Thèse n° 7805

A dynamical systems approach to synaptic consolidation and associations of concepts in hippocampus

Présentée le 21 mai 2021

Faculté des sciences de la vie
Laboratoire de calcul neuromimétique (SV/IC)
Programme doctoral en neurosciences

pour l'obtention du grade de Docteur ès Sciences

par

Chiara GASTALDI

Acceptée sur proposition du jury

Prof. J. Gräff, président du jury
Prof. W. Gerstner, directeur de thèse
Prof. N. Brunel, rapporteur
Prof. S. Romani, rapporteur
Prof. A. Mathis, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Your reason and your passion are the rudder
and the sails of your seafaring soul.
— Khalil Gibran

To my parents, Anna and Giorgio

Acknowledgements

First of all, I would like to thank my advisor, Wulfram Gerstner, for having supervised me during the Ph.D. Whether we were talking about work, or any other topics during the lunch breaks, exchanges with Wulfram have always been inspiring. He imprinted the LCN lab on the solid bases of personal freedom, intellectual rigor and mutual respect.
I am grateful to the jury experts, for their interest in my thesis: Sandro Romani, Nicolas Brunel and Alexander Mathis.
I also what to thank all the other people who have shared with me their path at the LCN. In particular, a big thanks goes to Tilo Schwalger, with whom I developed the largest part of my research. Thanks the other scientists I had the pleasure to collaborate with: Samuel Muscinelli, Emanuela De Falco and Rodrigo Quian Quiroga. Thanks to Johanni Brea for the precious scientific and philosophical discussions, besides having introduced me to Julialang. I am grateful to Vasikili Liakoni for being of great support during the final writing and to Sophia Becker for her valuable feedback. My thanks go to my colleagues of when I started my Ph.D., in particular, to Marco Lehmann for sharing the office, Florian Colombo for the priceless evenings sailing in Lac Leman, Adytia Gilra to be an inspirational example, Alexander Seeholzer, Dane Corneil, Mohammadjavad Faraji, Hesam Setareh. And thanks to current colleagues: Martin Barry, for the many Hopfield discussions and for the evenings at Totem, Alireza Modirshanechi, Valentin Schmutz, Guillaume Bellec, Bernd Illing, Christos Sourmpis, and Georgios Iatropoulos. A special thanks to Berfin Simsek for being 50% colleague and 100% friend. An affectionate thought goes to Noé Gallice. Finally, thanks to Rosa Ana Turielle for helping in all types of organization.
Thanks to my mentor Felix Schürmann, whose humanity I sincerely appreciated and to Stefano Fusi for giving me the chance to visit his lab.
I would like to thank all the wonderful friends with which I have shared the last five years, with whom I had the chance to discuss science and personal life. Among the many others who shared their time with me at the EPFL and outside, climbing, in the mountains, or at the Catalyst, I would like to mention a few. Violaine Bellée has been a wise friend since we arrived at the EPFL, an enthusiastic particle physicist and irreplaceable running partner. Madhu Raghunathan has psychologically supported me tirelessly and taught me about meditation. With Aiste Baleisyte and Loris Ercole I shared amazing outdoor days with equally amazing

conversations. I have shared a big part of our lives in Lausanne with Shriya Palchaudhuri and Edoardo Martino, in our beautiful flat and outside. Anna Pagani embodies my ecological awareness, I am grateful for our countless lunches at Esplanade. “Il vino buono sta nella botte piccola”, from neuroscience to spirituality, it is great to share with Silvana Pereira so much in a short time. Beer too is a good start for deep conversations, especially with Manu Condepudi. A warm thanks to Patrick Valeri and Caterina Bigoni, with whom I started the Ph.D. experience. Thanks to those who believed in the Geco project: Francesco Pennacchio, Evann Courdier, Sergio Hernandez, Benjamin Paccaud, Kelly Logan, Dhanesh Kothari, Stephan Kreuzberg, Gordon Dunn, Brigitte Burgisser and to my neighbor and friend Lea Ghisalberti in particular, who can bring a bright smile in all occasions.
Thanks to the friends who left Lausanne, but still feel so close: Chantal Gauthey, Emil Rotilio, Setayesh Radkani, Carlos Florensa and in particular Fiona Kirk, for the incredible mix of science, philosophy and hiking we have shared. And to Matteo Gambara, I am lucky to have such a genuine friend.
A heartfelt thanks goes to Enrico Manenti, who knows me better than anybody, and has supported me for years.
I would like to thank the friends from Brescia, Cecilia Musicco, for always being there, Luca Tonidandel and the amazing classe 5^aA of Liceo Calini: after over ten years, only a pandemic could stop our Christmas dinners.
I shared with Luca Binci the oddities and the joys of 2020. He has been my constant company during confinements and has supported me during the difficulties of the last year.
I am grateful to my parents Anna and Giorgio, for never doubting my study and career choices and believing in me more than I dared to. Their patient support has been my safe harbour during the difficult times. I am also thankful to my brother Alberto, for his passionate interest in my personal projects. Finally, thanks to my grandpa Angelo, for being my biggest fan.

Lausanne, April 27, 2021 Chiara Gastaldi

Abstract

In the present work, we approach two key aspects of memory formation: associative memory and synaptic consolidation.
The storage of associative memory is commonly related to the medial temporal lobe in humans. Experimental evidence shows that the memories of objects, people or places are represented in this brain area by cell assemblies that respond selectively to single concepts. Neurons forming such assemblies are called *concept cells*. Associations between different concepts are linked to concept cells shared between assemblies: we refer to the number of shared neurons as the *overlap* between memory engrams. The respective assemblies of two associated concepts (e.g. Hillary and Bill Clinton) share more neurons compared to the assemblies of two unrelated concepts (e.g. Hillary Clinton and the Eiffel tower). In particular, three characteristics of assemblies of concept cells are important for this work: (*a*) they exhibit a very low mean activity (about 0.2% of neurons respond to each concept), and (*b*) assemblies relative to strongly associated concepts share about 4% of their cells, (*c*) assemblies relative to non-associated concepts share less then 1% of their cells. This implies that the association between two concepts induces a higher level of overlap between the relative memory engrams.
In parallel, theoretical studies have shown that overlaps between memory engrams are fundamental in the process of free recall of sequences of words. These models assume that memory engrams have such a high mean activity that all assemblies are overlapping to a certain extent.
Associative memory is traditionally modeled through attractor neural networks. Memory engrams are represented by binary patterns of active/silent neurons. The engrams and their eventual overlaps are encoded into the synaptic weights structure. While there is extensive literature on independent low-activity patterns, only a few studies can be found on correlated patterns. Extending the existing theory to include correlation is a key missing point to answer questions such as: How can shared concept cells encode associations? Why are 4% of neurons shared and not more?
Using a mean-field approximation, we derive analytic equations for the network dynamics in the case of correlated patterns. Our results provide a theoretical framework that can explain the experimentally observed value of shared neurons. We find that for concepts represented by realistically sparse neural assemblies there are a minimal and a maximal fraction of shared neurons so that associations can be reliably coded. In the presence of a periodically modulated

signal, such as hippocampal oscillations, chains of associations can be recalled analogously to theories of free recall of lists of memorized words.
Finally, we compare the predicted number of concepts a neuron responds to with experimental data. We test different ways of constructing correlated patterns and confirmed the common opinion that information in the hippocampus is non-hierarchically organised.
In the second part of the thesis, we focus on one of the processes involved in learning new stimuli: synaptic consolidation. We propose a model based on two coupled dynamical variables: the fast synaptic weights and a slow internal synaptic mechanism. In classical experiments, the consolidation of the synapse is related to the stimulation frequency and number of repetitions. We show that it is exactly the time scale separation between the dynamics of the two variables that determines which combination of stimulation amplitude and frequency are suitable to elicit synaptic consolidation.

Keywords: concept cells, associations, attractor networks, dynamical systems, synaptic consolidation.

Compendio

Nel presente lavoro affronto due aspetti distinti della formazione della memoria: la memoria associativa e il consolidamento sinaptico.
Nell'uomo l'immagazzinamento della memoria associativa è comunemente collegato al lobo temporale medio. Evidenze sperimentali mostrano che i ricordi di oggetti, persone o luoghi sono rappresentati in quest'area del cervello da gruppi di cellule che rispondono selettivamente a singoli concetti. I neuroni che formano tali gruppi sono chiamati *cellule-concetto*.
L'associazione tra concetti diversi implica la condivisione di cellule-concetto tra gruppi di neuroni: il numero di neuroni condivisi tra le tracce di memoria è detto *sovrapposizione*. I gruppi di cellule relativi a due concetti associati (ad esempio, Hillary e Bill Clinton) condividono più neuroni rispetto a gruppi relativi a due concetti non associati (ad esempio Hillary Clinton e la Torre Eiffel).
In particolare, sono tre le caratteristiche dei gruppi di cellule-concetto rilevanti ai fini di questo lavoro: i gruppi sovrapposti (*a*) hanno un'attività media molto bassa (circa lo $0,2\%$ dei neuroni risponde a ciascun concetto), (*b*) i gruppi relativi a concetti associati condividono circa del circa 4% delle loro cellule tra loro e i gruppi relativi a concetti non associati (*c*) condividono meno del 1% delle loro cellule.
Parallelamente, studi teorici hanno dimostrato che le sovrapposizioni tra le tracce di memoria sono fondamentali nel processo di libero richiamo di sequenze di parole. Tali modelli teorici ipotizzano un'attività media dei gruppi così elevata che tutti i gruppi di concetti si sovrappongono in una certa misura.
La memoria associativa è tradizionalmente modellizata attraverso reti neurali attrattive. Le tracce della memoria e le loro eventuali sovrapposizioni sono codificate nella struttura dei pesi sinaptici. Le tracce della memoria sono rappresentate da patterns binari di neuroni, attivi o silenti. Sebbene esista una letteratura molto ampia su patterns indipendenti e a bassa attività, esistono pochi studi riguardanti patterns correlati. Per rispondere a domande come: in che modo le cellule concetto condivise possono codificare l'associazione? Perché il 4% dei neuroni è condiviso e non di più? È necassario estendere la teoria esistente per includere la correlazionetra patterns.
Usando un'approssimazione di campo medio, deriviamo equazioni analitiche che descrivono la dinamicha della rete neurale nel caso di patterns correlati. I nostri risultati forniscono un

quadro teorico che può spiegare il numero di neuroni condivisi osservato sperimentalmente. Mostriamo che per i concetti rappresentati da gruppi di neuroni realisticamente sparsi ci sono una frazione minima e una massima di neuroni condivisi, entro le quali le associazioni possano essere codificate in modo affidabile. In presenza di un segnale modulato periodicamente, simile alle oscillazioni ippocampali, catene di associazioni possono essere ricordate analogamente agli studi teorici del richiamo libero di di parole memorizzate.
Infine, confrontiamo il numero previsto di concetti a cui un neurone risponde con i dati sperimentali. Avendo testato diversi modi di costruire patterns correlati, confermiamo l'opinione comune che le informazioni nell'ippocampo non siano organizzate gerarchicamente.
Nella seconda parte della tesi, ci concentriamo su uno dei processi coinvolti nell'apprendimento di nuovi stimoli: il consolidamento sinaptico. Proponiamo un modello di consolidamento sinaptico basato su due variabili dinamiche accoppiate: i veloci pesi sinaptici e un lento meccanismo sinaptico interno. Negli esperimenti classici, il consolidamento della sinapsi è correlato alla frequenza di stimolazione e al numero di ripetizioni. Dimostriamo che è esattamente la differenza tra le velocità di evoluzione delle due variabili che determina le combinazioni di ampiezza e frequenza di stimolazione adatte a suscitare il consolidamento sinaptico.

Parole chiave: cellule-concetto, associazioni, reti di attrattori, sistemi dinamici, consolidamento sinaptico.

Contents

Acknowledgements **i**

Abstract (English/Italiano) **iii**

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**
1.1 Associative and episodic memory 1
1.1.1 The medial temporal lobe 2
1.1.2 Encoding concepts in the human medial temporal lobe 3
1.2 Computational models 5
1.2.1 Attractor neural networks 5
1.2.2 Models of synaptic consolidation 6
1.3 Thesis contribution 7

2 When shared concept cells support associations: theory of overlapping memory engrams **9**
2.1 Paper information 9
2.2 Introduction 11
2.3 Results 14
2.3.1 Critical fraction of shared neurons between memory engrams 15
2.3.2 Association chains 19
2.3.3 Comparison with experimental data 22
2.3.4 Robustness to heterogeneity 24
2.4 Discussion 24
2.4.1 Bio-plausibility and parameter choice 27
2.5 Methods 28
2.5.1 Model without adaptation and global feedback 29
2.5.2 Model with adaptation and global inhibitory feedback 29
2.5.3 Mean-field equations for two overlapping patterns 30

2.5.4 Excluding self-interaction . . . 31
2.5.5 Correlation between background patterns . . . 31
2.5.6 Algorithms to generate correlated patterns . . . 31
2.5.7 Experimental data . . . 32
2.5.8 Heterogeneous F-I curves . . . 33
2.6 Supplementary information . . . 33
2.6.1 Critical fraction of shared neurons between memory engrams . . . 34
2.6.2 Association chains . . . 34
2.6.3 Parameters choice . . . 39
2.6.4 Model without adaptation and global feedback . . . 40
2.6.5 Model with adaptation and global inhibitory feedback . . . 41
2.6.6 Non-dimensionalization of the model . . . 42
2.6.7 Review of attractor theory . . . 43
2.6.8 Dynamical mean-field equations . . . 46
2.6.9 Mean-field equations for two correlated patterns . . . 46
2.6.10 Excluding self-interaction . . . 49
2.6.11 Correlation between background patterns . . . 50
2.6.12 Mean-field dynamics in the presence of adaptation and global feedback 53
2.6.13 Stability of the fixed points . . . 54
2.6.14 The limit case $b \to \infty$: when the gain function is an Heaviside . . . 55
2.6.15 Robustness to heterogeneity . . . 57
2.6.16 Relation to experiments: algorithms to generate correlated patterns . . . 59
2.6.17 Comparing algorithm predictions with experimental data . . . 62
2.6.18 Comparing virtual experiments and expected distributions . . . 63
2.6.19 Numerical solutions . . . 64

3 Optimal Stimulation Protocol in a Bistable Synaptic Consolidation Model 69
3.1 Paper information . . . 69
3.2 Introduction . . . 70
3.3 Methods . . . 71
3.3.1 Choice of the model . . . 73
3.3.2 Simplification steps of the 2d-dynamics . . . 74
3.3.3 Nullclines and phase-plane analysis . . . 75
3.3.4 Symmetric changes of coupling coefficients reveal two bifurcations . . . 76
3.3.5 Asymmetric parameter choices shape the basins of attraction . . . 77
3.3.6 Numerical simulations . . . 80
3.4 Results . . . 80
3.4.1 Abstraction of the stimulation protocol . . . 81
3.4.2 One episode . . . 82
3.4.3 Repeated episodes . . . 84

3.5 Discussion 87
3.6 Supplementary information 91

4 Conclusions **95**
4.1 Summary of the main contributions 95
4.1.1 Concept association in the human MTL 95
4.1.2 Synaptic consolidation 96
4.2 Possible future developments 97
4.2.1 Synaptic plasticity 97
4.2.2 Increasing the bio-plausibility of the model 98
4.2.3 Large scale network simulation with diluted weights 99
4.2.4 Several time-scales 100

Bibliography **101**

Curriculum Vitae **111**

List of Figures

1.1 The human hippocampus and its internal structure. 2

2.1 *Caption on the next page* 13
2.1 Overlapping concepts can be retrieved separately and jointly. 14
2.2 Critical fraction $c_{\max}$ of shared neurons depends on the neuronal f-I curve but not on the memory load. 18
2.3 Chain of associations requires shared concepts cells. 20
2.4 Dependence of association chains on sparsity and neuronal parameters. 22
2.5 Comparison between theory prediction and experimental data on the number of concept a neuron responds to 23
2.6 The model is robust to heterogeneity of F-I curves. 25
2.7 The critical correlation is extracted from the bifurcation diagram 35
2.8 Evolution of $m^1(t)$ and $m^2(t)$ according to the mean-field dynamics for super-critical correlation. 36
2.9 Estimation of the correlation range in which retrieval of a chain of concepts is possible. 37
2.10 *Caption on the next page* 38
2.10 Dynamical mean-field solutions for m^1 and m^2 in the case of two dependent or independent patterns. 39
2.11 Retrieval dynamics in the presence of adaptation according to the mean-field equations 39
2.12 Extracting the critical correlation using parameters from (Pereira and Brunel, 2018) 40
2.13 Comparison between model prediction and data of the probability of finding a neuron responding to a given number of concepts 65

3.1 Schematic experimental setup and modeling framework. 72
3.2 Phase-plane diagram and basins of attractions for the symmetric case with equal coupling constants, $C_w = C_z = C$ 77
3.3 Bifurcations diagrams. 78
3.4 Asymmetric parameter choices. 79

3.5 Potentiation during a single episode. 83
3.6 Potentiation with repeated episodes. 85
3.7 Stimulation effort needed to achieve potentiation for $\tau_z/\tau_w = 7$ and $t_{\text{on}} = 0.01\tau_w$ 86
3.8 Synaptic weight as a function of time 92
3.9 Low-pass-filtered version of the synaptic weight as a function of time with . . . 93

List of Tables

2.1 Table of parameters, part1. 67
2.2 Table of parameters, part2. 68

3.1 Table of parameters 88

1 Introduction

1.1 Associative and episodic memory

Human memory includes a wide range of cognitive processes: we are able to memorize the map of a familiar environment or a repeated movement, so that we automatize quite complex tasks such as navigating in a complex environment, playing a piece of music on an instrument, or driving a car. However, when we think about memory, the first thing that comes to our mind is the ability of memorizing and recognising familiar people, places, or objects. It is common to use the term *concept* as a general term to indicate either a specific person or object or place (Quiroga et al., 2005). The memorization of concepts involves different brain areas that are specialized in specific sub-tasks (Shallice et al., 1994; Tulving, 2002; Cabeza et al., 2002). In this work, we are interested in two aspect of the memory of concepts: first the ability of remembering specific events, such as a birthday party or the moment of meeting a new person. This fundamental skill is usually referred to as *episodic memory* (Tulving, 1993). Secondly, the ability to learn and remember an association between two or more concepts (e.g. that certain person lives in that certain street). This ability is called *associative memory* (Anderson and Bower, 2014) and it is strongly related to episodic memories since the associations are learnt on the base of a specific episode and vice-versa episodic memory works with associations (Shing et al., 2010). For example, if we meet a new person in a specific environment we will tend to associate the person and the place where we met him/her: "meeting person X at place Y" is a single event. The functioning of episodic memory is also based on the overall knowledge of the context (Hemmer and Steyvers, 2009; Mizumori, 2013). Let's consider the following example: if we go to a new pizzeria and we get squared pizzas, we will easily remember that pizzeria as "the one that makes squared pizzas". The specific shape of the pizza is not an event itself, however the association between pizza and a round shape is already encoded in our memory as the sum of our past experiences in eating pizza. The new restaurant breaks our expectation of getting a round pizza and this constitutes the event which we easily remember. The two

examples, despite their triviality, exemplify the complexity of episodic memory, since forming episode-based associations requires the integration of pre-processed sensory information with emotional components (Allen et al., 2008; Murray et al., 2013). The emotional component of an event such as attention, surprise, fear or excitement plays a fundamental role in selecting the events worth remembering.

1.1.1 The medial temporal lobe

The Medial Temporal Lobe (MTL) is the area of the brain which includes the hippocampus and the para-hippocampal cortex and it is classically believed to be responsible for the formation of both new episodic memories (Rolls, 2018) and associative memory. Placed under the cerebral cortex, the hippocampus is part of the limbic system. The main components of the complex hippocampal structure are the dentate gyrus, and areas CA1 and CA3 (Fig. 1.1).

The hippocampus primarily receives inputs primarily from the entorhinal cortex into the dentate gyrus. The dentate gyrus projects into the CA3 area, which is characterised by the lack of internal topological organisation and by a dense internal recurrent connectivity. There is evidence that CA3 behaves as an auto-association (or attractor) network, meaning that it can reconstruct the full memory starting from a partial cue (Rolls, 2018; Rennó-Costa et al., 2014; Wills et al., 2005). CA3 projects into CA1 which projects back to the entorhinal cortex (Knierim, 2015; Hasselmo, 1999). However, the hippocampus also receives inputs from other areas related to the processing of emotions, the amygdala in particular (Felix-Ortiz and Tye, 2014), which projects into the ventral hippocampus, making it a particularly suited brain area for high-level integration of different information. Moreover, there is evidence of the role of the hippocampus in emotion regulations (Frühholz et al., 2014; Jacobson and Sapolsky, 1991; Koelsch et al., 2015).

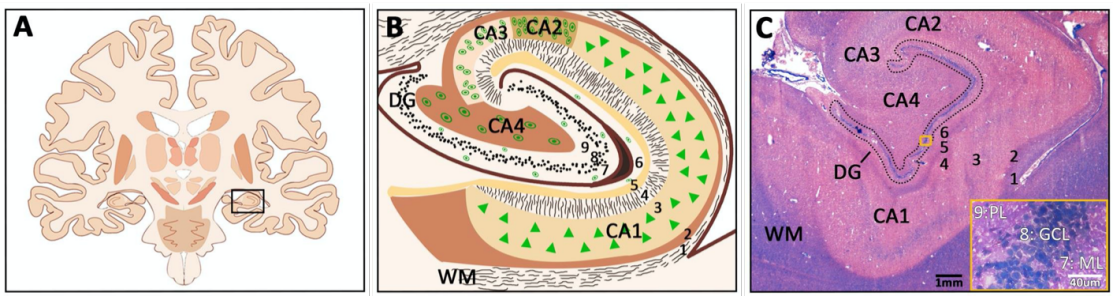


Figure 1.1 – The human hippocampus and its internal structure; A) Position of the hippocampus inside of the brain. B) Schematics of the hippocampal internal structure. C) A coronal human hippocampus section. Figure arranged from (Mendis, 2016).

In Chapter 2 we are mainly going to focus on the hippocampus. However, it is important to

underline that the hippocampus is not the only brain area involved in storing the memory of concepts. On the contrary, such a task involves several brain areas, engaging a quite large portion of the brain. While the hippocampus is fast in encoding new memories (Ison et al., 2015), it is believed that strong and long-lasting memories are only possible through the interaction between hippocampus and cortex (Eichenbaum et al., 1994). In particular, the hippocampus is necessary for the formation of new memories, while old memories are stored and recalled in cortical regions. This insight emerged from lesion studies such as the very famous case of the patient H.M., who got his hippocampus removed in a surgical operation in 1957, as an attempt to cure his severe epilepsy. While he was able to recall some long term memories acquired before the surgery, he was unable to form new ones, suggesting that the hippocampus is necessary for the formation of new memories, but old ones are stored in cortical regions.

There is evidence that the interaction between the hippocampus and the pre-frontal cortex (PFC) is essential for high-level behavior and memory processes (Miller et al., 2002; Zeithamova and Preston, 2010; Lavenex and Amaral, 2000). And it is well known that the inferotemporal (IT) cortex plays a fundamental role in face recognition and concept memorization and it projects directly to the MTL (Lavenex and Amaral, 2000).

In conclusion, the picture that emerges from hippocampal studies is that new memories of concepts, association between concepts or episodes, are first created in the hippocampus. The lack of hierarchical organisation in CA3 is believed to favor the fast creation of associations based on events. In fact, its lack of internal structure makes it equally easy to create an association between any pair of concepts. Later, only the memories and the associations that are relevant over a longer time get transferred to the cortex (Squire et al., 2015). On the contrary of CA3, the cortex is structured hierarchically, and here concepts are believed to be stored in an organised way (engrams of related concepts are physically close to each other).

1.1.2 Encoding concepts in the human medial temporal lobe

A long line of experimental studies has partially shined light on the fundamental role of the human medial temporal lobe (MTL) in storing episodic memories (Quiroga et al., 2005; Quiroga, 2012). Such experiments were conducted on patients with pharmacologically intractable epilepsy, to whom electrodes were implanted in order to find the seizure focus. Each electrode is able to record activity of individual neurons around its tip. Patients are awake during the operation and they are presented with pictures of familiar people, landmarks, celebrities and common animals. During these experiments, scientists have found neurons that respond selectively to specific concepts.

For this reasons, such neurons are called *concept cells*. Their response is invariant to the

specific visual field, for example, a “Sidney opera house” neuron would responds equally to an image of the famous opera house and the word “Sidney opera” written down. Each memory item is represented by a whole assembly of concept cells (Reddy and Thorpe, 2014), and many neurons are activated during the recall of each concept.

In a very recent study (Rey et al., 2020b), the response of concept cells was tested against different sensory stimuli, corresponding to the same concepts: the picture, the written word and the spoken word. It was found that the concept representation in the MTL is invariant within the same type of representation and largely invariant across representations. For example, different pictures of the same person elicit very similar neural responses (80% of overlap between the activated neurons). Analogously, a picture and the written name or a picture and the spoken name of the same person elicit similar neural responses (about 40% of overlap between the activated neurons in the case of picture and written word and about 50% in the case of picture and spoken word), thus concluding that representation of concepts in the human MTL is largely sensory invariant.

Such a high-level representation of concepts is typically human, and it does not apply to rodents, whose hippocampus seems to be more specialized in spatial representation, while the monkey hippocampus presents a mixed representation of space and concepts (Quiroga, 2019a).

More recent studies have provided evidence that, not only the concepts themselves, but also the associations between concepts are stored in the MTL. In (De Falco et al., 2016) the authors show that cell assemblies encoding associated concepts (such as Hillary and Bill Clinton) share more concept-cells than non associated cell assemblies, ans they proposed the idea that shared concept-cell support association between concepts. If that’s the case, it should be possible to increase the selectivity of a neuron (i.e. the number of concepts it responds to) by artificially creating the association between two concepts during the experiment. Let’s consider an Hillary Clinton concept-cell, we say that “Hillary Clinton” is the *preferred stimulus* for that cell. All other stimuli to which the cell does not respond to are called *non-preferred stimuli*. If an association is created between Hillary Clinton and a a non-preferred stimulus, let’s say the Eiffel tower, then we expect some of the Hillary-Clinton concept-cells to start responding to the Eiffel tower as well. The increase of neuron’s selectivity was indeed observed. In a study (Ison et al., 2015), after identifying the preferred stimulus of a concept cell (for example the picture of a celebrity), the scientists present to the patient a combined stimulus in which the preferred is presented together with a non-preferred stimulus (such as a famous landmark) in a single picture multiple times. Over the presentations, the recorded neuron acquires selectivity to the non preferred stimuli, confirming the idea that association between concepts is encoded in shared concept cells.

1.2 Computational models

In the present thesis, we approach two distinct aspects of memory formation: associative memory and synaptic consolidation. In the following, we introduce the two topics and how they can be approached using a dynamical system methodology.

1.2.1 Attractor neural networks

When modelling associative memory, Hopfield type attractor neural networks rule the field. While attractor neural networks are commonly refer to as "associative" memory network, they are more precisely *auto*-associative networks. Auto-associative networks can reconstruct the complete memory starting from a small cue. For example, let's suppose that the concept of "apple" is stored as a memory, and the network is given a picture of half an apple, or a picture of an apple partially covered by other objects: an auto-associative network is able to retrieve the complete memory of an apple. While there are in principle some differences between auto-associative and associative memory, in the "Thesis contribution" session we will make clear why attractor networks are the natural candidate to describe associative memory in the MTL too.

In attractor neural networks, each stored concept is represented by a binary memory pattern. Patterns represent memory engrams in the model. Going back to the previous example, let's suppose the concept of an apple is stored in the network, if neuron i belongs to the apple-cell-assembly than the apple-pattern is be 1 in position i and the other way around, if neuron j does not belong to the apple-cell-assembly than the the apple-pattern is 0 in position j.

Each pattern is an attractor of the network dynamics: if the network state correlates enough with one of the stored memory patterns, it converges to it, until the memory is fully retrieved. The *similarity measures*, also called *overlaps* in literature, measure the correlation between the network state and the stored patterns. Let's suppose the initial network state is a partially hidden apple, then in this state the similarity measure between the network state and the apple-memory-pattern will take some value between 0 and 1. During the process of retrieval of the full apple-concept, the similarity measure with the apple-pattern will gradually increase, until reaching the value of 1 corresponding to a complete retrieve of the concept. The ensemble of similarity measures to all stored patterns constitutes a system of macroscopic variables and their dynamical equations fully describe the network dynamics. The stable solutions of the similarity measures' dynamics constitute the stable states of an attractor network typically correspond to

1. a rest state, in which all neurons are inactive and no memory is recalled;
2. single retrieving states, in which only the neurons belonging to a specific memory are

activated. In this case, the network state strongly correlates only with the one pattern and we say that the memory has been retrieved;

3. one of the many *spurious states* which correspond to some combination of a few patterns being partially activated together.

Attractor neural networks are inspired by spin glasses theory, they initially shared the same statistical physics approach (Hopfield, 1982, 1984; Amit et al., 1985b,a; Mézard et al., 1987). The stable state solutions to the system are originally found in the very large network limit with the replica theory Amit et al. (1985b,a) or, alternatively, using the cavity method, also known as Thouless-Anderson-Palmer (TAP) equations (Mézard, 2017). Both methods rely on the hypothesis that the connection weights are symmetrical. However, the same solution can be found with a dynamical system approach (Shiino and Fukai, 1992). The latter method does not require the matrix of synaptic weights to be symmetric, opening the way to more bio-plausible asymmetric connections. Furthermore, it provides an interesting dynamical interpretation of the system. In fact, it allows to write exact dynamical mean-field equations for the similarity measures.

1.2.2 Models of synaptic consolidation

With synaptic consolidation, we indicate the mechanism that makes the effect of long term potentiation (LTP) persistent on the time scale of hours. Such mechanism acts locally, and it should not be confused with memory consolidation, which acts on the time scales of days and involves multiple brain areas.

In synaptic consolidation experiments, extracellular stimulation is given to both pre- and post-synaptic fibers. It is well known that, in synaptic pairing stimulation experiments, changes in synaptic efficacies are induced by the simultaneous activation of pre- and post-synaptic neurons. The persistence of such synaptic changes can be either short (in the order of tens of minutes) or long (in the order of several hours). When the changes are long-lasting, the synapse is said to be consolidated. Experiments show that, whether the synapse gets consolidated or not, strongly depends on the frequency of the stimulation protocol.

Synaptic consolidation models range from dynamical ones (Benna and Fusi, 2016) to probabilistic ones (Fusi et al., 2005), but they have one key ingredient in common: each model is the combination of several processes that act on different time scales. Indeed the process of synaptic consolidation involves a cascade of biochemical processes, characterised by different speeds and it can only be captured by models that take such heterogeneity into account.

1.3 Thesis contribution

In this work, I present two studies that tackle episodic and associative memory and synaptic consolidation using a similar mathematical approach.

