THE UNIVERSITY OF CHICAGO

UNSUPERVISED LEARNING OF NEURONAL REPRESENTATIONS IN BRAIN NETWORKS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
ULISES PEREIRA OBILINOVIC

CHICAGO, ILLINOIS
DECEMBER 2018

To Macarena and Sara

" ... unless it comes out of

your soul like a rocket,

unless being still would

drive you to madness or

suicide or murder,

don't do it.

unless the sun inside you is

burning your gut,

don't do it. "

-Charles Bukowski

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I first, want to thank the Department of Statistics at the University of Chicago for its vision and commitment to train in statistics and applied mathematics to graduate students with strong interest in interdisciplinary research. I want to thank Mihai Anitescu, Lek-Heng Lim and Rina Foygel Barber for teaching me with their example. To my friends Kushal Dey, Wooseok Ha, Vivak Patel and Ken Wong for our scientific discussions and friendship. To Laura Rigazzi for her crucial help in all the administrative hurdles. To my collaborators and colleagues Yonatan Aljadeff, Maxwell Gillett and Krithika Mohan. I have learned a lot by working with them. To David Freedman for being part of my committee, and introducing me to the wonders of cognitive neuroscience with his papers. I want to especially thank Yali Amit for being part of my committee, and for the last year and a half, my secondary advisor. For introducing me to the beautiful world of machine intelligence. For always believing in me, listening to me and supporting me. I really admire his intelligence, pragmatism, and attitude toward research. For teaching me to not content myself with incomplete explanations, and for pushing me to ask myself hard questions. I want to thank Nicolas Brunel, my advisor. Characters are not enough to express how impactful his mentorship, science, and vision of modern neuroscience has been to me. He has deeply influenced the course of my career by believing in my potential to do interesting science. He made a bet, and I can see today it was a long shot. However, he was always committed to support me, even when it was not clear I will succeed. He has been a superb mentor for me. This thesis would not be possible without him. We had memorable times doing this work together. He let me be independent to think, explore, fail and succeed without obstacles. But he guided me in every step. He had been instrumental in building confidence in my talents and passion for brain and intelligent systems. I want to thank my *nono* Yerko Obilinovic Marinovic who encouraged my curiosity about the world, and imprinted in me his atheistic interpretation of the biblical phrase 'Man does not live by bread alone'. I want to thank my mother Silvia Obilinovic Carvajal for

her deep and constant love and support. Lastly, I want to thank my family: to Neftali for his unconditional emotional support as a dog, and to my wife Macarena Galvez Barrera for being my companion and love in the *voragine* of life.

ABSTRACT

In this thesis, I show that a single class of unsupervised learning rules that can be inferred from *in vivo* data learns neuronal representations consistent with a wide range of datasets. Recurrent neuronal networks endowed with learning rules of this class represent memories as qualitatively different spatiotemporal attractors (i.e. fixed-point attractors, chaotic attractors or transient sequences of activity) depending on the stimuli statistics and learning rule. They match disparate observations from recordings in different species, brain regions and memory tasks, suggesting that memories are differentially represented in brain systems. This thesis provides a unified model for explaining the diversity in neuronal dynamics during memory retrieval.

CHAPTER 1

INTRODUCTION

1.1 Neuronal representations

Sensory experiences produce brain-wide activity changes. For example, exposure to narrative stories produce a semantic structure of activations across multiple cortical regions in humans (Huth et al. 2016). Exposure to natural images produce a hierarchy of activations in inferior temporal cortex in primates (Kiani et al. 2007, Kriegeskorte et al. 2008). Hippocampal neurons of rodents moving in an environment are activated in specific landmarks (O'Keefe & Dostrovsky 1971, O'keefe & Nadel 1978), while neurons in the entorhinal cortex get activated whenever the animal position coincides with intersection points on a grid that maps the environment (Hafting et al. 2005). These are just a few examples of the striking ability of brains to respond in a structured way to sensory stimulus. A natural question that comes to mind is if the neuronal responses are in anyway related with the *information* content of the sensory stimuli. A popular hypothesis underlying most of contemporary neuroscience is that salient information about the external world is *represented* in neuronal activity. In other words, activity patterns in networks in the brain encode information of the external world which can be then used for cognitive computations. A useful analogy for the non-expert reader of this hypothesis is the pointillism, in this painting technique popularized by the impressionists, complex scenes are painted using dots of different colors in the canvas as the one presented in Fig 1.1. Analogously to the distributed but structured ensemble of color dots in Fig 1.1, it is hypothesized that information in the brain is represented by distributed yet concerted single neuron activity in brain networks. This hypothesis is strongly supported by decades of neuronal recordings in brains of primates, rodents, cats, birds, fishes and flies during behavior. How neuronal representations are learned from experience and stored in brains as *memories*? This is the matter of this thesis.

Figure 1.1: A Sunday afternoon on the island of la Grande Jatte (207.6cm × 308cm). Painted by Georges Seurat on 1884-1886. Currently exhibited at the Art Institute of Chicago. This painting is a landmark example of the pointillist technique, here small dots of paint with distinct color are applied in a very large canvas to represent a complex scene. In this painting Seurat represents a typical Sunday afternoon in the Seine riverbank. Analogous to the pointillism, the theory of neuronal representations hypothesize that information in the brain is represented in the concerted activity of ensembles of neurons. As the thousands small color dots in the canvas that distributed yet coordinated lead to the Sunday afternoon scene in this painting, in the theory collective neuronal activity distributed in brain networks represents external (and internal) information.

1.2 Learning and memory of neuronal representations

Past experiences can be recalled on the basis of cognitive needs by brain systems. It has been hypothesized that when past experiences are recalled, neuronal representations are reactivated, conveying their past information. The reactivated neuronal representation corresponds to the neuronal substrate of a *remembered* memory. In other words, when a memory is remembered, patterns of neuronal activity correlated to the ones elicited by the memorandum that is being retrieved are reactivated. A large body of experimental data supports this hypothesis, one compelling example is the activity observed in CA3 hippocampal cells when rats change environments. Activity in CA3 is highly informative of the identity of the particular environment when rats are placed in it, even when the environment gets distorted or the geometry of two different environments is identical. This suggests that neuronal representations of environments in CA3 can be recalled on the basis of limited information about the environment (Leutgeb et al. 2004). In humans, semantic memories reactivate in or nearby areas corresponding to the sensory modalities involved in the recalled concept. It is believed that the reactivation of brain regions corresponding to different sensory modalities embodies semantic memories in the process of recalling (Binder & Desai 2011). In primates, it has been shown that learned categories of objects are represented in the neuronal activity of the prefrontal cortex. Stimuli varying its geometry within a given category elicit similar neuronal activity, suggesting that learned neuronal representations corresponding to categories are recalled from different but correlated stimuli (Freedman et al. 2001). These are just three handpicked examples from a large set of experimental work showing that in different brains and brain regions during retrieval, neuronal representations are reactivated for its use on cognitive demands.

How memories are learned from experience in brain networks? One of the candidate scenarios was first envisioned by Richard Semon (Semon 1909) and specifically refined for neuronal circuits by Donald Hebb (Hebb 1949). In this scenario, experienced items to be

memorized elicit patterns of activity in brain networks. These patterns of neuronal activity produce changes in the synaptic connectivity via activity-dependent synaptic plasticity, i.e. the cellular mechanism by which synaptic connections between neurons change depending on their activity, generating a distributed pattern of synaptic modifications. Therefore, the connectivity matrix gets structured according to the interplay between synaptic plasticity and the spatiotemporal statistics of neuronal activity patterns. The induced traces of synaptic modifications correspond to the *synaptic memory engram* of the stored memorandum. Analog to the painting shown in Fig 1.2 generated by layers of barrages of strokes in a canvas, in this theory, when a new item is memorized synaptic modifications change the connectivity matrix again *in top* of previous modifications. The overfall of sensory experiences leads to an *online* process in which the connectivity is modified continuously for learning new memories. Synaptic changes create a memory of the corresponding memorandum by fostering the corresponding neuronal representation in the network dynamics. After learning, upon a sensory cue correlated with the stored memorandum, the corresponding memory is retrieved by the activation of its neuronal representation. When no memory is retrieved, neuronal representations of memories are latent, existing only as synaptic memory engrams in the network connectivity.

In this thesis, the scenario described above is assumed as the working hypothesis. However, this *synaptocentric* scenario for learning and memory is currently a matter of scientific debate. An interesting complementary scenario is that both synapses and intrinsic single cell properties change during learning, corresponding in conjunction to a memory engram (Titley et al. 2017).

Figure 1.2: Convergence (237 cm × 390 cm). Painted by Jackson Pollock on 1952. Currently at the Albright-Knox Art Gallery. In this painting, Pollock paints a barrage of strokes of different colors in a canvas one after another in layers. The painting serves as a metaphor for the hypothetical mechanism of learning in brain networks described in section 1.2. As a set of the strokes of a particular color, synaptic modifications due to a given memorandum modify the network connectivity matrix. When a second memorandum is learned, the connectivity gets again modified adding new changes *in top* of the previous ones, as the second set of strokes of different colors in the painting. When more memories are learned, this process continues structuring the connectivity of brain networks as Pollock's strokes generate the final version of Convergence. Then memory engrams get intermingle in the connectivity matrix, and its information is distributed across the entire network.

1.3 Neurobiology of learning and memory

1.3.1 Synaptic plasticity

There is strong evidence supporting that the main cellular mechanism for learning new memories in brains is synaptic plasticity in the form of long term potentiation (LTP) and depression (LTD). LTP has been first studied in the hippocampal excitatory synapses, starting from the observation that a short high-frequency stimulation produced a long-lasting increase in synaptic strengths (Lomo 1966, Bliss & Lømo 1973). For almost 50 years scientists have dissected the molecular and cellular mechanisms involving LTP. The basic version of the mechanism is the following: glutamate, which is a neurotransmitter that is released from synaptic vesicles by a pre synaptic neuron, binds to both α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid and N-Methyl-D-aspartic acid post synaptic receptors (AMPAR and NMADR respectively). By binding to AMPAR, the influx of ions depolarizes the membrane potential of the post synaptic neuron, which might contribute to the production of a post synaptic action potential. On the other hand, by binding to NMDAR an influx of calcium is produced into the post synaptic neuron (specifically the dendrite). The calcium activates a massive holeoenzyme called CaMKII (Ca^{2+}/calmodulin-dependent protein kinase II), starting a complex cascade of phosphorylations ending in the increase of AMPAR, increasing with this the strength of the synapse (see Lisman et al. (2012), Herring & Nicoll (2016) for a review of the molecular and cellular mechanisms involved). The description above is an extremely simplified version of a complex phenomenon, which its properties, as well as its molecular and cellular details, vary across species, brain regions, cell types and developmental stages. However, a consistent finding is that synapses undergo long time changes in a post and pre synaptic activity-dependent fashion.

In brain slices experiments, it has been shown that synaptic plasticity depends on the timing between the pre and post synaptic spikes, not only leads to potentiation but also might

lead to depression, and also depends on firing rates and membrane potentials (Markram, Lübke, Frotscher & Sakmann 1997, Magee & Johnston 1997, Bell et al. 1997, Bi & Poo 1998, Sjöström et al. 2001, Abbott & Nelson 2000, Artola et al. 1990). Theoretical models have captured these observations with different degree of detail and biological realism (Kempter et al. 1999, Pfister & Gerstner 2006, Clopath & Gerstner 2010, Gjorgjieva et al. 2011, Graupner & Brunel 2012). However, it is unclear that the observations in experiments performed in brain slices (i.e. *ex vivo*) hold in alive animals in behaving conditions. In a recent work, researches have taken an alternative approach for capturing the activity dependence of synaptic changes during behavior, developing a statistical method for inferring learning rules from *in vivo* data (Lim et al. 2015). The inferred learning rules belong to a class in which the changes in synaptic strength (ΔJ_{ij}) depend as the product of two non-linear functions of the pre (r_i) and post (r_j) synaptic firing rate, i.e.:

$$\Delta J_{ij} \propto f(r_i)g(r_j). \tag{1.1}$$

In this thesis, this family of learning rules is explored. It is further assumed that f and g are non-decreasing.

1.3.2 Biological implementation of three classes of learning

Models of learning involving synaptic plasticity can be divided into three classes: supervised, reinforced and unsupervised. In models of the former class, synapses are updated according to the activity of the network and a *error signal* that carries information about the difference between the current network dynamics and the one that it is expected to learn by the network. This class of learning is one of the leading theories for learning in the Cerebellum, and has provided a normative explanation for the anatomical and synaptic organization of the cerebellar circuitry (Marr 1969, Albus 1971, Brunel et al. 2004, Bouvier et al. 2017). In models based of reinforcement learning, a *reward signal* guides learning towards what

the network is expected to learn. Reinforcement learning has been a successful theoretical framework to understand dopamine-mediated learning. A large body of data suggests that dopamine release neurons encode a reward prediction error, driving synaptic changes in cortical and sub-cortical regions (Glimcher 2011). Lastly, in the latter class of models learning occurs solely driven by external inputs. Synapses changes depending on external stimulation without an error signal. External inputs shape the network connectivity, sculpting the connections between neurons depending on the statistics of their neuronal responses. These models have reproduced key aspects data involving learning and retrieval in cortical areas, in particular, the prefrontal cortex (Amit 1995, Wang 2001, Brunel 2005) and the Hippocampus (Treves & Rolls 1992). The family of learning rules in Eq. (1.1), considered in this thesis, corresponds to this class of learning since no error signal is available for driving the synaptic changes.

1.3.3 Neuronal representations of memories

Persistent activity

How are memories represented in neuronal activity? Delay response tasks in primates have provided important experimental evidence regarding this question. In early (visual) versions of this class of experiments (Fuster et al. 1971, Fuster & Jervey 1981, Miyashita 1988), an image is presented in a screen to a monkey for a short period of time. After the presentation period, the image is withdrawn from the monkey's view for a delay period of the order of seconds. After the delay period, the monkey uses information about the image to perform a task. For example, deciding whether a second presented image match the previous one. This task is designed in such a way that for its successful performance information about the image has to be held in memory. Strikingly, persistent activity has been recorded during delay periods, i.e. stable elevated activity, in the prefrontal cortex (Fuster et al. 1971, Funahashi et al. 1989, Romo et al. 1999), parietal cortex (Koch & Fuster 1989*b*), inferior temporal

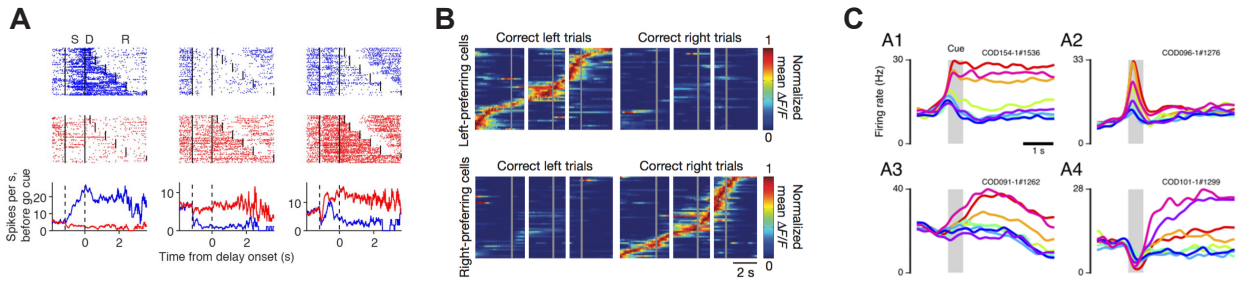

Figure 1.3: Persistent, sequential and heterogeneous delay activity. (A) Persistent activity for three representative neurons recorded in the mice anterior lateral motor cortex during an auditory delay response task. In this task a 3 or 12 kHz tone is presented to a mouse. After a delay period, the mouse has to leak a leak-port to the left or right depending on the frequency of the presented tone. For each trial, the duration of the delay period is randomly distributed according to an exponential distribution. The first and second rows show the spike raster plots corresponding to correct right (blue) and left (red) trials sorted by the delay period duration (ten trials per delay period duration). The last row shows the mean spike rate for right and left trials. Adapted from Inagaki et al. (2017). (B) Choice specific sequential activity recorded in the posterior parietal cortex during a navigational delay response task. In this task, a mouse navigates in a virtual reality maze while neuronal activity is being recorded using calcium imaging. A color in the landmark cues the mouse to turn right or left. After a delay period, the mouse has to turn according to the cue. The left and right columns in the figure correspond to correct left and right trials. The top and bottom rows correspond to the left- and right- preferring neurons respectively. Each row in the panels corresponds to the time-course of the normalized fluorescence for a single neuron. Adapted from Harvey et al. (2012). (C) Four representative neurons recorded in the prefrontal cortex of a monkey performing an oculomotor delay response task. In this task, a cue is presented in one of eight radial directions separated by 45°. After a delay period, the monkey has to saccade in the direction of the cue. Different panels correspond to different neurons, and different colors correspond to different directions (i.e. 0°, 45°,..., etc). Adapted from Murray et al. (2017), data from Funahashi et al. (1989) and Constantinidis et al. (2001).

cortex (Fuster & Jervey 1981, Miyashita 1988, Nakamura & Kubota 1995*a*) and other areas of the temporal lobe (Nakamura & Kubota 1995*a*). Recently, persistent activity has been observed during delay response tasks in rodents (Liu et al. 2014, Guo et al. 2014, Inagaki et al. 2017). It has been proposed that persistent activity constitute the neuronal correlate of memory (Goldman-Rakic 1995). That is, during retrieval, the neuronal representation of a memory corresponds to a stable pattern of activity in brain circuits. For an example of persistent activity see Fig. 1.3A.

Sequential activity

A qualitatively different type of neuronal dynamics, namely sequential activity, has also been observed relatively recently during delay response tasks. In this activity, neurons are active transiently for short periods of time in a sequence. An example of this activity has been observed during a navigational working memory task in the posterior parietal cortex of mice (Harvey et al. 2012). In this task, the mouse navigates in a virtual reality maze. During the presentation period, a color cue is presented. After a delay period, the mouse has to turn left or right depending on the cue. Interestingly, a choice specific set of neurons present sequential activity, see Fig. 1.3B. In recordings in the CA1 region of the hippocampus, choice specific sequences have also been observed when a rat runs in a wheel during delay periods (Pastalkova et al. 2008). Additionally, sequences also have been observed in tasks involving spatial navigation (Foster & Wilson 2006, Grosmark & Buzsáki 2016) and birdsong generation (Hahnloser et al. 2002, Amador et al. 2013, Okubo et al. 2015). It has been hypothesized that sequential activity also corresponds to a neuronal representation of memories. In this scenario, the information about the memorandum is hold in memory in the network's transient activity.

Heterogeneous activity

Heterogeneous time-varying fluctuations also have been observed during memory tasks in the prefrontal cortex. It has been reported during delay periods variability across trials for a single memorandum, strong temporal fluctuations and heterogeneity in the neuronal responses across neurons (Shafi et al. 2007, Lundqvist et al. 2016, Murray et al. 2017, Lundqvist et al. 2018) (see Fig. 1.3C for an example of heterogeneity in neuronal responses). It has been proposed that the observed heterogeneous activity corresponds to a qualitatively different neuronal representation of the retrieved memory from persistent or sequential activity (Murray et al. 2017, Druckmann & Chklovskii 2012).

1.4 Theoretical models for learning neuronal representations

Spatiotemporal dynamics of a neuronal network and its connectivity are deeply interlinked. Depending on their connectivity, neuronal networks have a plethora of qualitatively different types of dynamics as for example: fixed-point attractors (Hopfield 1982, Amit 1992), line attractors (Seung 1996), high dimensional attractors (Druckmann & Chklovskii 2012), chaotic attractors (Sompolinsky et al. 1988), sequential activity (Abeles 1991) and oscillations (Wilson & Cowan 1972). As discussed in section 1.2, in this thesis the underlying hypothesis is that neuronal representations of memories correspond to spatiotemporal patterns of activity. Attractor networks (Amit 1992) are one of the first theoretical instantiation of this idea. These network models have multiple stationary patterns of neuronal activity, i.e. fixed-point attractors. Each fixed-point attractor is correlated with a single memory, which corresponds to its neuronal representation. When a partial version of the stored memory is presented to the network, the state of the network goes to a region in the phase space where any point in this region evolves toward the fixed-point corresponding to the retrieved memory. For learning new memories, modifications of the connectivity according a particular synaptic plasticity rule creates a new fixed-point attractor representing the new memory.

The Hopfield model (Hopfield 1982) is the landmark model for attractor networks. In this model neurons are binary, and memories correspond to binary patterns learned using the covariance rule (Sejnowski 1977). Similar models to attractor networks have been proposed for learning sequences. In these models patterns of neuronal activity concatenated in time correspond to the neuronal representation of a memory. When a memory is retrieved, a partial version of the initial pattern in the sequence ignites the entire sequence of activity. New memories are learned in these models using an asymmetric version of the covariance rule (Sompolinsky & Kanter 1986, Kleinfeld 1986). Lastly, Tirozzi & Tsodyks have shown that chaotic attractors with associative memory properties are present for strong synapses and large number of patterns in the sparse version of the Hopfield model with analog neurons (Tirozzi & Tsodyks 1991). Therefore, in this model, chaotic attractors correspond to the neuronal representations of memories.

1.5 Overview

In the first chapter, I study a recurrent neuronal network endowed with a learning rule that belongs to the family described in Eq. (1.1) under an external dynamic stimulation. I show that depending on the stimulus properties, both sequential and persistent activity can be learned. This suggests that cortical circuits endowed with a single unsupervised learning rule can learn qualitatively different neuronal dynamics (i.e. persistent vs sequential activity) depending on the stimuli statistics. This chapter corresponds to a submitted paper which is currently posted on bioRxiv (Pereira & Brunel 2018*b*).

In the second chapter, I study a recurrent neuronal network constrained by *in vivo* inferior temporal cortex data (Woloszyn & Sheinberg 2012, Lim et al. 2015). The network presents attractor dynamics without any need for parameter tuning, reproducing landmark statistical properties of cortical neurons during delay response tasks. Additionally, I show that learning rules inferred from data (Lim et al. 2015) are close to maximizing the number

of stored patterns, suggesting that learning rules in ITC are optimized for storing a large number of memories as attractor states. Finally, I show that in a region of the parameter space memory states are chaotic, providing with this a new mechanism for explaining the heterogeneity observed during delay periods in the prefrontal cortex. This chapter corresponds to a published paper (Pereira & Brunel 2018*a*).

In the third chapter, I develop a general theory for the transition to chaos of memory states, and explore the effect of *online learning* of memories. I show that memory states can be fixed-point (newer memories) or chaotic attractors (older memories) depending on its age, leading to a continuum of different retrieval states with age-dependent spatiotemporal statistics. This chapter corresponds to a manuscript currently in preparation.

In the fourth chapter, I study a recurrent neuronal network in which sequences of patterns are learned. In this network, patterns are retrieved sequentially in the order that they were presented. I develop a theory for patterns with Gaussian statistics, obtaining dynamical equations for the transient correlation between the network activity and the stored patterns throughout the sequence. I compute the capacity of the network, that is the number of sequences that can be stored as a function of network size, and show that it grows linearly with network size. This result is comparable to that found in networks storing fixed-point attractors. This chapter corresponds to part of a manuscript currently in preparation.

CHAPTER 2

UNSUPERVISED LEARNING OF PERSISTENT AND SEQUENTIAL ACTIVITY

2.1 Contribution

The work presented in this chapter correspond to the submitted publication Pereira & Brunel (2018*b*). The authors are Ulises Pereira and Nicolas Brunel. U.P. and N.B. designed the research. U.P. and N.B. performed the research. U.P. and N.B. wrote the manuscript.

2.2 Introduction

Selective persistent activity (PA) has been observed in many neurophysiological experiments in primates performing delayed response tasks, in which the identity or spatial location of a stimulus must be maintained in working memory, in multiple cortical areas, including areas in the temporal lobe (Fuster et al. 1982, Miyashita 1988, Miyashita & Chang 1988, Sakai & Miyashita 1991, Nakamura & Kubota 1995*b*, Naya et al. 1996, Miller et al. 1996*a*, Erickson & Desimone 1999), parietal cortex (Koch & Fuster 1989*a*, Chafee & Goldman-Rakic 1998) and prefrontal cortex (Fuster et al. 1971, Funahashi et al. 1989, 1990, 1991, Miller et al. 1996*b*). More recently, selective persistent activity has also been observed in mice (Liu et al. 2014, Guo et al. 2014, Inagaki et al. 2017) as well as flies (Kim et al. 2017). It has been hypothesized that PA represents the mechanism at a network level of the ability to hold an item in working (*active*) memory for several seconds for behavioral demands. Theoretical studies support the hypothesis that persistent activity is caused by recurrent excitatory connections in networks of heavily interconnected populations of neurons (Amit et al. 1994, Durstewitz et al. 2000, Wang 2001, Brunel 2005). In these models, PA is represented as a fixed point attractor of the dynamics of a network that has multiple stable fixed points. The connectivity matrix in

such models has a strong degree of symmetry, with strong recurrent connections between subgroups of neurons which are activated by the same stimulus. This connectivity matrix can be learned by modifying recurrent connections in a network according to an unsupervised Hebbian learning rule (Mongillo et al. 2005, Litwin-Kumar & Doiron 2014*a*, Zenke et al. 2015).

Sequential activity (SA) has been also observed across multiples species in a number of behaviors such as spatial navigation (Foster & Wilson 2006, Harvey et al. 2012, Grosmark & Buzsáki 2016) and bird song generation (Hahnloser et al. 2002, Amador et al. 2013, Okubo et al. 2015). Furthermore, a large body of experimental evidence shows that SA can be learned throughout experience (Okubo et al. 2015, Grosmark & Buzsáki 2016). Several theoretical network models have been able to produce SA (Abeles 1991, Amari 1972, Kleinfeld & Sompolinsky 1988, Diesmann et al. 1999, Izhikevich 2006, Liu & Buonomano 2009, Fiete et al. 2010, Waddington et al. 2012, Cannon et al. 2015). In these models, the connectivity contains a feed-forward structure - neurons active at a given time in the sequence project in a feed-forward manner to the group of neurons which are active next. From a theoretical stand point, the mechanism to generate SA is fundamentally different from the one that generates PA. While SA usually corresponds to a path in the state space of the network, PA is identified as a fixed point attractor. Thus, SA has an inherent transient nature while PA is at least linearly stable in a dynamical system sense.

The question of how sequential activity can be learned in networks with plastic synapses has received increased interest in recent years. The models investigated can be roughly divided in two categories: models with supervised and unsupervised plasticity rules. In models with supervised plasticity rules, the synapses are updated according the activity of the network and an *error signal* that carries information about the difference between the current network dynamics and the one that it is expected to learn by the network (Sussillo & Abbott 2009, Memmesheimer et al. 2014, Laje & Buonomano 2013, Rajan et al. 2016).

In models with unsupervised plasticity rules, sequential dynamics is shaped by external stimulation without an error signal (Jun & Jin 2007, Liu & Buonomano 2009, Fiete et al. 2010, Waddington et al. 2012, Okubo et al. 2015, Veliz-Cuba et al. 2015). In those models SA is generated spontaneously, and the temporal statistics of the stimulation shapes the specific timing of the sequences.

Both experimental and theoretical work therefore suggest that neural networks in the brain are capable to learn PA and SA. One unresolved issue is whether the learning rules used by brain networks to learn PA are fundamentally different than the ones used to learn SA, or whether the same learning rule can produce both, depending on the statistics of the inputs to the network. Learning rules employed in theoretical studies to learn PA typically do not contain any temporal asymmetry, while rules used to learn SA need to contain such a temporal asymmetry.

Here, we hypothesize that a single learning rule is able to learn both, depending on the statistics of the inputs. We investigate what are the conditions for the plasticity mechanisms and external stimulation to learn PA or SA using unsupervised plasticity rules. We consider a model composed of multiple populations of excitatory neurons, each activated by a distinct stimulus. We consider a sequential stimulation protocol in which each population of neurons is stimulated one at a time, one after the other. This protocol is characterized by two parameters, the duration of stimulus presentations and the time interval between stimulations. This simple setting allows us to explore between the extremes of isolated stimulations with short or large duration and sequential stimulations close or far apart temporally. We use a rate model to describe the activity of populations of neurons (Wilson & Cowan 1972). The connectivity in this model represents the average of the synaptic connections between populations of neurons, allowing to investigate at a mesoscopic level the learning mechanisms of PA and SA. This model has the advantage of analytical tractability.

This paper is organized as follows: We first characterize the types of possible dynamics

observed in network with both feed-forward and recurrent connections, in the space of possible (fixed) connectivities. We then show that a network with plastic connections described by a unsupervised temporally asymmetric Hebbian plasticity rule stimulated sequentially does not stably learn PA and SA. We then explore two types of stabilization mechanisms: 1) synaptic normalization; 2) a multiplicative learning rule. We show that when a synaptic normalization mechanism is included, PA and SA cannot be learned stably during sequential stimulation. However, the addition of a modified multiplicative learning rule leads to successful learning of PA or SA, depending on the temporal parameters of external inputs, and the learning can be characterized analytically as a dynamical system in the space of fixed connectivities parametrized by the stimulus parameters.

2.3 Methods

2.3.1 Networks with fixed connectivity

We first consider three different n population rate models that share in common two connectivity motifs that have been classically considered a distinctive feature of PA and SA respectively: recurrent and feed-forward connections. The three network models considered are: 1) n excitatory neurons; 2) n excitatory neurons with shared inhibition; 3) n excitatory neurons with adaptation. The strength of the recurrent and feed-forward connections are w and s respectively. We used the current based version of the widely used firing rate model, which is equivalent to its rate based version (Miller & Fumarola 2012) with three different nonlinear transfer functions.

Network of excitatory neurons

The network consists in n excitatory populations connected by feed-forward and recurrent connections with strength w and s respectively as it is shown in Fig C.1A.I. The dynamics

is given by:

$$\begin{aligned} \tau \frac{du_1}{dt} &= I_1 - u_1 + w\phi(u_1) \\ \tau \frac{du_i}{dt} &= I_i - u_i + w\phi(u_i) + s\phi(u_{i-1}) \quad i = 2, \ldots, n \end{aligned} \tag{2.1}$$

where I_i represents the external input to neuron i, τ is the characteristic time scale for excitatory populations and $\phi(u)$ is the current to average firing rate transfer function (or f-I curve). The resulting average firing rates are denoted by $r_i \equiv \phi(u_i)$.

Network of excitatory neurons with shared inhibition

The network consist in n excitatory populations connected as in section 2.3.1, and a single inhibitory population fully connected with the excitatory populations. A schematic of the network architecture is shown in Fig C.1A.II. Assuming a linear inhibitory transfer function, the dynamics of the network is given by:

$$\begin{aligned} \tau \frac{du_1}{dt} &= I_1 - u_1 + w\phi(u_1) - w_{EI} u_I \\ \tau \frac{du_i}{dt} &= I_i - u_i + w\phi(u_i) + s\phi(u_{i-1}) - w_{EI} u_I \quad i = 2, \ldots, n \\ \tau_I \frac{du_I}{dt} &= -u_I + w_{IE} \sum_{j=1}^{n} \phi(u_j), \end{aligned} \tag{2.2}$$

where w_{EI} is the average inhibitory synaptic strength from inhibitory to excitatory populations, w_{IE} the average inhibitory synaptic strength from excitatory to inhibitory populations and τ_I the characteristic time scale of the inhibitory population. When $\tau_I \ll \tau$, then

$u_I \approx w_{IE} \sum_{i=1}^{N} \phi(u_i)$ and Eq. (2.2) becomes

$$\begin{aligned}
\tau \frac{du_1}{dt} &= I_1 - u_1 + w\phi(u_1) - \frac{w_I}{n} \sum_{j=1}^{n} \phi(u_j) \\
\tau \frac{du_i}{dt} &= I_i - u_i + w\phi(u_i) + s\phi(u_{i-1}) - \frac{w_I}{n} \sum_{j=1}^{n} \phi(u_j) \quad i = 2, \ldots, n,
\end{aligned} \tag{2.3}$$

where $w_I \equiv nW_{EI}W_{IE}$. See Fig. 2.12 in the Supplementary Material for the agreement between the full model described in Eq. (2.2) and its approximation in Eq. (2.3).

Network of excitatory neurons with adaptation

This network consist in n excitatory populations connected as in sections 2.3.1 and 2.3.1 plus an adaptation mechanism for each population. A schematic of the network architecture is shown in Fig C.1A.III. The dynamics of the network is given by:

$$\begin{aligned}
\tau \frac{du_1}{dt} &= I_1 - u_1 + w\phi(u_1) - a_1 \\
\tau \frac{du_i}{dt} &= I_i - u_i + w\phi(u_i) + s\phi(u_{i-1}) - a_i \quad i = 2, \ldots, n \\
\tau_a \frac{da_i}{dt} &= u_i - \beta a_i \quad i = 1, \ldots, n
\end{aligned} \tag{2.4}$$

where τ_a is the characteristic time scale of the adaptation mechanism, and β measures the strength of adaptation.

2.3.2 Transfer functions

For the fixed connectivity part of this study we used three different families of transfer functions. The sigmoidal transfer function is described by

$$\phi(u) = \frac{1}{2} \left(1 + \tanh[a(u+b)]\right). \tag{2.5}$$

This is a saturating monotonic function of the total input, and represents a normalized firing rate. This transfer function has been widely used in many theoretical studies in neuroscience (Gerstner et al. 2014, Ermentrout & Terman 2010), and have the advantage to be smooth. Furthermore, we have recently shown that such transfer functions provide good fits to *in vivo* data (Pereira & Brunel 2018*a*).

The second transfer function considered is piecewise linear:

$$\phi(u) = \begin{cases} 0 & \text{if } \theta > u \\ \nu(u-\theta) & \text{if } \theta \leq u \leq u_c \\ \nu(u_c-\theta) & u_c < u. \end{cases} \tag{2.6}$$

This is a piecewise linear approximation of the sigmoidal transfer function. Using this transfer function, the nonlinear dynamics of a network with a sigmoidal transfer function can be approximated and analyzed as a piecewise linear dynamical system.

The third transfer function used in this work is piecewise nonlinear (Brunel 2003)

$$\phi(u) = \begin{cases} 0 & \text{if } \tilde{\theta} > u \\ \tilde{\nu}\left(\frac{u-\tilde{\theta}}{\tilde{u}_c-\tilde{\theta}}\right)^2 & \text{if } \tilde{\theta} \leq u \leq \tilde{u}_c \\ 2\tilde{\nu}\sqrt{\frac{u-\tilde{\theta}}{\tilde{u}_c-\tilde{\theta}} - \frac{3}{4}} & \tilde{u}_c < u. \end{cases} \tag{2.7}$$

This transfer function combines several features that are present in more realistic spiking neuron models and/or real neurons: a supralinear region at low rates, described by a power law (Roxin et al. 2011), and a square root behavior at higher rates, as expected in neurons that exhibit a saddle node bifurcation to periodic firing (Ermentrout & Terman 2010). Examples of these three transfer functions are shown in Fig 2.1.

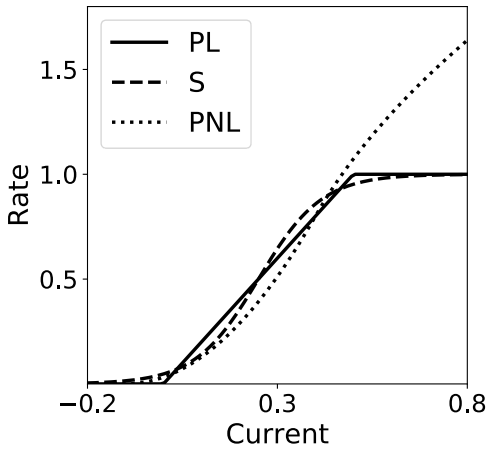

Figure 2.1: **Transfer Functions.** Piecewise linear (PL), sigmoidal (S) and piecewise non-linear (PNL) transfer functions. Parameters are the same as the ones used in Fig C.1.

2.3.3 Temporally asymmetric Hebbian plasticity rule

When a temporally asymmetric Hebbian plasticity rule is included (see sections 2.4.1-2.4.4 in Results), the dynamics of excitatory-to-excitatory connectivity obeys

$$\frac{d\mathbb{W}_{i,j}}{dt} = \frac{w_{max} f[r_i(t)] g[r_j(t-D)] - \mathbb{W}_{i,j}}{\tau_w[r_i(t), r_j(t-D)]}, \tag{2.8}$$

where $f(r)$ and $g(r)$ are sigmoidal functions given by

$$g(r) = \frac{1}{2}\left(1 + \tanh\left[a_{\text{pre}}(r - b_{\text{pre}})\right]\right) \tag{2.9}$$

$$f(r) = \frac{1}{2}\left(1 + \tanh\left[a_{\text{post}}(r - b_{\text{post}})\right]\right). \tag{2.10}$$

They describe the dependence of the learning rule on post and presynaptic firing rates, respectively (i.e. their dependence on $\phi(u_i)$ and $\phi(u_j)$), and are bounded by zero for small or negative values of the population synaptic current, and by one for large values (see Fig 2.4 A and B). Here w_{max} is the maximal synaptic efficacy; D is a temporal delay; and τ_w is an activity-dependent time constant of the plasticity rule. The learning time scale is given by

$$\tau_w[r_i(t), r_j(t-D)] = \tau_{\text{post}}[r_i(t)]\tau_{\text{pre}}[r_j(t-D)], \tag{2.11}$$

where

$$\tau_{\text{pre}}(r) = \tau_{\text{post}}(r) = \begin{cases} \infty & \text{if } r < r_w \\ \sqrt{T_w} & \text{if } r_w \leq r. \end{cases} \tag{2.12}$$

Here r_w and T_w are the plasticity threshold (see dashed line in Fig 2.4A-C) and time scale respectively. The time scale T_w is chosen to be several order of magnitude slower than the population dynamics (see Table 2.3). When pre and/or post-synaptic currents are below a plasticity threshold r_w, the activity-dependent time constant τ_w becomes infinite, and therefore no plasticity occurs. When both are above r_w, then the activity-dependent time constant τ_w is equal to T_w, and plasticity is ongoing. Thus, with this rule strong, long and/or contiguous in time enough stimuli produce lasting modifications in the synaptic weights. Otherwise, no learning occur.

2.3.4 Synaptic normalization

When a synaptic normalization mechanism is included (see section 2.4.2 in Results), in addition to the Hebbian plasticity rule described in section 2.3.3, in our network simulations, at each time step we subtracted the average synaptic change to each incoming synapse to a given neuron. This average is taken over all the incoming synapses to a particular neuron. This simulation scheme ensures that the sum of the incoming synaptic weights to each neuron remains constant, i.e.

$$\sum_{j=1}^{n} \mathbb{W}_{i,j} = C \qquad i = 1, 2, \ldots, n. \tag{2.13}$$

2.3.5 Multiplicative homeostatic plasticity rule

We implement a modified version of the multiplicative homeostatic rule proposed in Renart et al. (2003), Toyoizumi et al. (2014) (see sections 2.4.3 and 2.4.4 in Results). The rule is implemented in addition to the Hebbian plasticity rule described in the section 2.3.3. In this rule an homeostatic variable H_i slowly controls the firing rate of neuron i by scaling its synaptic weights multiplicatively. The synaptic weights will be given by

$$\mathbf{W}_{i,j}(t) = H_i(t)\mathbb{W}_{i,j}(t). \tag{2.14}$$

The variable $\mathbb{W}_{i,j}(t)$ is governed by the Hebbian plasticity rule described by Eqs (2.8-2.12). The dynamics for H_i is given by

$$\tau_H \dot{H}_i = \left(1 - \frac{r_i(t)}{r_0}\right) H_i - H_i^2, \tag{2.15}$$

where $r_0 = \phi(u_0)$ is a parameter that controls the *average* firing rate of population i and τ_H is the characteristic time scale of the learning rule. Note that because of the quadratic term in the r.h.s. of Eq. (2.15), this rule does not in general keep the firing rates at a fixed value, and therefore this rule is not strictly speaking homeostatic. However, we keep this terminology due to the similarity with the standard homeostatic rule that does not include this quadratic term.

2.3.6 Learning dynamics under noisy stimulation

In the last section of the Results, we include noise in the population dynamics in order to asses the robustness of the learning process (see section 2.4.4 in Results). The equations used to describe the dynamics of the network with Hebbian and homeostatic plasticity are given by

$$
\begin{aligned}
\tau \dot{u}_i &= \sigma\eta_i + I_i + \sum_{j=1}^{n} H_i \mathbb{W}_{i,j} r_j - \frac{W_I}{n} \sum_{i=1}^{n} \phi(u_i) \\
\dot{\mathbb{W}}_{i,j} &= \frac{w_{max} f[r_i(t))] g[r_j(t-D)] - \mathbb{W}_{i,j}}{\tau_w(r_i(t), r_j(t-D))} \\
\tau_H \dot{H}_i &= \left(1 - \frac{r_i(t)}{r_0}\right) H_i - H_i^2,
\end{aligned} \tag{2.16}
$$

where $r_i(t) = \phi(u_i(t))$ for $i = 1, 2, \ldots, n$ and η_i is a Gaussian white noise.

2.3.7 Sequential stimulation

During the learning protocol excitatory populations are stimulated sequentially once at a time for a period T and a time delay Δ. The stimulation can be implemented as a sequence of vectors presented to the entire the network (i.e. $I\vec{e}_1, I\vec{e}_2, \ldots, I\vec{e}_n$), each vector corresponds to the canonical base in $\mathbb{R}^n$ scaled by a stimulation amplitude I. This sequence of stimulation is repeated k times. To prevent a concatenation between the first and the last population stimulated, the period between each repetition k is much longer than T and Δ and any time constant of the network. Each stimulus in the sequence has the same magnitude, that is larger than the learning threshold (i.e. $r_w < I$). A schematic diagram of the stimulation protocol is shown in Fig 2.5 A.

2.4 Persistent and sequential activity in networks with fixed connectivity

To better understand the dependence of PA and SA generation on network connectivity, we consider first a simple n population rate model with fixed feed-forward and recurrent connectivity (see Fig C.1A). This architecture possesses the two connectivity motifs that have been classically considered the hallmarks of PA and SA — recurrent and feed-forward

connections — in a space of parameters that is low dimensional enough to be suitable for full analytical treatment. In this model, the dynamics of the network is characterized by the synaptic inputs u_i to each population of the network ($i = 1, \ldots, n$) whose dynamics obey the system of ordinary differential equations in Eq. (2.1). Note that we use here the *current based* formulation of the firing rate equations, that has been shown to be equivalent to the *rate based* formulation (Miller & Fumarola 2012).

In this model, we identify the regions in the connectivity parameter space where SA, PA or decaying sequences of activity (dSA) are generated. We start with a piecewise linear transfer function with slope ν, and compute the bifurcation diagram that gives the boundaries for qualitatively different dynamics in the parameter space (see Fig C.1B and section 2.6.2 in the Supplementary Material for mathematical details). We find that robust SA can be generated provided recurrent connections are smaller than the inverse of the slope ν, and the feed-forward connections are strong enough, $w < 1/\nu < w + s$. For large values of w ($w > 1/\nu$), the dynamics converge to a fixed point where $0 \leq p \leq n$ populations are in a high rate state, where p depends on the initial conditions. When both recurrent and feed-forward connections are weak enough (i.e. $w + s < 1/\nu$) the activity decays to zero firing rate fixed point, after a transient in which different populations are transiently activated - a pattern which we term decaying sequence of activity or dSA.

This picture is qualitatively similar when other types of nonlinear transfer functions are used (see Methods and Fig 2.1 for the transfer functions used in this paper). The saturation nonlinearity of the transfer function is key to generate long lasting (non-attenuated) SA even when the number of populations is large. In a linear network, sequential activity would increase without bound for an increasing number of populations participating in the SA (see Fig C.1B, dashed lines and section 2.6.2 in the Supplementary Material for mathematical details). During sequential activity, each population is active for a specific time interval. We used the analytical solution of the linearized system (see Eq. 2.22) to show that the duration

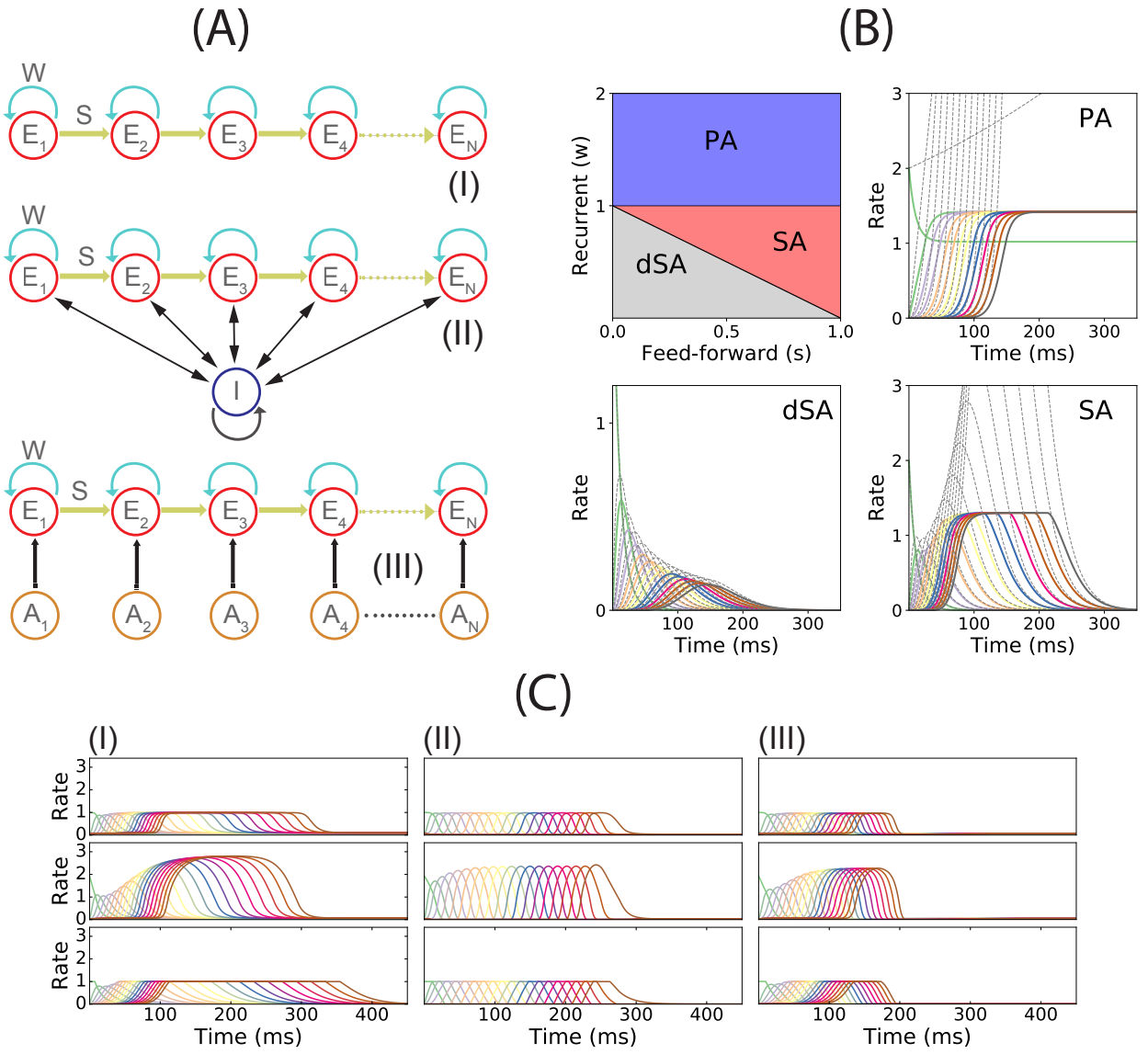

Figure 2.2: **PA and SA generation in a network with fixed connectivity. (A)** Three models of recurrent and feed-forward connected populations: (I) pure excitatory, (II) excitatory with shared inhibition and (III) excitatory with adaptation. **(B)** Phase diagram for model (I) using a piecewise linear transfer function (top-left plot) and examples of the dynamics corresponding to the three phases. Dashed lines correspond to the dynamics for the same network but using a linear transfer function. **(C)** SA generation for models (I), (II) and (III) using sigmoidal (first row), piecewise nonlinear (second row) and piecewise linear (third row) transfer functions. Parameters used in panels B,C can be found in Table 2.1.

of this active interval scales as the squared root of the position of the population along the sequence. This implies that for long lasting SA the fraction of active populations will increase with time (see Fig C.1B). This feature is not consistent with experimental evidence that shows that the width of the bursts of activity along the sequence is approximately constant in time (Hahnloser et al. 2002, Harvey et al. 2012). In the model, we can prevent this phenomenon by including negative feedback mechanisms to our network architecture, either global inhibition (see Fig C.1A.II) or adaptation (see Fig C.1 A.III). We found that in both cases the network robustly generates PA and SA in which the fraction of active populations is approximately constant in time. These results were also qualitatively similar when different saturation nonlinearities in the transfer function were considered (see Fig C.1C).

