_i\right) \delta\left(\sum_i w_i \rho_i - \lambda N\right)} \rangle \tag{4.21}$$

with V defined in eq.(4.5). It can be rewritten,

$$\langle P(W_{i_0})\rangle = \lim_{n\to 0}\langle V^{n-1}\frac{\int\left(\prod_{i=1}^{N} dw_i\right)\delta(W_{i_0}-w_{i_0})\delta\left(\sum_i w_i\rho_i-\lambda N\right)\prod_{\mu=1}^{P}\Theta(\Delta^\mu-\kappa)}{\int\left(\prod_{i=1}^{N} dw_i\right)\delta\left(\sum_i w_i\rho_i-\lambda N\right)}\rangle \tag{4.22}$$

As in the previous section, we introduce n replicas of the system where one of the replica has the particularity of displaying the term $\delta(W_{i_0}-w_{i_0})$, we compute the disorder average for n an integer, and then perform the limit $n\to 0$.

Following the same calculations has above, we can derive

$$\langle P(W_{i_0})\rangle = \int Dt\frac{\exp\left(\left(\hat{Q}-\frac{\hat{q}}{2}\right)W_{i_0}^2-\left(\hat{M}-t\sqrt{\hat{q}}-E\rho_{i_0}\right)W_{i_0}\right)}{\int_0^{+\infty}dw\exp\left(\left(\hat{Q}-\frac{\hat{q}}{2}\right)w^2-\left(\hat{M}-t\sqrt{\hat{q}}-E\rho_{i_0}\right)w\right)}\Theta(W_{i_0}) \tag{4.23}$$

Where the conjugate parameters are solutions of the same system of saddle point equations (4.39)-(4.47) mentioned above.

At maximal capacity, this distribution is a truncated gaussian with a mean that depends on the cost ρ_i,

$$\begin{aligned}\langle P(W_{i_0})\rangle &= \delta(W_{i_0})\,H\left[-(B+D\rho_{i_0})\right]+\\ &\quad\Theta(W_{i_0})\,\frac{1}{\sqrt{2\pi}W_s}\exp\left[-\frac{1}{2W_s^2}\left(W_{i_0}+(B+D\rho_{i_0})W_s\right)^2\right]\end{aligned} \tag{4.24}$$

with $W_s=\frac{\sqrt{C}}{A}$ and A,B,C,D the solution of the saddle point equations at maximal capacity.

The distribution of all the weights, independently of the neuron index i can be obtained by integrating over the distribution of costs. In the subcritical case ($\alpha<\alpha_c$) the probability $P(W)$ to find a weight at a value W at the end of learning is

$$P(W)=\int d\rho P(\rho)\int Dt\frac{\exp\left(\left(\hat{Q}-\frac{\hat{q}}{2}\right)W^2-\left(\hat{M}-t\sqrt{\hat{q}}-E\rho\right)W\right)}{\int_0^{+\infty}dw\exp\left(\left(\hat{Q}-\frac{\hat{q}}{2}\right)w^2-\left(\hat{M}-t\sqrt{\hat{q}}-E\rho\right)w\right)}\Theta(W) \tag{4.25}$$

In the case we are considering where they are distributed normally with a variance σ and average 1, the distribution of weights at maximal capacity is still a truncated Gaussian

$$P(W) = \delta(W) H\left(-\frac{B+D}{\sqrt{1+D^2\sigma^2}}\right) + \Theta(W) \frac{\exp\left(-\frac{1}{2}\frac{(W+(B+D)W_s)^2}{W_s^2(1+D^2\sigma^2)}\right)}{\sqrt{2\pi W_s^2(1+D^2\sigma^2)}} \tag{4.26}$$

This expression is valid for every values of λ, the dependence on this parameter comes from the values of the order parameters obtained by solving the saddle point equations that depend on λ.

In figure 4.1B, we plot the density of the connectivity, defined as the fraction of W's that are non-zero after learning. For the same parameters that are used in figure 4.1A. Again for $\lambda = \frac{\theta}{f} = 1$, we find that half of the synapses have values zero (Brunel et al., 2004). For $\lambda < \frac{\theta}{f}$, the density is decreasing. Blue points with error bars show the results of simulations. In figure 4.1C, we plot the distributions of weights obtained from equations (4.25) or (4.26) when the network is at maximal capacity α_c or below, for $\lambda = 0.7$. When the network is loaded up to its critical capacity, the distribution is composed of a peak at zero and a truncated Gaussian. For sub-critical loading, there is no concentration of synapses at 0, but many weights will have low values.

4.5 Comparison with experimental data

4.5.1 Networks at maximal capacity

We now compare properties of the connectivity at maximal capacity with the connectivity measured by Markov et al. (2012) when measuring the weights between cortical areas (see the chapter 'Introduction' for a more detailed presentation of these data). Doing so, we interpret each unit of the theoretical network as representing a cortical area. A pattern of activity then models a cortical state with a fraction f of the cortical areas active. And the weight between two cortical areas i and j corresponds to the FLN from j to i, the ratio of neurons that are counted in j when injecting tracers in area i to the total number of neurons counted in all areas when injecting in i. We want to assess whether the measured statistics of cortical connectivity is consistent with the connectivity of a network maximizing the number of stable states that can be learned. To do so, we choose a particular distance vector $\vec{\rho}$ where each element ρ_i is drawn independently from a gaussian distribution of mean $\mu = 1$ and standard deviation $\sigma = \frac{10.11mm}{26.57mm} = 0.38$. This choice is determined by the fact that the distribution of distances between cortical areas is well fitted by a Gaussian distribution with $\mu = 26.57mm$ and $\sigma = 10.11mm$ (Ercsey-Ravasz et al., 2013). With this choice the distance constraint corresponds

to a cost ρ_j equals to the distance between areas. For this $\vec{\rho}$, the global distribution of weights can be computed with equation (4.26).