The first study (presented in Chapter 2) targets directly the question of how the MTL can reliably encode the association between memorised concepts. We propose a model that provides a possible explanation of how memory association is stored and recalled in the human MTL. The model is based on a recurrent neural network and it is particularly suitable to understand the area CA3 of the hippocampus, since it is a highly recurrent network (each neuron receives inputs from about 30'000 others (Andersen et al., 2006)) and it has been shown to have an auto-associative or attractor behavior (Rolls, 2007). Most theoretical models for associative memory only consider independent memory engrams with a relative high fraction of active neurons, while experimental studies in the MTL measured

1. a low fraction of active neurons of about 0.2% (Ison et al., 2015) and
2. a fraction of shared neurons between associated engrams significantly above chance.

This makes the introduction of overlapping low-activity memory patterns a key missing piece of the theory. Using a mean-field type of approximation, we derive analytic equations for the network dynamics in the case of overlapping patterns in order to answer the following questions: How do shared neurons encode associations? How many neurons can be shared?

We show that, for memory engrams realistically sparse, there is a minimal and a maximal critical fraction of shared neurons such that associations are reliably encoded. If the the fraction of shared neurons is below the minimal critical value, the paradigm of free recall of words can not be reproduced. Whereas, if the fraction of shared neurons is bigger than the maximal critical value, the associated concepts can not be recalled individually. Moreover, overlapping memory engrams can be constructed in several ways, for example we can start from the hypothesis that associations have a hierarchical structure or not. To determine which construction is the most realistic, we compare predicted number of concepts a neuron respond to with data (De Falco et al., 2016). The non-hierarchical construction for overlapping engrams is the best fit to the experimental data.

If only two patterns are stored it is possible to fully describe the system with the phase-plane analysis of the similarity measure relative to the two stored patterns. This is a very convenient tool to identify the fixed points of the systems, corresponding to stable or unstable states. The stable states have a physical interpretation and they correspond to the rest state, the two single retrieving state and one spurious state, in which both memories are partially recalled. Starting from the phase-plane, it is possible to track the fixed points as a function of the correlation

between the two stored patterns. At some critical values of correlation the retrieval state might appear or disappear: there is a phase transition. This is equivalent to construct bifurcation diagrams (each bifurcation corresponding to a phase transition) to identify the correlation regime that corresponds to biologically plausible physical phases.

In the second study (presented in Chapter 3), we move to a more local aspect of memory formation: synaptic consolidation. Synaptic consolidation can be observed in both hippocampus and cortex. Our work is inspired by in vitro experiments in the hippocampus and the cortex of mice. It is worth noticing, however, that synaptic consolidation has been especially studied in the hippocampus where a particular role is played by acetylcholine (Hasselmo, 1999), a neurotransmitter that alters the available pathways between Hippocampal sub-areas. We proposed a mathematical model for synaptic consolidation. Experiments to obtain synaptic consolidation in vitro show that repeated stimulation is key to a long-lasting synaptic potentiation, however, why is that so? Why a repeated stimulus is preferred to a prolonged one and what is the role of the stimulation frequency? We propose a phenomenological model that clarifies the role of the stimulation protocol in the slice experiments. We look at the direct consequences of time scale separation. We simplify the system down to two coupled dynamical variables that symbolize a fast and a slow consolidation mechanism. The system is then studied with phase-plane analysis and we show that it can operate in different regimes depending on the stimulation parameters, in particular the amplitude of the stimulation and the repetition frequency.

2 When shared concept cells support associations: theory of overlapping memory engrams

2.1 Paper information

The research and the paper writing was developed by Chiara Gastaldi*[1], Tilo Schwalger[2], Emanuela De Falco[3], Rodrigo Q. Quiroga[4] and Wulfram Gerstner[1].

At the time of submission, authors are affiliated to the following institutions [1]School of Computer and Communication Sciences and School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL, Switzerland
[2]Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany
[3] School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne EPFL, Switzerland
[4]Centre for Systems Neuroscience, University of Leicester, LE1 7HA Leicester, UK

We intend to submit the present paper as soon as possible to Nature Communications.

In this chapter we propose a preprint version of the paper, i.e. the version of the manuscript which will soon be submitted to the journal (before peer-reviewing).

The code necessary to solve the mean-field equation and produce all figures will be published on GitHub by CG, following the publication of the paper.

Abstract

Memory engrams have been suggested to be the neural basis for encoding acquired concepts into assemblies of neurons. In Human Medial Temporal Lobe, neurons forming such assemblies are called concept cells. Neurons that are shared between two assemblies of concept cells

have been hypothesized to encode the association between the two concepts. Here we test this hypothesis in a computational model of attractor neural networks. We find that for concepts encoded in realistically sparse neural assemblies there is a minimal fraction $c_{\min}$ of neurons shared between assemblies below which associations can't be reliably coded; and a maximal fraction $c_{\max}$ of shared neurons above which single concepts can no longer be retrieved. In the presence of a periodically modulated background signal, such as hippocampal oscillations, chains of associations can be recalled in a way similar to theories of free recall of lists of memorized words. Model predictions on the number of concepts a neuron responds to are compared with experimental data.

Keywords:

Concept cells | Association | Attractor Neural Network | Hippocampus

Author contributions

CG, TS, WG and RQQ contributed to the conception of the study. CG implemented the code needed to produce all figures. CG and TS contributed to the formulation of the mean-field theory. TS and CG contributed to the non-dimentionalization of the model. EDF and RQQ provided the experimental data. EDF and CG contributed to data analysis. WG, CG contributed to develop different algorithms to fit the experimental data. CG wrote the first draft of the manuscript. All authors contributed to correcting and improving the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgements

We thanks Valentin Marc Schmutz, Martin Barry, Alireza Modirshanechi and Johanni Brea for useful comments and discussions.

Funding

This research was supported by the Swiss National Science Foundation, grant agreement 200020_184615 and by the European Union Horizon 2020 Framework Program under grant agreement no. 785907 (HumanBrain Project, SGA2).

2.2 Introduction

Human memory exploits associations between concepts. If you visited a famous place with a friend, a postcard of that place will remind you of him or her. The episode "with my friend at this place" has given rise to an association between two existing concepts: before the trip (the episodic event), you already knew your friend (first concept) and had seen a photograph of the place (second concept), but only after the trip, you associate these two concepts.

Concepts are encoded in the human Medial Temporal Lobe (MTL) by neurons, called "concept cells", that respond selectively and invariantly to stimuli representing a specific person or a specific place (Quiroga et al., 2005; Ison and Quiroga, 2008; Quiroga, 2019a). Each concept is thought to be represented by an assembly of concept cells that increases their firing rates simultaneously. The fraction γ of neurons in the human MTL which is involved in the representation of each concept is estimated to be be $\gamma \sim 0.23\%$ (Waydo et al., 2006). Under the assumption that each memory item is represented by the activation of a fixed, but random, subset of active neurons, a single concept is expected to activate γN neurons and two arbitrary concepts are expected to share $\gamma^2 N$ units, where N is the total number of neurons in the relevant brain areas.

Experimental studies have shown that single neurons can became responsive to new concepts while learning pairs of associations (Ison et al., 2015). Moreover, it has been estimated that assemblies representing two arbitrary concepts share less than 1% of neurons, whereas assemblies representing previously associated concepts share up to $4-5\%$ of neurons (De Falco et al., 2016) indicating that an increased fraction of shared neurons supports the association between concepts (De Falco et al., 2016; Rey et al., 2018, 2020a).

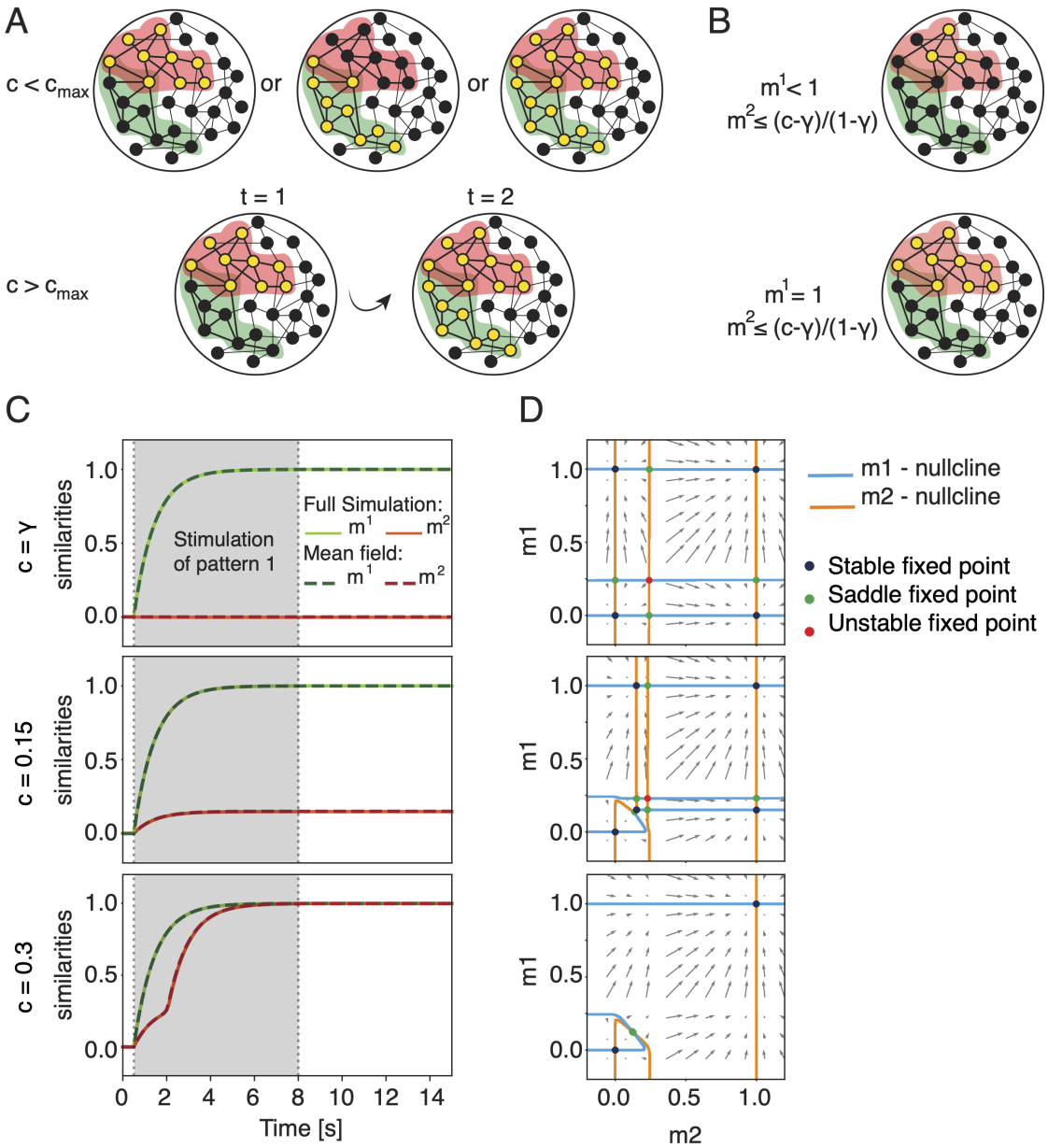
Intuitively, with the presence of shared neurons, the activation of a first assembly (e.g., a place) makes it more likely to activate a second assembly (e.g., a person). However, this interpretation poses several theoretical questions.

First, for the brain to function correctly as a memory network, it must remain possible to recall the two associated concepts separately (e.g. place without your friend), and not automatically the two together. However, if the concepts share too many neurons it becomes likely that the two memory items can no longer be distinguished, but are merged into a single, broader concept with a larger number of active neurons. We therefore ask as first question: what is the maximally allowed fraction $c_{\max}$ of shared neurons between two assemblies before the possibility of separate memory recall breaks down? Shared concept cells can be visualised as an overlap between two memory engrams. Below the maximal fraction $c_{\max}$ of shared neurons, each of the associated patterns can be recalled as a separate memory pattern, as schematically illustrated in Fig. 2.1A.

Second, instead of a static recall of one or the other concept, or of the two associated concepts together, we are also able to to organize and recall temporal chain of memory items. There is still no clear evidence on whether human MTL cells encode memories using sequential coding or rather pattern co-activations(Quiroga, 2020). Nonetheless, associative chains are directly linked to the psychophysics of free memory recall task (Romani et al., 2013; Recanatesi et al., 2015, 2017; Naim et al., 2020). In this context, we ask a second question: if each concept is represented by a small fraction γ of active neurons, is there a minimal fraction of shared neurons necessary to enable a reliable activation of a memory chain across the associated concepts?

Associative memory in recurrent networks, such as the area CA3 of the Hippocampus, has been modeled with attractor neural network (Hopfield, 1982; Weisbuch and Fogelman-Soulié, 1985; Amit and Amit, 1992; Kanter and Sompolinsky, 1987; Tsodyks and Feigel'man, 1988) where each memory item is encoded as a memory engram (Tonegawa et al., 2015; Josselyn and Tonegawa, 2020) in a fixed random subset of neurons (called 'pattern' in the theoretical literature (Tsodyks and Feigel'man, 1988)) such that no pattern has an overlap above chance with another one. Animal studies provide evidence of attractor dynamics in area CA3 (Rennó-Costa et al., 2014; Wills et al., 2005). The few theoretical studies that considered overlapping memory engrams (Bös et al., 1988; Boboeva et al., 2018) in the past focused on overlaps arising from a hierarchical organization of memories. Whereas such a hierarchical approach is suitable for modeling memory representation in the cortex, we are interested in modeling MTL, and in particular area CA3 of the Hippocampus, where experimentally no hierarchical or topographical organization has been observed (De Falco et al., 2016). In experiments, episodic associations between *arbitrary* different concepts (such as a person and a place) - and shared neurons in the corresponding assemblies - can be induced by joint presentation of images representing the different concepts (Ison et al., 2015). Inspired by these experiments, we create pairwise associations between a number of concepts by artificially introducing shared concept cells in the model. We will talk about "overlapping engrams" if the fraction of shared concept cells is significantly beyond the number $N\gamma^2$ of cells that are shared by chance.

Part 1 of the RESULT addresses the first question by developing a theory of stationary activity in a network where assemblies share concept cells; Part 2 addresses the second question and studies the dynamics of association chains in networks of adaptive neurons. Finally, Part 3 and 4 look at the bio-plausibility of the model, first by comparing model predictions to experimental data and second, by testing the model's robustness to different types of heterogeneity.



Caption on the next page

Figure 2.1 *(previous page)* – Overlapping concepts can be retrieved separately and jointly. A) Activation of concepts (schematic). Black filled circles = inactive neurons. Yellow filled circles = active neurons. Colored halos (red, green) represents assignment to a specific concept. When the fraction of shared neurons is small (top row, $c < c_{\max}$) the two concepts can be recalled separately or together. If the number of shared concept cells is too large (bottom row, $c > c_{\max}$), the recall of a first concept (red) at time $t = 1$ leads inevitably to the activation of the second associated concept (green). In this regime, it is impossible to reliably recall a single concept. B) Similarity measure. If only a subset of neurons belonging to the first memory engram is activated (top), the configuration exhibits similarities $m^1 < 1$ and $m^2 \leq \frac{c-\gamma}{1-\gamma}$. If the first memory is fully recalled, while memory two is not (bottom), the similarity measures are $m^1 = 1$ and $m^2 \sim 0$. C) Dynamics of the similarities for different fractions of shared neurons. The similarities m^1 (green) and m^2 (red) as a function of time in a full network simulation (solid lines) are compared to predictions of mean-field theory (dashed lines). External stimulation is given to the units belonging to concept $\mu = 1$ during a stimulation period (in grey), long enough to trigger the activation of the corresponding assembly of concept cells. If $c > c_{\max}$, the assembly corresponding to concept $\mu = 2$ gets activated without receiving any stimulation. D) Three phase-planes of the dynamics of similarity variables m^1 and m^2 for different values of fraction of shared neurons c. Arrows indicate direction and speed of increase or decrease of the similarity variables. Intersections of blue and orange lines (the 'nullclines' of the two variables m^1, m^2) indicate fixed points, with a stability encoded by color (legend). For $c = 0.3$ only two stable fixed points are left. Parameters: $\hat{h}_0 = 0.25$, $\hat{b} = 100$, $r_{\max} = 1$, $\hat{\tau} = \tau = 1$, $\alpha = 0$, $\gamma = 0.002$. For simulations: $N = 10000$, $P = 2$.

2.3 Results

The two questions introduced above can be summarized into a more general one: What is the role of concept cells shared between stored memory engrams during the retrieval of concepts? To answer this question, we consider an attractor neural network of N neurons in which P engrams $1 \leq \mu \leq P$ are stored in the form of binary random patterns (Rey et al., 2018). The pattern $\vec{\xi}^{\mu} = \{\xi_i^{\mu} \in \{0,1\}; 1 \leq i \leq N\}$ with pattern index μ represents one of the stored memory engrams: a value $\xi_i^{\mu} = 1$ indicates that neuron i is part of the stored memory engram and therefore belongs to the assembly of concept μ, while a value of $\xi_i^{\mu} = 0$ indicates that it does not. A network that has stored P memory engrams is said to have a memory load of $\alpha = P/N$.

Since concept-cells in human Hippocampus form sparse neural assemblies with a sparseness parameter $\gamma \sim 0.23\%$ (Waydo et al., 2006), we pay particular attention to the case of sparse memory engrams. In other words, an arbitrary neuron i has a low probability $\gamma = \text{Prob}(\xi_i^{\mu} = 1) \ll 1$ to participate in the assembly of concept cells corresponding to memory engram μ.

The attractor neural network is implemented in a standard way (Pereira and Brunel, 2018;

Hopfield, 1984). Each neuron, $i = 1, \ldots, N$, is modelled by a firing rate model (Hopfield, 1984)

$$\tau \frac{dr_i}{dt} = -r_i + \phi(h_i), \tag{2.1}$$

where $r_i(t)$ is the firing rate of neuron i and $\phi(h) = r_{\max}/\{1 + \exp[-b(h - h_0)]\}$ is the sigmoidal transfer function, or frequency-current (f-I) curve, characterized by the firing threshold h_0, the maximal steepness b, and the maximal firing rate $r_{\max}$. The patterns $\vec{\xi}^{\mu}$ are encoded in the synaptic weights w_{ij} via a Hopfield-Tsodyks connectivity for sparse patterns so that the average of synaptic weights across a large population of neurons vanishes (Tsodyks and Feigel'man, 1988).

In attractor neural network models, memory engrams μ induce stable values $r^*_{\mu,i}$ of the neuronal firing rates during the retrieval of a stored concept. In mathematical terms, to each engram μ corresponds a fixed point $\vec{r}^*_{\mu}$ in such a way that the firing rate $r^*_{\mu,i}$ of neuron i is high if $\xi^{\mu}_i = 1$ and low if $\xi^{\mu}_i = 0$. When the network state $\vec{r}(t)$ is initialized close enough to the stored memory μ, the attractor dynamics drives the network to the retrieval state $\vec{r}^*_{\mu}$ characterized by persistent activity of all those neurons that belong to the assembly of concept μ.

The similarity between the momentary network state and a stored memory μ is defined as

$$m^{\mu}(t) = \frac{1}{N\gamma(1-\gamma)r_{\max}} \sum_{j=1}^{N} \left(\xi^{\mu}_j - \gamma\right) r_j(t). \tag{2.2}$$

The similarity measures the correlation between the firing rates $\{r_j(t)\}_{j=1,\ldots,N}$ and the stored patterns $\vec{\xi}^{\mu}$ such that if memory concept μ is retrieved, then $m^{\mu} \sim 1$ (schematics in Fig. 2.1B), and, if no memory is recalled (*resting state*), then $m^{\mu} \sim 0$ for all μ.

2.3.1 Critical fraction of shared neurons between memory engrams

We start with the case where only the first two assemblies of concept cells (say, one representing a person and the other one a place) are allowed to share neurons above chance level, i.e. the patterns $\vec{\xi}^1$ and $\vec{\xi}^2$ of the corresponding memory engrams are overlapping. The *total number n of shared neurons* in a network of size N depends on the *fraction of shared neurons c* between the two patterns and the sparsity parameter γ, i.e., $n = \gamma c N$.

Let us imagine to gradually increase the fraction of shared neurons between the first two memory engrams. At the lowest end, $c = \gamma$, the patterns $\vec{\xi}^1$ and $\vec{\xi}^2$ are independent, and hence cell assemblies 1 and 2 share only a fraction of neurons corresponding to chance level. It is well known, that in this case, each memory engram generates a separate attractive fixed-point of the network dynamics (Tsodyks and Feigel'man, 1988) implying that the two corresponding concepts can be retrieved separately. However, experimental data reports that, for associated

concepts, the fraction of shared neurons $c \sim 4-5\%$ (De Falco et al., 2016) is much larger than the chance level $\gamma \sim 0.23\%$. This observation suggests that the patterns $\vec{\xi}^1$ and $\vec{\xi}^2$ of two associated memory engrams have a fraction of shared neurons larger than chance level, $c > \gamma$. On the other hand, in the (trivial) limit case of large fraction of shared neurons $c \to 1$, the two memory engrams and hence the two cell assemblies share all neurons, and it is clearly impossible to retrieve one memory without the other.

To study the critical fraction of shared neurons $c_{\max}$ at which independent memory recall breaks down, we use a mean-field approach for large networks and work in the limit $N \to \infty$. In this limit, it is possible to fully describe the network dynamics using the similarities m^μ as the relevant macroscopic variables. Since we are interested in the retrieval process of concepts μ =1 and 2, we assume the similarity of the present network state with other memories $\mu > 2$ to be close to zero: we will refer to these non-activated memories as "background patterns". Under these assumptions, we find dynamical mean-field equations that capture the network dynamics through the similarity variables m^1 and m^2.

$$\tau \frac{dm^1}{dt} = -m^1 + F_1(m^1, m^2) \tag{2.3a}$$

$$\tau \frac{dm^2}{dt} = -m^2 + F_2(m^1, m^2) \tag{2.3b}$$

where the explicit form of the functions F_1 and F_2 is given in Eq. (2.11) of METHODS. Equation (2.3) represents a two-dimensional dynamical systems which can be analyzed using phase-plane analysis. Figure 2.1D shows three phase-planes in the $m^1 - m^2$ space, each for a different value of the fraction of shared neurons. The m^1- or m^2-nullclines solve $dm^1/dt = 0$ or $dm^2/dt = 0$ in Eq. (2.3a) and (2.3b), respectively. The intersections between the m^1- and m^2-nullcline are equilibrium solutions, or fixed points, of the mean-field dynamics and are color-coded according to their stability. For $c = \gamma$, we identify four stable fixed points: the resting state $(m^1, m^2) = (0,0)$, two single-retrieval states $(m^1, m^2) = (1,0)$ and $(m^1, m^2) = (0,1)$ corresponding to the retrieval of concept $\mu = 1$ and the retrieval of concept $\mu = 2$, respectively. Finally, there is a symmetric state which corresponds to the activation of both concepts simultaneously, $(m^1 = m^2 \lesssim 1)$.

If the fraction of shared neurons c exceeds γ, the nullclines are deformed and the fixed points are shifted. The critical value $c = c_{\max}$ of the fraction of shared neurons is reached if the two single-retrieval states merge with their nearby saddle points and disappear. To compute the numerical value of the critical fraction of shared neurons, we extract it from the bifurcation diagram (Fig. S1). For fractions of shared neurons $c > c_{\max}$ only two stable fixed points are left, the resting state and the symmetric state in which assemblies of both concepts are activated together: this symmetric state is the theoretical description of the state that we qualitatively predicted above where the activation of a first concept leads inevitably to the activation of the

second, overlapping one (2.1C, bottom).

In the limit of infinite steepness $b \to \infty$, vanishing load $\alpha = 0$ and vanishing sparseness $\gamma \to 0$, the value $c^0_{\max}$ of the critical fraction of shared neurons can be calculated analytically. Since this value provides an upper bound of the critical fraction of shared neurons for arbitrary b, we have the inequality (Fig. 2.2A)

$$c_{\max} \leq c^0_{\max} \equiv \gamma + (1-\gamma)\frac{h_0}{A\, r_{\max}}, \tag{2.4}$$

where A is a constant that regulates the strength of the synaptic connections. Further analysis (see METHODS) shows that the stationary states of the mean-field dynamics depend – apart from the parameters γ, c and α related to the patterns – only on two dimensionless parameters: the rescaled firing threshold $\hat{h}_0 = h_0/(A\, r_{\max})$ and the rescaled steepness $\hat{b} = b \cdot (A\, r_{\max})$. We find that the critical fraction of shared neurons $c_{\max}$ increases with both $\hat{h}_0$ and $\hat{b}$ (Fig.2.2A).

We proceed by studying how the critical fraction of shared neurons varies as a function of the memory load $\alpha = P/N$ (Fig. 2.2B). As we increase the load, we observe that the critical fraction of shared neurons decreases, but very slowly so (solid grey line). This weak dependence on network parameters is robust against two variations of the network where (i) self-interaction of neurons is excluded; or (ii) the $P-2$ background patterns are also overlapping in pairs, e.g., pattern 3 is overlapping with pattern 4, 5 with 6, etc. For both modifications, the mean-field equations look slightly different (METHODS) but neither modification leads to a significant change of the critical fraction of shared neurons $c_{\max}$ (Fig.2.2A). In a network that has stored a total of P memory engrams, the critical fraction of shared neurons could potentially depend on the group size p of patterns that are all overlapping with each other. So far we have considered $p = 2$. We extended the mean-field approach to the case of three and four overlapping patterns (SI) by rewriting and adapting Eq. (2.11) to the case of three and four overlapping patterns. Again we find that the critical correlation is not significantly influenced by the group size p of overlapping patterns (Fig. 2.2B). The group size p can be large provided that the total number of patterns P does not exceed the memory capacity of the network. For example, the critical fraction of shared neurons of a network that has stored a single subgroup of 64 patterns is similar to that of a network that has stored 32 subgroups of 2 patterns. A fraction of shared neurons $c < \gamma$ corresponds to memory engrams that share less neurons than chance level. While this regime does not seem biologically plausible, we checked that there is a value $c_{\min} < \gamma$ below which memory retrieval is no longer possible.

Finally, we compare the mathematical results of the mean-field theory with simulations in a network of $N = 10000$ neurons. In line with our main theory, we allow only pattern 1 and 2 to be overlapping. After initializing the network in its resting state, we give an external stimulus to all those neurons that belong to the assembly of concept $\mu = 1$ in order to activate them. As

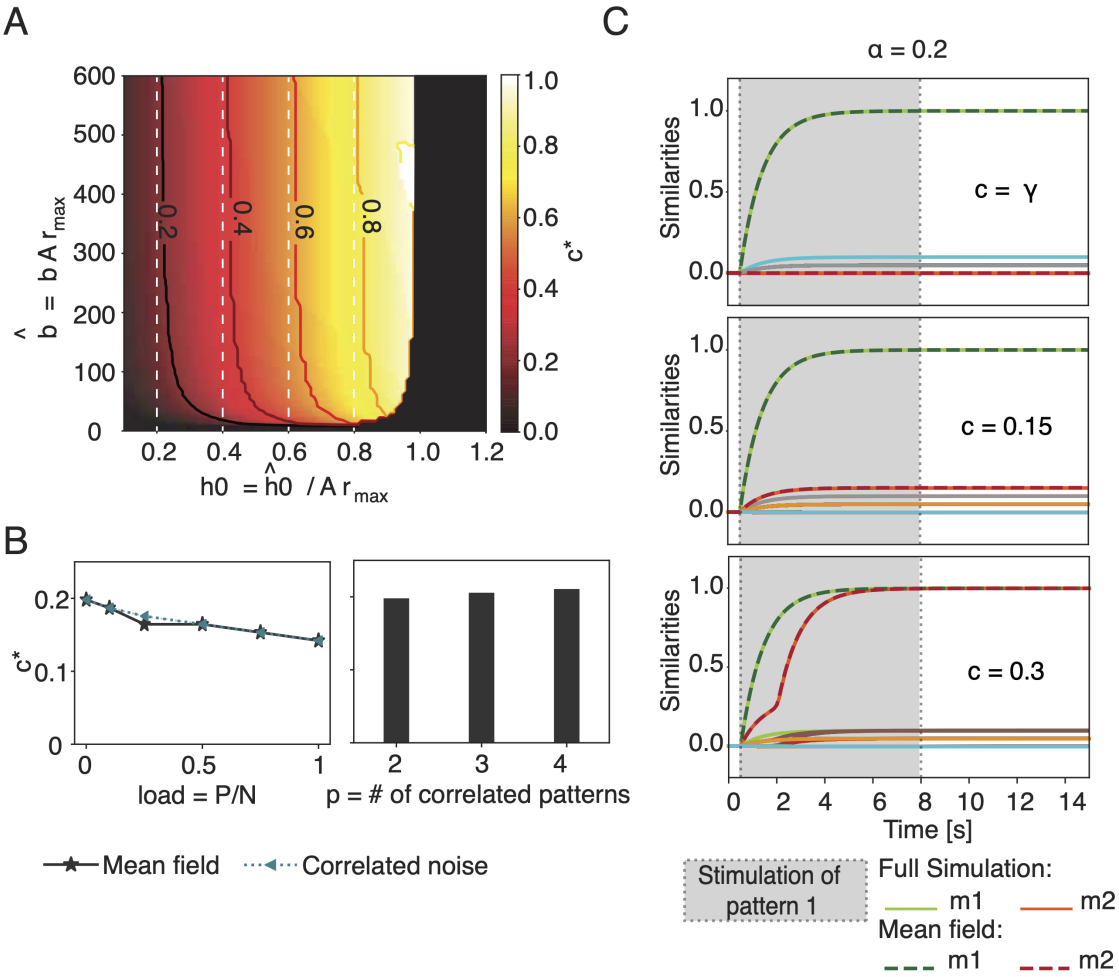


Figure 2.2 – Critical fraction $c_{\max}$ of shared neurons depends on the neuronal f-I curve but not on the memory load. A) Critical fraction ($c_{\max}$, color code) as a function of the parameters $\hat{b} = b\,Ar_{\max}$ (steepness) and $\hat{h}_0 = h_0/(Ar_{\max})$ (firing threshold) of the f-I curve. The level set is added for the indicated values of $c_{\max}$. In the black area the resting state is the only stable solution. Vertical white dashed lines indicate the asymptotic value $c^0_{\max}$ for different values of $\hat{h}_0$. B) Critical fraction $c_{\max}$ of shared neurons as a function of the memory load $\alpha = P/N$ (left graph) without (solid grey line) or with overlaps in pairs of two of the $P-2$ background patters (dashed green line); and as a function of the number p of correlated patterns (histogram, right graph). C) As in Fig. 2.1C, but with a large number of background patterns ($\alpha = 0.2$). Network activity exhibits only small similarity with background patterns (diversely colored lines) but large similarity with the stimulated pattern $\mu = 1$. Parameters (unless specified): $\gamma = 0.002$, $\hat{b} = 100$, $\hat{h}_0 = 0.25$; $\alpha = 0$ in A-B. For simulations in C: $N = 10000$, $p = 2$, $\gamma = 0.002$.

predicted by the mean-field theory, for a fraction of shared neurons smaller than the critical value, concept 2 is not activated, while for values $c > c_{\max}$, the activity of the cells representing

concept 1 triggers, after some delay, activity in the cells representing concept 2. The lag in the activation decreases with increasing C. While in Fig. 2.1C the network has stored a very small number of memory engrams ($\alpha = 0$), the network in Fig. 2.2C has stored 2000 memory engrams ($\alpha = 0.2$), yet the time course of activation of concepts is similar.