We now turn our attention to the network of excitatory neurons with global inhibition (Fig C.1 A.II), since inhibition is likely to be the dominant source of negative feedback in local cortical circuits. Inhibitory interneurons are typically faster than excitatory neurons (McCormick et al. 1985). For the sake of simplicity we set the inhibitory population dynamics as instantaneous compared with the excitatory timescale. Our numerical simulations confirm that this approximation preserves all the qualitative features of the dynamics with finite inhibitory time constants, up to values of $\tau_I = 0.5\tau$ (see Fig. 2.12 in the Supplementary Material). Using this approximation, the connectivity of the network is equivalent to a recurrent and feed-forward architecture plus a uniform matrix whose elements are $w_I \equiv nw_{EI}w_{IE}$. We obtained the bifurcation diagram for such a network with a piecewise linear transfer function (see section 2.6.4 in the Supplementary Material). This new bifurcation diagram shows qualitative differences with the pure excitatory network bifurcation diagram (see Fig 2.3). First, a qualitatively different behavior arises, where SA ends in persistent activity (region SA/PA). Second, the PA region breaks down in $n(n+1)/2$ square regions of size $w_I/n \times w_I/n$. Each region is characterized by a minimum and maximum number of populations active during PA. The lower left corner of each squared region is $(i_{min}\left(\frac{w_I}{n}\right), 1+$

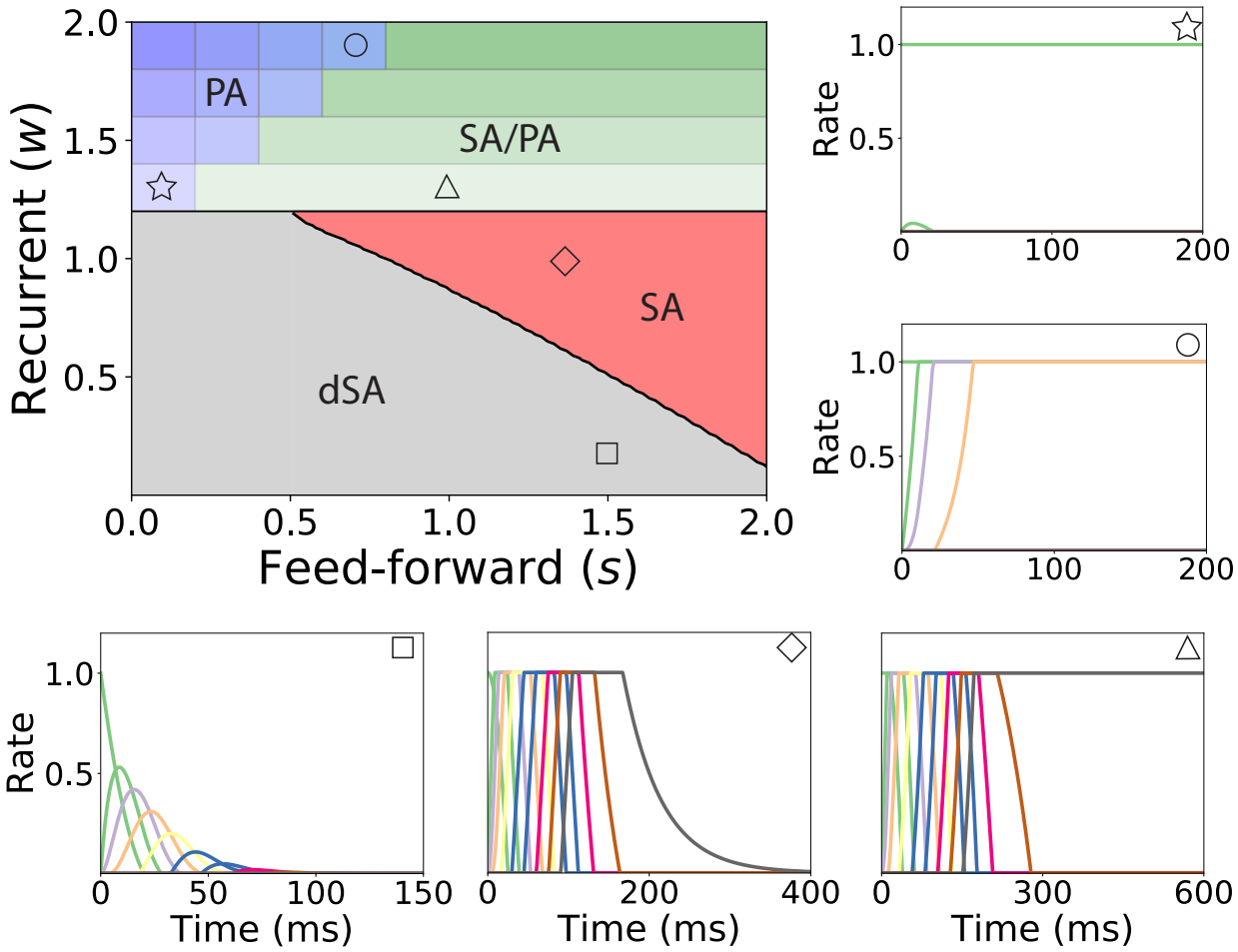

Figure 2.3: **Bifurcation diagram for feed-forward-recurrent connected network of excitatory populations with shared inhibition**. Top left plot: Bifurcation diagram in the s-w plane, showing qualitatively different regions: dSA (gray), SA (red), SA/PA (green) and PA (blue). The PA region is divided in sub-regions which are distingushed by the maximum and minimum number of populations active during PA (see text). The SA/PA region is also subdivided into sub-regions characterized by a different number of the maximum number of populations active in PA at the end of the sequence. Regions are separated by black lines and sub-regions are separated by gray lines. Five plots encompassing the bifurcation diagram show examples of the dynamics observed in its four qualitatively different regions. Initial condition: first population active at the maximum rate, while the rest is silent. The location in the corresponding regions of the parameter space are indicated with the symbols on the top right of the surrounding plots. Parameters can be found in Table 2.2.

$i_{max}\left(\frac{w_I}{n}\right)$) with $i_{min}, i_{max} = 1, 2, \ldots, n$ (see Fig 2.3, different regions in graded blue), where i_{min} and i_{max} correspond to the minimum and maximum number of population active during PA within this squared region when just the first population is initialized in the active state (Fig 2.3 top and middle right plots). Therefore, the number of possible patterns of PA increases with the strength of the recurrent connections and decreases with strength of the feed-forward connections. On the other hand, the SA/PA is divided in n qualitatively different rectangular regions of size $\left(\frac{w_I}{n}\right) \times [1 - j_{SA/PA}\left(\frac{w_I}{n}\right)]$ with $j_{SA/PA} = 1, 2, \ldots, n$, where $j_{SA/PA}$ corresponds to the number of populations that ends in PA after SA elicited by stimulating the first population in the sequence (Fig 2.3 bottom right plot). Then for a given strength of the recurrent connectivity w^* above $1 + \left(\frac{w_I}{n}\right)$, the critical feed-forward strength s_c that separates the PA and SA/PA regions is

$$s_c = \frac{w_I}{n}\left\lceil \frac{\left(w^* - 1 - \frac{w_I}{n}\right) n}{w_I} \right\rceil, \tag{2.17}$$

where $\lceil \cdot \rceil$ is the ceiling function. Similarly, for a given strength of the feed-forward connection s^* above $\frac{w_I}{n}$, the critical recurrent strength separating SA/PA and PA is

$$w_c = \frac{w_I}{n}\left\lceil \frac{\left(s^* - \frac{w_I}{n}\right) n}{w_I} \right\rceil. \tag{2.18}$$

Lastly, we find that the SA region is shrunk compared with the pure excitatory network, and that the dSA region is wider.

2.4.1 Unsupervised temporally asymmetric Hebbian plasticity rule

Let us consider now a fully connected network of n excitatory populations with plastic synapses and global fixed inhibition. The plasticity rule for the excitatory-to-excitatory connectivity is described by Eq. (2.8). Using this learning rule, with fixed pre and post activity, the connectivity tends asymptotically to a separable function of the pre and post

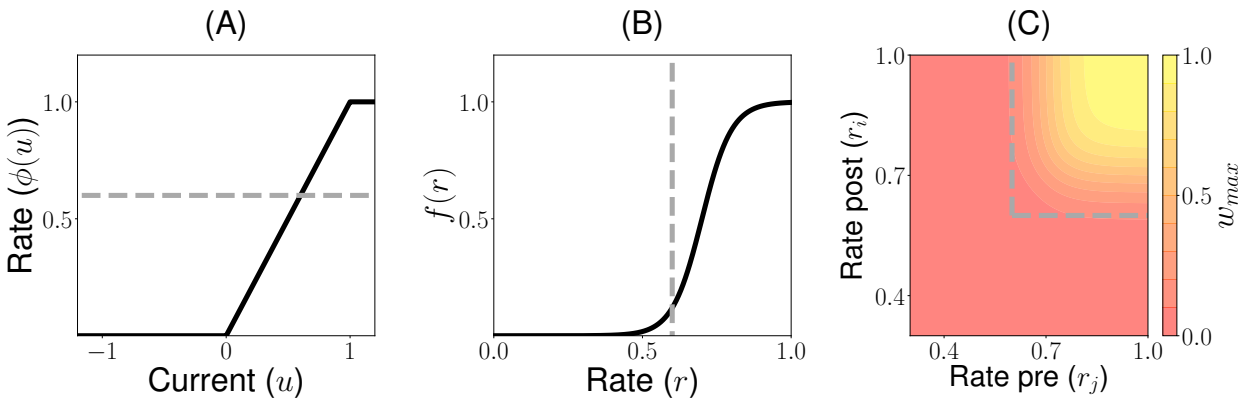

Figure 2.4: **Unsupervised Hebbian learning rule**: **(A)** Piecewise linear transfer function. The dashed gray horizontal line indicates the plasticity threshold r_w. **(B)** Post synaptic dependence on the rates of the stationary connectivity function, $f(r)$. The vertical dashed gray line indicates the plasticity threshold. **(C)** Contour plot of the stationary connectivity function, $w_{max} f(r_i) g(r_j)$. The dashed gray box indicates the plasticity threshold. Parameters can be found in Table 2.3.

synaptic activity. The functions $f(r)$ and $g(r)$ are bounded by zero for small or negative values of the population synaptic current, and by one for large values (see Fig 2.4 A and B). This learning rule is a generalization of classic Hebbian rules like the covariance rule (Dayan & Abbott 2001), with a non-linear dependence on both pre and post-synaptic firing rates.

The delay D in the learning rule leads to a temporal asymmetry (Blum & Abbott 1996, Gerstner & Abbott 1997, Veliz-Cuba et al. 2015). This delay describes the time it takes for calcium influx through NMDA receptors to reach its maximum (Sabatini et al. 2002, Graupner & Brunel 2012). When this learning rule operates and the network is externally stimulated, the connectivity changes depending on the interaction of the input, the network dynamics and the learning rule. Due to the relaxational nature of Eq. (2.8), for long times with no external stimulation the connectivity matrix will converge to a stationary rank-1 matrix with entries of the form $f(r_i^*)g(r_j^*)$, where $\vec{r}^* = \phi(\vec{u}^*)$ is the stationary firing rate vector, independent of all inputs presented in the past. Therefore, stimuli learned in the connectivity matrix will be erased by the background activity of the network for long times after stimulation. To prevent this inherent forgetting nature of the learning rule we introduce

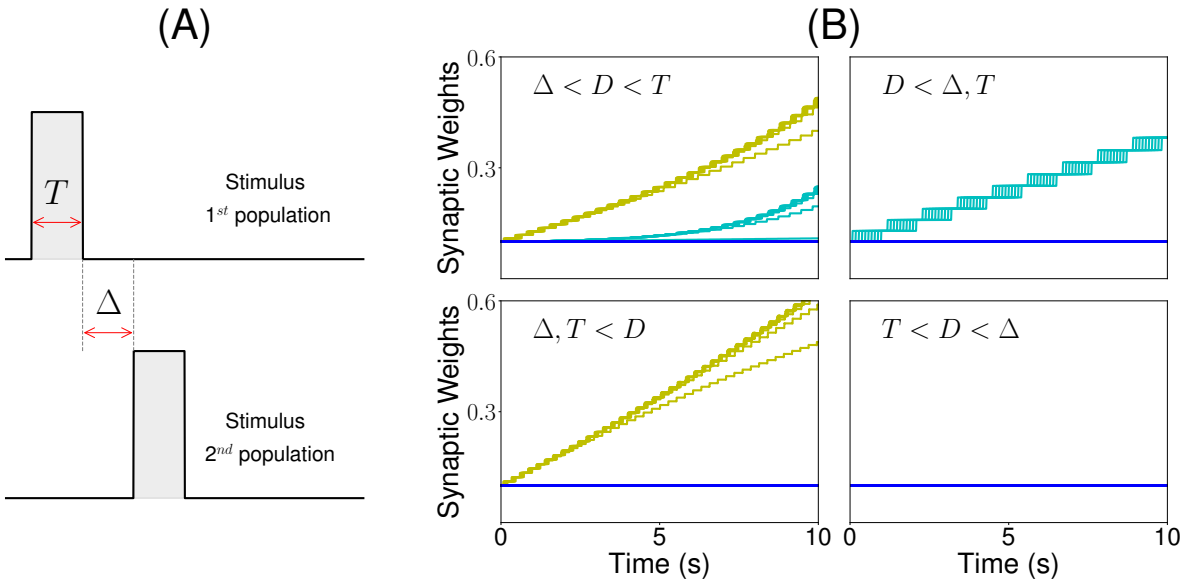

Figure 2.5: **Sequential stimulation and initial synaptic weights dynamics**. **(A)** Schematic diagram showing stimulation protocol for two populations. Population 1 is first stimulated for some time T. Then, after an inter-stimulation Δ time, population 2 is stimulated for the same duration T. **(B)** The weight dynamics is shown for four different stimulation regimes. Top-left: $\Delta < D < T$; top-right: $D < \Delta, T$; bottom-left: $T, \Delta < D$; bottom-right: $T < D < \Delta$. Cyan: recurrent connections; Yellow/Green: feed-forward; Blue: all other connections. Parameters can be found in Tables 2.3,2.4.

an activity-dependent plasticity time scale in Eqs. (2.11,2.12). Thus, when pre and/or post-synaptic currents are below a plasticity threshold r_w, the time scale becomes infinite, and therefore no plasticity occurs. When both are above r_w, then the time constant is given by T_w (see equation (2.12) and Fig 2.4). Lastly, the time scale T_w of these changes are chosen to be several order of magnitude slower than the population dynamics, consistent with the time it takes ($\sim$ 1 minute or more) for plasticity to be induced in standard synaptic plasticity protocols (see e.g. Markram, Lübke, Frotscher & Sakmann (1997), Bi & Poo (1998), Sjöström et al. (2001), but see Bittner et al. (2017)).

Our goal is to understand the conditions for a sequential stimulation to lead the network dynamics to PA or SA, depending of the temporal characteristics of the stimulus, when this plasticity rule is introduced. Here we consider a simple stimulation protocol where each population in the network is stimulated sequentially one population at a time (see Fig 2.5 A). In this protocol, population 1 is first stimulated for some time T. Then, after an inter-stimulation time Δ, population 2 is stimulated for the same duration T. The other populations are then stimulated one at a time (3, 4, ..., n) using the same protocol. The amplitude of the stimulation is fixed such that the maximum of the current elicited in each population is greater than the plasticity threshold of the learning rule. The time interval between each repetition of the sequence is much longer than T and Δ and any time constant of the network. When the duration of each stimulation is larger than the synaptic delay (i.e. $D < T$), recurrent connections increase, since the Hebbian term driving synaptic changes ($f[r_i(t)]g[r_i(t-D)]$, where i is the stimulated population) becomes large after a time D after the onset of the presentation. When the inter-stimulation time is smaller that the synaptic delay (i.e. $\Delta < D$), then the the feed-forward connections increase, since the Hebbian term driving synaptic changes ($f[r_{i+1}(t)]g[r_i(t-D)]$) is large in some initial interval during presentation of stimulus $i+1$.

As a result, there are four distinct regions of interest depending on the relative values

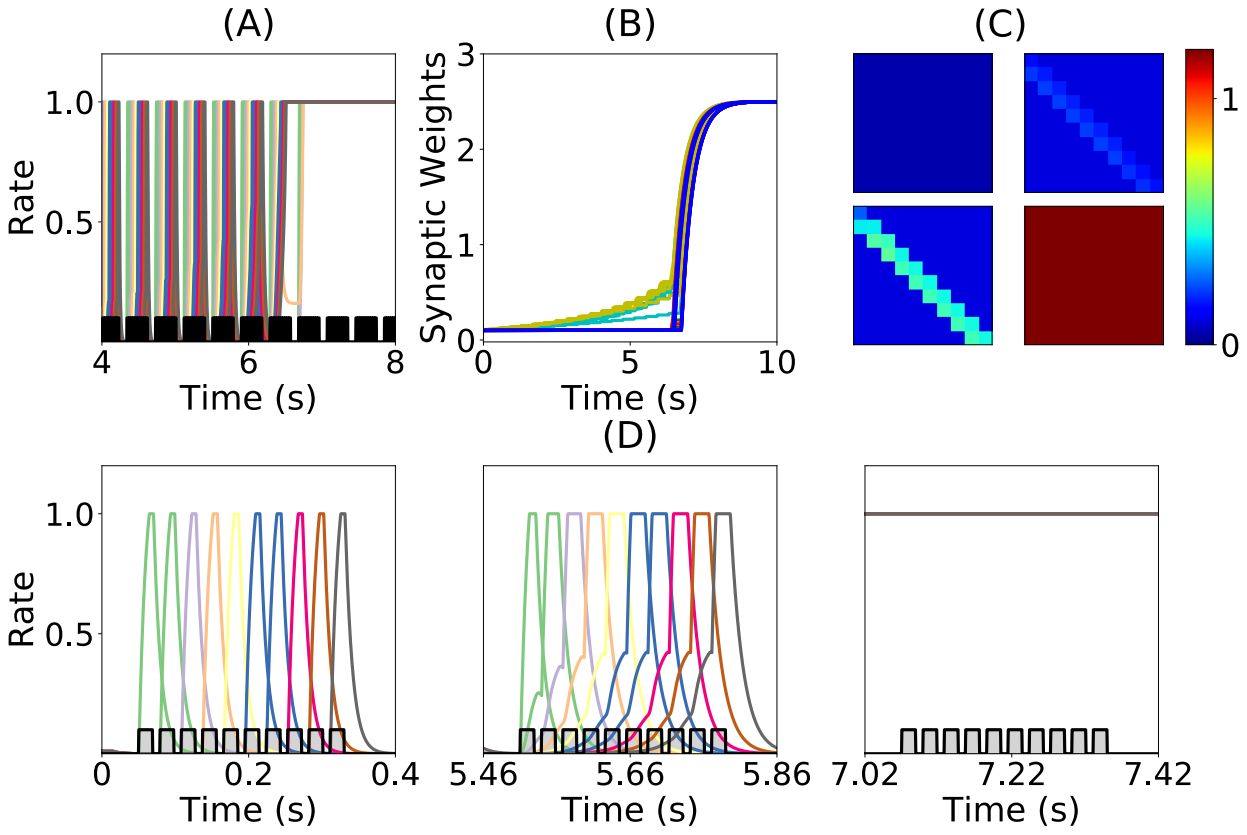

Figure 2.6: **Runaway instability of the unsupervised Hebbian learning rule**. **(A)** Population dynamics during 10s of sequential stimulation with $T = 19$ms and $\Delta = 10$ms. After about 6s, all populations become active at maximal rates. **(B)** Synaptic weights dynamics during stimulation. Color code as in Fig 2.4D. **(C)** Connectivity matrix at different stimulation times. From left to right and from top to bottom: 0s, 3s, 6s and 9s. **(D)** Three examples of population dynamics during a single sequential stimulation at 0s, 5.46s and 7.02s respectively. Note the buildup of activity preceding each stimulus presentation because of the build-up in the feedforward connectivity at 5.46s. In A and D the black and gray traces indicate a scaled version of the stimulus. Parameters can be found in Tables 2.3,2.4.

of the Δ and T with respect to the synaptic delay D. When T is larger than the synaptic delay, and Δ is smaller than the synaptic delay, both recurrent and feed-forward connections increase. When T is larger than the synaptic delay and Δ is much larger than D, only the recurrent connections increase. When Δ is smaller than the synaptic delay and T is much smaller, only the feed-forward connections increase. Lastly, when Δ is larger and T is smaller than D no changes in the connectivity are observed. The initial temporal evolution of both recurrent and feed-forward weights in representative examples of the four regions is presented in Fig 2.5 B. We chose not to study the region corresponding to $2T + \Delta < D$ here, which is a region where 'feed-forward' connections involving non-nearest neighbor populations can also increase during learning.

We found that this learning rule is in general unstable for long sequential stimulation when both feed-forward and recurrent connections increase during the stimulation (i.e. $\Delta < D < T$) to values large enough to produce persistent activity states. This is a consequence of the classic instability observed with Hebbian plasticity rules, where a positive feedback loop between the increase in synaptic connectivity and increase in firing rates leads to an explosive increase in both (Dayan & Abbott 2001). Larger feed-forward and recurrent connections lead to an increase in number of populations active at the same time during stimulation (see Fig 2.6 A and D) which produce an increase of the overall connectivity by the synaptic plasticity rule (Fig 2.6 B and C). This leads to an increase in the overall activity producing longer periods of PA during stimulation until a fixed point where many populations have high firing rates is reached, and the connectivity increases exponentially to its maximum value (see Fig 2.6 B and C). By increasing the plasticity threshold, it is possible to increase the number of stimulations (and consequently the strength of the feed-forward and recurrent connections) where the network's activity is stable. However, this does not solve the problem, since the instability on the weights eventually occurs but for a larger number of stimulations and stronger synaptic weights. In order to prevent this instability, we investigate in the next

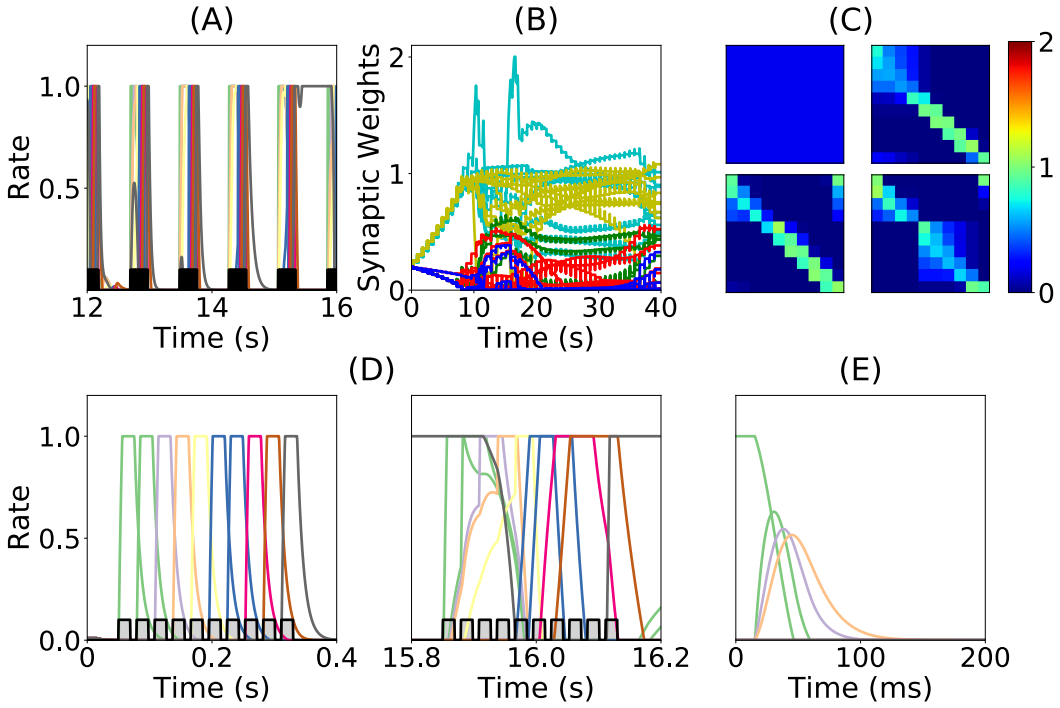

Figure 2.7: **Heterogeneous synaptic dynamics for Hebbian plasticity and synaptic normalization**. **(A)** Population dynamics during 10s of sequential stimulation with $T = 19$ms and $\Delta = 10$ms. **(B)** Synaptic weights dynamics during stimulation. Cyan: recurrent connections; Light Yellow/Green: feed-forward; Red: feed-backward; Blue: feed-second-forward; Green: feed-second-backward. **(C)** Connectivity matrix at different stimulation times. From left to right and from top to bottom: 0s, 13.8s, 27.6s and 41.5s. **(D)** Two examples of population dynamics during a single sequential stimulation at 0s and 15.8s respectively. In A and D the black and gray traces indicate a scaled version of the stimulus. **(E)** Network dynamics after learning for the initial condition where the first population is active at high rate and the rest silent. Parameters can be found in Tables 2.3,2.4.

sections two different stabilization mechanisms: synaptic normalization and homeostatic plasticity. Throughout this paper, for testing whether PA, SA, SA/PA or dSA is learned, after sequential stimulation we stimulate the first population and then check whether the network recalls the corresponding type of activity (see Fig 2.3).

2.4.2 Synaptic normalization

The first mechanism we consider is synaptic normalization. This mechanism is motivated by experimental evidence of conservation of total synaptic weight in neurons (Royer & Paré

2003, Bourne & Harris 2011). In our model, we enforce that the sum of the incoming synaptic weights to a given population is fixed throughout the dynamics (see Eq. 2.13 in Methods). This constraint prevents the growth of all the synaptic weights to their maximum value during sequential stimulation due to the Hebbian plasticity, as is described in the previous section. This leads to an heterogeneous dynamics in the synaptic weights where they strongly fluctuate in time during the stimulation period, see Fig 2.7B. We find that the network does not reach a stable connectivity structure, and that the connectivity after the stimulation markedly depends on the specific moment when stimulation ended for a large range of stimulation parameters.

At the initial stages of the stimulation, feed-forward and recurrent connections grow, while the rest of the synaptic connections decrease at the same rate (see Fig 2.7 B). When the feed-forward and recurrent connections are large enough for producing persistent activity, co-activation between a population(s) undergoing persistent activity and the population active due to the stimulation (which are not necessarily adjacent in the stimulation sequence, see Fig 2.7A,D) produce an increase in feed-back and upper triangular connections that are different than feed-forward and recurrent (see Fig 2.7B). In turn, feed-forward and recurrent connections decrease due to the synaptic normalization mechanism. This leads to complex dynamics in the synaptic weights, in which the connections sustaining co-active neuronal assemblies learned via Hebbian plasticity are depressed due to the interplay between synaptic normalization and sequential stimulation. This then leads to the formation of new assemblies due to the interplay of Hebbian plasticity and sequential stimulation.

During stimulation, the feed-forward and recurrent connectivity studied in the first section increase first, leading then in a second stage to clustered connectivities with strong bi-directional connections (see Fig 2.7C). Therefore, neither persistent nor sequential activity can be learned consistently after long times (see Fig 2.7E). Moreover, it is not clear whether neural circuits can use the observed complex synaptic dynamics to store retrievable

information about the external stimuli. Thus, we find that synaptic normalization is not sufficient in this case to stabilize learning dynamics and to lead to a consistent retrieval of PA or SA. We checked that this finding is robust to changes in parameters, in particular the sum of incoming synaptic weights. In the next section we consider a second stabilization mechanism, namely Homeostatic plasticity.

2.4.3 Multiplicative homeostatic plasticity

Homeostatic plasticity is another potential stabilization mechanism that has been characterized extensively in experiments (Turrigiano et al. 1998, Turrigiano 2017). The interplay between homeostatic plasticity and Hebbian plasticity has recently been the focus of multiple theoreotical studies (Renart et al. 2003, Toyoizumi et al. 2014, Keck et al. 2017). Here, we study the effect of multiplicative homeostatic and Hebbian plasticity for learning SA and PA. We consider a model for homeostatic plasticity in which the overall connectivity at each time $\mathbf{W}_{i,j}(t)$ is given by the multiplication of two synaptic variables with different time scales as is shown in Eq. (2.14). In this equation, the fast plastic variable $\mathbb{W}_{i,j}(t)$ (time scale of seconds) is governed by Hebbian plasticity, see Eq. (2.8). On the other hand, the slow (with a time scale of tens to hundred of seconds) homeostatic variable $H_i(t)$ scales the incoming weights to population i, ensuring that the network maintain low average firing rates on long time scales. Its dynamics of the homeostatic variable is given by Eq. (2.15). This is a modification of the standard homeostatic learning rule (Renart et al. 2003, Toyoizumi et al. 2014), that does not include the quadratic term in the r.h.s. of Eq. (2.15). The equation proposed in (Toyoizumi et al. 2014) stabilizes the network's activity during stimulation, preventing the runaway of the firing rates and synaptic weights. Scaling down the overall connectivity during stimulation prevents co-activation of multiple populations, and lead to stable learning, see Fig 2.13D and E. However, in the network's steady state (i.e. when times longer than the time scale of the homeostatic variable have passed without any stimulation), if the equation

proposed in (Toyoizumi et al. 2014) is used, then each connection will be proportional to the factor $\frac{\phi^{-1}(r_0)}{r_0}$ multiplied by a number of order one (see section 2.6.5 and 2.6.5 in the Supplementary Material for a general discussion and the corresponding mathematical details respectively). This implies that the steady state connectivity after learning will depend sensitively on the choice of the value of the objective background firing rate (i.e. r_0) and the specific functional form of the transfer function (i.e. $\phi(u)$). Due to the transfer function non-linearity, small changes in r_0 might produce large values for the factor $\frac{\phi^{-1}(r_0)}{r_0}$ and therefore very strong connections for the steady state connectivity (see Fig 2.13). This is due to the fact that steady state large values in the homeostatic variable H scale up the connectivity learned via Hebbian plasticity in a multiplicative fashion, see Eq. (2.14). In practice, PA is retrieved almost always independently of the type of stimulation presented during learning, and in the absence of the quadratic term in Eq. (2.15) no temporal attractor other than PA can be learned. This problem can be prevented by the introduction of a quadratic term in the original homeostatic rule (see section 2.6.5 in the Supplementary Material). Note that with this quadratic term, the homeostatic plasticity rule does not exactly achieve a given target firing rate, and therefore is not strictly speaking 'homeostatic'. However, since it is variant of the classic linear homeostatic rule, we have chosen to stick with this terminology.

We explore the role of this multiplicative homeostatic learning rule for learning both PA and SA. During sequential stimulation, the average firing rate is higher than the background objective firing rate r_0, and the homeostatic variables decrease to values that are smaller than 1, see Fig 2.8 A and C. As a result, during sequential stimulation the dynamics of the homeostatic variable will be dominated by the linear version of the homeostatic learning rule proposed in (Toyoizumi et al. 2014), since $H_i^2 \ll 1$. Then, the small values that the homeostatic variables take during the sequential stimulation scale down the increasing values of the recurrent and feed-forward connections due to Hebbian plasticity. This produces a weak excitatory connectivity during a repeated sequential stimulation (see Fig 2.8 C), prevent-

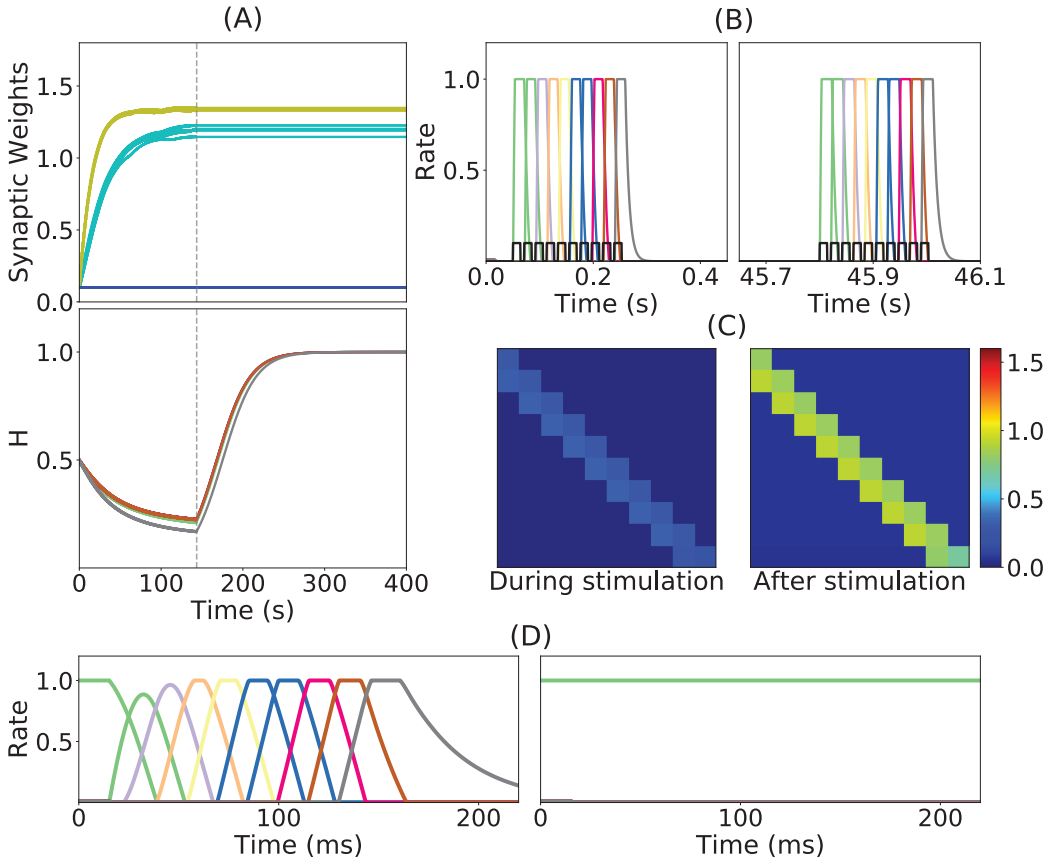

Figure 2.8: **Learning dynamics in a network with Hebbian and multiplicative homeostatic plasticity**: **(A)** Top: synaptic weights dynamics during and after stimulation. Cyan: recurrent; Yellow: feed-forward; Blue: all other connections. Bottom Homeostatic variables in excitatory populations. neuron i. Gray vertical dashed line indicate the end of the sequential stimulation. **(B)** Neuron dynamics during stimulation for two different periods of time. **(C)** Snapshots of the connectivity matrix $\mathbf{W}_{i,j}(t)$ at the end of the sequential stimulation (left) and 60s after the end of the sequential stimulation (right). **(D)** Network dynamics after learning following an initial condition where the first population is active at high rate while all others are silent for two different stimulation parameters, for two stimulation parameters, one that generates SA (left), the other PA (right). Parameters can be found in Tables 2.3,2.4.

ing activation of spurious populations during stimulation (see Fig 2.8 B), even though the strength of recurrent and feed-forward connections learned via Hebbian plasticity are strong enough to produce PA or SA, since these connections are *masked* by the homeostatic variable. When the network returns to the steady state after sequential stimulation, the homeostatic variables return to values $H_i \sim \mathcal{O}(1)$ (see section 2.6.5 in the Supplementary Material for the mathematical details), and the recurrent and feed-forward connections learned via Hebbian plasticity are *unmasked*. This mechanism stabilizes learning, allowing the network to stably learn strong recurrent and feed-forward connections, consistent with SA or PA dynamics (see Fig 2.8D).

The weakening of recurrent connections during sequential stimulation allows us to derive an approximate analytical description of the temporal evolution of the synaptic connectivity with learning. Since the net current due to connections between populations is very small, each population dynamics is well approximated by an exponential rise (decay) toward the stimulation current (background current) provided inhibition is weak enough (see Fig 2.9). By using this approximation we build a mapping that yields the value of the recurrent and feed-forward synaptic strengths as a function of stimulation number k, stimulation period, T, and delay, Δ (see Eqs. (2.50,2.51) in 2.6.6 of Supplementary Material). This mapping provides a fairly accurate match of both the dynamics of the synaptic weights and the final steady state connectivity matrix in the case of weak inhibition (see Fig 2.10A, corresponding to $w_I = 1$) and a less accurate match for stronger inhibition (see Fig 2.10B, $w_I = 2$). This is expected since our theoretical analysis neglects the effect of inhibition during learning (see section 2.6.6 of Supplementary Material). The mapping derived for evolution of the synaptic weights during sequential stimulation corresponds to a dynamical system in the (s, w) phase space that depends on the stimulus parameters (Δ, T) and the initial connectivity. The final connectivity corresponds to the fixed point of these dynamics (see Eqs. (2.52,2.53) in section 2.6.6 of Supplementary Material).

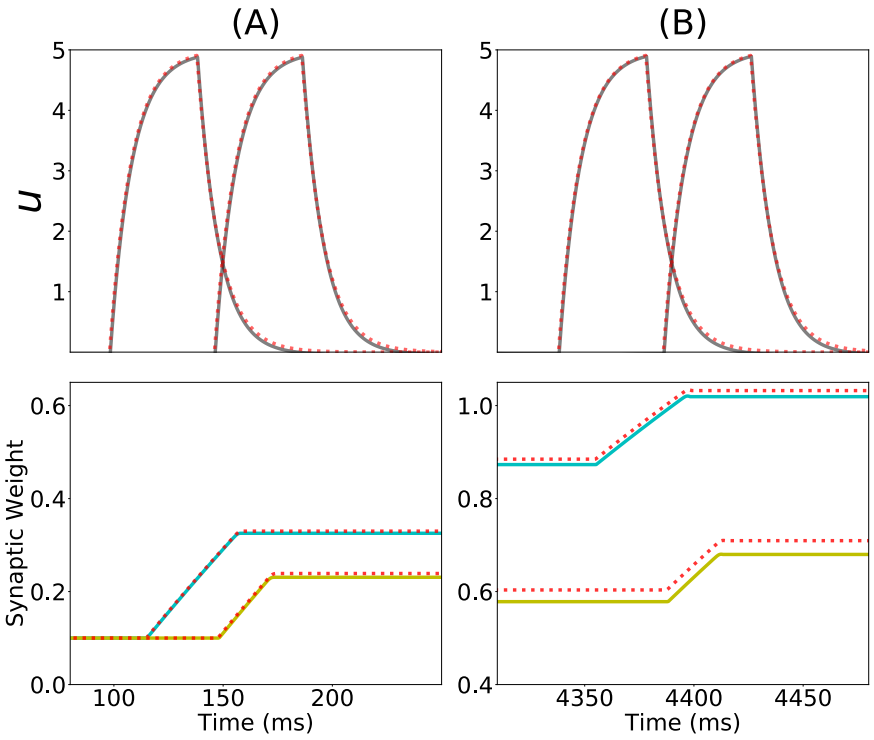

Figure 2.9: **Analytical approximation of the dynamics of the network with Hebbian and multiplicative homeostatic plasticity**: (**First row**) Current dynamics for the second and third populations in a network of 20 populations during one presentation of the sequence. The dashed red line shows the analytical approximation for the dynamics during stimulation (Eq. 2.42 in section 2.6.6 of Supplementary Material). (**Second row**) Dynamics of the recurrent synaptic strength within the second population (cyan), and the 'feed-forward' synaptic strength from the second to the third population (yellow) during the same presentation of the sequence. The dashed red line shows the analytical approximation for the synaptic weight dynamics (Eq. (2.44,2.48) in section 2.6.6 of Supplementary Materials). (**A**) and (**B**) correspond to the first and the fifth presentation of the stimulation sequence respectively. Parameters can be found in Tables 2.3,2.4.

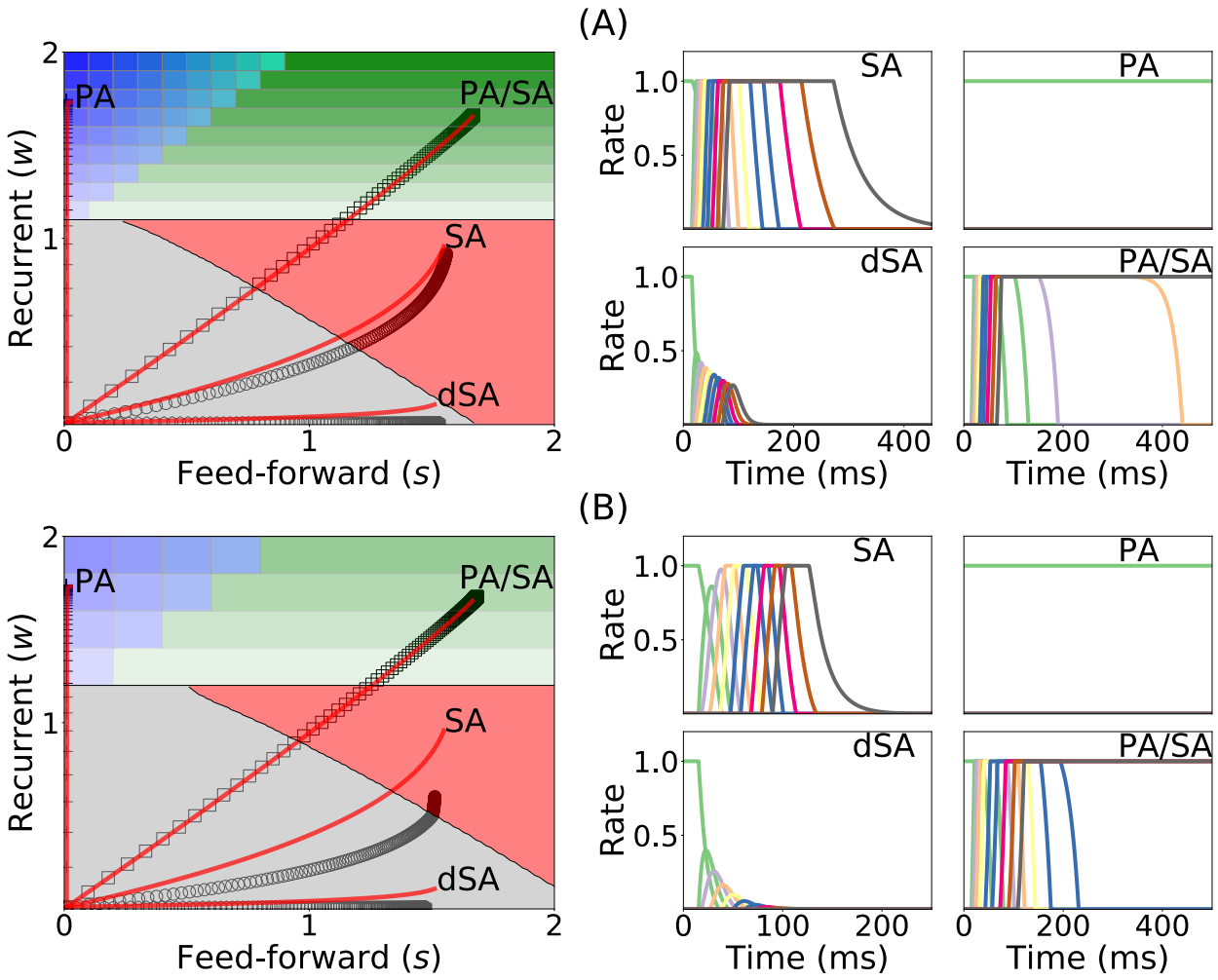

Figure 2.10: **Changes in recurrent and feed-forward synaptic strengths with learning, for different sequences with different temporal parameters**. (**Left**) Dynamics of recurrent and feed-forward connections in the (s, w) parameter space during sequential stimulation for four different values of Δ and T. Black circles (SA), plus signs (PA), hexagons (dSA), and squares (PA/SA) show the simulated dynamics for $(T, \Delta) = \{(7, 14), (50, 40), (5, 13), (20, 8.5)\}$ (in ms) respectively. Red traces indicate the approximated dynamics derived in section 2.6.6 of Supplementary Material. (**Right**) Rates dynamics after many presentations of the sequence. The first population was initialized at high rates, the others at low rates. (**A**) and (**B**) correspond to $w_I = 1$ and $w_I = 2$ respectively. Parameters can be found in Tables 2.3,2.4.

Fig 2.10 shows that depending on the temporal characteristics of the input sequence, the network can reach any of the four qualitatively different regions of the phase diagrams in a completely unsupervised fashion. For values of Δ that are smaller than the synaptic delay D and T on the order or larger than D, the network generates SA. For values of T approximately larger than D and for Δ small enough, the dynamics lead to SA/PA. Lastly PA is obtained for large enough Δ and T. These observations match with the intuition that stimulations long enough but far delayed in time leads to learning of PA and that stimulations contiguous in time but short enough leads to SA. Stimulations between these two conditions (long and contiguous) leads to a combination of both dynamics, i.e. SA/PA, as shown in Fig 2.10.

2.4.4 Learning and retrieval is robust to noise

Under *in vivo* conditions neural systems operate with large amount of variability in their inputs. In order to assess the effect of highly variable synaptic input current during learning and retrieval, we add a mean zero uncorrelated white noise to the dynamics when both Hebbian learning and homeoestatic plasticity are included in the network, as described in Eq. (2.16). We found that both the synaptic weights dynamics during learning and the retrieved spatiotemporal dynamics after learning are robust to noise (see Fig 2.11), even when the amplitude of the noise is large (i.e. inputs with values equal to the standard deviation of the noise lead to a population to fire at 30% of the maximum firing rate). During sequential stimulation, the learning dynamics is marginally altered for both weak and strong inhibition (compare Fig 2.11 with Fig 2.10). Importantly, the synaptic weights reach very similar stationary values compared with the case without noise. After learning, even though the rates stochastically fluctuate in time, the retrieved spatiotemporal attractors (i.e. PA, SA, dSA or PA/SA) are qualitatively similar as in the case without noise (compare Fig 2.11 with Fig 2.10). One qualitative difference in the case with external noise, is that in both

SA and PA/SA dynamical regimes random inputs lead to a repetition of the full or partial learned sequence. Altogether, this simulations show that the network can robustly learn and retrieve qualitatively the same spatiotemporal attractors in the presence of external noise.

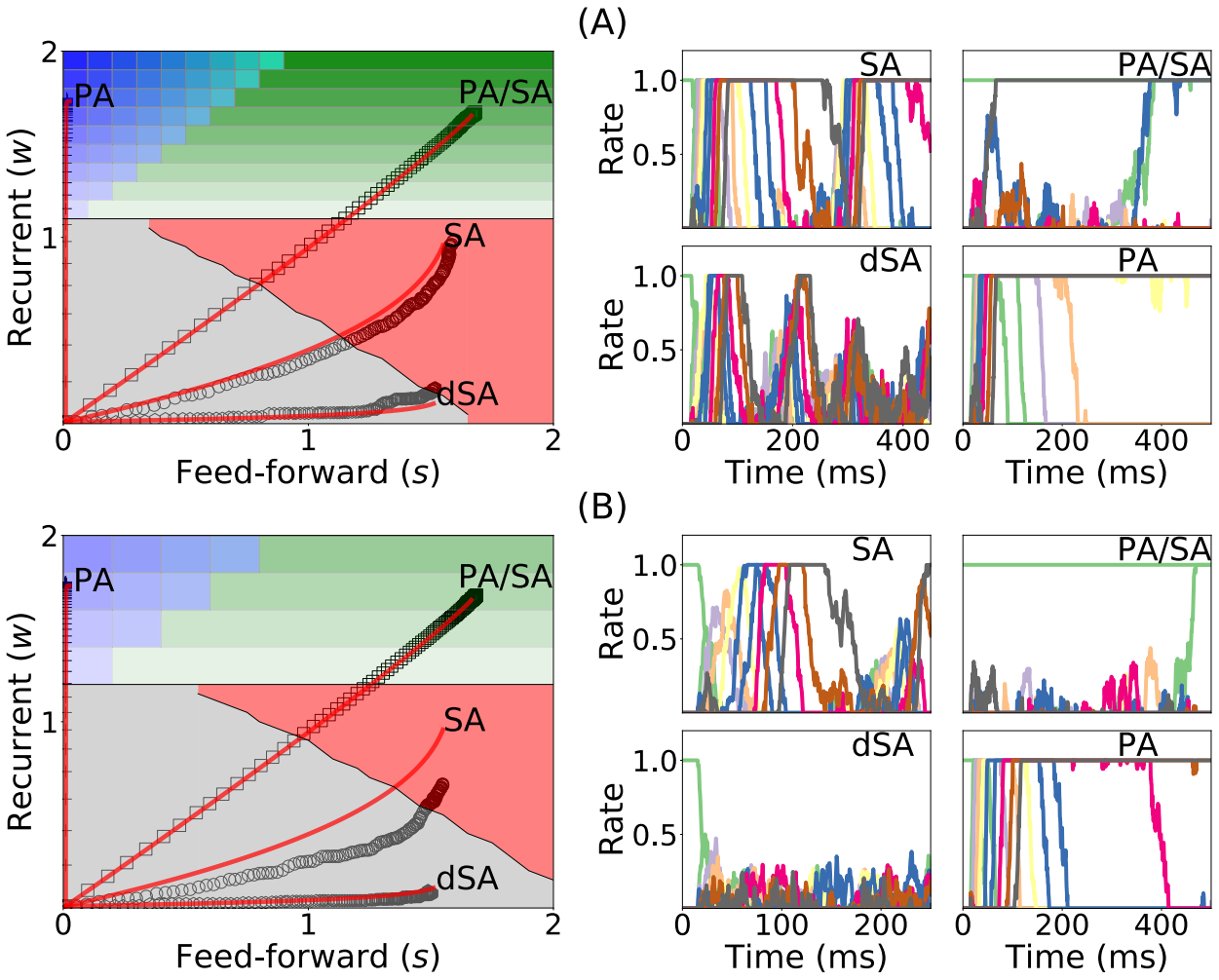

Figure 2.11: **Learning dynamics under noisy stimulation.** Same as in Fig 2.10, but in the presence of a white noise input current, with mean 0 and standard deviation of 0.3 (i.e. $\sigma = 0.3$ in Eq. (2.16)).

2.5 Discussion

We have shown that under sequential stimulation a network with biologically plausible plasticity rules can learn both PA or SA depending on the stimulus parameters. Two plasticity

mechanisms are needed: 1) Hebbian plasticity with temporal asymmetry; 2) a stabilization mechanism which prevents the runaway of synaptic weights while learning. When unsupervised Hebbian plasticity is present alone the network fails to stably learn PA or SA, while including multiplicative homeostatic plasticity stabilizes learning. For stable learning, we show that the learning process is described by a low dimensional autonomous dynamical system in the space of connectivities, leading to a simplified description of unsupervised learning of PA and SA by the network from external stimuli. Depending on the stimulus parameters, the network is flexible enough to learn selectively both types of activity by repeated exposure to a sequence of stimuli, without need for supervision. This suggests that cortical circuits endowed with a single learning rule can learn qualitatively different neural dynamics (i.e. persistent vs sequential activity) depending on the stimuli statistics.

Using the full characterization of the bifurcation diagram in the space of fixed feed-forward and recurrent connections developed here, we mapped the evolution of the connectivity during stimulation in the bifurcation diagram. We analytically and numerically showed that the synaptic weights evolve in the feed-forward–recurrent synaptic connections space until they reach their steady state (when the number of sequential stimulations is large). The specific point of the steady state in the bifurcation diagram depends solely on the stimulation parameters — stimulation period T and time delay Δ– and the connectivity initial conditions. We found that stimulations with long durations and large delays generically leads to the formation of PA, whereas stimulations with long enough durations and short delays leads to the formation of SA. Thus, persistent stimulation leads to persistent activity while sequential stimulation leads to sequential activity.