Figure 4.2A shows how the theoretical and simulated distributions of non-zero weights depend on the parameters λ for $f = 0.5$, $\theta = 0.5$, $\kappa = 0$. To build these distributions, we proceed as follows. For the simulated distributions (blue histograms), we simulate learning in 50 perceptrons (with $N = 100$) subject to the distance constraint (4.4), from which we obtain sets of N weights. In the experimental data, the values of the logarithm (in base 10) of the FLN range from $\sim 10^{-6} - 10^{0}$, and by definition the sum of the FLN from the areas $j = 1...N$ to a given injected area i is 1. We reproduced these features by normalizing the obtained weights and applying a cut-off (setting low weights at 0) such that the obtained weights have a logarithm that range between $\sim 10^{-6} - 10^{0}$ and sum to 1. To obtain the theoretical distributions (black curves), we apply the same procedure with weights drawn from the distribution (4.26).
The histograms we have obtained should be compared with the left panel of figure 4.3A, the distribution of non-zero FLN obtained experimentally. For our model, at maximal capacity, the distribution of non-zero weights is peaked and concentrated at rather high values, around $\log_{10}(FLN) \simeq 10^{-2}$, compared to the measured experimental data for which the distribution is much broader and peaked around 10^{-3}. We have explored how this depends on the other parameters f, θ and κ and always found such distributions, concentrated at FLN around 10^{-2}. Such concentration of FLN can be paralleled to the critical distribution of figure 4.1C where most of the weights are of the same order than the mean weight. It can be interpreted by saying that at optimal capacity, each non zero weight has a contribution to the total input sent to the output neuron that is of the same order. Too large weights being detrimental because they tend to drive the output neuron above threshold in patterns that should produce a 0 output, and weights too low being useless in driving the output neurons above threshold in patterns that should produce a 1 output.

Another feature of the experimentally measured cortical connectivity matrix that is not captured by the distribution of FLN is the density of the binary cortical connectivity matrix (the number of non-zero entries in the cortical connectivity matrix). In figure 4.2B, we show how the density, i.e. the fraction of non-zero weights, depends on the distance between two areas (which depends on the index i in equation (4.24)) for different values of λ. The average density is shown in figure 4.1B. The evolution of these histograms, from nearly flat to a more and

more pronounced decrease of density with distance, shows that what we called the 'distance constraint' indeed favors connections between areas that are close to each others. Even though the general trend (see figure 4.3) of the dependence with distance can be captured, the theoretical and simulated density is always too low (for satisfying density profiles). Indeed, the experimental measures of connectivity between cortical areas found an average density of 66%. For all the parameters tested, at maximal capacity, the density we have found was always smaller than 50%, as illustrated by figure 4.1B.

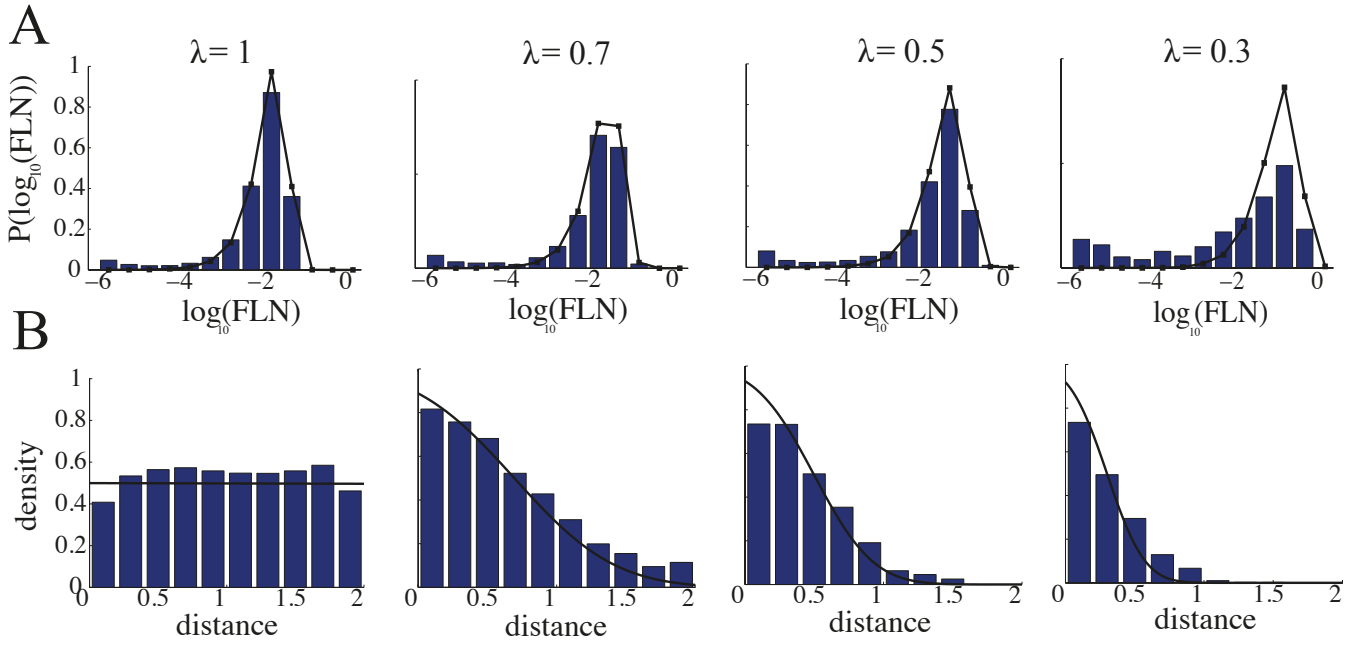

Figure 4.2: *Dependence on λ of the distributions of non-zero weights, and network densities as a function of the distance between units at maximal capacity. A- Distributions of the logarithms of the FLN (see text for definition), for different values of λ. Blue histograms are the results of simulating perceptrons reaching maximal capacity with $N = 100$ and the gaussian cost vector. Black lines are obtained by applying the definition of the FLN for weights drawn according to the theoretical distribution* (4.26). *B- Density versus distance (see text for the definition of distance). Blue histograms are simulated data and black lines theoretical predictions. Parameters are $f = 0.5$, $\theta = 0.5$ and $\kappa = 0$.*

4.5.2 Networks below maximal capacity

As suggested by figure 4.1C, the connectivity for $\alpha < \alpha_c$ contains much more low weights, which should lead to broaden the distribution of FLN. Also the theory predicts that only a vanishingly small proportion of weights are 0 at the end of learning, which should lead to densities of 100%. We performed the same kind of simulations than the one described before with $\alpha < \alpha_c$, to see whether we could find connectivities that better match the experimental measurements. In figure 4.3A-B we show the distribution of FLN, the dependence with distance of the density and the average density for simulations with $f = 0.5$, $\theta = 0.5$,

$\kappa = 0$, $\lambda = 0.4$ and $\alpha = \frac{\alpha_c}{2} = 0.10$. The results are now much more similar to the experimental connectivity, except the average density 48% which is still lower than the experimentally observed 66% (we do not obtain 100% densities probably because of the procedure used to transform the obtained weights into FLN's). Note that the parameters of the model leading to these features have been found 'manually'. We have found a single best set of parameters, although the entire parameter space has not been explored.