2.3.2 Association chains

Neurons shared between memory engrams have been proposed to be the basis for the recall of a memorized list of words (Romani et al., 2013; Recanatesi et al., 2015; Katkov et al., 2015; Recanatesi et al., 2017; Naim et al., 2020). In order to translate this idea to chains of associated concepts (Fig. 2.3A), we follow earlier work (Romani et al., 2013; Recanatesi et al., 2015; Katkov et al., 2015; Recanatesi et al., 2017; Naim et al., 2020) and add two ingredients to the model of the previous subsection. First, we add a global inhibitory feedback current, whose strength $J_0(t)$ is periodically modulated by oscillations mimicking Hippocampal oscillatory activity. The oscillations provide a clock signal that triggers transitions between overlapping concepts. The importance of the global feedback signal is discussed later. Second, we add to each neuron i an adaptation current $\theta_i(t)$ in order to prevent the network state to return, after a transition, to the previous concept. With this extended model, the network state hops from one concept to the next (Fig. 2.3B). The next concept to be activated is the one which shares the most concept cells with the currently active one, similar to previous models (Romani et al., 2013; Recanatesi et al., 2015; Katkov et al., 2015; Recanatesi et al., 2017; Naim et al., 2020). Transitions are repeated, but after some time the network state returns to one of the already retrieved memories, leading to a periodic cycle of patterns (Romani et al., 2013) (Fig. 2.3B). In our theory, the number of shared concept cells is identical between all pairs of concepts within the same subgroup, so that the order of the recalled concepts depends on the initial condition. In network simulations where concepts are represented by sparse memory engrams ($\gamma = 0.002$), we allow a subgroup of $p = 2, 4$ or 16 memory engrams to share a fraction of neurons of $c = 0.2$. If the subgroup of overlapping engrams is small ($p = 2, 4$), all patterns are retrieved, while for a large group of overlapping engrams ($p = 16$) the cycle closes once a subgroup of the overlapping memory engrams has been retrieved (Fig. 2.3B).

In previous studies (Romani et al., 2013; Recanatesi et al., 2015; Katkov et al., 2015; Recanatesi et al., 2017; Naim et al., 2020), each memory engram involved a large fraction ($\gamma = 10\%$) of neurons so that transitions could rely on the number of units shared by chance. However, given that the value of the sparsity in MTL is much smaller ($\gamma \sim 0.23\%$), it is natural to ask whether the number of neurons shared by chance ($c = \gamma$) is sufficient to induce a sequence of memory retrievals. Our simulations indicate that this is not the case (Fig. 2.3C): decreasing γ leads to a gradual disruption of the sequential retrieval dynamics. Thus, in a network storing assemblies with a realistically low sparsity $\gamma \sim 0.002$, memory engrams with a fraction of

shared neurons above chance are necessary for the retrieval of chains of concepts.

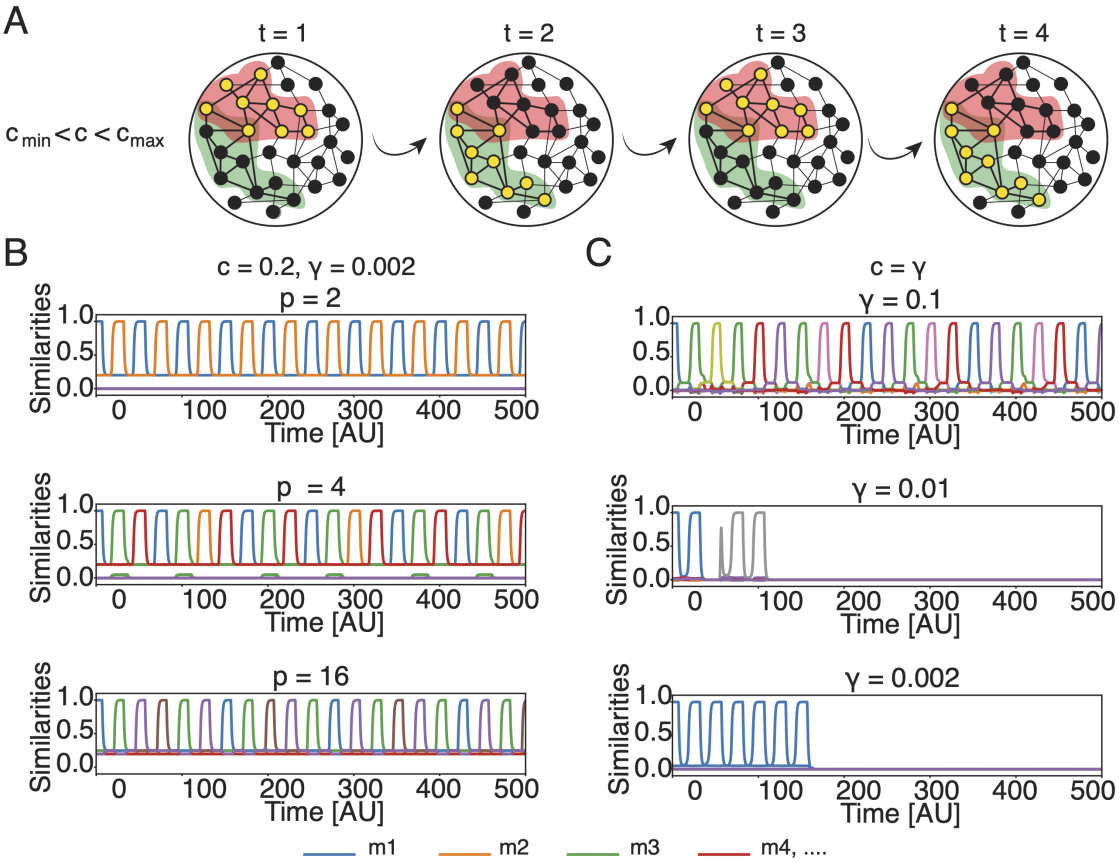


Figure 2.3 – Chain of associations requires shared concepts cells. A) Schematic of a chain of association cycling between two concepts. Assignment of cells to assemblies is indicated by halos' color. Filled black circles indicate inactive neurons and filled white circles indicate active neurons. The schematics corresponds to the upper line of part B. B) Full network simulation for engrams overlapping above chance level ($c = 0.2 > \gamma$) with low sparsity ($\gamma = 0.002$). Each line corresponds to the similarity m^{μ} with one of the stored memory engrams as a function of time. A subgroup of p engrams is overlapping (top to bottom: p =2,4,16. If the network state is initialized to retrieve one of the overlapping concepts, other concepts within the subgroup are retrieved later. C) Same as in B, but memory engrams are independent ($c = \gamma$) and only share cells by chance. By decreasing their mean activity γ, the retrieval dynamics of a chain of memories is disrupted. The match between mean-field theory and simulations is shown in Fig. S5. Parameters: $N = 10000$, $P = 16$, $\hat{b} = 100$, $\tau_\theta = 45$, $T = 0.015$, $T_{J_0} = 25$.

To better understand the role of overlaps between engrams for the formation of association chains, we extend the mean-field dynamics to include the global feedback with periodic modulation $J_0(t)$. Since simulations indicate that overlaps are necessary, we want to estimate

the minimal and maximal fractions of shared neurons required to enable association chains. Intuitively, the fractions of shared neurons should be large enough to allow the assembly of the next concept to be activated, but not too large in order to keep concepts separated - and avoid the merging of concepts discussed above in Figs. 2.1 and 2.2. Because, in our model, the strength of the global inhibition $J_0(t)$ exhibits a slow periodic modulation, we consider the dynamics as quasi-static and consider the two limit cases, one when J_0 is at its maximum and another one when J_0 is at its minimum. For our parameter setting, when $J_0(t)$ is clamped at its **minimum**, the network possesses three stable states: the resting state and the two single retrieval states. For a successful association chain, we need that concepts can be retrieved separately. The fraction of shared neurons, $c_{\max}$, that makes the two single retrieval states disappear therefore sets the upper bound of the useful correlation range. The parameter $c_{\max}$ is analogous to $c_{\max}$ in the previous section, but evaluated in the presence of global inhibition. Global inhibition increases competition between activated neurons so that, compared to Fig.2.1 two concepts cannot be retrieved simultaneously, but only one or the other or none.

Next, we consider the situation when the global inhibition is clamped at its maximum and find the minimal fraction such that the system has, besides the resting state, a second stable fixed point for $m^1 = m^2 > 0$ where the assemblies of both the previous and next concept are simultaneously active at low firing rates. Since this state is necessary to enable the transition, we call it the *transition state*. If the transition state is present, the network could, once global inhibition decreases, either return from the transition state to the previous concept, or jump to the next one (Fig. 2.4B, right side). However, in the presence of adaptation (which is not included in the phase plane picture of Fig. 2.4), the transition to the next concept is systematically favored because neurons participating in the assembly of the earlier concept are fatigued. Thus, the lower bound of fraction of shared neurons $c_{\min}$ is the smallest overlap such that the transition state exists, and the existence of the transition state is a necessary condition for the formation of temporal association chains. Since in the mean-field limit, the transition state appears only for non-vanishing c, a fraction of shared neurons above chance level is needed to allow the hopping between concepts. In Fig. 2.4C we show the dependence of the maximal and minimal fraction of shared concept cells upon the sparsity γ and the steepness b: in both cases the dependence is not strong, but sparser networks lead to a slightly smaller range of the admissible fraction c of shared neurons supporting association chains. Importantly, the minimal fraction of shared neurons between concepts necessary to make association chains possible is significantly above the fraction of neurons that are shared by chance in a neural network where memory engrams correspond to a random subset of neurons.

In conclusion, we have shown the need for correlations between memory engrams – equivalent to a number of shared concept cells significantly above chance level – to explain the free memory recall paradigm in recurrent networks such as the human MTL where each engram

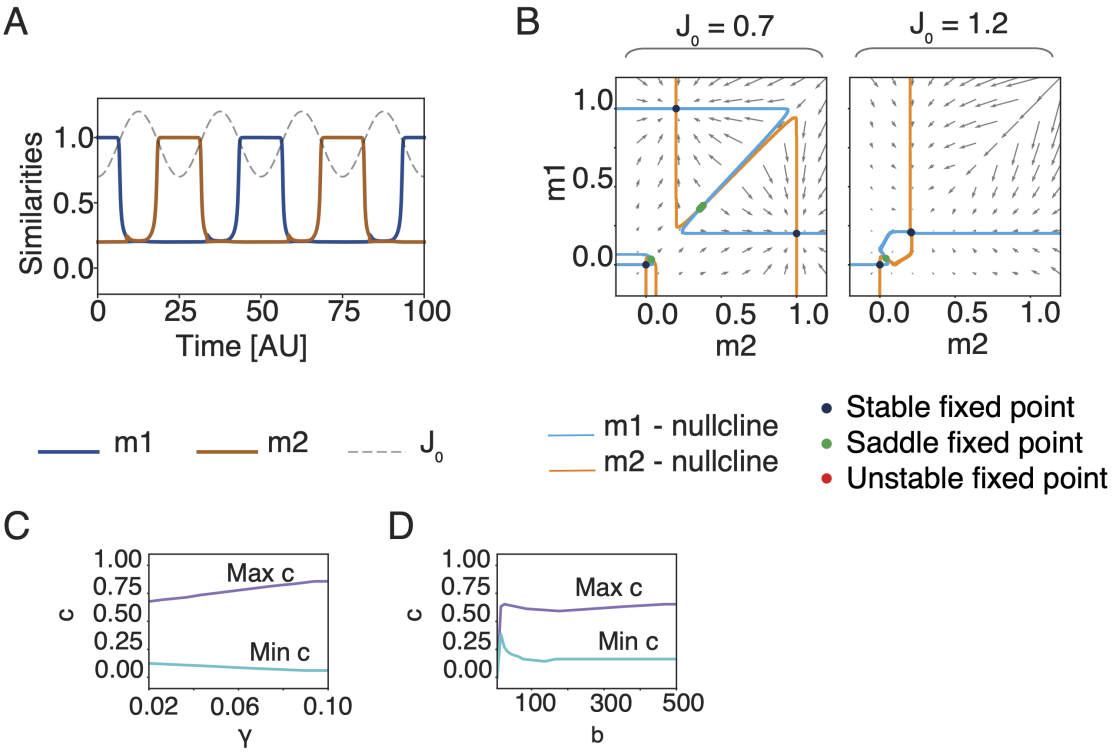


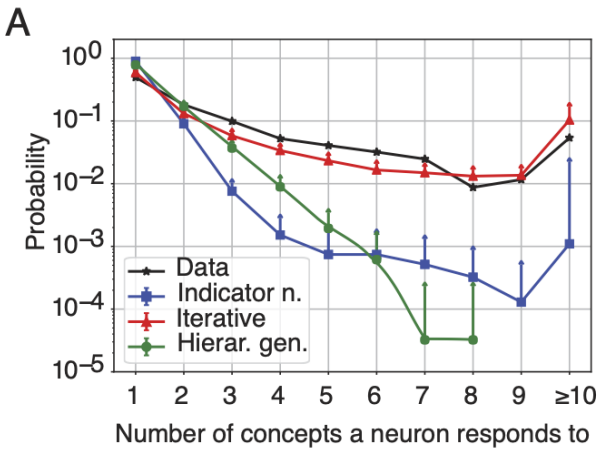
Figure 2.4 – Dependence of association chains on sparsity and neuronal parameters. A) Dynamical mean-field solutions for m^1 and m^2 in the case of two correlated patterns. B) Phase planes corresponding to the minimum ($J_0 = 0.7$) and maximum ($J_0 = 1.2$) value of inhibition in the case of two associated patterns. C) Minimal and maximal fraction of shared concept cells as a function of the sparsity γ and F) of the steepness b. Parameters (unless specified): $\gamma = 0.002$, $\hat{b} = 100$, $c = 0.2$, $\theta_i = 0$ for every i.

involves only a small fraction of neurons. The dynamics of the mean-field theory leads to predictions that are confirmed by our simulations.

2.3.3 Comparison with experimental data

We compare the predictions of our model with experimental data using a previously published dataset of human concept cells (De Falco et al., 2016). The dataset contains the activity of 4066 neurons recorded from the human MTL during the presentation of several visual stimuli, from which we can extract the experimental probability that a single neuron responds to exactly k different concepts (Fig. 2.5A, black stars). From the probability distribution, we observed the existence of neurons responding to a large number of concepts (10 or more). We will refer to those neurons as “hub neurons”. We want to investigate whether, and under which conditions, our model accounts for the existence of such hub neurons. In the model, the

probability that a neuron responds to a certain number of stimuli depends on how the pairwise overlaps between memory engrams is built. Here we test three different algorithms to generate correlated patterns in a network of $N = 100'000$ neurons. All algorithms generate random patterns with sparsity $\gamma = 0.2\%$, therefore they all assign, on average, the same number of active neurons per pattern, when an isolated pattern is considered (Fig. 2.5B, $p = 1$). Moreover, all algorithms generate an identical pairwise fraction of shared neurons between patterns ($c = 0.04$) , meaning that a subgroup of only two correlated patterns involves the same number of neurons (Fig. 2.5B, $p = 2$). However, given the different procedures used to generate pairwise correlations (METHODS), each algorithm uses a different number of neurons when it generates subgroups of $p = 4$ or $p = 16$ patterns. The indicator neuron model assumes that a small



B

Algorithm	p = 1	p = 2	p = 4	p = 16
Hierarch. gen. model	200	392	753	2398
Indicator neuron model	200	392	773	3001
Iterative model	200	392	753	2414

Figure 2.5 – Comparison between theory prediction and experimental data on the number of concept a neuron responds to. A) Probability that a neuron responds to a given number of concepts: comparison between data and 3 different algorithms: the indicator neurons, the iterative correlation and the hierarchical generative model. Hub neurons (right-most data point) are defined as those responding to ≥ 10 concepts). Each algorithm was run 40 times to generate the mean and error bars (only upward bars are displayed, corresponding to one standard deviation). B) For each of the three correlation algorithms, we generated three subgroups of patterns containing $p = 16$, $p = 4$, or $p = 2$ patterns, and an isolated pattern ($p = 1$). The table gives the number of neurons needed to represent each subgroup in a neural network of $100'000$ neurons if patterns have sparsity $\gamma = 0.002$ and a pairwise fraction of shared neurons $c = 0.04$.

subset of neurons indicates to which group of concepts a stimulus belongs (akin to assigning a title to each subgroup) whereas the iterative correlation model builds pairwise correlations as new stimuli are included into a subgroup without assuming any hierarchical arrangement. Both the indicator neuron model and the iterative correlation model predict the existence of hub neurons, supporting the validity of our approach to model networks of concept cells. However, the iterative correlation model better fits the data (Fig. 2.5.C) and it is the only one

that predicts the correct probability of finding hub neurons. Since the iterative correlation model is not based on a hierarchical generation of patterns, this suggests that the MTL encodes large subgroups of memory engrams in a non-hierarchical way.

2.3.4 Robustness to heterogeneity

Because biological neural networks present different forms of heterogeneity, we have checked our model's robustness to (i) the heterogeneity of F-I curves and (ii) dilution of the number of synaptic connections.

In the experimental data set, each neuron is characterized by different baseline firing rates and maximal rates in response to the preferred stimulus. We therefore introduce in our model heterogeneous F-I curves characterised by a minimum and a maximum firing rates $(r_{\min})_i$ and $(r_{\max})_i$ respectively and renormalize the network dynamics appropriately (METHODS). Figure 2.6A is equivalent to Fig. 2.2C but with heterogeneous F-I curves and shows that the full network simulations are robust to heterogeneous F-I curves when recalling a single memory. Similarly Fig. 2.6B is the equivalent of Fig. 2.3B, proving that associations chains do not get disrupted by the introduction of heterogeneous F-I curves.

Secondly, we allow the weight matrix to be diluted. Whereas so far we have assumed an "all-to-all" connectivity, we now introduce the dilution coefficient d, which indicates the fraction of synaptic connections we actually allow out of all possible ones. Simulations in Fig. 2.6C show that the model is robust for $d = 0.8$, i.e. after dropping 20% of all possible synaptic connections in combination with an appropriate rescaling of the average connection strength. We emphasize that for sparsely connected networks, the memory load α is $\alpha = P/M$ where M is the mean number of connections per neuron (see SI for details).

2.4 Discussion

Our theory and simulations show that overlaps between memory engrams are a suitable mechanism for encoding the association between concepts. Our results bridge between observations and theories from four different fields: first, experimental observations in the human MTL indicating that learned associations are encoded by neurons shared between assemblies of concept cells (Quiroga et al., 2005; Quiroga, 2012, 2019b,a; De Falco et al., 2016); second, experimental observations of memory engrams where each memory engram consists of a set of neurons activated during a specific memory (Tonegawa et al., 2015; Josselyn and Tonegawa, 2020); third, the theory of association chains used to explain free memory recall (Romani et al., 2013; Recanatesi et al., 2015, 2017; Naim et al., 2020); fourth the classic theory of attractor neural networks (Hopfield, 1982, 1984); and fifth the experimental observation

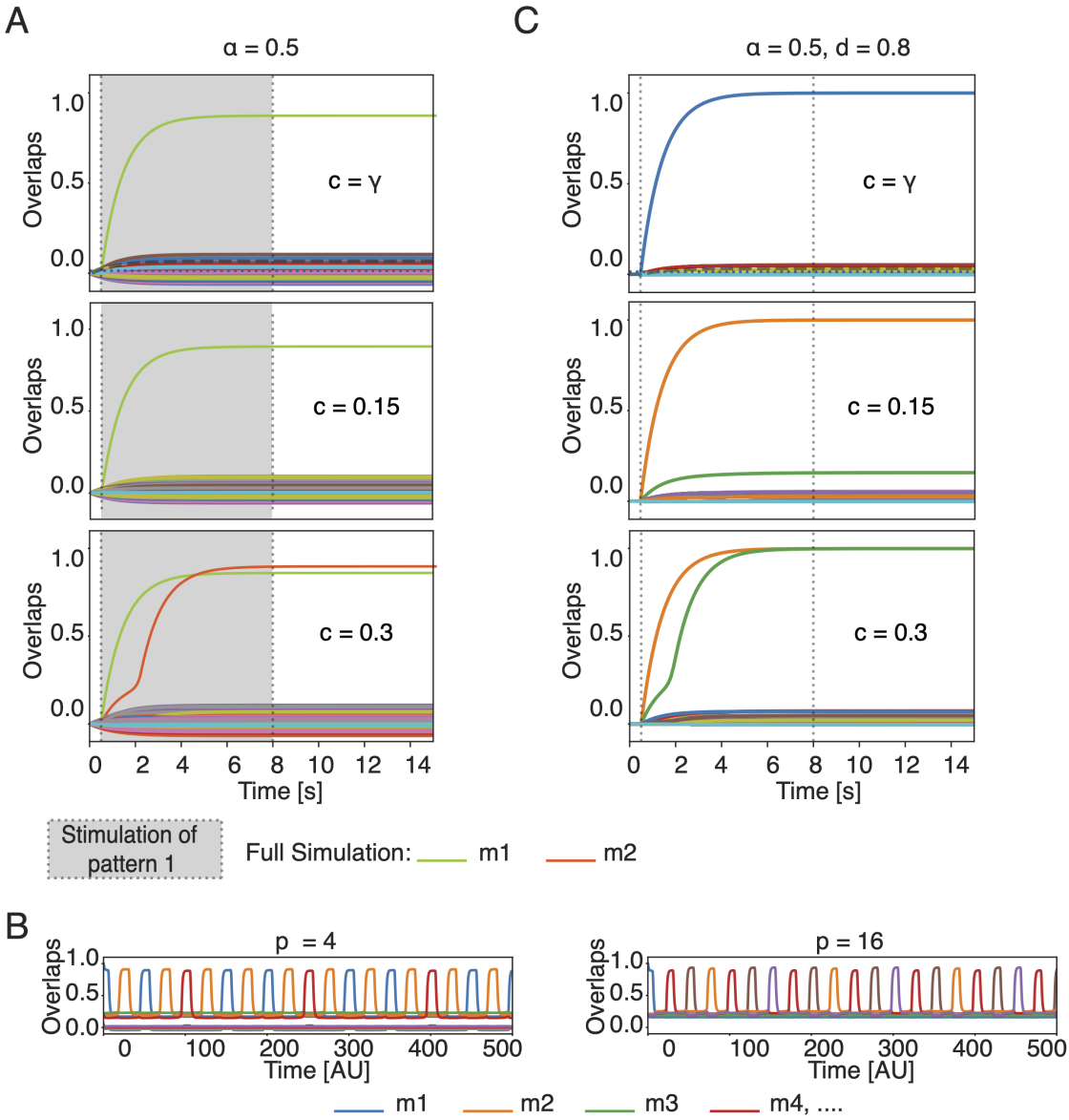


Figure 2.6 – The model is robust to heterogeneity of F-I curves. Full network simulations A) in absence of adaptation, equivalent to Fig. 2.2C, and B) in presence of adaptation and periodic inhibition. C) The model is robust to the dilutions of the synaptic connections. Full network simulations equivalent to Fig. 2.2C

of the number of concepts a neuron responds to. Our main result is that, in networks were concepts are encoded by sparse assemblies, the number of shared concept cells must be above chance level but below a maximum in order to enable a reliable implementation of association chains. In particular, with 4-5% overlap between memory assemblies as reported in the human

MTL (De Falco et al., 2016), association chains are possible for broad range of parameters of f-I curves; see Figs 4C,D,E. To show this result, we have extended the classical mean-field formalism (Amit and Amit, 1992) to the case in which memory engrams are overlapping. Overlaps can be across pairs or larger subgroups of concepts.

We have shown how overlaps between memory engrams influence the retrieval of patterns in two different setups. In the first set-up, we have analyzed the recall of a single concept that shares concept cells with a second concept and computed the critical fraction of shared neurons of memory engrams above which concepts can no longer be recalled individually but effectively act as one unified enlarged concept. Despite the fact that memory patterns in the area CA3 of the Hippocampus are known to be very sparse, (for human hippocampus it is debated whether dentate gyrus provides additional orthogonalization (see (Marr and Thach, 1969; Marr et al., 1971) vs (Quiroga, 2020)) Hebbian learning could induce overlaps between a small number of specific memories engrams by shared concept cells (De Falco et al., 2016). The existence of a critical maximal fraction of shared neurons implies that Hebbian learning must work with an intrinsic control mechanism so as to avoid unwanted merging of separate concepts.

The second set-up is that of association chains that could form the basis of a "stream of thought" where the direction of transitions from one concept to the next is based on learned associations, e.g., as experienced in episodic memory. Our network dynamics is inspired by the model of Romani and Tsodyks (Romani et al., 2013) and follow-up studies from the same group (Recanatesi et al., 2015, 2017; Naim et al., 2020). These studies give a theoretical explanation of the number of words (and their distribution) typically recalled after memorizing a list of words. Transitions between words depend on the number of shared units between the stored memory engrams. Even though in the Romani-Tsodys model memory engrams are independent, finite size effects make some pairs of engrams share neurons above chance level which enables sequential recall in the presence of a periodic background input. In contrast to their work on list of words, we ask if such a network is suitable to model association chains in the human MTL where memory representation is extremely sparse ($\gamma \approx 0.23\%$). We find that, in the limit of large networks, neurons shared by chance are not enough to reliably induce the retrieval of a chain of concepts. Therefore, we extend our mean-field theory to the case of time-dependent periodic background input and prove that overlaps between memory engrams can restore the stream of thought dynamics in both the high sparsity and big networks regimes. For sparse memory engrams, encoding associations with overlaps ensures that only learned associations, based on real-life episodes, influence the recall dynamics, excluding the possibility of randomly formed associations.

In our model, we trigger transitions in the association chain by an oscillatory background similar to the Romani-Tsodyks model (Romani et al., 2013), but other mechanisms for triggering

transitions are also possible. For example, in the Russo-Treves model (Russo et al., 2008; Russo and Treves, 2012; Akrami et al., 2012) the switch from one pattern to the next is triggered by two adaptation mechanisms that act on different time scales without the need of periodic inhibition.

2.4.1 Bio-plausibility and parameter choice

Attractor networks are suitable candidate models for associative memory because they present two features: (i) memory retrieval after stimulation with a partial cue and (ii) sustained activity after a stimulus has been removed. Many steps have been taken to increase the bio-plausibility of attractor models, beyond the Hopfield model (Hopfield, 1982, 1984). Two steps have been the storage of $\{0,1\}$-patterns of sparsity $\gamma \ll 1$ (Tsodyks and Feigel'man, 1988) and the switch from all-to-all connectivity in rate networks to random connectivity in spiking networks (Amit and Brunel, 1997). Moreover, the requirement of symmetric connections is also obsolete, because the same mean-field equations classically derived with the replica theory (Amit et al., 1987) or cavity method (Mézard et al., 1987; Shamir and Sompolinsky, 2000), can alternatively be derived with dynamical systems arguments that do not require the weight matrix to be symmetric (Shiino and Fukai, 1992). In this work, we store sparse binary $\{0,1\}$-patterns and use the dynamical approach to derive the mean-field equations of the system. Even though we use a fully connected and symmetric synaptic weight matrix in part 1 and 2 of RESULTS, these characteristics are not a necessary requirement in any of the derivations and we show in Part 3 of the RESULTS that the model is robust to a random (non-symmetric) dilution of the synaptic weight matrix.

The maximal number of patterns that can be stored in an attractor neural networks has attracted a lot of research (Amit et al., 1985b; Amit and Amit, 1992; Tsodyks and Feigel'man, 1988). However, does the hippocampus actually operate in the regime of high memory load? Let's estimate an upper bound for the load α in the area CA3 of the Hippocampus as follows: an adult human can recognize at most about $P \approx 3 \cdot 10^4$ objects (Biederman, 1987) (which is also about the number of words a native English speaker knows according to The Economist, Lexical Facts). In CA3 there are about 2.7×10^6 neurons (and 34×10^6 in the whole Hippocampus) (Harding et al., 1998), each receiving inputs from about M =30000 other neurons (Andersen et al., 2006). For sparsely connected networks the load is defined by comparing the number of patterns P with the number of connections M per neuron $\alpha = P/M$. Hence we estimate a load $1 < \alpha < 5$ if concepts are stored in area CA3 – and our theory incorporates such a high load. For full connectivity ($M = N$) and sparse memory engrams ($\gamma \ll 1$), the maximum capacity predicted by the mean-field theory is larger than $\alpha = 1$ (Amit et al., 1987; Tsodyks and Feigel'man, 1988; Weisbuch and Fogelman-Soulié, 1985), it is, in principle, possible to store more patterns than neurons.

In the first section of the RESULTS, we find that the maximal number of neurons which two concepts can share before they effectively merge into a single concept mainly depends on two dimensionless parameters: the rescaled threshold $\hat{h}_0 = h_0/(Ar_{\max})$ and the rescaled steepness $\hat{b} = Ar_{\max}b$. Such parameters have so far not being estimated for the human CA3 area of the Hippocampus or for the MTL in general. A technique for extracting these parameters *in-vivo* is proposed inc(Pereira and Brunel, 2018) where they were able to fully characterize the gain function of pyramidal neurons of the macaque's IT cortex.

In the absence of a similar quantification for the human Hippocampus, we calculate in Fig. 2.2b the critical number of shared neurons as a function of $\hat{h}_0$ and $\hat{b}$ over a plausible range of the two parameters. Secondly, real neurons are characterized by different levels of responsiveness: we modeled such variety with heterogeneous F-I curves and showed that the model is robust to it.

Finally, by comparing the experimental measured number of concepts a neurons responds to and model predictions we see that the model can predict the number of hub-neurons quite accurately. The choice of how to build engrams overlaps plays a key role in fitting the experimental data and confirms the idea that memory engrams in the Hippocampus are not hierarchically organised.

2.5 Methods

We consider an attractor neural network of N rate units with firing rates r_i, in which P memory engrams are stored. Each engram $1 \leq \mu \leq P$ consists of a binary random pattern $\xi_i^\mu \in \{0,1\}$ where the index i runs over all neurons $1 \leq i \leq N$. For fixed μ, the set of numbers $\{\xi_i^\mu\}^{i=1,..,N}$ is in the following called the pattern μ. Each neuron follows the rate dynamics of Eq. (2.1). The connection matrix w_{ij} contains the synaptic weights from neurons j to neurons i and is defined as (Tsodyks and Feigel'man, 1988; Pereira and Brunel, 2018)

$$w_{ij} = \frac{A}{N\gamma(1-\gamma)} \sum_{\mu}^{P} \left(\xi_j^\mu - \gamma\right)\left(\xi_i^\mu - \gamma\right) \tag{2.5}$$

where the constant A can be interpret as the global scale of "connection strength". Subtracting in Eq. (2.5) the expected number of ones in the random patterns, $\langle \xi_i^\mu \rangle = \gamma$, guarantees that $\langle w_{ij} \rangle = 0$, where $\langle . \rangle$ indicates expectation over the random numbers that make up the patterns.