2.5.1 Learning of sequences in networks

A growing number of network models have been shown to be able to learn sequential activity. Models with supervised learning can reproduce perfectly target sequences through

minimization of a suitable error function (Sussillo & Abbott 2009, Memmesheimer et al. 2014, Laje & Buonomano 2013, Rajan et al. 2016), but the corresponding learning rules are not biophysically realistic.

Other investigators have studied how unsupervised learning rules leads to sequence generation. Early models of networks of binary neurons showed how various prescriptions for incorporating input sequences in the connectivity matrix can lead to sequence generation (see Kuhn & van Hemmen (1991)) - or, sometimes, both sequence generation or fixed point attractors depending on the inputs (Herz et al. 1988). The drawback of these models is that they separated a learning phase in which recurrent dynamics was shut down in order to form the synaptic connectivity matrix, and a retrieval phase in which the connectivity matrix does not change anymore.

Our model removes this artificial separation, since both plasticity rule and recurrent dynamics operate continuously, both during learning and recall. However, we found that there needs to be a mechanism to attenuate recurrent dynamics during learning for it to be stable. The mechanism we propose rely on a modified version of a standard homeostatic rule. Other mechanisms have been proposed, such as neuromodulators that would change the balance between recurrent and external inputs during presentation of behaviorally relevant stimuli (Hasselmo 2006).

The cost of not having supervision is that the network can only learn the temporal order of the presented stimuli, but not their precise timing. Veliz-Cuba et al (Veliz-Cuba et al. 2015) have recently provided a model which bear strong similarities with our model (rate model with unsupervised temporally asymmetric Hebbian plasticity rule), but includes in addition a short-term facilitation mechanism that allows the network to learn both order and precise timing of a sequence presented in input. However, their mechanisms requires precise fine tuning of parameters.

Models with temporally asymmetric Hebbian plasticity have also been investigated in

the context of the hippocampus (Abbott & Blum 1996, Gerstner & Abbott 1997, Mehta et al. 1997, Jahnke et al. 2015, Chenkov et al. 2017, Theodoni et al. 2017). In such models, feed-forward connectivity is learned through multiple visits of neighboring place fields, and sequential activity ('replays') can be triggered using appropriate inputs mimicking sharp-wave ripples. Other models use unsupervised Hebbian plasticity but qualitatively distinct mechanisms to generate sequential activity. In particular, several studies showed that sequences can be generated spontaneously from unstructured input noise (Fiete et al. 2010, Okubo et al. 2015). Murray and Escola (Murray & Escola 2017) showed that sequences can be generated in networks of inhibitory neurons with anti-Hebbian plasticity, and proposed that this mechanism is at work in the striatum.

2.5.2 Stabilization mechanisms

Consistent with many previous studies (Dayan & Abbott 2001), we have shown that a network with unsupervised Hebbian plasticity under sequential stimulation leads to a runaway of the synaptic weights. This instability is due to a positive feed-back loop generated by the progressive increase of network activity leading to a progressive increase in average synaptic strength when PA or SA are being learned. One possible solution for this problem was first proposed in the context of attractor neural network models (Amit et al. 1985, Amit & Fusi 1994, Tsodyks & Feigel'Man 1988). In these models, patterns are learned upon presentation during a *learning phase* where synapses are plastic but there is no ongoing network dynamics. After the *learning phase*, the learning of attractors is tested in a *retrieval phase*, where the network dynamics is ongoing but synaptic plasticity is not present. Therefore, by compartmentalizing in time dynamics and learning, the network dynamics does not lead to changes in the synaptic weights during retrieval, and conversely, changes in synaptic weights do not lead to changes in the dynamics during learning. This separation prevents the observed runaway of the synaptic weights due to unsupervised Hebbian plasticity.

However, it is unclear whether such compartmentalization exists in cortical networks. In this work, we explored the alternative scenario, in which both plasticity and dynamics happen concurrently during learning and retrieval (see also Mongillo et al. (2005), Litwin-Kumar & Doiron (2014*a*), Zenke et al. (2015) for a similar approach in networks of spiking neurons). We found that adding multiplicative homeostatic plasticity to unsupervised Hebbian plasticity leads to stable learning of PA and SA. During sequential stimulation, the increase in co-activation between multiple populations due to recurrent and feed-forward connections learned via unsupervised Hebbian plasticity is prevented by suppressing its effect in the network dynamics. Homeostatic plasticity scales down the overall connectivity producing a weakly connected network. PA and SA is prevented to occur during stimulation, which weakens the positive feed-back loop generated by the increase in co-activations of neuronal populations. After learning, the dynamic variables of the Homeostatic plasticity rule reach a steady state with values similar of what they where before stimulation (see Fig 2.8 A) and the connectivity learned via unsupervised Hebbian plasticity can lead to retrieval of PA and SA upon stimulation (see Fig 2.8 C). The homeostatic variable reaches its steady state at a value close to one, and the connectivity recovers, *unmasking* the feed-forward and recurrent learned architecture. We have also tried other stabilization mechanisms such as inhibitory to excitatory plasticity (Vogels et al. 2011) instead of homeostatic plasticity. In this case we found that stable learning of PA and SA is possible, but for distinct sets of network and stimulation parameters (data not shown).

As explained in Zenke & Gerstner (2017), Zenke et al. (2017), in order to prevent the runaway of the synaptic weights produced by Hebbian plasticity, the time-scale of any compensatory mechanism should be of the same order or faster than the Hebbian time-scale. For multiplicative homeostatic plasticity, the time-scale of the homeostatic variable H_i is dependent on the firing rate of neuron i and the target firing rate (i.e. $\phi(u_i)/\phi(u_0)$). When the network firing rate is close to the target firing rate the homeostatic learning rule is slow, and

the homeostatic mechanism seldom play a role in the dynamics. On the other hand, for high firing rates the homeostatic plasticity time-scale becomes faster, preventing the runaway of the synaptic weights. There is currently an ongoing debate about whether the time-scales of compensatory processes used in theoretical studies, as the ones used here, are consistent with experimental evidence (see e.g. Zenke & Gerstner (2017), Zenke et al. (2017)).

2.6 Appendix

2.6.1 Parameters values

For the networks with fixed connectivity the parameters used in Fig C.1 and 2.3 are summarized in the Table 2.1 and 2.2 respectively. For networks with plastic connectivity the parameters used in Fig 2.4-2.10 are summarized in Table 2.3. The sequential stimulation parameters used in Fig 2.5-2.10 are summarized in Table 2.4.

2.6.2 Bifurcation diagram for a network of excitatory neurons with recurrent and feed-forward connectivity

Let us consider a network composed of an arbitrary number of excitatory populations. For tractability we will use the piecewise linear transfer function, see Eq (2.6). We want to study the conditions for an initial stimulus to the first population to: 1) propagate throughout the network without decaying; 2) grow until all populations are active at its maximum firing rate; 3) decay. Now consider a stimulus to the network such that all the populations are inactive, except the first (i.e. $\theta \leq u_1(0) \leq u_c$ and $u_j(0) = 0 \quad \forall j \neq 1$). For the first population, the dynamics reads

$$u_1(t) = \frac{w\nu\theta}{w\nu - 1} + \left(u_1(0) - \frac{w\nu\theta}{w\nu - 1} \right) e^{-\frac{(1-w\nu)}{\tau}t}.$$

Table 2.1: **Parameters used in Fig C.1.**

	(B)	**(C) I**	**(C) II**	**(C) III**
n	14	20	20	20
w	-	0.05	0.35	0.25
s	-	0.6	0.6	0.45
τ	10ms	10ms	10ms	10ms
w_{EI}	-	-	0.08	-
w_{IE}	-	-	1	-
τ_I	-	-	5ms	-
β	-	-	-	0.8
τ_a	-	-	-	80ms
a	-	6	6	6
b	-	-0.25	-0.25	-0.25
ν	1	2	2	2
θ	0	0	0	0
u_c	1	0.5	0.5	0.5
$\tilde{\nu}$	-	0.8	0.8	0.8
$\tilde{\theta}$	-	-0.1	-0.1	-0.1
$\tilde{u}_c$	-	0.5	0.5	0.5

Table 2.2: **Parameters used in Fig 2.3.** With $n = 10$ and $w_I = 2$.

	dSA	**SA**	**SA/PA**	**PA (bottom)**	**PA (top)**
w	0.2	1.01	1.3	1.9	1.3
s	1.5	1.39	1.	0.7	0.15

Assuming $\theta = 0$ and $u_1(0) = 1$, and defining

$$\begin{aligned} a &\equiv \frac{w\nu - 1}{\tau} \\ b &\equiv \frac{s\nu}{\tau}, \end{aligned}$$

we first compute the dynamics for the linear range of the transfer function, that is, assuming that the dynamics elicited is kept within the interval (θ, u_c) for the first K populations (i.e.

Table 2.3: **Network parameters used in Fig 2.4-2.10.** *Values of all the entries in the initial matrix $\mathbb{W}$.

	Fig 2.4	**Fig 2.5**	**Fig 2.6**	**Fig 2.7**	**Fig 2.8**	**Fig 2.9**	**Fig 2.10**
n	10	10	10	10	10	20	20
w_I	1	1	1	1	4.3	1	1
w_{max}	1.5	1.5	2.5	2.5	1.45	2.4	2.4
T_w	-	400ms	400ms	400ms	400ms	400ms	400ms
$\mathbb{W}_{i,j}(0)$*	-	0.1	0.1	0.2	0.1	0.1	0.1
a_{pre}	10	10	10	10	10	10	10
b_{pre}	0.7	0.7	0.7	0.7	0.7	0.7	0.7
a_{post}	10	10	10	10	10	10	10
b_{post}	0.7	0.7	0.7	0.7	0.7	0.7	0.7
D	-	15.3ms	15.3ms	15.3ms	15.3ms	15.3ms	15.3ms
r_w	0.6	0.6	0.6	0.6	0.6	0.6	0.6
r_0	-	-	-	-	0.05	0.01	0.01
τ_H	-	-	-	-	20s	20s	20s
$\sum_{j=1}^{n} \mathbb{W}_{i,j}$	-	-	-	2	-	-	-

populations k such that $k \leq K \leq n$). The populations dynamics is given by

$$\tau \frac{du_j}{dt} = -u_j + w\nu u_j + s\nu u_{j-1} \qquad j \leq K, \tag{2.19}$$

which leads to an analytical solution for the dynamics of each population

$$u_{k+1}(t) = be^{-at} \int_0^t u_k(t')e^{at'}dt' \qquad k \leq K.$$

Using an inductive argument we obtain that

$$u_{k+1}(t) = \frac{(bt)^{k-1}}{(k-1)!}e^{at} \qquad k = 1, \ldots, K, \tag{2.20}$$

and in the limit $k \to \infty$ and with the condition $a < b$, we obtain

$$\lim_{k\to\infty} u_{k+1}(t) = e^{at} \lim_{k\to\infty} \frac{(bt)^{k-1}}{(k-1)!}.$$

Table 2.4: **Stimulation parameters Fig 2.4-2.10.**

	I	T (ms)	Δ (ms)
Fig 2.5 (top-left)	1.25	18	8
Fig 2.5 (top-right)	1.8	20.5	80
Fig 2.5 (bottom-left)	1.8	7	9
Fig 2.5 (bottom-right)	1.8	1	50
Fig 2.6	1.3	19	10
Fig 2.7	2.2	19	10
Fig 2.8 A-C and D (left)	3.5	14	7
Fig 2.9	5	40	8
Fig 2.10 (top-left)	5.5	7	14
Fig 2.10 (top-right)	5.5	50	40
Fig 2.10 (bottom-left)	5.5	5	13
Fig 2.10 (bottom-right)	5.5	20	8.5

Since $b > 0$, we have that

$$\lim_{k\to\infty} \frac{(bt)^{k-1}}{(k-1)!} \leq \lim_{k\to\infty} \sum_{l=0}^{k} \frac{(bt)^l}{l!} = e^{bt},$$

then

$$\lim_{k\to\infty} u_k(t) \leq e^{(a+b)t}.$$

For the sequence to decay away, it is sufficient to impose that

$$w + s < \frac{1}{\nu},$$

since populations will receive inputs from previous populations that decrease with the position in the feed-forward connectivity. On the other hand, for $w + s, w > \frac{1}{\nu}$ the activity of $0 \leq p \leq n$ populations depending on the initial condition will grow until they reach an state in which they are active at their maximum firing rate. However, for $w + s > \frac{1}{\nu} > w$, the dynamics for first population decays exponentially towards zero

$$u_1(t) \approx e^{at},$$

but for the rest $n-1$ populations any input will lead to an increase in their activity since $w+s > \frac{1}{\nu}$. In this regime, after the stimulation of the first population, the feed-forward input from the first population to the second transiently decreases, producing a transient increase in firing rate of the second population. This produces a sequence of transient increase in activity along the feed-forward connectivity. Importantly, neurons later in the connectivity will received feed-forward inputs for longer times. To see that, let us consider the fact that the r.h.s. of Eq. (2.20) always has a maximum. Its maximum is achieved when $t_{max} = -k/a$. The value of the maximum is

$$u_{k+1}(t_{max}) = \frac{(bk)^k}{(-a)^k k!} e^{-k}.$$

For $k \to \infty$ we can use the Stirling approximation for the factorial, obtaining

$$u_{k+1}(t_{max}) \sim \sqrt{2\pi k} e^{k(\log(kb/(-a)) - \log(k/e) - 1)}, \tag{2.21}$$

which is equivalent to

$$u_{k+1}(t_{max}) \sim \sqrt{2\pi k} \left(-\frac{b}{a}\right)^k \qquad \text{for } k \to \infty.$$

Then, for $a < 0$ and $-b/a > 1$ (i.e. $w + s > \frac{1}{\nu} > w$) populations whose activities are in the linear range of the transfer function (i.e. $k \leq K$) present an increasing maximum activity with its position in the feed-forward connectivity. Which implies an increasingly stronger feed-forward input with the population's position. If we standarize the population's activity $u_k(t)$ with the total area under the dynamics (i.e. $\int_0^\infty dt u_k(t) = \frac{b^{k-1}}{(-a)^k}$) defining $\tilde{u}_k(t) \equiv u_k(t) / \left(\frac{b^{k-1}}{(-a)^k}\right)$, then an approximation for the time that a population is active in the sequence is

$$\sqrt{\int_0^\infty t^2 \tilde{u}_k(t) - \left(\int_0^\infty dt \tilde{u}_k(t) t\right)^2} = -\frac{1}{a}\sqrt{k} = \frac{\tau}{1 - w\nu}\sqrt{k}. \tag{2.22}$$

Therefore, the time that a population is active in the sequence scales with the squared root of the position of the population in the feed-forward connectivity (i.e. $\sqrt{k}$, for $k \leq K$). In fact, we found that this scaling also holds for populations whose activities are larger than the upper bound of the transfer function (i.e. u_c), see Fig C.1.

2.6.3 Instantaneous shared inhibition approximation

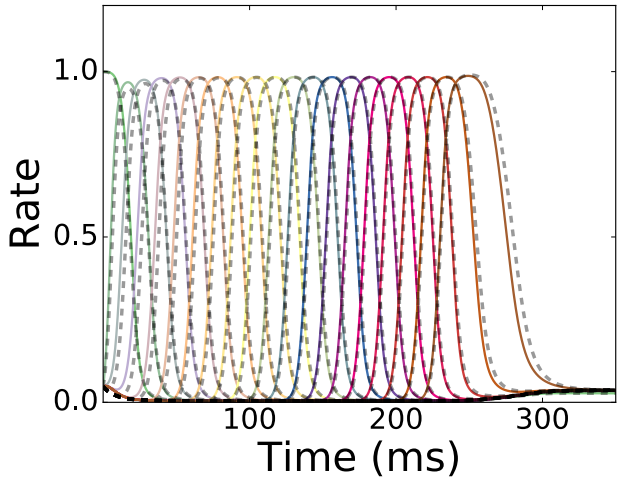

Figure 2.12: **Shared Inhibition vs Instantaneous shared inhibition approximation**. In color lines it is shown SA generation for a network of 20 neurons with fixed feed-forward and recurrent connections and shared inhibition. In grey dashed lines it is shown the instantaneous shared inhibition approximation.

2.6.4 Bifurcation diagram for a network of excitatory neurons with recurrent and feed-forward connections and shared Inhibition

Here we derive the PA-PA/SA, PA-(SA or dSA), and PA/SA-(SA or dSA) boundaries in the bifurcation diagram of a network of excitatory neurons with recurrent and feed-forward connections and shared Inhibition. Let us consider Eq. (2.3) on the main text for the

dynamics of a network of excitatory populations with shared inhibition. In addition, let us assume a piecewise linear transfer function such that $0 \leq \phi(u) \leq 1$ and $u_c = 1$. For the beginning of the analysis let us assume that $s = 0$ and m populations are in a high rate state, that is $\phi(u_k) = 1$, whereas all the rest are in the low activity state, i.e. $\phi(u_j) = 0 \quad j \neq k$. Then if $\phi(u_{k_j}) = 1$ for $k_1, k_2, \cdots, k_m$, a necessary condition for the high rate state to be a fixed point is

$$u_{k_j} = w - \frac{mw_I}{n} > 1 \quad j = 1, \ldots, m. \tag{2.23}$$

If we only want to have at most m populations in the high rates state as a fixed point then

$$u_{k_j} = w - \frac{(m+1)w_I}{n} < 1 \quad j = 1, \ldots, m, \tag{2.24}$$

which implies

$$1 + \frac{mw_I}{n} \leq w \leq 1 + \frac{(m+1)w_I}{n}. \tag{2.25}$$

Let us now consider $s \neq 0$. If we want to have at most r contiguous populations connected via feed-forward connections in the high rate, a necessary condition is that the first population in the architecture needs to be able to sustain PA when r populations are active, i.e.

$$u_k = w - \frac{rw_I}{n} > 1. \tag{2.26}$$

The second necessary condition is that the last of the r population *down stream* in the recurrent-feed-forward connected architecture does not die out to the low rate state due to inhibition, when this population is in the low activity state and the population before is in high rate state, i.e.

$$u_{k+r-1} = s - \frac{(r-1)w_I}{n} > 0. \tag{2.27}$$

If we want at most r populations active, we need to impose that the population right

after remains in the low rate state

$$u_{k+r} = s - \frac{rw_I}{n} < 0, \tag{2.28}$$

which is equivalent to

$$\frac{(r-1)w_I}{n} \leq s \leq \frac{rw_I}{n}. \tag{2.29}$$

Then if w fulfills Eq (2.25) and we stimulate all the populations of the network, we have that at most m populations remain active. On the other hand, if Eqs. (2.26,2.27) hold, and we stimulate just the first population in the network then at least r contiguous populations remain active in PA. Lastly if we activate the first population and all the rest are in the low activity state, and

$$\begin{aligned} &\frac{rw_I}{n} \leq s \\ &1 + \frac{rw_I}{n} \leq w \leq 1 + \frac{(r+1)w_I}{n}. \end{aligned} \tag{2.30}$$

Consequently, the next r populations go to the high rate activity state. Due to shared inhibition, the first population decreases to the low rate state, since Eqs. (2.30) holds and it is the population that receives less current because it lacks feed-forward inputs. The decrease in the shared inhibitory input due to the activity decay of the first population leads to the $(r+2)^{\text{th}}$ population to increase its activity toward a high rate activity state. This consequently produces that the second population decay to the low rate state due to a new increase in shared inhibition. This process iterates producing a sequence that stabilizes when the last population and the $r-1$ populations before this one are in the high rate state. We call SA/PA to this sequential activity that ends in persistent activity.

2.6.5 Multiplicative homeostatic plasticity

Brief discussion

As is mentioned in the main text, the overall steady state connectivity (i.e. $\mathbf{W}$) is very sensitive to differences in values of the connectivity learned via Hebbian plasticity (i.e. $\mathbb{W}$) when a linear version of homeostatic learning rule used in the main text (i.e. Eq. (2.15) on the main text without the quadratic term in H) is used. This can be intuitively understood analyzing a one population network with connectivity strength w. When the linear version of homeostatic plasticity is used the H nullcline is vertical (see Fig 2.13 A). This produces that slight changes in the connectivity strength of the excitatory population dramatically change the value of the steady state homeostatic variable (fixed point of the dynamics). If a quadratic nonlinearity is included in the learning rule, the H nullcline is now a straight line with slope $-\frac{1}{u_0}$ and intercept 1 (see Fig 2.13 B). As a consequence, the steady state homeoestatic variable is close to 1 (i.e. $H \approx 1$) provided w is not very large (see Fig 2.13 C). This analysis is generalized for an arbitrary number of excitatory populations undergoing sequential stimulation in the next section. Therefore, the linear version of the homeostatic plasticity rule, in general leads to an steady state connectivity uniformly strong disregarding the sequential stimulation parameters (see Fig 2.13 F). And even though Hebbian plasticity ensures non-uniformity in the learned connectivity (see Fig 2.13 D and F), very strong recurrent and feed-forward connections due high values for the homeostatic variables usually leads to the retrieval of PA when the network is perturbed from the background state (see Fig 2.3 main text). Preventing with this that differences in connectivities learned due to different stimuli to be reflected in different learned dynamics (i.e. PA, SA, PA/SA or dSA). In practice, PA is initially retrieved almost always independent of the type of stimulation presented when the linear version of the homeostatic learning rule leads to large values of the homeostatic variable. Additionally, strong excitatory connectivity due to large values of

the homeostatic variables might produce *spikes* on the homeostatic variables and synaptic weights learned via Hebbian plasticity uniformly shooting down the overall connectivity (see Fig 2.13 D-F). These phenomena can be interpreted as a *forgetting* of the stimulus learned during sequential stimulation via Hebbian plasticity, since is prevented the retrieval of any temporal attractor other than PA or the background state. If a quadratic term is introduced (as in Eq. (2.15) on the main text) and: 1) the recurrent and feed-forward connections learned via Hebbian plasticity are not large; 2) the background activity is within the sub-linear region of the transfer function. Then $H_i \approx 1$ and the connectivity in the steady state is approximately the connectivity learned via Hebbian plasticity (i.e. $\mathbf{W}_{i,j} \approx \mathbb{W}_{i,j}$). In the next section we provide a mathematical proof to this assertion.

Mathematical analysis

Let us assume that after r repetitions of the sequential stimulation described in the main text, there is an increase in the synaptic weights to a final value $\mathbb{W}_{i,j}^{(r)}$ due to Hebbian plasticity. We will also assume that the sequential stimulation is in the range of parameters for T and Δ where only recurrent and feed-forward connections increase due stimulation. Using the plasticity rule proposed in Toyoizumi et al. (2014) (i.e. linear version of Eq. (9) on the main text) the fixed points for the network's dynamics after r sequential stimulation are given by

$$
\begin{aligned}
u_i^* &= u_0 \qquad i = 1, \ldots, n \\
H_1^{(r)} &= \frac{u_0}{\phi(u_0)\mathbb{W}_{1,1}^{(r)} - w_{EI}\phi(u_I^*)} \\
H_i^{(r)} &= \frac{u_0}{\phi(u_0)(\mathbb{W}_{i,i}^{(r)} + \mathbb{W}_{i-1,i}^{(r)}) - w_{EI}\phi(u_I^*)} \qquad i = 2, \ldots, n \\
u_I^* &= \phi(u_I^*) + \phi(u_0) n w_{IE}.
\end{aligned}
$$

Where the target firing rate is $r_0 = \phi(u_0)$. Then, the connectivity matrix for the excitatory populations is given by

$$\begin{aligned}
\mathbf{W}_{1,1}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1-\frac{w_{EI}\phi(u_I^*)}{\mathbb{W}_{1,1}^{(r)}\phi(u_0)}}\right) \\
\mathbf{W}_{i,i}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1+\frac{\mathbb{W}_{i-1,i}^{(r)}}{\mathbb{W}_{i,i}^{(r)}}-\frac{w_{EI}\phi(u_I^*)}{\mathbb{W}_{i,i}^{(r)}\phi(u_0)}}\right) \qquad i=2,\ldots,n \\
\mathbf{W}_{i-1,i}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1+\frac{\mathbb{W}_{i,i}^{(r)}}{\mathbb{W}_{i-1,i}^{(r)}}-\frac{w_{EI}\phi(u_I^*)}{\mathbb{W}_{i-1,i}^{(r)}\phi(u_0)}}\right) \qquad i=2,\ldots,n.
\end{aligned}$$

Considering that

$$\begin{aligned}
\frac{w_{EI}\phi(u_I^*)}{\mathbb{W}_{i-1,i}^{(r)}\phi(u_0)} &= \frac{w_{EI}}{\mathbb{W}_{i-1,i}^{(r)}}\left(\frac{u_I^*}{\phi(u_0)}-nw_{IE}\right) \\
&= -\frac{nw_{EI}w_{IE}}{\mathbb{W}_{i-1,i}^{(r)}}\left(1-\frac{u_I^*}{n\phi(u_0)w_{IE}}\right) \\
&= -\frac{w_I}{\mathbb{W}_{i-1,i}^{(r)}}\left(1-\frac{u_I^*}{n\phi(u_0)w_{IE}}\right),
\end{aligned}$$

and assuming that $w_{EI} \sim O\left(\frac{1}{\sqrt{n}}\right)$ and $w_{IE} \sim O\left(\frac{1}{\sqrt{n}}\right)$, we have that for large n

$$\frac{w_{EI}\phi(u_I^*)}{\mathbb{W}_{i-1,i}^{(r)}\phi(u_0)} \approx -\frac{w_I}{\mathbb{W}_{i-1,i}^{(r)}}. \tag{2.31}$$

Which leads to

$$\begin{aligned}
\mathbf{W}_{1,1}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1+\frac{w_I}{\mathbb{W}_{i-1,i}^{(r)}}}\right) \\
\mathbf{W}_{i,i}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1+\frac{\mathbb{W}_{i-1,i}^{(r)}}{\mathbb{W}_{i,i}^{(r)}}+\frac{w_I}{\mathbb{W}_{i-1,i}^{(r)}}}\right) \qquad i=2,\ldots,n \\
\mathbf{W}_{i-1,i}^{(r)} &= \left(\frac{u_0}{\phi(u_0)}\right)\left(\frac{1}{1+\frac{\mathbb{W}_{i,i}^{(r)}}{\mathbb{W}_{i-1,i}^{(r)}}+\frac{w_I}{\mathbb{W}_{i-1,i}^{(r)}}}\right) \qquad i=2,\ldots,n.
\end{aligned}$$

Assuming that the sequential stimulation parameters are such that the recurrent and feed-forward connections learned have the same order of magnitude

$$\frac{\mathbb{W}_{i,i}^{(r)}}{\mathbb{W}_{i-1,i}^{(r)}} \sim O(1),$$

we obtain

$$\mathbf{W}_{i-1,i}^{(r)} \sim \mathbf{W}_{i,i}^{(r)} \sim O\left(\frac{u_0}{\phi(u_0)}\right). \tag{2.32}$$

Therefore, after sequential stimulation the connectivity matrix is weakly dependent of the synaptic weights learned via Hebbian plasticity and proportional to the quotient of the target firing rate synaptic input current u_0 and the corresponding target firing rate $r_0 = \phi(u_0)$. If we now consider the homeostatic learning rule in Eq. (2.15) on the main text, we have that

the fixed points for the dynamics of the network are given by

$$u_1^* = \frac{\mathbb{W}_{1,1}^{(r)}\phi(u_1^*)\left(u_0 + w_{EI}\phi(u_I^*)\right)}{u_0 + \mathbb{W}_{1,1}^{(r)}\phi(u_1^*)} - w_{EI}\phi(u_I^*) \tag{2.33}$$

$$u_i^* = \frac{\left(\mathbb{W}_{i,i}^{(r)}\phi(u_i^*) + \mathbb{W}_{i+1,i}^{(r)}\phi(u_{i-1}^*)\right)\left(u_0 + w_{EI}\phi(u_I^*)\right)}{u_0 + \mathbb{W}_{i,i}^{(r)}\phi(u_i^*) + \mathbb{W}_{i+1,i}^{(r)}\phi(u_{i-1}^*)} - w_{EI}\phi(u_I^*) \quad i = 2, \ldots, N \tag{2.34}$$

$$H_1^{(r)} = \frac{1 + w_{EI}\frac{\phi(u_I^*)}{u_0}}{1 + \mathbb{W}_{1,1}^{(r)}\frac{\phi(u_1^*)}{u_0}} \tag{2.35}$$

$$H_i^{(r)} = \frac{1 + w_{EI}\frac{\phi(u_I^*)}{u_0}}{1 + \mathbb{W}_{i,i}^{(r)}\frac{\phi(u_i^*)}{u_0} + \mathbb{W}_{i+1,i}^{(r)}\frac{\phi(u_{i-1}^*)}{u_0}} \qquad i = 2, \ldots, N \tag{2.36}$$

$$u_I^* = \phi(u_I^*) + w_{IE}\sum_{j=1}^{N}\phi(u_j^*). \tag{2.37}$$

Then, using approximation in Eq. (2.31), the connectivity matrix for the excitatory populations is

$$\begin{aligned}
\mathbf{W}_{1,1}^{(r)} &= \frac{\mathbb{W}_{1,1}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right)}{1 + \mathbb{W}_{1,1}^{(r)}\frac{\phi(u_1^*)}{u_0}} \\
\mathbf{W}_{i,i}^{(r)} &= \frac{\mathbb{W}_{i,i}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right)}{1 + \mathbb{W}_{i,i}^{(r)}\frac{\phi(u_i^*)}{u_0} + \mathbb{W}_{i+1,i}^{(r)}\frac{\phi(u_{i-1}^*)}{u_0}} \qquad i = 2, \ldots, n \\
\mathbf{W}_{i+1,i}^{(r)} &= \frac{\mathbb{W}_{i+1,i}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right)}{1 + \mathbb{W}_{i,i}^{(r)}\frac{\phi(u_i^*)}{u_0} + \mathbb{W}_{i+1,i}^{(r)}\frac{\phi(u_{i-1}^*)}{u_0}} \qquad i = 2, \ldots, n.
\end{aligned}$$

For large values of the learned recurrent and feed-forward connections (i.e. $\mathbb{W}_{i,i}^{(r)}, \mathbb{W}_{i+1,i}^{(r)} \to \infty$), the fixed point for each neuron currents becomes $u_i^* = u_0 \quad i = 1, 2, \ldots, n$, see Eq (2.33). Assuming that as $\mathbb{W}_{i,i}^{(r)}/\mathbb{W}_{i+1,i}^{(r)} \to \alpha^{(r)} < \infty$ for all r. Then, when the recurrent and feed-forward connections learned via Hebbian plasticity are large, the overall steady state

connectivity is approximated by

$$\begin{aligned}
\mathbf{W}_{1,1}^{(r)} &\approx \frac{u_0}{\phi(u_0)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \\
\mathbf{W}_{i,i}^{(r)} &\approx \frac{u_0}{\phi(u_0)\left(1+\frac{1}{\alpha^{(r)}}\right)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \qquad i = 2,\ldots,n \\
\mathbf{W}_{i+1,i}^{(r)} &\approx \frac{u_0}{\phi(u_0)(1+\alpha^{(r)})}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \qquad i = 2,\ldots,n.
\end{aligned}$$

In this limit, as in the linear version of the homeostatic learning rule analyzed above, the recurrent and feed-forward synaptic weights learned are scaled by the $\frac{u_0}{\phi(u_0)}$ quotient. This means that the final connectivity after learning is not strongly dependent of the history of stimulation. However, for weak synaptic weights learned (i.e. $\mathbb{W}_{i,i}^{(r)}, \mathbb{W}_{i+1,i}^{(r)} \ll \frac{u_0}{\phi(u_0)}$) we have that at the lowest order

$$\begin{aligned}
\mathbf{W}_{1,1}^{(r)} &\approx \mathbb{W}_{1,1}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \\
\mathbf{W}_{i,i}^{(r)} &\approx \mathbb{W}_{i,i}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \qquad i = 2,\ldots,N \\
\mathbf{W}_{i+1,i}^{(r)} &\approx \mathbb{W}_{i+1,i}^{(r)}\left(1 - w_I\frac{\phi(u_0)}{u_0}\right) \qquad i = 2,\ldots,N
\end{aligned}$$

If we assume that target synaptic input current of the homeostatic learning rule (i.e. u_0) is within the sub-linear region of the transfer function, then $w_I\frac{\phi(u_0)}{u_0} \ll 1$. This leads to

$$\begin{aligned} \mathbf{W}^{(r)}_{1,1} &\approx \mathbb{W}^{(r)}_{1,1} \\ \mathbf{W}^{(r)}_{i,i} &\approx \mathbb{W}^{(r)}_{i,i} \qquad i = 2, \ldots, n \\ \mathbf{W}^{(r)}_{i+1,i} &\approx \mathbb{W}^{(r)}_{i+1,i} \qquad i = 2, \ldots, n. \end{aligned} \tag{2.38}$$

Therefore, we conclude that when recurrent and feed-forward connections learned via Hebbian plasticity are such that $\mathbb{W}^{(r)}_{i,i}, \mathbb{W}^{(r)}_{i+1,i} \ll \frac{u_0}{\phi(u_0)}$ and the background activity u_0 is within the sub-linear region of the transfer function. Then the steady state overall connectivity in a network with both Hebbian and homeostatic plasticity (i.e.$\mathbf{W}$) is approximately the connectivity learned via Hebbian plasticity (i.e.$\mathbb{W}$).

2.6.6 Approximation for the synaptic weights dynamics during repeated sequential stimulation

In this section we obtain an approximation for the synaptic weights dynamics during the sequential stimulation protocol for a network with Hebbian and homeostatic plasticity. First we will approximate the increase in the synaptic weights after a single stimulation by approximating the time that the neuron's current u_i is above the learning threshold $u_w \equiv \phi^{-1}(r_w)$. During sequential stimulation protocol the effective connectivity is very weak due to the homeostatic plasticity (i.e. $\mathbf{W}_{i,j} = H_i \mathbb{W}_{i,j} \ll 1$). Then, neglecting the effect of inhibition, the dynamics of each population can be approximated by

$$\tau \dot{u}_i \approx I - u_i \quad i = 1, \ldots, n.$$

During the stimulation period T populations dynamics reads

$$u_i(t) = I\left(1 - e^{-\frac{t}{\tau}}\right) \quad t \in [0, T].$$

Its final value right after the stimulation is

$$u_i(T) = I(1 - e^{-\frac{T}{\tau}}),$$

and after the stimulation the population current decays as

$$u_i = u_i(T)e^{-\frac{t}{\tau}} = I(1 - e^{-\frac{T}{\tau}})e^{-\frac{t}{\tau}}.$$

Then the approximate time that takes each population to reach the learning threshold from resting is

$$\tau_{u_0,u_w} \equiv -\tau \ln\left(1 - \frac{u_w}{I}\right). \tag{2.39}$$

On the other hand the approximate time that takes to each population to decay to the learning threshold from its maximum activity after stimulation ($u_{max} = I(1-e^{-\frac{T}{\tau}})$) is given by

$$u_w = I(1 - e^{-\frac{T}{\tau}})e^{-\frac{\tau_{u_{max},u_w}}{\tau}},$$

which leads to

$$\tau_{u_{max},u_w} = -\tau \ln\left(\frac{u_w}{I\left(1 - e^{-\frac{T}{\tau}}\right)}\right). \tag{2.40}$$

Hence, an approximation for the time that each population spends above the learning threshold is

$$\tau_{u_w} = T - \tau_{u_0,u_w} + \tau_{u_{max},u_w} \tag{2.41}$$

The population dynamics above the plasticity threshold u_w when is stimulated at time

t_k can be approximated by:

$$u_i(t-t_k) = \begin{cases} I\left(1-e^{-\frac{t-t_k}{\tau}}\right) & t-t_k \in [\tau_{u_0,u_w}, T] \\ I(1-e^{-\frac{T}{\tau}})e^{-\frac{t-t_k}{\tau}} & t-t_k \in [T, \tau_{u_w}] \end{cases} \tag{2.42}$$

Let us first consider the increase in the recurrent connections. First define $\tilde{t}_k \equiv t - t_k$ and

$$\Omega_{i,j}(a,b) = f[\phi(u_i(a)))]g[\phi(u_j(b-D))].$$

In order to compute the increment in the synaptic weight $\mathbb{W}_{i,i}$ we need to solve

$$\frac{d\mathbb{W}_{i,i}}{d\tilde{t}_k} = \frac{\Omega_{i,i}(\tilde{t}_k, \tilde{t}_k - D) - \mathbb{W}_{i,j}}{\tau_w}, \tag{2.43}$$

for

$$\tilde{t}_k \in [D + \tau_{u_0,u_w}, T + \tau_{u_{max},u_w}].$$

We obtain

$$\mathbb{W}_{i,i}(\tilde{t}_k) = e^{-\frac{\tilde{t}_k - D - \tau_{u_0,u_w}}{\tau_w}} \left(\mathbb{W}_{i,i}(D + \tau_{u_0,u_w}) + \frac{1}{\tau_w} \int_{D+\tau_{u_0,uw}}^{\tilde{t}_k} dt \Omega_{i,i}(t_k, t_k - D) e^{\frac{t-D-\tau_{u_0,u_w}}{\tau_w}} \right), \tag{2.44}$$

for

$$\tilde{t}_k \in [D + \tau_{u_0,u_w}, T + \tau_{u_{max},u_w}].$$

To compute the increase on feed-forward connections after one stimulation, let us define $\tilde{t}_k^i \equiv t - t_k^i$ as the time elapsed after stimulation of neuron i. Due to the nature of the sequential stimulation, we have the following relation $\tilde{t}_k^{i+1} = \tilde{t}_k^i - T - \Delta$. Then in order to

compute the increment in the synaptic weight $\mathbb{W}_{i+1,i}$ we need to solve the following equation

$$\frac{d\mathbb{W}_{i+1,i}}{d\tilde{t}_k^{i+1}} = \frac{\Omega_{i+1,i}(\tilde{t}_k^{i+1}, \tilde{t}_k^i - D) - \mathbb{W}_{i,j}}{\tau_w}, \tag{2.45}$$

for

$$\begin{aligned} \tilde{t}_k^{i+1} &\in [\tau_{u_0,u_w}, T + \tau_{u_{max},u_w}] \\ \tilde{t}_k^i &\in [D + T + \Delta, D + T + \Delta + \tau_{u_{max},u_w}]. \end{aligned}$$

Considering that the upper boundary for $\tilde{t}_k^{i+1}$ should be such that:

$$\text{Max}(\tilde{t}_k^{i+1}) + T + \Delta - D = T + \tau_{u_{max},u_w}$$

we obtain

$$\frac{d\mathbb{W}_{i+1,i}}{d\tilde{t}_k^{i+1}} = \frac{\Omega_{i+1,i}(\tilde{t}_k^{i+1}, \tilde{t}_k^{i+1} + T + \Delta - D) - \mathbb{W}_{i+1,i}}{\tau_w}, \tag{2.46}$$

for

$$\tilde{t}_k^{i+1} \in [\tau_{u_0,u_w}, D + \tau_{u_{max},u_w} - \Delta].$$

We then obtain the following approximation for the dynamics of the feed-forward connections

$$\begin{aligned} \mathbb{W}_{i+1,i}(\tilde{t}_k^{i+1}) &= e^{-\frac{\tilde{t}_k^{i+1} - \tau_{u_0,u_w}}{\tau_w}} \mathbb{W}_{i+1,i}(\tau_{u_0,u_w}) \\ &+ \frac{e^{-\frac{\tilde{t}_k^{i+1} - \tau_{u_0,u_w}}{\tau_w}}}{\tau_w} \int_{\tau_{u_0,u_w}}^{\tilde{t}_k^{i+1}} dt \Omega_{i+1,i}(t, t + T + \Delta - D) e^{\frac{t - \tau_{u_0,u_w}}{\tau_w}} \end{aligned} \tag{2.47}$$

for

$$\tilde{t}_k^{i+1} \in [\tau_{u_0,u_w}, D + \tau_{u_{max},u_w} - \Delta].$$

These approximations appear to be accurate for the dynamics during stimulation, as Fig 2.9 shows. Using Eqs. (2.44,2.48), we can write iterative equations for the final recurrent and feed-forward synaptic weights for sequential stimulation $k+1$ as

$$\frac{\mathbb{W}_{i,i}^{k+1}}{e^{-\frac{T+\tau_{umax,uw}-D-\tau_{u0,uw}}{\tau_w}}} = \mathbb{W}_{i,i}^{k} + \frac{\int_{D+\tau_{u0,uw}}^{T+\tau_{umax,uw}} dt\Omega_{i,i}(t,t-D)e^{\frac{t-D-\tau_{u0,uw}}{\tau_w}}}{\tau_w} \tag{2.48}$$

$$\frac{\mathbb{W}_{i+1,i}^{k+1}}{e^{-\frac{D+\tau_{umax,uw}-\Delta-\tau_{u0,uw}}{\tau_w}}} = \mathbb{W}_{i+1,i}^{k} + \frac{\int_{\tau_{u0,uw}}^{D+\tau_{umax,uw}-\Delta} dt\Omega_{i+1,i}(t,t+T+\Delta-D)e^{\frac{t-\tau_{u0,uw}}{\tau_w}}}{\tau_w}, \tag{2.49}$$

where $\mathbb{W}_{i,i}^{k}$ and $\mathbb{W}_{i+1,i}^{k}$ are the recurrent and feed-forward connections after sequential stimulation k. Defining

$$\begin{aligned}
\Gamma_l^{rec} &\equiv D+\tau_{u0,uw} \\
\Gamma_u^{rec} &\equiv T+\tau_{umax,uw} \\
\Gamma_l^{ff} &\equiv \tau_{u0,uw} \\
\Gamma_u^{ff} &\equiv D+\tau_{umax,uw}-\Delta,
\end{aligned}$$

and iterating Eqs. (2.48,2.49) we obtain

$$\begin{aligned}
\mathbb{W}_{i,i}^{k+1} &= e^{-\frac{k(\Gamma_u^{rec}-\Gamma_l^{rec})}{\tau_w}}\mathbb{W}_{i,i}^{0} \\
&\quad + \left(\sum_{j=1}^{k} e^{-\frac{j(\Gamma_u^{rec}-\Gamma_l^{rec})}{\tau_w}}\right)\frac{1}{\tau_w}\int_{\Gamma_l^{rec}}^{\Gamma_u^{rec}} dt\Omega_{i,i}(t,t-D)e^{\frac{t-\Gamma_l^{rec}}{\tau_w}}
\end{aligned} \tag{2.50}$$

$$\begin{aligned}
\mathbb{W}_{i+1,i}^{k+1} &= e^{-\frac{k(\Gamma_u^{ff}-\Gamma_l^{ff})}{\tau_w}}\mathbb{W}_{i+1,i}^{0} \\
&\quad + \left(\sum_{j=1}^{k} e^{-\frac{j(\Gamma_u^{ff}-\Gamma_l^{ff})}{\tau_w}}\right)\frac{1}{\tau_w}\int_{\Gamma_l^{ff}}^{\Gamma_u^{ff}} dt\Omega_{i+1,i}(t,t+T+\Delta-D)e^{\frac{t-\Gamma_l^{ff}}{\tau_w}}.
\end{aligned} \tag{2.51}$$

For a large number of repetitions of the sequential stimulation (i.e. $k \to \infty$) the stationary recurrent and feed-forward connections are given by

$$\mathbb{W}^{\infty}_{i,i} = \frac{1}{\tau_w}\left(\frac{e^{-\frac{\Gamma^{rec}_u-\Gamma^{rec}_l}{\tau_w}}}{1-e^{-\frac{\Gamma^{rec}_u-\Gamma^{rec}_l}{\tau_w}}}\right)\int_{\Gamma^{rec}_l}^{\Gamma^{rec}_u} dt\Omega_{i,i}(t,t-D)e^{\frac{t-\Gamma^{rec}_l}{\tau_w}} \tag{2.52}$$

$$\mathbb{W}^{\infty}_{i+1,i} = \frac{1}{\tau_w}\left(\frac{e^{-\frac{\Gamma^{ff}_u-\Gamma^{ff}_l}{\tau_w}}}{1-e^{-\frac{\Gamma^{ff}_u-\Gamma^{ff}_l}{\tau_w}}}\right)\int_{\Gamma^{ff}_l}^{\Gamma^{ff}_u} dt\Omega_{i+1,i}(t,t+T+\Delta-D)e^{\frac{t-\Gamma^{ff}_l}{\tau_w}} \tag{2.53}$$

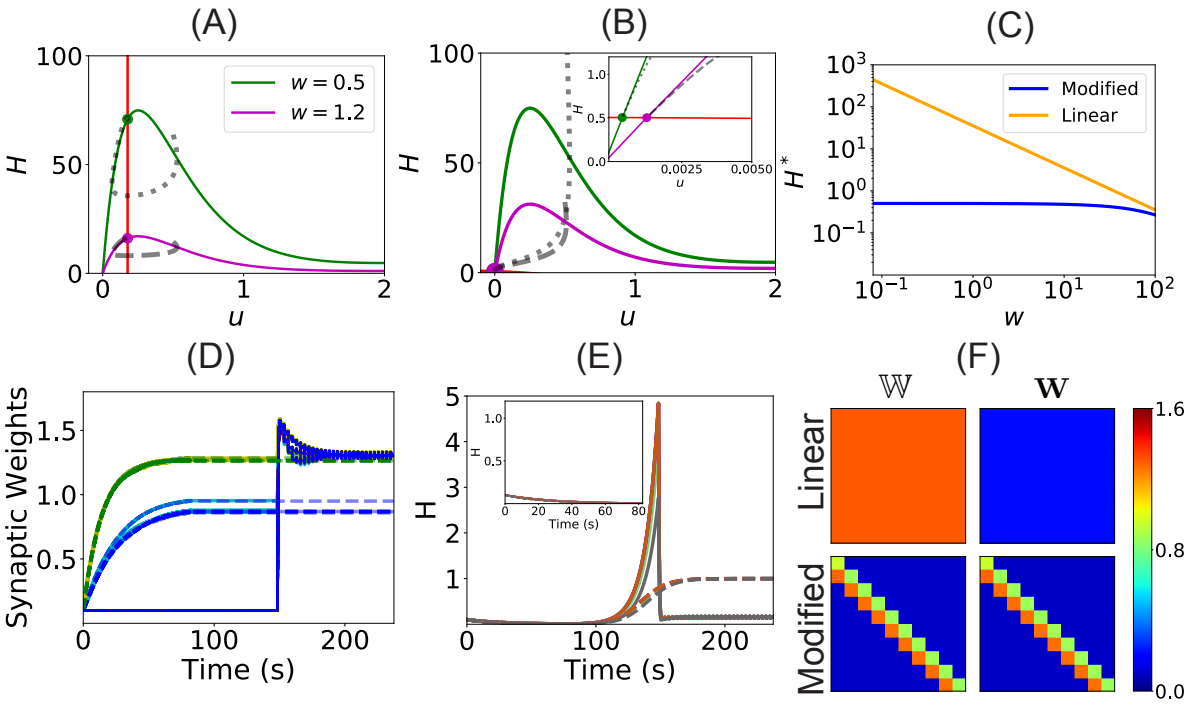

Figure 2.13: **Linear and nonlinear multiplicative homeostatic plasticity**. **(A and B)** Phase plane ($u - H$ plane) of a single population network with fixed recurrent connectivity (w), fast linear inhibition ($w_I = 0.1$) and homeostatic plasticity. Red: H nullcline. Green and purple: u nullcline for $w = \{0.5, 1.2\}$ respectively. Dashed lines show two orbits of the network's dynamics for a single initial condition with $w = 1.2$ and 0.5, respectively. The fixed points are indicated with solid circles using the same color code. In A, the linear version of the homeostatic learning rule is used (i.e. Eq. (2.15) main text without the quadratic term on the r.h.s.). In B, the homeostatic learning rule in Eq. (2.15) of the main text is used (nonlinear version). Inset of B shows a zoom in the region of the fixed points. **(C)** Stationary state value of H (fixed point of the dynamics, i.e. H such that $\dot{H} = 0$ and $\dot{u} = 0$) vs w for the single population network with fixed recurrent connectivity and fast linear inhibition studied in A and B. Orange: linear version of the homeostatic learning rule; Blue: nonlinear version. **(D-F)** Learning dynamics for a network with $n = 10$ populations, fast linear shared inhibition, Hebbian plasticity, linear or nonlinear homeostatic learning rule. The network is stimulated for the first 82s, and then the dynamics freely evolves toward its steady state. **(D)** Synaptic weights learned via Hebbian plasticity during stimulation (i.e. $\mathbb{W}$). Cyan: recurrent; Yellow: feed-forward; Red: feed-backward; Blue: feed-second-forward; Green: feed-second-backward connections. Solid and dashed lines correspond to a network with Hebbian plasticity plus the linear and nonlinear version of the homeostatic learning rule respectively. **(E)** Dynamics of the homeostatic variable. The color code is the one used in Fig 2.8. Solid and dashed lines correspond to the linear and nonlinear version of the homeostatic learning rule respectively. **(F)** Steady state connectivity matrix corresponding to the connectivity learned via Hebbian plasticity $\mathbb{W}$ (first column) and the overall connectivity $\mathbf{W}$ (second column) for the linear (first row) and nonlinear (second row) version of the homeostatic learning rule.

CHAPTER 3

ATTRACTOR DYNAMICS IN NETWORKS WITH LEARNING RULES INFERRED FROM *IN VIVO* DATA

3.1 Contribution

The work presented in this chapter correspond to the publication Pereira & Brunel (2018*a*). The authors are Ulises Pereira and Nicolas Brunel. U.P. and N.B. designed the research. U.P. and N.B. developed the mathematical theory. U.P. performed the analytical calculations and numerical simulations. U.P. analyzed the data. U.P. and N.B. wrote the manuscript.

3.2 Introduction

Attractor networks have been proposed as models of learning and memory in the cerebral cortex (Hopfield 1982, Amit 1992, 1995, Brunel 2005). In these models, synaptic connectivity in a recurrent neural network is set up in such a way that the network dynamics have multiple attractor states, each of which represents a particular item that is stored in memory. Each attractor state is a specific pattern of activity of the network, that is correlated with the state of the network when the particular item is presented through external inputs. The attractor property means that the network converges to the stored pattern, even if the external inputs are correlated to, but not identical, to the pattern, a necessary requirement for an associative memory model. In many of these models, the appropriate synaptic connectivity is assumed to be generated thanks to a 'Hebbian' learning process, according to which synaptic efficacies are modified by the activity of pre and post-synaptic neurons (Hebb 1949).