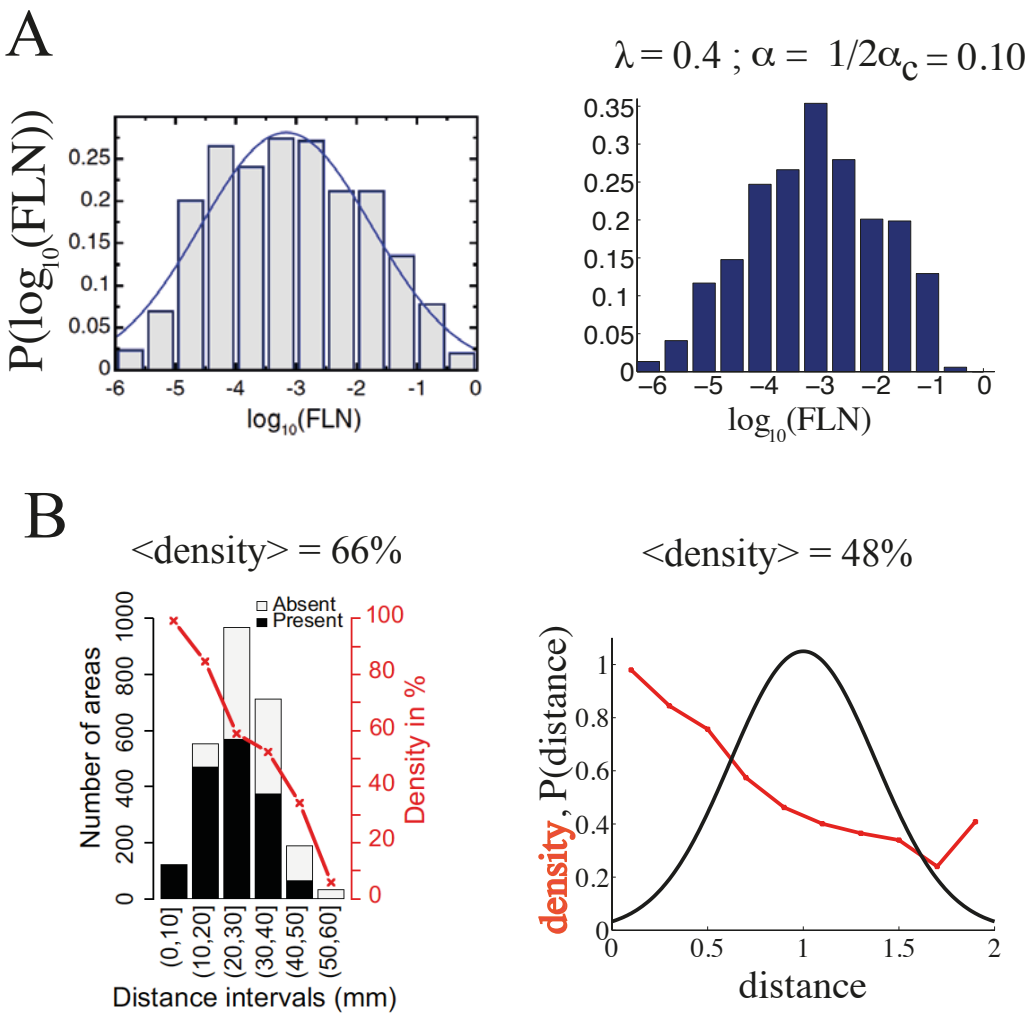

Figure 4.3: *Features of the experimentally measured cortical connectivity matrix compared to a network loaded below its maximal capacity. A- Distribution of the logarithm of the experimental FLN (left, taken from Markov et al. (2012)) and simulated FLN (right). B- Red lines: dependence of network density (i.e. the probability to find a non-zero connection between two areas) as a function of the distance separating to areas as seen in the experiments (left, taken from Markov et al. (2013)) and in the simulations. The black histogram on the left panel shows how the distance between pairs of areas is distributed across the cortical surface. The black line on the right shows the similar distribution we have implemented in our simulations. All the simulations are done with the same parameters* $f = 0.5$, $\theta = 0.5$, $\kappa = 0$, $\lambda = 0.4$ *and* $\alpha = 0.10$

4.6 Discussion

We have studied a perceptron with positive weights where the amplitude of each weight W_i from an input unit i to the output unit is associated with a specific cost ρ_i. We have interpreted this model by identifying each unit i of the perceptron with a cortical area, each weight W_i as a measure of the number of synaptic connections from an area i to the output area, ρ_i as the distance from area i to the output area. In fact the network of cortical areas is a recurrent network, but learning fixed points in recurrent networks with N neurons can be reduced to learning N perceptrons with N input units (Brunel, 2003; Chapeton et al., 2012). Modeling the activity of an entire cortical area with a binary variable is quite bold, as obviously the patterns of activation that a cortical area exhibit are much more numerous than two. However, in order to interpret the available data of Markov et al. (2011) that quantifies connectivity between cortical areas and discards finer structure of cortical connectivity (Pucak et al., 1996), we have chosen such a minimal model.
In a first attempt, we have compared the theoretical distributions of connection weights that optimize storage capacity with the experimentally observed distribution. While the experimental distribution is quite wide, the optimal theoretical distributions are much more narrow (figure 4.2A). Also we have found that the density (fraction of non-zero connections in the network) is always too low for the model (figure 4.1B) compare to the experimental value of 66% (Markov et al., 2011).
In a second attempt, we have compared the experimental data and the theoretical distribution of weights for networks below the optimal capacity. As expected, in this case the distributions of weights get wider and we were able to find a set of network parameters for which the distribution of weights as well as the dependence of the density with distance fit reasonably well the experimental data. Note however that the obtained average density is still lower than the observed one. For these parameters, the network is able to sustain 10 patterns of activity with approximately half of the areas active and the other half silent. Thus we have shown that some statistics of the connectivity matrix of the network of cortical areas is consistent with a simple model with 10 states that are able to self-sustain, however it is not granted that the connectivity, in its detail, supports such states.