2.5.1 Model without adaptation and global feedback

For deriving the results in Figs. 2.1 - 2.2, the total input driving neuron i is

$$h_i(t) = \sum_j w_{ij} r_j(t) + I_i(t) = \sum_p (\xi_i^\mu - \gamma) m^\mu(t) + I_i(t) \tag{2.6}$$

where I_i is the external input. The similarity measure (also called "overlap" in the attractor network literature) m^μ measures the similarity (correlation) of the current network state with pattern μ; cf. Eq. (2.2). In Figs. 1C 2C and 6 the external input $I_i = I^{\text{ext}} \xi_i^1$ is positive during stimulation for all neurons that belong to the assembly of pattern $\mu = 1$.

2.5.2 Model with adaptation and global inhibitory feedback

For Fig. 2.3 of RESULTS, we added adaptation and a global inhibitory feedback to the model as described in previous studies (Romani et al., 2013; Recanatesi et al., 2015, 2017; Naim et al., 2020). Specifically, we add two negative feedback terms to the input potential:

$$h_i(t) = \sum_{j=1}^{N} w_{ij} r_j(t) - \theta_i(t) - \frac{J_0(t)}{\gamma N} \sum_{j=1}^{N} r_j(t), \tag{2.7}$$

First, the variable $\theta_i(t)$ models neuron-specific firing-rate adaptation via the first-order kinetics

$$\tau_\theta \frac{d\theta_i}{dt} = -\theta_i + D_\theta r_i. \tag{2.8}$$

Here, τ_θ is the adaptation time constant and D_θ determines the strength of adaptation. Note that this adaptation model with a hyperpolarizing feedback current is equivalent to a model in which adaptation is implemented as an increase in the threshold $h_0 + \theta_i(t)$.

Second, the global inhibitory feedback term proportional to $J_0(t)$ (third term in (2.7)) provides a clock signal that triggers transitions betweeen attractors. Importantly, inhibition proportional to the summed activity of the network units penalizes network configurations with many active neurons and therefore reduces stability of the fused state where two memories are merged during retrieval. Here, the strength $J_0(t)$ of the global feedback is modulated periodically between values 0.7 and 1.2 with a sinusoidal time course of period T_{J_0} that sets the time scale of transitions between memories.

Note that the model without adaptation and global feedback is a special case of the full model and found by setting $D_\theta = 0$ and $J_0(t) \equiv 0$.

2.5.3 Mean-field equations for two overlapping patterns

We introduce correlation between two of the stored patterns. Without loss of generality, we take patterns $\vec{\xi}^1$ and $\vec{\xi}^2$ to be correlated, while all other $P-2$ patterns are independent. We define the correlation between the two patterns as the Pearson correlation coefficient (covariance/variance):

$$C = \frac{\mathrm{Cov}(\xi_i^1, \xi_i^2)}{\mathrm{Var}(\xi_i^\mu)} = \frac{P_{11} - \gamma^2}{\gamma(1-\gamma)}, \tag{2.9}$$

where $P_{11} = P(\xi_i^1 = 1, \xi_i^2 = 1) = \langle \xi_i^1 \cdot \xi_i^2 \rangle$ is the joint probability of a neuron to be selective to both patterns. We generate correlated patterns with mean activity $\langle \xi_i^1 \rangle_i = \langle \xi_i^2 \rangle_i = \gamma$ and correlation coefficient C, using the procedure described in SI. The fraction c of shared neurons is related to C by the identity $c = C(1-\gamma) + \gamma$.

We are interested in the retrieval dynamics of the correlated patterns $\vec{\xi}^1$ and $\vec{\xi}^2$. The input h_i now depends on the similarities m^1 and m^2:

$$\hat{h}_i(m^1, m^2) = \left(\xi_i^1 - \gamma\right) m^1 + \left(\xi_i^2 - \gamma\right) m^2 + \sqrt{\alpha R(m^1, m^2)} Z_i + I_i \tag{2.10}$$

with

$$\hat{\tau} \frac{dm^1}{dt} = -m^1 + F_1(m^1, m^2) \tag{2.11a}$$

$$\hat{\tau} \frac{dm^2}{dt} = -m^2 + F_2(m^1, m^2) \tag{2.11b}$$

$$R(m^1, m^2) = \frac{p}{(1-q)^2} \tag{2.11c}$$

$$q = \sum_{x^1=0,1} \sum_{x^2=0,1} P_{x^1,x^2} \int \hat{\phi}'(\hat{h}_{x^1 x^2}(m^1, m^2, z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} \tag{2.11d}$$

$$p = \sum_{x^1=0,1} \sum_{x^2=0,1} P_{x^1,x^2} \int \hat{\phi}^2(\hat{h}_{x^1 x^2}(m^1, m^2, z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}}, \tag{2.11e}$$

where

$$F_\mu(m^1, m^2) = \sum_{x^1=0,1} \sum_{x^2=0,1} \frac{x_\mu - \gamma}{\gamma(1-\gamma)} P_{x^1 x^2} \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \hat{\phi}(\hat{h}_{x^1 x^2}(m^1, m^2, z)) \tag{2.12}$$

and

$$\hat{h}_{x^1 x^2}(m^1, m^2, z) = \sum_{\nu=1,2} \left(x_\nu - \gamma\right) m^\nu + I(x^1, x^2) + \sqrt{\alpha R_h(m^1, m^2)} z. \tag{2.13}$$

For details see SI.

A completely analogous procedure can be used to generate more than two correlated binary patterns. In Fig. 2.2B we generate a subgroup of $p = 2, 3, 4$ correlated patterns and compute their critical correlation by solving the system of equations that is analogous to (2.11). To

generate Fig. 2.3C-D we use the mean-field dynamics in the presence of adaptation and global feedback, analogously to Eq. (2.56). All details are provided in the SI.

2.5.4 Excluding self-interaction

In order to make the network more bio-plausible, we can consider that a neuron does not send direct input to itself. The effect of excluding the self-interaction term in Eq. (2.10) is captured by a correction term to be included on the right-hand-side of Eq. (2.13) (Shiino and Fukai, 1992):

$$\frac{q\alpha\hat{\phi}(\hat{h})}{(1-q)}. \tag{2.14}$$

2.5.5 Correlation between background patterns

In Fig. 2.2B we explore the possibility that the maximal fraction of shared neurons correlation $c_{\max}$ might be influenced by the presence of correlation between pairs of background patterns. Moreover, the hypotheses that there are many subgroups of correlated patterns seems more biologically plausible. If we let the background patterns to be correlated in subgroups of 2 patterns each, the variable R in the mean-field equations of Eq. (2.11) needs to be replaced by

$$R' = \frac{p}{\left[(1-q)^2 - (qC')^2\right]^2} \times \tag{2.15}$$

$$\left\{\left[(1-q)^2 + (qC')^2\right](1+C^2) + 4(1-q)qC'\right\} \tag{2.16}$$

where $C' = \frac{P_{11}-\gamma^2}{\gamma(1-\gamma)}$ is the correlation between pairs of patterns. The detailed derivation is provided in the SI.

Furthermore, Eq. (2.16) can be extended to the case in which background patterns share correlation C between non-overlapping subgroups of exactly p patterns (the complete derivation is provided in the SI).

2.5.6 Algorithms to generate correlated patterns

In this section we describe how a single subgroup of K overlapping patterns with sparseness γ and overlap $c$is created according to three different algorithms. The first two models are variations of a hierarchical construction for correlated patters: in both cases, the correlated patters are generated from a common parent patter. In the hierarchical generative model only the active units of the parent pattern can be inherited by the correlated children patterns, while in the indicator neuron model, children patterns can have active units, which do not belong to the parent pattern. Finally in the iterative model, no structure is assumed and the

subgroup of correlated patterns is built pattern by pattern, following an iterative algorithm. Details and the theoretical probability distribution associated to the algorithms are given in the SI.

Hierarchical generative model. We start by creating a "parent" pattern which is not part of the subgroup. The parent pattern has sparseness $\lambda = \gamma/c$, i.e., $\text{prob}(\xi_i^{\text{parent}} = 1) = \lambda$. We proceed to create the actual patterns by copying the ones of the parent pattern with probability c, while the zeros stay untouched: $\text{prob}(\xi_i^{\mu} = 1|\xi_i^{\text{parent}} = 1) = c$, $\text{prob}(\xi_i^{\mu} = 1|\xi_i^{\text{parent}} = 0) = 0$.

Indicator neuron model. We generate with probability λ a small subset of indicator neurons. This subset gives a parent pattern of indicator neurons: $\text{prob}(\xi_i^{\text{parent}} = 1) = \lambda = \frac{c-\gamma^2}{1+c-2\gamma}$. In a network of N neurons, we select $n_{ind} = \lambda N$ neurons as indicator neurons. To create a pattern μ, we flip the bits of the parent pattern with probability ϵ (i.e. a bit is flipped either from 0 to 1 or from 1 to 0 with probability ϵ.) The desired values of λ and ϵ are obtained by expressing the correlation coefficient C and the sparseness γ as a function of them and reversing the formula.

Iterative model. We start by generating the first pattern with γN active neurons. Neurons that have not been selected for constructing any pattern are classified as "untouched neurons". For the construction of each of the following patterns, from 2 to p (where p is the number of correlated patterns), we select randomly cN neurons from each of the already created patterns. While building pattern from 2 to p, we count the amount of already shared neurons between the pattern under construction and the one we are picking the shared units from. Therefore, we pick cN neurons minus the number of already shared neurons from each of the already built patterns. Finally we pick the remaining neurons to reach the target of γN active neurons from the untouched ones.

2.5.7 Experimental data

The experimental dataset of Fig. 2.5 comes from a previous publication (De Falco et al., 2016). The data was recorded from patients implanted with chronic depth electrodes in the MTL for the monitoring of epileptic seizures. Micro-wires recorded the localized neural activity; spike detection and sorting allowed to identify the activity of 4066 single neurons. During recordings, patients were shown different pictures of known people and places repeated several times. For each neuron, the stimuli eliciting a response were identified using a statistical criterion based on the modulations of firing rate during stimulus presentation compared to baseline epochs. For additional details on the dataset and data processing refer to the original publication.

2.5.8 Heterogeneous F-I curves

The F-I function of model neurons is neuron-specific and re-written as

$$\phi_i(x) = \frac{(r_{\max})_i - (r_{\min})_i}{1 + e^{-\hat{b}(x-(\hat{h}_0)_i)}} + (r_{\min})_i, \tag{2.17}$$

where the values of $(r_{\min})_i$ and $(r_{\max})_i$ are randomly sampled for each neuron from a Gaussian distribution with mean and standard deviation $\mu_{\min}$, $\sigma_{\min}$ and $\mu_{\max}$, $\sigma_{\max}$ respectively. The parameter $(\hat{h}_0)_i$ is then defined as $h_0((r_{\max})_i - (r_{\min})_i)$, where h_0 is a global constant. Finally in the firing rate equation, we re-scale the firing rates as follows:

$$r_i \to \mathrm{Max}\left[0, \frac{r_i - (r_{\min})_i}{(r_{\max})_i - (r_{\min})_i}\right]. \tag{2.18}$$

In Fig. 2.6 we choose the parameters $\mu_{\min} = 0$, $\mu_{\max} = 1$, and $\sigma_{\min} = \sigma_{\max} = 0.1$.

2.6 Supplementary information

Experimental studies measure the fraction of shared memory cell c. Such fraction of shared neurons can be interpreted as the probability that a neuron that already responds to one concept also responds to a second one. Formally, the fraction of shared neuron is the *conditional* probability $\mathrm{Prob}(x^2 = 1|x^1 = 1)$, where x^μ are binary variables that indicate is a neuron belongs to the cell assembly representing concept μ. In other words, c is the fraction of shared neurons *relative* to the fact that we are considering only neurons already belonging to one assembly. We call overlap between two cell assemblies n the *absolute* number of neurons that are shared between two cell assemblies: $n = \sum_i^N \xi_i^\mu \xi_i^\nu = c\gamma N$, where γ is the probability for a neuron to take part in a memory assembly. Note that $c\gamma$ is the *joint* probability that a neuron belongs to two memory assemblies: $P_{11} = \mathrm{Prob}(x^2 = 1, x^1 = 1)$. If two memory assemblies are not associate, than the chance they share a neuron is the probability that a neuron is picked twice to be part of a memory assembly, i.e. γ^2. We refer to this scenario saying that the number of shared neurons is at chance level $n = \gamma^2 N$ (and therefore $c = \gamma$).

From a formal point a view, the number of shared units is closely related to the Person's correlation coefficient between two memory patterns:

$$C = \frac{\mathrm{Cov}(\xi_i^1, \xi_i^2)}{\mathrm{Var}(\xi_i^\mu)} = \frac{P_{11} - \gamma^2}{\gamma(1-\gamma)}, \tag{2.19}$$

The fraction of shared neurons c can be related to the Person's correlation coefficient C uniquely: $c = n/(\gamma N) = \gamma + (1-\gamma)C$. In the following text, we show the derivation of the mean-field equations for overlapping memory engrams always referring to C simply as correlation.

2.6.1 Critical fraction of shared neurons between memory engrams

In the METHODS, we introduce the concept of critical fraction of shared neurons, as the fraction of shared neurons below which each pattern has a separated basin of attraction and above which the basin of attraction merge, or in other words, the memories are indistinguishable. As mentioned in the introduction each fraction of shared neurons corresponds to a specific correlation value, so we can equivalently speak of critical correlation between memory patterns. In order to compute numerically the critical correlation, we use the bifurcation diagram in Fig. 2.7B: the fixed points in the phase-plane are projected on the m^1-axis and their positions are plotted as C increases. From the bifurcation diagram we can extract the value $C_{\max}$ at which the single retrieval states merge with the saddle points and disappear. Thus, at $C = C_{\max}$ we have a saddle-node bifurcation.

The value of the critical correlation $C^0_{\max}$ an be calculated analytically in the limit of infinite steepness $b \to \infty$, vanishing load $\alpha = 0$, vanishing sparseness and load, $\gamma \to 0$ and $\alpha = 0$. This value matches the one extracted from the bifurcation plot in Fig. 2.7C.

In the super critical regime $C > C_{\max}$, the activation of patter one triggers, with some delay, the activation of the correlated pattern 2, as shown in Fig. 2.1D. The activation delay of pattern 2 can be quantified using the dynamical mean-field in Eq. (2.56). In Fig. 2.8A we show the evolution of the system in the phase plane, before, during and after the stimulation respectively. In Fig. 2.8B, we can see the activation of m^2 due to the super-critical correlation with m^1 as predicted by the mean-field theory. We choose to quantify the time delay between the activation of pattern one and that of pattern two, by comparing the $m^1(t)$ and $m^2(t)$ lines in Fig. 2.8B when they cross the value $m^1 = m^2 = \hat{h}_0$: the time gap within the two is delay. Indeed, the dimensionless parameter $\hat{h}_0$ marks the point where the $m^2(t)$ curve becomes steeper, or, in the phase-planes, the ghost of the fixed point corresponding to the single retrieval state.

2.6.2 Association chains

We estimate the range of correlations so that the association chains are possible, i.e. correlation should be large enough to trigger the activation of the next pattern, but not so large that the basin of attraction of the single patterns merge. The strength of the global inhibition $J_0(t)$ varies slowly between its maximum and its minimum. Note that, when $J_0(t)$ is clamped at its minimum in Fig. 2.10B,D,F,H left hand side, the double retrieving state is not present (as it was in Fig. 2.1C), this is a consequence of the introduction of the periodic inhibitory proportional to the mean neural activity. The value of correlation, $\max(C)$, that makes the two single retrieval states disappear can be read off from the bifurcation diagram in Fig. 2.9A. It sets the upper bound of the useful correlation range and strongly depends on the value of mininal global inhibitory feedback, $\min(J_0)$. When we want to estimate the smallest value of correlation,

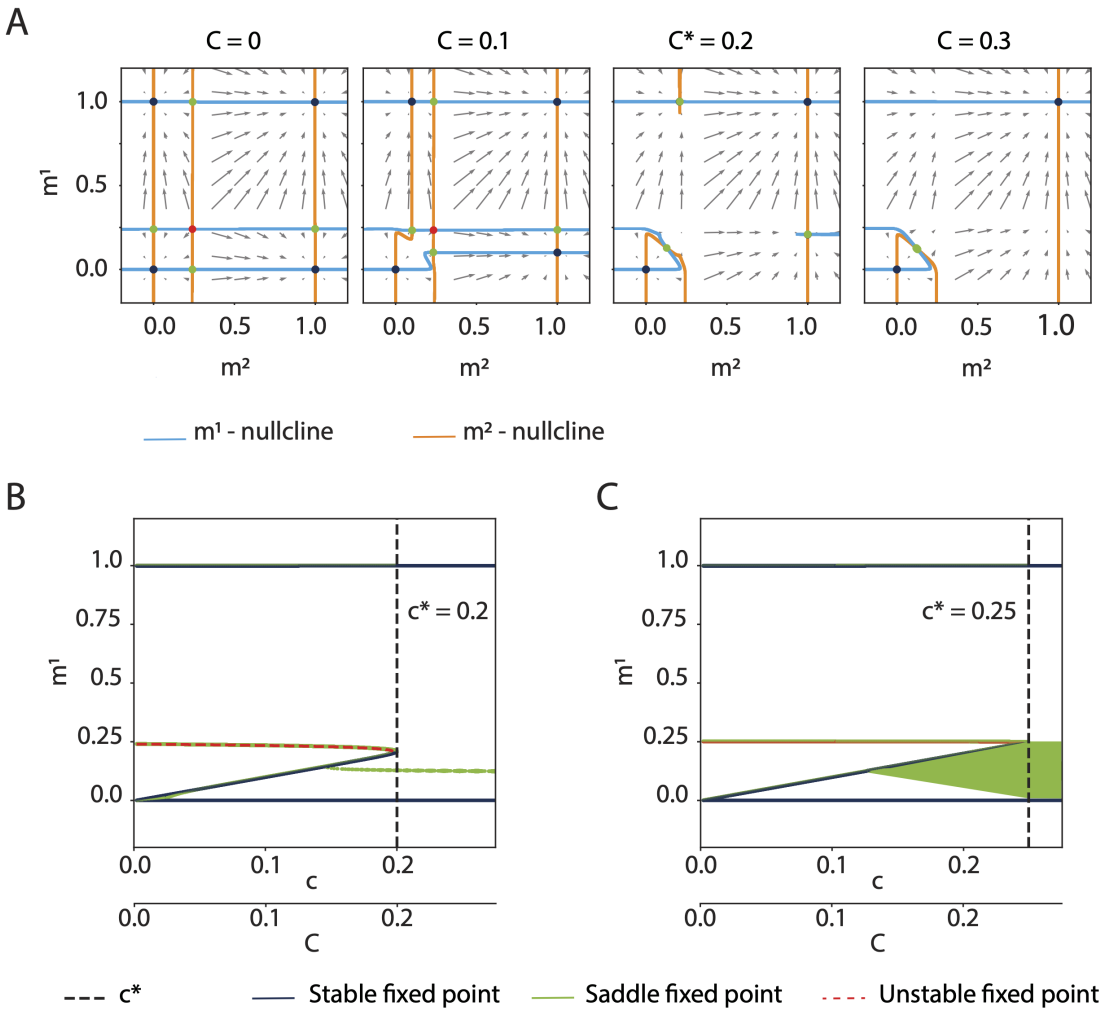


Figure 2.7 – A) Four phase-planes of the dynamics of variables m^1 and m^2 for different values of correlation C. Fixed points are color-coded by their stability: blue = stable, green = saddle and red = unstable. B) Bifurcation diagram. The projection of the fixed points position on m^1 is plotted against C. The critical correlation $C_{\max}$ is highlighted by the black dashed line. C) Same as B, but in the limit $\hat{b} \to \infty$, which leads to $C \to \hat{h}_0$. Parameters: $\gamma = 0.002$, $\hat{b} = 100$, $\hat{h}_0 = 0.25$, $\alpha = 0$.

min(C) so that hopping between attractors is possible, we consider the situation when the global inhibition is clamped at its maximum and find the minimal correlation such that the system exhibits a *transition state*. The transition state is visible in Fig. 2.10D,H right hand side, but it is not present in Fig. 2.10F, since $C = 0$. Thus, the lower bound of correlation min(C) is estimated by the left end of the stable diagonal branch of fixed points in the bifurcation diagram (Fig. 2.9B).

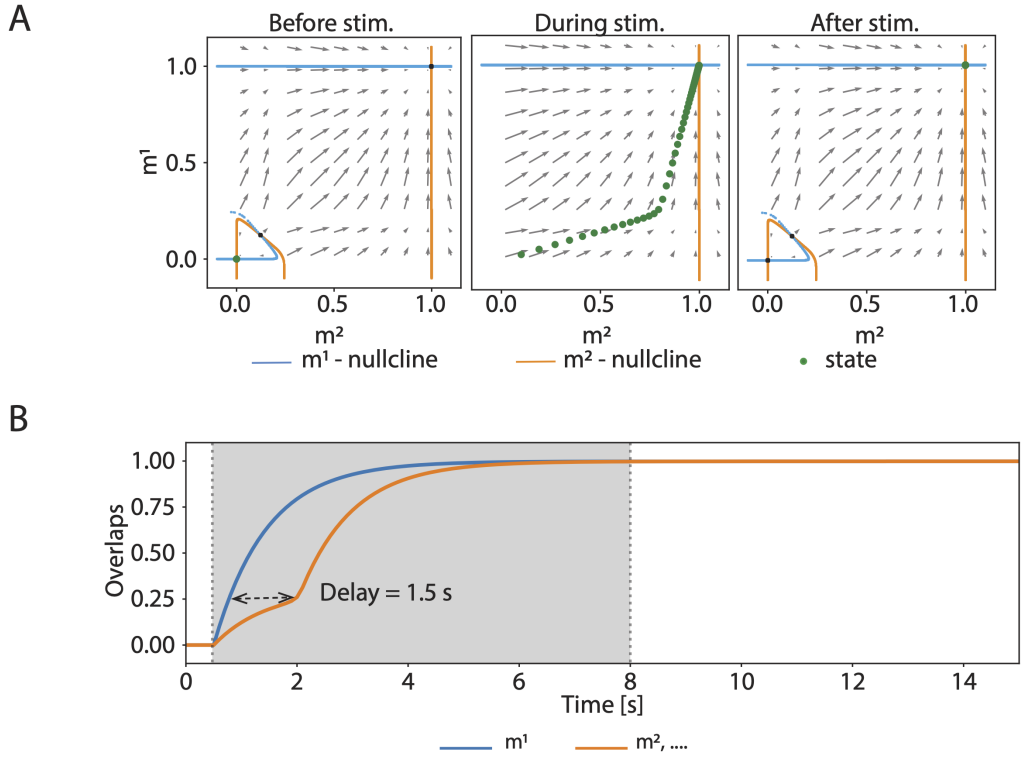


Figure 2.8 – Evolution of $m^1(t)$ and $m^2(t)$ according to the mean-field dynamics for super-critical correlation. The system is initialized in the rest state. During the stimulation period (0.5−8s) $m^1(t)$ receives external input. A) The system state is plotted in the phase-plane before, during and after stimulation respectively. B) The delay between the activation of $m^1(t)$ and $m^2(t)$ is highlighted. Parameters: $\gamma = 0.002$, $\hat{b} = 100$, $\hat{h}_0 = 0.25$, , $r_{\max} = 1$, $\hat{\tau} = \tau = 1 r_{\max}$, $\alpha = 0$, $C = 0.2$.

In Fig. 2.10 we compare the effect of different sparseness and correlations on the when hopping from one attractor to the next. For very sparse patterns, $\gamma = 0.002$, the transition is sharper (Fig. 2.10H) and we observe the same dynamics (Fig. 2.10G). While chains of associations are not possible for $C = 0$. Indeed, in Fig. 2.10B we assume $\vec{\xi}^1$ and $\vec{\xi}^2$ to be independent ($C = 0$) and the state is never able to leave the basin of attraction of pattern 1. The reason is clear from the phase-planes in Fig. 2.10A, when J_0 is at its maximum, m^1 is decreased in value until a stable fixed point, where still $m^2 = 0$: the second pattern is not even partially activated, and when the inhibition decreases, the system state falls back into the first pattern basin of attraction. At the contrary, when enough correlation is added, any activation of ξ^1 implies a partial activation of ξ^2. When J_0 reaches an high value, the system is pushed in a neutral state where both $m^1 = m^2 \sim C$. At the subsequent decrease of J_0, the system might fall in either of the two single retrieval states, but adaptation breaks the symmetry and pushes the system

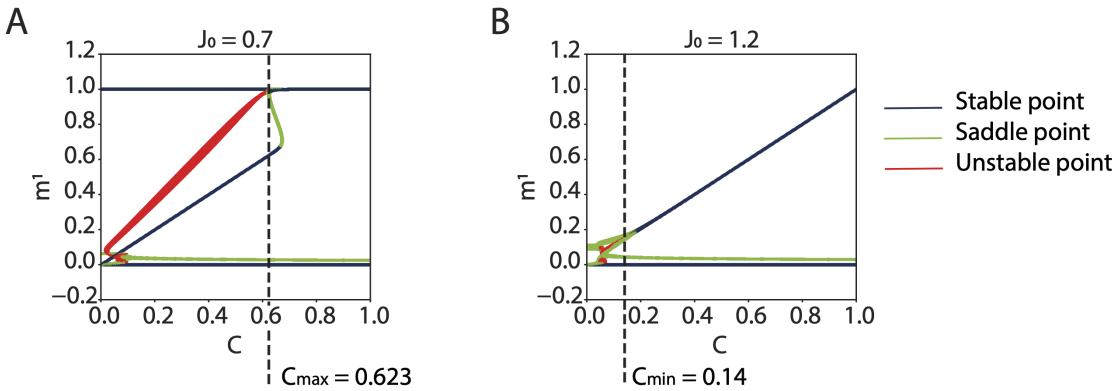
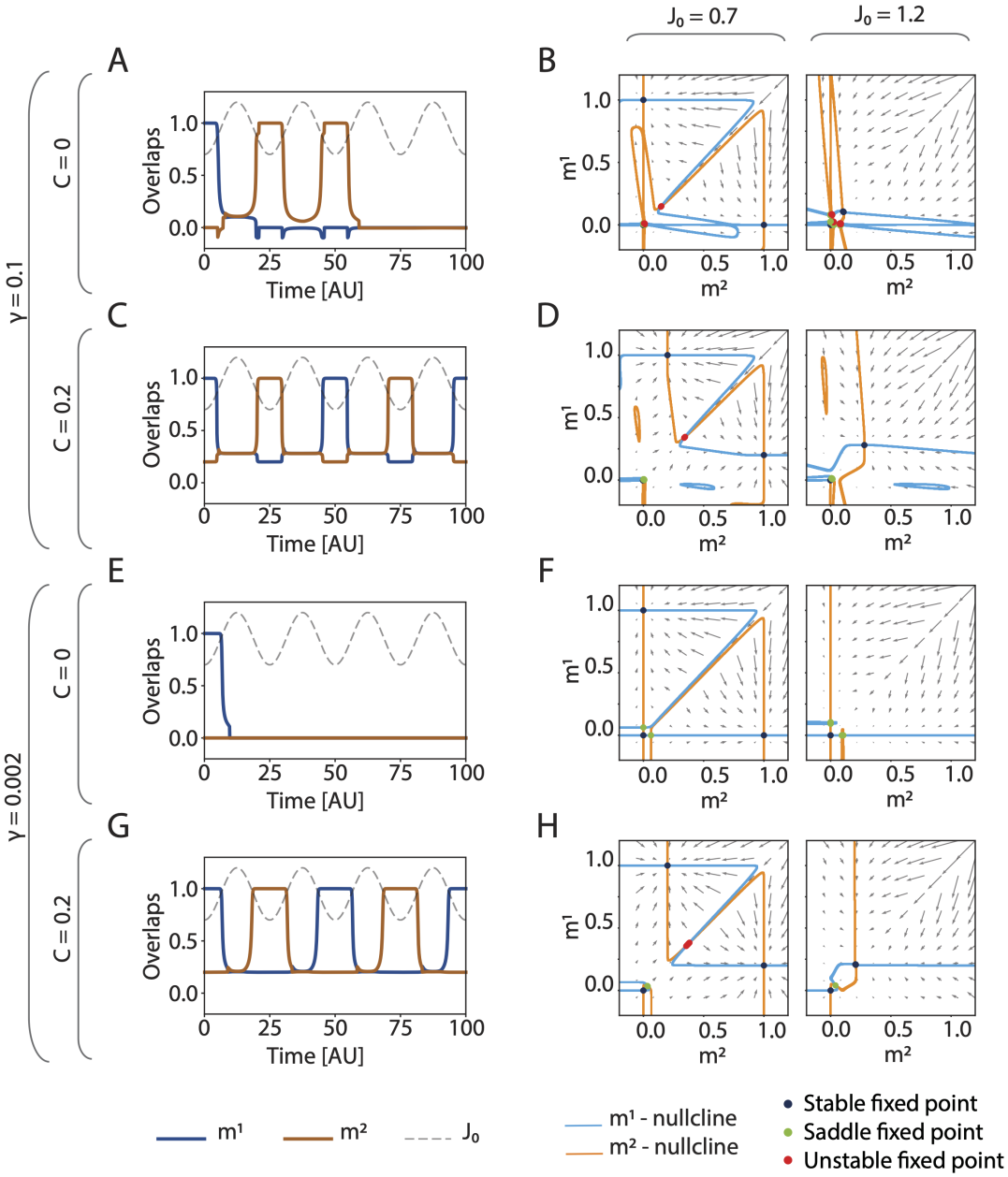


Figure 2.9 – Estimation of the correlation range in which retrieval of a chain of concepts is possible. A) Estimation of the maximum correlation, which correspond to the loss of the two single retrieval states, when J_0 is lowest. B) Estimation of the minimum correlation, which corresponds to the creation of the stable fixed point at $m^1 = m^2 > 0$, when the inhibition J_0 is at its maximum. In both A and B adaptation is frozen and $\theta = 0$. Parameters: $\gamma = 0.002$, $\alpha = 0$, $\hat{b} = 50$, $\hat{h}_0 = 0$, $\min(\hat{J}_0) = 0.7$, $\min(\hat{J}_0) = 1.2$.

towards the pattern that was not activated yet.



Caption on the next page

Figure 2.10 *(previous page)* – A) Dynamical mean-field solutions for m^1 and m^2 in the case of two independent patterns. B) Phase planes corresponding to the minimum ($J_0 = 0.7$) and maximum ($J_0 = 1.2$) value of inhibition in the case of two independent patterns. C,D) Same as A and B, but for correlated patterns $C = 0.2$. Parameters in A - D: $\gamma = 0.1$, $\alpha = 0$, $b = 100$. E, F) Same as C and D but in the low activity regime and for independent patterns. G, H) Same as C and D but in the low activity regime and for correlated patterns. Parameters in E - H: $\gamma = 0.002$, $\alpha = 0$, $b = 100$, $\tau_\theta = 45$, $T = 0.015$, $T_{J_0} = 25$. For the dynamics: resolution = 200, factor =1. For the phase-planes: resolution = 1000, factor = 1, upper bound = 1.2, lower bound = -0.2.