These models have been successful in reproducing qualitatively several landmark observations in delayed response tasks experiments in monkeys (Fuster et al. 1971, Miyashita 1988, Funahashi et al. 1989, Goldman-Rakic 1995) and rodents (Liu et al. 2014, Guo et al. 2014,

Inagaki et al. 2017). In some of the monkey experiments, animals are trained to perform a task in which they have to remember for short times the identity or the location of a visual stimulus. These tasks share in common a presentation period during which the monkey is subjected to an external stimulus, and a delay period during which the monkey has to maintain in working memory the identity of the stimulus, which is needed to solve the task after the end of the delay period. One of the major findings of these experiments is the observation of selective persistent activity during the delay period in a subset of recorded neurons in many cortical areas, in particular in prefrontal cortex (Fuster et al. 1971, Funahashi et al. 1989, Romo et al. 1999), parietal cortex (Koch & Fuster 1989*b*), inferior temporal cortex (Fuster & Jervey 1981, Miyashita 1988, Nakamura & Kubota 1995*a*) and other areas of the temporal lobe (Nakamura & Kubota 1995*a*). In those neurons, the firing rate does not decay to baseline during the delay period, but it is rather maintained at higher than baseline levels. Furthermore, this increase in firing rate is selective, i.e. it occurs only for a subset of stimuli used in the experiment. Selective persistent activity is consistent with attractor dynamics in a recurrent neural network, whose synaptic connectivity is shaped by experience dependent synaptic plasticity (Amit 1995, Wang 2001, Brunel 2005).

The attractor network scenario was originally instantiated in highly simplified fully connected networks of binary neurons (Amari 1972, Hopfield 1982). While theorists have since strived to incorporate more neurophysiological realism into associative memory models, using e.g. asymmetric and sparse connectivity (Derrida et al. 1987), sparse coding of memories (Tsodyks & Feigel'Man 1988, Tsodyks 1988), online learning (Mézard et al. 1986*a*, Parisi 1986, Amit & Fusi 1994), spiking neurons (Gerstner & van Hemmen 1992, Treves 1993, Amit & Brunel 1997, Brunel & Wang 2001, Lansner 2009), there is still a large gap between these models and experimental data. First, none of the existing models use patterns whose statistics is consistent with data. Most models use bimodal distributions of firing rates, with neurons either 'activated' by a stimulus or not, while there is no indication of such a bi-

modality in the data. Second, the connectivity matrices used in these models are essentially engineered (and sometimes highly fine-tuned) such as to produce attractor dynamics, but are totally unconstrained by data. Third, the attractor network scenario has been challenged by the observation of a high degree of irregularity and strong temporal variations in the firing rates of many neurons, which seem hard to reconcile with fixed point attractors (Druckmann & Chklovskii 2012, Barak et al. 2013, Murray et al. 2017).

A recent study (Lim et al. 2015) provides us with the tools to potentially bridge these gaps. It used data from experiments in which neuronal activity is recorded in IT cortex in response to large sets of novel and familiar stimuli (Woloszyn & Sheinberg 2012). The distribution of neuronal responses to novel stimuli allows the inference of the distribution of firing rates of neurons in stimuli that are being memorized. This distribution is close to a lognormal, at odds with bimodal distributions of firing rates used in the vast majority of theoretical studies (for a few exceptions, see Treves (1990*a*,*b*), Festa et al. (2014)). Comparison between the distributions of responses to novel and familiar stimuli allows the inference of the dependence of the learning rule on post-synaptic firing rates. The inferred learning rule is Hebbian, but shows two major differences with classic rules such as the covariance rule (Sejnowski 1977): (1) The post-synaptic dependence of the rule is dominated by depression, such that the vast majority of external inputs leads to a net decrease in total synaptic inputs to a neuron with learning, leading to a sparser representation of external stimuli; (2) The dependence of the rule on post-synaptic firing rates is highly non-linear, as in the Bienenstock-Cooper-Munro rule (Bienenstock et al. 1982).

These results beg the question of whether associative memory can emerge in networks whose distributions of firing rates and learning rules are consistent with data. We therefore set out to study a recurrent network model in which distributions of external inputs, single neuron transfer function and learning rule are all inferred from ITC data (Lim et al. 2015). We show that: (1) learning rules inferred from visual responses in ITC lead to attractor

dynamics, without any need for parameter adjustment or fine tuning; (2) Activity in the delay period is graded, with broad distributions of firing rates; (3) Learning rules inferred from data are close to maximizing the number of stored patterns, in a space of unsupervised Hebbian learning rules with sigmoidal dependence on pre and post-synaptic firing rates; (4) In a large parameter region, our model presents irregular temporal dynamics during retrieval states that strongly resembles the temporal variability observed during delay periods. In this region, retrieval states are chaotic attractors that maintain a positive overlap with the corresponding stored memory, and the network performs as a associative memory device with fluctuations internally generated by the chaotic dynamics.

3.3 The model

We model local cortical circuits in IT cortex by a recurrent network composed of 'firing rate' units (Hopfield 1984). The network is composed of N neurons whose firing rates are described by analog variables r_i, where $i = 1, 2, \ldots, N$ represents the neuron index, as a simplified model for a local network in ITC (see Fig. 3.1 for a schematic depiction of the network). Firing rates obey standard rate equations (Grossberg 1969, Hopfield 1984)

$$\tau \dot{r}_i = -r_i + \phi \left(I_i + \sum_{i \neq j}^{N} J_{ij} r_j \right), \tag{3.1}$$

where τ is the time constant of firing rate dynamics, ϕ is the input-output single neuron transfer function (or f-I curve), I_i are the external inputs to neuron i, and J_{ij} is the strength of the synapse connecting neuron j to neuron i.

The connectivity matrix is sparse, and existing connections are shaped by external inputs ('patterns') through a non-linear unsupervised Hebbian synaptic plasticity rule. In this rule, external synaptic inputs ξ_i^μ to neuron i during presentation of pattern μ ($i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, p$) are generated randomly and independently from a Gaussian distribution (see

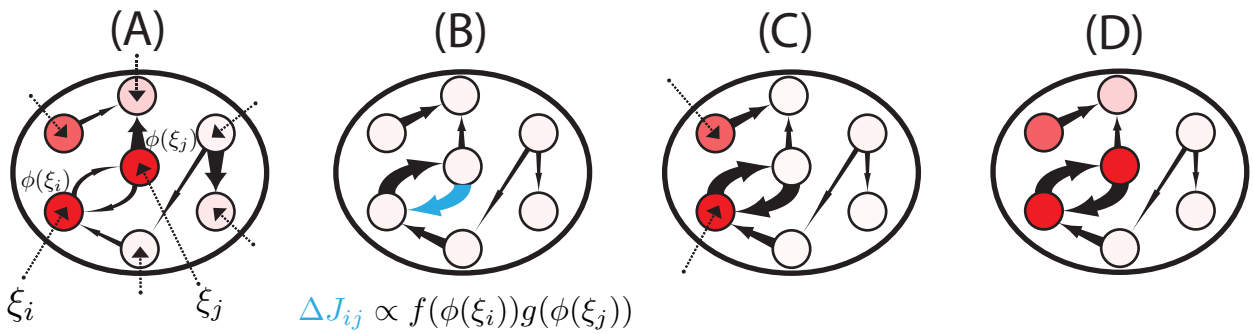

Figure 3.1: Learning and retrieval in recurrent neural networks with unsupervised Hebbian learning rules. (**A**) When a novel pattern is presented to the network, synaptic inputs to each neuron in the network (ξ_l, for neurons $l = 1, \ldots, N$) are drawn randomly and independently from a Gaussian distribution. Synaptic inputs elicit firing rates through the static transfer function, i.e. $\phi(\xi_l)$. Some neurons respond strongly (red circles), others weakly (white circles). (**B**) The firing rate pattern produced by the synaptic input currents modifies the network connectivity according to an unsupervised Hebbian learning rule. The connection strength is represented by the thickness of the corresponding arrow (the thicker the arrow the stronger the connection). (**C**) After learning, a pattern of synaptic inputs that is correlated but not identical to the stored pattern is presented to the network. (**D**) Following the presentation, the network goes to an attractor state which strongly overlaps with the stored pattern (compare with panel A), which indicates the retrieval of the corresponding memory.

Fig. 3.1 A,B and Methods). The assumption of independence of the patterns is consistent with the data (see Fig. 3.2). The external inputs shape the connectivity matrix through the firing rates $\phi(\xi_i^\mu)$ generated by such inputs, and through two non-linear functions f and g that characterize the dependence of the learning rule on the post-synaptic rate (f) and pre-synaptic rate (g), respectively. When p patterns are learned by the network, the final connectivity after learning gets structured as

$$J_{ij} = \frac{Ac_{ij}}{cN} \sum_{k=1}^{p} f\left[\phi(\xi_i^k)\right] g\left[\phi(\xi_j^k)\right], \tag{3.2}$$

where c_{ij} is a sparse random (Erdos-Renyi) structural connectivity matrix ($c_{ij} = 1$ with probability c, $c_{ij} = 0$ with probability $1-c$, where $c \ll 1$). This synaptic connectivity matrix can be obtained by a learning rule that changes the synaptic connectivity matrix by a factor $\Delta J_{ij} \propto f\left[\phi(\xi_i^\mu)\right] g\left[\phi(\xi_j^\mu)\right]$ when a pattern μ is presented to the network, starting from an

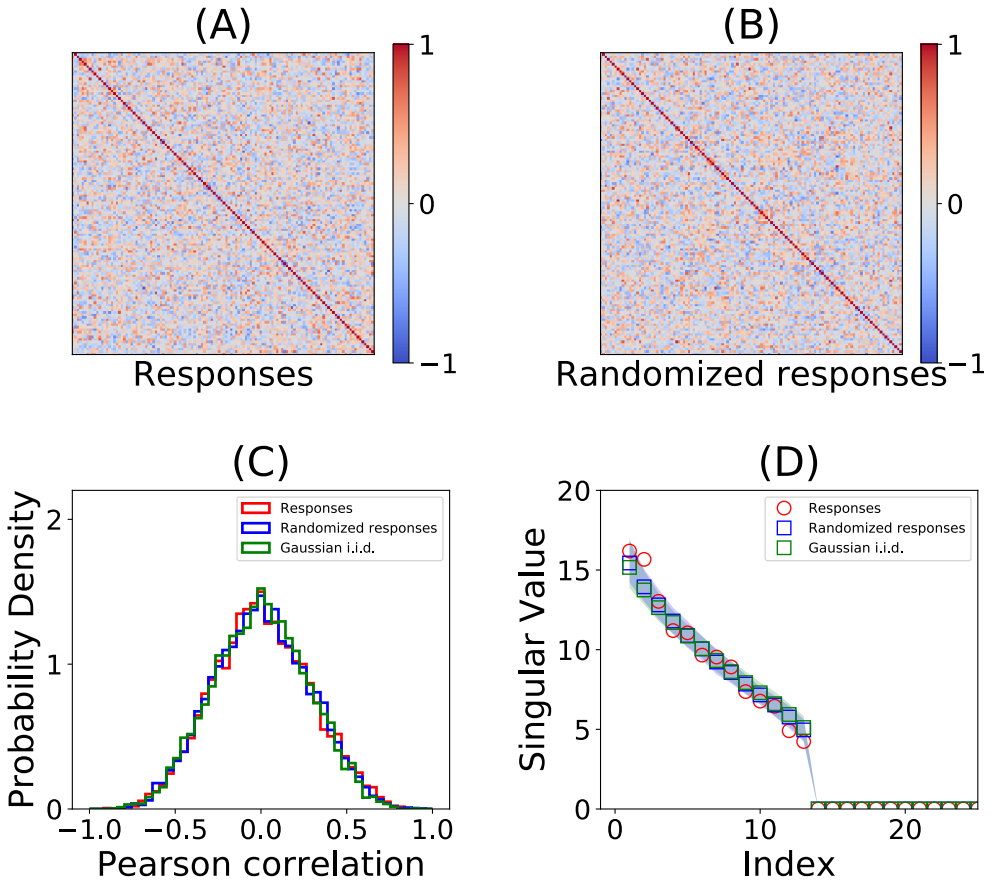

Figure 3.2: Correlations between input currents corresponding to familiar images. Our theory relies on the assumption that stored patterns are uncorrelated. When correlations between stored patterns are included, the storage capacity of our network drastically decreases. To test this assumption, we computed correlations between input currents corresponding to familiar images. Using the transfer function inferred from novel stimuli (see Fig. 3.3B), we computed for each neuron the input currents that elicit the firing rate responses to each of the 125 familiar images. We then computed the correlations between input currents corresponding to different familiar images. **(A)** Correlation matrix for the input currents corresponding to the 125 familiar images across the 14 putative excitatory neurons considered in this study. **(B)** Correlation matrix for the input currents corresponding to the 125 familiar images when the identity of the images are randomly shuffled for each neuron. **(C)** Histogram of the correlation values for the original correlation matrix in panel A (red), the correlation matrix from the randomized responses in panel B (blue) and from a correlation matrix of input currents drawn independently from a Gaussian distribution with zero mean and unit variance (i.e. Gaussian i.i.d.; green). The diagonal terms were excluded. **(D)** Largest 20 singular values for correlation matrices in panel A (red), B (green) and from the Gaussian i.i.d. input currents (blue). The opaque blue and green shaded areas correspond to the 95% confidence interval for the singular values across 200 realizations of the correlation matrices of the randomized responses and the Gaussian i.i.d. input currents respectively.

initial *tabula rasa* $J_{ij} = 0$, and neglecting the contributions of recurrent connections during learning. This rule is a generalization of Hebbian rules used in classic models such as the Hopfield model (Hopfield 1982) or the Tsodyks-Feigel'man model (Tsodyks & Feigel'Man 1988), with two important differences: patterns have a Gaussian distribution instead of binary; and the dependence of the rule on firing rates is non-linear instead of linear. In the following, the patterns that have shaped the connectivity matrix will be termed 'familiar' while all other random patterns presented to the network will be termed 'novel'.

3.4 Inferring transfer function and learning rule from data

The model defined by Eqs. (4.1,3.2) depends on three functions ϕ, f and g that define the single neuron transfer function and synaptic learning rule, respectively. How to choose these functions? We used a method that was recently introduced by Lim et al. (2015) to infer the tranfer function (ϕ) and the post-synaptic dependence of the learning rule f from electrophysiological data recorded in ITC (Woloszyn & Sheinberg 2012). The transfer function ϕ is obtained by finding the function that maps a standard Gaussian distribution to the empirical distribution of visual responses of neurons to a large set of novel stimuli (see Methods). The post-synaptic dependence of the learning rule f was obtained from the differences between the distribution of visual responses to familiar and novel stimuli, under the assumption that changes in such distributions are due to changes in synaptic connectivity in recurrent ITC circuits. Note that only the function f, and not g, can be inferred from data - this is due to the fact that the mean inputs to a neuron are proportional to $f\left[\phi(\xi_i^k)\right]$ while the function g only appears in an integral (see Methods, Eq. (3.51)). Therefore, the knowledge of how the mean inputs change with learning as a function of its firing rate allows us to infer f but not g. As an additional step to the procedure described by Lim et al. (2015), we fitted the resulting functions ϕ and f using sigmoidal functions (see Methods and Fig. 3.3, i.e. $\phi_i(\xi) = r_m/(1 + e^{-\beta(\xi - h_0)})$ and $f(r) = \frac{1}{2}\left[2q_f - 1 + \tanh(\beta_f(r - x_f))\right]$

respectively). These sigmoidal functions provided good fits to the data (see Fig. 3.3A-C, that shows fits of three representative ITC neurons; and Fig. A.1-A.3 for all neurons in the data set). This fitting procedure gave us for each neurons three parameters of the transfer function: the maximal firing rate r_m (median: $r_m = 76.2$Hz), a measure of the slope at the inflection point β_T (median: $\beta_T = 0.82$), and the threshold (current at the inflection point, median: $h_0 = 2.46$ - see Fig. 3.3D for a boxplot of these parameters). It also gives us for each neuron three parameters characterizing the function f: the threshold x_f (median: 26.6 Hz), slope at the inflection point β_f (median: 0.28 s) and saturation q_f (median: 0.83). Finally, the fitting procedure also gives us the learning rate A (median: 3.55).

A number of features of these fitted functions are noteworthy: First, the vast majority of the visual responses of neurons are in the supralinear part of the transfer function, and therefore far from saturation. This is consistent with many studies showing supra-linear transfer functions at low firing rates, both *in vitro* (Rauch et al. 2003) and *in vivo* (Anderson et al. 2000). Second, this has the consequence that the distribution of visual responses are strongly right-skewed, and in fact close to lognormal distributions, consistent with multiple observations *in vivo* (Hromadka et al. 2008, Roxin et al. 2011, Buzsaki & Mizuseki 2014, Lim et al. 2015). Third, the function f is strongly non-linear, and the threshold between depression and potentiation occurs at a firing rate that is much higher than the mean rate, leading to depression of the mean synaptic inputs to a neuron for the vast majority of shown stimuli. Fourth, the average of the function f across the distribution of patterns is negative, which leads to a decrease of the average visual response with familiarity (Lim et al. 2015).

The only parameters that are left unconstrained by data are two parameters characterizing the function g. In most of the following, we will take those parameters to be identical to the corresponding parameters of the function f (i.e. $x_g = x_f$ and $\beta_g = \beta_f$; note that q_g is fixed by the condition that the average of the function g across the distribution of patterns is zero, see Methods). We will also explore the space of values of x_g and β_g (see below).

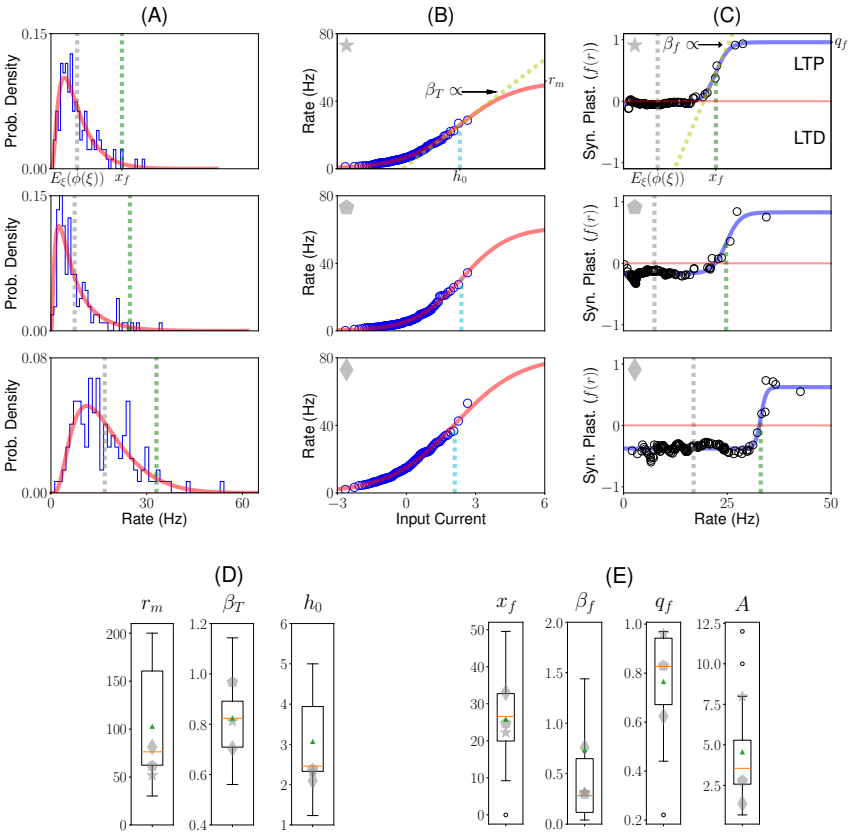

Figure 3.3: Inferring transfer function and learning rule from ITC data. (**A**) Distributions of firing rates in response to novel stimuli, for three different ITC neurons. Blue histogram: histogram of experimentally recorded visual responses. Red: Distribution of firing rates obtained from passing a standard normal distribution through the sigmoidal transfer function shown in B. Gray vertical line: average firing rate. Green vertical line: learning rule threshold x_f (see C). (**B**) Static transfer function ϕ derived from the distribution of visual responses for novel stimuli (see A), assuming a Gaussian distribution of inputs (see (Lim et al. 2015) and Methods) for the same three neurons shown in A. The data (blue circles) was fitted using a sigmoidal function (red line; see Methods, Eq. (3.48)), defined by three parameters: the current h_0 that leads to half the maximal firing rate (cyan dashed lines), a slope parameter β_T (dashed yellow line in top plot), and maximal firing rate r_m. (**C**) Dependence of the synaptic plasticity rule on the postsynaptic firing rate as a function of firing rate (i.e. $f(r)$). The data (black circles) was fitted with a sigmoidal function (blue line; see Methods, Eq. (3.53)), defined by three parameters: maximum potentiation q_f; threshold x_f (see green dashed line); and slope parameter β_f (dashed yellow line in top plot). On the right axis is indicated the maximum potentiation of the fit q_f. (**D**) Boxplot for the fitted parameters r_m, β_T and h_0 of the transfer function. (**E**) Boxplot for the fitted parameters x_f, β_f, q_f of the dependence of the synaptic plasticity rule on the postsynaptic firing rate, and A, the learning rate. The red line and green triangle indicate the median and the mean of the fitted parameters, respectively. Gray symbols indicate the parameters of the three neurons shown in A,B,C.

3.5 Dynamics of the network following presentation of familiar and novel stimuli

Having specified the model, we now turn to the dynamics of the network described by Eqs. (4.1,3.2), whose parameters are set to the median best-fit parameters according to the procedure described above. In particular, we ask whether the model exhibits attractor dynamics. To address this question, we used both numerical simulations of large networks (see Methods) and mean field theory (MFT - see Methods). For the MFT, we assume that both the number of neurons and stored patterns are large (i.e. more specifically the limit $p, N \to \infty$), while the number of stored patterns p divided by the average number of synapses per neuron (Nc), $\alpha \equiv p/Nc$ remains of order one. We call α the *memory load* of the network. The results of the MFT only depend on N, c and p via this quantity (see Methods). From our MFT analysis, we obtain mathematical expressions for two 'order parameters' that describe how network states are correlated (or not) with stored patterns. We are specifically interested here in the situation when the network state is correlated with one of the stored patterns (e.g. following the presentation of this particular pattern).

The first order parameter describes the 'overlap' m between the current state of the network (described by the vector of firing rates r_i, for $i = 1, 2, \ldots, N$) and the pattern of interest (see methods for the mathematical definition of m). When m is of order 1, this indicates that the corresponding pattern is retrieved from memory. Consequently, each pattern stored in memory can be retrieved by initializing the network dynamics with a configuration that is close to that particular pattern, and letting the network evolve towards its attractor state. In this case, giving a partial cue to the network leads the dynamics towards an attractor state correlated with the stored pattern, a signature of associative memory. The other order parameter M describes the interference due to the other stored patterns in the connectivity matrix; it is proportional to the average squared firing rates of the network (see Methods). Equations for the order parameters as a function of α, ϕ, f and

g are given in Methods.

The results of the simulation of a particular realization of a network of $N = 50,000$ neurons with $c = 0.005$ (an average of 250 connections per neuron) storing $p = 30$ patterns ($\alpha = 0.12$), and the comparison with the results from MFT are shown in Fig. 3.4. In the simulations, the network was initialized in a state which was uncorrelated with all the stored patterns. For these parameters, the network converged to a 'background' state in which all neurons fire at low rates (average 7.98/s, standard deviation 2.92/s). Upon presentation of a novel stimulus (Fig. 3.4A), neurons were driven to stimulus-specific firing rates, with a distribution of firing rates that was close to a lognormal distribution (Fig. 3.4C), similar to experimental observations (Lim et al. 2015). The distribution is close to lognormal because the distribution of inputs to neurons is Gaussian, and the neuronal transfer function is close to being exponential at low rates (see Methods). After the end of the presentation of the stimulus, the network came back to its initial background state (Fig. 3.4A). Upon presentation of a familiar stimulus (Fig. 3.4D), the statistics of neuronal responses differed markedly from the response to novel stimuli: a few neurons responded at higher rates, but the majority of neurons responded at lower rates compared to a novel stimulus. The distribution of visual responses for familiar stimuli had consequently a lower mean compared to the distribution of responses for novel stimuli but a larger tail at high rates (compare Fig. 3.4C and F). These two features were consistent with data recorded in ITC by multiple groups (Li et al. 1993, Kobatake et al. 1998, Logothetis et al. 1995, Freedman et al. 2006, Woloszyn & Sheinberg 2012).

After removal of a familiar stimulus, the network no longer came back to the initial background state, but rather converged to an attractor state that was strongly correlated with the shown stimulus (Fig. 3.4D), as shown by the strong overlap between the network state and the shown pattern (see blue curve in Fig. 3.4E). A small fraction of neurons exhibited persistent activity at high rates (4.3% of the neurons are above half maximal

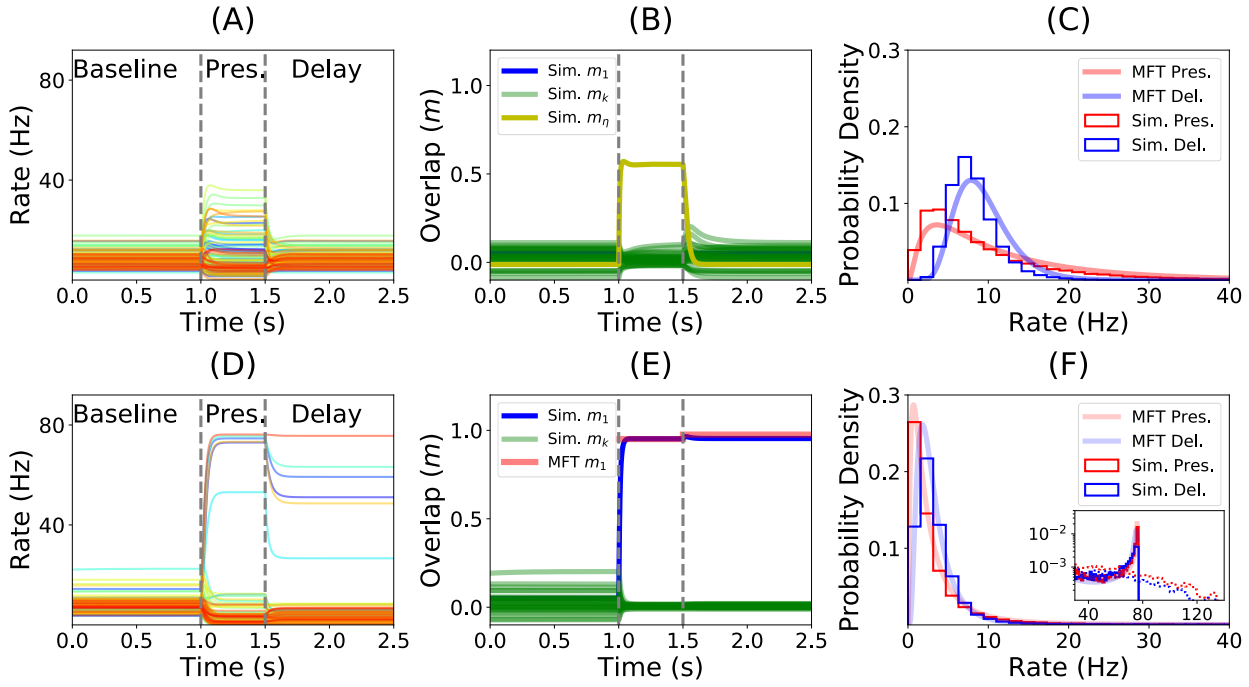

Figure 3.4: Dynamics of the network before, during and after the presentation of novel (top row) and familiar (bottom row) stimuli, mimicking the initial part of a trial of a delay match to sample (DMS) experiment. (**A**) Firing rate of a randomly sampled subset of 100 neurons of a simulated network before, during and after the presentation of a novel stimulus. Vertical dashed lines indicate the beginning and the end of the presentation. Note that the firing rates of all neurons decay to baseline following removal of the stimulus. (**B**) Dynamics of the overlaps with the stored patterns. Green traces show overlaps computed numerically from the network simulation corresponding to each of the stored patterns. The yellow trace shows the overlap of the network state with the shown novel pattern. (**C**) Distribution of firing rates during the presentation (red) and delay (blue) periods. Smooth curves correspond to the predictions of the MFT, histograms are obtained from network simulations. (**D**) Similar to A, except that the shown stimulus is familiar. Note that this time firing rates do not decay to baseline during the delay period, but to a value that is strongly correlated (but not identical) to the visual response. (**E**) Dynamics of overlaps when a familiar stimulus is presented. The blue trace shows the numerically computed overlap with the pattern presented during the presentation period. The red trace shows the corresponding overlap computed from MFT. (**F**) Distribution of firing rates during the presentation (red) and delay (blue) periods in response to the presentation of a familiar stimulus. The vast majority of the neurons fire in the 0-10Hz range. A closer inspection of the tail of the distribution shows a tiny peak close to saturation in homogeneous networks (full lines), while this peak disappears when the heterogeneity in maximal firing rates is included (dashed lines).

rate), but most neurons remained at low rates during the simulated delay period (Fig. 3.4F). The distribution of firing rates was again similar to a lognormal distribution at low rates, but the tail of the distribution was shaped by neuronal saturation and therefore exhibited a tiny peak close to maximal firing rates. Both overlap with presented pattern and distributions of firing rates could be computed by the MFT and were in close agreement with network simulations (Fig. 3.4E and F). When the heterogeneity on the neuronal saturation is included into our model by randomly selecting maximal firing rates for each neuron from a log-normal distribution that fits the empirical distribution of the best-fit maximal firing rates (see Fig. 3.3E), the peak at maximal firing rate disappears. Thus, in a heterogeneous network, distributions of firing rates during both presentation and delay periods become unimodal (Fig. 3.4F dashed lines).

Thus, our network behaved as an associative memory when constrained by ITC data, without any need for parameter variation or fine tuning. Furthermore, in addition to reproducing the distributions of visual responses for both novel and familiar stimuli seen experimentally, it also exhibited qualitatively some of the main features observed both during spontaneous and delay activity in IT cortex: broad distribution of firing rates in both spontaneous and delay period activity, and small fraction of neurons firing at elevated rates during persistent activity (Miyashita 1988, Nakamura & Kubota 1995*a*).

3.6 Storage capacity, and its dependence on g

We now turn to the question of the storage capacity of the network, i.e. how many different patterns can be stored in the connectivity matrix. The calculation of the storage capacity of associative memory models such as the Hopfield model was one of the first successful applications of statistical physics to theoretical neuroscience (Amit et al. 1987). One of the main findings of such models is that the number of patterns that can be stored scales linearly with the number of plastic connections per neuron, i.e. the maximal value of α is of order

1. This maximal storage capacity α_c has been computed in many variants of the Hopfield model (see e.g. Amit (1992)). To compute the storage capacity of our network, we found numerically the largest value of α for which retrieval states (i.e. states with positive overlap with one of the stored patterns, $m > 0$) exist. Fig. 3.5A shows how the overlap in retrieval states m varies as a function of the storage load α, computed using both MFT (solid line) and simulations (symbols with errorbars) when parameters of the functions ϕ and f are taken to be the median best-fit parameters, and those of the function g (except q_g, that is set by the balance condition, Eq. 3.56) are taken to be identical to f. It shows that m gradually decreases with α, due to more 'noise' in the retrieval due to other stored patterns, until it drops abruptly to zero at a value of $\alpha_c = 0.56$. This value is remarkably close to the maximal capacity of the sparsely connected Hopfield model of binary neurons storing binary patterns, for which $\alpha_c = 0.64$ (Derrida et al. 1987).

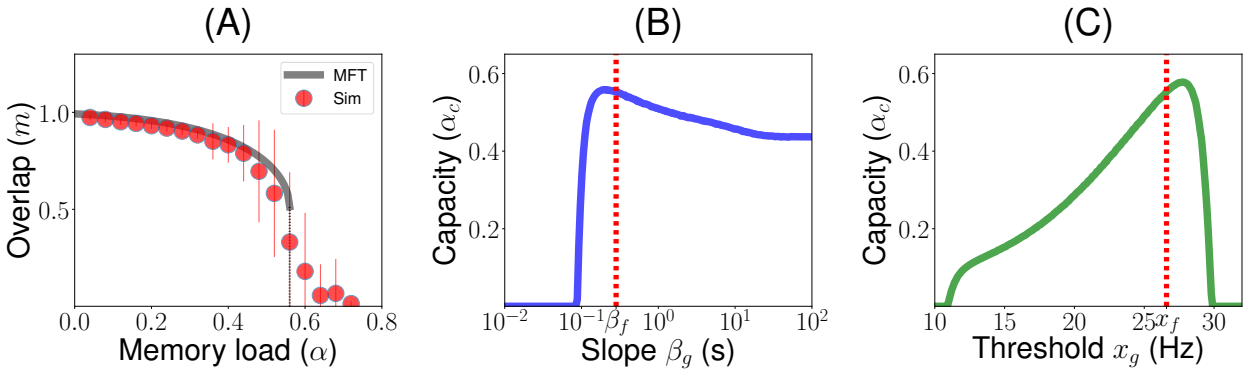

Figure 3.5: Storage capacity of the network, and its dependence on g. (**A**) Overlap as a function of memory load α (number of patterns stored divided by average number of connections per neuron). Grey: MFT. Red circles: Numerical simulations (average and standard deviations computed from 100 realizations with $N = 5 \cdot 10^4$). The overlap stays positive until $\alpha \sim 0.56$. Parameters of g are chosen to be identical to those of f. (**B**) Capacity vs β_g. The capacity is maximized for $\beta_g \sim \beta_f$ (dashed red line $\beta_g = \beta_f$). (**C**) Capacity vs x_g. The capacity is close to being maximized for $x_f \sim x_g$ (dashed red line $x_g = x_f$). Other parameters as in Fig. 3.4.

We then explored how the capacity depends on the parameters of the function g, that describes the dependence of the learning rule on the presynaptic firing rate. Fig. 3.5B and

C show that the capacity is close to being maximized when these parameters match those of the function f, i.e. $x_g = x_f$ and $\beta_g = \beta_f$. Fig. 3.5B shows that the capacity is non-zero only when the g is sufficiently non-linear, i.e. $\beta_g > 0.1$. It peaks around $\beta_g = \beta_f$, but remains high in the $\beta_g \to \infty$ limit when the function g becomes a step function. Fig. 3.5C shows that the capacity is non-zero only in a finite range of x_f, between 10 and 30/s. It shows again that capacity peaks when x_g is close to x_f.

3.7 Learning rules inferred from ITC data are close to maximizing memory storage

The storage capacity of the network with median parameters is in the same range or higher than the capacity of classic associative memory models of binary neurons - for instance, the Hopfield model has a capacity of $\alpha_c \sim 0.14$ (Amit et al. 1987), while its sparsely connected variant has a capacity of $\alpha_c \sim 0.64$ (Derrida et al. 1987). The next question we addressed is how this capacity depends on the parameters of this learning rule. We have already discussed above the dependence of the capacity on x_g and β_g. Here, we explore the dependence on the four remaining parameters characterizing the learning rule - A, x_f, β_f and q_f. Using MFT, we explored systematically the space of these four parameters, and plot in Fig. 3.6 all possible cuts of this four dimensional space, in which 2 of the 4 parameters are varied, while the other 2 are set to the median values. In all these plots, the maximal capacity α_c is plotted as a function of two parameters, using a gray scale (white indicate high capacity, black low capacity). The yellow dashed line indicates the line for which the function f is 'balanced' (i.e. its average across the distribution of patterns is zero). It marks the border between a depression-dominated region, for which learning leads to a decrease in average responses, and a potentiation-dominated region, for which learning leads to an increase of such responses. The red cross mark indicates the median parameters, while the dashed red rectangle indicates the interquartile range.

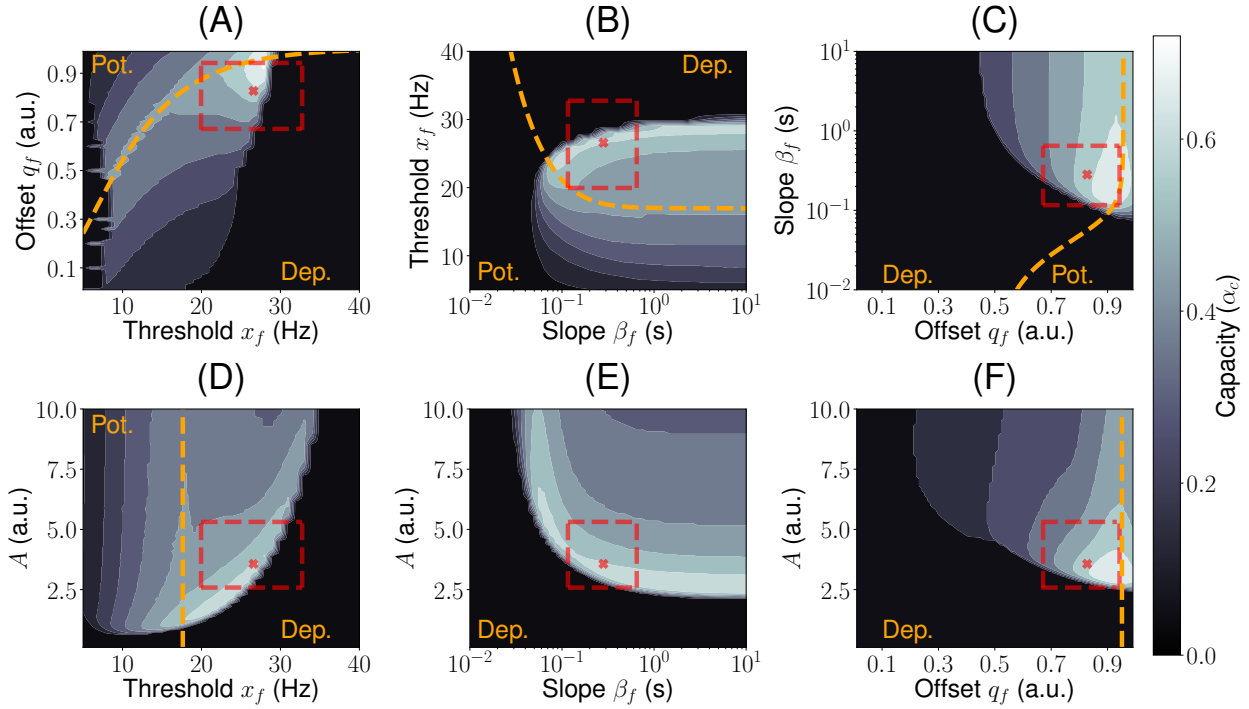

Figure 3.6: Inferred learning rules from ITC are close to maximizing memory storage. Contour plots for the capacity of the network as a function of two parameters. In each plot, two parameters are set to the median best-fit parameters, and the other two are varied. The yellow dashed line indicates the curve where potentiation and depression are balanced in average (i.e. $\int d\xi \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} f(\phi(\xi))) = 0$). It separates the potentiation (i.e. $\int d\xi \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} f(\phi(\xi))) > 0$) and depression (i.e. $\int d\xi \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} f(\phi(\xi))) < 0$) regions. The parameter region corresponding to the interquartile range is indicated with a red dashed rectangle. The median best-fit parameters are shown as a red cross mark. The parameters of g: $x_g = x_f$ and $\beta_g = \beta_f$.

Fig. 3.6 shows that the median parameters are close to maximizing storage capacity. In fact, we found that the maximal capacity over this space is $\alpha_c \approx 0.85$ (see Fig. 3.7 and 3.8 for details). These figures show also that most (but not all) of the interquartile range lie in a high-capacity region. It also shows that some parameter variations lead to little changes in capacity, while others lead to a drastic drop. Decreasing the learning strength A from its optimal value leads to an abrupt drop in capacity, while increasing it leads to a much gentler decrease (see Fig. 3.6D-F). A similar effect is observed for the slope of f; decreasing the slope (i.e. making f more linear) leads to an abrupt decrease in capacity, while increasing it beyond the median value leads to very little change in capacity (see Fig. 3.6B-D). Thresholds x_f for which high capacities are obtained are much higher than the mean response to novel visual stimuli (Fig. 3.6A,B and D), leading to a sparsening of the representations of the patterns by the network. Finally, the optimal offset is close to the 'balanced' line, but slightly on the depression-dominated region, as the median parameter (Fig. 3.6A,C and F).

3.8 A chaotic phase with associative memory properties

Are fixed point attractors the only possible dynamical regime in this network? Firing rate models with asymmetric connectivity have been shown to exhibit strongly chaotic states (Sompolinsky et al. 1988, Tirozzi & Tsodyks 1991). Varying parameters of the learning rule, we found parameter regions in which background and/or retrieval fixed point attractor states destabilize and the network settle into strongly chaotic states. Fig. 3.9A shows an example of such chaotic states, obtained for the median parameters as in Fig. 3.4, except for the learning rate which is three times its median best-fit value ($A = 10.65$). For such parameters, the background state is strongly chaotic. Presentation of a familiar stimulus leads to a transition to another chaotic state, in which all neurons fluctuate chaotically around stimulus-specific firing rates, such that the mean overlap with the corresponding pattern remains high (see Fig. 3.9 B). Remarkably, chaotic retrieval states remain strongly

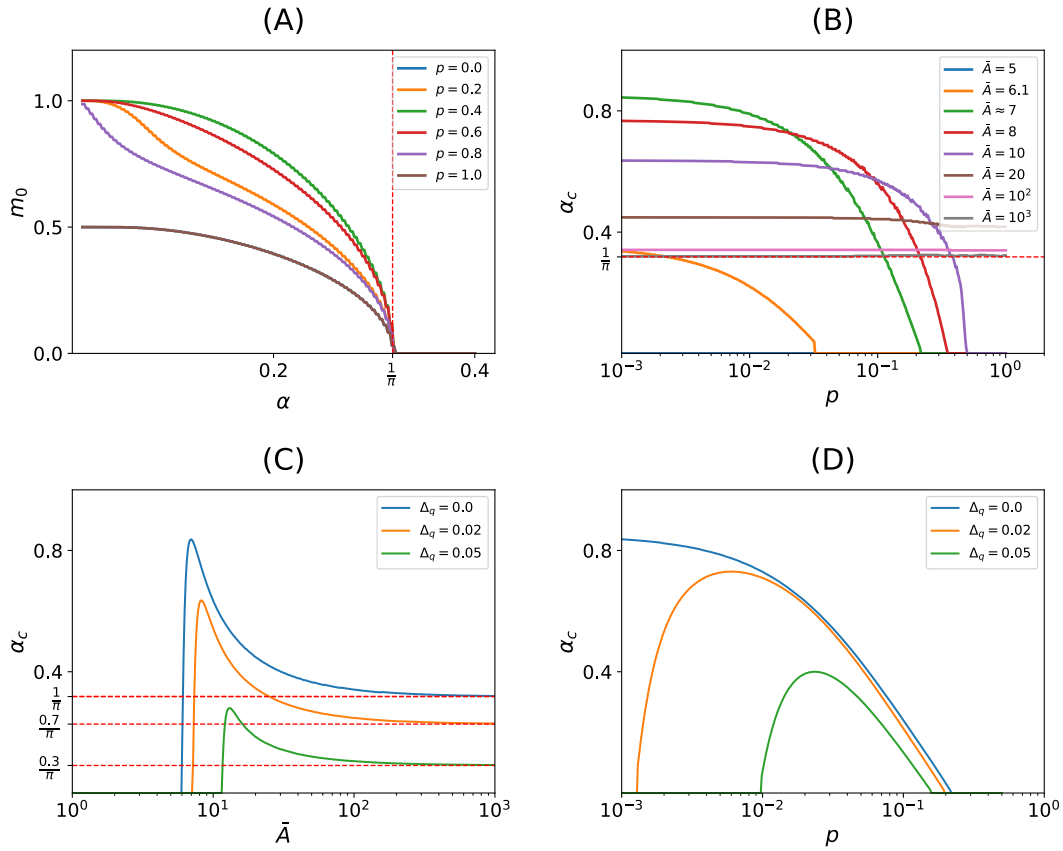

Figure 3.7: MFT limits and capacity. (**A**) Overlap vs load α in the limit $q_g \to q_f$ and $\bar{A} \to \infty$ (see (3.39) and (3.40) on Methods) for $p = 0, 0.2, 0.4, 0.6, 0.8, 1$ ($p = 1 - q_g$). In this limit, the mean field equations (see Eq. (20) and (21) in Tsodyks (1988)) with $\theta_0 = 0$ are recovered. $\alpha_c \approx \frac{\theta_0^2}{2p\log(1/p)}$ is not attainable since $\theta_0 = 0$. The capacity is $\alpha_c = 1/\pi$ for all p. (**B**) Capacity vs p for $\bar{A} = 5, 6.1, 6.95, 8, 10, 20, 100, 1000$ (see Eq. (3.39) and (3.40) on Methods). For a fixed $\bar{A} \sim \mathcal{O}(1)$, capacity is maximized in the sparse coding limit (i.e. $p \to 0$). $\bar{A} \approx 6.95$ and $p \to 0$ leads to the maximal capacity in the $\bar{A}$-p plane, with $\alpha_c \approx 0.85$ (see green curve). For $\bar{A} \to \infty$, which implies $\frac{r_m}{p(1-p)} \ll A$, the capacity is $\alpha_c = 1/\pi$ for all p (see gray curve and dashed red line) as shown in panel A. (**C**) Capacity vs $\bar{A}$ for $p = 10^{-3}$ and $\Delta_q = q_g - q_f = 0, 0.02, 0.05$ (see Eq. (33) and (34) on section 3 of Methods S1). For $\Delta_p = 0$ the mean field equations are the same as in panel B with $p = 10^{-3}$, showing the maximal capacity $\alpha_c \approx 0.85$ at $\bar{A} \approx 6.95$. Increasing Δ_q, which implies $q_f < q_g$, produces a rapid decrease in the capacity (see orange and green curves). For $\bar{A} \to \infty$ the capacity decreases rapidly as $\alpha_c = \eta^2/\pi$ (see dashed red lines) as shown on Methods (see Eq. (3.44)). (**D**) Capacity vs p for $\bar{A} = 6.95$ and $\Delta_q = 0, 0.02, 0.05$ (see Eq. (3.37) and (3.38) on Methods). As in panel C, the capacity decreases rapidly as Δ_q increases. For $0 < \Delta_q$ the maximal capacity occurs at non-zero p (the capacity curve becomes concave). This is similar to what is observed for sigmoidal f and g (see Fig. 3.6), where the maximal capacity is obtained for a finite threshold and therefore a non-zero value of p.

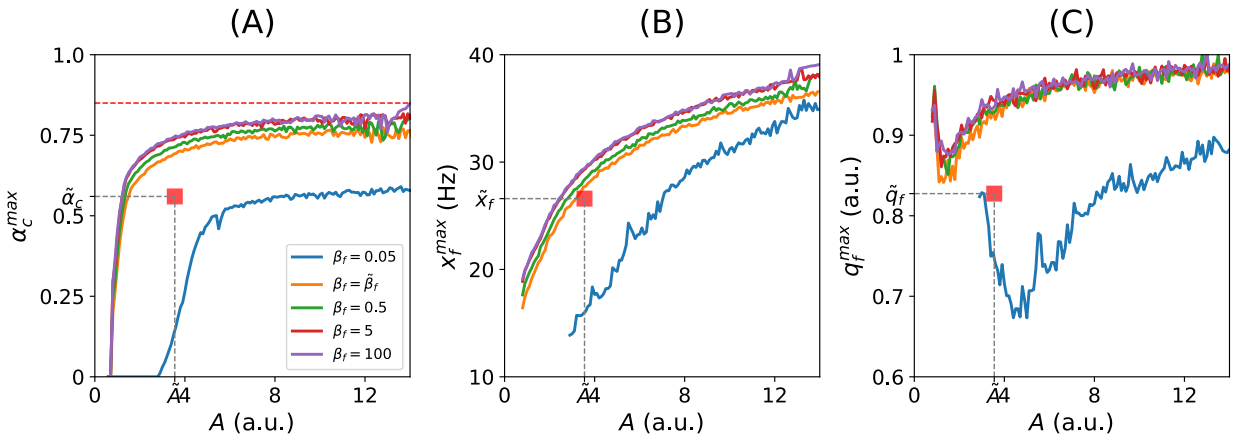

Figure 3.8: Maximal capacity in the (x_f, β_f, q_f, A) parameter space. Numerical search of the maximal capacity using the Nelder-Mead algorithm for sigmoidal f and g (see Eq. (3.13) and (3.13)) with $x_g = x_f$ and $\beta_g = \beta_f$. For four representative values of the slope β_f ($\beta_f = \{0.05, \tilde{\beta}_f, 0.5, 5\}$ where $\tilde{\beta}_f$ is the median of the best-fit slopes), the maximal capacity was searched in the (x_f, q_f) parameter space for fixed values of A in a grid. Starting with the largest value of A in the grid, we used the previous maximal capacity point (x_f^{max}, q_f^{max}) as the initial condition for searching the maximal capacity of the next value of A. In this way, we smoothly followed the maximal capacity in the (x_f, q_f, A) parameter space. The maximal capacity increases monotonically with β_f. (**A**) Maximal capacity (α_c) vs A for different values of β_f. The maximal capacity approaches asymptotically the value found in Fig. 3.7.B (i.e. $\alpha_c \approx 0.85$, see dashed red line). The capacity for the median best-fit parameters (see red square) is smaller but comparable with the maximal capacity in the (x_f, β_f, q_f, A) parameter space. (**B**) Optimal threshold (x_f) vs A. Red square: Median best-fit threshold. (**C**) Optimal saturation (q_f) vs A. Red square: Median best-fit saturation.

correlated with the corresponding patterns (see Fig. 3.9B), so that the network can still perform as an associative memory in spite of the chaotic fluctuations of network activity. Interestingly, the storage capacity for such parameters is larger than the capacity estimated from the static MFT (see Fig. 3.9C).

In such chaotic retrieval states, single neuron activity exhibit strong firing rate fluctuations which vary from trial to trial (see thin colored lines in Fig. 3.9D-F showing three randomly selected neurons), but trial-averaged firing rates show systematic temporal patterns. For instance, the activity of the neuron shown in Fig. 3.9D ramps up in the first second of the delay period, before this activity plateaus at a rate of about 20/s. The neuron shown in Fig. 3.9F shows a rapid activity increase during the presentation period, followed by a trough, followed by a second increase during the delay period. These temporal patterns of the trial-averaged firing rate, together with a strong irregularity within trials, are reminiscent of observations by multiple groups in primate PFC during delay periods (Shafi et al. 2007, Brody et al. 2003, Murray et al. 2017).