4.7 Methods

Here we first give more details on how to express $\langle V^n \rangle$ as defined in equation (4.7), to the simpler expression (4.15). Similar calculations can be found in Gardner

(1988); Hertz et al. (1991); Brunel et al. (2004). We then explicit the system of saddle point equations from which we can extract all the relevant quantities allowing to describe storage capacity and network connectivity.

We first start by performing the average over the distribution of patterns of activity to store. Introducing integral representation of the Heaviside function

$$\theta(z-\kappa) = \int_{\kappa}^{+\infty} dy \int \frac{dx}{2\pi} \exp(ix(y-z)) \tag{4.27}$$

we can rewrite

$$\begin{aligned}
&\langle \prod_{\mu=1}^{P} \prod_{\alpha=1}^{n} \Theta\left(\Delta^{\mu,\alpha} - \kappa\right)\rangle = \\
&\langle \Big[\int_{\kappa}^{+\infty} \prod_{\alpha} dy^{\alpha} \int \prod_{\alpha} \frac{dx}{2\pi} \exp\Big(i \sum_{\alpha=1}^{n} x^{\alpha}(y^{\alpha} - \frac{f\zeta}{\sqrt{N}} \sum_{j=1}^{N} (w_j^{\alpha} - \frac{\theta}{f})) \\
&- \frac{f(1-f)}{2N} \sum_{\alpha} (x^{\alpha})^2 \sum_{j} (w_j^{\alpha})^2 - \frac{f(1-f)}{N} \sum_{\alpha<\beta} x^{\alpha} x^{\beta} \sum_{j} w_j^{\alpha} w_j^{\beta} \Big) \Big]^{P} \rangle_{\zeta}
\end{aligned} \tag{4.28}$$

Where $\langle . \rangle_{\zeta}$ is the average over the distribution of the output unit. We have not performed explicitly this average for formal concision. This expression can be simplified by introducing the order parameters $q^{\alpha\beta}, Q^{\alpha}, M^{\alpha}$ defined in (4.8),(4.9),(4.10). The expression of $\langle V^n \rangle$ can then be rewritten, using integral representation of the δ-functions and introducing the conjugate parameters defined in (4.11),(4.12),(4.13)

$$\begin{aligned}
\langle V^n \rangle &= \int \prod_{\alpha} dM^{\alpha} d\hat{M}^{\alpha} dQ^{\alpha} d\hat{Q}^{\alpha} dE^{\alpha} \prod_{\alpha<\beta} dq^{\alpha\beta} d\hat{q}^{\alpha\beta} \langle \exp NF' \rangle_{\zeta} \\
&\Big[\int \prod_{\alpha} \frac{dE^{\alpha}}{2\pi i} \exp NH \Big]^{-1}
\end{aligned} \tag{4.29}$$

with

$$F' = Z_1' + Z_2' + Z_3' \tag{4.30}$$

where

$$\begin{aligned}
Z_1' &= \frac{P}{N} \ln \Big[\int_{\kappa}^{+\infty} \prod_{\alpha} dy^{\alpha} \int \prod_{\alpha} \frac{dx^{\alpha}}{2\pi} \exp\Big(i \sum_{\alpha} x^{\alpha}(y^{\alpha} - f\zeta M^{\alpha}) - \\
&f(1-f)(\frac{1}{2} \sum_{\alpha} (x^{\alpha})^2 Q^{\alpha} + \sum_{\alpha<\beta} x^{\alpha} x^{\beta} q^{\alpha\beta}) \Big) \Big]
\end{aligned} \tag{4.31}$$

$$Z_2' = \frac{1}{N}\ln\Big[\int\prod_{j,\alpha}dw_j^\alpha\prod_\alpha\exp\Big(-E^\alpha\sum_j w_j^\alpha\rho_j+\hat{Q}^\alpha\sum_j(w_j^\alpha)^2-\hat{M}^\alpha\sum_j w_j^\alpha\Big)$$
$$\prod_{\alpha<\beta}\exp\Big(\hat{q}^{\alpha\beta}\sum_j w_j^\alpha w_j^\beta\Big)\Big] \tag{4.32}$$

$$Z_3' = -\sum_{\alpha<\beta}q^{\alpha\beta}\hat{q}^{\alpha\beta}-\sum_\alpha\Big(Q^\alpha\hat{Q}^\alpha-\frac{\theta}{f}\hat{M}^\alpha-\lambda E^\alpha\Big)+o(1) \tag{4.33}$$

and

$$H = \frac{1}{N}\ln\left[\int\prod_{\alpha,j}dw_j^\alpha\prod_\alpha\exp\left(-E^\alpha\sum_j w_j^\alpha\rho_j\right)\right]+\sum_\alpha E^\alpha \tag{4.34}$$

The expressions of the Z''s can be simplified under the assumption of a replica symmetric solution $q^{\alpha\beta}=q$, $Q^\alpha=Q$, $M^\alpha=M$, $E^\alpha=E$, $\hat{q}^{\alpha\beta}=\hat{q}$, $\hat{Q}^\alpha=\hat{Q}$, $\hat{M}^\alpha=\hat{M}$, and by using the following identities

$$e^{\frac{1}{2}qu^2}=\int Dte^{t\sqrt{q}u} \text{ with the notation } \int Dt=\int_{-\infty}^{+\infty}dt\frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} \tag{4.35}$$

and

$$\ln\int Dtf(t)^n \underset{n\ll 1}{\simeq} n\int Dt\ln f(t) \tag{4.36}$$

It gives

$$\begin{aligned} Z_1' &= \frac{n}{N}\sum_{j=1}^N\int Dt\ln\left[\int_0^{+\infty}dw\exp\left(-(\hat{M}-t\sqrt{\hat{q}}-E\rho_i)w-\frac{1}{2}(\hat{q}-2\hat{Q})\right)\right] \\ &= n\int d\rho P(\rho)\int Dt\ln\left[\int_0^{+\infty}dw\exp\left(-(\hat{M}-t\sqrt{\hat{q}}-E\rho)w-\frac{1}{2}(\hat{q}-2\hat{Q})\right)\right] \end{aligned} \tag{4.37}$$

with $P(\rho)$ the probability density function from which the elements of $\vec{\rho}$ are drawn.

$$Z_2' = n\int Dt\ln H\left(\frac{\kappa-f\zeta M+t\sqrt{f(1-f)q}}{\sqrt{f(1-f)(Q-q)}}\right) \tag{4.38}$$

Putting everything together into the expression of $\langle V^n\rangle$ we get equation (4.15).