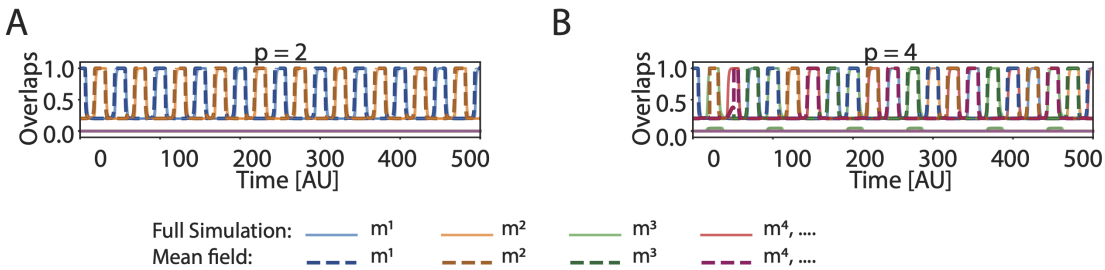


Figure 2.11 – Retrieval dynamics in the presence of adaptation according to the mean-field equations (dashed lines), and comparison with Fig. 2.3B (shaded solid lines) A) Only two patterns are correlated. B) Four patterns are correlated.Fig Parameters: $N = 10^4$, $P = 16$ in full network simulations and $\alpha = 0$ in mean-field. $\gamma = 0.002$, $\tau_\theta = 45$, $T = 0.015$, $T_{J_0} = 25$ in both.

In Fig. 2.11 we compare full network simulation with dynamical mean-field for $p = 2$ and $p = 4$. The mean-field and the full simulation match .

2.6.3 Parameters choice

We have discussed that the critical correlation between patterns depends on two dimensionless parameters: the rescaled threshold $\hat{h}_0 = h_0/(Ar_{\max})$ and the rescaled steepness $\hat{b} = Ar_{\max}b$. While these parameters have so far not being estimated for human Hippocampus, in (Pereira and Brunel, 2018) where they were able to fully characterize the gain function of pyramidal neurons of the macaque's IT cortex. We though it would be interesting, to compute the critical correlation for this set of parameters. First, in (Pereira and Brunel, 2018) they input is filtered by the combination of two Sigmoids, we have chosen the parameters such that our single Sigmoidal gain function fits the combination of Sigmoids. Moreover, in (Pereira and Brunel, 2018) they assume patterns to follow a Gaussian distribution $N(0,1)$, and our theory assumed binary patterns. To match, we have chosen $\gamma = N(0,1)(h_0)$, where h_0 is estimated from the fitted gain function. We obtained $\gamma = 0.0375$. In Fig. 2.12A we show the phase-plane for $C = 0$ (or equivalently $c = \gamma$) and in Fig. 2.12B we provide the bifurcation diagram from which the

critical correlation is extracted (equivalent to that in Fig. 2.7B).

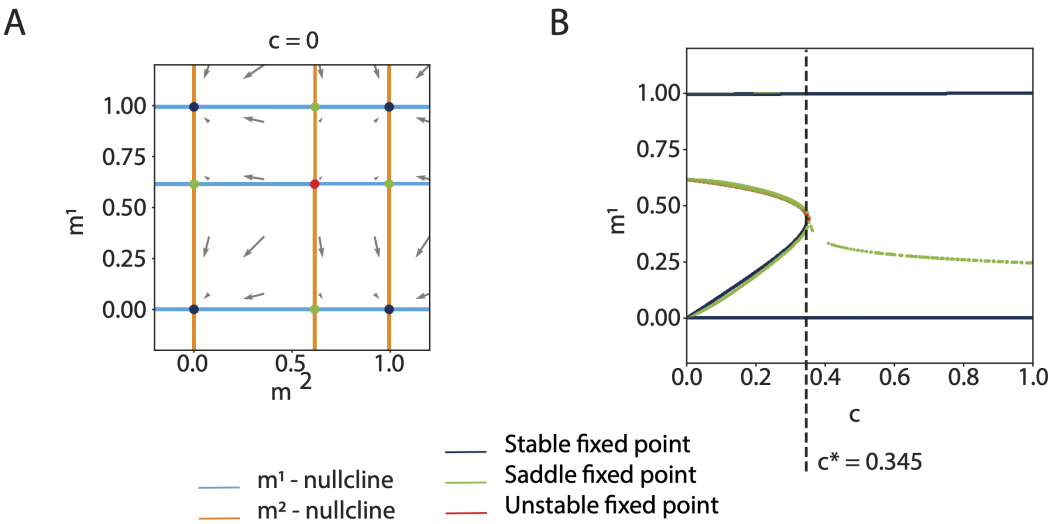


Figure 2.12 – Equivalent of Fig 2.7 but with the parameters extracted from (Pereira and Brunel, 2018). In A and B the gain function parameters are taken as those of function ϕ in (Pereira and Brunel, 2018): $A = 3.55$, $r_{\max} = 76.2$, $b = 0.82$, $h_0 = 2.46$. On the other hand, in C and D I estimated the parameters of a Sigmoid function that fits the function $f(\phi)$ in (Pereira and Brunel, 2018) as follows: $A = 3.55$, $r_{\max} = 0.83$, $b = 4.35$, $h_0 = 1.7$. In all plots $\gamma = 0.001$. A and C) The phase-plane for $c = 0$ shows the position of fixed points. B and D) Bifurcation diagram and critical fraction of shared neurons according to different parameter choices.

Methods

In this section, we will present the details and the derivation of both the full network simulations and the mean-field model. We partially repeat information present in the main text, with the goal of providing here a complete and self consistent presentation of the methods used.

2.6.4 Model without adaptation and global feedback

In the figures relative to the Section "Critical fraction of shared neurons between memory engrams" of the main paper and the SI, we consider an attractor neural network of N rate units r_i, in which P binary memory patterns are stored $\{\xi_i^\mu\}_{i=1,..,N}^{\mu=1,..,P}$. Each neuron follows the Wilson-Cowan dynamics (Wilson and Cowan, 1972):

$$\tau \frac{dr_i}{dt} = -r_i + \phi(h_i), \tag{2.20}$$

where h_i is the input term

$$h_i = \sum_{j \neq i}^{N} w_{ij} r_j + I_i(t) \tag{2.21}$$

which include the sum over the recurrent input from other neurons and a possible external input $I_i(t)$. The input is filtered with gain function ϕ, which is chosen to be a sigmoid:

$$\phi(h) = \frac{r_{\max}}{1 + e^{-b(h-h_0)}}. \tag{2.22}$$

The parameters that define the gain function can be interpreted as follows: $r_{\max}$ is the maximal firing rate, b is the steepness of the gain function and h_0 is the bias which is commonly interpreted as firing threshold. While the firing threshold is actually equal to h_0 for $b \to \infty$, at finite b the firing threshold is effectively a bit smaller. The connection matrix, w_{ij}, contains the synaptic weights between neurons i and j is defined as in Eq. (2.5), which we repeat for convenience (Pereira and Brunel, 2018)

$$w_{ij} = \frac{A}{N\gamma(1-\gamma)} \sum_{\mu}^{P} \left(\xi_j^\mu - \gamma\right)\left(\xi_i^\mu - \gamma\right) \tag{2.23}$$

where the constant A can be interpret as “connection strength". In order for the weights to be $\langle w_{ij} \rangle = 0$, we subtract the mean activity of patterns $< \xi_i^\mu >= \gamma$. In order to measure how close is the network state to recalling a pattern, we defined the overlaps m^μ which are a measure of the correlation between the network state and the store patterns:

$$m^\mu(t) = \frac{1}{N\gamma(1-\gamma) r_{\max}} \sum_{j=1}^{N} \left(\xi_j^\mu - \gamma\right) r_j(t). \tag{2.24}$$

2.6.5 Model with adaptation and global inhibitory feedback

For Figs. of sections “Association chains”, we added adaptation and a global periodic inhibitory to the model, following previous works (Romani et al., 2013; Recanatesi et al., 2015, 2017; Naim et al., 2020), as introduced in Eq. (2.7):

$$h_i = \sum_{j=1}^{N} w_{ij} r_j(t) - \theta_i(t) - \frac{J_0(t)}{\gamma N} \sum_{j=1}^{N} r_j(t), \tag{2.25}$$

The second term $-\theta_i(t)$, is a *local* neuron-specific feedback that models firing-rate adaptation as follows (as in Eq. (2.8)):

$$\tau_\theta \frac{d\theta_i}{dt} = -\theta_i + D_\theta r_i, \tag{2.26}$$

where τ_θ is the adaptation time constant and D_θ determines the strength of adaptation.

The term $-J_0(t)/(\gamma N)\sum_{j=1}^{N} r_j(t)$ is the *global* inhibition which provides a clock signal that triggers the transitions betweeen attractors. The strength of the global feedback, $J_0(t)$, is modulated periodically in time:

$$J_0 = \frac{1}{2}(\max(J_0) - \min(J_0))\sin\left(t\frac{(2\pi)}{T_{J_0}} - \pi/2\right) + \frac{1}{2}(\max(J_0) + \min(J_0)) \tag{2.27}$$

2.6.6 Non-dimensionalization of the model

The calculations below are considerably simplified if the model is made dimension-free. We take into account that $r_{\max}$ has units of $1/\text{time}$ and the parameter A has units of $\text{current}\cdot\text{time}$ and measure in the following time in units of r_{max}^{-1} and current input in units of Ar_{max}.

Model without adaptation and global feedback. Using the dimensionless quantities

$$\hat{h}_i = \frac{h_i}{Ar_{\max}}, \quad \hat{h}_0 = \frac{h_0}{Ar_{\max}}, \quad \hat{b} = bAr_{\max}, \quad \hat{r}_i = \frac{r_i}{r_{\max}} \tag{2.28}$$

$$\hat{w}_{ij} = \frac{w_{ij}}{A}, \quad \hat{\tau} = \tau r_{\max}, \quad \hat{I}_i(t) = \frac{I_i(t)}{Ar_{\max}}, \tag{2.29}$$

the dimension-free model without adaptation reads

$$\hat{\tau}\frac{d\hat{r}_i}{d\hat{t}} = -\hat{r}_i + \hat{\phi}(\hat{h}_i), \text{ with } \hat{h}_i = \sum_{j=1}^{N} \hat{w}_{ij}\hat{r}_j + \hat{I}_i \tag{2.30}$$

with a transfer function $\hat{\phi}(\hat{h}) = 1/\{1 + \exp[-\hat{b}(\hat{h} - \hat{h}_0)]\}$.

Model with adaptation and global feedback. Introduction of further dimensionless quantities

$$\hat{\tau}_\theta = \tau_\theta r_{\max}, \quad \hat{\theta} = \frac{\theta}{Ar_{max}}, \quad \hat{D}_\theta = \frac{D_\theta}{A}, \quad \hat{J}_0(\hat{t}) = \frac{J_0(\hat{t}/r_{\max})}{A} \tag{2.31}$$

leads to the non-dimensional model with adaptation

$$\hat{\tau}\frac{d\hat{r}_i}{d\hat{t}} = -\hat{r}_i + \hat{\phi}(\hat{h}_i), \tag{2.32}$$

$$\hat{\tau}_\theta\frac{d\hat{\theta}_i}{d\hat{t}} = -\hat{\theta}_i + \hat{D}_\theta\hat{r}_i \tag{2.33}$$

with input

$$\hat{h}_i = \sum_{j=1}^{N} \hat{w}_{ij}\hat{r}_j - \hat{\theta}_i - \frac{\hat{J}_0(t)}{\gamma N}\sum_{j=1}^{N} \hat{r}_j(t) + \hat{I}_i. \tag{2.34}$$

To lighten the notation, we will omit the hats on top of the dimensionless quantities for the rest of this paper, while we keep the strict distinction between quantities with and without hat in the Results section.

2.6.7 Review of attractor theory

Starting from the overlap definition Eq. (2.24), we can write equations for the overlaps variables. We follow an approach well known in literature (Tsodyks and Feigel'man, 1988; Shiino and Fukai, 1992) and focus on the model without adaptation and global feedback. The only time dependent variable in the definition of m^μ is the network state r_j, so that:

$$\hat{\tau}\frac{dm^\mu}{dt} = \frac{1}{N\gamma(1-\gamma)}\sum_{j=1}^{N}\left(\xi_j^\mu - \gamma\right)\tau\frac{d\hat{r}_j}{dt}. \tag{2.35}$$

By inserting the expression for the single neuron dynamics Eq. (2.20) and recognizing the overlap definition Eq. (2.24), we obtain:

$$\hat{\tau}\frac{dm^\mu}{dt} = -m^\mu + \bar{m}^\mu. \tag{2.36}$$

with

$$\bar{m}^\mu = \frac{1}{N\gamma(1-\gamma)}\sum_{j=1}^{N}\left(\xi_j^\mu - \gamma\right)\hat{\phi}\left(\hat{h}_j\right). \tag{2.37}$$

The input term h_j, Eq.2.21, can also be re-written as a function of the overlaps $m^1, ..., m^P$, by using the definition of the weights w_{ij}, Eq. (2.5), and that of the overlaps:

$$\hat{h}_i = \sum_{\mu=1}^{P}\left(\xi_i^\mu - \gamma\right)m^\mu + I_i. \tag{2.38}$$

In what follows we are interested in finding stable solutions of Eq. (2.36), $m^\mu = \bar{m}^\mu$. We are interested in studying the pattern retrieval and, without loss of generality, we consider the retrieval of pattern 1. To this end, we assume that among all m^μ, only m^1 is significantly larger than zero. Under this assumption we can re-write the input term h_i isolating the contribution from m^1

$$\hat{h}_i = \left(\xi_i^1 - \gamma\right)m^1 + \sum_{\mu=2}^{P}\left(\xi_i^\mu - \gamma\right)m^\mu + I_i. \tag{2.39}$$

We call the patterns that are not recalled "background patterns"; in our case, these are all patterns for which $\mu \geq 2$. The second term on the r.h.s of Eq. 2.39 represents the contribution from the background patterns causing some degree of heterogeneity of the input potential for

neurons with the same selectivity to pattern 1. For large P, this heterogeneity can be captured by replacing the term $\sum_{\mu\neq 1}^{P}\left(\xi_i^\mu-\gamma\right)m^\mu$ by a Gaussian random variable with mean zero and variance

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}\sum_{\mu=2}^{P}\sum_{\nu=2}^{P}\left(\xi_i^\mu-\gamma\right)\left(\xi_i^\nu-\gamma\right)m^\mu m^\nu=\gamma(1-\gamma)\sum_{\nu=2}^{P}(m^\nu)^2 \tag{2.40}$$

To obtain the result in Eq. (2.40), we used the assumption that patterns ξ_i^μ and ξ_i^ν are uncorrelated, and the fact that only the term for $\mu\neq\nu$ survives, in fact, $\left\langle(\xi_i^\mu\xi_i^\nu+\gamma^2-\gamma\xi_i^\nu-\gamma\xi_i^\mu)\right\rangle_i=\delta_{\mu\nu}\gamma(1-\gamma)$. Here and in the following, the brackets $\langle x_i\rangle$ of a variable x_i denotes the population average $\langle x_i\rangle=\frac{1}{N}\sum_{i=1}^{N}x_i$. In the next passages we compute $(m^\mu)^2,\mu\neq 1$, in the large network limit $N\to\infty$. We expand Eq. (2.37) around $m^\mu=0$ up to first order in m^μ,

$$\bar{m}^\mu\approx\frac{1}{\gamma(1-\gamma)N}\sum_{j=1}^{N}\left[\left(\xi_j^\mu-\gamma\right)\hat{\phi}(\hat{h}_j)\Big|_{m^\mu=0}+\left(\xi_j^\mu-\gamma\right)^2m^\mu\hat{\phi}'(\hat{h}_j)\Big|_{m^\mu=0}\right] \tag{2.41}$$

and collect on the left hand side the terms multiplied by m^μ,

$$m^\mu\left(1-\sum_{j=1}^{N}\frac{\left(\xi_j^\mu-\gamma\right)^2\hat{\phi}'(\hat{h}_j)}{\gamma(1-\gamma)N}\right)=\frac{1}{\gamma(1-\gamma)N}\sum_{j=1}^{N}\left(\xi_j^\mu-\gamma\right)\hat{\phi}(\hat{h}_j). \tag{2.42}$$

On the left hand side of the last expression, we can make some simplification, considering that ξ_j^μ is uncorrelated with $\phi'(h_j)$, in the $N\to\infty$ limit

$$\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\frac{\left(\xi_j^\mu-\gamma\right)^2\hat{\phi}'(\hat{h}_j)}{\gamma(1-\gamma)}=\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\hat{\phi}'(\hat{h}_j)=\left\langle\hat{\phi}'(\hat{h}_i)\right\rangle_i. \tag{2.43}$$

We can therefore define the quantity $q:=\left\langle\hat{\phi}'(\hat{h}_i)\right\rangle_i$ as the expectation of $\hat{\phi}'(\hat{h}_i)$ over neurons. As a consequence, m^μ can be written as

$$m^\mu=\frac{1}{\gamma(1-\gamma)(1-q)N}\sum_{j=1}^{N}\left(\xi_j^\mu-\gamma\right)\hat{\phi}(\hat{h}_j). \tag{2.44}$$

We can finally compute the squared of m^ν, using Eq. (2.44):

$$(m^\nu)^2=\frac{1}{\gamma^2(1-\gamma)^2(1-q)^2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\xi_i^\mu-\gamma\right)\left(\xi_j^\mu-\gamma\right)\hat{\phi}(\hat{h}_i)\hat{\phi}(\hat{h}_j), \tag{2.45}$$

$$=\frac{1}{\gamma^2(1-\gamma)^2(1-q)^2N^2}\sum_{i=1}^{N}\left(\xi_i^\mu-\gamma\right)^2[\hat{\phi}(\hat{h}_i)]^2, \tag{2.46}$$

$$=\frac{p}{\gamma(1-\gamma)(1-q)^2N}. \tag{2.47}$$

where $p := \langle \hat{\phi}^2(\hat{h}_i) \rangle_i$. Similarly as in Eq. (2.40), we used that in the double sum $\sum_{i=1}^N \sum_{j=1}^N$, only the term $i = j$ survives. The standard deviation of the neuron-to-neuron variability (heterogeneity), Eq. (2.40), is

$$\sigma = \sqrt{\alpha R}, \qquad R := \frac{p}{(1-q)^2}. \tag{2.48}$$

Finally, for $\mu = 1$, we have an a mean-field expression for m^1:

$$m^1 = \frac{1}{\gamma(1-\gamma)} \left\langle \left(\xi_i^1 - \gamma\right) \hat{\phi}(\hat{h}_i) \right\rangle_i \tag{2.49}$$

with input potentials

$$\hat{h}_i = \left(\xi_i^1 - \gamma\right) m^1 + \sqrt{\alpha R} Z_i + I(\xi_i^1) \tag{2.50}$$

and Gaussian random variables $Z_i \sim N(0,1)$.

The population averages $\langle \cdot \rangle$ can be treated as expectations over the independent random variables ξ_i^1 and Z_i. On the one hand, ξ_i^1 is a Bernoulli variable such that $\xi_i^1 = 1$ with probability $P_1 = \gamma$ and $\xi_i^1 = 0$ with probability $P_0 = 1 - \gamma$. On the other hand, Z_i is a standard normal random variable with probability density $p_Z(z) = \exp(-z^2/2)/\sqrt{2\pi}$. We can therefore rewrite the population average in Eq. (2.49) explicitly resulting in

$$m^1 = \frac{1}{\gamma(1-\gamma)} \sum_{k=0,1} P_k \left(k - \gamma\right) \int \hat{\phi}(\hat{h}_k(m^1, z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} \tag{2.51a}$$

where

$$\hat{h}_k(m,z) = \left(k - \gamma\right) m + \sqrt{\alpha R(m)} z + I_k. \tag{2.51b}$$

$$R(m) = \frac{p(m)}{[1 - q(m)]^2} \tag{2.51c}$$

$$q(m) = \sum_{k=0,1} P_k \int \hat{\phi}'(\hat{h}_k(m,z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} \tag{2.51d}$$

$$p(m) = \sum_{x^1=0,1} P_k \int \hat{\phi}^2(\hat{h}_k(m,z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}}. \tag{2.51e}$$

where is x^1 is a binomial random variable which can assume values $1, 0$ with probability $P(x^1)$, where $P(0) = P_0 = (x^1 - \gamma)$ and $P(1) = P_1 = \gamma$.

2.6.8 Dynamical mean-field equations

In the case we assume there are no background patterns, it's possible to write a dynamcial mean field equation for the retrieving of pattern 1:

$$\hat{\tau}\frac{dm^1}{dt} = -m^1 + \frac{1}{\gamma(1-\gamma)}\left\langle\left(\xi_i^1-\gamma\right)\hat{\phi}(\hat{h}_i)\right\rangle_i \tag{2.52}$$

with input potential

$$\hat{h}_i = \left(\xi_i^1-\gamma\right)m^1 + I(\xi_i^1). \tag{2.53}$$

This approach is not directly extensible to the presence of background patterns, since in passage Eq. (2.41) we use the assumption that overlaps m^μ reached their equilibrium value $\bar{m}^\mu$. However, in what follows we write dynamical mean field equations were the overlaps with retrieved patterns evolve over time and those with background patters are always at their equilibrium value. While this is not formally correct, it gives results in agreement with full network simulations (Fig. 2.2C). In other words the mean-field dynamics in Fig. 2.2C is correct before the stimulus onset and after the system has retrieved pattern 1, but not formally correct in the transient. Moreover, we argued in the discussion that assuming a small or even negligible network load $\alpha \geq 0$ is a bio-plausible assumption for the human MTL. Dynamical mean-field equation are formally exact for $\alpha = 0$.

2.6.9 Mean-field equations for two correlated patterns

Taking the results derived in the previous section as a starting point, we now allow two of the stored patterns to be correlated. Without loss of generality, we choose patterns $\vec{\xi}^1$ and $\vec{\xi}^2$ to be correlated, while all other $P-2$ patterns are independent. The correlation between the two patterns is defined as the Pearson correlation coefficient (covariance/variance), in Eq. (2.9), which we rewrite here for convenience:

$$C = \frac{\mathrm{Cov}(\xi_i^1,\xi_i^2)}{\mathrm{Var}(\xi_i^\mu)} = \frac{P_{11}-\gamma^2}{\gamma(1-\gamma)}, \tag{2.54}$$

where $P_{11} = P(\xi_i^1 = 1, \xi_i^2 = 1) = \langle \xi_i^1 \cdot \xi_i^2 \rangle$ is the joint probability of a neuron to be selective to both patterns. In order to generate correlated patterns with mean activity $\langle \xi_i^1 \rangle_i = \langle \xi_i^2 \rangle_i = \gamma$ and correlation coefficient C, we use the indicator neuron model with $\epsilon = \Omega$ described later in the SI. However it is important to note that, if we consider only two correlated patterns, it does not matter which algorithm we choose, as it will be clear in the "Relation to experiments" session.

As in the previous section we are interested in the retrieval dynamics of pattern $\vec{\xi}^1$. However, given the correlation with pattern $\vec{\xi}^2$, we cannot neglect the overlap of the network state with

$\vec{\xi}^2$. The derivation of the system of mean-field equations in case two correlated pattern case in Eq. (2.11) is analogous to that described in the section above. The input term h_i has now two non-negligible terms, both from $\vec{\xi}^1$ and $\vec{\xi}^2$:

$$\hat{h}_i(m^1, m^2) = \left(\xi_i^1 - \gamma\right) m^1 + \left(\xi_i^2 - \gamma\right) m^2 + \sqrt{\alpha R(m^1, m^2)} Z_i + I_i \tag{2.55}$$

$$\hat{\tau} \frac{dm^1}{dt} = -m^1 + \frac{1}{\gamma(1-\gamma)} \left\langle \left(\xi_i^1 - \gamma\right) \hat{\phi}(\hat{h}_i) \right\rangle_i \tag{2.56a}$$

$$\hat{\tau} \frac{dm^2}{dt} = -m^2 + \frac{1}{\gamma(1-\gamma)} \left\langle \left(\xi_i^2 - \gamma\right) \hat{\phi}(\hat{h}_i) \right\rangle_i \tag{2.56b}$$

$$q = \left\langle \hat{\phi}'(\hat{h}_i) \right\rangle_i \tag{2.56c}$$

$$p = \left\langle \hat{\phi}^2(\hat{h}_i) \right\rangle_i \tag{2.56d}$$

$$R(m^1, m^2) = \frac{p}{(1-q)^2} \tag{2.56e}$$

Analogously to Eq. (2.51) we can compute the population averages in Eq. (2.56) explicitly leading to

$$\hat{\tau} \frac{dm^1}{dt} = -m^1 + F_1(m^1, m^2) \tag{2.57a}$$

$$\hat{\tau} \frac{dm^2}{dt} = -m^2 + F_2(m^1, m^2) \tag{2.57b}$$

where the nonlinear functions F_1 and F_2 are given by ($\mu = 1, 2$)

$$F_\mu(m^1, m^2) = \sum_{x^1=0,1} \sum_{x^2=0,1} \frac{x_\mu - \gamma}{\gamma(1-\gamma)} P_{x^1 x^2} \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \hat{\phi}(\hat{h}_{x^1 x^2}(m^1, m^2, z)) \tag{2.57c}$$

with

$$\hat{h}_{x^1 x^2}(m^1, m^2, z) = \sum_{\nu=1,2} \left(x_\nu - \gamma\right) m^\nu + I(x^1, x^2) + \sqrt{\alpha R_h(m^1, m^2)} z. \tag{2.57d}$$

This function can be interpreted as the mean-field input potential of a neuron with selectivity $\xi_i^1 = x^1$ and $\xi_i^2 = x^2$, background variability $Z_i = z$, in the case when the network has overlap m^1 and m^2 with patterns 1 and 2, respectively. The last term in Eq. (2.57d) captures the influence of background patterns on the mean-field dynamics of $m^1(t)$ and $m^2(t)$. This influence is quantified by the function $R_h(m^1, m^2)$ representing the mean squared overlap of the system with the background patterns $\mu = 3, \ldots, P$.

We used a subscript h for $R_h(m^1, m^2)$ to indicate that R depends functionally on the mean-field potential $\hat{h}_{x^1x^2}(m^1, m^2, z)$. This functional is given by

$$R_h(m^1, m^2) = \frac{p}{(1-q)^2} \tag{2.57e}$$

$$q = \sum_{x^1=0,1} \sum_{x^2=0,1} P_{x^1,x^2} \int \hat{\phi}'(\hat{h}_{x^1x^2}(m^1, m^2, z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} \tag{2.57f}$$

$$p = \sum_{x^1=0,1} \sum_{x^2=0,1} P_{x^1,x^2} \int \hat{\phi}^2(\hat{h}_{x^1x^2}(m^1, m^2, z)) e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}}. \tag{2.57g}$$

The mean-field input potentials $\hat{h}_{x^1x^2}(m^1, m^2, z)$, $x^1, x^2 \in \{0, 1\}$, needed in Eq. (2.57c) are obtained from the self-consistent solution of the functional equations (2.57d)–(2.57g), details are in Section “Numerical solutions".

Equations (2.57) simplify significantly for $\alpha = 0$, which is the papramter choice of most figures, so it is worth writing explicitly the m^1 and m^2 dynamics in the case of negligible load:

$$\hat{\tau}\frac{dm^1}{dt} = -m^1 + \frac{1}{\gamma(1-\gamma)} \{ P_{11}(1-\gamma)\hat{\phi}\left[(1-\gamma)(m^1+m^2) + I_1 + I_2\right] + \tag{2.58a}$$

$$P_{10}(1-\gamma)\hat{\phi}\left[(1-\gamma)m^1 - \gamma m^2 + I_1\right] - \tag{2.58b}$$

$$P_{01}\gamma\hat{\phi}\left[-\gamma m^1 + (1-\gamma)m^2 + I_2\right] - P_{00}\gamma\hat{\phi}\left[-\gamma(m^1+m^2)\right]\}, \tag{2.58c}$$

$$\hat{\tau}\frac{dm^2}{dt} = -m^2 + \frac{1}{\gamma(1-\gamma)} \{ P_{11}(1-\gamma)\hat{\phi}\left[(1-\gamma)(m^1+m^2) + I_1 + I_2\right] - \tag{2.58d}$$

$$P_{10}\gamma\hat{\phi}\left[-\gamma m^1 + (1-\gamma)m^2 + I_1\right] + \tag{2.58e}$$

$$P_{01}(1-\gamma)\hat{\phi}\left[(1-\gamma)m^1 - \gamma m^2 + I_2\right] - P_{00}\gamma\hat{\phi}\left[-\gamma(m^1+m^2)\right]\}. \tag{2.58f}$$

$$\tag{2.58g}$$

The same procedure can be generalized to generate several correlate binary patterns, as in Fig.2.2B. The generalization is straightforward, we can re-write the system in Eq. (2.57) with one dynamical equations for each correlated pattern and add the relative terms in the input $\hat{h}(x^1, .., x^\mu, z)$. Finally we need the joint probabilities P_{x^1,x^2,x^3} and P_{x^1,x^2,x^3,x^4}. The general formula for the joint probability is given in Eq. (2.113). For instance, for three correlated

patterns, the mean-field dynamics analogue to Eq. (2.57) is given by

$$\hat{\tau}\frac{dm^1}{dt} = -m^1 + \frac{1}{\gamma(1-\gamma)}\left\langle\left(\xi_i^1-\gamma\right)\hat{\phi}(\hat{h}_i)\right\rangle_i \tag{2.59a}$$

$$\hat{\tau}\frac{dm^2}{dt} = -m^2 + \frac{1}{\gamma(1-\gamma)}\left\langle\left(\xi_i^2-\gamma\right)\hat{\phi}(\hat{h}_i)\right\rangle_i \tag{2.59b}$$

$$\hat{\tau}\frac{dm^3}{dt} = -m^3 + \frac{1}{\gamma(1-\gamma)}\left\langle\left(\xi_i^2-\gamma\right)\hat{\phi}(\hat{h}_i)\right\rangle_i \tag{2.59c}$$

$$q = \left\langle\hat{\phi}'(\hat{h}_i)\right\rangle_i \tag{2.59d}$$

$$p = \left\langle\hat{\phi}^2(\hat{h}_i)\right\rangle_i \tag{2.59e}$$

$$R(m^1,m^2,m^3) = \frac{p}{(1-q)^2} \tag{2.59f}$$

where

$$\hat{h}_i(m^1,m^2,m^3) = \left(\xi_i^1-\gamma\right)m^1 + \left(\xi_i^2-\gamma\right)m^2 + \left(\xi_i^3-\gamma\right)m^3 + \sqrt{\alpha R(m^1,m^2,m^3)}Z_i + I_i. \tag{2.60}$$

2.6.10 Excluding self-interaction

In Section "Review: mean-field equations for independent patterns" we show the derivation of the mean-field equations for the retrieval of one pattern in an attractor neural network. However in literature often times some corrections are added in order to keep into account that neurons do not self-interact. This corrections are to both make the network more bio-plausible (indeed it would be hard to justify such a self-interaction biologically) and to avoid the creation of local minima around the attractors corresponding to the stored patterns. Indeed, not excluding the term $w_{ii}r_j$ into the input h_i allows some units to self-stabilize. The effect of excluding the self interaction term on input terms in Eqs. (2.51a) and (2.57d) is capture by a the correction term (Shiino and Fukai, 1992):

$$\frac{q\alpha\hat{\phi}(\hat{h})}{\left(1-q\right)}. \tag{2.61}$$

For example, Eq. (2.57d) would become

$$\hat{h}(x^1,x^2,z) = \left(x^1-\gamma\right)m^1 + \left(x^2-\gamma\right)m^2 + \frac{q\alpha\phi(h)}{\left(1-q\right)} + \sqrt{\alpha r}z + I(x^1,x^2). \tag{2.62}$$

The input term in Eq. (2.62), is used in Fig. 2.2B, left hand side.