To check whether these states are truly chaotic, we computed the temporal evolution of the distance between two network states with slightly different initial conditions (see Methods). Fig. 3.9G shows that an initial distance between two initial conditions of $4.5 \cdot 10^{-6}$Hz exponentially grows and then plateaus to an average of $\sim$ 13Hz. This sensitivity to initial conditions, and initial exponential growth of the distance between perturbed and unperturbed network states is the defining feature of a chaotic system (Guckenheimer & Holmes 2013). The divergence of the network states starts to be noticeable in the single neuron dynamics in about $\sim$1s (see Fig. 3.9H). However, the overlap with the stored pattern remains high in both networks states (see Fig. 3.9I). Therefore, despite the growth of the distance between the two network states, their dynamics keep aligned to the 1-dimensional subspace (of the full N-dimensional network space) spanned by the retrieved memory, providing a low dimensional representation of each memory.

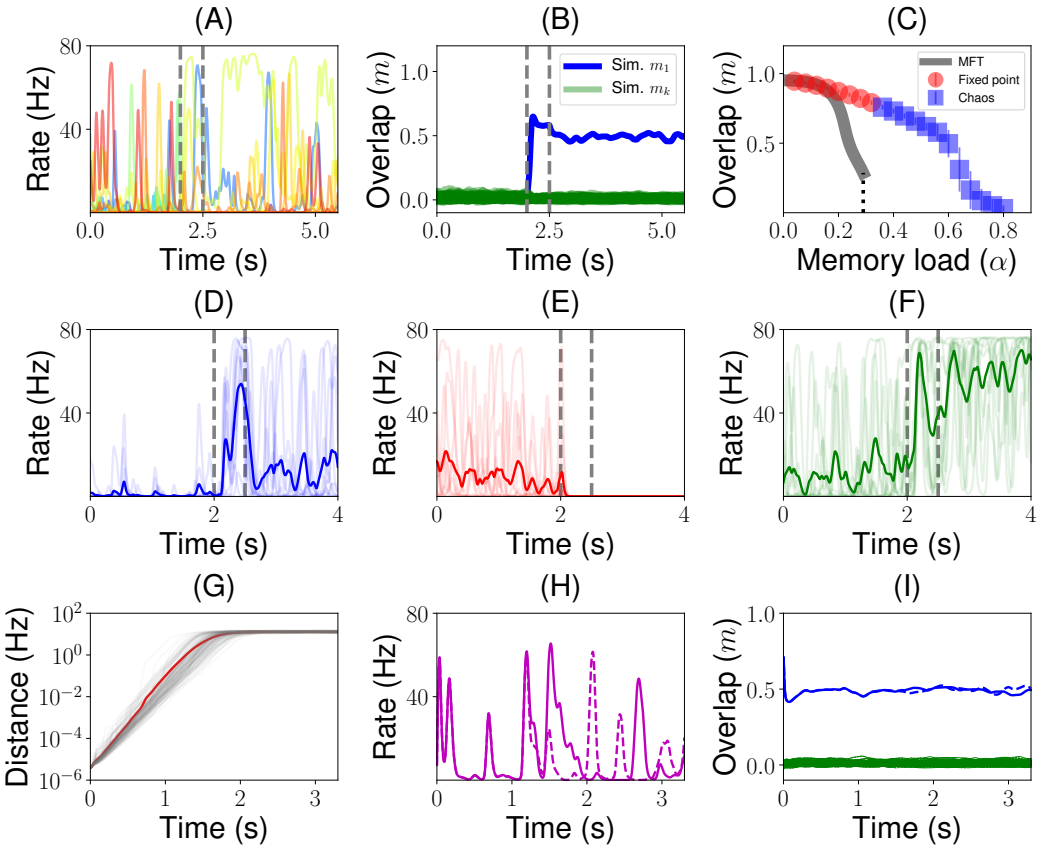

Figure 3.9: Chaotic background and retrieval states, for a network with parameters as in Fig. 3.4, except for the learning rate ($A = 10.65$) and memory load ($\alpha = 0.48$ in all panels except in C). (**A**) Firing rate dynamics for a randomly sampled subset of 10 neurons of a simulated network when a familiar stimulus (i.e. one of the stored patterns) is presented. (**B**) Dynamics of the overlaps before, during and after the presentation of a familiar stimulus. Green traces shown all the overlaps computed numerically from the network simulation corresponding to each of the stored patterns except the one with the presented pattern, shown in blue. (**C**) Overlap vs memory load. Gray curve: MFT. Red circles: simulations in which the dynamics converge to fixed point attractors. Blue square: simulations in which the dynamics converge to chaotic states. (**D-F**) Dynamics of the firing rate of three example neurons in 10 different trials (random initial conditions - transparent traces). Trial-averaged firing rate (over 20 trials) is shown with an opaque trace. (**G**) Light gray traces: exponential initial growth followed by saturation of the distance between pairs of retrieval states corresponding to the same stored pattern but slightly different initial conditions (see Methods). Red curve: average distance between pairs of retrieval states with slightly different initial conditions. (**H**) Firing rate of a single neuron starting from two slightly different initial conditions (continuous vs dashed). (**I**) Overlaps with the retrieved pattern (blue) and all other stored patterns (green) again for a pair of initial conditions (continuous vs dashed). As in Fig. 3.4, in A, B and D-F vertical dashed lines indicate the beginning and the end of the presentation period.

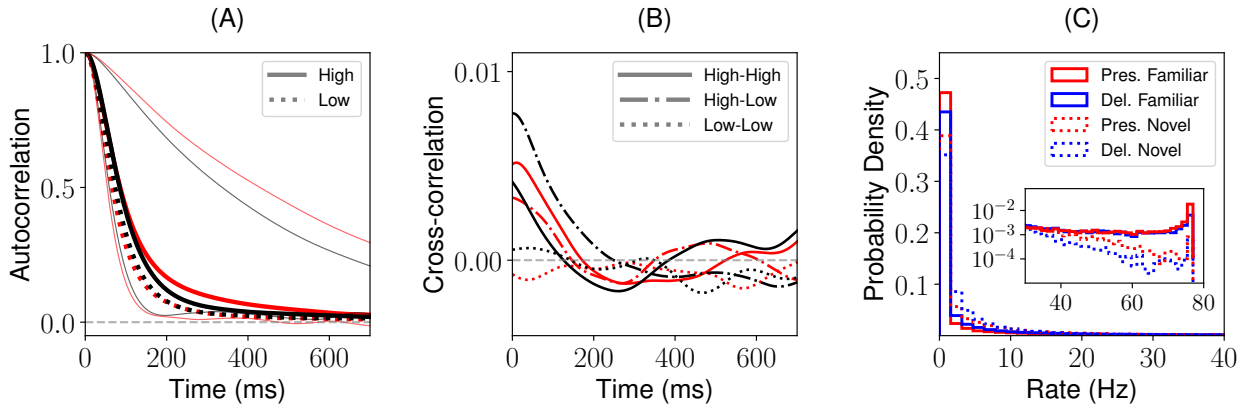

Figure 3.10: Statistical properties of the chaotic background and retrieval states, for a network with parameters as in Fig 3.9. (A) Red: background state. Black: retrieval state. Thick traces: mean autocorrelation (AC) functions across 100 randomly sampled neurons with mean firing rate between 1Hz and half of the maximal firing rate (low mean firing rates; dashed) and between half of the maximal firing rate and 65Hz (high mean firing rates; solid). Light traces: AC function for neurons with the fastest and slowest decays, showing a broad range of individual AC timescales. (B) Mean cross-correlation (CC) functions across 200 randomly chosen pairs of neurons with high (i.e. high-high), low (i.e. low-low) and with one neuron high and the other low (i.e. high-low) mean firing rates. Same color code than panel A. (C) Distribution of mean firing rates during the presentation (red) and delay (blue) periods for novel (dashed) and familiar (solid) stimuli.

Across neurons, for both the background and retrieval state, the chaotic fluctuations in the rates have a distinctive times scale of about 100ms (see Fig. 3.10A). However, there is a broad diversity of time scales for individual neurons, ranging from about $\sim$ 50ms to $\sim$ 500ms (see Fig. 3.10A, light traces). Neurons are weakly correlated, for both background and retrieval states (see Fig. 3.10B). Lastly, the distributions of the mean firing rates are qualitatively similar to the ones described for the fixed-point attractor scenario (compare Fig. 3.4C and F with Fig. 3.10C), but with a higher proportion of neurons at very low rates.

3.9 Methods

3.9.1 Static mean field theory

The Model

We consider a network of N neurons with firing rates represented by a vector of analog variables $\vec{r}$. Standard normal patterns of current $\{\vec{\xi}^k\}_{k=1}^p$ with $\xi_i^k \overset{iid}{\sim} \mathcal{N}(0,1)$ are imprinted in the connectivity matrix as the corresponding firing rates elicited by these current patterns, neglecting contributions of the recurrent connections. Hence, the firing rate patterns corresponding to these current patterns are given by $\phi(\xi_i^k)$, where ϕ is the static transfer function of single neurons. In other words, the stored firing rate patterns are standard normal patterns of current $\{\vec{\xi}^k\}_{i=1}^p$ passed through the static transfer function ϕ. Note that in the limit where h_0 is large (see Fig 3.3 B and Eq. (3.48)), these firing rate patterns become distributed according to a log-normal distribution, since the transfer function is exponential in that limit. The rate dependent learning rule is given by two firing rate dependent functions: 1) g which characterizes the dependence on the firing rate of the pre synaptic neuron; 2)f which characterizes the dependence on the firing rate of the post synaptic neuron. With this learning rule, assuming a linear summation of terms corresponding to the different patterns,

as in the Hopfield model (Hopfield 1982) and many of its generalizations, the connectivity matrix is given by

$$J_{ij} = \frac{Ac_{ij}}{cN} \sum_{k=1}^{p} f[\phi(\xi_i^k)] g[\phi(\xi_i^k)], \tag{3.3}$$

where c_{ij} is a sparse directed Erdős-Rényi structural connectivity with each synapse present with probability c, and the pair of functions f and g define together the learning rule. This is a generalization of classical Hebbian learning rules such as the covariance (Sejnowski 1977) and BCM (Bienenstock et al. 1982) since the synaptic strength of the connections between pre and post synaptic neurons is proportional to the product of two functions of their activities. This feature allows a nonlinear dependence of the synaptic strength with the pre and post synaptic activity, but maintains the separability of the learning rule. The operation of f and g under a vector $\vec{r}$, i.e. $f(\vec{r})$ or $g(\vec{r})$, is element-wise. We assume that

$$\int_{-\infty}^{\infty} \mathcal{D}z g(\phi(z)) = 0 \tag{3.4}$$

which ensures that the average change in connection strength due to learning of a single pattern is zero. This could be enforced by a homeostatic mechanism that controls the mean changes in the incoming inputs due to learning (Toyoizumi et al. 2014, Vogels et al. 2011). In our model we assume that both functions f and g are bounded above and below by q_f/q_g and $q_f - 1/q_g - 1$, respectively, where $0 < q_f < 1$, $0 < q_g < 1$. The constant A in Eq. (3.3) controls the strength of the changes in the connectivity due to the learning rule.

The firing rate $r_i(t)$ of each neuron evolve according to standard rate equations (Grossberg 1969, Hopfield 1984), i.e.

$$\tau \dot{r}_i = -r_i + \phi\left(I_i + \sum_{i \neq j}^{N} J_{ij} r_j\right). \tag{3.5}$$

Thus, the steady or attractor state for the dynamics is given by

$$r_i = \phi\left(\sum_{i\neq j}^{N} J_{ij} r_j\right) \quad i = 1, \ldots, N. \tag{3.6}$$

Order parameters - delay period

Throughout this chapter, we will perform a mean field analysis of the steady states of the network in the limits N, cN and p going to infinity, $1 \ll Nc \ll N$ and $p = \alpha/cN$ where α remains of order 1. We consider exclusively steady states that are correlated with a single pattern $\vec{\xi}^1$ but uncorrelated with all other patterns $\vec{\xi}^\mu$ for $\mu > 1$. States with a non-zero correlation with one of the patterns are termed 'retrieval states', while the state with no correlation with any of the patterns is termed 'background state'. The steady state $\vec{r}$ given by Eq. (3.6) depends on the pattern being retrieved $\vec{\xi}^1$ (the 'signal') but also on two sources of frozen noise: 1) the disorder due to the random patterns stored in the connectivity; 2) the disorder given by the structural connectivity C (where C is a binary matrix with entries $c_{ij} \in \{0, 1\}$). The goal of the mean-field analysis is to compute whether and how the network state $\vec{r}$ is correlated with $\vec{\xi}^1$, together with other quantities of interest such as the distribution of firing rates.

The first step in the mean field analysis consists in computing the statistics of the synaptic inputs,

$$h_i = I_i + \sum_{i\neq j}^{N} J_{ij} r_j, \tag{3.7}$$

where the connectivity matrix J_{ij} is given by Eq. (3.3). We first start by the situation in which there are no external inputs, $I_i = 0$. In a delay match to sample experiment, this describes the intervals before presentation of the stimulus, and after this presentation (delay period)

To compute the statistics of synaptic inputs, it is useful to separate the contribution due

to the first pattern ξ_i^1 that the network is trying to retrieve, with the contributions of all other patterns, which will act as noise on the retrieval of the first pattern,

$$h_i = Af(\xi_i^1)\frac{1}{cN}\sum_j c_{ij}g(\phi(\xi_j^1))r_j + Y_i \tag{3.8}$$

where Y_i describes the 'noise' term,

$$Y_i = \frac{A}{cN}\sum_{\mu>1}\sum_j c_{ij}f(\xi_i^\mu)g(\phi(\xi_j^\mu))r_j$$

In the large cN limit, due to the law of large numbers, the first term in Eq. (3.8) converges in probability to $Af(\xi_i^1)q$, where q is given by

$$q = \frac{1}{N}\sum_i g(\phi(\xi_i^1))r_i. \tag{3.9}$$

q is our first order parameter (recall that c_{ij} and ξ_j are independent). It describes how correlated the network state is with a non-linear transformation of the stored pattern ξ_i^1, $g(\phi(\xi_i^1))$. This is a natural generalization of the overlap defined in classical models (Amit et al. 1985) for networks with generalized Hebbian learning rules.

It is instructive to consider first the case in which ξ^1 is the only stored pattern in the connectivity matrix. In this case, the synaptic input to neuron i is uniquely determined by the learning rate A, the post-synaptic function f taken at the firing rate induced by the pattern $\phi(\xi_i^1)$, and q. To compute q, we can use Eq. (3.9), replace r_i by $\phi(h_i)$ where $h_i = Af(\xi_i^1)q$, and replace $1/N\sum_i$ by an integral over the distribution of ξ_i,

$$q = \int \mathcal{D}\xi g(\phi(\xi))\phi(Af(\phi(\xi))q), \tag{3.10}$$

where $\mathcal{D}\xi$ denotes the Gaussian measure $d\xi e^{-\xi^2/2}/\sqrt{2\pi}$. Eq. (3.10) can be solved to obtain

the possible values of q given f, g and A. Note that $q = 0$ (corresponding to the background state) is always a solution to this equation, due to Eq. (3.4).

In the case in which many patterns are stored in the connectivity matrix, we need to compute the statistics of the noise term Y_i. In the large p, N limits, this term becomes distributed according to a Gaussian distribution with zero mean (since the average of $g(\phi(\xi))$ over the distribution of ξs is zero) and a variance given by

$$\mathrm{Var}(Y_i) = \alpha\gamma M$$

where

$$\gamma \equiv A^2 \int_{-\infty}^{\infty} \mathcal{D}\xi f^2(\phi(\xi))) \int_{-\infty}^{\infty} \mathcal{D}\xi g^2(\phi(\xi))), \tag{3.11}$$

and M is our second order parameter, which is equal to the average squared firing rate over the network,

$$M = \frac{1}{N}\sum_i r_i^2. \tag{3.12}$$

In this calculation the independence between ξ_j from r_j is assumed. The final step is to compute the order parameters self-consistently. For this, we use the fact that Y_i is a Gaussian random variable with zero mean and variance $\alpha\gamma M$, replace r_i by $\phi(qAf(\phi(\xi_i^1)) + Y_i)$ in Eqs. (3.9,3.12) and replace the sums over i by a double integral over the distributions of ξ_i and Y_i, leading to

$$q = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y g(\phi(z))\phi(qAf(\phi(z)) + \sqrt{\alpha\gamma M}y) \tag{3.13}$$

$$M = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y \phi^2(qAf(\phi(z)) + \sqrt{\alpha\gamma M}y). \tag{3.14}$$

The overlap m, which corresponds to the correlation between $g(\phi(\xi))$ and the firing rates

r, is given by

$$m = \frac{q}{(M - R^2)\sqrt{\int_{-\infty}^{\infty} \mathcal{D}z g(\phi(z))^2}}, \tag{3.15}$$

where R is the mean firing rate in the attractor state given by

$$R = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{D}z \mathcal{D}y \phi(qAf(\phi(z)) + \sqrt{\alpha\gamma M} y). \tag{3.16}$$

Distributions of firing rates - delay period

To compute the distribution of firing rates, we use the fact that the distribution of synaptic inputs conditioned on the pattern being retrieved is Gaussian,

$$p\left(h|\xi^1 = z\right) = \mathcal{N}\left(Af(\phi(z))q, \alpha\gamma M\right), \tag{3.17}$$

where the order parameters q and M are determined by the self-consistent equations (3.13) and (3.14).

Using the fact that the transfer function is non-decreasing, we obtain the distribution of steady state firing rates conditional to the pattern $\vec{\xi}^1$ presented during the delay period

$$p_r(r|\xi^1 = z) = \frac{1}{\sqrt{2\pi\alpha\gamma M}} \exp\left(-\frac{\left(\phi^{-1}(r) - Af(z)q\right)^2}{2\alpha\gamma M}\right) \frac{d\phi^{-1}(r)}{dr}. \tag{3.18}$$

From this conditional probability distribution, we obtain the marginal distribution of firing rates at the steady state, r,

$$p_r(r) = \int_{-\infty}^{\infty} \mathcal{D}z \frac{1}{\sqrt{2\pi\alpha\gamma M}} \exp\left(-\frac{\left(\phi^{-1}(r) - Af(z)q\right)^2}{2\alpha\gamma M}\right) \frac{d\phi^{-1}(r)}{dr}. \tag{3.19}$$

Order parameters and distributions of firing rates - presentation period

A similar analysis can be done in the situation when an external stimulus is presented to the network. We consider here two scenarios, one in which the presented stimulus is one of the stored patterns, $I_i = \xi_i^1$ (a 'familiar' stimulus), and the other in which the stimulus is uncorrelated with the stored patterns (a 'novel' stimulus).

In the 'novel' case, the synaptic inputs are

$$h_i = I_i + Y_i \tag{3.20}$$

where the external stimulus $\{I_i\}$ is independently sampled from a normal distribution with mean zero and variance I_0 (i.e. $I_i \overset{iid}{\sim} \mathcal{N}(0, I_0^2)$), where I_0 is the amplitude of the stimulation. For consistency reasons we use $I_0 = 1$ in all the results shown in this chapter, but show here calculations for arbitrary I_0. The stimulus $\vec{I}$ is independent of all the previous patterns learned $\{\vec{\xi}^k\}_{k=1}^p$. Therefore, the synaptic inputs are the sum of two uncorrelated Gaussian random variables, one with variance I_0^2, the other with variance $\alpha\gamma M$. Hence, they are distributed according to a Gaussian of variance $\sqrt{I_0^2 + \alpha\gamma M}$.

Since the stimulus is uncorrelated with all stored patterns, the overlap q is equal to zero, while the other order parameter M is given by

$$M = \int_{-\infty}^{\infty} \mathcal{D}z \phi^2(\sqrt{I_0^2 + \alpha\gamma M} z). \tag{3.21}$$

The distribution of firing rates during the presentation period for a novel stimulus is a distribution of a Gaussian of mean zero and variance $\sqrt{I_0^2 + \alpha\gamma M}$ passed through the non-linear function ϕ and is therefore given by

$$p_{\text{pres}}^{\text{nov}}(r) = \frac{1}{\sqrt{2\pi(I_0^2 + \alpha\gamma M)}} \frac{d\phi^{-1}(r)}{dr} \exp\left(-\frac{(\phi^{-1}(r))^2}{2(I_0^2 + \alpha\gamma M)} \right). \tag{3.22}$$

In the 'familiar' case, the synaptic inputs during presentation of the pattern become

$$h_i = I_0 \xi_i^1 + qAf(\phi(\xi_i^1)) + Y_i \tag{3.23}$$

where the first term in the r.h.s. of Eq. (3.23) is due to the external input, and the two other terms are identical to the situation analyzed in the previous section. Again, we use in all results shown in this chapter $I_0 = 1$ but show the calculations for arbitrary I_0.

The distribution of the synaptic inputs, conditioned on the pattern ξ_i^1, has now a mean $I_0\xi_i^1 + qAf(\phi(\xi_i^1))$, and a variance $\alpha\gamma M$. This leads to the following equations for the order parameters q and M,

$$q = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y g(\phi(z))\phi(I_0 z + Af(\phi(z))q + \sqrt{\alpha\gamma M}y) \tag{3.24}$$

$$M = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y \phi^2(I_0 z + Af(\phi(z))q + \sqrt{\alpha\gamma M}y), \tag{3.25}$$

while the distribution of firing rates is

$$p_{\text{pres}}^{\text{fam}}(r) = \frac{1}{\sqrt{2\pi\alpha\gamma M}} \frac{d\phi^{-1}(r)}{dr} \int_{-\infty}^{\infty} \mathcal{D}z \exp\left(-\frac{(\phi^{-1}(r) - I_0 z - Af(\phi(z))q)^2}{2\alpha\gamma M}\right). \tag{3.26}$$

MFT when f and g are step functions

Here we take f and g to be step functions (i.e. $\beta_f, \beta_g \to \infty$) with the same threshold, i.e.:

$$f(\eta) = \begin{cases} q_f & \eta \geq x_f \\ -(1-q_f) & \eta < x_f \end{cases} \tag{3.27}$$

and

$$g(\eta) = \begin{cases} q_g & \eta \geq x_f \\ -(1-q_g) & \eta < x_f. \end{cases} \tag{3.28}$$

The condition $\int_{-\infty}^{\infty} \mathcal{D}\xi g(\phi(\xi))) = 0$ implies that

$$q_g = \int_{-\infty}^{x_f} dr \frac{\frac{d\phi^{-1}(r)}{dr}}{\sqrt{2\pi}} e^{-\frac{(\phi^{-1}(r))^2}{2}}.$$

The mean field equations simplify to

$$\begin{aligned} q &= q_g(1-q_g)\left\{\int_{-\infty}^{\infty} \mathcal{D}y\phi\left(A\sqrt{\tilde{\gamma}}\left[\left(\frac{q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right)\right. \\ &\quad \left. - \int_{-\infty}^{\infty} \mathcal{D}y\phi\left(A\sqrt{\tilde{\gamma}}\left[-\left(\frac{1-q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right)\right\} \qquad (3.29) \\ M &= (1-q_g)\int_{-\infty}^{\infty} \mathcal{D}y\phi^2\left(A\sqrt{\tilde{\gamma}}\left[q\left(\frac{q_f}{\sqrt{\tilde{\gamma}}}\right) + \sqrt{\alpha M}y\right]\right) \\ &\quad + q_g\int_{-\infty}^{\infty} \mathcal{D}y\phi^2\left(A\sqrt{\tilde{\gamma}}\left[-\left(\frac{1-q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right) \qquad (3.30) \end{aligned}$$

where

$$\tilde{\gamma} = \int_{-\infty}^{\infty} \mathcal{D}\xi\left(g(\phi(\xi))^2\right)\int_{-\infty}^{\infty} \mathcal{D}\xi\left(f(\phi(\xi))^2\right) = q_g(1-q_g)\left[q_f^2(1-q_g) + (1-q_f)^2 q_g\right].$$

Defining

$$\begin{aligned} m_0 &\equiv \frac{q}{r_m q_g(1-q_g)} \qquad (3.31) \\ M_0 &\equiv \frac{M}{r_m^2} \qquad (3.32) \\ \bar{A} &\equiv A r_m\sqrt{\tilde{\gamma}} \qquad (3.33) \\ \psi(x) &\equiv \frac{\phi(x)}{r_m} \qquad (3.34) \\ p &\equiv 1-q_g \qquad (3.35) \\ \eta &\equiv \sqrt{\frac{q_g(1-q_g)}{q_f^2(1-q_g) + (1-q_f)^2 q_g}}, \qquad (3.36) \end{aligned}$$

we obtain

$$m_0 = \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[q_f\eta m_0 + \sqrt{\alpha M_0}y\right]\right) - \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[-(1-q_f)\eta m_0 + \sqrt{\alpha M_0}y\right]\right) \tag{3.37}$$

$$\begin{aligned} M_0 &= p\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[q_f\eta m_0 + \sqrt{\alpha M_0}y\right]\right) \\ &\quad +(1-p)\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[-(1-q_f)\eta m_0 + \sqrt{\alpha M_0}y\right]\right). \end{aligned} \tag{3.38}$$

When $q_f = q_g$, the mean field equations read

$$m_0 = \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[(1-p)m_0 + \sqrt{\alpha M_0}y\right]\right) - \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[-pm_0 + \sqrt{\alpha M_0}y\right]\right) \tag{3.39}$$

$$\begin{aligned} M_0 &= p\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[(1-p)m_0 + \sqrt{\alpha M_0}y\right]\right) \\ &\quad +(1-p)\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[-pm_0 + \sqrt{\alpha M_0}y\right]\right). \end{aligned} \tag{3.40}$$

Solutions to this equation are numerically explored in Fig. 3.7C and D.

In the limit $\bar{A} \to \infty$, the function $\psi(\bar{A}x)$ become a step (Heaviside) function, $\psi(\bar{A}x) \to 1$ if $x > 0$, 0 otherwise. Consequently, the mean field equations become

$$m_0 = \Phi\left(\frac{-(1-p)m_0}{\sqrt{\alpha M_0}}\right) - \Phi\left(\frac{pm_0}{\sqrt{\alpha M_0}}\right) \tag{3.41}$$

$$M_0 = p\Phi\left(\frac{-(1-p)m_0}{\sqrt{\alpha M_0}}\right) + (1-p)\Phi\left(\frac{pm_0}{\sqrt{\alpha M_0}}\right), \tag{3.42}$$

where $\Phi(x) = \int_x^{\infty} dx e^{-x^2/2}/\sqrt{2\pi}$. These equations are identical to equations (20) and (21) derived by Tsodyks (1988) in a sparsely connected network of binary 0,1 neurons (with a threshold θ_0) storing binary random patterns with coding level p, with $\theta_0 = 0$. Note that the full equations derived by Tsodyks can be recovered when the threshold of the transfer function scales as $h_0 = \bar{A}\theta_0$.

Using these equations, Tsodyks found that the capacity diverges in the sparse coding limit as $\alpha_c \approx \frac{\theta_0^2}{2p\log(1/p)}$(Tsodyks 1988). In our network, the capacity cannot diverge in the $p \to 0$ limit due to the fact that $\theta_0 = 0$, since h_0 is a fixed parameter and therefore does not scale with $\bar{A}$. However, optimizing the threshold of the transfer function together with the parameters of the learning rule would allow one to reach the same scaling as the one obtained by Tsodyks (1988). This would require setting $h_0 = \bar{A}\theta_0$.

To obtain the capacity of our network, i.e. the largest value of α for which we can find a solution of Eqs. (3.41,3.42) with $m_0 > 0$, we analyze the Jacobian of the right side of equations (3.41) and (3.42) in the limit $m_0 \to 0^+$ (i.e. when the overlap approaches to zero) which gives

$$\mathbb{J} = \begin{pmatrix} \frac{1}{\sqrt{\pi\alpha}} & 0 \\ 0 & 0 \end{pmatrix}.$$

By you doing a linear expansion around $m_0 = 0^+$, we study the stability of retrieval states close to capacity. For equations (3.41) and (3.42) to have a stable solution in the limit $m_0 \to 0$, the eigenvalues of the Jacobian have to be less than one. This leads to the maximal capacity

$$\alpha_c = \frac{1}{\pi} \approx 0.318, \tag{3.43}$$

for all p.

Since the trace of the Jacobian is zero at the critical point, then the phase transition is of the second order (see Fig. 3.7 A and B). The parameter p has no effect on the capacity for this limit and the capacity is much lower than what has been found for the best-fit median parameters. For $q_f \neq q_g$, it is straightforward to show that the capacity is

$$\alpha_c = \frac{\eta^2}{\pi} \tag{3.44}$$

for all p.

This is always lower or equal than what is found in Eq. (3.43) since $\max_{q_f \in [0,1]}(\eta) = 1$ with $\text{argmax}_{q_f \in [0,1]}(\eta) = q_g = 1 - p$.

3.9.2 Simulations

For most simulations shown in this chapter, the probability of connections was set to 0.5% (i.e. $c = 0.005$) and the number of neurons to $N = 50000$, which implies an average number of connections per neuron of $Nc = 250$. The choice of a low connection probability was motivated by the fact that the MFT is exact in the sparse connectivity limit (see static mean field theory and Derrida et al. (1987), Kree & Zippelius (1987)). We have also simulated networks with with various values of N and c (see Fig. 3.11). These simulations show that our theory gives good quantitative predictions for denser connectivities. The single neuron time constant was chosen as $\tau = 20ms$, similar to time constants of single neurons (McCormick et al. 1985) and synapses (Destexhe et al. 1998), and with the decay time constant of cortical activity as measured *in vivo* (Reinhold et al. 2015). Open source built-in linear algebra methods in scipy and numpy Python packages suited for sparse matrices were used to generate the connectivity matrix. For simulating the networks dynamics, the Euler method was used with a time step size of 0.5ms. For a few parameter sets, we checked that results are unchanged when a smaller value of $dt = 0.1$ms is used. In the simulations, the background state was sometimes unstable, and the dynamics in this case converged to one of the 'memory states'. This tended to happen in particular for small values of α.

In Fig. 3.9 G-I, the Runge-Kutta fourth-order method with $dt = 0.1$ms was used. In Fig. 3.10 the auto- and cross-correlation functions are computed over 100 realizations of a 8s network simulation. For retrieval states, in each realization the input current is given by the current corresponding to the stored pattern plus a random vector whose entries are i.i.d. random Gaussian variables with zero mean and S.D. 0.2. For the background state, the initial condition of the dynamics are the firing rates obtained from passing an i.i.d. standard

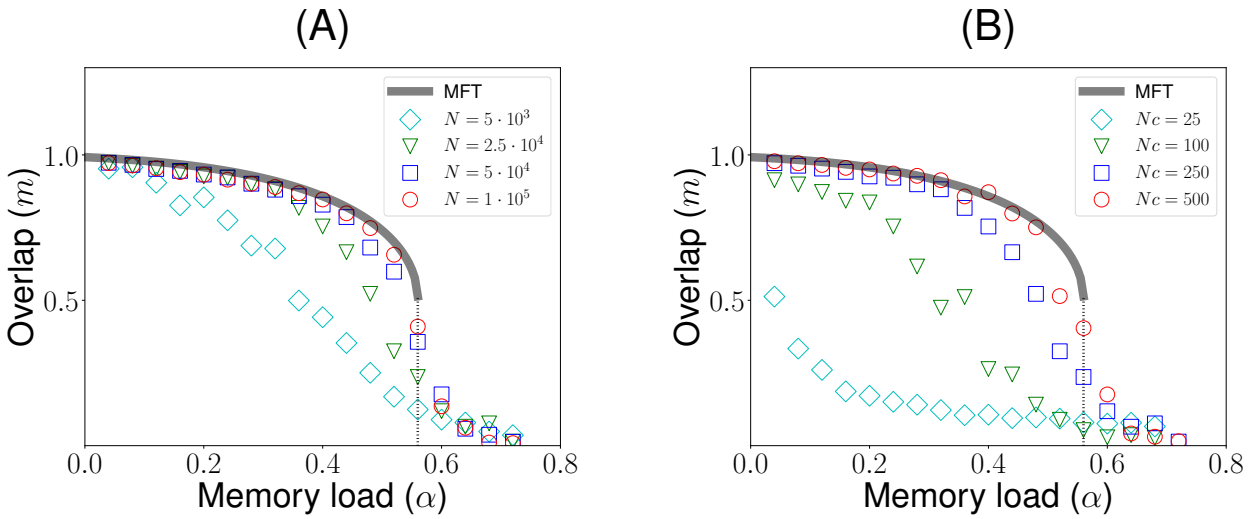

Figure 3.11: Finite size effects. Our MFT is valid in the large population size ($N \to \infty$) and large mean in-degree ($Nc \to \infty$) limit, such that the number of neurons is much larger than the average number of synapses per neuron ($1 \ll Nc \ll N$). Here we explore the effects of varying the population size (N) and mean in-degree (Nc) on the overlap in retrieval states. (**A**) Overlap as a function of memory load α for various values of N, at fixed $cN = 250$. Grey curve: MFT. Symbols: average overlap (computed from 50 realizations) in simulations with population sizes $N = 5000, 25000, 50000, 100000$. (**B**) Overlap as a function of memory load α for various mean in-degrees, for $N = 25000$. Grey curve: MFT. Symbols: average overlap in simulations with average mean in-degrees $Nc = 25, 100, 250, 500$. Parameters for ϕ, f and g are chosen as in Fig. 3.5A on the main text.

normal vector through the transfer function ϕ. The first second of simulation is not taken into account to compute auto and cross-correlation functions. Only neurons with mean firing rates between 1Hz and 65Hz are selected in order to avoid numerical artifacts arising from neurons whose mean firing rates stay close to zero or to the maximum firing rate during most of the simulation.

To measure the sensitivity of the network dynamics to small perturbations, we choose two slightly different initial conditions and follow the dynamics of the network following both initial conditions, to investigate whether these two initial conditions converge to the same state (indicating non-chaotic dynamics), or vice versa diverge exponentially (indicating chaotic dynamics). These two slightly different initial conditions are generated as follows

$$\vec{r}_k^{(1)}(0) = \phi(\vec{\xi}^k) \tag{3.45}$$

$$\vec{r}^{(2)}(0) = \phi(\vec{\xi}^k) + \vec{\eta}\frac{\delta}{\|\vec{\eta}\|_2}. \tag{3.46}$$

where the index k corresponds to one of the p stored patterns (i.e. $k \in \{1, 2, \ldots, p\}$), $\delta = 10^{-3}$ is the distance between the initial conditions and $\vec{\eta}$ is an independent and identically distributed Gaussian vector. Thus, $\vec{r}_k^{(1)}(0)$ is the firing rate produced by the k^{th} stored pattern, while $\vec{r}^{(2)}(0)$ is a slightly perturbed version of this pattern. We define the distance between the two network states during the time evaluation of the dynamics by

$$d_k(t) = \frac{\left\|\vec{r}_k^{(1)}(t) - \vec{r}_k^{(2)}(t)\right\|_2}{\sqrt{N}}. \tag{3.47}$$

This distance gives the typical difference between the firing rates of a single neuron between two network states produced by slightly different initial conditions at time t, for the retrieval state corresponding to pattern k, and has units of Hz.

3.9.3 Data analysis

We reanalyze the data recorded by Luke Woloszyn and David Sheinberg (Woloszyn & Sheinberg 2012) using the method described in Lim et al. (2015). This data consists in trial-averaged firing rates of individual neurons in ITC (in a time window between 75 ms and 200 ms after stimulus onset) in response to 125 novel and 125 familiar stimuli measured, during a passive fixation task. We focused on the 30 putative excitatory neurons whose distributions of visual responses for novel and familiar stimuli were significantly different, using the Mann-Whitney U test at 5 significance level. In these neurons, the postsynaptic dependence of the learning rule, was inferred using the method described in Lim et al. (2015). In this subset of neurons, we focused on 14 excitatory neurons, the ones that show negative input changes for low firing rates and positive input changes for high firing rates. For these 14 neurons, the transfer function ϕ, and the postsynaptic dependence of the learning rule, f, are inferred using the method described in Lim et al. (2015).

The first step is to infer the transfer function ϕ. We assume that inputs to neurons during presentation of novel stimuli have a Gaussian distribution. The transfer function is then obtained as the function ϕ that maps a standard Gaussian to the empirical distribution of firing rates for novel stimuli (Lim et al. 2015). In practice, the function is obtained by building a quantile-quantile plot between the distribution of firing rates for novel stimuli and the assumed standard normal distribution of inputs (see Fig. 3.3 A and B; Fig. A.1 and A.2). The obtained transfer function (blue circles in Fig. 3.3) was fitted with the sigmoidal function

$$\phi_i(\xi) = \frac{r_m^{(i)}}{1 + e^{-\beta_T^{(i)}(\xi - h_0^{(i)})}} \tag{3.48}$$

where $r_m^{(i)}$ is the maximal firing rate, $\beta_T^{(i)}$ measures the slope at the inflection point, and $h_0^{(i)}$ is the location of this inflection point. h_0 is also the current leading to half maximal firing rate. These parameters were obtained by minimizing the squared error. We thus obtained

for each of the 14 neurons the best estimators $r_m^{(i)}$, $\beta_T^{(i)}$ and $h_0^{(i)}$ with $i = 1, 2, \ldots, 14$ whose statistics are summarized in Fig. 3.3D.

The next step is to infer the postsynaptic dependence of the learning rule, f. For this, we use the difference between the distributions of visual responses to novel and familiar stimuli (Lim et al. 2015). In the model, learning of a novel stimulus defined by inputs ξ_i^k that leads to firing rates $r_i^k = \phi(\xi_i^k)$ leads to changes in recurrent inputs, due to changes in synaptic inputs

$$\Delta J_{ij} = \frac{A c_{ij}}{cN} f(r_i^k) g(r_j^k) \tag{3.49}$$

This leads to a change in total inputs to neurons that is proportional to

$$\Delta h_i = A f(r_i^k) \frac{1}{cN} \sum_j c_{ij} g(r_j^k) r_j^k \tag{3.50}$$

In the large N limit, Eq. (3.50) becomes

$$\Delta h_i = A f(r_i^k) \int_{-\infty}^{\infty} \mathcal{D}z g(\phi(z)) \phi(z). \tag{3.51}$$

where $\mathcal{D}z$ is the standard Gaussian measure, $\mathcal{D}z = dz e^{-z^2/2}/\sqrt{2\pi}$. Eq. (3.51) give us the relationship between changes of total inputs to a neuron with learning of a particular stimulus, and the firing rate of the neuron upon presentation of that stimulus for the first time. This relationship can be inferred from the data by computing the difference between the quantile function of visual responses to familiar stimuli and the quantile function of visual responses to novel stimuli, and by plotting this difference as a function of visual response to novel stimuli (Lim et al. 2015). We then fitted the input change with a sigmoidal function given by

$$\Delta h_i^{fit}(r) = \frac{C^{(i)}}{2} \left[2q_f^{(i)} - 1 + \tanh(\beta_f^{(i)}(r - x_f^{(i)})) \right]. \tag{3.52}$$

where $C^{(i)}$ gives the amplitude of the total changes, q_f^i measures the vertical offset of the

curve (for $q_f = 1$, Δh is non-negative at all rates, while for $q_f = 0$ it is non-positive at all rates), $\beta_f^{(i)}$ measures the slope at the inflection point, and $x_f^{(i)}$ is the rate at the inflection point. In the following, we refer to $x_f^{(i)}$ as the threshold since it is typically very close to the rate at which Δh changes sign. For each of the 14 neurons, the parameters $C^{(i)}$, $q_f^{(i)}$, $\beta_f^{(i)}$ and $x_f^{(i)}$ with $i = 1, 2, \ldots, 14$ were estimated by minimizing the squared error. The inferred function f for each neuron is given by

$$f_i(r) = \frac{\Delta h_i^{fit}(r)}{C^{(i)}} = \frac{1}{2}\left[2q_f^{(i)} - 1 + \tanh(\beta_f^{(i)}(r - x_f^{(i)}))\right]. \tag{3.53}$$

The parameter A is then obtained as

$$A^{(i)} = \frac{C^{(i)}}{\int_{-\infty}^{\infty} \mathcal{D}z g(\tilde{\phi}(z))\tilde{\phi}(z)}, \tag{3.54}$$

where $\tilde{\phi}$ is the sigmoidal transfer function in Eq. (3.54) whose parameters are the medians of the fitted parameters. The function g was also chosen to be a sigmoid, given by

$$g(r) = \frac{1}{2}\left[2q_g - 1 + \tanh(\beta_g(r - x_g))\right], \tag{3.55}$$

with q_g set such that the average change in connection strength due to learning of a single pattern is zero, i.e.

$$\int_{-\infty}^{\infty} \mathcal{D}z g(\tilde{\phi}(z)) = 0. \tag{3.56}$$

Note that g is unconstrained by data. For most of the paper, we set the slope and the threshold for g to the median of the fitted parameters for f, i.e. $\beta_g = \tilde{\beta}_f$ and $x_g = \tilde{x}_f$. We also explored how the capacity depends on β_g and x_g, as shown in Fig. 3.5.

3.10 Discussion

We have shown that a learning rule inferred from data generate attractor dynamics, without any need for parameter adjustment or tuning, except for the condition that the dependence of the learning rule on the presynaptic rate should be 'balanced' (i.e. have a zero average over the distribution of visual responses, see below). Furthermore, this rule produces a storage capacity that is close to the maximal capacity, in the space of unsupervised Hebbian learning rules with sigmoidal dependence on both pre and post-synaptic firing rates. Remarkably, similar to the learning rules inferred from ITC recordings, learning rules derived from memory storage maximization depress the bulk of the distribution of the learned inputs (those that lead to low to intermediate firing rates) while potentiating outliers (those that lead to high rates), leading to a sparse representation of stored memories. The attractor states generated by our model are characterized by graded activity with a continuous range of firing rates (Treves 1990*a*,*b*, Festa et al. 2014). Most of the distribution lies in the low rate region of the neuronal transfer function, leading to a strongly skewed distribution, with a small fraction of neurons firing at higher rates. These observations are consistent with the available data in ITC during delay match to sample experiments (Miyashita 1988, Nakamura & Kubota 1995*a*).

For a range of parameters values consistent with learning rules inferred from data, our model presents irregular temporal dynamics for retrieval states, similar to the temporal and across trial variability observed during delay periods in multiple studies (Murray et al. 2017). In this regime, retrieval states are chaotic, yet they maintain non-zero overlap with the corresponding memories. Thus, the network performs robustly as an associative memory device, even though strong fluctuations are internally generated by its own chaotic dynamics.

3.10.1 Distribution of firing rates

Our model naturally gives rise to highly skewed distributions of firing rates, consistent with those that have been observed during presentation of visual stimuli in ITC (Lehky et al. 2011, Lim et al. 2015) and during delay periods of DMS tasks (Miyashita 1988, Nakamura & Kubota 1995*a*). By construction of the model, it also reproduces the decrease in the mean response with familiarity, and the increase in selectivity with familiarity. Our model shows for most of the explored parameter space a weak bimodality in the distribution of firing rates due to neuronal saturation in response to familiar stimuli, with a tiny peak close to neuronal saturation, when the network is homogeneous. When heterogeneity in maximal firing rates is implemented in the network, the peak at high firing rates disappears and the distribution of firing rates becomes unimodal.

3.10.2 Learning rule

The learning rule we have used in our network model was inferred from ITC data (Lim et al. 2015). It is an unsupervised Hebbian rule, as it only depends on the pre and post-synaptic firing rates, and it leads to potentiation for large pre and post-synaptic rates. As other popular examples of Hebbian rules such as the covariance rule (Sejnowski 1977) or the BCM rule (Bienenstock et al. 1982), it is separable in pre and post-synaptic rates. Unlike the covariance rule, but similar to other Hebbian rules (Bienenstock et al. 1982, Senn et al. 2001, Pfister & Gerstner 2006), it is strongly non-linear as a function of the post-synaptic firing rate. It reproduces some of the phenomenology of the dependence of synaptic plasticity on pre and post-synaptic firing rates in cortical slices; in particular, large pre and post-synaptic firing rates lead to LTP (Sjöström et al. 2001). Large pre-synaptic firing rate in conjunction with low post-synaptic firing rate, lead to depression, consistent with 'pairing' experiments in which LTD is triggered by pre-synaptic activity, together with intermediate values of the membrane potential (Ngezahayo et al. 2000). Plasticity at low

pre-synaptic firing rates could be due to plasticity mechanisms leading to 'normalization' or homeostasis. Indeed, our plasticity rule could be written as $\Delta J_{ij} = \Delta J_{ij}^{Hebb} + \Delta J_{ij}^{hom}$ where $\Delta J_{ij}^{Hebb} = Af(r_i)(g(r_j) - g(0))$, $\Delta J_{ij}^{hom} = Af(r_i)g(0)$. The 'homeostatic' component ΔJ_{ij}^{hom} leads to a decrease in the efficacy of all synapses onto a post-synaptic neuron when the neuron is firing at high rates, while it leads to an increase when the neuron fires at low rates (since $g(0) < 0$). Note that such a homeostatic mechanism would also automatically lead to a 'balanced' dependence of the rule of the pre-synaptic firing rate, which is necessary for the network to be able to store a large nuber of patterns. The analysis described in the Supplementary Material shows that if g has a non-zero average, then the mean of the noise term due to other patterns stored in the connectivity matrix would no longer be zero, but rather scale as $\alpha cN\langle g\rangle$, where $\langle g\rangle$ is the average of g over the distribution of visual responses. This has the consequence that the network would be able to store only a finite number of patterns. A precise balance could be restored by the homeostatic mechanism mentioned above - for a non-zero $\langle g\rangle$, this homeostatic term would become $\Delta J_{ij}^{hom} = Af(r_i)(g(0) - \langle g\rangle)$, which would ensure that the average synaptic strength (and consequently mean firing rate) onto a neuron remains constant with learning.

The synaptic connectivity matrix we used is assumed to be generated through multiple presentations of initially novel patterns. The simplest implementation of this plasticity rule consists in adding a term ΔJ_{ij} to the current matrix, as described above, but only when a novel pattern is presented to the network. This would require a novelty detector that would gate plasticity, perhaps through neuromodulators. An interesting hypothesis is that novelty detection could be generated by the network itself, through its mean activity (which is significantly higher for novel than for familiar stimuli). This novelty signal could in principle then be used to trigger learning.

To derive the learning rule, we used a subset of the data recorded by Woloszyn & Sheinberg (2012), i.e. excitatory neurons that show negative changes at low rates and positive

changes at high rates. Those neurons are approximately half (14/30) of the neurons that showed significant differences between the distributions of visual responses for familiar and novel stimuli. Out the remaining 16 neurons, 10 showed negative changes for all rates, while 6 showed the opposite pattern of positive changes for all rates. This heterogeneity in inferred learning rules could be due to a heterogeneity in neuronal properties - for instance, it could be that the 'putative' excitatory neurons recorded in this study form a heterogeneous group of cells, some of which might actually be inhibitory. Consistent with this, some inhibitory neuron classes have electrophysiological properties (and in particular, spike width) that are closer to pyramidal cells that to fast-spiking interneurons. Another possibility is that part of the apparent heterogeneity stems form the same underlying learning rule, but with heterogeneous parameters. For instance, inferred learning rules with negative changes at all rates are consistent with a sigmoidal post-synaptic dependence f, but with a high threshold x_f that lies above the range of firing rates elicited in that particular experiment. Elucidating which of these scenarios hold in IT cortex will need recordings from more neurons, as well as recordings of single neurons with more stimuli.

Our approach is complementary to other studies that have inferred learning rules from *in vitro* studies, and then shown that these rules lead to attractor dynamics in large networks of spiking neurons (Litwin-Kumar & Doiron 2014*b*, Zenke et al. 2015). In contrast to these studies, we showed that a network with a learning rule inferred from *in vivo* data can achieve a high storage capacity, and generate graded distributions of firing rates during visual presentation and delay periods. An important difference between the studies of Litwin-Kumar & Doiron (2014*b*) and Zenke et al. (2015) is that they used an online learning rule that is constantly active, while our connectivity matrix is assumed to be frozen following the learning process. It will be interesting to investigate whether, and in which conditions spike-timing and voltage based learning rules used in such studies can produce a firing rate dependence that is consistent with the rule used here.

3.10.3 Time-varying neural representations

In recent years, the standard attractor network scenario has been challenged by multiple observations of strong variability and non-stationarity during the delay period in prefrontal cortex (Compte et al. 2003, Shafi et al. 2007, Barak et al. 2010, Barak & Tsodyks 2014, Kobak et al. 2016, Murray et al. 2017). Statistical analysis of recordings in this area during two different working memory tasks has shown that variability observed during delay periods is consistent with static coding of the stimulus kept in memory (Murray et al. 2017). Various models have been proposed to account for variability and/or non-stationarity (Barbieri & Brunel 2007, Mongillo et al. 2008, Lundqvist et al. 2010, Mongillo et al. 2012, Druckmann & Chklovskii 2012).

Here we propose an alternative mechanism where chaotic attractors with associative memory properties naturally generate the time-varying irregular activity observed during delay periods in associative memory tasks. In this state, chaotic attractors correspond to internal representations of stored memories. Each chaotic attractor state maintains a positive overlap with the corresponding stored memory. In this scenario, the network performs as an associative memory device where temporal variability is generated internally by chaos. This model naturally exhibits the combination of strong temporaly dynamics yet stable memory encoding which has been demonstrated in PFC by various groups (Druckmann & Chklovskii 2012, Murray et al. 2017). It will be interesting to compare this model to existing data, using for instance methods used in Murray et al. (2017).

There has been a longstanding debate whether the type of chaotic states seen in firing rate models can be seen also in spiking network models under the form of 'rate chaos'. Recent studies indicate that this type of chaos can be observed provided coupling is sufficiently strong, as in firing rate models Ostojic (2014), Harish & Hansel (2015), Kadmon & Sompolinsky (2015). Thus, it is reasonable to expect that the type of retrieval chaotic states we observed in our network can also be realized in networks of spiking neurons.