The macroscopic quantities describing the network are the parameters that maximize the function F. Indeed, when applying the steepest descent method to the integral in (4.15) in the limit $N \to +\infty$, only the values that maximize F contribute to the final value of $\langle V^n \rangle$. These values are given by the following equations

$$\frac{\theta}{f} = \int d\rho P(\rho) \int Dt \frac{\int_0^{+\infty} dw w \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)}{\int_0^{+\infty} dw \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)} \tag{4.39}$$

$$Q = \int d\rho P(\rho) \int Dt \frac{\int_0^{+\infty} dw w^2 \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)}{\int_0^{+\infty} dw \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)} \tag{4.40}$$

$$q = \int d\rho P(\rho) \int Dt \frac{\int_0^{+\infty} dw (w^2 - w\frac{t}{\sqrt{\hat{q}}}) \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)}{\int_0^{+\infty} dw \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)} \tag{4.41}$$

$$\lambda = \int d\rho P(\rho) \rho \int Dt \frac{\int_0^{+\infty} dw w \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)}{\int_0^{+\infty} dw \exp\left(-\frac{1}{2}\left[\hat{q} - 2\hat{Q}\right] w^2 + (t\sqrt{\hat{q}} - \hat{M} - E\rho)w\right)} \tag{4.42}$$

$$0 = \sum_{\zeta=\pm 1} p_\zeta \zeta \int Dt \frac{G(a_\zeta(t))}{H(a_\zeta(t))} \tag{4.43}$$

$$\hat{Q} = \frac{\alpha}{2} \sum_{\zeta=\pm 1} p_\zeta \int Dt \frac{a_\zeta(t)}{(Q-q)} \frac{G(a_\zeta(t))}{H(a_\zeta(t))} \tag{4.44}$$

$$\hat{q} = \alpha \sum_{\zeta=\pm 1} p_\zeta \int Dt \left(\frac{t}{\sqrt{q(Q-q)}} + \frac{a_\zeta(t)}{Q-q} \right) \frac{G(a_\zeta(t))}{H(a_\zeta(t))} \tag{4.45}$$

$$a_\zeta(t) = \sqrt{\frac{q}{Q-q}}(t - \tau_\zeta) \tag{4.46}$$

$$\tau_\zeta = -\frac{\kappa - \zeta f M}{\sqrt{qf(1-f)}} \tag{4.47}$$

In the limit $q \to Q$, these can be simplified by introducing A, B, C, D such that

$$\begin{aligned} 2\hat{Q} \sim \hat{q} \sim \frac{C}{(Q-q)^2} \quad ; \quad \hat{q} - 2\hat{Q} \sim \frac{A}{Q-q} \\ \hat{M} \sim \frac{B\sqrt{C}}{Q-q} \quad ; \quad E \sim \frac{D\sqrt{C}}{Q-q} \end{aligned} \tag{4.48}$$

It gives

$$\frac{\theta}{f} = \frac{\sqrt{C}}{A}\int d\rho P(\rho)\left[G(B+D\rho)-(B+D\rho)H(B+D\rho)\right] \tag{4.49}$$

$$Q = \frac{C}{A^2}\int d\rho P(\rho)\left[\left(1+(B+D\rho)^2\right)H(B+D\rho)-(B+D\rho)G(B+D\rho)\right] \tag{4.50}$$

$$A = \int d\rho P(\rho) H(B+D\rho) \tag{4.51}$$

$$\lambda = \frac{\sqrt{C}}{A}\int d\rho P(\rho)\rho\left[G(B+D\rho)-(B+D\rho)H(B+D\rho)\right] \tag{4.52}$$

$$0 = \sum_{\zeta=\pm 1} p_\zeta \zeta (G(\tau_\zeta)-\tau_\zeta H(\tau_\zeta)) \tag{4.53}$$

$$C = \alpha_c Q \sum_{\zeta=\pm 1} p_\zeta \left((1+\tau_\zeta^2)H(\tau_\zeta)-\tau_\zeta G(\tau_\zeta)\right) \tag{4.54}$$

$$A = \alpha_c \sum_{\zeta=\pm 1} p_\zeta H(\tau_\zeta) \tag{4.55}$$

These equations can then be solved numerically to give the storage capacity of the network α_c and to extract the order parameters that allow to express the distributions of weights. Note that calculations for the distributions of weights is almost identical to the one we have just described.

Chapter 5

Discussion

We have studied models of ANNs describing cortical circuits at different scales. In chapter 2, we have considered fully connected networks that can be thought of as an approximation of local cortical networks at a scale of a few hundred microns (Hellwig, 2000; Kalisman et al., 2005). We have computed the storage capacity of networks with discrete states synapses in different learning scenarios, a first one where patterns are presented only once and a second one where patterns are presented multiple times. We have interpreted this second learning scenario as describing learning in local cortical circuits. Naturally we want to ask whether the properties of the connectivity of this model are consistent with known facts about connectivity of cortical networks. In our model synapses can be in a finite number of states. It has been argued that it would be difficult to build synapses that can take continuous values with known bio-physical processes (see e.g. Brunel 2003), but to the best of our knowledge there is no direct evidence that cortical synapses can be described by discrete variables. Nonetheless, for hippocampal circuits where it is easier to induce plasticity, studies using minimal stimulation have provided evidence that synapses could be described by binary variables (Petersen et al., 1998; O'Connor et al., 2005). In the mouse somato-sensory cortex, two neighboring cells touch each other on average at ~ 5 different locations Kalisman et al. (2005). We have modeled this feature by considering synapses made of 5 binary contacts that are independently modified during learning, we have seen that this feature could increase the storage capacity of local cortical networks, compare to models with a single binary synaptic contact. A measure of connectivity that has been well documented is the connection probability between pyramidal cells, which has consistently been found to be low, around 10%. In our cortical model with 5 independent binary synaptic contacts, parameters optimizing capacity lead to a connection probability around 80%. These parameters are the one saturating the storage capacity of the network, thus we expect that each memory is associ-

ated with a small basin of attraction. In the framework developed by Gardner, it has been shown that connection probability decreases when stored patterns are required to have larger basins of attraction (Brunel et al., 2004). It would be interesting to introduce such a requirement about the basins of attractions in our model, and see how storage capacity and connection probability are modified. Also, as we have seen, the learning rule we have studied is not optimal in terms of storage capacity (Gutfreund and Stein, 1990). For binary synapses and standard coding level ($f = \frac{1}{2}$) a network optimizing storage capacity has a connection probability $< 30\%$. It would be interesting to know how this value changes for discrete synapses with more than two states and sparser more realistic coding levels, and then see whether a biologically plausible learning rule could be found that reaches this capacity.