2.6.11 Correlation between background patterns

In this section, we provide the derivation needed for Fig. 2.2B, left hand side. We let the background patterns to be correlated and study how the mean-field equations for the system change.

To start with, let's suppose that each pattern is correlated with just one other, so pattern ξ^{μ} is only correlated with $\xi^{\mu'}$. What changes, compared to the derivation in Section "Review: mean-field equations for independent patterns" is the variance of the heterogeneity term in Eq. (2.40) , $\langle\sigma^2\rangle = \langle\sum_{\mu\nu}\left(\xi_i^{\mu}\xi_i^{\nu} - \gamma\xi_i^{\nu} - \gamma\xi_i^{\mu} + \gamma^2\right)\rangle\, m^{\mu}m^{\nu}$, where patterns are pair-wise correlated. We obtain:

$$\left\langle \sum_{\mu\nu}\left(\xi_i^{\mu}\xi_i^{\nu} - \gamma\xi_i^{\nu} - \gamma\xi_i^{\mu} + \gamma^2\right)\right\rangle = \delta_{\mu\nu}\gamma(1-\gamma) + \delta_{\mu,\mu'}(P_{11} - \gamma^2), \tag{2.63}$$

where the second term at the right hand side captures the effect of correlation. Background patterns can still be approximated by a Gaussian variable in the large network limit, in this case with variance:

$$\langle\sigma^2\rangle = \sum_{\nu\neq 1}^{P}\left[\gamma(1-\gamma)(m^{\nu})^2 + (P_{11} - \gamma^2)m^{\nu}m^{\nu'}\right]. \tag{2.64}$$

In order to compute Eq. (2.64), we need to derive $(m^{\nu})^2$ and $m^{\nu}m^{\nu'}$. In what follows, we use the same definition of q and p as in Eq. (2.51). Let's start from writing m^{ν} at the first order Taylor expansion for m^{ν} and $m^{\nu'}$ both small:

$$\begin{aligned} m^{\nu} \sim & \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu} - \gamma\right)\hat{\phi}(\hat{h}_i) + \\ & \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu} - \gamma\right)^2\hat{\phi}'(\hat{h}_i)m^{\nu} + \\ & \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu} - \gamma\right)\left(\xi_i^{\nu'} - \gamma\right)\hat{\phi}'(\hat{h}_i)m^{\nu'} \end{aligned} \tag{2.65}$$

Then, following analogous passages as Eq. (2.41-2.44) we can obtain the expressions:

$$(1-q)m^{\nu} = \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu} - \gamma\right)\hat{\phi}(\hat{h}_i) + q\frac{P_{11} - \gamma^2}{\gamma(1-\gamma)}m^{\nu'}, \tag{2.66a}$$

$$(1-q)m^{\nu'} = \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu'} - \gamma\right)\hat{\phi}(\hat{h}_i) + q\frac{P_{11} - \gamma^2}{\gamma(1-\gamma)}m^{\nu}, \tag{2.66b}$$

where P_{11} is the joint probability defined in Eq. (2.109). Eq. (2.66) is a linear symmetric system

of the form:

$$Dm^{\nu} = B + Cm^{\nu'} \tag{2.67a}$$

$$Dm^{\nu'} = B' + Cm^{\nu} \tag{2.67b}$$

where $B = \frac{1}{\gamma(1-\gamma)N}\sum_{i=1}^{N}\left(\xi_i^{\nu}-\gamma\right)\hat{\phi}(\hat{h}_i)$, $D = (1-q)$ and $C = q\frac{P_{11}-\gamma^2}{\gamma(1-\gamma)}$. System Eq. (2.66) has solutions

$$m^{\nu} = \frac{DB + CB'}{D^2 - C^2} \tag{2.68a}$$

$$m^{\nu'} = \frac{DB' + CB}{D^2 - C^2}. \tag{2.68b}$$

We are now ready to write the expressions for $(m^{\nu})^2$ and $m^{\nu}m^{\nu'}$:

$$\begin{aligned}(m^{\nu})^2 =& \frac{D^2B^2 + C^2(B')^2 + 2DCBB'}{(D^2-C^2)^2},\\ m^{\nu}m^{\nu'} =& \frac{D^2BB' + CD(B')^2 + CDB^2 + C^2BB'}{(D^2-C^2)^2},\end{aligned} \tag{2.69}$$

where B and B' are analogous to the term on right hand side of Eq. (2.42): $(B')^2 = B^2 = \frac{p}{N^2\gamma(1-\gamma)}$. The last missing piece is the cross term BB' which can also be calculated analogously to Eq. (2.47):

$$\begin{aligned}BB' &= \frac{1}{\left[N\gamma(1-\gamma)\right]^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\left(\xi_i^{\nu}-\gamma\right)\left(\xi_j^{\nu'}-\gamma\right)\hat{\phi}(\hat{h}_i)\hat{\phi}(\hat{h}_j)\right] =\\ &= \frac{1}{\left[N\gamma(1-\gamma)\right]^2}\sum_{i=1}^{N}\left[(P_{11}-\gamma^2)\right]\hat{\phi}^2(\hat{h}_i) =\\ &= \frac{P_{11}-\gamma^2}{N\left[\gamma(1-\gamma)\right]^2}p.\end{aligned} \tag{2.70}$$

In the last passage we used the fact that $\hat{\phi}(\hat{h}_i)$ and $\hat{\phi}(\hat{h}_j)$ are independent thank to the Taylor expansion in passage Eq. (2.65). Plugging the expressions for $(m^{\nu})^2$ and $m^{\nu}m^{\nu'}$ into Eq. (2.64), we can obtain the variance

$$\begin{aligned}\langle\sigma^2\rangle =& \frac{p}{(D^2-C^2)^2\left[\gamma(1-\gamma)\right]^2}\alpha\cdot\\ &\cdot\left\{\gamma(1-\gamma)\left[(D^2+C^2)\gamma(1-\gamma)+2DC(P_{11}-\gamma^2)\right]+\right.\\ &\left.+(P_{11}-\gamma^2)\left[2CD\gamma(1-\gamma)+(D^2+C^2)(P_{11}-\gamma^2)\right]\right\}\end{aligned} \tag{2.71}$$

Finally we can write the expression for the effective r', under the effect of pairwise correlation

between background patterns:

$$r' = \frac{p}{\left[(1-q)^2 - (qC)^2\right]^2} \left\{\left[(1-q)^2 + (qC)^2\right](1+C^2) + 4(1-q)qC\right\} \tag{2.72}$$

where $C = \frac{P_{11}-\gamma^2}{\gamma(1-\gamma)}$ is the correlation between couples of patterns. The so obtained expression for r' can be substituted to that of system Eq. (2.56).
Eq. (2.72) can be proven as follows:

$$\begin{aligned} r' =& \frac{p}{\left[(D^2-C^2)\gamma(1-\gamma)\right]^2} \left\{\gamma(1-\gamma)\left[(D^2+C^2)\gamma(1-\gamma) + 2DC(P_{11}-\gamma^2)\right] + \right. \\ & \left. (P_{11}-\gamma^2)\left[2DC\gamma(1-\gamma) + (D^2+C^2)(P_{11}-\gamma^2)\right]\right\} = \\ =& \frac{p}{(D^2-C^2)^2}\left\{(D^2+C^2) + 2DCC + (D^2+C^2)C^2\right\} = \\ =& \frac{p}{\left[(1-q)^2-(qC)^2\right]^2}\left\{\left[(1-q)^2+(qC)^2\right](1+C^2) + 4(1-q)qC\right\}. \quad \square \end{aligned} \tag{2.73}$$

The derivation of Eq. (2.72) can be extended to the case in which background patterns share correlation C between non overlapping groups of exactly p patterns. To do so, we need to extend the system in Eq. (2.67b) linearly:

$$M = \begin{pmatrix} D & -C & -C & \cdots & -C \\ -C & D & -C & \cdots & -C \\ -C & -C & D & \cdots & -C \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -C & -C & -C & \cdots & D \end{pmatrix} \cdot \begin{pmatrix} m^{\nu} \\ m^{\nu'} \\ m_{\nu''} \\ \vdots \\ m_{\nu^{n'}} \end{pmatrix} = \begin{pmatrix} B \\ B' \\ B'' \\ \vdots \\ B^{n'} \end{pmatrix} \tag{2.74}$$

where M is a $p \times p$ matrix. In order to find the solution $\vec{m}^{\nu} = M^{-1}\vec{B}$ of system Eq. (2.74) we need to invert the matrix M. Indeed matrices with the symmetry properties of M

$$\{M\}_{ij} = \begin{cases} D & \text{if } i = j \\ -C & \text{if } i \neq j \end{cases} \tag{2.75}$$

are invertible, with inverse matrix:

$$\left\{M^{-1}\right\}_{ij} = \frac{1}{Z}\begin{cases} D+C-1 & \text{if } i = j \\ -1 & \text{if } i \neq j \end{cases} \tag{2.76}$$

where

$$Z = \frac{-D+(n-1)C}{C(D+C)^2}. \tag{2.77}$$

In order to derive Eq. (2.76) we can rewrite the matrix M as $M = A - C\nu\nu^T$, where A is diagonal

with entries $A_{i,i} = D - C$ and v is a column vector of all ones. If M and A are both invertible, we can use the Sherman-Morrison formula:

$$M^{-1} = \left(A - Cvv^T\right)^{-1} = A^{-1} - \frac{-CA^{-1}vv^TA^{-1}}{1 - Cv^TA^{-1}v}. \tag{2.78}$$

Since A is diagonal, then $(A^{-1})_{i,i} = (A_{i,i})^{-1} = \frac{1}{D+C}$. Then

$$\left\{M^{-1}\right\}_{ij} = \begin{cases} \frac{1}{D+C} - \frac{1}{c(D-C)^2} & \text{if } i = j \\ -\frac{1}{c(D-C)^2} & \text{if } i \neq j \end{cases} \tag{2.79}$$

where the constant $c = -\frac{1}{C} + n\frac{1}{D+C}$. Terms can be re-arranged to obtain Eq.(2.76).

As a final note, we consider the case in which all patterns are equally correlated, then

$$\left\langle \sigma^2 \right\rangle = \left\langle \frac{P}{N}\gamma(1-\gamma)(m^\nu)^2 + \frac{P^2}{N}(P_{11} - \gamma^2)m^\nu m^\mu \right\rangle \tag{2.80}$$

This variance can not be computed since the two terms in the brackets scale differently with P and N. We can conclude that it is not possible to allow all stored patterns to be correlated with arbitrary correlation C if we fix the pattern mean activity to γ.

2.6.12 Mean-field dynamics in the presence of adaptation and global feedback

In order to derive the mean-field equations for the model with adaptation and global feedback, we consider the simplest case, in which only two patterns are correlated ($\vec{\xi}^1$ and $\vec{\xi}^2$) while all the others are independent. Analogously to Section "Mean field equations for two correlated patterns", we can group neurons into four homogeneous populations (in the presence of background patterns, the neural populations can be slightly in-homogeneous): neurons that are selective to both patterns ($\xi_i^1 = \xi_i^2 = 1$), neurons selective to pattern 1 but not 2 ($\xi_i^1 = 1, \xi_i^2 = 0$), neurons selective to pattern 2 but not 1 ($\xi_i^1 = 0, \xi_i^2 = 1$) and neurons that are selective to neither pattern 1 or 2 ($\xi_i^1 = \xi_i^2 = 0$). The probability for a neuron to belong to population (x^1, x^2), i.e. $\xi_i^1 = x^1$ and $\xi_i^2 = x^2$, is the joint probability P_{x^1,x^2} in Eq. (2.109). Furthermore each population (x^1, x^2) is characterized by a different firing threshold $\theta_{x^1,x^2}(t)$. Following a similar derivation to that that of to Eq. (2.57), we obtain the six-dimensional

mean-field dynamics:

$$\hat{\tau}\frac{dm^1}{dt} = -m^1 + F_1(m^1, m^2, \{\hat{\theta}_{x^1x^2}\}), \tag{2.81a}$$

$$\hat{\tau}\frac{dm^2}{dt} = -m^2 + F_2(m^1, m^2, \{\hat{\theta}_{x^1x^2}\}), \tag{2.81b}$$

$$\hat{\tau}_\theta\frac{d\hat{\theta}_{x^1x^2}}{dt} = -\theta_{x^1x^2} + \hat{\theta}_0 + \hat{D}_\theta\left\langle\hat{\phi}_{x^1x^2}(m^1, m^2, \hat{\theta}_{x^1x^2})\right\rangle \quad x^1, x^2 \in \{0,1\}. \tag{2.81c}$$

Here, we have introduced the nonlinear functions

$$F_\mu(m^1, m^2, \{\hat{\theta}_{x^1x^2}\}) = \sum_{x^1=0,1}\sum_{x^2=0,1} P_{x^1,x^2}\frac{x^\mu-\gamma}{\gamma(1-\gamma)}\left\langle\hat{\phi}_{x^1x^2}(m^1, m^2, \hat{\theta}_{x^1x^2})\right\rangle \tag{2.81d}$$

($\mu = 1,2$), and the mean firing-rate u of population (x^1, x^2)

$$u := \left\langle\hat{\phi}_{x^1x^2}(m^1, m^2, \hat{\theta}_{x^1x^2})\right\rangle = \int \hat{\phi}(\hat{h}_{x^1x^2}(m^1, m^2, \hat{\theta}_{x^1x^2}, z))e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}}, \tag{2.81e}$$

the mean-field input potential

$$\begin{aligned}\hat{h}_{x^1,x^2}(m^1, m^2, \{\hat{\theta}_{k^1k^2}\}, z) = &\left(x^1-\gamma\right)m^1 + \left(x^2-\gamma\right)m^2 + \sqrt{\alpha R}z \\ &-\hat{\theta}_{x^1x^2} - \frac{\hat{J}_0(t)}{\gamma}\sum_{k^1=0,1}\sum_{k^2=0,1} P_{k^1,k^2}\left\langle\hat{\phi}_{x^1x^2}(m^1, m^2, \hat{\theta}_{k^1k^2})\right\rangle.\end{aligned} \tag{2.81f}$$

and the mean squared overlap of background patterns R given by

$$R = \frac{p}{(1-q)^2} \tag{2.81g}$$

$$q = \sum_{x^1=0,1}\sum_{x^2=0,1} P_{x^1,x^2}\int \hat{\phi}'(\hat{h}(x^1, x^2, z), \hat{\theta}_{x^1x^2})e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}} \tag{2.81h}$$

$$p = \sum_{x^1=0,1}\sum_{x^2=0,1} P_{x^1,x^2}\int \hat{\phi}^2(\hat{h}_{x^1,x^2}(m^1, m^2, \{\hat{\theta}_{x^1x^2}\}, z), \hat{\theta}_{x^1x^2})e^{-\frac{z^2}{2}}\frac{dz}{\sqrt{2\pi}} \tag{2.81i}$$

In order to obtain $\left\langle\hat{\phi}_{x^1x^2}(m^1, m^2, \hat{\theta}_{x^1x^2})\right\rangle$ in Eq. (2.81c) and (2.81d), Eqs. (2.81e) – (2.81i) need be solved self-consistently (for more details, see Section “Numerical Solutions”).

2.6.13 Stability of the fixed points

In order to compute the stability of the fixed points in Fig. 2.7, we compute the eigenvalues of the Jacobian matrix J of the $m^1 - m^2$ dynamics at the point location in the $m^1 - m^2$ plane. The Jacobian matrix is symmetric and the three independent entries are computed from Eq. (2.56)

as:

$$J_{11}(m^1, m^2) = -1 + \frac{A}{\gamma(1-\gamma)} \left\langle \left(\xi_i^1 - \gamma\right)^2 \phi'(h_i) \right\rangle_i \tag{2.82}$$

$$J_{12}(m^1, m^2) = J_{21}(m^1, m^2) = \frac{A}{\gamma(1-\gamma)} \left\langle \left(\xi_i^1 - \gamma\right)\left(\xi_i^2 - \gamma\right) \phi'(h_i) \right\rangle_i \tag{2.83}$$

$$J_{22}(m^1, m^2) = -1 + \frac{A}{\gamma(1-\gamma)} \left\langle \left(\xi_i^2 - \gamma\right)^2 \phi'(h_i)) \right\rangle_i \tag{2.84}$$

In the numerical computation of the J, we exploited the symmetries under exchange of m^1 and m^2, for example $J_{22}(m^1, m^2) = J_{11}(m^2, m^1)$ and so on.

Analogously to the system in Eq. (2.56), also the Jacobian matrix can be adapted to the case of 3 or 4 correlated pattern, using the joint probabilities in Eq. (2.113) and the generic forms

$$J_{\mu,\mu}(m^\mu, m^\mu) = -1 + \frac{A}{\gamma(1-\gamma)} \left\langle \left(\xi_i^\mu - \gamma\right)^2 \phi'(h_i) \right\rangle_i, \tag{2.85}$$

$$J_{\mu,\nu}(m^\mu, m^\nu) = \frac{A}{\gamma(1-\gamma)} \left\langle \left(\xi_i^\mu - \gamma\right)\left(\xi_i^\nu - \gamma\right) \phi'(h_i) \right\rangle_i \tag{2.86}$$

2.6.14 The limit case $b \to \infty$: when the gain function is an Heaviside

In the limit $b \to \infty$, the gain function can be approximated by an Heaviside function $\phi(h) = r_{\max}\Theta(h - h_0)$ which leads to some simplifications in the explicit writing of the mean-field system Eq. (2.56). First of all, we can rewrite $\phi^2(h) = r_{\max}^2\Theta(h - h_0)$ and $\phi'(h) = r_{\max}\delta(h - h_0)$, where $\delta(x)$ is the Dirac delta function. In the next passage the erfc function come at hands. Erfc is defined as $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$, where erf is the error function and we use the following identity, which follows directly from the definition:

$$\int_c^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} = \frac{1}{2}\mathrm{erfc}\left(\frac{c}{\sqrt{2}}\right). \tag{2.87}$$

The identity in Eq. (2.87) allows to rewrite the system Eq. (2.56) as:

$$\tau \frac{dm^1}{dt} = -m^1 + \frac{1}{2\gamma(1-\gamma)} \left\langle \left(\xi_i^1 - \gamma\right) \operatorname{erfc}\left(\frac{h_0 - h_i}{A\sqrt{2\alpha r}}\right) \right\rangle_i \tag{2.88a}$$

$$\tau \frac{dm^2}{dt} = -m^2 + \frac{1}{2\gamma(1-\gamma)} \left\langle \left(\xi_i^2 - \gamma\right) \operatorname{erfc}\left(\frac{h_0 - h_i}{A\sqrt{2\alpha r}}\right) \right\rangle_i \tag{2.88b}$$

$$q = \int \sum_{x^1} \sum_{x^2} P(x^1, x^2) \frac{r_{\max}}{\sqrt{2\pi}} e^{-\frac{(h_i)^2}{A\sqrt{2\alpha r}}} \tag{2.88c}$$

$$p = \frac{r_{\max}^2}{2} \left\langle \operatorname{erfc}\left(\frac{h_0 - h_i}{A\sqrt{2\alpha r}}\right) \right\rangle_i \tag{2.88d}$$

$$r = \frac{p}{(1 - Aq)^2} \tag{2.88e}$$

In Eq. (2.88), the input term $h(x^1, x^2, z))$ is the same as Eq. (2.57d) and the expectations over i are computed as in Eq. (2.56).

If we consider the case where neural self-interaction is excluded, an extra correction term should be added to the input $h(x^1, x^2, z))$ and its limit for $b \to \infty$ reads as follows:

$$\frac{A^2 q \alpha \phi(h)}{\left(1 - Aq\right)} \xrightarrow[b \to \infty]{} \frac{A \alpha r_{\max}}{2}. \tag{2.89}$$

Finally, in order to derive Eq. (2.4) let's consider the retrieving state of pattern 1 (that of pattern 2 is symmetric with respect to the $m^1 - m^2$ axis in absence of external input): in this state, $m^1 = 1$, and m^2 depends on the correlation C, as it emerges from Fig. 2.7A, however what is the the exact value? It can be computed analytically in limit, $b \to \infty$ and $\gamma \to 0$. In this limit, we can make the following approximations

$$\gamma(1-\gamma) \sim \gamma, \tag{2.90a}$$

$$P_{11} = \gamma^2 + \gamma(1-\gamma)C \sim \gamma C, \tag{2.90b}$$

$$1 - \gamma \sim 1, \tag{2.90c}$$

$$\gamma \sim 0, \tag{2.90d}$$

and the identity $P_{01} = \epsilon(1-\epsilon)$ (as derived in Eq. (2.109)). In these approximations, the m^2 value in the fixed points is given by

$$m^2 = \frac{1}{r_{\max}\gamma} \{P_{11}\phi\left[Ar_{\max}(m^1 + m^2)\right] + P_{01}\phi\left[Ar_{\max}m^2\right]\}, \tag{2.91}$$

or equivalently

$$m^2 = \frac{1}{r_{\max}\gamma}\{\gamma C\phi\left[Ar_{\max}(m^1+m^2)\right] + \epsilon(1-\epsilon)\phi\left[Ar_{\max}m^2\right]\}. \tag{2.92}$$

We use the $b \to \infty$ and the assumption $m^1 = 1$:

$$m^2 = C\Theta\left(Ar_{\max}(1+m^2) - h_0\right) + \frac{\epsilon(1-\epsilon)}{\gamma}\Theta\left(Ar_{\max}m^2 - h_0\right). \tag{2.93}$$

Since $h_0 < 1$ and $m^2 \geq 0$, the term $\Theta\left(Ar_{\max}(1+m^2) - h_0\right) = 1$. Therefore, $m^2 = C$ if $Ar_{\max}m^2 < h_0$ (confront to the bifurcation diagram in Fig. 2.7C). In the limit case where $m^2 \to \frac{h_0}{Ar_{\max}}$ we obtain:

$$C_{\max} \leq C^0_{\max} \equiv \frac{h_0}{Ar_{\max}} = \hat{h}_0. \tag{2.94}$$

2.6.15 Robustness to heterogeneity

Heterogeneous F-I curves

In this section we introduce heterogeneous F-I curves. Each neuron is characterised by a baseline firing rate $(r_{\min})_i$ a maximum firing rates $(r_{\max})_i$. The sigmoidal F-I function in Eq. 2.22 has to be rewritten as

$$\phi(h) = \frac{r_{\max}}{1+e^{-b(h-h_0)}} \to \phi_i(x) = \frac{(r_{\max})_i - (r_{\min})_i}{1+e^{-\hat{b}(x-(\hat{h}_0)_i)}} + (r_{\min})_i. \tag{2.95}$$

We define two Gaussian distribution $N(\mu_{r_{\min}}, \sigma_{r_{\min}})$ and $N(\mu_{r_{\max}}, \sigma_{r_{\max}})$from which we sample $(r_{\min})_i$ and $(r_{\max})_i$ respectively. Since negative values of firing rates do not have a physical meaning we set the minimum of $(r_{\min})_i$ to 0. Similarly, we do not want to allow the maximal firing rate to be too low, so we set the minimum value of the re-scaled firing rate $(\hat{r}_{\max})_i$ to 0.5. Finally, parameter $(\hat{h}_0)_i$ is defined as $h_0((r_{\max})_i - (r_{\min})_i)$, where h_0 is a model's parameter. Finally in Eq. 2.20, we re-scale the firing rates as follows:

$$r_i \to \text{Max}\left[0, \frac{(r_i - (r_{\min})_i}{(r_{\max})_i - (r_{\min})_i}\right]. \tag{2.96}$$

Diluted weight matrix

We define an attractor neural network of N units, where each unit receives input from K others. The probability of having a connection between two units is $d = M/N$. The load of the network is defined as $\alpha = P/N$, where P is the total number of patterns. We also assume that $A/d =$ constant (to be introduced into the dimensional analysis). The input term Eq. 2.21 is

filtered with gain function ϕ, which is chosen to be a sigmoid as in Eq. 2.22. The connection matrix, w_{ij}, contains the synaptic weights between neurons i and j, but, compared to Eq. 2.5, connections are diluted with probability d as defined in (Pereira and Brunel, 2018)

$$w_{ij} = \frac{A}{N\gamma(1-\gamma)} \frac{d_{ij}}{d} \sum_{\mu}^{P} \left(\xi_i^\mu - \gamma\right)\left(\xi_j^\mu - \gamma\right) \tag{2.97}$$

where d_{ij} is 1 with probability M/N and 0 otherwise and the constant A can be interpret as "connection strength". In order for the weights to have expectation $\langle w_{ij} \rangle = 0$, we subtract the mean activity of patterns $< \xi_i^\mu >= \gamma$. Using the similarity measure introduced in Eq. 2.24, the input terms h_j can also be re-written as a function of the overlaps $m^1, ..., m^P$, by using the definition of the weights w_{ij}, Eq. (2.97), and that of the overlaps.

$$\begin{aligned} h_i = \sum_j^N w_{ij} r_j &= \frac{A}{Nd\gamma(1-\gamma)} \sum_j^N d_{ij} \sum_\mu^P \left(\xi_i^\mu - \gamma\right)\left(\xi_j^\mu - \gamma\right) r_j = \\ &= \frac{A}{Nd\gamma(1-\gamma)} \sum_j^N d_{ij} \left(\xi_i^1 - \gamma\right) r_j + \frac{A}{Nd\gamma(1-\gamma)} \sum_j^N d_{ij} \sum_{\mu=2}^P \left(\xi_j^\mu - \gamma\right)\left(\xi_i^\mu - \gamma\right) r_j \end{aligned} \tag{2.98}$$

where we have separate the "signal" related to the first pattern being retrieved and a noise term Y_i. We write $h_i = Am^1 + Y_i$, with

$$Y_i = \frac{A}{Nd\gamma(1-\gamma)} \sum_j^N (1-d_{ij}) \left(\xi_i^1 - \gamma\right) r_j + \frac{A}{Nd\gamma(1-\gamma)} \sum_j^N d_{ij} \sum_{\mu=2}^P \left(\xi_j^\mu - \gamma\right)\left(\xi_i^\mu - \gamma\right) r_j \tag{2.99}$$

Since the terms d_{ij} and $\left(\xi_i^1 - \gamma\right)$ are independent, we have $\langle Y_i \rangle$.

We assume Y_i to be distributed like a Gaussian with variance

$$\begin{aligned} \langle\langle \sigma^2 \rangle\rangle_i = \langle\langle (Y_i)^2 \rangle\rangle_i &= \frac{1}{N} \sum_i^N \frac{A^2}{N^2 d^2 \gamma^2 (1-\gamma)^2} \sum_{\mu\neq 1}^P \sum_{\nu\neq 1}^P \left\langle \left(\xi_i^\mu - \gamma\right)\left(\xi_i^\nu - \gamma\right) \sum_j \sum_k \left(\xi_j^\mu - \gamma\right)\left(\xi_k^\nu - \gamma\right) \right\rangle d_{ij} d_{ik} r_j r_k = \\ &= \frac{1}{N} \sum_i^N \frac{A^2}{N^2 d^2 \gamma^2 (1-\gamma)^2} \sum_{\mu\neq 1}^P \left(\xi_i^\mu - \gamma\right)^2 \sum_j d_{ij} \left(\xi_j^\mu - \gamma\right)^2 r_j^2 \end{aligned} \tag{2.100}$$

in the last passage, we used the fact that $\left\langle \left(\xi_j^\mu - \gamma\right)\left(\xi_k^\nu - \gamma\right) \right\rangle = \delta_{jk}\gamma(1-\gamma)$ and $d_{ij}^2 = d_{ij}$. We then apply the same independence argument as used for the signal term and obtain

$$\langle \sigma^2 \rangle = \frac{A^2 r_{\max}^2}{d^2} \gamma(1-\gamma) d \sum_\mu (m^\mu)^2 \tag{2.101}$$

From now on the passages are the same as in the SI, except maybe the correction term for

excluding self-interaction, which I should recompute.

The final difference in the equations is that the term $\sqrt{\alpha' r}z$, where $\alpha' = P/N$ should be substituted with $\sqrt{\alpha d r}z$. The two terms however are equivalent since $\alpha' = \alpha d$.

2.6.16 Relation to experiments: algorithms to generate correlated patterns

In this section we describe how a single subgroup of K patterns with sparseness γ is created according to three different algorithms. Patterns belonging to the same subgroup correspond to associated concepts and share pair-wise a fraction of neurons c. For the hierarchical generative model and the indicator neuron model, we associate the algorithm to the theoretical probability distribution for a neuron to respond exactly to k concepts out of K.

Hierarchical generative model.

We start by creating a "parent" pattern which is not part of the subgroup. The parent pattern has sparseness $\lambda = \gamma/c$: $\text{prob}(\xi_i^{\text{parent}} = 1) = \lambda$. We proceed to create the actual patterns by copying the ones of the parent pattern with probability c, while the zeros stay untouched, following the conditional probabilities

$$\text{prob}(\xi_i^\mu = 1 | \xi_i^{\text{parent}} = 1) = c, \tag{2.102a}$$
$$\text{prob}(\xi_i^\mu = 1 | \xi_i^{\text{parent}} = 0) = 1 - c, \tag{2.102b}$$
$$\text{prob}(\xi_i^\mu = 1 | \xi_i^{\text{parent}} = 0) = 0, \tag{2.102c}$$
$$\text{prob}(\xi_i^\mu = 0 | \xi_i^{\text{parent}} = 0) = 1. \tag{2.102d}$$
$$\tag{2.102e}$$

This ensures that the patterns ξ_i^μ have the right sparseness and fraction of pair-wise shared neurons. The sparseness can be checked as follows:

$$\text{prob}(\xi_i^\mu = 1) = \lambda c = \gamma, \tag{2.103a}$$
$$\text{prob}(\xi_i^\mu = 0) = \lambda(1 - c) + (1 - \lambda) = 1 - \gamma. \tag{2.103b}$$

On the other hand, the fraction of pair-wise shared neurons is given by the conditional probability that a neuron is part of pattern ν given that is it part of pattern μ:

$$\text{prob}(\xi_i^\nu = 1 | \xi_i^\mu = 1) = c + (1 - c)\delta^{\mu\nu}. \tag{2.104}$$

Hence the fraction of shared neurons as it should be. More generically, the theoretical probability (or the expectation) that a neuron participates in k patterns out of K is

$$P^K(k) = \frac{K!}{(K-k)!k!}\lambda c^k(1-c)^{K-k} + (1-\lambda)\delta_{k0}. \tag{2.105}$$

Indicator neuron model.