3.10.4 Optimality criteria for information storage

Here, we have argued that learning rules that are inferred from electrophysiological recordings in ITC of behaving primates are close to optimizing information storage, in the space of unsupervised Hebbian learning rules that have a sigmoidal dependence on both pre and post-synaptic firing rates. Such learning rules are appealing because synapses do not need to know anything beyond the firing rates of pre and post-synaptic neurons to form memories, two quantities that are easily available at a synapse. However, one cannot exclude that the dependence of plasticity on neuronal activity takes other forms than the one investigated here. In particular, a potentially more powerful approach proposed by Gardner (1987) relies in maximizing the number of attractors in the space of all possible synaptic matrices. Unsurprisingly, this approach leads in general to a larger capacity than the ones that can be achieved by unsupervised Hebbian rules, but it turns out that in sparse coding limit, the covariance rule reaches asymptotically the Gardner bound (Tsodyks & Feigel'Man 1988, Tsodyks 1988). These results have been obtained in networks of binary neurons, and it remains to be investigated whether similar results could be obtained in networks of analog firing rate neurons. An additional challenge in comparing the two approaches in such networks is that the stored attractors are in our case not identical to the pattern that was initially shown to the network, while in the standard Gardner approach, the two were constrained to be identical.

Another motivation for considering the Gardner approach is provided by a recent study that showed that synaptic connectivity in a network of excitatory binary neurons that maximizes storage capacity in the space of all possible matrices reproduces a number of basic experimental facts on cortical excitatory connectivity (Brunel 2016): Low connection probability (Markram, Lübke, Frotscher, Roth & Sakmann 1997, Sjöström et al. 2001, Lefort et al. 2009), in spite of full potential connectivity (Kalisman et al. 2005); And strong overrepresentation of bidirectionnally connected pairs of neurons compared to a random Erdos-

Renyi network (Sjöström et al. 2001). In contrast with the network studied by Brunel (2016), the synaptic connectivity of the model proposed here has the unrealistic feature that it does not obey Dale's law. One could reconcile the present model with cortical connectivity by using a connectivity matrix that is a rectified version of Eq. (3.2) - such a connectivity matrix would then obey Dale's law, be sparse and be more symmetric than a random Erdos-Renyi network, making it therefore consistent with slice data. Such a generalization is beyond the scope of the present paper and will be the subject of a future study.

Altogether, our results strongly reinforce the link between attractor network theory and electrophysiological data during delayed response tasks in primates. Furthermore, they suggest that learning rules in association cortex are close to maximizing the number of possible internal representations of memories as attractor states.

CHAPTER 4

MEMORY AND CHAOS IN NEURONAL NETWORKS

4.1 Contribution

The work presented in this chapter is part of a manuscript in preparation for publication. The authors are Ulises Pereira, Yonatan Aljadeff and Nicolas Brunel. U.P. and N.B. designed the research. U.P. and Y.A. performed the mean field theory (MFT) calculations. U.P. performed the the numerical solutions to the MFT and the network simulations. U.P. wrote the manuscript with inputs from Y.A. and N.B..

4.2 Introduction

Attractor networks are an influential theory for memory storage in brain systems (Hopfield 1982, Amit et al. 1985, Amit 1992, Brunel 2005). In this theory, memories correspond to fixed-point attractors states, which are stable patterns of network activity *representing* the stored memoranda. When a memory is learned, changes in the connectivity through synaptic plasticity driven by an external input to the network produce a distributed connectivity pattern of synaptic modifications. These changes in the connectivity create a fixed-point attractor corresponding to the *neural representation* of the learned memorandum. In the attractor state, the network activity is correlated with, but not identical to, the original external input to the network. Upon an external cue correlated with the stored memorandum that is being retrieved, the network autonomously relaxes to the corresponding attractor state, and the identity of the memorandum can be easealy decoded by downstream circuitry.

The theory parsimoniously reproduces selective persistent activity (Goldman-Rakic 1995), i.e. stable elevated activity during delay periods, which is a salient feature observed in neural recordings in monkeys (Fuster et al. 1971, Miyashita 1988, Funahashi et al. 1989, Goldman-Rakic 1995) and rodents (Liu et al. 2014, Guo et al. 2014, Inagaki et al. 2017) during delay

response tasks. However, the role of persistent activity as the activity subserving mnemonic representations has been recently challenged, and with this the role of attractor networks as viable theory for memory storage. In prefrontal cortex, during delay periods, it has been observed high degree of temporal irregularity, variability across trials for single memoranda and heterogeneity across neurons (Compte et al. 2003, Shafi et al. 2007, Barak et al. 2010, Barak & Tsodyks 2014, Kobak et al. 2016, Murray et al. 2017). There is an ongoing debate on whether this kind of activity is consistent with attractor dynamics (Lundqvist et al. 2018, Constantinidis et al. 2018). Various models have been proposed to account for this extra variability in attractor networks (Barbieri & Brunel 2007, Mongillo et al. 2008, Lundqvist et al. 2010, Mongillo et al. 2012, Druckmann & Chklovskii 2012). Recently, we have proposed a new alternative scenario to account for the observed variability in an attractor network whose learning rules are inferred from *in vivo* data (Pereira & Brunel 2018*a*). In this scenario, chaotic attractors (in contrast to fixed-point attractors as in classical attractor networks (Hopfield 1982, Amit et al. 1985, Tsodyks & Feigel'Man 1988)) correspond to neural representations of stored memories. Neural activity presents strong temporal fluctuations that are internally generated by the network's chaotic dynamics, but maintains a positive correlation with the stored pattern. Therefore, the network behaves as an associative memory device in which chaotic attractors correspond to internal representations of memories. Using a dynamic mean field theory (DMFT) (Sompolinsky et al. 1988, Crisanti & Sompolinsky 2018) Tirozzi and Tsodyks predicted the existence of this chaotic associative memory phase in the sparse version of the Hopfield model (Tirozzi & Tsodyks 1991). The transition to chaos is extensive, i.e. all fixed-point attractor memory states transition to chaos at once, which it has been found to be also the case in networks constrained by *in vivo* data (Pereira & Brunel 2018*a*). Furthermore, consistent of what Tirozzi and Tsodyks predicted for the sparse Hopfield model (Tirozzi & Tsodyks 1991), it has also been found in this model that the capacity (i.e. the maximum number of memory states the network can

store) for chaotic attractors is larger than what is predicted by a static mean field theory (SMFT) for fixed-point attractors as memory states (Pereira & Brunel 2018*a*). However, a theory for chaotic memory states in networks constrained by data is lacking.

A well known phenomena in attractor networks is catastrophic forgetting (Amit et al. 1985). It refers to the observation that when the number of stored patterns is larger than the network's capacity, all memories are forgotten and consequently no memory can be retrieved from the network. The basic recipe for catastrophic forgetting is based in the statistical symmetry between stored patterns: when all patterns are identical and independently distributed as in classical attractor network models (Hopfield 1982, Amit et al. 1985, Tsodyks & Feigel'Man 1988) forgetting one pattern is statistically equivalent to forgeting all. For large networks, their behavior converges to their average behavior, and when one pattern is forgotten then all patterns are forgotten at once. The recipe for fixing catastrophic forgetting is also well known: by introducing a *forgetting process* the notion of age breaks the statistical symmetry between patterns and newer patterns are remembered while older patterns are forgotten in an *online* process of learning (and forgetting) (Parisi 1986, Amit & Fusi 1994). Recently, the question of the *optimal forgetting process* for maximizing capacity have been explored for networks of binary neurons, and optimal forgetting kernels and bounds for the memory storage have been derived (Amit & Huang 2010, Huang & Amit 2011, Lahiri & Ganguli 2013, Benna & Fusi 2016). Importantly, the memory states analyzed in the above studies are fixed-point attractors, and the effect of online learning for networks endowed with chaotic memory states is unknown. Furthermore, a theory for attractor networks in such scenario is lacking.

In this paper we provide general theory for a family of attractor networks with unsupervised Hebbian learning rules as the ones inferred from *in vivo* data (Lim et al. 2015, Pereira & Brunel 2018*a*) and online learning of memories. In section 4.3 we introduce the family of attractor network models. In section 4.4 we provide a dynamic mean field theory for the

network's dynamics. In section 4.5 and 4.6 we derived the general curves for the transition to chaos and capacity, and show that memory states lay in a continuum of different statistical properties depending on age. In section 4.7 we recapitulate the results of Tirozzi and Tsodyks (Tirozzi & Tsodyks 1991), providing numerical solutions for the mean field equations and contrasting the results with simulations of large networks. In section 4.8, we show that when forgetting is included in this model memories stored as both fixed point and chaotic attractors co-exist. Depending on the pattern age, its retrieval state is a fixed-point (newer patterns) or chaotic attractor (older patterns), leading to a continuum of different retrieval states. Additionally, we found the optimal forgetting time-scale for an exponential forgetting kernel.

4.3 The Model

In this model, the network is composed of N neurons whose input current are described by analog variables h_i, where $i = 1, 2, \ldots, N$ represents the neuron index. The instantaneous firing rates of neurons are given by the the input-output single neuron transfer function (or f-I curve) ϕ. Input currents obey the standard current-based version of the rate equations (Grossberg 1969, Hopfield 1984) (which are equivalent to the rate-based version, see Miller & Fumarola (2012))

$$\dot{h}_i = -h_i + \sum_{i \neq j}^{N} J_{ij} \phi(h_j). \tag{4.1}$$

Here J_{ij} is the strength of the synapse connecting neuron j to neuron i. The connectivity matrix is sparse, and existing connections are shaped by external inputs ('patterns') through a non-linear unsupervised Hebbian synaptic plasticity rule. In this rule, firing rate patterns η_i^μ of neuron i during presentation of pattern μ ($i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, p$) are generated randomly and independently from some distribution p_η (i.e. $\eta_j^\mu \overset{iid}{\sim} p_\eta(\eta)$). The

firing rate patterns shape the connectivity matrix through two non-linear functions f and g that characterize the dependence of the learning rule on the post-synaptic rate (f) and pre-synaptic rate (g), respectively. Patterns recently learned are strongly imprinted in the connectivity than older patterns according to a 'forgetting kernel', $\Theta(\mu)$, in analogy to what has been proposed in palimpsest models for attractor neural networks of binary neurons (Mézard et al. 1986*b*, Parisi 1986, Amit & Fusi 1994, Romani et al. 2008, Amit & Huang 2010, Huang & Amit 2011, Dubreuil et al. 2014). The idea is that recent stored patterns partially erase the traces of older ones in the connectivity matrix, and the function $\Theta(\mu)$ gives to what degree a particular pattern of age μ has been forgotten. When p patterns are learned by the network, the connectivity after learning gets structured as

$$J_{ij} = \frac{Ac_{ij}}{K} \sum_{\mu=1}^{p} \Theta(\mu) f\left[\eta_i^\mu\right] g\left[\eta_j^\mu\right], \tag{4.2}$$

where c_{ij} is a sparse random (Erdos-Renyi) structural connectivity matrix ($c_{ij} = 1$ with probability K/N, $c_{ij} = 0$ with probability $1-K/N$). The sparsity in the connectivity models the low connection probabilities reported in cortical ($\sim$10%) (Mason et al. 1991, Markram, Lübke, Frotscher, Roth & Sakmann 1997, Holmgren et al. 2003, Thomson & Lamy 2007, Lefort et al. 2009) and hippocampal ($\sim$1%) (Guzman et al. 2016) microcircuits. The learning rule is a generalization of the unsupervised Hebbian rule used in chapter 3 (compare Eq. (4.2) with Eq. (3.2)) with two important differences: 1) The assumption of starting from an initial *tabula rasa* connectivity $J_{ij} = 0$ is not necessary, and can be dropped. Depending of the learning kernel, the synaptic connectivity matrix can be obtained by learning a continuous stream of patterns (i.e. online learning) where recent ones ones can be retrieved and older ones are forgotten; 2) The distribution of firing rates patterns p_η is left unspecified. The model in chapter 3 is a particular case when $\eta = \phi(\xi)$ and ξ is a standard normal random variable (see section 3.9.3, Eq. (3.48)), and therefore $p_\eta(x) = (e^{-(\phi^{-1}(x))^2/2}/\sqrt{2\pi})(d\phi^{-1}(x)/dx)$. Besides the addition of a forgetting process $\Theta(\mu)$, further differences with classic models such as the

Hopfield model (Hopfield 1982) or the Tsodyks-Feigel'man model (Tsodyks & Feigel'Man 1988) are: patterns can have a continuous distribution instead of binary; and the dependence of the rule on firing rates is non-linear instead of linear.

In chapter 3 we have shown that both the transfer function (ϕ) and the post-synaptic dependence of the learning rule f can be inferred from electrophysiological data (see section 3.9.3 and (Lim et al. 2015)). As in chapter 3, we constrain g by the condition that the average of the function g across the distribution of patterns is zero, i.e. $\langle g(\eta)\rangle_\eta = 0$, which ensures that the average change in connection strength due to learning of a single pattern is zero. This could be enforced by a homeostatic mechanism that controls the mean changes in the incoming inputs due to learning (Toyoizumi et al. 2014, Vogels et al. 2011).

4.4 Dynamic mean field theory

In the limit of infinitely large number of neurons (i.e. $N \to \infty$), synapses per neuron (i.e. $K \to \infty$), and strongly sparsely connected network (i.e. $K/N \to 0$) a dynamical mean field theory can be developed using functional integration as in Sompolinsky & Zippelius (1982), Kree & Zippelius (1987), Tirozzi & Tsodyks (1991). In this limit Eq. (4.1) is reduced to

$$\dot{h}_i(t) = -h(t) + \mu_i + \rho_i(t), \tag{4.3}$$

where

$$\mu_i = A \sum_{\mu=1}^{s} f(\eta_i^\mu)\Theta(\mu)m_\mu, \tag{4.4}$$

corresponds to the average input current to neurons i. As in classical mean field theories for attractor neuronal network models (Amit et al. 1985, Tsodyks & Feigel'Man 1988) we define the order parameters

$$m_\mu \equiv \langle g(\eta^\mu)\phi(h)\rangle_{h,\eta^\mu}. \tag{4.5}$$

Here the average is over the distribution of the synaptic input current h and of the stored pattern η^μ. Eq (4.5) corresponds to the overlap between the network state and a non-linear transformation of the stored pattern. In our theory we assumed that the overlaps do not depend on time (i.e. $m_\mu(t) = m_\mu$), which is trivially true for fixed-point attractor memory states, and a good approximation for chaotic attractor memory states. These order parameters are a natural extension for analog neurons and nonlinear learning rules of the overlaps used in classic attractor neural network models. In our theory the number of retrieved patterns s are of order $s \sim \mathcal{O}(1)$, and therefore just a finite number of patterns have an non-negligible overlap with neural activity. The variable $\rho_i(t)$ is a random gaussian field with zero mean and auto-covariance given by

$$\mathrm{Cov}\left(\rho_i(t), \rho_i(t+\tau)\right) = \gamma A^2 \kappa \langle \phi(h(t))\phi(h(t+\tau))\rangle_h, \tag{4.6}$$

where

$$\kappa = \frac{1}{K}\sum_{\mu=1}^{p} \Theta^2(\mu). \tag{4.7}$$

We explore the scenario in which an infinite stream of pattern is presented to the network, and therefore $p \to \infty$. Each pattern is presented once for learning, and then gradually forgotten due to the learning of the subsequent patterns in the stream. We define the learning rule dependent constant $\gamma = \langle f^2(\eta)\rangle\langle g^2(\eta)\rangle$. The dynamics of the field can be approximated by the following time dependent Gaussian random field

$$\dot{h}_i = -h_i + A f(\eta_i^\mu)\Theta(\mu) m_\mu + A\sqrt{\gamma\kappa}\, y(t). \tag{4.8}$$

Where $y(t)$ is a gaussian random field with auto-covariance function

$$C(\tau) = \langle y(t)y(t+\tau)\rangle_y = \langle \phi(h(t))\phi(h(t+\tau))\rangle_h, \tag{4.9}$$

which is calculated self-consistently. By defining the local currents $u_i(t) = h_i(t) - Af(\eta_i^\mu)\Theta(\mu)m_\mu$ then Eq (4.8), (4.5) and (4.9) can be re-written as

$$\begin{aligned} \dot{u} &= -u + A\sqrt{\gamma\alpha}y(t) && (4.10)\\ m_\mu &= \langle g(\eta)\phi(u(t) + Af(\eta)\Theta(\mu)m_\mu)\rangle_{u,\eta}, && (4.11) \end{aligned}$$

and

$$C(\tau) = \langle \phi\left(u(t) + Af(\eta)\Theta(\mu)m_\mu\right)\phi\left(u(t+\tau) + Af(\eta)\Theta(\mu)m_\mu\right)\rangle_{u,\eta}. \tag{4.12}$$

In Eq. (4.10) we assume a translation invariance of the auto-covariance. As in (Sompolinsky et al. 1988) we introduce the local-field auto-covariance function

$$\Delta(\tau) = \langle u(t)u(t+\tau)\rangle_u. \tag{4.13}$$

Analogous to the derivation in Crisanti & Sompolinsky (2018), Schücker et al. (2016) we derive a self-consistent equation for the local-field auto-covariance

$$\frac{d^2\Delta(\tau)}{d^2\tau} = \Delta(\tau) - A^2\gamma\kappa C(\tau). \tag{4.14}$$

See appendix B for a version of the derivation. Analogously to Sompolinsky et al. (1988), Tsodyks & Feigel'Man (1988), the auto-covariance in Eq. (4.12) can be written as

$$C(\tau) = \int D\eta Dz \left[\int Dx\phi \left(A \left[\sqrt{\Delta_0 - |\Delta(\tau)|}x + \sqrt{|\Delta(\tau)|}z + f(\eta)\Theta(\mu)m_\mu \right] \right) \right]^2, \quad (4.15)$$

where $D\eta = p_\eta(\eta)d\eta$ (distribution of the stored patterns). For the proof of this, analogous equations to Eqs. (B.7,B.8) for $u_i(t)$ should be inserted in Eq. (4.12). For the equation above we further assume that $0 \leq \Delta(\tau)$. We also re-scaled $\Delta(\tau)$ as $\Delta(\tau) \to A^2\Delta(\tau)$ for ease some of the algebra. In analogy with Sompolinsky et al. (1988) Eq. (4.14) can be re-written as

$$\frac{d^2\Delta}{d^2\tau} = -\frac{\partial V(\Delta, \Delta_0)}{\partial \Delta}, \quad (4.16)$$

by defining the following potential

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + \frac{\alpha\gamma}{A^2} \int D\eta Dz \left[\int Dx\Phi \left(A \left[\sqrt{\Delta_0 - |\Delta|}x + \sqrt{|\Delta|}z + f(\eta)\Theta(\mu)m_\mu \right] \right) \right]^2, \quad (4.17)$$

where $\Phi(x) = \int_0^x dr\phi(r)$. Notice that analogous to m_μ, the auto-covariance of the local fields Δ_0 also depends on the age of the retrieved memory μ, however we choose to not make explicit this dependency in order to simplify the notation.

4.5 Transition to chaos

In this section we determine the location in the parameters space where fixed-point attractors transition to chaotic attractors. We distinguish two qualitatively different attractor states depending on the overlap with the stored memory: 1) states with order one overlap (i.e. $m_\mu \sim \mathcal{O}(1)$) we call memory states; 2) states with negligible overlaps (i.e. $m_\mu \ll 1 \quad \forall\mu$) we call the background state.

4.5.1 Transition to chaos of fixed-point attractor memory states

For fixed-point attractors there are no temporal fluctuations in the input currents. Then the auto-covariance of the local fields in Eq. (4.13) is equal to the variance of the local currents at all times (i.e. $\Delta(\tau) = \Delta_0$), which leads to

$$m_\mu = \int D\eta Dx g(\eta)\phi\left(A\left[\sqrt{\Delta_0}x + f(\eta)\Theta(\mu)m_\mu\right]\right) \tag{4.18}$$

$$\Delta_0 = \gamma\kappa\int D\eta Dx \phi^2\left(A\left[\sqrt{\Delta_0}x + f(\eta)\Theta(\mu)m_\mu\right]\right). \tag{4.19}$$

The above equation give the overlap with the memory μ for fixed-point attractors. These fixed-point memory states may become chaotic depending on the parameters. In this scenario, the model presents chaotic dynamics with associative memory properties. Importantly, in this model the chaotic properties of the attractors depends on the age of the patterns. As we will discuss in the next sections, recent memory states are fixed-point and older memory states are chaotic. Analogous to Sompolinsky et al. (1988), to find the transition to chaos of memory states, it is necessary to find the point in parameter space where the static solution $\Delta(\tau) = \Delta_0$ becomes unstable. At this point the auto-covariance of the local-field $\Delta(\tau)$ transition from stationary to time-depend, and in the large K limit our theory predicts that the network becomes chaotic. Since the dependence on time of the auto-covariance of the local fields is ruled by the newton equation in Eq. (4.14), finding the transition point is equivalent to find the critical point Δ_0^{chaos} where the potential in Eq. (4.17) changes its concavity. After this point, solutions for the auto-covariance of the local field starting at Δ_0 relax to $\lim_{\tau\to\infty}\Delta(\tau) \equiv \Delta_1$. The transition point is given by

$$A^2\gamma\kappa\int D\eta Dz\left\{\phi'\left(A\left[\sqrt{\Delta_0}z + f(\eta)\Theta(\mu)m_\mu\right]\right)\right\}^2 = 1. \tag{4.20}$$

Equation (4.20) in addition to Eqs. (4.18,4.19) describe the curve in the parameter space

that separates fixed-point from chaotic memory states.

4.5.2 Transition to chaos of the background state

In the background state, all the overlaps with the stored memories are zero, i.e. $m_\mu = 0$, in the thermodynamic limit. The critical line in the space of parameters for its transition to chaos is given by

$$A^2\gamma\kappa \int Dz \left\{\phi^{'}\left(A\sqrt{\Delta_0}z\right)\right\}^2 = 1 \tag{4.21}$$

$$\Delta_0 = \gamma\kappa \int Dz\phi^2\left(A\sqrt{\Delta_0}z\right) \tag{4.22}$$

4.6 Capacity

4.6.1 Capacity for chaotic memory states

Analogously to the static mean field theory derived in section 3.9.1 of chapter 3, the capacity of fixed-point attractor states are given by the curve in the parameter space when the overlap is zero, that is the smaller μ in which $m_\mu = 0$ is the only solution of Eqs. (4.18,4.19). In this scenario the underling assumption is that memory states are fixed-points. However, in this model chaotic memory states may undergo a transition to having zero overlap with the stored pattern. To calculate the capacity for chaotic attractors, we first assume that the network is in a regime in which the static solution is no longer stable. That is when the potential defined in Eq. (4.17) is no longer convex (i.e. $\frac{\partial^2 V}{\partial^2 \Delta} > 0$), and the auto-covariance of the local currents in Eq. (4.13) are time dependent. Additionally, as is explained in the previous section, a chaotic solution will have an aperiodic decreasing solution for the potential. This correspond to the condition $\lim_{\tau\to\infty} V(\Delta(\tau)) = V(\Delta_0)$, which is equivalent to

$$-\frac{\Delta_0^2}{2}+\frac{\kappa\gamma}{A^2}\int \mathcal{D}\eta Dx\Phi^2\left(A\left[\sqrt{\Delta_0}x+f(\eta)\Theta(\mu)m_\mu\right]\right)=-\frac{\Delta_1^2}{2}+$$
$$\frac{\kappa\gamma}{A^2}\int D\eta Dz\left[\int Dx\Phi\left(A\left[\sqrt{\Delta_0-|\Delta_1|}x+\sqrt{|\Delta_1|}z+f(\eta)\Theta(\mu)m_\mu\right]\right)\right]^2. \tag{4.23}$$

Here Δ_1 corresponds to $\Delta(t)\xrightarrow{t\to\infty}\Delta_1$. Therefore, $\partial V/\partial\Delta|_{\Delta=\Delta_1}=0$, which is equivalent to

$$\Delta_1=\kappa\gamma\int D\eta Dz\left[\int Dx\phi\left(A\left[\sqrt{\Delta_0-|\Delta_1|}x+s\sqrt{|\Delta|}z+f(\eta)\Theta(\mu)m_\mu\right]\right)\right]^2. \tag{4.24}$$

Lastly, Eq. (4.36) for the overlap also holds. Therefore, Eqs (4.36,4.23,4.24) above give the overlap curve for chaotic attractors.

4.7 The sparsely connected Hopfield model

In this chapter we will briefly recapitulate the results of Tirozzi and Tsodyks (Tirozzi & Tsodyks 1991) where they study a sparse version of the Hopfield model for analog neurons. Using the generating functional method, it can be shown that that in the highly sparse limit the theory presented in (Tirozzi & Tsodyks 1991) is exact (Kree & Zippelius 1987). The connectivity in this network is given by

$$J_{ij}=\frac{Ac_{ij}}{Nc}\sum_{\mu=1}^{p}\eta_i^\mu\eta_j^\mu. \tag{4.25}$$

Here $\eta_i^k\in\{-1,1\}$ and iid with probability 0.5. The dynamics of the network is given by Eq. (4.1) with $\phi(x)=\tanh(x)$. In this model $\Theta(\mu)=1$, then

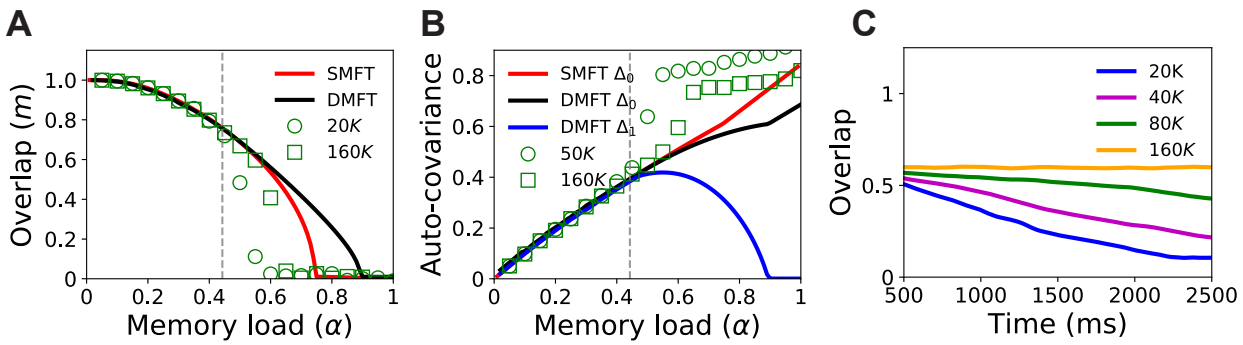

Figure 4.1: DMFT vs SMFT for sparsely connected Hopfield model. (A) Overlap vs memory load. Circle and square markers correspond to the average overlap calculated from network simulations of $N = 50000$ and $N = 160000$ neurons. The average was taken over 1000 times steps. The dashed line corresponds transition to chaos memory load. (B) Δ_0 and Δ_1 vs memory load. (C) Average overlap vs time for a memory load of $\alpha = 0.6$ and network sizes of $N = 20000, 40000, 80000, 160000$. The average is taken over 12 network realizations but for $N = 160000$ which corresponds to just one realization. The displayed dynamics is for the 500-2500ms time interval. In A-C the sparsity level is $c = 1/\sqrt{N}$ and $A = 5.5$.

$$\kappa = \lim_{K\to\infty} \frac{1}{K} \sum_{\mu=1}^{\alpha K} \Theta^2(\mu) = \alpha. \tag{4.26}$$

Furthermore, for this network $\gamma = 1$. Lastly, the mean field Eq. (4.4) is given by

$$\rho_i = A\xi_i m. \tag{4.27}$$

Here the index of the pattern μ is omitted since when any of the patterns is retrieved it produces the same mean field. Plugging-in these parameters in the equations of sections **??**, **??** and **??** we obtain the mean field equations in (Tirozzi & Tsodyks 1991). As in Hopfield model, when the memory load α increases the overlap m decreases due to the increase in the number of stored patterns, increasing in turn the variance of the fields Δ_0 (see Fig 4.1 A and B respectively). Interestingly, the capacity computed using the SMFT is smaller than the capacity computed using the DMFT (compare red and black lines in Fig 4.1 A). Since in the SMFT memory states are assumed to be fixed-points attractors while in the DMFT

may be both fixed-point and chaotic attracors, the disagreement between the two theories begins after the transition to chaos (dashed lines in Fig 4.1).The agreement between the numerical simulations is good for high overlaps (low loads) and deteriorates rapidly close to the transition to chaos. Increasing the network size and sparsity improves the agreement between the theory and numerical simulations, suggesting these are finite size effects.

We numerically solved the mean field equations for the transitions described in sections 4.5 and 4.6, obtaining the complete network's bifurcation diagram (see Fig 4.2). For small values A (i.e. weak connectivity) there are no memory states (red region in Fig 4.2). For larger values of A and low memory loads, the background and the memory states are fixed-point attractors (region bellow the red dashed line and blue region in Fig 4.2 respectively). When the memory load increases, the background state transition to chaos (region above the red dashed line in Fig 4.2) and memory states are fixed-point attractors (blue region in Fig 4.2). Larger memory loads lead to the memory states to transit to chaos (green region in Fig 4.2) reaching a phase when the dynamics is chaotic but the network retains a finite overlap with the stored memory. Finally, if the memory load further increases the network reaches its capacity and then memories are forgotten (gray region in Fig 4.2) .

4.8 Fixed-point and chaotic attractors co-exist due to forgetting

Here we consider a scenario in which random binary patterns $\{\eta_i^k\}_{k=1}^p$ are stored by a network using a Hebbian learning rule, with $\eta_i^k \in \{-1, 1\}$ and iid with probability 0.5. We assume the following forgetting kernel:

$$\Theta(\mu) = e^{-\frac{\mu}{\tau_f Nc}} \left(\frac{\mu}{Nc} + 1\right)^a. \tag{4.28}$$

Here patterns indexes begin at $\mu = 0$. We choose this kernel in order to explore monotonically decreasing ($a \leq 0$) and non-monotonic ($0 < a$) forgetting scenarios. Notice that

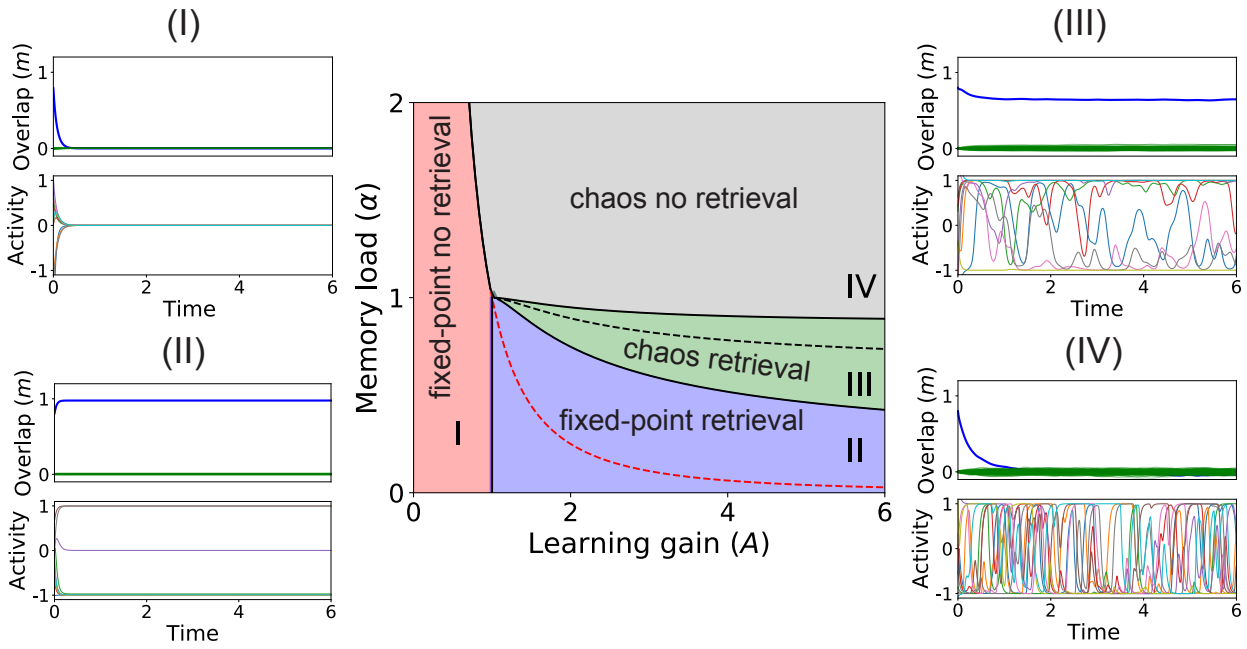

Figure 4.2: Center: Bifurcation diagram for the sparsely connected Hopfield model. Surrounding plots: Overlaps (top row) and activations (bottom row) for retrieval states of networks with parameters indicated with markers in the left panel. The rest parameter values are $A = 5.5$, $N = 50000$ and $c = 0.005$.

when $a = 0$ and $\tau_f \to \infty$ this model is equivalent to the sparse Hopfield model presented in the previous section. The connectivity is given by

$$J_{ij} = \frac{Ac_{ij}}{Nc} \sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu e^{-\frac{\mu}{\tau_f Nc}} \left(\frac{\mu}{Nc} + 1\right)^a. \tag{4.29}$$

The mean field in Eq. (4.4) is given by

$$\mu_i = A\eta_i^\mu e^{-\frac{\mu}{\tau_f Nc}} \left(\frac{\mu}{Nc} + 1\right)^a m_\mu. \tag{4.30}$$

The auto-covariance function in Eq. (4.7) is given by

$$\mathrm{Cov}\left(\rho_i(t), \rho_i(t+\tau)\right) = \frac{A^2}{Nc} \sum_{\mu=1}^{p} \Theta(\rho)^2 C(\tau). \tag{4.31}$$

In the limit $Nc \to \infty$ and $p \to \infty$ we obtain

$$\frac{1}{Nc} \sum_{\mu=0}^{p} \Theta(\mu)^2 \xrightarrow[Nc\to\infty]{p\to\infty} e^{\frac{2}{\tau_f}} \int_1^\infty dx e^{-\frac{2x}{\tau_f}} x^{2a} \equiv \Gamma(a, \tau_f) \tag{4.32}$$

Notice that here we used the fact that $\tanh(x)$ is an odd function and patterns are binary $\{-1, 1\}$. Then Eq. (4.8) for this network reads

$$\dot{h}_i = -h_i + A\eta_i^s e^{-\frac{s}{\tau_f}} (s+1)^a m + A\sqrt{\Gamma(a,\tau)} y(t). \tag{4.33}$$

Here $s = \mu/Nc$ which is the continuous version of μ when $Nc \to \infty$. As is described in section 4.4 (see Eq. (4.9)) $y(t)$ is a gaussian random field with auto-covariance function given by

$$C(\tau) = \int Dz \left[\int Dx\phi\left(A\left[\sqrt{\Delta_0 - \Delta(\tau)}x + \sqrt{\Delta(\tau)}z + e^{-\frac{s}{\tau_f}} (s+1)^a m\right]\right)\right]^2, \tag{4.34}$$

The potential defined in Eq. (4.17) in this case is given by

$$V(\Delta,\Delta_0) = -\frac{\Delta^2}{2} + \Gamma(a,\tau)A^2 \int Dz \left[\int Dx\Phi\left(A\left[\sqrt{\Delta_0-\Delta}x + \sqrt{\Delta}z + e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right)\right]^2 . \tag{4.35}$$

4.8.1 Transitions

Overlap of fixed-point attractors with memories

For fixed-point attractors Eqs. (4.18,4.19) become

$$m = \int Dx\phi\left(A\left[\sqrt{\Delta_0}x + e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right) \tag{4.36}$$

$$\Delta_0 = \Gamma(a,\tau)\int Dx\phi^2\left(A\left[\sqrt{\Delta_0}x + e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right). \tag{4.37}$$

Notice the overlap curve depends on the age s of the pattern.

Transition to Chaos for Fixed-Point Attractors

Let us write first the second derivative of the potential

$$-1 + \Gamma(a,\tau)A^2 \int Dz \left[\int Dx\phi'\left(A\left[\sqrt{\Delta_0-\Delta}x + \sqrt{\Delta}z + e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right)\right]^2 \tag{4.38}$$

For a fixed-point attractor the equation above becomes

$$\Gamma(a,\tau)A^2\int Dz\left\{\phi'\left(A\left[\sqrt{\Delta_0}z+e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right)\right\}^2-1. \tag{4.39}$$

As explained in section 4.5, we need to find the value of Δ_0 where the potential changes from convex to concave. In this case, Eq. (4.20) becomes

$$\int Dz\left\{\phi'\left(A\left[\sqrt{\Delta_0}z+e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right)\right\}^2=\frac{1}{\Gamma(a,\tau)A^2}. \tag{4.40}$$

Transition to Chaos for Retrieval Fixed-points ($0<m$)

For retrieval states the equations for finding the critical line in the space of parameters are

$$\int Dz\left\{\phi'\left(A\left[\sqrt{\Delta_0}z+e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right)\right\}^2=\frac{1}{\Gamma(a,\tau)A^2} \tag{4.41}$$

$$m=\int Dx\phi\left(A\left[\sqrt{\Delta_0}x+e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right) \tag{4.42}$$

$$\Delta_0=\Gamma(a,\tau)\int Dx\phi^2\left(A\left[\sqrt{\Delta_0}x+e^{-\frac{s}{\tau_f}}(s+1)^a m\right]\right). \tag{4.43}$$

Transition to Chaos for Background Fixed-point ($m=0$)

For the background state (i.e. $m=0$) the equations for finding the critical line in the space of parameters are

$$\Gamma(a,\tau)A^2\int Dz\left\{\phi'\left(A\sqrt{\Delta_0}z\right)\right\}^2=1 \tag{4.44}$$

$$\Delta_0=\Gamma(a,\tau)\int Dz\phi^2\left(A\sqrt{\Delta_0}z\right). \tag{4.45}$$

4.8.2 Capacity

Analogous to the capacity calculation in the previous section, in this section we want to find the pattern age s^c where the network cannot work as an associative memory device. Equalizing the potential we obtain

$$-\frac{\Delta_0^2}{2}+\frac{\Gamma(a,\tau)}{A^2}\int Dx\Phi^2\left(A\left[\sqrt{\Delta_0}x+e^{-\frac{s}{\tau_f}}\,(s+1)^a\,m\right]\right)=-\frac{\Delta_1^2}{2}+$$
$$\frac{\Gamma(a,\tau)}{A^2}\int Dz\left(\int Dz\Phi\left(A\left[\sqrt{\Delta_0-\Delta_1}x+\sqrt{\Delta_1}z+e^{-\frac{s}{\tau_f}}\,(s+1)^a\,m\right]\right)\right)^2. \quad (4.46)$$

The derivative of the potential equal to zero becomes

$$\Delta_1=\Gamma(a,\tau_f)\int Dx\left(\int Dz\phi\left(A\left[\sqrt{\Delta_0-\Delta_1}x+\sqrt{\Delta_1}z+e^{-\frac{s}{\tau_f}}\,(s+1)^a\,m\right]\right)\right)^2. \quad (4.47)$$

Notice Eq. (4.47) is zero for $\Delta_1=0$ since $\phi(x)=-\phi(-x)$. Then the capacity is given by Eq. (4.47)

$$\Gamma(a,\tau_f)=\frac{(A\Delta_0^c)^2}{2\left[\int Dx\Phi^2\left(A\left[\sqrt{\Delta_0^c}x\right]\right)-\left(\int Dx\Phi\left(A\left[\sqrt{\Delta_0^c}x\right]\right)\right)^2\right]}, \quad (4.48)$$

obtaining

$$Ae^{-\frac{s_c}{\tau_f}}\,(s_c+1)^a\int Dx\phi\left(A\left[\sqrt{\Delta_0^c}x\right]\right)=1. \quad (4.49)$$

Then, Eqs, (4.48,4.49) provide the capacity curve (τ_f^c,a^c,s^c).

We numerically solved the mean field equations for the transitions described above, obtaining the complete network's bifurcation diagram (see Fig 4.3). For small value of τ (i.e.

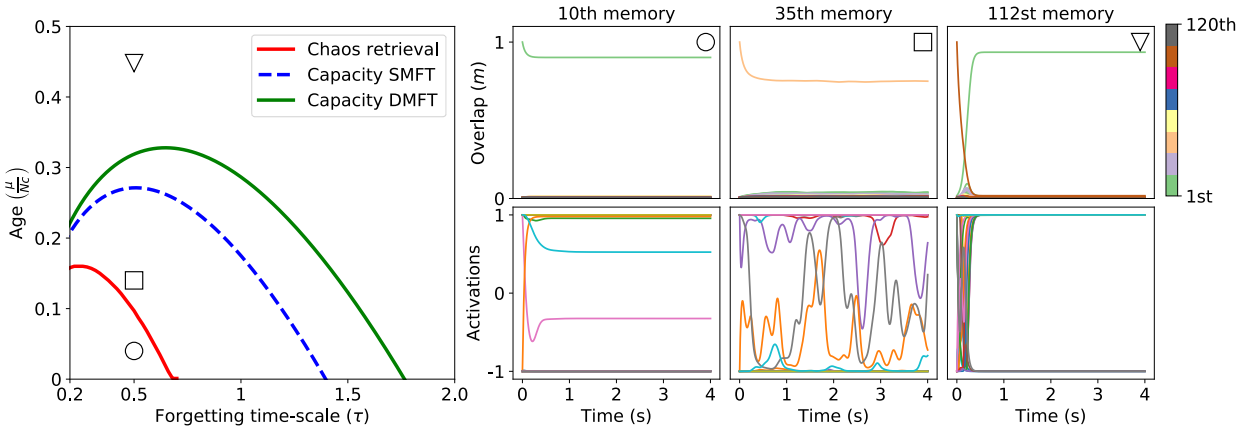

Figure 4.3: Left: Bifurcation diagram for the sparsely connected Hopfield model with exponential forgetting. Right: Retrieval states for the 10th, 35th and 112th memories for the same realization of the connectivity matrix.

fast forgetting) and age s (i.e. newer patterns), memory states are fixed-point attractors (see the region the below red line in the left panel of Fig 4.3). For example, the 10th stored memory in Fig 4.3 corresponds to a fixed-point. Older memory states are chaotic, the transition transition line between fixed-point and chaotic memory states is given by Eqs (4.41-4.43), see Fig. (4.3) red line. Above this line memory states are chaotic attractors, as for example the 35th stored memory in Fig 4.3, . When the age of the pattern further increases above the capacity line given by Eqs. (4.46,4.47) (green line in left panel of Fig 4.2) memories cannot be retrieved, as for example the 112th memory in Fig 4.3. For larger values of τ (i.e. slow forgetting) all memory states are chaotic for this particular value of A.

4.8.3 Optimal forgetting

We study the optimal forgetting time scale τ for the exponential forgetting kernel $exp(-s/\tau)$.

Large gain limit

Let us start for the case $A \to \infty$ making the following approximations

$$\phi(Ax) \to \begin{cases} -1 & x < 0 \\ 1 & 0 \leq x \end{cases} \tag{4.50}$$

and

$$\Phi(Ax) \approx \begin{cases} -Ax & x < 0 \\ Ax & 0 \leq x. \end{cases} \tag{4.51}$$

For the SMFT we get the following MF equations:

$$m = \psi\left(-\frac{me^{-\frac{s}{\tau}}}{\sqrt{\Delta_0}}\right) + \psi\left(\frac{me^{-\frac{s}{\tau}}}{\sqrt{\Delta_0}}\right) \tag{4.52}$$

$$\Delta_0 = \frac{\tau}{2}. \tag{4.53}$$

Here $\psi(x) = \int_x^\infty dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}$. Applying a derivative to equation (4.52) and setting $m = 0$ we get

$$1 = \frac{2e^{-\frac{s}{\tau}}}{\sqrt{2\pi\frac{\tau}{2}}} \implies \tau = \frac{4e^{-\frac{2s}{\tau}}}{\pi}$$

Hence, the capacity curve for $A \to \infty$ is given by

$$s^c = -\frac{\tau}{2}\log\left(\frac{\pi\tau}{4}\right). \tag{4.54}$$

And the optimal τ is given by

$$\tau_{max}^{fixed-points} = \frac{4}{e\pi} \approx 0.47 \tag{4.55}$$

For chaotic attractors we first derivate equation (4.52) and setting $m = 0$ we get

$$\Delta_0 = \frac{2e^{-\frac{2s}{\tau}}}{\pi}. \tag{4.56}$$

By calculating these integrals

$$\begin{aligned} \int_{-\infty}^{\infty} Dx\Phi^2(A\sqrt{\Delta_0}x) &= A^2\Delta_0 \\ \int_{-\infty}^{\infty} Dx\Phi(A\sqrt{\Delta_0}x) &= A\sqrt{\Delta_0}\frac{2}{\sqrt{2\pi}}. \end{aligned}$$

Then Eq. (4.48) becomes

$$\tau = \frac{2e^{\frac{-s}{2\tau}}}{\pi - 2}.$$

Hence, the optimal τ for $A \to \infty$ predicted for chaotic attractors is given by

$$s^c = -\frac{\tau}{2}\log\left(\frac{(\pi-2)\,\tau}{2}\right). \tag{4.57}$$

And the optimal critical τ is given by

$$\tau_{max}^{chaos} = \frac{2}{e(\pi-2)} \approx 0.64. \tag{4.58}$$

In Fig. 4.4 the results above are contrasted with numerical solutions of the mean field equations for finite values of A.

4.9 Discussion

Attractor networks (Hopfield 1982, Amit et al. 1985, Amit 1992) are a class of recurrent connected networks that have been influential in neuroscience by providing a mechanistic model for associative memory. In this class of models, memory states correspond to fixed-

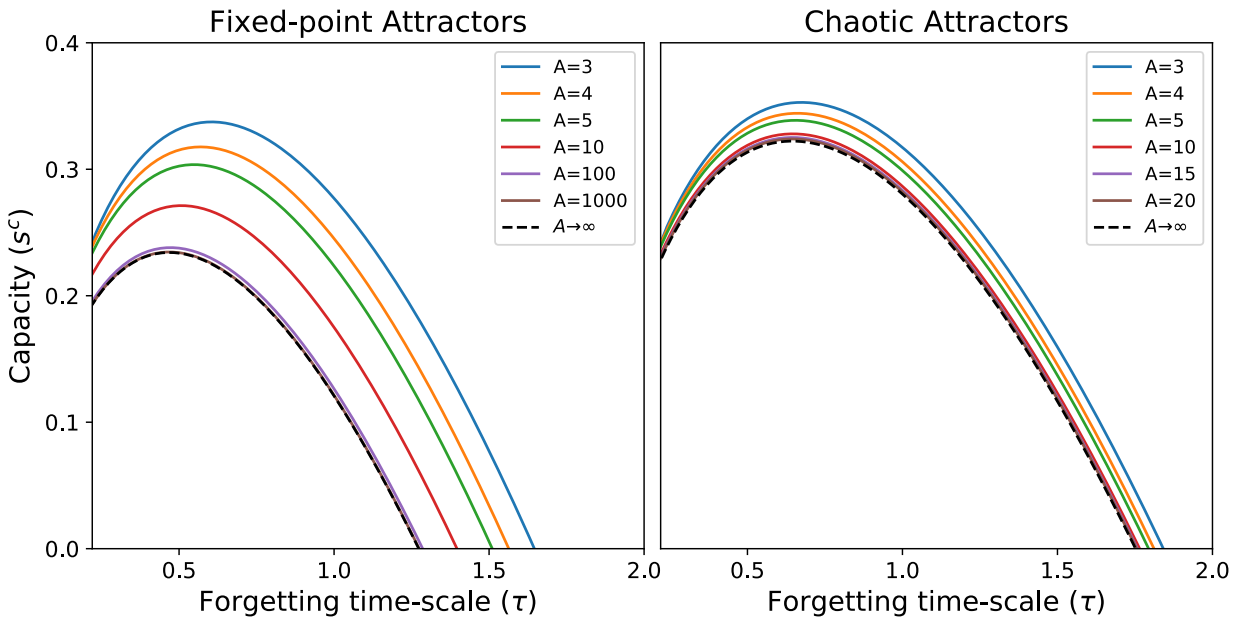

Figure 4.4: Capacity vs Forgetting time-scale. Left: Capacity calculated from the static MFT, see Eq. (4.36) and (4.40). Right: Capacity calculated from the dynamic MFT, see Eq. (4.48) and (4.49). In dashed black are the analytical capacity curves for the static and dynamic MFT, see Eq. (4.54) and (4.57) respectively.

point attractors in the network dynamics. When a memory is retrieved, the network reaches a fixed-point attractor and its activity is constant in time and correlated with the retrieved memory. Randomly connected recurrent networks (Sompolinsky et al. 1988, Van Vreeswijk et al. 1996, Brunel 2000) have been also influential in neuroscience by providing a network mechanism for explaining the strong temporal variability observed in cortical networks. In these networks, the activity fluctuate chaotically, but the scenario in which memories are stored as chaotic attractors have been seldom explored. Here we connect these two class of models, showing that in attractor networks memory states can be both fixed-point and chaotic attractors depending on parameters. Strikingly, we show that when the online learning scenario is considered, the network presents a continuum of memory states in which fixed-points and chaotic attractors co-exist.

CHAPTER 5
UNSUPERVISED LEARNING OF SEQUENTIAL ACTIVITY WITH TEMPORALLY ASYMMETRIC HEBBIAN LEARNING RULES

5.1 Contribution

The work presented in this chapter is part of a manuscript in preparation for publication. The authors are Maxwell Gillet, Ulises Pereira and Nicolas Brunel. M.G., U.P. and N.B. designed the research. U.P. and M.G. performed the mean field theory and capacity calculations. M.G. performed the numerical simulations and data comparison. M.G., U.P. and N.B. wrote the manuscript.

5.2 Introduction

Sequential activity has been observed across multiples species in a number of behaviors such as spatial navigation (Foster & Wilson 2006, Harvey et al. 2012, Grosmark & Buzsáki 2016) and bird song generation (Hahnloser et al. 2002, Amador et al. 2013, Okubo et al. 2015). Experimental evidence shows that sequential activity can be learned throughout experience (Okubo et al. 2015, Grosmark & Buzsáki 2016). Several theoretical network models have been able to produce sequential activity (Abeles 1991, Amari 1972, Kleinfeld & Sompolinsky 1988, Diesmann et al. 1999, Izhikevich 2006, Liu & Buonomano 2009, Fiete et al. 2010, Waddington et al. 2012, Cannon et al. 2015). In these models, the connectivity contains a feed-forward structure - neurons active at a given time in the sequence project in a feed-forward manner to the group of neurons which are active next.