In the third chapter, we have studied memory properties of modular networks where modules are fully-connected networks connected to each others via diluted long-range connections. As in the previous studies of similar models (O'Kane and Treves, 1992; Mari and Treves, 1998) we have found that these networks are able to store a number of bits per synapse of order one, with a storage capacity that interpolates between fully-connected networks and diluted networks. We have studied two different kinds of models, a first one where patterns of activity of active modules are chosen randomly and another kind where patterns are organized in categories. For patterns organized in categories, the network has an information capacity of order one when the macroscopic coding level scales as $\frac{1}{M}$ -where M is the number of modules. In this case the number of patterns P stored in the network can then scale as $P \propto M \frac{N^2}{(\ln N)^2}$ -where N is the number of neurons in each module-, i.e. the network stores a number of patterns proportional to size of the network, as in Mari and Treves (1998).

If these networks are required to correct macroscopic errors (e.g. single modules in a local attractor not consistent with the whole pattern) some constraints on the sources of reverberating activity have to be satisfied. Namely, the component of the reverberating activity supported by long-range connections has to be larger than the local component. In terms of network connectivity this means that a given neuron, belonging to a module m active in a pattern $\vec{\Xi}^{\mu_0}$, receives less connections from neurons in m than from neurons belonging to other modules active in $\vec{\Xi}^{\mu_0}$.

What kind of cortical circuits can be well modeled by these networks ? It is tempting to interpret modules as the patches identified by connectivity studies like the one of Pucak et al. (1996). In this case, modeling a patch by a fully connected network might not be ideal as these patches are on average $1.7mm \times 0.25mm$,

which is larger than the size of the local networks described in chapter 2, at most $0.5mm \times 0.5mm$. However these last dimensions have been inferred from studies in sensory cortices of rodents, it might be that similar studies in pre-frontal cortices of monkeys would lead to larger dimensions for networks that can reasonably be approximated by fully-connected models. In our models, the patterns of activity also have a modular structure. With an fMRI study on humans, Huth et al. (2012) have shown how visually perceived objects are represented on the cortical surface. Small 'patches' of sizes of the order of the millimeter (close to the spatial resolution of the fMRI machine) are activated when a given object is presented. One could speculate that these 'patches' of activity coincide with the patches of connectivity mentioned above. It would not be that surprising, for instance it has been shown in visual cortex of tree shrew that interconnected patches have similar preferred orientations (Bosking et al., 1997). In order to see whether these networks support ANN dynamics, it would be interesting to record the activity of a large piece of cortex with a spatial resolution of the order of the millimeter, while a subject is performing a working memory task, to see if there is some sort of distributed persistent activity.

If a network of interconnected patches of cortex operates as an ANN and is able to correct for macroscopic errors, our study predicts that injection of retrograde tracers in a single neuron (Rancz et al., 2011) would label less neurons in the local patch to which the neuron belongs than in other patches representing similar items. For instance this could be done in the patches of cortex representing faces that have been well identified (Tsao et al., 2003). Note that an analysis of Stepanyants et al. (2009) shows that in the visual cortex of mice, 36% of the connections to a neuron originate inside a cylinder of radius 1 mm, the other inputs being mediated by long-range connections. This is consistent with the constraint we have established on the minimal amount of long-range connections, although the origin of these connections would need to be known to be sure the constraint is satisfied.

Recently, quantitative data about the connectivity between cortical areas have been published (Markov et al., 2011). Could this connectivity sustain ANN dynamics at the scale of the whole cortex ? To study this question, models like the one studied in chapter 3 seem inappropriate since they do not incorporate the heterogeneity in the distances between modules which is present between cortical areas, while Markov et al. (2013) have shown that distance strongly modulates connectivity between areas.

Experimental measures of connectivity in the cerebellum or local cortical networks are consistent with perceptron or attractor network models working at op-

timal storage capacity (Brunel et al., 2004; Chapeton et al., 2012). Guided by a similar idea, we have studied the connectivity of a minimal model of the network of cortical areas where the activity of each area is described by a binary unit and where the amplitude of the connection between two areas has a cost that depends on distance. Using Gardner's approach, we have computed statistical properties of the connectivity matrix of such networks. In the framework of our simple model, statistical properties of the experimental connectivity matrix could not be well fitted by statistical properties of a network working at an optimal capacity. However, we have found some parameters for a network below maximal capacity that fit the experimental measurements rather well. For these parameters, the model can sustain 10 patterns via reverberating activity through connections between cortical areas. Even though the model shares similar statistics of connectivity with the real network, it is not granted the real network can sustain patterns of activity. Indeed, connection weights drawn randomly from the theoretical distribution would not necessarily lead to attractor states. It would be interesting to implement the detailed measured connectivity matrix in a network model and study the fixed points of this network. A first difficulty arises from the fact that this connectivity matrix is only partially known (about 30% of the full matrix has been characterized). Instead, one could focus on the 29×29 sub-network of cortical areas whose connectivity is fully known, and identify which fixed points could be implemented in this network. It would also be interesting to compare measures of functional connectivity established from resting state activity and the anatomical connectivity in the 29×29 sub-network, and maybe use functional connectivity to guess the anatomical connections that have not yet been explored.

Bibliography

Abeles, M. (1991). *Corticonics*. New York: Cambridge University Press.

Alvarez, P. and Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences*, 91(15):7041Ð7045.

Amaldi, E. (1991). On the complexity of training perceptrons. *Artificial Neural Networks*, 1:55–60.

Amit, D. and Treves, A. (1989). Associative memory neural network with low temporal spiking rates. *Proc. Natl. Acad. Sci. U.S.A.*, 86:7871–7875.

Amit, D. J. (1989). *Modeling brain function*. Cambridge University Press.

Amit, D. J. and Brunel, N. (1997a). Dynamics of a recurrent network of spiking neurons before and following learning. *Network*, 8:373–404.