To create a subgroup of pair-wise associated patterns using indicator neurons (i.e. neurons that indicate the subgroup), we proceed in three steps:

1) generate with probability λ a small subset of indicator neurons for this subgroup. This subset gives a parent pattern of indicator neurons:

$$\text{prob}(\xi_i^{\text{parent}} = 1) = \lambda_{ind} = \frac{c\gamma - \gamma^2}{(1-\epsilon)^2 - 2\gamma(1-\epsilon) + c\gamma}. \tag{2.106}$$

In a network of N neurons, $n_{ind} = \lambda_{ind}N$ are indicator neurons.

2) To create each pattern μ of the subgroup, copy indicator neurons with probability $(1-\epsilon)$:

$$\text{prob}(\xi_i^{\mu} = +1|\xi_i^{parent} = 1) = 1-\epsilon \tag{2.107}$$

3) Add random neurons (with probability Ω) to pattern μ

$$\text{prob}(\xi_i^{\mu} = 1|\xi_i^{\text{parent}} = 0) = \Omega = \frac{\gamma - \lambda_{ind}(1-\epsilon)}{1-\lambda_{ind}}. \tag{2.108}$$

This last probability can also be interpreted as the probability of flipping a 0 from the parent pattern when creating the correlated patterns.

With this construction, the total number of neurons that are active in pattern μ is $\lambda_{ind}N(1-\epsilon) + (1-\lambda_{ind})N\frac{\gamma-\lambda_{ind}(1-\epsilon)}{1-\lambda_{ind}} = N\gamma$ as it should be. The value of λ_{ind} is chosen in order to ensure that the fraction of pair-wise shared neurons is c. Indeed we found it by solving $c\gamma = \lambda(1-\epsilon)^2 + (1-\lambda)\Omega$.

In this work, we always choose ϵ such that $\epsilon = \Omega$, For specific case $\epsilon = \Omega$, it is possible to derive ϵ directly from the correlation C and the sparsity γ.

We create a "parent" pattern $\vec{\xi}^0$ with mean activity $\langle \xi_i^0 \rangle_i = \lambda$. Starting from $\vec{\xi}^0$ we create $\vec{\xi}^1$ and $\vec{\xi}^2$, each unit i has probability ϵ of being the equal to ξ_i^0 and probability $1-\epsilon$ of being flipped compared to ξ_i^0. All other patterns $\xi^\mu, \mu = 3, ..., P$ are sorted independently from a Bernoulli distribution with probability $P(\xi_i^\mu = 1) = \gamma$. The joint probabilities $P_{kl} = P(\xi_i^1 = k, \xi_i^2 = l)$ can

be computed as functions of the probabilities λ and ϵ:

$$P_{11} = \lambda\epsilon^2 + (1-\lambda)(1-\epsilon)^2, \tag{2.109a}$$
$$P_{10} = P_{01} = \lambda\epsilon(1-\epsilon) + (1-\lambda)\epsilon(1-\epsilon) = \epsilon(1-\epsilon), \tag{2.109b}$$
$$P_{00} = \lambda(1-\epsilon)^2 + (1-\lambda)\epsilon^2. \tag{2.109c}$$
$$\tag{2.109d}$$

Note that by this procedure we only obtain non-negative correlations $C \in [0,1]$.

Using P_{11} from Eq. (2.109), we can express C as

$$C(\lambda,\epsilon) = \frac{P_{11}-\gamma^2}{\gamma(1-\gamma)} = \frac{(1-\lambda)\left[\lambda\epsilon^2 + (1-\epsilon)^2\right]}{\gamma(1-\gamma)}. \tag{2.110}$$

Similarly, the mean activity of the correlated patterns can be expressed as a function of λ and ϵ as

$$\gamma(\lambda,\epsilon) = \langle \xi_i^1 \rangle_i = \langle \xi_i^2 \rangle_i = \lambda\epsilon + (1-\lambda)(1-\epsilon). \tag{2.111}$$

So far, we showed how to generate correlated patterns given the probabilities λ and ϵ. Conversely, how do we choose λ and ϵ given the mean activity γ and the correlation C, $C \geq 0$? To this end, we invert the above relations in order to solve for $\lambda(C,\gamma)$ and $\epsilon(C,\gamma)$:

$$\lambda = \frac{\gamma+\epsilon-1}{2\epsilon-1}, \tag{2.112a}$$
$$2\epsilon^3 - 3\epsilon^2 + \left[1 + 2\gamma(1-\gamma)(1-\hat{C})\right]\epsilon - \gamma(1-\gamma)(1-\hat{C}) = 0, \tag{2.112b}$$

Eq. (2.112b) has up to three solutions, we chose those that are real and in the range $[0,1]$.

The same procedure can be generalized to generate several correlate binary patterns. The general formula for the joint probability can be written as follows:

$$P_{x^1,\ldots,x^n} = \lambda\epsilon^a(1-\epsilon)^b + (1-\lambda)\epsilon^b(1-\epsilon)^a, \tag{2.113}$$

where $a = \sum_{\mu=1}^n x^\mu$ is the number of x^μ variables taking value 1 and $b = n - a$ is the number of x^μ variables taking value 0. The value of the joint probabilities in Eq. (2.113) is invariant under permutation of the x^μ.

Iterative correlation model.

In this subgroup construction, we do not define any parent pattern. We define the number of active neurons as γN and the number of pair-wise shared neurons as $\gamma c N$.

1) We define the set of "untouched neurons", which counts all neurons at the beginning of the procedure

2) We create pattern 1 by randomly sample γN neurons and exclude the sampled neurons from the untouched ones.

3) For every pattern ξ^{μ} with μ from 2 to K, compare it with each of the already created patterns. Let's suppose we are comparing the new pattern μ with the already formed pattern ν. a) check how many neurons are in common between the two. b) sample from pattern ν the remaining neurons needed to reach γcN shared neurons.

4) Complete pattern μ by adding neurons from the untouched ones until reaching γN active units. 5) Remove the units used in point 4 from the untouched ones.

It is important to underline the necessity of point 3.a. To make it clearer, let's consider the case we are building a subgroup of 3 patterns. We build the first one as in point 2. When we build pattern two starting from scratch, it does not share any neuron with pattern 1, so we just sample γcN from pattern 1 and $\gamma N(1-c)$ from the untouched neurons. Now we move to pattern 3. As before, it does not share neurons from pattern 2, so we pick γcN from it. Now we compare pattern 3 with pattern 1: it ca be that between the neurons we picked from pattern 2 some belong to pattern 1 as well, that's why we need to adjust the number of neurons to pick in order to preserve the correct amount of pair-wise correlation.

When the subgroup size K is big however it is still possible to exceed the correct fraction of shared neurons between some of the patterns that get build the last. Let's suppose we are creating a subgroup of size $K = 16$, I start by applying point 3 of the algorithm between pattern 16 and 15, then pattern 16 and 14 and so on. It can be that when we get to the point of picking neurons from pattern 4, 3, 2, 1 we take some neurons that also belong to pattern 15 but they are not the ones we picked in the in the previous iteration and thus get accepted. This created an higher correlation between the last built patterns in large subgroups. We checked that this does not influence significantly the average pairwise correlation during the virtual experiments described in the next section.

2.6.17 Comparing algorithm predictions with experimental data

We use a data set containing the activity of human MTL neurons (De Falco et al., 2016). Data were collected in 100 recording sessions with epileptic patients. In each recording session several stimuli were presented to the patient. The association between each pair of stimuli was estimated using a web-based association score.

In order to compare the predictions of the algorithms with the data, we try to reproduce

the real data by running virtual experiments based on the three algorithms presented in the previous section. In each virtual experiment we replicate the conditions of the real experiment as follows. For each real experimental session, we first extract the number of responsive neurons in each session. We then group the presented stimuli into clusters based on an association matrix derived from the web-association scores. To do so, we use an hierarchical agglomerating clustering algorithm with threshold equal to the mean of the association matrix for the session. Such clusters indicate the amount and the size of the patterns subgroups we have to build for the corresponding virtual experiment.

We can then proceed with the virtual experiment: in each session we a) build subgroups of patterns in the same number and size as the clusters of stimuli for each of the three algorithms and then b) sample a neuron at the time and count to how many patterns does it respond to. c) Finally, the count of how many stimuli a neuron responds to that of other sessions. We sample neurons until we match the number of responsive neurons with that of the real experimental session. Each virtual experiment counts $N = 10^5$ neurons and it is run 40 times and plot in Fig. 2.5A the normalised mean and standard deviation.

We choose to ignore non-responding neurons in our analysis, since it is likely that the proportion of non-responsive neurons compared to that of responsive ones is largely underestimated in the experiment (non-responsive neurons are more likely to remain silent during the experiment and not to be recorded at all).

2.6.18 Comparing virtual experiments and expected distributions

It is also possible to compare the virtual experiments with the theoretical distributions in Eq. (2.105) and (2.113). Eq. (2.105) and (2.113) provide the probability that a neuron is selective to k out of K patterns if a single subgroup of stimuli is stored in the network. But how do we combine such probabilities when several subgroups of patterns are stored in the network? We define $\Psi_s(k)$ the probability that a neuron responds to exactly k patterns in session s. We know from the previous session the number and sizes G_j of subgroups present in each session. Then

$$\Psi_s(k) = \sum_{j=k}^{\max K} G_j P^j(k) \zeta_j \tag{2.114}$$

where $\max K$ is the biggest between all subgroup sizes K_j and $\zeta_j = 1 - P^j(0)$ is the probability that a neuron takes part into the subgroup j. The formula Eq. 2.114 is valid in the assumption that subgroups are strictly disjoint, meaning that we assume that the same can not take part into encoding patterns belonging to different subgroups. This assumption is not true for the way we algorithmically build subgroups patterns in the virtual experiments, however dropping it make the expression for $\Psi_s(k)$ not treatable. Finally the probabilities $\Psi_s(k)$ from

each session must be combined into the final distribution $\Psi_{\text{final}}(k)$:

$$\Psi_{\text{final}}(k) = \frac{\sum_s N_s^{\text{sample}} \Psi_s(k)}{\sum_s \sum_k N_s^{\text{sample}} \Psi_s(k)} = \frac{N_1^{\text{sample}} \Psi_1(k) + N_2^{\text{sample}} \Psi_2(k) + \ldots}{N_{\text{tot}}^{\text{sample}}} \tag{2.115}$$

where N_s^{sample} is the amount of responsive neurons measured in each experimental session and N_{tot} is the total amount of measured responsive neurons. In the last passage, note that $\sum_s \sum_k N_s^{\text{sample}} \Psi_s(k) = \sum_s N_s^{\text{sample}} \sum_k \Psi_s(k) = N_{\text{tot}}$, since $\sum_k \Psi_s(k)$ in every session s. The comparison between the theoretical distributions, the virtual experiments and the experimental data is shown in Fig. 2.13. The virtual experiments are the same a in Fig. 2.6: we re-run the experiment 40 times and took the average (main points) and standard deviation (error bars). The small mismatch between the theoretical predictions and virtual experiments is due to the fact that in the theoretical prediction we do not allow the same neurons to take part to two or more subgroups of concepts, while there is no such a restriction in the virtual experiment. Theory prediction and mean of the virtual experiments are really close, proving that only very few neurons take part in encoding different subgroups.

2.6.19 Numerical solutions

Two correlated patterns: finding the fixed points

The system in Eq.(2.56) is solved numerically to obtain the fixed nullclines, points, and flux arrows, plotted in Fig. 2.7. Fixed points are obtain through a grid search in the three-dimensional space spanned by m^1, m^2 and R. For each value of $R_{\text{val}} \in [0, \max(R)]$ and $m^1_{\text{val}}, m^2_{\text{val}} \in \big[\text{Lower bound}, \text{Upper bound}\big]$, Eq. (2.57d)–(2.57g) are solved. We call the value of R obtained by Eq. (2.57d) $R_{\text{reconstructed}}$. If R_{val} and $R_{\text{reconstructed}}$ are close enough, namely

$$|R_{\text{val}} - R_{\text{reconstructed}}| < \text{ correction-constant} \cdot \text{step}. \tag{2.116}$$

The quantity called “step" is the step size of the linear space we used to span R,

$$\text{step} = \frac{\max(R)}{\text{Resolution}} \tag{2.117}$$

The correction constant can increase or decrease the range in which we accept a value R_{val} as a valid solution: it is equal to 1 in most cases, but can be chose to be a bit bigger than one to avoid counting the same fixed point too many times. The values of R_{val} that satisfy Eq. (2.116) are then used to solve Eqs. (2.57), providing the values $m^1_{\text{reconstructed}}$ and $m^2_{\text{reconstructed}}$. Analogously to before, we find the solutions of Eqs. (2.57) comparing the values m^1_{val} and m^2_{val}

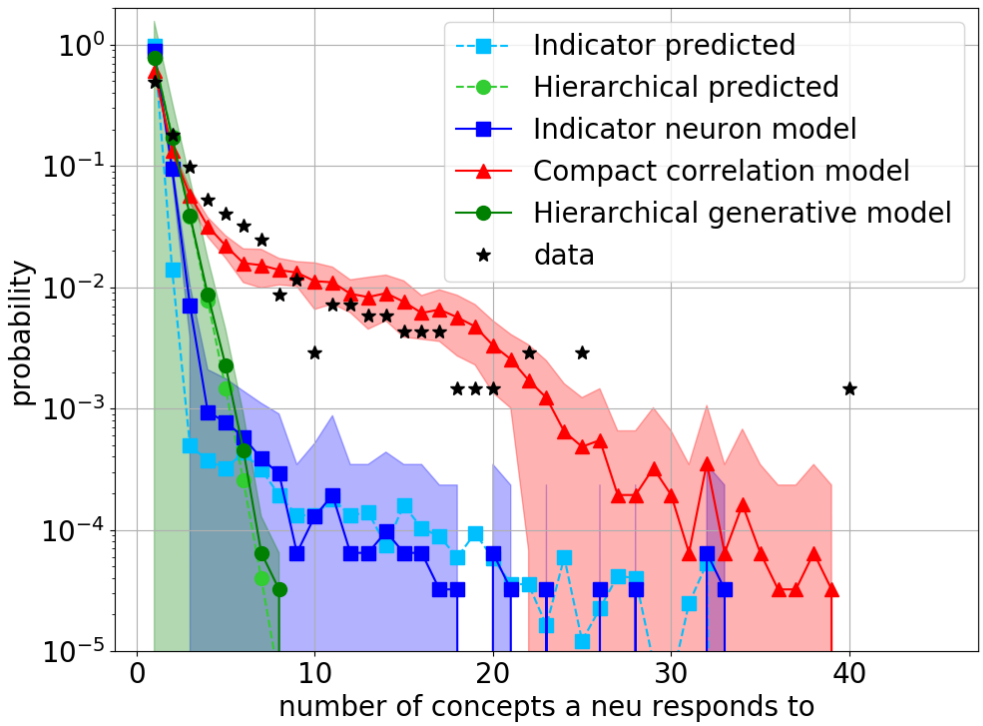


Figure 2.13 – Comparison between model prediction and data. Probability of finding a neuron responding to a given number of concepts as measured from experimental data (black stars), predicted by the three algorithms (as in Fig. 2.6, the area between error bar of one standard deviation is shaded) and theoretically forecast for the indicator neuron model (light blue) and for the hierarchical generative model (light green) obtained from Eq. (2.115).

with the recomputed counterparts $m^1_{\text{reconstructed}}$ and $m^2_{\text{reconstructed}}$ as follows

$$\left| m^{\mu}_{\text{val}} - m^{\mu}_{\text{reconstructed}} \right| < \text{ correction-constant} \cdot \text{step}, \tag{2.118}$$

where the step is defined as

$$\text{step} = \frac{|\text{Upper bound} - \text{Lower bound}\,|}{\text{Resolution}}. \tag{2.119}$$

Two correlated patterns with adaptation and periodic inhibition

In order to solve the dynamical equations of the mean-field in the presence of adaptation and global inhibition (as done in Fig. 2.10 and 2.11) we compute at each point in time

$\phi_{\bar{x^1}x^2}(m^1, m^1, \Theta_{x^1x^2})$, p, q, R and $J_0(t)$. In particular, the four $\phi_{\bar{x^1}x^2}(m^1, m^1, \Theta_{x^1x^2})$ are solved first and recursively since they are functions of themselves. We then update $\Theta_{x^1x^2}(t)$ with Euler method. Finally, we compute $m^1(t)$, $m^2(t)$. In order to compute $m^1(t)$, $m^2(t)$, we make a time-scale separation argument. We assume that $m^1(t)$ and $m^2(t)$ dynamics are much faster than $\Theta_{x^1x^2}(t)$ and $J_0(t)$, $\tau << T_{J_0} < T_\theta$. According to this approximation, at each point in time we let $m^1(t)$ and $m^2(t)$ reach their equilibrium values given the current $\Theta_{x^1x^2}(t)$ and $J_0(t)$. In other words, at each point in time, we consider all dynamical quantities frozen, than let $m^1(t)$ and $m^2(t)$ evolve according to their dynamics (we use Euler method) until convergence, and finally update the other quantities.

To find the fixed points in Fig. 2.10, we proceed like in the non-adaptive case: we do a grid search in the space spanned by m^1, m^2 and R. For each solution of R, $\phi_{\bar{x^1}x^2}(m^1, m^1, \Theta_{x^1x^2})$ are computed recursively. Finally, for the obtained values of R and $\phi_{\bar{x^1}x^2}(m^1, m^1, \Theta_{x^1x^2})$, the solutions of m^1 and m^2 are found.

Excluding self-interaction: a numerical approximation

In Fig. 2.2B, we compute the critical correlation for non-zero network load, $\alpha > 0$, in the case we consider the correction to exclude self interaction. To find the numerical solutions of the fixed points, we have approximated the input term $h(x^1, x^2, z)$ to the first order in z as follows:

$$h(x^1, x^2, z) = \langle h(x^1, x^2, z)\rangle_z + A\sqrt{\alpha r}z. \tag{2.120}$$

Then the quantity

$$\langle h(x^1, x^2, z)\rangle_z = Ar_{\max}\left(x^1 - \gamma\right)m^1 + Ar_{\max}\left(x^2 - \gamma\right)m^2 + \left\langle \frac{A^2 q\alpha\phi(h(x^1, x^2, z))}{\left(1 - Aq\right)}\right\rangle_z \tag{2.121}$$

can be approximated by

$$\langle h(x^1, x^2, z)\rangle_z \sim Ar_{\max}\left(x^1 - \gamma\right)m^1 + Ar_{\max}\left(x^2 - \gamma\right)m^2 + \frac{A^2 q\alpha\phi(\langle h(x^1, x^2, z)\rangle_z)}{\left(1 - Aq\right)} \tag{2.122}$$

which is equivalent to take the order 0 term into the Taylor expansion of $h(x^1, x^2, z)$ for small z.

Stability of the fixed points

The stability of the fixed points in Figs. 2.7,2.8 and 2.9 is obtained by computing the Jacobian matrix of the differential equations for m^1 and m^2 from Eq. (2.56a-b) respectively. Analogously, the stability of the fixed points in Figs. 2.9, 2.10 are obtained by computing the Jacobian matrix of the differential equations for m^1 and m^2 from Eq. (2.81ca-b).

Table 2.1 – Table of parameters, part1. For convenience the section number was dropped, so that Fig. 2.1, is referred to as simply Fig.1. Meaning of the abbreviations used: MF: obtained from mean field equations, FS: full network simulation, var. = this parameter is variable and specified in the figure.

Table of parameters								
Param.	Fig 1C/7A	Fig 8B-C	Fig 2A	Fig 2B	Fig 1D-2C	Fig 8A	Fig 3	Fig 9
γ	0.002	0.002	0.002	0.002	0.002	0.002	var.	0.002
$\hat{b}$	100	100	var.	100	100	100	100	100
$\hat{h}_0$	0.25	0.25	var.	var.	0.25	0.25	0.	0.
$r_{\max}$	1	1	1	1	1	1	1	1
τ	1	1	1	1	1	1	1	1
A	1	1	1	1	1	1	1	1
α	0	0	0	var.	0, 0.2	0	~0	0
C	var.	var.	-	-	var.	0.3	var.	var.
T_θ	-	-	-	-	-	-	45	∞
T_{J_0}	-	-	-	-	-	-	25	∞
Model	MF	MF	MF	var.	FS/MF	MF	FS	MF
Upper bound	1.2	1.05	1.05	1.05	-	1.05	-	1.2
Lower bound	-1.2	-0.05	-0.05	-0.05	-	-0.05	-	-0.2
max(R)	0.03	0.03	0.03	0.03	-	0	-	0
Resolution	1000	1000	100	100	-	1000	-	500
Corr. Const.	1	1	1.1	1.1	-	1	-	1
Size	1000	-	50	50	-	-	-	500
dt	-	-	-	-	0.1	0.1	0.1	-

When the steepness of the gain function is very high, $b > 1000$, we approximate the gain function with an Heaviside. The system Eq. (2.56) as well as the Jacobian matrix are rewritten in a simpler way for $b \to \infty$ as can be found in Eq. (2.88).

In the numerical computation of the Jacobian matrix computed in Section "Stability of the fixed points", we exploited the symmetries under exchange of m^1 and m^2, for example $J_{22}(m^1, m^2) = J_{11}(m^2, m^1)$ and so on.

Table 2.2 – Table of parameters, part2. For convenience the section number was dropped, so that Fig. 2.1, is referred to as simply Fig.1. Meaning of the abbreviations used: MF: obtained from mean field equations, FS: full network simulation, var. = this parameter is variable and specified in the figure.

Table of parameters					
Param.	Fig 4/10	Fig 11	Fig 12	Fig 5/13	Fig 6
γ	var.	0.002	0.001	0.002	0.002
$\hat{b}$	100	100	4.35	- 100	
$\hat{h}_0$	0.	0.	1.7	-	var.
$r_{\max}$	1	1	0.83	-	var.
τ	$\to 0$	$\to 0$	1	-	1
A	1	1	3.55	-	1
α	0	0	0	-	0.5
C	var.	0.2	var.	c = 0.04	var.
T_θ	45	45	-	-	-
T_{J_0}	25	25	-	-	-
Model	MF	MF/FS	MF	-	FS
Upper bound	1.2	-	1.2	-	-
Lower bound	-0.2	-	-0.2	-	-
max(R)	0	-	0	-	-
Resolution	1000	-	500	-	-
Corr. Const.	1	-	1	-	-
Size	-	-	200	-	-
dt	0.1	0.1	-	-	0.1

3 Optimal Stimulation Protocol in a Bistable Synaptic Consolidation Model

3.1 Paper information

The research and the paper writing was developed by Chiara Gastaldi, Samuel Muscinelli and Wulfram Gerstner.

At the time of publishing all authors were affiliated to the School of Computer and Communication Sciences and School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

The paper was published on the 13 November 2019 by Frontiers in Computational Neuroscience, Volume 13, Article 78. doi: 10.3389/fncom.2019.00078.

The paper version proposed here corresponds to the postprint version, identical to the published one.

Abstract

Synaptic changes induced by neural activity need to be consolidated to maintain memory over a timescale of hours. In experiments, synaptic consolidation can be induced by repeating a stimulation protocol several times and the effectiveness of consolidation depends crucially on the repetition frequency of the stimulations. We address the question: is there an understandable reason why induction protocols with repetitions at some frequency work better than sustained protocols – even though the accumulated stimulation strength might be exactly the same in both cases? In real synapses, plasticity occurs on multiple time scales from seconds (induction), to several minutes (early phase of long-term potentiation) to hours and days (late phase of synaptic consolidation). We use a simplified mathematical model of just two times scales to elucidate the above question in a purified setting. Our mathematical results show

that, even in such a simple model, the repetition frequency of stimulation plays an important role for the successful induction, and stabilization, of potentiation.

Keywords:

Synaptic consolidation, Plasticity, LTP, stimulation frequency, bistability.

Author Contributions

All authors contributed to the conception of the study. CG implemented the code needed to produce all figures, except for Figure 3.3. SM performed the bifurcation analysis presented in Figure 3.3 and contributed to the revision and improvement of the figures. CG and SM wrote the first draft of the manuscript. All authors contributed to correcting and improving the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

Acknowledgments

The authors would like to thank Tilo Schwalger for useful comments and discussions.

Funding

This research was supported by the Swiss national science foundation, grant agreement 200020_165538 and by the HBP.

3.2 Introduction

Synaptic plasticity, i.e. the modification of the synaptic efficacies due to neural activity, is considered the neural correlate of learning (Hebb, 1949; Martin et al., 2000; Caroni et al., 2012; Nabavi et al., 2014; Hayashi-Takagi et al., 2015; Holtmaat and Caroni, 2016). It involves several biochemical mechanisms which interact on multiple timescales. The induction protocols for short-term plasticity (STP, on the order of hundreds of milliseconds) (Turrigiano et al., 1996; Markram et al., 1998) and for the early phase of long-term potentiation or depression (LTP or LTD, on the order of minutes to hours) (Levy and Stewart, 1983; Brown et al., 1989; Artola et al., 1990; Bliss and Collingridge, 1993; Markram et al., 1997; Sjöström et al., 2001) are well established and have led to numerous models (Bienenstock et al., 1982; Gerstner et al., 1996; Van Rossum et al., 2000; Song et al., 2000; Senn et al., 2001; Shouval et al., 2002; Rubin et al., 2005; Pfister and Gerstner, 2006; Brader et al., 2007; Graupner and Brunel, 2007; Clopath

et al., 2010; Gjorgjieva et al., 2011; Nicolas and Gerstner, 2016). On the other hand, various experiments have shown that the further evolution of synaptic efficacies on the timescale of hours depends in a complex way on the stimulation protocol (Frey and Morris, 1997; Dudai and Morris, 2000; Nader et al., 2000; Redondo and Morris, 2011). This phenomenon is called *synaptic* consolidation, to be distinguished from *memory* consolidation, which is believed to take place through the interaction between hippocampus and cortex and which occurs on an even longer timescale (Hasselmo, 1999; Roelfsema, 2006; Brandon et al., 2011). Such a richness of plasticity mechanisms across multiple timescales has been hypothesized to be fundamental in explaining the large storage capacity of memory networks (Fusi et al., 2005; Benna and Fusi, 2016).

Synaptic consolidation is often studied in hippocampal or cortical slices, in which it is induced by extra-cellular stimulation of afferent fibers with short current pulses (Frey and Morris, 1997; Sajikumar and Frey, 2004a,b). Experimental protocols are typically organized in multiple repetitions of stimulation episodes, with variable repetition frequency and duration of each episode (Fig. 3.1A). The dependence of the consolidation dynamics on the parameters of the experimental protocol is complex and has remained elusive. Both the intra-episode pulse frequency and the inter-episode delay play an important role in determining whether a synapse gets potentiated or not after the stimulation (Larson and Munkácsy, 2015; Kumar and Mehta, 2011). Furthermore, recent evidence suggests the existence of optimal parameters to achieve consolidation (Larson and Munkácsy, 2015; Kumar and Mehta, 2011). Existing models succeeded in reproducing experimental results on early and late LTP (Clopath et al., 2008; Barrett et al., 2009; Ziegler et al., 2015; Kastner et al., 2016), by a mathematical description of the interaction of different synaptic mechanisms. However, the complexity of those models prevents a complete characterization of the dynamics, that links stimulation protocols to synaptic consolidation. Here we address the following question: why is the temporal structure of stimulation, i.e. the timing of repetitions, so important for synaptic consolidation?

We introduce a phenomenological model of synaptic consolidation (Fig. 3.1B,C,D) in which, as suggested by experiments (Petersen et al., 1998; O'Connor et al., 2005; Bosch et al., 2014), both model variables are bistable. We find that, despite the simplicity of our model, potentiation of a synapse depends in a complex way on the temporal profile of the stimulation protocol. Our results suggest that not just the total number of stimulation pulses, but also the precise timing within an episode and across repetitions of episodes are important, in agreement with anecdotal evidence that changes in protocols can have unexpected consequences.

3.3 Methods

In what follows, we introduce the synaptic consolidation model that we analyze in the Results section. Since describing the details of molecular interactions inside a synapse as a system of

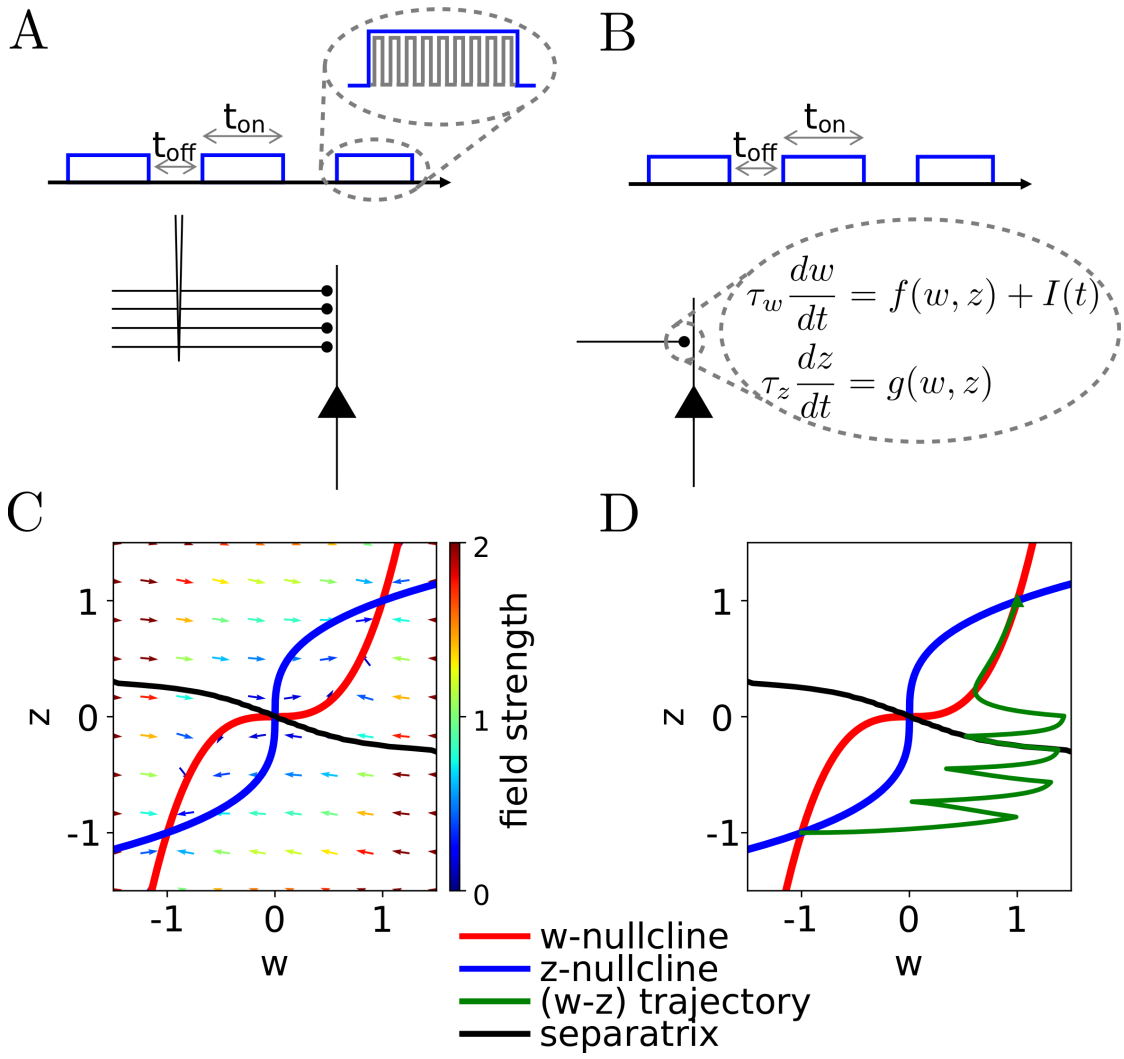


Figure 3.1 – **Schematic experimental setup and modeling framework. A:** Schematic of extracellular stimulation in experiments. The plasticity-inducing stimulus consists of several episodes of duration t_{on} with inter-episode interval t_{off}. Zoom: Each episode contains several high-frequency pulses. **B:** Schematic of single-synapse consolidation model. The synapse is described by a weight variable w with time constant τ_w and a slower consolidation variable z with time constant $\tau_z \geq \tau_w$. Each episode corresponds to a rectangular plasticity-inducing stimulus $I(t)$. **C:** Phase-plane for a specific choice of $f(w,z)$ and $g(w,z)$, $I(t) = 0$ and $\tau_z = 7\tau_w$. The fixed points in $(w,z) = (-1,-1)$ and $(w,z) = (1,1)$ are stable and correspond to an unpotentiated and potentiated synapse, respectively. The black line separates the basins of attraction of the two stable fixed points. **D:** Evolution of the system dynamics in the phase-plane. The system is initialized in the unpotentiated state and it evolves under the effect of a plasticity-inducing stimulus made of three pulses.

differential equations (Bhalla and Iyengar., 1999; Lisman and Zhabotinsky, 2001) would be far too complicated for our purpose, we aim to capture the essential dynamics responsible for synaptic consolidation with an effective low-dimensional dynamical system. In this view, variables are mathematical abstractions that represent the global state of a network of biochemical molecules inside a synapse, e.g. during a transition from one metastable configuration to another (Bosch et al., 2014).