As we have described in chapter 2, models for learning sequential activity in networks with plastic synapses can be roughly divided in two categories: models with supervised and

unsupervised plasticity rules. In models with supervised plasticity rules, the synapses are updated according the activity of the network and an *error signal* that carries information about the difference between the current network dynamics and the target dynamics (Sussillo & Abbott 2009, Memmesheimer et al. 2014, Laje & Buonomano 2013, Rajan et al. 2016). However, it is not clear that in cortex such error signal drives synaptic modifications, and learning of sequences may occur without supervision by the solely exposure of the network to sensory inputs. In models with unsupervised plasticity rules, sequential dynamics is shaped by external stimulation without an error signal (Jun & Jin 2007, Liu & Buonomano 2009, Fiete et al. 2010, Waddington et al. 2012, Okubo et al. 2015, Veliz-Cuba et al. 2015). In those models sequential activity is generated spontaneously, and the temporal statistics of the stimulation shapes the specific timing of the sequences. While these networks possess a high degree of biological realism, few quantitative results exist governing their storage and retrieval properties. Here we study a network of rate neurons in which sequences are learned without supervision from external inputs. In our model, sequential activity is learned by an asymmetric Hebbian learning rule that transforms temporally ordered random input patterns into synaptic weight updates. Learned patterns can be sequentially retrieved in the order that they were presented in an stereotypical and reliable manner. Importantly, during retrieval the network presents transient sequential dynamics both in its correlation with the stored patterns and activations. We developed a mean field theory for stored patterns with Gaussian statistics, obtaining dynamical equations for the transient correlation between the network activity and the stored patterns throughout the sequence. We compute the sequential capacity of these networks, that is the number of sequences that can be stored as a function of network size, and show that it grows linearly with network size, comparable to that found in networks storing fixed-point attractors.

5.3 The model

We consider a learning process that converts successive patterns of stimulation into synaptic weight changes. In our setting the network learns P sequences of S input patterns as we describe below. At time t an input pattern that belongs to the pth sequence is presented, eliciting a corresponding pattern of neural activity $\xi_i^{p,1}$ at neuron i (see Fig 5.1 left column). At time $t+1$ an uncorrelated input pattern that also belongs to the pth sequence is presented, eliciting the pattern of neural activity $\xi_i^{p,2}$ (see Fig 5.1 middle column). For each successive pair of presented inputs in a sequence, the strength J_{ij} of a synaptic connection from neuron j to neuron i is modified according to a temporally asymmetric Hebbian learning rule. In this rule, synaptic connections are modified according to the product of two functions of the pre and postsynaptic firing rates:

$$\Delta J_{ij} \propto f(\xi_i^{p,2}) g(\xi_j^{p,1}), \tag{5.1}$$

see Fig 5.1 right column. As in chapter 3, the functions f and g correspond to the post and pre synaptic dependence of the learning rule respectively. If the presynaptic neuron activation is $\xi_j^{p,1}$ when the first pattern is presented, and the postsynaptic neuron activation is $\xi_i^{p,2}$ for the the next pattern, then the synapse between neuron i and j is potentiated (depressed) according to Eq. (5.1) (see Fig 5.1 right column). For simplicity, as in chapter 3, we assume that learning starts from a *tabula rasa*, i.e. $J_{ij} = 0$. After learning S sequences of P patterns each the connectivity is sculptured by the learning process taking the form:

$$J_{ij} = \frac{c_{ij}}{Nc} \sum_{p=1}^{P} \sum_{\mu=1}^{S} f(\xi_i^{p,\mu+1}) g(\xi_j^{p,\mu}). \tag{5.2}$$

Here c_{ij} is a Bernoulli random variable with probability c encoding the presence or absence of a synaptic connection, N is the number of neurons and Nc represents the average in-degree of a neuron. We are agnostic about the source of these patterns. They

may originate from external inputs projecting to the network, or from internally-generated fluctuations. As in chapter 2, firing rates obey standard rate equations

$$\frac{dr_i}{dt} = -r_i + \phi\left(\sum_{j\neq i} J_{ij} r_j\right). \tag{5.3}$$

When the network is initialized with the first pattern in a given sequence, it presents a transient sequential dynamics. Interestingly, single neurons take approximately the same sequence of values that the learned patterns throughout the dynamics. For example, as shown in Fig 5.2a, neuron i takes values that are often close $\xi_i^{p,1}, \xi_i^{p,2}, \cdots, \xi_i^{p,S}$ when the network is initialized with pattern $\vec{\xi}^{p,1}$. The transient dynamics elicited is robust against perturbations in the initial conditions (see Fig 5.2b). The correlations between the network activity and the learned patterns (i.e. overlaps) throughout the sequence also depict a transient sequential dynamics. Unlike the dynamics of single neuron, the sequential dynamics of the overlaps is characterized by the rise of one overlap after another in a stereotyped sequence (see Fig 5.2c). This is consistent with the fact that single neurons take approximately the same corresponding values of the learned patterns throughout the sequence.

5.4 Gaussian patterns

5.4.1 Mean field theory

In this section we will derive a mean field theory for a network where stored patterns are Gaussian and the learning rule is linear, i.e. $f(x) = x$ and $g(y) = y$. As is described above, after learning the connectivity matrix is given by

$$J_{ij} = \frac{c_{ij}}{Nc} \sum_{p=1}^{P} \sum_{\mu=1}^{S} \xi_i^{p,\mu+1} \xi_j^{p,\mu}. \tag{5.4}$$

Here μ corresponds to the index of a particular concatenated pair of patterns, i.e.

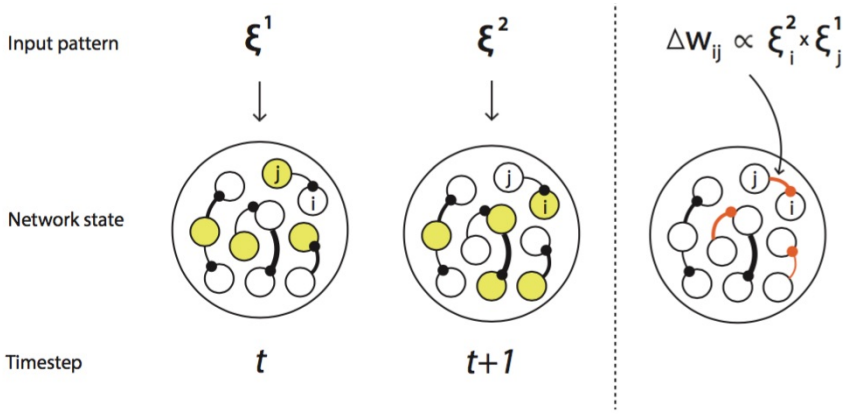

Figure 5.1: Learning and retrieval in recurrent neural networks with asymmetric unsupervised Hebbian learning rules. At time t a novel pattern is presented to the network, synaptic inputs to each neuron in the network (ξ_l^1, for neurons $l = 1, \ldots, N$) are drawn randomly and independently from a Gaussian distribution. Some neurons respond strongly (yellow circles) and other weakly (white circles). At the next time $t + 1$ a different pattern with the same statistics is presented to the network, and a different assembly of neurons than at time t is activated. Activity that is contiguous in time produced by the synaptic input currents modifies the network connectivity according to an asymmetric unsupervised Hebbian learning rule. Connections between neurons that are activated contiguously in time get modified (see red arrows). The connection strength is represented by the thickness of the corresponding arrow (the thicker the arrow the stronger the connection).

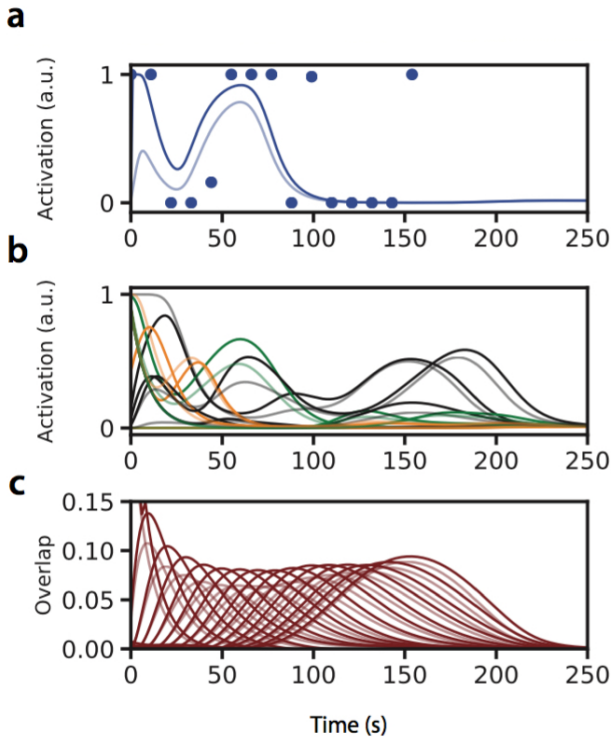

Figure 5.2: Sequence retrieval. a,b. Representative sample of single unit activity. Solid lines represent the trajectory of single unit activity in time. Discrete points correspond to stored sequential patterns. c. Overlap of network activity with each stored pattern. Light-colored lines show activity in response to a perturbation at the start of the trial.

$\xi_i^{p,\mu+1}\xi_j^{p,\mu}$, while p corresponds to the index of a particular sequence of concatenated patterns of length S, i.e. $\xi_i^{p,2}\xi_j^{p,1} + \xi_i^{p,3}\xi_j^{p,2} + \cdots + \xi_i^{p,S+1}\xi_j^{p,S}$. The patterns are identically and independently distributed (i.i.d.) as $\xi_i^{p,\mu} \overset{iid}{\sim} \mathcal{N}(0,1)$. The input current to neuron i at a given time t is given by the synaptic currents contributed by all the connected neurons:

$$h_i(t) = \sum_{j \neq i} J_{ij} r_j(t). \tag{5.5}$$

In this analysis we assume the dynamics starts with an initial condition that is correlated with the first pattern of sequence p, i.e. $\vec{\xi}^{p,1}$. The input current can be re-written as

$$h_i(t) = \sum_{\mu=1}^{S} \xi_i^{\mu+1,p} \frac{1}{Nc} \sum_{j \neq i}^{N} c_{ij} \xi_j^{\mu,p} r_j(t) + Y_i(t) \tag{5.6}$$

where Y_i describes the 'noise' term,

$$Y_i(t) = \frac{1}{Nc} \sum_{l \neq p}^{P} \sum_{\mu=1}^{S} \xi_i^{l,\mu+1} \sum_{j \neq i}^{N} c_{ij} \xi_j^{l,\mu} r_j(t). \tag{5.7}$$

In the large cN limit, due to the law of large numbers, the first term in Eq. (5.6) converges in probability to

$$\sum_{\mu=1}^{S} \xi_i^{\mu} q_{\mu}^{p}(t), \tag{5.8}$$

where the q_μ^ps are given by

$$q_{\mu}^{p}(t) = \frac{1}{N} \sum_{j=1}^{N} \xi_j^{\mu,p} r_j(t). \tag{5.9}$$

Here $\{q_\mu^p(t)\}_{\mu=1}^S$ are our first S order parameters. They described how correlated the network state is with the stored patterns $\vec{\xi}^{1,p}, \vec{\xi}^{2,p}, \cdots, \vec{\xi}^{S,p}$ respectively. We assume that the network state is uncorrelated with the rest of stored patterns since $q_\mu^l(t) \sim O(1/\sqrt{N})$

for $l \neq p$. Then the 'noise term' Y_i has mean zero and variance

$$Var\left(Y_i\right) = \alpha M(t), \tag{5.10}$$

where the sequential load is defined by

$$\alpha \equiv \frac{ps}{Nc}, \tag{5.11}$$

and M, the mean of the squared firing rate, is an additional order parameter defined by

$$M(t) = \frac{1}{N}\sum_{j=1}^{N} r_j^2(t). \tag{5.12}$$

In this theory we assume that the number of stored patterns is much larger than the number of patterns in a sequence, i.e. $s \ll \alpha Nc$. Then, we can approximate the dynamics in Eq. (5.3) as

$$\frac{dr_i}{dt} = -r_i + \phi\left(\sum_{\mu=1}^{S} \xi_i^{\mu+1} q_\mu(t) + \sqrt{\alpha M(t)}y_i\right). \tag{5.13}$$

Since all sequences are statistically equivalent we dropped the index p corresponding to the particular sequence of concatenated patterns. The variable y_i corresponds to the quenched noise produced by the stored patterns that do not belong to the sequence that is being retrieved (i.e. sequence p). By the central limit theorem the variable y_i is approximatly i.i.d. normally distributed across neurons, i.e. $y_i \overset{iid}{\sim} \mathcal{N}(0,1)$. For simplicity, we take a *static mean field theory* approach, where y_i is assumed to be constant in time, and therefore its auto-covariance is equal to its variance. Using equation (5.9) we get the following dynamical equations for the overlaps

$$\frac{dq_l}{dt} = -q_l + \int \mathcal{D}\vec{\xi}\mathcal{D}z\xi^l\phi\left(\sum_{\mu=1}^{S}\xi_i^{\mu+1}q_\mu(t) + \sqrt{\alpha M(t)}y_i\right) \quad l = 2,\ldots,S, \tag{5.14}$$

where $\mathcal{D}z = e^{-x^2/2}/\sqrt{2\pi}$ and $\mathcal{D}\vec{\xi} = \prod_{i=2}^{S}\mathcal{D}\xi^i$. Now we define

$$r_l^2(t) = \sum_{k\neq l}^{S} q_k^2(t) + \alpha M(t). \tag{5.15}$$

Since the stored patterns are gaussian we write Eq. (5.14) as

$$\frac{dq_l}{dt} = -q_l + \int D\xi^l\mathcal{D}z\xi^l\phi\left(\xi^l q_{l-1}(t) + r_{l-1}(t)x\right) \quad l = 2,\ldots,S. \tag{5.16}$$

Notice that ξ^l and x are independent standard normal random variables. Using the transformation

$$\begin{aligned} v &= \frac{\xi^l q_{l-1} + x r_{l-1}}{\sqrt{q_{l-1}^2 + r_{l-1}^2}} \\ u &= \frac{\xi^l r_{l-1} - x q_{l-1}}{\sqrt{q_{l-1}^2 + r_{l-1}^2}}, \end{aligned}$$

where u and v are also uncorrelated standard normal random variables, equation (5.16) becomes

$$\frac{dq_l}{dt} = -q_l + q_{l-1}G(\|\vec{q}(t)\|_2^2, M(t)) \quad l = 2,\ldots,S, \tag{5.17}$$

where we define

$$G(\|\vec{q}(t)\|_2^2, M(t)) \equiv \frac{\int \mathcal{D}v v\phi\left(v\sqrt{\|\vec{q}(t)\|_2^2 + \alpha M}\right)}{\sqrt{\|\vec{q}(t)\|_2^2 + \alpha M(t)}}. \tag{5.18}$$

Notice that the dynamical equation for the first overlap (i.e. q_1) is given by

$$\frac{dq_1}{dt} = -q_1. \tag{5.19}$$

Then by defining the 'delay line' matrix as

$$L = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \ddots & \cdots & \vdots \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{5.20}$$

We finally can write equation (5.18) in a vectorial form

$$\frac{d\vec{q}}{dt} = -\vec{q} + G(\|\vec{q}(t)\|_2^2, M(t))L\vec{q}. \tag{5.21}$$

Now we derive an approximate dynamical equation for M. From Eq. (5.3) we obtain the following two equations:

$$\frac{dr_i^2}{dt} = -2r_i^2 + 2r_i\phi\left(\sum_{j\neq i} J_{ij}r_j\right) \tag{5.22}$$

$$\left(\frac{dr_i}{dt}\right)^2 = r_i^2 + \phi^2\left(\sum_{j\neq i} J_{ij}r_j\right) - 2r_i\phi\left(\sum_{j\neq i} J_{ij}r_j\right). \tag{5.23}$$

Considering the following fact

$$\frac{1}{2}\frac{d^2 r_i^2}{dt^2} = \left(\frac{dr_i}{dt}\right)^2 + r_i \frac{d^2 r_i}{dt^2}, \tag{5.24}$$

and adding-up Eq. (5.22) and Eq. (5.23) we get

$$\frac{1}{2}\frac{d^2 r_i^2}{dt^2} + \frac{dr_i^2}{dt} - r_i \frac{d^2 r_i}{dt^2} = -r_i^2 + \phi^2 \left(\sum_{j\neq i} J_{ij} r_j\right). \tag{5.25}$$

From Eq. (5.3) we have that

$$r_i \frac{d^2 r_i}{dt^2} = -\frac{r_i^2}{dt} + r_i \phi^{'} \left(\sum_{j\neq i} J_{ij} r_j\right) \sum_{j=1}^{N} J_{ij} \dot{r}_j. \tag{5.26}$$

Then Eq. (5.25) becomes

$$\frac{1}{2}\frac{d^2 r_i^2}{dt^2} + 2\frac{dr_i^2}{dt} - r_i \phi^{'} \left(\sum_{j\neq i} J_{ij} r_j\right) \sum_{j=1}^{N} J_{ij} \dot{r}_j = -r_i^2 + \phi^2 \left(\sum_{j\neq i} J_{ij} r_j\right). \tag{5.27}$$

By averaging Eq (5.27) similarly as it was done for Eq. (5.21) we obtain

$$\frac{1}{2}\frac{d^2 M}{dt^2} + 2\frac{dM}{dt} - \left\langle r_i \phi^{'} \left(\sum_{j\neq i} J_{ij} r_j\right) \sum_{j=1}^{N} J_{ij} \dot{r}_j \right\rangle = -M + \int \mathcal{D}v \phi^2 \left(v\sqrt{\|\vec{q}(t)\|_2^2 + \alpha M}\right). \tag{5.28}$$

By approximating the third term in the l.h.s as the product of independent terms, i.e.

$$\left\langle \phi^{'} \left(\sum_{j\neq i} J_{ij} r_j\right) \sum_{j=1}^{N} J_{ij} \dot{r}_j \right\rangle \approx \left\langle \phi^{'} \left(\sum_{j\neq i} J_{ij} r_j\right) \right\rangle \left\langle r_i \sum_{j=1}^{N} J_{ij} \dot{r}_j \right\rangle, \tag{5.29}$$

we approximate Eq. (5.30) as

$$\frac{1}{2}\frac{d^2M}{dt^2} + 2\frac{dM}{dt} - \int \mathcal{D}v\phi'\left(v\sqrt{\|\vec{q}(t)\|_2^2 + \alpha M}\right)\left(\sum_{\mu=1}^{P} q_\mu(t)\dot{q}_\mu(t)\right) = -M + \int \mathcal{D}v\phi^2\left(v\sqrt{\|\vec{q}(t)\|_2^2 + \alpha M}\right). \tag{5.30}$$

Our mean field theory gives good quantitative predictions for the dynamics of the overlaps when it is compared with numerical simulations a large networks (see Fig 5.2a). Interestingly, the network can stored and successfully retrieve more than one sequence (see Fig 5.2 a). In the next section, we will calculate the maximum number of sequences that a network can store and successfully retrieve depending on the network parameters. We call this quantity sequential capacity.

5.4.2 Sequential capacity

We define the sequential capacity as the maximum number of sequences the network can store without decaying to zero in the limit of infinitely long sequences (i.e. $S \to \infty$) and time (i.e. $t \to \infty$) when the network is initialized with the first pattern in the sequence (as in Figs 5.2 and 5.3). The intuition for this definition is the following: for very long sequences, if the network is below capacity, it can be still retrieved after a long time since maintains finite overlaps with the stored patterns. On the other hand, if the network is above capacity, the overlaps die away after some time and the retrieval of the sequence is not possible. For finding the capacity of the network we will study the squared norm of the overlaps $\|\vec{q}(t)\|_2^2$. If this quantity is finite, there is a set of overlaps that are also finite. On the other hand, if this quantity is zero, all the overlaps are also zero. Therefore, the minimal value of the sequential load α in which $\|\vec{q}(t)\|_2^2 = 0$ corresponds to the network capacity, analogous to the capacity for attractor neuronal networks (Amit et al. 1985). Using Eq. (5.21) and noticing that for $P \to \infty$ we have that $L^T L^s = L^{s-1}$, then dynamical equations for the norm read

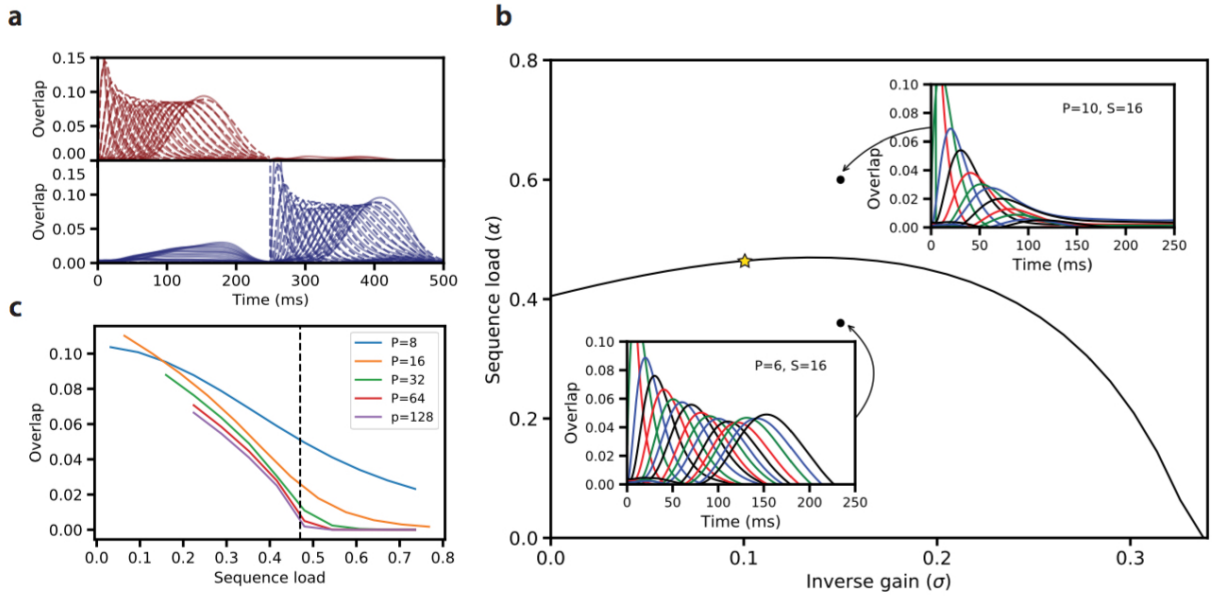

Figure 5.3: Capacity. a. Sequential activations of two discrete sequences. Solid lines are full network simulations, dashed lines are simulations of the mean-field description. b. Capacity (black line) as a function of the gain of the neural transfer function (all other parameters fixed). Insets display representative activity for network parameters above and below capacity curve, where solid lines are full network simulations. c. The maximal overlap with the final pattern in the stored sequence, for parameters corresponding to the yellow star in panel (b). The vertical dashed lines marks the predicted capacity.

$$
\begin{aligned}
\frac{d}{dt}\|\vec{q}\|_2^2 &= -2\|\vec{q}(t)\|_2^2 + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L\vec{q} + \vec{q}^T L^T\vec{q}\right)\\
\frac{d}{dt}\vec{q}^T L\vec{q} &= -2\vec{q}^T L\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^2\vec{q} + \vec{q}^T L^T L\vec{q}\right)\\
&= -2\vec{q}^T L\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^2\vec{q} + \|\vec{q}\|_2^2\right)\\
\frac{d}{dt}\vec{q}^T L^2\vec{q} &= -2\vec{q}^T L^2\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^3\vec{q} + \vec{q}^T L^T L^2\vec{q}\right)\\
&= -2\vec{q}^T L^2\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^3\vec{q} + \vec{q}^T L\vec{q}\right)\\
\frac{d}{dt}\vec{q}^T L^3\vec{q} &= -2\vec{q}^T L^3\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^4\vec{q} + \vec{q}^T L^T L^3\vec{q}\right)\\
&= -2\vec{q}^T L^3\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\cdot\left(\vec{q}^T L^4\vec{q} + \vec{q}^T L^2\vec{q}\right)\\
\vdots &= \quad \vdots
\end{aligned}
$$

By considering the fact that $(I-L)^{-1} = I + L + L^2 + L^3 + \cdots$, we then add the above equations obtaining:

$$
\frac{d}{dt}\vec{q}^T(I-L)^{-1}\vec{q} = -2\vec{q}^T(I-L)^{-1}\vec{q} + G(\|\vec{q}\|_2^2, M;\alpha)\left[2\vec{q}^T(I-L)^{-1}\vec{q} - \|\vec{q}\|_2^2 + \vec{q}^T L^T\vec{q}\right]. \tag{5.31}
$$

For very long times, the steady state of Eq. (5.31) is given by

$$
G(\|\vec{q}\|_2^2, M;\alpha)\left[\|\vec{q}\|_2^2 - \vec{q}^T L^T\vec{q}\right] = 2\vec{q}^T(I-L)^{-1}\vec{q}\left[G(\|\vec{q}\|_2^2, M;\alpha) - 1\right] \tag{5.32}
$$

Since $(I-L)^{-1}$ is the lower triangular matrix

$$
(I-L)^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix},
$$

then it is positive definite, which is by definition equivalent to

$$0 < \vec{q}^T (I - L)^{-1} \vec{q}.$$

On the other hand, for a infinitely long sequence we have that

$$\vec{q}^T L^T \vec{q} \leq \|\vec{q}\|_2^2. \tag{5.33}$$

Therefore, for the equality in Eq. (5.32) to hold if

$$G(\|\vec{q}\|_2^2, M; \alpha) < 1, \tag{5.34}$$

then $\vec{q} = 0$. In other words, if Eq. (5.34) holds then sequences decay after some finite time, and therefore the network is above capacity. Then the capacity curve is given by

$$G(0, M; \alpha) = 1. \tag{5.35}$$

At capacity the critical load α_c is given by

$$\frac{\int \mathcal{D}v v\phi\left(v\sqrt{\alpha_c M}\right)}{\sqrt{\alpha_c M}} = 1. \tag{5.36}$$

On the other hand, using Eq. (5.30), the value of M is given by

$$M = \int \mathcal{D}v \phi^2\left(v\sqrt{\alpha_c M}\right). \tag{5.37}$$

By solving both Eqs (5.36,5.37) we obtain the network capacity α_c. Our theory shows a good quantitative agreement with numerical simulations of large networks (see Fig 5.3b,c). The agreement is increasingly accurate as the size of the sequences increases (see Fig 5.3c).

5.5 Discussion

We have shown that the family of unsupervised Hebbian learning rules previously described for learning attractors (Pereira & Brunel 2018*a*) learn sequential activity when a temporal asymmetry in the learning rule is introduced (i.e. $J_{ij} \propto f(\xi_i^{p,s+1})g(\xi_j^{p,s})$ instead of $J_{ij} \propto f(\xi_i^{p,s})g(\xi_j^{p,s})$). This asymmetry naturally arises when a temporal delay as the time it takes for calcium influx through NMDA receptors to reach its maximum (Sabatini et al. 2002, Graupner & Brunel 2012) is considered (see Veliz-Cuba et al. (2015), Abbott & Blum (1996), Gerstner & Abbott (1997), Mehta et al. (1997), Jahnke et al. (2015), Chenkov et al. (2017), Theodoni et al. (2017), Pereira & Brunel (2018*b*) for models with temporally asymmetric Hebbian learning rules). When this delay is much slower than the external stimulus driving the network dynamics, patterns of activity of pre and post synaptic neurons in delayed times are approximately uncorrelated, and Hebbian learning rules take an asymmetric form as in Eq. (5.2).

The asymmetric learning rule analyzed in section 5.4 is well suited for storing sequences, since its capacity scales with the network size. In contrast, in appendix C we show its symmetric version, i.e. the covariance rule, leads to zero capacity for attractor states correlated with a single pattern. In this case, attractor states are correlated with multiple patterns, and the retrieval of a single memory is not possible.

This learning rule recapitulates two important features of the sequential activity observed in cortex: 1) stereotyped cue dependent sequential activity (see Fig 5.2 and 5.3b); 2) robust to perturbations sequential activity (see Fig 5.2). Remarkably, the network dynamics can be analyzed by a mean field theory, finding a low-dimensional description for the sequential dynamics (the dynamics is described by $S + 1$ degrees of freedom instead of the original N) in terms of the overlaps with the learned patterns. We show that the overlaps obey a non-linear feed-forward dynamical system, and the network dynamics is effectively feed-forward in the linear space spanned by the patterns in the learned sequence (space spanned

by $\vec{\xi}^{p,1}, \vec{\xi}^{p,2}, \cdots, \vec{\xi}^{p,S}$). Using this theory, we compute the sequential capacity of the network, showing that it grows linearly with network size, comparable to what is found in networks storing fixed-point attractors.

CHAPTER 6

CONCLUSIONS

In this thesis, I show that neural representations of memories in brain networks can be learned as qualitatively different spatiotemporal attractors by a single class of unsupervised learning rules in recurrent neuronal networks. Depending on the learning rule and the statistical properties of the inputs, neural representations of memories can be fixed-point attractors, chaotic attractors or sequences of activity. This model reproduces a wide range of data sets and provides an unified framework for understanding unsupervised learning of memories in brain networks. In the next sections, I will discuss outstanding questions and future directions

6.1 Possible functional relevance of different neuronal representations

What is the advantage (if any?) of having different representations of memories in brain networks? Memory capacity for fixed-point attractors, chaotic attractors and sequences scale linearly with the average number of synaptic connections. Therefore, in terms of memory capacity, there is no qualitative difference between the three types of neural representations. However, memories are encoded differently for fixed-point and chaotic attractors from sequences. For fixed-point and chaotic attractors the code is static, that is, the activity of the network lies in a single linear subspace which corresponds to the optimal decoder sub-space during retrieval. In the case of sequences, the optimal decoder sub-space changes dynamically, and the network optimally encodes different patterns at different times. Functionally, the static code is optimal for holding a single item in memory while the dynamic code is optimal encoding information concatenated with a certain timing. These different codes might be more favorable for different functions. For example, for encoding episodic memories a

static code might be better suited, since the memory needs to be held for a period of time for cognitive use. On the other hand, for encoding memories of motor actions, a dynamic code might be a better strategy, since it can encode the specific sequence of actions and the timing between them.

6.2 Online learning of memories in cortex

Neuronal responses of excitatory neurons to familiar images in the inferior temporal cortex (ITC) have lower mean firing rates but higher maximum firing rates than to novel (Woloszyn & Sheinberg 2012). These differences can be accounted by learning in the ITC recurrent microcircuit (Lim et al. 2015, Pereira & Brunel 2018*a*). The learning rules are inferred from neuronal responses to a large number of familiar and novel images (Lim et al. 2015, Pereira & Brunel 2018*a*). In these data, for a novel image to become familiar it is shown to a monkey more than 5000 times. However, is still unknown how the dynamics of the neuronal responses changes across presentations as well as the underpinning learning rule. Preliminary data sheds light upon this question showing that learning occurs within 2-4 training sessions (i.e. 70-140 presentations) (Mohan & Freedman 2018). Interestingly, very recently, similar timescales for learning familiar images in V2 (Huang et al. 2018) have been observed. I participate in a collaborative research project led by professors Yali Amit, Nicolas Brunel, and David Freedman with the aim to uncover the multiscale dynamics during visual recognition and memory in cortical circuits. Our objective is to infer *presentation-dependent* learning rules from *in vivo* recordings in ITC, similar to the online learning rules proposed in chapter 4. In these inferred learning rules, patterns presented many times to the network are gradually learned depending on the number of presentations, becoming progressively familiar from novel. We plan to implement the inferred online learning rules in an attractor neuronal network model analogously as in Pereira & Brunel (2018*a*). The objective is to reproduce the dynamics of the changes of firing rates across presentations, as well as exploring the

consequences for learning attractors in a network with online learning rules inferred from data.

6.3 Diversity of time scales in the prefrontal cortex

In chapters 3 and 4 we have shown that fixed-point attractors transition to chaotic attractors for strong synapses and high memory loads. They retain the information of the corresponding stored memories, and the network performs as an associative memory device with internally generated variability. We have proposed this scenario as an alternative mechanism for explaining the strong heterogeneity and temporal variability observed during delay response tasks in the prefrontal cortex (PFC). This scenario is consistent with previous studies showing that the coding of memories in the PFC is static (Murray et al. 2017), as is discussed in section 6.1. However, a quantitative comparison contrasting this model with available data is still lacking. Recent data from two different groups have shown that neurons in the prefrontal cortex show a diversity of time scales (see Fig 6.1) with distributions close to a log-normal (Cavanagh et al. 2018, Wasmuht et al. 2018). Interestingly, slow timescales neurons are more infromative about the retrieved memoranda than fast timescale neurons. Additionally, the coding of the memoranda by slow timescale neurons seems to be a combination of static and dynamic coding. In the model studied in chapters 3, the distribution of timescales is skewed similar to a log-normal distributions observed in Cavanagh et al. (2018), Wasmuht et al. (2018) but with narrower spread (compare Fig 6.1 A and B with Fig 6.1 C). This result is in apparent contradiction with the predictions of the theory in chapter 4, where the single neuron autocorrelation function is the same for all neuron up to a difference of order $1/\sqrt{N}$. In fact, we have shown that for the parameters inferred from data used in chapter 3 the network is in a mixed state in which has a large overlap with the retrieved memory and very small (but not negligible) overlap with all other stored memories. This mixed state is a consequence of strong finite size effects, and presents interesting properties

summarized as following: 1) Much larger capacities than what is predicted by the theory in chapter 4; 2) Non-self-averaging autocorrelation function (these results are not shown in this thesis). In the future, I would like to perform a quantitative comparison of the network model in chapter 3 using similar analyses as in Murray et al. (2017), Cavanagh et al. (2018), Wasmuht et al. (2018). Additionally, it will be ideal to also perform these analyses for the delay activity of the same neurons where the learning rules and transfer functions of the model in chapter 3 are inferred.

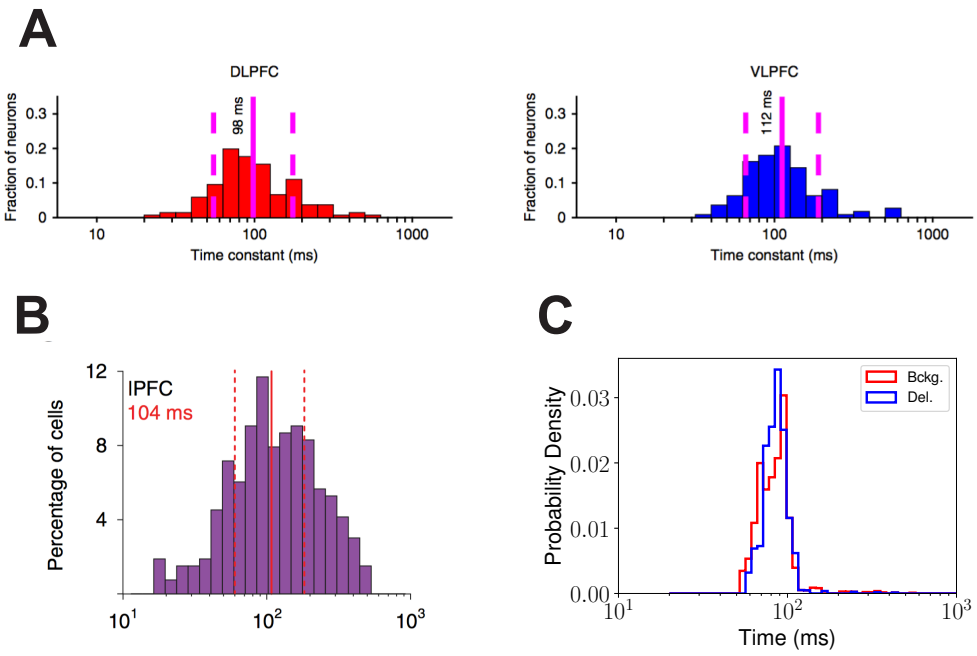

Figure 6.1: Diversity of time scales in PFC and in a chaotic attractor network model. (A) Distribution of time scales of dorsolateral and ventrolateral PFC neurons, adapted from Cavanagh et al. (2018). (B) Distribution of time scales in lateral PFC neurons, adapted from Wasmuht et al. (2018). (C) Distribution of time scales for 200 neurons in the attractor neuronal network model in chapter 3 for parameters shown in Figs 3.9 and 3.10. Time scales were computed as in Cavanagh et al. (2018).

6.4 Reinforcement learning of sequences

In chapter 5 we show that patterns of activity can be learned using an asymmetric unsupervised Hebbian learning rule. Since the stored patterns are random, the neuronal activations throughout the sequence are unstructured (see Fig 5.1). The sequences match well activity observed in posterior parietal cortex, hippocampus, and HVC. Nevertheless, when this model is assigned with the task of matching a particular sequence it fails. The reason is that in this model learning is unsupervised, therefore no error signal provides information to the network for precisely matching the target activity sequence. Supervised settings have been very successful for learning given sequences (Sussillo & Abbott 2009, Rajan et al. 2016). However the learning rules used (Sussillo & Abbott 2009, DePasquale et al. 2018, Rumelhart et al. 1985) are not biologically realistic. Furthermore, most models lack a theoretical understanding of their capacity and robustness. In numerical experiments, I have explored introducing arbitrary correlations between patterns in the model discussed in chapter 5 for matching a particular sequence of activity, with anecdotal success. An interesting scenario to explore is to combine unsupervised learning as in 5 with reinforcement learning in order to learn particular sequences of activity. The basic idea is the following: 1) random patterns of activity are learned using the unsupervised learning setting studied in chapter 5; 2) these patterns are then refined in a reinforcement learning setting using neuromodulator dependent learning rules (Frémaux et al. 2010, Frémaux & Gerstner 2016, Kuśmierz et al. 2017) for learning a particular sequence. The advantage of this model is twofold: 1) the unsupervised and reinforcement learning rules are biologically plausible and it is likely that the two class of learning happen concurrently in a single microcircuit in the cortex; 2) the network model is amenable to be theoretically analyzed using mean field techniques as in chapter 5.

Appendices

APPENDIX A

ATTRACTOR DYNAMICS IN NETWORKS WITH LEARNING RULES INFERRED FROM *IN VIVO* DATA

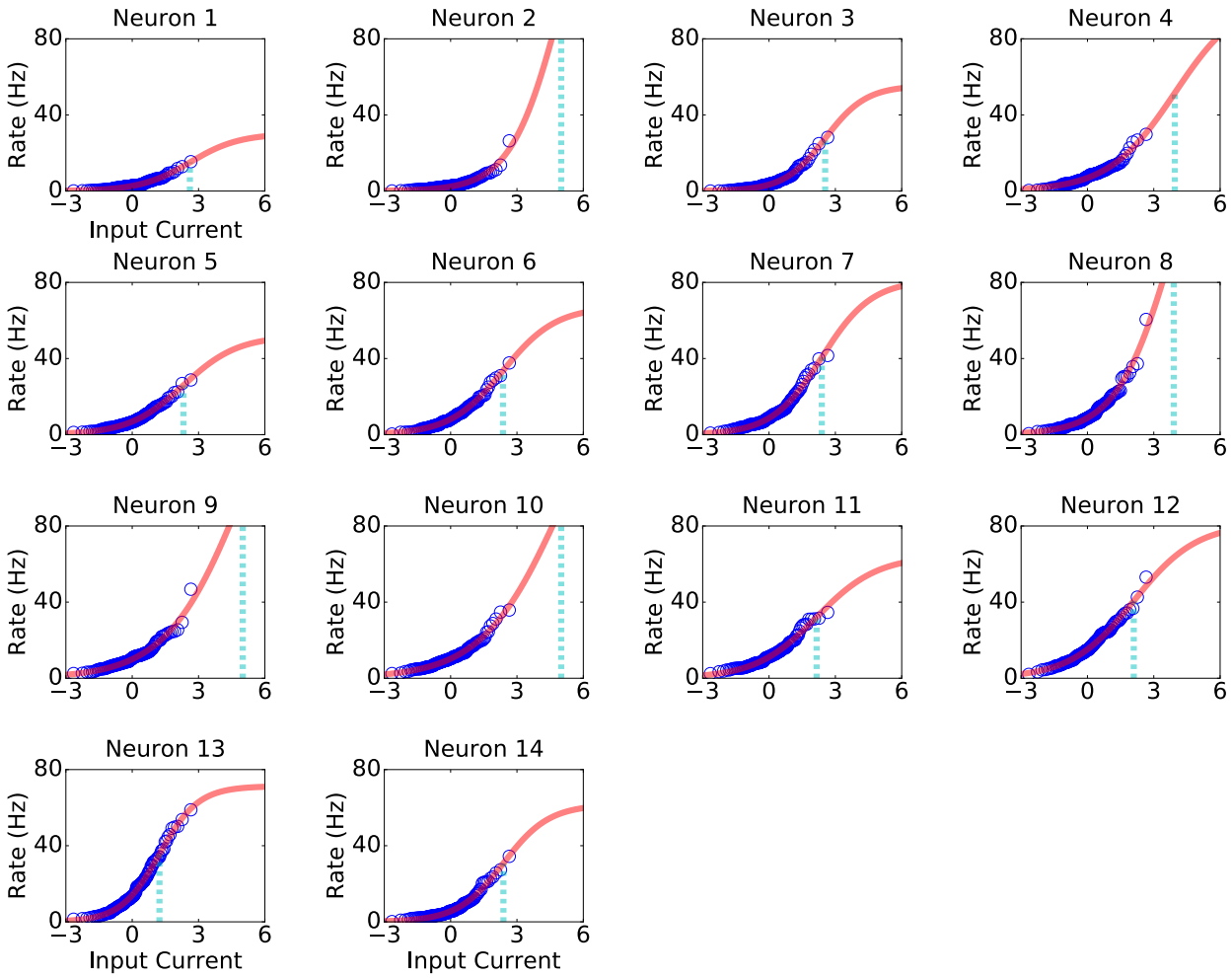

Figure A.1: Inferred static transfer functions. The static transfer function ϕ is derived from the distribution of visual responses for novel stimuli for 14 different ITC neurons using the procedure described in Lim et al. (2015). The data (blue circles) was fitted using a sigmoidal function (red line; see Methods, Eq. (3.48)). Cyan vertical dashed lines indicate the parameter h_0 of the sigmoidal fit. For details, see Methods main text.

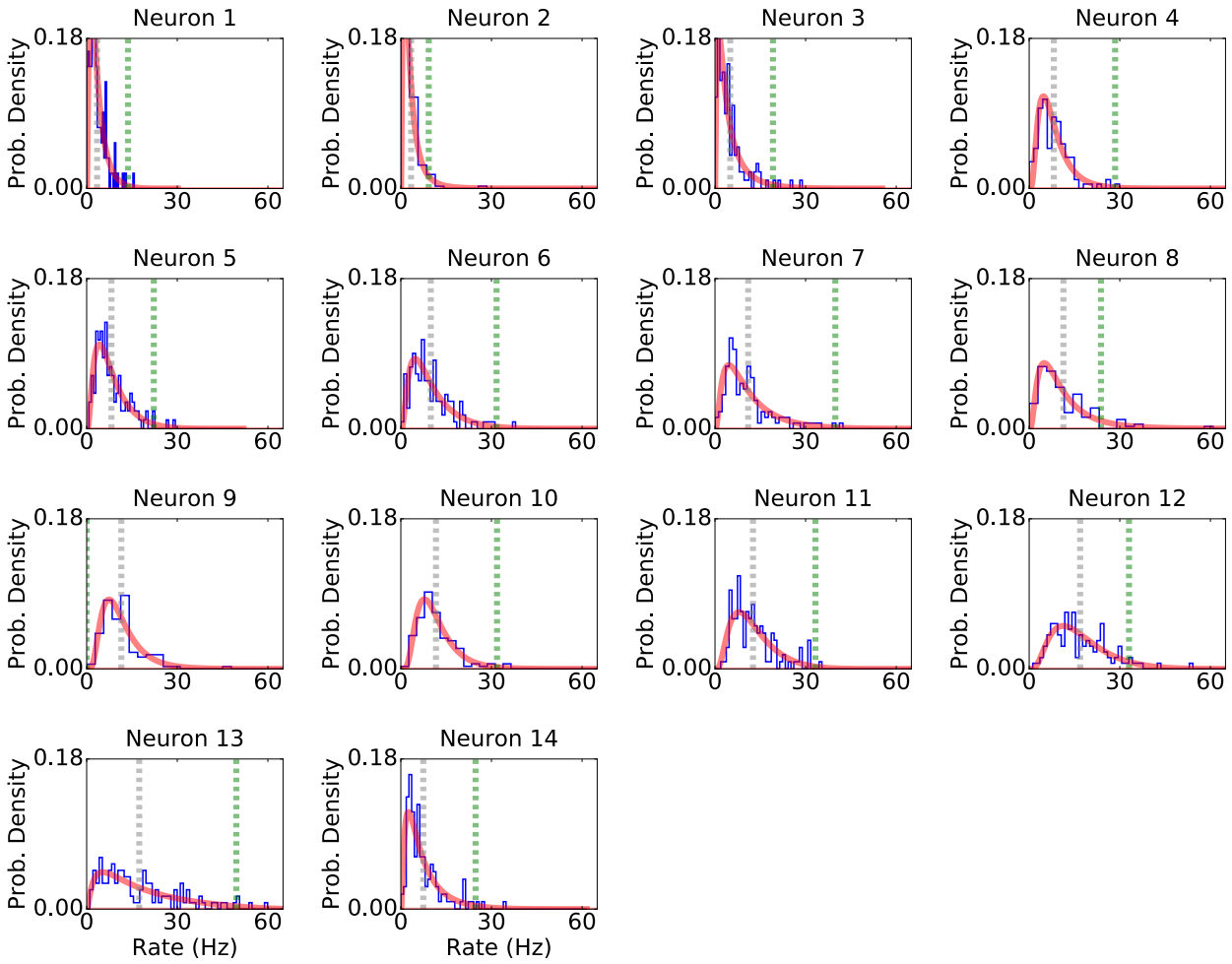

Figure A.2: Distributions of firing rates for novel stimuli. Distributions of firing rates in response to 125 novel stimuli for 14 ITC neurons. The firing rate histogram (blue) is plotted together with the distribution of firing rates (red line) obtained when standard normal patterns of current (i.e. $\xi \sim \mathcal{N}(0,1)$) are transformed using the static sigmoidal transfer function fitted in Fig. A.1 (i.e. $\phi(\xi)$; see Eq. (3.48) in Methods). The gray and green vertical dashed lines indicate the mean of the fitted firing rate distribution and the parameter x_f of the plasticity rule (see Fig. A.3).

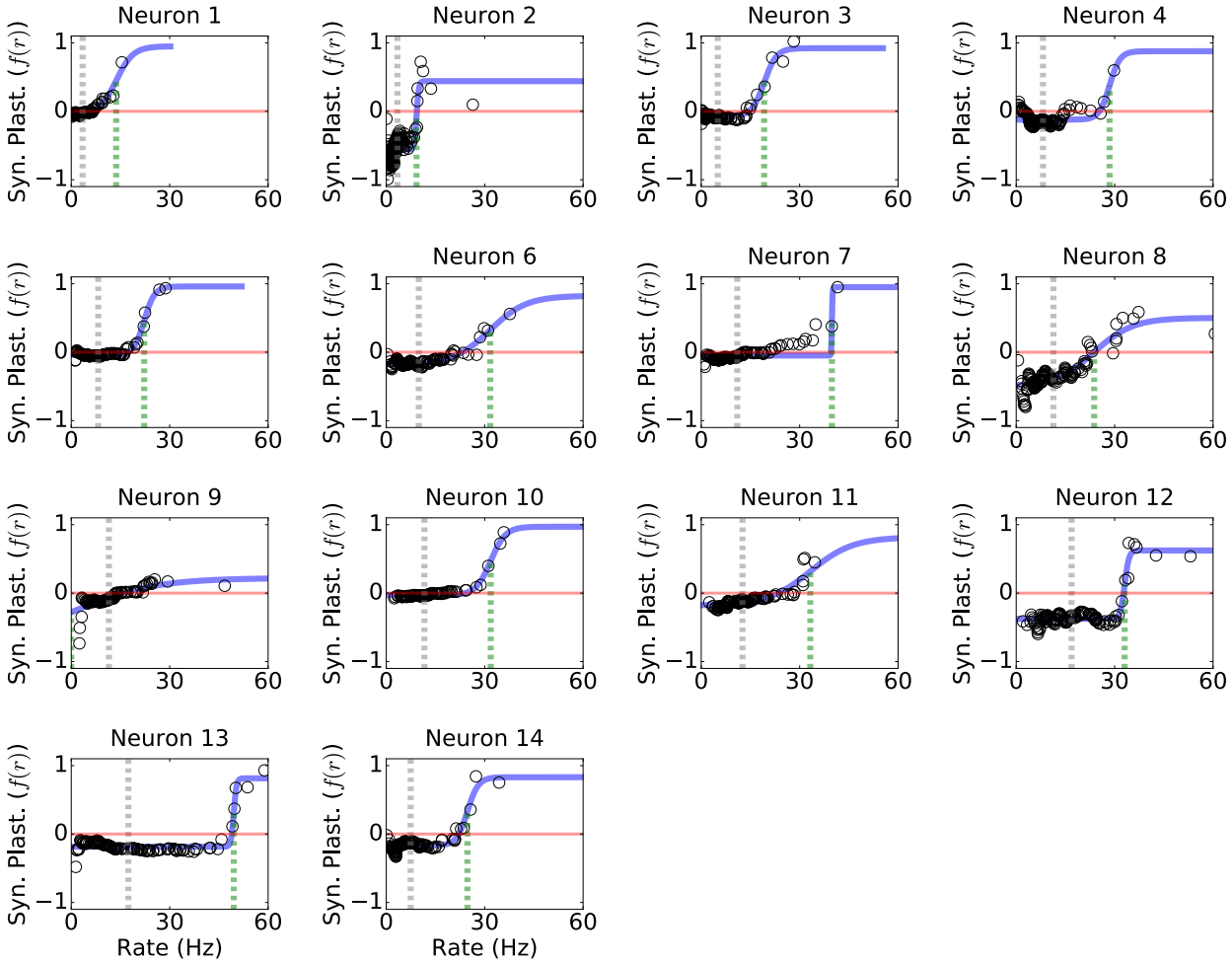

Figure A.3: Inferred dependence on the postsynaptic firing rate of the learning rule. The dependence of synaptic plasticity rule on the postsynaptic firing rate (i.e. $f(r)$) is inferred for 14 ITC neurons. The data is indicated with black circles and the sigmoidal fit with a blue line. The red line indicates the threshold between long term potentiation (LTP) and long term depression (LTD). As in Fig. A.2, the gray and green vertical dashed lines indicate the mean of the fitted firing rate distribution and the parameter x_f of the learning rule.