Amit, D. J. and Brunel, N. (1997b). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7:237–252.

Amit, D. J. and Fusi, S. (1994). Dynamic learning in neural networks with material sysnapses. *Neural Computation*, 6:957–982.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1531.

Amit, D. J. and Tsodyks, M. V. (1991). Quantitative study of attractor neural network retrieving at low spike rates II: Low-rate retrieval in symmetric networks. *Network*, 2:275.

Amit, Y. and Huang, Y. (2010). Precise capacity analysis in binary networks with multiple coding level inputs. *Neural Comput*, 22:660–688.

Baldassi, C., Braunstein, A., Brunel, N., and Zecchina, R. (2007). Efficient supervised learning in networks with binary synapses. *Proc. Natl. Acad. Sci. U.S.A.*, 104:11079–11084.

Barbour, B., Brunel, N., Hakim, V., and Nadal, J. (2007). What can we learn from synaptic weight distributions? *Trends Neurosci.*, 30:622–629.

Barrett, A. B. and van Rossum, M. C. (2008). Optimal learning rules for discrete synapses. *PLoS Comput. Biol.*, 4:e1000230.

Bathellier, B., Ushakova, L., and Rumpel, S. (2012). Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron*, 76(2):435–449.

Battaglia, F. P. and Treves, A. (1998). Attractor neural networks storing multiple space representations: A model for hippocampal place fields. *Phys. Rev. E*, 58:7738–7753.

Batuev, A., Pirogov, A., and Orlov, A. (1979). Unit activity of the prefrontal cortex during delayed alternation performance in monkey. *Acta physiologica Academiae Scientiarum Hungaricae*, 53(3):345.

Bienenstock, E., Cooper, L., and Munro, P. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2:32–48.

Booth, V. and Rinzel, J. (1995). A minimal, compartmental model for a dendritic origin of bistability of motoneuron firing patterns. *J Comput Neurosci*, 2:299–312.

Bosking, W., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J Neurosci.*, 17:2112–2127.

Braitenberg, V. and Schütz, A. (1991). *Anatomy of the cortex*. Springer-Verlag.

Brunel, N. (1994). Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory. *Journal of Physics A: Mathematical and General*, 27(14):4783.

Brunel, N. (2000a). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.*, 8:183–208.

Brunel, N. (2000b). Persistent activity and the single cell f-I curve in a cortical network model. *Network*, 11:261–280.

Brunel, N. (2003). Network models of memory. In *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School*, pages 407–476.

Brunel, N., Carusi, F., and Fusi, S. (1998). Slow stochastic Hebbian learning of classes in recurrent neural networks. *Network*, 9:123–152.

Brunel, N., Hakim, V., Isope, P., Nadal, J. P., and Barbour, B. (2004). Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron*, 43:745–57.

Chapeton, J., Fares, T., LaSota, D., and Stepanyants, A. (2012). Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. *Proceedings of the National Academy of Sciences*, 109(51):E3614–E3622.

Coolen, A. C. C. and Sherrington, D. (1993). Dynamics of fully connected attractor neural networks near saturation. *Physical review letters*, 71(23):3886Ð3889.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans*, EC-14:326.

da Fonseca, C. M. (2007). The characteristic polynomial of some perturbed tridiagonal k-toeplitz matrices. *Applied Mathematica Sciences*, 1:59–67.

DeFelipe, J., Conley, M., and Jones, E. G. (1986). Long-range focal collateralization of axons arising from corticocortical cells in monkey sensory-motor cortex. *J Neurosci.*, 6:3749–3766.

Delord, B., Klaassen, A. J., Burnod, Y., Costalat, R., and Guigon, E. (1997). Bistable behaviour in a neocortical neurone model. *Neuroreport*, 8:1019–23.

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

Diesman, M., Gewaltig, M., and Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402:529–533.

Dipoppa, M. and Gutkin, B. (2013). Flexible frequency control of cortical oscillations enables computations required for working memory. *Proc.Natl.Acad.Sci.USA*, 110:12828–12833.

Ercsey-Ravasz, M., Markov, N., Lamy, C., VanÊEssen, D., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). A predictive network model of cerebral cortical connectivity based on a distance rule. *Neuron*, 80(1):184–197.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1(1):1–47.

Fusi, S. and Abbott, L. F. (2007). Limits on the memory storage capacity of bounded synapses. *Nat. Neurosci.*, 10:485–493.

Fusi, S., Drew, P. J., and Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, 45:599–611.

Fuster, J. M. and Alexander, G. (1971). Neuron activity related to short-term memory. *Science*, 173:652–654.

Ganguli, S., Huh, D., and Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.*, 105:18970–18975.

Gardner, E. and Derrida, B. (1988). Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271.

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

Garey, M. R. and Johnson, D. S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Gerstner, W. and van Hemmen, J. L. (1992). Associative memory in a network of 'spiking' neurons. *Network*, 3:139–164.

Gilbert, C. D. and Wiesel, T. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci.*, 9:2432–2442.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, 61(4):621–634.

Gutfreund, H. and Stein, Y. (1990). Capacity of neural networks with discrete synaptic couplings. *J. Phys. A: Math. Gen.*, 23:2613–2630.

Harvey, C. D., Coen, P., and Tank, D. W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–68.

Hebb, D. O. (1949). *Organization of behavior*. New York: Wiley.

Hellwig, B. (2000). A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological cybernetics*, 82(2):111Ð121.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City.

Hertz, J. and Prügel-Bennett, A. (1996). Learning synfire chains: turning noise into signal. *International journal of neural systems*, 7(04):445–450.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conductance and excitation in nerve. *J. Physiol.*, 117:500–544.

Holmgren, C., Harkany, T., Svennenfors, B., and Zilberter, Y. (2003). Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J. Physiol.*, 551:139–153.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, 79:2554–2558.

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

Huang, Y. and Amit, Y. (2011). Capacity analysis in multi-state synaptic models: a retrieval probability perspective. *J Comput Neurosci*, 30:699–720.

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Izhikevich, E., Gally, J., and Edelman, G. (2004). Spike-timing dynamics of neuronal groups. *Cereb. Cortex*, 14:933–944.

Johansson, C. and Lansner, A. (2007). Imposing biological constraints onto an abstract neocortical attractor network model. *Neural Comput.*, 19(7):1871–1896.

Káli, S. and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature neuroscience*, 7(3):286–294.

Kalisman, N., Silberberg, G., and Markram, H. (2005). The neocortical microcircuit as a tabula rasa. *Proc Natl Acad Sci U S A*, 102(3):880–885.

Kirkwood, A. and Bear, M. (1994). Hebbian synapses in visual cortex. *J Neurosci.*, 14:1634–1645.

Knoblauch, A., Palm, G., and Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289Ð341.

Kouachi, S. (2006). Eigenvalues and eigenvectors of tridiagonal matrices. *Electronic Journal of Linear Algebra*, 15:115–133.

Krauth, W. and Mezard, M. (1989). Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066.

Kropff, E. and Treves, A. (2005). The storage capacity of potts models for semantic memory retrieval. *J. Stat. Mech.*, 8:P08010.

Leibold, C. and Kempter, R. (2008). Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. *Cereb. Cortex*, 18:67–77.

Levy, N., Horn, D., Meilijson, I., and Ruppin, E. (2001). Distributed synchrony in a cell assembly of spiking neurons. *Neural Networks*, 14(6):815–824.

Lim, S. and Goldman, M. S. (2012). Noise tolerance of attractor and feedforward memory models. *Neural computation*, 24(2):332–390.

Lim, S. and Goldman, M. S. (2013). Balanced cortical microcircuitry for maintaining information in working memory. *Nature neuroscience*.

Lisman, J. E. and Idiart, M. (1995). Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science*, 267(5203):1512–1515.

Liu, J. K. and Buonomano, D. V. (2009). Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner. *Journal of Neuroscience*, 29(42):13172–13181.

Loewenstein, Y. and Sompolinsky, H. (2003). Temporal integration by calcium dynamics in a model neuron. *Nature neuroscience*, 6(9):961–967.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput*, 14(11):2531–2560.

Machens, C. K., Romo, R., and Brody, C. D. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science*, 307(5712):1121–1124.

Major, G. and Tank, D. (2004). Persistent neural activity: prevalence and mechanisms. *Current opinion in neurobiology*, 14(6):675–684.

Mari, C. F. (2004). Extremely dilute modular neuronal networks: Neocortical memory retrieval dynamics. *Journal of Computational Neuroscience*, 17(1):57–79.

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

Markov, N., Misery, P., Falchier, A., Lamy, C., Vezoli, J., Quilodran, R., Gariel, M., Giroud, P., Ercsey-Ravasz, M., Pilaz, L., et al. (2011). Weight consistency specifies regularities of macaque cortical networks. *Cerebral Cortex*, 21(6):1254.

Markov, N. T., Ercsey-Ravasz, M., Lamy, C., Gomes, A. R. R., Magrou, L., Misery, P., Giroud, P., Barone, P., Dehay, C., Toroczkai, Z., et al. (2013). The role of long-range connections on the specificity of the macaque interareal cortical network. *Proceedings of the National Academy of Sciences*, 110(13):5187–5192.

Markov, N. T., Ercsey-Ravasz, M. M., Ribeiro Gomes, A. R., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M. A., Sallet, J., Gamanut, R., Huissoud, C., Clavagnier, S., Giroud, P., Sappey-Marinier, D., Barone, P., Dehay, C., Toroczkai, Z., Knoblauch, K., Van Essen, D. C., and Kennedy, H. (2012). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex*, 24(1):17–36.

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and beyond*. World Scientific: Singapore.

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

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, 319:1543.

Nadal, J.-P. (1991). Associative memory: on the (puzzling) sparse coding limit. *J. Phys. A: Math. Gen.*, 24:1093–1101.

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

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

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

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Experimental Brain Research*, 34:171–175.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press (Oxford).

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

Petersen, C. C., Malenka, R. C., Nicoll, R. A., and Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proc.Natl.Acad.Sci.USA*, 95:4732–4737.

Pucak, M. L., Levitt, J. B., Lund, J. S., and Lewis, D. A. (1996). Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *J. Comp. Neurol.*, 338:360–376.

Rancz, E. A., Franks, K. M., Schwarz, M. K., Pichler, B., Schaefer, A. T., and Magrie, T. W. (2011). Transfection via whole-cell recording in vivo: bridging single-cell physiology, genetics and connectomics. *Nature Neuroscience*, 14:527–532.

Renart, A., Brunel, N., and Wang, X.-J. (2003). *Mean-field theory of recurrent cortical networks: from irregularly spiking neurons to working memory*, chapter 15, pages 431–490. CRC Press, Boca Raton.

Renart, A., Moreno-Bote, R., Wang, X.-J., and Parga, N. (2007). Mean-driven and fluctuation-driven persistent activity in recurrent networks. *Neural Comput*, 19:1–46.

Roudi, Y. and Latham, P. (2007). A balanced memory network. *PLoS Comput. Biol.*, 3:1679–1700.

Roxin, A. and Fusi, S. (2013). Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Computational Biology*, 9(7):e1003146.

Samsonovich, A. and McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, 17:5900–5920.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *J. Math. Biol.*, 4:303–.

Seung, H. S. (1996). How the brain keeps the eyes still. *Proc Natl Acad Sci USA*, 93:13339–13344.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32:1149–1164.

Sompolinsky, H. (1986). Neural networks with nonlinear synapses and a static noise. *Phys. Rev. A*, 34:2571–2574.

Song, S., Sjostrom, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol*, 3:e68.

Stepanyants, A., Martinez, L. M., Ferecsko, A. S., and Kisvarday, Z. F. (2009). The fractions of short-and long-range connections in the visual cortex. *Proceedings of the National Academy of Sciences*, 106(9):3555Ð3560.

Tkavcik, G., Prentice, J. S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32):14419Ð14424.

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

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995.

Tsodyks, M. (1990). Associative memory in neural networks with binary synapses. *Mod. Phys. Lett. B*, 4:713–.

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

van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724–1726.

van Vreeswijk, C. and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural Computation*, 10:1321–1371.

van Vreeswijk, C. and Sompolinsky, H. (2005). Irregular activity in large networks of neurons. In Chow, C., Gutkin, B., Hansel, D., Meunier, C., and Dalibard, J., editors, *Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School 2003*. Elsevier.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.*, 24:455–463.

White, O., Lee, D., and Sompolinsky, H. (2004). Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14).

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

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12:1–24.

Wilson, M. and McNaughton, B. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261:1055–1058.