APPENDIX B

LOCAL-FIELD AUTO-COVARIANCE CALCULATION

B.1 Local-field auto-covariance calculation

Let us consider the auto-covariance of the fields in Eq. (4.13)

$$\Delta(\tau) = \mathrm{Cov}(h(t)h(t+\tau)). \tag{B.1}$$

Using the properties of the auto-covariance we obtain

$$\ddot{\Delta}(\tau) = \mathrm{Cov}(\dot{h}(t)\dot{h}(t+\tau)). \tag{B.2}$$

In our dynamic mean field theory, the dynamics of the network is approximated by a random gaussian field given by Eq. (4.8), i.e.:

$$\dot{h_i} = -h_i + Af(\eta_i^{\mu})\Theta(\mu)m_{\mu} + A\sqrt{\gamma\kappa}y(t), \tag{B.3}$$

where

$$C(\tau) = \mathrm{Cov}_y(y(t)y(t+\tau)) = \mathrm{Cov}_h(\phi(h(t))\phi(h(t+\tau))). \tag{B.4}$$

Here for simplicity we will set $A = 1$ and $\Theta(\mu) = 1$. The later implies that $\kappa = \alpha = p/Nc$. By using Eq. (B.3,B.4) we obtain

$$
\begin{aligned}
\ddot{\Delta}(\tau) &= \mathrm{Cov}_h([-h_i(t) + f(\eta_i)m + \sqrt{\alpha\gamma}y(t)]\,[-h_i(t+\tau) + f(\eta_i)m + \sqrt{\alpha\gamma}y(t+\tau)]) \\
&= \Delta(\tau) + \mathrm{Var}_\eta(f(\eta))m^2 + \alpha\gamma C(\tau) - m\mathrm{Cov}_h(h_i(t), f(\eta_i)) - m\mathrm{Cov}_h(h_i(t+\tau), f(\eta_i)) \\
&\quad -\sqrt{\alpha\gamma}\mathrm{Cov}_h(y_i(t+\tau), h(t)) - \sqrt{\alpha\gamma}\mathrm{Cov}_h(y_i(t), h(t+\tau)) \\
&\quad +m\mathrm{Cov}_h(y_i(t+\tau), f(\eta_i)) + m\mathrm{Cov}_h(y_i(t), f(\eta_i)) \qquad \text{(B.5)}
\end{aligned}
$$

In our theory, the random variable $y(t)$ represents the variability in the synaptic input current. For large networks, the synaptic input currents are uncorrelated with the particular pattern that is being retrieved. Therefore we have that $\mathrm{Cov}_h(y_i(t+\tau), f(\eta_i)) = \mathrm{Cov}_h(y_i(t), f(\eta_i)) = 0$. On the other hand,

$$
\begin{aligned}
\mathrm{Cov}_h(y_i(t+\tau), h(t)) &= \mathrm{Cov}_h(y_i(t+\tau), \sqrt{\alpha\gamma}y(t) + f(\eta_i) - \dot{h}_i(t)) \\
&= \sqrt{\alpha\gamma}C(\tau) - \mathrm{Cov}_h(y_i(t+\tau), \dot{h}_i(t)).
\end{aligned}
$$

Similarly

$$
\begin{aligned}
\mathrm{Cov}_h(y_i(t), h(t+\tau)) &= \mathrm{Cov}_h(y_i(t), \sqrt{\alpha\gamma}y(t+\tau) + f(\eta_i) - \dot{h}_i(t+\tau)) \\
&= \sqrt{\alpha\gamma}C(\tau) - \mathrm{Cov}_h(y_i(t), \dot{h}_i(t+\tau)).
\end{aligned}
$$

Lastly, considering

$$
\begin{aligned}
\mathrm{Cov}_h(y_i(t), \dot{h}_i(t+\tau)) + \mathrm{Cov}_h(y_i(t+\tau), \dot{h}_i(t)) &= \mathrm{Cov}_h(y_i(t), \dot{h}_i(t+\tau)) + \mathrm{Cov}_h(y_i(t'), \dot{h}_i(t'-\tau)) \\
&= \mathrm{Cov}_h(y_i(t), \dot{h}_i(t+\tau)) - \mathrm{Cov}_h(y_i(t'), \dot{h}_i(t'+\tau)) \\
&= 0,
\end{aligned}
$$

then Eq. (B.5) becomes

$$\ddot{\Delta}(\tau) = \Delta(\tau) + \mathrm{Var}_\eta(f(\eta))m^2 - \alpha\gamma C(\tau) - m\mathrm{Cov}_h(h_i(t), f(\eta_i)) - m\mathrm{Cov}_h(h_i(t+\tau), f(\eta_i)). \tag{B.6}$$

In Eq. (B.3) the synaptic input currents $h_i(t)$ are described by a gaussian random field, therefore can be written as

$$h_i(t) = \sqrt{\Delta_0 - |\Delta(\tau)|}x + \mathrm{sgn}(\Delta(\tau))\sqrt{|\Delta(\tau)|}z + f(\eta_i)m \tag{B.7}$$

$$h_i(t+\tau) = \sqrt{\Delta_0 - |\Delta(\tau)|}y + \sqrt{|\Delta(\tau)|}z + f(\eta_i)m, \tag{B.8}$$

where, x, y, z are independent standard normal random variables. This implies

$$\langle h_i(t)^2 \rangle = \Delta_0 \tag{B.9}$$

$$\langle h_i(t+\tau)^2 \rangle = \Delta_0 \tag{B.10}$$

$$\langle h_i(t)h_i(t+\tau) \rangle = \Delta(\tau). \tag{B.11}$$

$$\tag{B.12}$$

Then $\mathrm{Cov}_h(h_i(t), f(\eta_i)) = \mathrm{Cov}_h(h_i(t+\tau), f(\eta_i)) = m\mathrm{Var}_\eta(f(\eta))$, obtaining

$$\ddot{\Delta}(\tau) = \Delta(\tau) - \mathrm{Var}_\eta(f(\eta))m^2 - \alpha\gamma C(\tau). \tag{B.13}$$

Finally, by doing the following translation $\Delta(\tau) \to \Delta(\tau) - \mathrm{Var}_\eta(f(\eta))m^2$ we obtain Eq. (4.14).

APPENDIX C

UNSUPERVISED LEARNING OF SEQUENTIAL ACTIVITY WITH TEMPORALLY ASYMMETRIC HEBBIAN LEARNING RULES

C.1 Mixed States

Here we show that recurrent networks endowed with the covariance rule (Sejnowski 1977) learn attractor states correlated with multiple memories (i.e. mixed state) when the stored patterns are normally distributed. In this state, the retrieval of a single memory is not possible. This results was first found in the Hopfield model by Amit et al. (1985).

C.1.1 Pure state

We will first start analyzing the case of just one condensed pattern. When the network is in its steady state the incoming current to neuron i is given by

$$h_i = \frac{1}{Nc} \sum_{j \neq i}^{N} \sum_{k=1}^{p} c_{ij} \xi_i^k \xi_j^k r_j. \tag{C.1}$$

The mean field over the disorder produced by all the patterns and the structural connectivity (i.e. $C, \vec{\xi^2}, \ldots, \vec{\xi^p}$), conditional on the first pattern (i.e. $\vec{\xi^1} = \vec{z}$) is given by

$$\mathbb{E}_{\xi} \left(h_i | \vec{\xi} = \vec{z} \right) = z_i \mathbb{E}_{\xi, \xi^1} \left(\xi^1 r \right). \tag{C.2}$$

The conditional variance of the field over the disorder produced by all the patterns and the structural connectivity, conditional on the first pattern is given by

$$\mathrm{Var}_{\xi} \left(h_i | \vec{\xi} = \vec{z} \right) = \alpha \mathbb{E}_{\xi, \xi^1} \left(r^2 \right). \tag{C.3}$$

For computing Eq.(C.3) we use the fact that $c \ll 1$ to neglect correlations between neurons. As in the main text, we refer to the parameter $\alpha \equiv \frac{p}{Nc}$ as the memory load, which corresponds to the number of pattern per average number of synapses. We define the following order parameters

$$m = \mathbb{E}_{\xi,\xi^1}\left(\xi^1 r(\xi,\xi^1)\right) \tag{C.4}$$

$$M = \mathbb{E}_{\xi,\xi^1}\left(r^2(\xi,\xi^1)\right). \tag{C.5}$$

Where m, that we call the overlap, corresponds to the covariance between the first pattern and the steady state of the network, while M corresponds to the second moment of the steady state of the network. We compute the order parameters self-consistently by using the mean Eq. (C.2) and variance Eq. (C.3) of the field

$$m = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y z F(mz + \sqrt{\alpha M}y) \tag{C.6}$$

$$M = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y F^2(mz + \sqrt{\alpha M}y). \tag{C.7}$$

Where $\mathcal{D}z = dze^{-\frac{z^2}{2}}/\sqrt{2\pi}$ and similarly for $\mathcal{D}y$. By defining the following quantity $b^2 \equiv m^2 + \alpha M$ these equations simplify to

$$b = \int_{-\infty}^{\infty} \mathcal{D}v v F(bv) \tag{C.8}$$

$$M = \int_{-\infty}^{\infty} \mathcal{D}v F^2(bv). \tag{C.9}$$

Computing the order parameters m and M can be done as following: first b is calculated

self-consistently by using Eq. (C.8); second M is calculated by using Eq. (C.9) and finally m is calculated by

$$m = \sqrt{b^2 - \alpha M}. \tag{C.10}$$

When we compare our mean field equations with numerical simulations we find that the retrieval of one pattern (i.e. the steady state of the network is correlated with just one pattern) is not possible for a large parameter exploration. In contrast, our numerical simulations show that only *mixed states* where the steady of the network is correlated with a finite number of patterns is prevalent. In Fig. C.1 A it is shown the overlaps dynamics of a network with two-patterns-stored connectivity. After transients, the steady state of the network is correlated with both patterns for the four realizations depicted. In the next section, we will show that this is generic for any realization as it is shown in Fig. C.1 B.

C.1.2 Finite Number of Condensed Patterns

To understand this discrepancy between our previous MFT and the numerical simulations, we now assume that the steady state of the networks is correlated with the q first pattern learned with q finite i.e. $m_1, m_2, \ldots, m_q$. In other words, we assume that the steady state of the networks depends on the q first pattern learned, and on the rest uncondensed $p - q$ patterns only depend indirectly through the field h that they produce. Since in our analysis, the number of patterns is assumed to be large $q \ll p$. Then the steady state of the network in this case is approximated by

$$r = F(h|\xi^1, \ldots, \xi^q). \tag{C.11}$$

Then, the conditional mean field over the disorder produced by the last $(p - q)$ patterns and the structural connectivity (i.e. $C, \vec{\xi^{q+1}}, \ldots, \vec{\xi^p}$), conditional on the first q pattern (i.e.

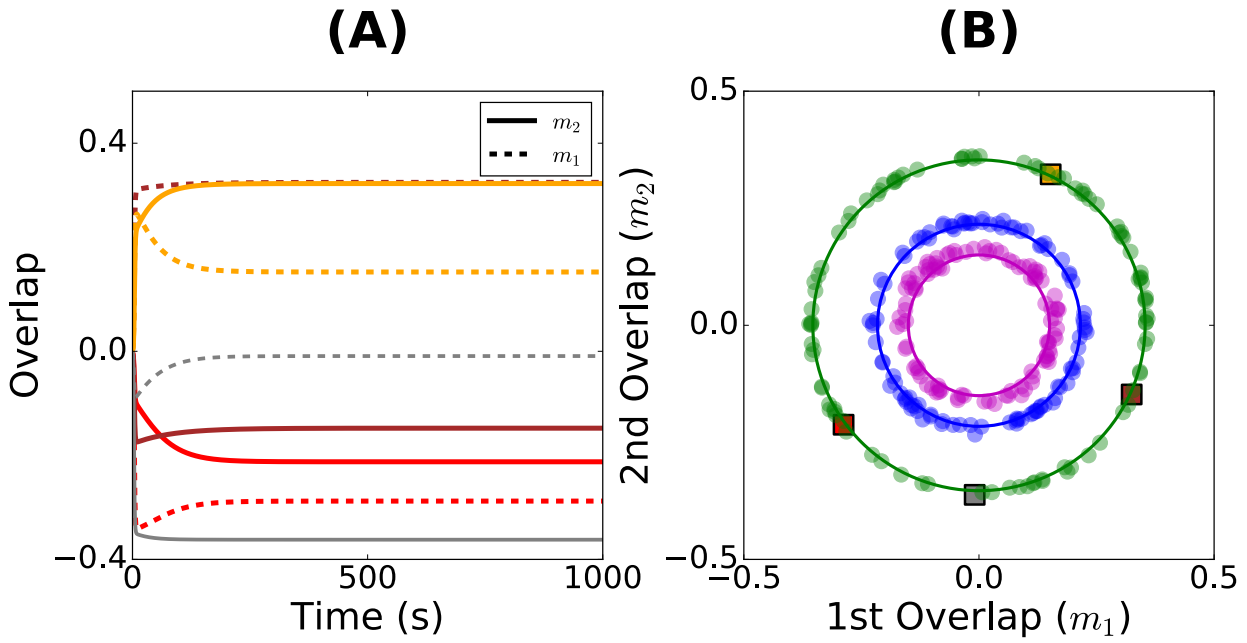

Figure C.1: (**A**) Numerically computed overlaps vs time for a two-patterns-stored connectivity (i.e. $p = 2$). Four different realizations of the network are shown in yellow, gray, red and maroon. In dashed and continuous lines are shown respectively the overlaps with the fist and second pattern. In these four realizations, after transients, the steady state of the network is correlated with both patterns. (**B**) In solid circles, it is shown the numerically computed overlaps after transients placed in the m_1-m_2 plane for one hundred realizations of the network. Circumferences with the radius given by Eq. C.19-C.21 are the manifolds where lie the overlaps in the m_1-m_2 plane predicted by our MFT for a two-patterns-stored connectivity. In green, blue and magenta solid circles (numerical simulations) and circumferences (MFT) are shown the results for three different parameters used $\beta = 10$ and $h_0 = 0$, $\beta = 5$ and $h_0 = 0$ and $\beta = 5$ and $h_0 = 0.15$ respectively, with $r_{\max} = 1$. In yellow, gray, red and maroon squares are placed in the m_1-m_2 plane the overlaps depicted in Fig. C.1 A. For these simulations the network parameters were $c = 0.005$ and $N = 5 \cdot 10^5$.

$\vec{\xi^1} = \vec{z}_1, \vec{\xi^2} = \vec{z}_2, \ldots, \vec{\xi^q} = \vec{z^q}$) is given by

$$\mathbb{E}_\xi \left(h_i | \vec{\xi^1} = \vec{z^1}, \ldots, \vec{\xi^q} = \vec{z^q} \right) = \sum_{k=1}^{q} m_k z_i^k. \tag{C.12}$$

With the order parameter m_k defined by

$$m_k = \mathbb{E}_{\xi^1, \ldots, \xi^q, \xi}(\xi^k r). \tag{C.13}$$

On the other hand, the conditional variance of the field over the disorder produced by the last $(p-q)$ patterns and the structural connectivity (i.e. $C, \vec{\xi^{q+1}}, \ldots, \vec{\xi^p}$), conditional on the first q pattern (i.e. $\vec{\xi^1} = \vec{z}_1, \vec{\xi^2} = \vec{z}_2, \ldots, \vec{\xi^q} = \vec{z^q}$) is given by

$$\mathrm{Var}_\xi \left(h_i | \vec{\xi^1} = \vec{z^1}, \ldots, \vec{\xi^q} = \vec{z^q} \right) = \alpha M. \tag{C.14}$$

With the order parameter M given by

$$M = \mathbb{E}_{\xi^1, \ldots, \xi^q, \xi} \left(r^2(\xi^1, \ldots, \xi^q, h) \right). \tag{C.15}$$

Using the central limit theorem, we approximate the distribution of the field over the disorder produced by the last $(p-q)$ patterns and the structural connectivity conditional to the q condensed patterns to

$$p(h | \xi^1 = z^1, \ldots, \xi^q = z^q) = \mathcal{N} \left(\sum_{l=1}^{q} z^l m_l, \sqrt{\alpha M} \right), \tag{C.16}$$

with $\xi^k \overset{iid}{\sim} \mathcal{N}(0,1)$. By using the fact in the steady state $r_i = F(h_i)$, we write the self-consistent mean field equations for the order parameters as following

$$m_k = \mathbb{E}_{\xi^1,\ldots,\xi^q,y}(\xi^k F(m_1\xi^1 + m_2\xi^2 + \cdots + m_q\xi^q + \sqrt{\alpha M}y)) \quad k = 1,\ldots,q \tag{C.17}$$

$$M = \mathbb{E}_{\xi^1,\ldots,\xi^q,y}(F^2(m_1\xi^1 + m_2\xi^2 + \cdots + m_q\xi^q + \sqrt{\alpha M}y)), \tag{C.18}$$

where y is a standard normal random variable. These $q+1$ equations can be reduced to three equations given by

$$b = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dv e^{-\frac{v^2}{2}} vF(bv) \tag{C.19}$$

$$M = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dv e^{-\frac{v^2}{2}} F^2(bv) \tag{C.20}$$

$$\sum_{l=1}^{q} m_l^2 = b^2 - \alpha M. \tag{C.21}$$

Equations (C.19) and (C.20) are equivalent to equations (C.8) and (C.9) obtained in the one-condensed-pattern case analyzed in the previous section. On the other hand, Eq. (C.10) is the *one-pattern version* of Eq. (C.21). This analysis shows that for the covariance rule, retrieval states which are correlated with a finite number of patterns exist. Moreover, there is a continuum of such states, that lie on a manifold described by the surface of the hypersphere $\sum_{l=1}^{q} m_l^2 = b^2 - \alpha M$. Thus, 'pure' retrieval states (i.e. states correlated with just a single stored pattern) are only marginally stable. In finite networks, numerical simulations find only *mixed states*, consistent with symmetry breaking that lead to a discrete set of mixed states as the only possible attractors of the system. Therefore, with this rule retrieval of a single memory is not possible. In figure Fig. C.1 B it is shown the circumference where the overlaps are predicted to lie by our MFT (i.e. Eq. C.19-C.21) for a two-patterns-stored network. There is a good agreement between our theory and numerical simulations for multiples realizations of the network.

REFERENCES

Abbott, L. F. & Blum, K. I. (1996), 'Functional significance of long-term potentiation for sequence learning and prediction', *Cereb. Cortex* **6**, 406–416.

Abbott, L. F. & Nelson, S. B. (2000), 'Synaptic plasticity: taming the beast', *Nature neuroscience* **3**(11s), 1178.

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

Albus, J. S. (1971), 'A theory of cerebellar function', *Mathematical Biosciences* **10**(1-2), 25–61.

Amador, A., Perl, Y. S., Mindlin, G. B. & Margoliash, D. (2013), 'Elemental gesture dynamics are encoded by song premotor cortical neurons', *Nature* **495**(7439), 59–64.

Amari, S.-I. (1972), 'Learning patterns and pattern sequences by self-organizing nets of threshold elements', *Computers, IEEE Transactions on* **100**(11), 1197–1206.

Amit, D., Gutfreund, H. & Sompolinsky, H. (1987), 'Statistical mechanics of neural networks near saturation', *Annals of physics* **173**(1), 30–67.

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

Amit, D. J. (1995), 'The hebbian paradigm reintegrated: local reverberations as internal representations', *Behav. Brain Sci.* **18**, 617.

Amit, D. J. & Brunel, N. (1997), 'Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex.', *Cerebral cortex* **7**(3), 237–252.

Amit, D. J., Brunel, N. & Tsodyks, M. (1994), 'Correlations of cortical hebbian reverberations: theory versus experiment', *The Journal of neuroscience* **14**(11), 6435–6445.

Amit, D. J. & Fusi, S. (1994), 'Learning in neural networks with material synapses', *Neural Computation* **6**(5), 957–982.

Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985), 'Spin-glass models of neural networks', *Physical Review A* **32**(2), 1007.

Amit, Y. & Huang, Y. (2010), 'Precise capacity analysis in binary networks with multiple coding level inputs', *Neural computation* **22**(3), 660–688.

Anderson, J. S., Lampl, I., Gillespie, D. C. & Ferster, D. (2000), 'The contribution of noise to contrast invariance of orientation tuning in cat visual cortex', *Science* **290**, 1968–1972.

Artola, A., Bröcher, S. & Singer, W. (1990), 'Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex', *Nature* **347**(6288), 69.

Barak, O., Sussillo, D., Romo, R., Tsodyks, M. & Abbott, L. F. (2013), 'From fixed points to chaos: three models of delayed discrimination', *Prog. Neurobiol.* **103**, 214–222.

Barak, O. & Tsodyks, M. (2014), 'Working models of working memory', *Curr. Opin. Neurobiol.* **25**, 20–24.

Barak, O., Tsodyks, M. & Romo, R. (2010), 'Neuronal population coding of parametric working memory', *J. Neurosci.* **30**, 9424–9430.

Barbieri, F. & Brunel, N. (2007), 'Irregular persistent activity induced by synaptic excitatory feedback', *Frontiers in Computational Neuroscience* **1**, 5.

Bell, C. C., Han, V. Z., Sugawara, Y. & Grant, K. (1997), 'Synaptic plasticity in a cerebellum-like structure depends on temporal order', *Nature* **387**(6630), 278.

Benna, M. K. & Fusi, S. (2016), 'Computational principles of synaptic memory consolidation', *Nature neuroscience* **19**(12), 1697.

Bi, G.-q. & Poo, M.-m. (1998), 'Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type', *The Journal of neuroscience* **18**(24), 10464–10472.

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

Binder, J. R. & Desai, R. H. (2011), 'The neurobiology of semantic memory', *Trends in cognitive sciences* **15**(11), 527–536.

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. (2017), 'Behavioral time scale synaptic plasticity underlies ca1 place fields', *Science* **357**(6355), 1033–1036.

Bliss, T. V. & Lømo, T. (1973), 'Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path', *The Journal of physiology* **232**(2), 331–356.

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

Bourne, J. N. & Harris, K. M. (2011), 'Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal ca1 dendrites during ltp', *Hippocampus* **21**(4), 354–373.

Bouvier, G., Aljadeff, J., Clopath, C., Bimbard, C., Nadal, J.-P., Brunel, N., Hakim, V. & Barbour, B. (2017), 'Cerebellar learning using perturbations', *bioRxiv* p. 053785.

Brody, C. D., Hernández, A., Zainos, A. & Romo, R. (2003), 'Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex', *Cerebral cortex* **13**(11), 1196–1207.

Brunel, N. (2000), 'Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons', *Journal of computational neuroscience* **8**(3), 183–208.

Brunel, N. (2003), 'Dynamics and plasticity of stimulus-selective persistent activity in cortical network models', *Cerebral Cortex* **13**(11), 1151–1161.

Brunel, N. (2005), Network models of memory, *in* C. Chow, B. Gutkin, D. Hansel, C. Meunier & J. Dalibard, eds, 'Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School 2003', Elsevier, pp. 407–476.

Brunel, N. (2016), 'Is cortical connectivity optimized for storing information?', *Nature neuroscience* .

Brunel, N., Hakim, V., Isope, P., Nadal, J.-P. & Barbour, B. (2004), 'Optimal information storage and the distribution of synaptic weights: perceptron versus purkinje cell', *Neuron* **43**(5), 745–757.

Brunel, N. & Wang, X. J. (2001), 'Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition', *J. Comput. Neurosci.* **11**, 63–85.

Buzsaki, G. & Mizuseki, K. (2014), 'The log-dynamic brain: how skewed distributions affect network operations', *Nat. Rev. Neurosci.* **15**, 264–278.

Cannon, J., Kopell, N., Gardner, T. & Markowitz, J. (2015), 'Neural sequence generation using spatiotemporal patterns of inhibition', *PLoS Comput Biol* **11**(11), e1004581.

Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. (2018), 'Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex', *Nature communications* **9**(1), 3498.

Chafee, M. V. & Goldman-Rakic, P. S. (1998), 'Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task', *J. Neurophysiol.* **79**(6), 2919–2940.

Chenkov, N., Sprekeler, H. & Kempter, R. (2017), 'Memory replay in balanced recurrent networks', *PLoS Comput. Biol.* **13**, e1005359.

Clopath, C. & Gerstner, W. (2010), 'Voltage and spike timing interact in stdp–a unified model', *Frontiers in synaptic neuroscience* **2**, 25.

Compte, A., Constantinidis, C., Tegnér, J., Raghavachari, S., Chafee, M., Goldman-Rakic, P. S. & Wang, X.-J. (2003), 'Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task', *J. Neurophysiol.* **90**, 3441–3454.

Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. (2001), 'Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex', *Journal of Neuroscience* **21**(10), 3646–3655.

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M. & Arnsten, A. F. (2018), 'Persistent spiking activity underlies working memory', *Journal of Neuroscience* **38**(32), 7020–7028.

Crisanti, A. & Sompolinsky, H. (2018), 'Path integral approach to random neural networks', *arXiv preprint arXiv:1809.06042* .

Dayan, P. & Abbott, L. F. (2001), *Theoretical neuroscience*, Vol. 806, Cambridge, MA: MIT Press.

DePasquale, B., Cueva, C. J., Rajan, K., Abbott, L. et al. (2018), 'full-force: A target-based method for training recurrent networks', *PloS one* **13**(2), e0191527.

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

Destexhe, A., Mainen, Z. F. & Sejnowski, T. J. (1998), Kinetic models of synaptic transmission, *in* C. Koch & I. Segev, eds, 'Methods in Neuronal Modeling', 2nd edn, MIT press, Cambridge, MA, pp. 1–25.

Diesmann, M., Gewaltig, M.-O. & Aertsen, A. (1999), 'Stable propagation of synchronous spiking in cortical neural networks', *Nature* **402**(6761), 529–533.

Druckmann, S. & Chklovskii, D. B. (2012), 'Neuronal circuits underlying persistent representations despite time varying activity', *Current Biology* **22**(22), 2095–2103.

Dubreuil, A. M., Amit, Y. & Brunel, N. (2014), 'Memory capacity of networks with stochastic binary synapses', *PLoS computational biology* **10**(8), e1003727.

Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. (2000), 'Neurocomputational models of working memory', *Nature neuroscience* **3**, 1184–1191.

Erickson, C. A. & Desimone, R. (1999), 'Responses of macaque perirhinal neurons during and after visual stimulus association learning', *J. Neurosci.* **19**(23), 10404–10416.

Ermentrout, G. B. & Terman, D. H. (2010), *Mathematical foundations of neuroscience*, Vol. 35, Springer Science & Business Media.

Festa, D., Hennequin, G. & Lengyel, M. (2014), Analog memories in a balanced rate-based network of e-i neurons, *in* Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems 27', Curran Associates, Inc., pp. 2231–2239.

Fiete, I. R., Senn, W., Wang, C. Z. & Hahnloser, R. H. (2010), 'Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity', *Neuron* **65**(4), 563–576.

Foster, D. J. & Wilson, M. A. (2006), 'Reverse replay of behavioural sequences in hippocampal place cells during the awake state', *Nature* **440**(7084), 680–683.

Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. (2001), 'Categorical representation of visual stimuli in the primate prefrontal cortex', *Science* **291**(5502), 312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. (2006), 'Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex', *Cerebral Cortex* **16**(11), 1631–1644.

Frémaux, N. & Gerstner, W. (2016), 'Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules', *Frontiers in neural circuits* **9**, 85.

Frémaux, N., Sprekeler, H. & Gerstner, W. (2010), 'Functional requirements for reward-modulated spike-timing-dependent plasticity', *Journal of Neuroscience* **30**(40), 13326–13337.

Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. (1989), 'Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex', *Journal of neurophysiology* **61**(2), 331–349.

Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. (1990), 'Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms', *Journal of Neurophysiology* **63**(4), 814–831.

Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. (1991), 'Neuronal activity related to saccadic eye movements in the monkey's dorsolateral prefrontal cortex', *Journal of neurophysiology* **65**(6), 1464–1483.

Fuster, J. M., Alexander, G. E. et al. (1971), 'Neuron activity related to short-term memory', *Science* **173**(3997), 652–654.

Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1982), 'Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks', *Experimental neurology* **77**(3), 679–694.

Fuster, J. M. & Jervey, J. P. (1981), 'Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli', *Science* **212**(4497), 952–955.

Gardner, E. (1987), 'Maximum storage capacity in neural networks', *EPL (Europhysics Letters)* **4**(4), 481.

Gerstner, W. & Abbott, L. (1997), 'Learning navigational maps through potentiation and modulation of hippocampal place cells', *Journal of computational neuroscience* **4**(1), 79–94.

Gerstner, W., Kistler, W. M., Naud, R. & Paninski, L. (2014), *Neuronal dynamics: From single neurons to networks and models of cognition*, Cambridge University Press.

Gerstner, W. & van Hemmen, J. L. (1992), 'Associative memory in a network of ?spiking?neurons', *Network: Computation in Neural Systems* **3**(2), 139–164.

Gjorgjieva, J., Clopath, C., Audet, J. & Pfister, J.-P. (2011), 'A triplet spike-timing–dependent plasticity model generalizes the bienenstock–cooper–munro rule to higher-order spatiotemporal correlations', *Proceedings of the National Academy of Sciences* **108**(48), 19383–19388.

Glimcher, P. W. (2011), 'Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis', *Proceedings of the National Academy of Sciences* **108**(Supplement 3), 15647–15654.

Goldman-Rakic, P. S. (1995), 'Cellular basis of working memory', *Neuron* **14**(3), 477–485.

Graupner, M. & Brunel, N. (2012), 'Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location', *Proceedings of the National Academy of Sciences* **109**(10), 3991–3996.

Grosmark, A. D. & Buzsáki, G. (2016), 'Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences', *Science* **351**(6280), 1440–1443.

Grossberg, S. (1969), 'On learning, information, lateral inhibition, and transmitters', *Mathematical Biosciences* **4**(3-4), 255–310.

Guckenheimer, J. & Holmes, P. (2013), *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Vol. 42, Springer Science & Business Media.

Guo, Z. V., Li, N., Huber, D., Ophir, E., Gutnisky, D., Ting, J. T., Feng, G. & Svoboda, K. (2014), 'Flow of cortical activity underlying a tactile decision in mice', *Neuron* **81**(1), 179–194.

Guzman, S. J., Schlögl, A., Frotscher, M. & Jonas, P. (2016), 'Synaptic mechanisms of pattern completion in the hippocampal ca3 network', *Science* **353**(6304), 1117–1123.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. (2005), 'Microstructure of a spatial map in the entorhinal cortex', *Nature* **436**(7052), 801.

Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. (2002), 'An ultra-sparse code underliesthe generation of neural sequences in a songbird', *Nature* **419**(6902), 65–70.

Harish, O. & Hansel, D. (2015), 'Asynchronous rate chaos in spiking neuronal circuits', *PLoS Comput Biol* **11**(7), e1004266.

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

Hasselmo, M. E. (2006), 'The role of acetylcholine in learning and memory', *Curr. Opin. Neurobiol.* **16**, 710–715.

Hebb, D. (1949), *The organization of behavior: A neuropsychological theory*, John Wiley.

Herring, B. E. & Nicoll, R. A. (2016), 'Long-term potentiation: from camkii to ampa receptor trafficking', *Annual review of physiology* **78**, 351–365.

Herz, A., Sulzer, B., Kühn, R. & van Hemmen, J. L. (1988), 'The Hebb rule: Storing static and dynamic objects in an associative neural network', *Europhys. Lett.* **7**, 663–669.

Holmgren, C., Harkany, T., Svennenfors, B. & Zilberter, Y. (2003), 'Pyramidal cell communication within local networks in layer 2/3 of rat neocortex', *The Journal of physiology* **551**(1), 139–153.

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

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.

Hromadka, T., Deweese, M. R. & Zador, A. M. (2008), 'Sparse representation of sounds in the unanesthetized auditory cortex', *PLoS Biol.* **6**, e16.

Huang, G., Ramachandran, S., Lee, T. S. & Olson, C. R. (2018), 'Neural correlate of visual familiarity in macaque area v2', *Journal of Neuroscience* pp. 0664–18.

Huang, Y. & Amit, Y. (2011), 'Capacity analysis in multi-state synaptic models: a retrieval probability perspective', *Journal of computational neuroscience* **30**(3), 699–720.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. (2016), 'Natural speech reveals the semantic maps that tile human cerebral cortex', *Nature* **532**(7600), 453.

Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. (2017), 'Discrete attractor dynamics underlying selective persistent activity in frontal cortex', *Biorxiv* p. 203448.

Izhikevich, E. M. (2006), 'Polychronization: computation with spikes', *Neural computation* **18**(2), 245–282.

Jahnke, S., Timme, M. & Memmesheimer, R. M. (2015), 'A Unified Dynamic Model for Learning, Replay, and Sharp-Wave/Ripples', *J. Neurosci.* **35**, 16236–16258.

Jun, J. K. & Jin, D. Z. (2007), 'Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity', *PLoS One* **2**(8), e723.

Kadmon, J. & Sompolinsky, H. (2015), 'Transition to chaos in random neuronal networks', *Physical Review X* **5**(4), 041030.

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

Keck, T., Toyoizumi, T., Chen, L., Doiron, B., Feldman, D. E., Fox, K., Gerstner, W., Haydon, P. G., Hübener, M., Lee, H.-K. et al. (2017), 'Integrating hebbian and homeostatic plasticity: the current state of the field and future research directions', *Phil. Trans. R. Soc. B* **372**(1715), 20160158.

Kempter, R., Gerstner, W. & Van Hemmen, J. L. (1999), 'Hebbian learning and spiking neurons', *Physical Review E* **59**(4), 4498.

Kiani, R., Esteky, H., Mirpour, K. & Tanaka, K. (2007), 'Object category structure in response patterns of neuronal population in monkey inferior temporal cortex', *Journal of neurophysiology* **97**(6), 4296–4309.

Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. (2017), 'Ring attractor dynamics in the drosophila central brain', *Science* **356**(6340), 849–853.

Kleinfeld, D. (1986), 'Sequential state generation by model neural networks', *Proceedings of the National Academy of Sciences* **83**(24), 9469–9473.

Kleinfeld, D. & Sompolinsky, H. (1988), 'Associative neural network model for the generation of temporal patterns. theory and application to central pattern generators.', *Biophysical Journal* **54**(6), 1039.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X. L., Romo, R., Uchida, N. & Machens, C. K. (2016), 'Demixed principal component analysis of neural population data', *Elife* **5**.

Kobatake, E., Wang, G. & Tanaka, K. (1998), 'Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys', *Journal of Neurophysiology* **80**(1), 324–330.

Koch, K. & Fuster, J. (1989*a*), 'Unit activity in monkey parietal cortex related to haptic perception and temporary memory', *Experimental Brain Research* **76**(2), 292–306.

Koch, K. W. & Fuster, J. M. (1989*b*), 'Unit activity in monkey parietal cortex related to haptic perception and temporary memory', *Exp. Brain Res.* **76**, 292–306.

Kree, R. & Zippelius, A. (1987), 'Continuous-time dynamics of asymmetrically diluted neural networks', *Phys Rev A Gen Phys* **36**, 4421–4427.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K. & Bandettini, P. A. (2008), 'Matching categorical object representations in inferior temporal cortex of man and monkey', *Neuron* **60**(6), 1126–1141.

Kuhn, R. & van Hemmen, J. L. (1991), Temporal association, *in* E. Domany, J. L. van Hemmen & K. Schulten, eds, 'Models of Neural Networks', Springer, pp. 221–285.

Kuśmierz, Ł., Isomura, T. & Toyoizumi, T. (2017), 'Learning with three factors: modulating hebbian plasticity with errors', *Current opinion in neurobiology* **46**, 170–177.

Lahiri, S. & Ganguli, S. (2013), A memory frontier for complex synapses, *in* 'Advances in neural information processing systems', pp. 1034–1042.

Laje, R. & Buonomano, D. V. (2013), 'Robust timing and motor patterns by taming chaos in recurrent neural networks', *Nature neuroscience* **16**(7), 925–933.

Lansner, A. (2009), 'Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations', *Trends Neurosci.* **32**(3), 178–186.

Lefort, S., Tomm, C., Sarria, J.-C. F. & Petersen, C. C. (2009), 'The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex', *Neuron* **61**(2), 301–316.

Lehky, S. R., Kiani, R., Esteky, H. & Tanaka, K. (2011), 'Statistics of visual responses in primate inferotemporal cortex to object stimuli', *J. Neurophysiol.* **106**, 1097–1117.

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

Li, L., Miller, E. K. & Desimone, R. (1993), 'The representation of stimulus familiarity in anterior inferior temporal cortex', *Journal of neurophysiology* **69**(6), 1918–1929.

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L. & Brunel, N. (2015), 'Inferring learning rules from distributions of firing rates in cortical neurons', *Nature neuroscience* .

Lisman, J., Yasuda, R. & Raghavachari, S. (2012), 'Mechanisms of camkii action in long-term potentiation', *Nature reviews neuroscience* **13**(3), 169.

Litwin-Kumar, A. & Doiron, B. (2014*a*), 'Formation and maintenance of neuronal assemblies through synaptic plasticity', *Nature communications* **5**.

Litwin-Kumar, A. & Doiron, B. (2014*b*), 'Formation and maintenance of neuronal assemblies through synaptic plasticity', *Nat Commun* **5**, 5319.

Liu, D., Gu, X., Zhu, J., Zhang, X., Han, Z., Yan, W., Cheng, Q., Hao, J., Fan, H., Hou, R. et al. (2014), 'Medial prefrontal activity during delay period contributes to learning of a working memory task', *Science* **346**(6208), 458–463.

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

Logothetis, N. K., Pauls, J. & Poggio, T. (1995), 'Shape representation in the inferior temporal cortex of monkeys', *Current Biology* **5**(5), 552–563.

Lomo, T. (1966), Frequency potentiation of excitatory synaptic activity in dentate area of hippocampal formation, *in* 'Acta Physiologica Scandinavica', BLACKWELL SCIENCE LTD PO BOX 88, OSNEY MEAD, OXFORD OX2 0NE, OXON, ENGLAND, p. 128.

Lundqvist, M., Compte, A. & Lansner, A. (2010), 'Bistable, irregular firing and population oscillations in a modular attractor memory network', *PLoS Comput. Biol.* **6**, e1000803.

Lundqvist, M., Herman, P. & Miller, E. K. (2018), 'Working memory: Delay activity, yes! persistent activity? maybe not', *Journal of Neuroscience* **38**(32), 7013–7019.

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J. & Miller, E. K. (2016), 'Gamma and beta bursts underlie working memory', *Neuron* **90**(1), 152–164.

Magee, J. C. & Johnston, D. (1997), 'A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons', *Science* **275**(5297), 209–213.

Markram, H., Lübke, J., Frotscher, M., Roth, A. & Sakmann, B. (1997), 'Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex.', *The Journal of physiology* **500**(2), 409–440.

Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. (1997), 'Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps', *Science* **275**(5297), 213–215.

Marr, D. (1969), 'A theory of cerebellar cortex', *The Journal of physiology* **202**(2), 437–470.

Mason, A., Nicoll, A. & Stratford, K. (1991), 'Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro', *Journal of Neuroscience* **11**(1), 72–84.

McCormick, D. A., Connors, B. W., Lighthall, J. W. & Prince, D. A. (1985), 'Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex', *Journal of neurophysiology* **54**(4), 782–806.

Mehta, M. R., Barnes, C. A. & McNaughton, B. L. (1997), 'Experience-dependent, asymmetric expansion of hippocampal place fields', *Proc. Natl. Acad. Sci. USA* **94**, 8918–8921.

Memmesheimer, R.-M., Rubin, R., Ölveczky, B. P. & Sompolinsky, H. (2014), 'Learning precisely timed spikes', *Neuron* **82**(4), 925–938.

Mézard, M., Nadal, J.-P. & Toulouse, G. (1986*a*), 'Solvable models of working memories', *J. Physique* **47**, 1457–.

Mézard, M., Nadal, J. & Toulouse, G. (1986*b*), 'Solvable models of working memories', *Journal de physique* **47**(9), 1457–1462.

Miller, E. K., Erickson, C. A. & Desimone, R. (1996*a*), 'Neural mechanisms of visual working memory in prefrontal cortex of the macaque', *J. Neurosci.* **16**(16), 5154–5167.

Miller, E. K., Erickson, C. A. & Desimone, R. (1996*b*), 'Neural mechanisms of visual working memory in prefrontal cortex of the macaque', *The Journal of Neuroscience* **16**(16), 5154–5167.

Miller, K. D. & Fumarola, F. (2012), 'Mathematical equivalence of two common forms of firing rate models of neural networks', *Neural computation* **24**(1), 25–31.

Miyashita, Y. (1988), 'Neuronal correlate of visual associative long-term memory in the primate temporal cortex', *Nature* **335**(6193), 817–820.

Miyashita, Y. & Chang, H. S. (1988), 'Neuronal correlate of pictorial short-term memory in the primate temporal cortex', *Nature* **331**(6151), 68–70.

Mohan, K. & Freedman, D. (2018), Private Communication.

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

Mongillo, G., Curti, E., Romani, S. & Amit, D. J. (2005), 'Learning in realistic networks of spiking neurons and spike-driven plastic synapses', *European Journal of Neuroscience* **21**(11), 3143–3160.

Mongillo, G., Hansel, D. & van Vreeswijk, C. (2012), 'Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission', *Phys. Rev. Lett.* **108**, 158101.

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R. & Wang, X.-J. (2017), 'Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex', *Proceedings of the National Academy of Sciences* **114**(2), 394–399.

Murray, J. M. & Escola, G. S. (2017), 'Learning multiple variable-speed sequences in striatum via cortical tutoring', *Elife* **6**, e26084.

Nakamura, K. & Kubota, K. (1995*a*), 'Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task', *Journal of neurophysiology* **74**(1), 162–178.

Nakamura, K. & Kubota, K. (1995*b*), 'Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task', *J. Neurophysiol.* **74**(1), 162–178.

Naya, Y., Sakai, K. & Miyashita, Y. (1996), 'Activity of primate inferotemporal neurons related to a sought target in pair-association task', *Proc. Natl. Acad. Sci. U.S.A.* **93**(7), 2664–2669.

Ngezahayo, A., Schachner, M. & Artola, A. (2000), 'Synaptic activity modulates the induction of bidirectional synaptic changes in adult mouse hippocampus', *J. Neurosci.* **20**, 2451–2458.

O'Keefe, J. & Dostrovsky, J. (1971), 'The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat.', *Brain research* .

O'keefe, J. & Nadel, L. (1978), *The hippocampus as a cognitive map*, Oxford: Clarendon Press.

Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. (2015), 'Growth and splitting of neural sequences in songbird vocal development', *Nature* **528**(7582), 352–357.

Ostojic, S. (2014), 'Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons', *Nature neuroscience* **17**(4), 594–600.

Parisi, G. (1986), 'A memory which forgets', *Journal of Physics A: Mathematical and General* **19**(10), L617.

Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. (2008), 'Internally generated cell assembly sequences in the rat hippocampus', *Science* **321**(5894), 1322–1327.

Pereira, U. & Brunel, N. (2018*a*), 'Attractor dynamics in networks with learning rules inferred from in vivo data', *Neuron* **99**(1), 227–238.

Pereira, U. & Brunel, N. (2018*b*), 'Unsupervised learning of persistent and sequential activity'.

Pfister, J.-P. & Gerstner, W. (2006), 'Triplets of spikes in a model of spike timing-dependent plasticity', *Journal of Neuroscience* **26**(38), 9673–9682.

Rajan, K., Harvey, C. D. & Tank, D. W. (2016), 'Recurrent network models of sequence generation and memory', *Neuron* **90**, 1–15.

Rauch, A., Camera, G. L., Lüscher, H.-R., Senn, W. & Fusi, S. (2003), 'Neocortical pyramidal cells respond as integrate-and-fire neurons to *in vivo*-like input currents', *J. Neurophysiol.* **90**, 1598–1612.

Reinhold, K., Lien, A. D. & Scanziani, M. (2015), 'Distinct recurrent versus afferent dynamics in cortical visual processing', *Nat. Neurosci.* **18**, 1789–1797.

Renart, A., Song, P. & Wang, X.-J. (2003), 'Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks', *Neuron* **38**(3), 473–485.

Romani, S., Amit, D. J. & Amit, Y. (2008), 'Optimizing one-shot learning with binary synapses', *Neural computation* **20**(8), 1928–1950.

Romo, R., Brody, C. D., Hernández, A. & Lemus, L. (1999), 'Neuronal correlates of parametric working memory in the prefrontal cortex', *Nature* **399**, 470–474.

Roxin, A., Brunel, N., Hansel, D., Mongillo, G. & van Vreeswijk, C. (2011), 'On the distribution of firing rates in networks of cortical neurons', *The Journal of neuroscience* **31**(45), 16217–16226.

Royer, S. & Paré, D. (2003), 'Conservation of total synaptic weight through balanced synaptic depression and potentiation', *Nature* **422**(6931), 518–522.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1985), Learning internal representations by error propagation, Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sabatini, B. L., Oertner, T. G. & Svoboda, K. (2002), 'The life cycle of ca 2+ ions in dendritic spines', *Neuron* **33**(3), 439–452.

Sakai, K. & Miyashita, Y. (1991), 'Neural organization for the long-term memory of paired associates.', *Nature* **354**(6349), 152–155.

Schücker, J., Goedeke, S., Dahmen, D. & Helias, M. (2016), 'Functional methods for disordered neural networks', *arXiv preprint arXiv:1605.06758* .

Sejnowski, T. J. (1977), 'Storing covariance with nonlinearly interacting neurons', *Journal of mathematical biology* **4**(4), 303–321.

Semon, R. (1909), 'Die nmemischen empfindungen.[mnemic psychology]', *Leipzig: Wilhelm Engelmann* .

Senn, W., Markram, H. & Tsodyks, M. (2001), 'An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing', *Neural Comput* **13**, 35–67.

Seung, H. S. (1996), 'How the brain keeps the eyes still', *Proceedings of the National Academy of Sciences* **93**(23), 13339–13344.

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J. & Bodner, M. (2007), 'Variability in neuronal activity in primate cortex during working memory tasks', *Neuroscience* **146**(3), 1082–1108.

Sjöström, P. J., Turrigiano, G. G. & Nelson, S. B. (2001), 'Rate, timing, and cooperativity jointly determine cortical synaptic plasticity', *Neuron* **32**(6), 1149–1164.

Sompolinsky, H., Crisanti, A. & Sommers, H.-J. (1988), 'Chaos in random neural networks', *Physical Review Letters* **61**(3), 259.

Sompolinsky, H. & Kanter, I. (1986), 'Temporal association in asymmetric neural networks', *Physical review letters* **57**(22), 2861.

Sompolinsky, H. & Zippelius, A. (1982), 'Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses', *Physical Review B* **25**(11), 6860.

Sussillo, D. & Abbott, L. F. (2009), 'Generating coherent patterns of activity from chaotic neural networks', *Neuron* **63**(4), 544–557.

Theodoni, P., Rovira, B., Wang, Y. & Roxin, A. (2017), 'Theta-modulation drives the emergence of network-wide connectivity patterns underlying replay in aa model of hippocampal place cells'.

Thomson, A. M. & Lamy, C. (2007), 'Functional maps of neocortical local circuitry', *Frontiers in neuroscience* **1**, 2.

Tirozzi, B. & Tsodyks, M. (1991), 'Chaos in highly diluted neural networks', *EPL (Europhysics Letters)* **14**(8), 727.

Titley, H. K., Brunel, N. & Hansel, C. (2017), 'Toward a neurocentric view of learning', *Neuron* **95**(1), 19–32.

Toyoizumi, T., Kaneko, M., Stryker, M. P. & Miller, K. D. (2014), 'Modeling the dynamic interaction of hebbian and homeostatic plasticity', *Neuron* **84**(2), 497–510.

Treves, A. (1990*a*), 'Graded-response neurons and information encodings in autoassociative memories', *Physical Review A* **42**(4), 2418.

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

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

Treves, A. & Rolls, E. T. (1992), 'Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network', *Hippocampus* **2**(2), 189–199.

Tsodyks, M. (1988), 'Associative memory in asymmetric diluted network with low level of activity', *EPL (Europhysics Letters)* **7**(3), 203.

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

Turrigiano, G. G. (2017), 'The dialectic of hebb and homeostasis', *Phil. Trans. R. Soc. B* **372**(1715), 20160258.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. (1998), 'Activity-dependent scaling of quantal amplitude in neocortical neurons', *Nature* **391**(6670), 892.

Van Vreeswijk, C., Sompolinsky, H. et al. (1996), 'Chaos in neuronal networks with balanced excitatory and inhibitory activity', *Science* **274**(5293), 1724–1726.

Veliz-Cuba, A., Shouval, H. Z., Josić, K. & Kilpatrick, Z. P. (2015), 'Networks that learn the precise timing of event sequences', *Journal of computational neuroscience* **39**(3), 235–254.

Vogels, T., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. (2011), 'Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks', *Science* **334**(6062), 1569–1573.

Waddington, A., Appleby, P. A., De Kamps, M. & Cohen, N. (2012), 'Triphasic spike-timing-dependent plasticity organizes networks to produce robust sequences of neural activity', *Frontiers in computational neuroscience* **6**.

Wang, X.-J. (2001), 'Synaptic reverberation underlying mnemonic persistent activity', *Trends in neurosciences* **24**(8), 455–463.

Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K. & Stokes, M. G. (2018), 'Intrinsic neuronal dynamics predict distinct functional roles during working memory', *Nature communications* **9**(1), 3499.

Wilson, H. R. & Cowan, J. D. (1972), 'Excitatory and inhibitory interactions in localized populations of model neurons', *Biophysical journal* **12**(1), 1.

Woloszyn, L. & Sheinberg, D. L. (2012), 'Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex', *Neuron* **74**(1), 193–205.

Zenke, F., Agnes, E. J. & Gerstner, W. (2015), 'Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks', *Nature communications* **6**.

Zenke, F. & Gerstner, W. (2017), 'Hebbian plasticity requires compensatory processes on multiple timescales', *Phil. Trans. R. Soc. B* **372**(1715), 20160259.

Zenke, F., Gerstner, W. & Ganguli, S. (2017), 'The temporal paradox of hebbian learning and homeostatic plasticity', *Current Opinion in Neurobiology* **43**, 166–176.