3.3.1 Choice of the model

A one-dimensional dynamical system is not expressive enough to capture experimental data. Indeed, in a one-dimensional differential equation, it would be sufficient to know the instantaneous state of a single variable of the synapse (such as the weight) to predict its evolution, while this is not the case in experiments. As a natural step toward more complexity, we consider a general autonomous two-dimensional model

$$\begin{aligned} \frac{dw}{dt} &= f(w,z) \\ \frac{dz}{dt} &= g(w,z) \quad , \end{aligned} \tag{3.1}$$

where w represents the measured efficacy of a synaptic contact point (e.g. the amplitude of the EPSP caused by pre-synaptic spike arrival), while z is an abstract auxiliary variable. For simplicity, both variables will be considered unit-less. We choose the functions f and g, such that

$$\begin{aligned} \tau_w \frac{dw}{dt} &= -K_w(w-w_0)(w+w_0)w + C_w\left(z - \frac{z_0}{w_0}w\right) + I \\ \tau_z \frac{dz}{dt} &= -K_z(z-z_0)(z+z_0)z + C_z\left(w - \frac{w_0}{z_0}z\right) \quad , \end{aligned} \tag{3.2}$$

where $(w,z) = \pm(w_0,z_0)$ are the stable fixed points of the two-dimensional system in the presence of a fixed coupling $C_w \geq 0$, $C_z \geq 0$ and in the absence of a drive, i.e. $I = 0$. In our simulations, we always choose $w_0 = z_0 = 1$. For $K_w \neq 0$ and $K_z \neq 0$, we could divide Eq. 3.2 by K_w and K_z to further reduce the numbers of parameters. However, we will stick to a notation with explicit K_w and K_z since we do not want to exclude the choice $K_w = 0$ or $K_z = 0$. Without loss of generality, we will choose $K_{w,z} \in \{0,1\}$, i.e either zero or unity. Note that the choice $K_z = 0$ implies that the dynamics of the auxiliary variable z are linear, while $K_z = 1$ implies full non-linearity. The choice of the model is explained in the next section.

3.3.2 Simplification steps of the 2d-dynamics

In this section we present the arguments leading from Eq. 3.1 to Eq. 3.2. Readers not interested in the details may jump to the next sections. One way to tackle the very general system in Eq. (3.1) is to perform a Taylor expansion around $w = 0$ for the first equation

$$\frac{dw}{dt} = A(z) + B(z) \cdot w + C(z) \cdot w^2 + D(z) \cdot w^3 + \dots \tag{3.3}$$

and around $z = 0$ for the second one

$$\frac{dz}{dt} = A'(w) + B'(w) \cdot z + C'(w) \cdot z^2 + D'(w) \cdot z^3 \dots \quad . \tag{3.4}$$

An expansion up to the third order enables us to implement the bistable dynamics (Petersen et al., 1998; O'Connor et al., 2005) of single contact points. Bistability requires the system to have at least two stable fixed points at finite value. This condition cannot be met by degree 1 or degree 2 polynomials since they can have at most one *stable* fixed point. Therefore bistability requires a polynomial of degree 3 or higher in at least one equation. To be more general, we will consider a system in which both polynomials are of degree 3. We restrict our analysis to the situation in which we have linear coupling between the two variables, of the form $A(z) = A_0 + A_1 \cdot z$, $B(z) = B$, $C(z) = C$ and $D(z) = D$. Analogously, in the second equation we set $A'(w) = A'_0 + A'_1 \cdot w$, $B'(w) = B'$, $C'(w) = C'$ and $D'(w) = D'$.

Bistability is be obtained with a negative coefficient of the third power in both equations. Before we start the analysis, we rewrite Eqs. 3.3 and 3.4 in a more symmetric form. To do so we proceed in three steps. (i) Assuming that the degree 3 polynomial has three real roots, we rewrite our system in the more intuitive form

$$\begin{aligned} \tau_w \frac{dw}{dt} &= -K_1(w - w_1)(w - w_2)(w - w_3) + C_1 z \\ \tau_z \frac{dz}{dt} &= -K_2(z - z_1)(z - z_2)(z - z_3) + C_2 w \quad , \end{aligned} \tag{3.5}$$

where C_1 and C_2 are coupling constants and the roots w_1, w_2, w_3 correspond to the fixed points of the equations in the uncoupled case ($C_1 = C_2 = 0$). The parameters τ_w and τ_z can be interpreted as time constants since they do not influence the location of the fixed points but only the speed of the dynamics. K_1 and K_2 are two positive constants that scale the whole polynomial, while C_1 and C_2 are positive constants that control the amount of coupling between the two variables. If we exclude the coupling terms, each equation corresponds to an over-damped particle moving in a double-well potential (Strogatz, 2014). The parameters K_1, K_2, τ_w, τ_z, C_1, C_2, w_1, w_2, w_3 are simple transformations of the parameters A_0, A_1, B, C, D, A'_0, A'_1, B', C', D' of the original system. For example $K_1 = D$. (ii) In order to further simplify our study, we assume that in both equations one of the three roots is zero, one is positive and

one negative, equally distant from zero. Following (Zenke et al., 2015a), we add a plasticity induction term to the first equation that describes the drive provided by an LTP induction protocol. The equations now read

$$\begin{aligned} \tau_w \frac{dw}{dt} &= -K_1(w-\bar{w})(w+\bar{w})w + C_1 z + I \\ \tau_z \frac{dz}{dt} &= -K_2(z-\bar{z})(z+\bar{z})z + C_2 w \quad . \end{aligned} \tag{3.6}$$

In the absence of coupling, the double well potential related to Eq. 3.6 has minima in $w = \pm\bar{w}$, $z = \pm\bar{z}$ and a local maximum in $w = 0$ ($z = 0$). Notice that this seems to imply that a synaptic weight can take both positive and negative values, which is biologically implausible. However, this choice simplifies the calculations without loss of generality, since it is always possible to go back to a system with positive weights by applying a coordinate translation.

(iii) In the absence of a drive ($I = 0$), the system has eight free parameters, which all influence the location of the fixed points. In a final transformation step we rewrite Eq. 3.6 such that the location of two stable fixed points becomes independent of the coupling constants C_1 and C_2. The reason for doing this is that the stable fixed points of the system are easier to access experimentally than other constants. In particular, the value of w at the stable fixed point should be related to the synaptic weight measured experimentally. We, therefore, rewrite the system in the form of Eq. 3.2, where w_0 and z_0 are the absolute values of the stable fixed point and the parameters can be mapped from Eq. 3.6 to Eq. 3.2, for example, $K_w = K_1$ and $C_w = (K_1\bar{w}^2 - K_1 w_0^2)\, w_0/z_0$ and analogously for C_z and K_z.

3.3.3 Nullclines and phase-plane analysis

Since the system is two-dimensional, it can be studied using phase-plane analysis, following a well established tradition in computational neuroscience (Rinzel and Ermentrout, 1998; Ermentrout, 1996, 2002; Wilson and Cowan, 1972). The fixed points of the system are graphically represented by the intersections of the nullclines (i.e. the curves defined by either $\frac{dw}{dt} = 0$ or $\frac{dz}{dt} = 0$), which in our system are:

$$\begin{aligned} w-\text{nullcline:}\ \ z &= \frac{z_0}{w_0} w + \frac{K_w}{C_w}(w-w_0)(w+w_0)w - \frac{I}{C_w} \\ z-\text{nullcline:}\ \ w &= \frac{w_0}{z_0} z + \frac{K_z}{C_z}(z-z_0)(z+z_0)z \quad . \end{aligned} \tag{3.7}$$

The maximum number of fixed points for the system in Eq. 3.2 can be easily computed. To do so, consider a more general form of two nullclines:

$$\begin{aligned} w-\text{nullcline:}\ \ z &= P_n(w) \\ z-\text{nullcline:}\ \ w &= Q_m(z) \quad , \end{aligned} \tag{3.8}$$

where $P_n(z)$ is a polynomial of degree n in w and, analogously , $Q_m(w)$ is a polynomial of degree m in z; cf. Eq. 3.7. To find the fixed points of the system Eq. (3.8) we need to solve:

$$w = Q_m(P_n(w)) \quad . \tag{3.9}$$

Eq. (3.9) is a polynomial equation of degree $n \cdot m$ in w and therefore it allows a number of real solutions s, $0 \le s \le n \cdot m$. Applying this formula to our case, we find that we can have a maximum of nine fixed points.

In order to reduce the number of parameters from 8 to 4, we first consider the symmetric case (section 3.3.4) in which the two equations have the same parameters. Moreover, since we make the choice $z_0 = w_0 = 1$, the actual number of free parameters is three. In the next section, we show the effect of changing the coupling coefficients. Then, we briefly comment on the effect of the time constants and of a constant plasticity-inducing stimulus I. We will move to the analysis of the asymmetric cases in section 3.3.5.

3.3.4 Symmetric changes of coupling coefficients reveal two bifurcations

We study the case of symmetric coupling $C_w = C_z = C$ and analyze how a change of coupling strength influences the dynamics of the system. As an aside, we note that for symmetric coupling we can define a pseudopotential (Cohen and Grossberg, 1983)

$$V(w,z) = \frac{K_w}{4}w^4 + \frac{K_z}{4}z^4 - \frac{1}{2w_0}\left(K_w w_0^3 - z_0 C\right)w^2 - \frac{1}{2z_0}\left(K_z z_0^3 - w_0 C\right)z^2 - Cwz + I \tag{3.10}$$

in which the dynamical variables move according to $\tau_w \frac{dw}{dt} = -\frac{\partial V}{\partial w}$ and $\tau_z \frac{dz}{dt} = -\frac{\partial V}{\partial z}$.

We fix $\tau_w = \tau_z$, $K_w = K_z = 1$, $I = 0$, $w_0 = z_0 = 1$ and vary C in Eq. 3.2. In the case $C = 1$, the system is in a rather simple regime: there are two stable fixed points in $(w,z) = (-1,-1)$ and $(w,z) = (1,1)$ and a saddle fixed point at the origin (Fig. 3.2). The basins of attraction of the stable fixed points are separated by the $z = -w$ diagonal.

If we decrease the coupling C, we encounter two bifurcations. A first pitchfork bifurcation takes place at $C = 1/2$, when the two nullclines are tangent to each other in the saddle point. Beyond the bifurcation point of the coupling coefficient, we observe the creation of two additional saddle points (Fig. 3.2B). The stability properties, the location and the basins of attraction of

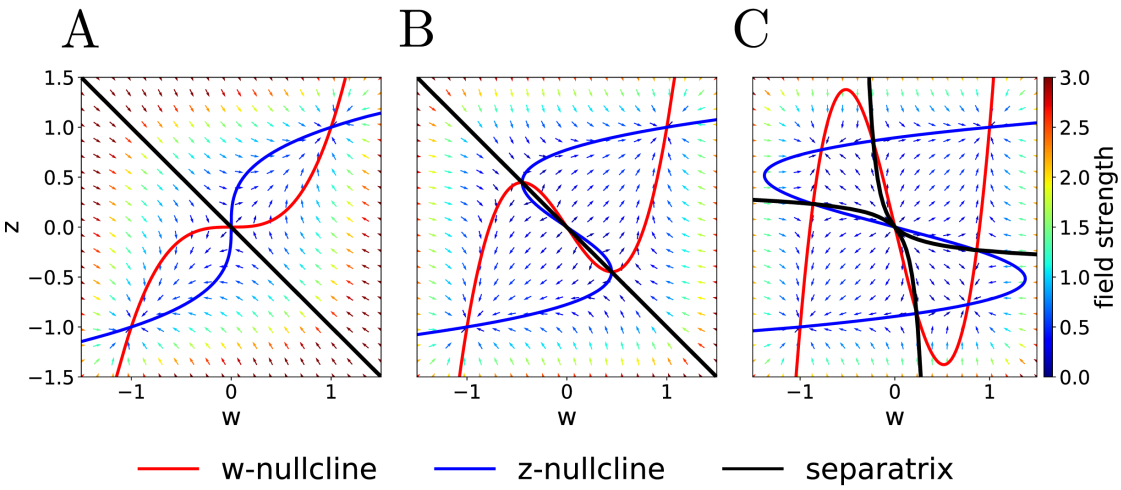


Figure 3.2 – **Phase-plane diagram and basins of attractions for the symmetric case with equal coupling constants,** $C_w = C_z = C$. The plasticity-inducing stimulus is null, $I = 0$. **A:** $C = 1$, phase-plane with field arrows. The color of the arrows is proportional to the field strength. $w-$ and $z-$ nullclines are indicated in red and blue respectively. The line that separates the two basins of attraction is indicated in black. **B:** Same as A, but $C = 0.4$. Compared to A, we notice the creation of two saddle points. **C:** Same as A, but $C = 0.2$. The maximum number of fixed points is achieved. In this case we have four basins of attraction.

the other two fixed points remain unchanged, but the local field strength changes, as shown by the colored arrows. The second pitchfork bifurcation takes place at $C = 1/3$. For this coupling value, each of the two new saddle points splits into a stable fixed point and two further saddle points. Therefore, for very weak coupling we observe four basins of attractions, whose shape is shown in Fig. 3.2C. The stability of the fixed points in $(w, z) = (-1, -1)$ and $(w, z) = (1, 1)$ is not affected by the bifurcations.

On the other hand, if we increase the coupling coefficient to a value $C > 1$, then the two nullclines will progressively flatten, but the location of the three fixed points is unchanged with respect to the case $C = 1$.These observations have been summarized in the bifurcation diagram of Fig. 3.3A. We observe that there are actually three pitchfork bifurcations, but that two of them are degenerate since they happen for the same value of C.

3.3.5 Asymmetric parameter choices shape the basins of attraction

As a more general case, we consider asymmetric coupling C or timescale τ. When the coupling coefficients are asymmetric, we can plot the position of the bifurcation points in the $C_w - C_z$ plane, (Fig. 3.3B). The choice $C_w = C_z$ of the previous section corresponds to the dashed gray line. We notice that in the asymmetric case it is possible to have three distinct bifurcations (for

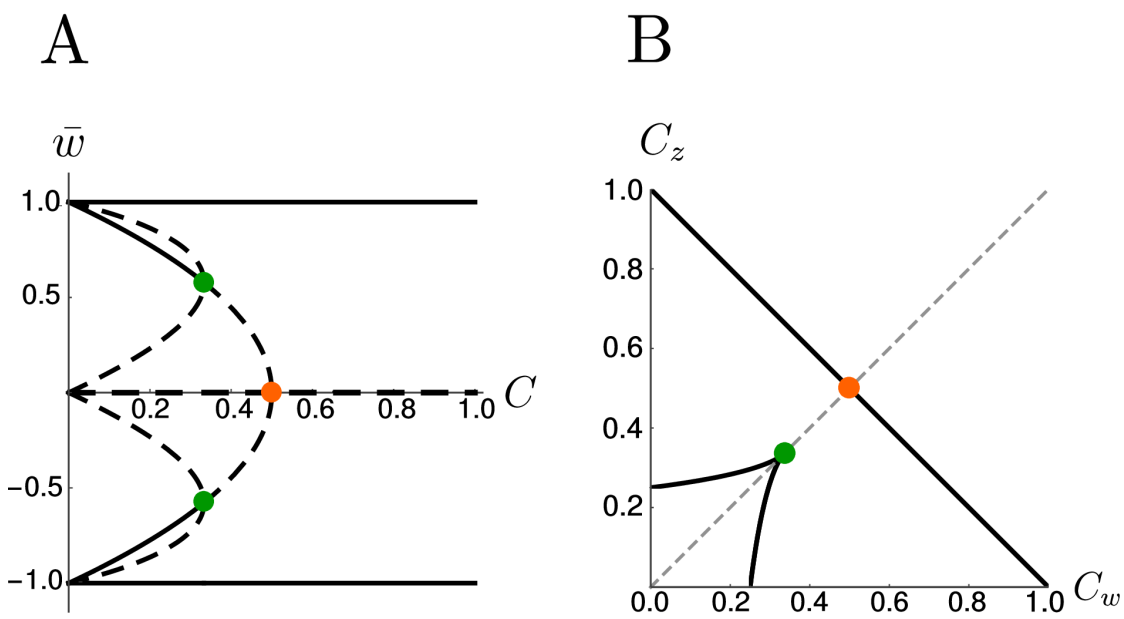


Figure 3.3 – **Bifurcations diagrams. A:** Fixed points in the symmetric case. Dashed lines indicate unstable fixed points while continuous lines indicate stable fixed points. Orange and green dots indicate bifurcation points. **B:** Bifurcation points in the $C_w - C_z$ plane (black) for the general (asymmetric) case. The dashed gray line corresponds to $C_w = C_z$. The orange and green dots indicate the corresponding bifurcations in A. Note that, in B, the bifurcation at $C_w = C_z = \frac{1}{3}$ (green dot) is a degenerate point.

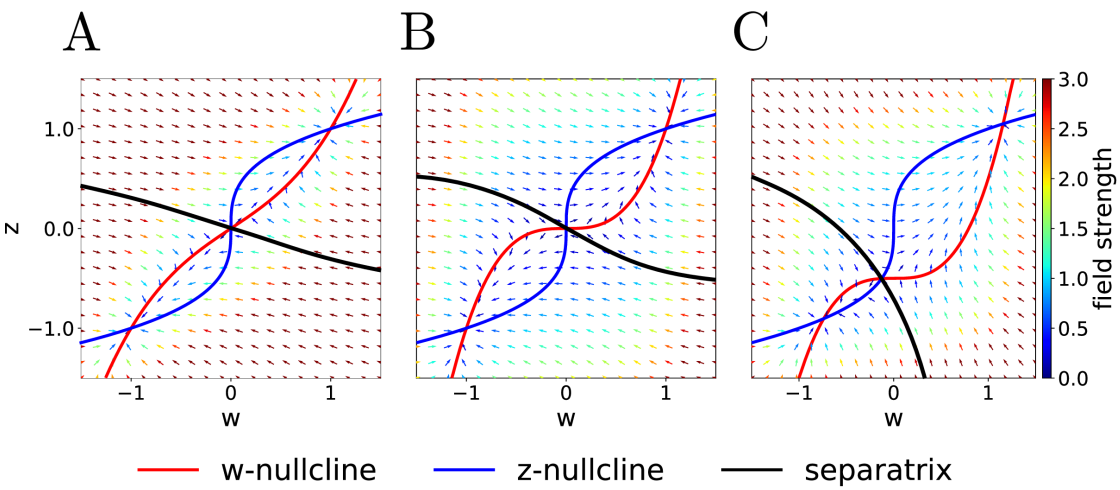


Figure 3.4 – **Asymmetric parameter choices. A:** In the case $C_w = 3 > C_z = 1$, the curvature of the $w-$nullcline (red) is smaller than that of the $z-$nullcline (blue) and the basins of attraction are deformed compared to Fig. 3.2A. ($\tau_z = \tau_w = 1$) **B:** For $\tau_z/\tau_w = 3$ and $C_w = C_z = 1$, nullclines are not affected (compare to Fig. 3.2A) but the basins of attraction are. **C:** For $I = 0.5$ (all other parameters set to 1), the basin of attraction of the fixed point at $(-1,-1)$ is smaller than of the fixed point at $(1,1)$.

example, we can fix $C_w = 0.3$ and decrease C_z, from 1 to 0). We find that, for $C_w + C_z > 1$, the number of fixed points is always three and no bifurcation is possible. On the other hand, if $C_w + C_z < 1$, the system enters in the regime with minimum five fixed points. Moreover, we can analytically compute the bifurcation value of one coupling constant, given the other. An asymmetric choice $C_w \neq C_z$ influences the shape of the basins of attraction (Fig. 3.4A).

If we keep $C_w = C_z$ but consider instead $\tau_z > \tau_w$, the system in Eq. 3.2 may be interpreted as two different molecular mechanisms that act on different timescales. For example, the variable z can be interpreted as a tagging mechanism or a consolidation variable while w is the weight variable or amplitude of a post-synaptic potential. A comparison of Fig. 3.2A and Fig. 3.4B shows that the changes in τ do not affect the nullclines but change the flow field and the basin of attraction.

Another way by which we can introduce asymmetry in the system is by adding a plasticity-inducing stimulus I. It follows from Eq. 3.7 that a value $I > 0$ will cause a down shift of the $w-$nullcline. The case of $C_w = C_z = 1$, $\tau_w = \tau_z = 1$ s , $K_w = K_z = 1$ and $I > 0$ is shown in Fig. 3.4C. A plasticity-inducing stimulus $I > 0$ also implies a reduction of the basin of attraction of the lower stable fixed point in favor of an increase of the basin of attraction of the upper stable fixed point. For high values of I, the basin of attraction of the lower fixed point disappears via a steady state bifurcation. Therefore, when $I > 0$ is large enough, the system is forced to move to the upper fixed point that can be interpreted as a potentiated state of the synapse.

Analogously, when $I < 0$, the attraction basin of the lower fixed point is enlarged and leads, eventually, to a bifurcation in which the upper fixed point and the saddle point are lost.

A possible generalization of the model would be to consider the coupling coefficients C_w and C_z as dynamical variables, as it has been explored in previous work (Ziegler et al., 2015). In these models, the coupling parameters C_w and C_z of the two dynamical variables alternates between $C_w = 0$ and $C_z = 1$ or $C_w = 1$ and $C_z = 0$, implementing a write-protection mechanism. The price we pay is the introduction of additional differential equations and parameters for the dynamics of the coupling coefficients. In the specific implementation of (Ziegler et al., 2015), the dynamical coupling is controlled by a low-pass filter of the plasticity-inducing stimulus I and the concentration of neuromodulators on plasticity.

3.3.6 Numerical simulations

All figures were obtained using Python 2.7, except for the bifurcation plot in Fig. 3, which was created with Wolfram Mathematica. In the phase-plane plots, the separatrix between the basins of attraction was obtained doing a mesh-grid search: we initialized the dynamical system Eq. 3.2 in each point of a 100×100 grid in the w, z space ($w, z \in [-1.5, 1.5]$) and checked to which stable fixed point it converges. Therefore we interpolated the separation line. The trajectory of the system in the phase-plane was obtained by solving the system in Eq. 3.2 using the Runge–Kutta 4 method with integration step $dt = 0.01$. In Figs. 3.6-3.7, we inject an external stimulus into the dynamical equations. The system trajectory is always initialized in the depotentiated state $(-1, -1)$ and the simulation is stopped when the trajectory enters into the basin of attraction of the potentiated state $(1, 1)$. The position of the stable fixed points depends on the choice $w_0 = z_0 = 1$, which we made for simplicity. In fact, we can remap the values of the synaptic weight w the desired (positive range) with an affine transformation, without loss of generality.

3.4 Results

The two-dimensional model, introduced Methods section, predicts a complex dependence of the synaptic consolidation dynamics upon the parameters of the experimental protocol. This complex dependence has similarities with the behavior observed in experiments (Sajikumar et al., 2005; Larson and Munkácsy, 2015); cf. Fig. 3.1. First, we describe how we abstract the experimental protocol into a time-dependent plasticity-inducing stimulus $I(t)$. Then, we show the response of our model to different stimulation protocols. In our model, the plasticity-inducing stimulus $I(t)$ drives the synaptic weight w via a non-linear equation characterized by a time constant τ_w. The weight w is coupled to a second variable z with time constant τ_z (Eq. 3.2). The variable z is an abstract description of the complex metastable states (potentiated or

unpotentiated) caused by consolidation (Redondo and Morris, 2011; Bosch et al., 2014). After an analysis of a single rectangular stimulation (one episode), in section 3.4.2, we will move to the more realistic case of repetitive stimulation across multiple episodes. Throughout the results section, we will focus on synaptic potentiation. Since the self-interaction term in Eq. 3.2 is symmetric with respect to $w = z = 0$, synaptic depression of a potentiated state is the mirror image of synaptic potentiation of a unpotentiated state.

3.4.1 Abstraction of the stimulation protocol

In their seminal work, (Bliss and Lømo, 1973) showed that repeated high-frequency stimulation of afferent fibers can lead to long-lasting synaptic potentiation. In later work it was shown that low-frequency stimulation can lead to long-lasting synaptic depression (Bashir and Collingridge, 1994). In order to keep the analysis transparent, we use a time-dependent, real-valued quantity $I(t)$ as an abstraction for such experimental protocols. In what follows, we will refer to $I(t)$ as to the plasticity-inducing stimulus. Note that, we do not perform an explicit mapping from the electrical current used in LTP experiments for the stimulation of pre-synaptic fibers onto the plasticity-inducing stimulus $I(t)$ that influences the dynamics of Eq. 3.2. A precise mapping would require additional assumptions on (i) how extra-cellular stimulation triggers axonal spikes in multiple fibers, (ii) how pre-synaptic spike arrivals cause post-synaptic firing and (iii) how pre- and post-synaptic neural activity leads, potentially via a Hebbian model, to the induction of early-LTP. This means that, in principle, the model's dynamics is rich enough to reproduce the four classical synaptic-consolidation experiments (Frey and Morris, 1997; Nader et al., 2000), however, we would need to set at least four free parameters, corresponding to the amplitudes of the external input I, needed for strong and weak LTP and LTD. Instead, we model a set of extra-cellular high-frequency pulses as a single rectangular plasticity-inducing stimulus of positive amplitude (Fig. 3.1B). The larger the stimulation frequency, the larger the amplitude of $I(t)$. Analogously, a set of extra-cellular current pulses at low frequency is modeled as a single negative rectangular plasticity-inducing stimulus. The compression of multiple extra-cellular pulses into a single rectangular episode $I(t)$ is justifiable since the time between single pulses, even in the case of low-frequency stimulation, is very short compared to the timescale of plasticity. This implies that multiple short pulses in experiments can be well approximated by a single episode, described by one prolonged rectangular stimulus in our model (Fig. 3.1A and B). In agreement with well-established plasticity models (Bienenstock et al., 1982; Senn et al., 2001; Pfister and Gerstner, 2006; Clopath et al., 2010; Gjorgjieva et al., 2011), we use $I > 0$ to describe a high-frequency stimulation since a positive I leads to potentiation, see Methods and Fig. 3.4C. Conversely, a negative I favors depotentiation. On the other hand, experiments that involve global variables, such as cross-tagging (Sajikumar and Frey, 2004b,a), can not be explained by our model.

3.4.2 One episode

We consider the case in which our two-variable synapse model is stimulated with a single rectangular plasticity-inducing stimulus $I(t)$ of variable amplitude and duration t_{on} (Fig. 3.5A). Experimentally, this would correspond to single-episode, high-frequency protocols of variable stimulation intensity (i.e. pulse frequency) and duration. For each choice of duration and amplitude, we initialize the system in the unpotentiated state, defined by the initial value $(w, z) = (-1, -1)$ and we numerically integrate the system dynamics until convergence. We then measure the final state of the synapse, i.e. whether it converged to the potentiated or returned to the unpotentiated state. In Fig. 3.5 we plot the curve that separates the region of the parameter space that yields potentiation (shaded area) from the one that does not. Different curves correspond to different time constants τ_w and τ_z of the synaptic variables w and z in Eq. 3.2.

Fig. 3.5C illustrates a rather intuitive result, i.e. if the amplitude of the plasticity-inducing stimulus is increased, the duration needed for potentiation decreases. Moreover, if the amplitude is too small, we cannot achieve potentiation, even for an infinite pulse duration. The limit of infinite pulse duration is in the following called the "DC" limit. The effect of DC-stimulation can be easily understood from a phase-plane analysis (Fig. 3.4). Indeed, the introduction of a constant term $I > 0$ in the w equation (DC term), yields a shift in the w-nullcline vertically downward. However, if the term is too small to cause the loss of the low fixed point, potentiation cannot be achieved (Fig. 3.4C).

The separation curves in Fig. 3.5C indicate that the minimal duration of an episode necessary for potentiation decreases as the intensity of the plasticity-inducing stimulus increases. We wondered whether the relevant parameter for potentiation is the area under the rectangular plasticity-inducing stimulus. To study this, we performed a similar analysis, with the amplitude of the plasticity-inducing stimulus and its area as independent variables. For each choice of area and amplitude, the duration of the episode is given by $t_{\text{on}} = \text{area}/\text{amplitude}$ (see Fig. 3.5B). The results are shown in Fig. 3.5D. If there were a regime in which the relevant parameter is the area of the pulse, then the curve separating parameters of successful from unsuccessful potentiation would be horizontal. However, we find a near-horizontal curve only for $\tau_z = \tau_w$, limited to the high-amplitude region. For $\tau_z > \tau_w$ we find the existence of an optimal value of the amplitude that yields potentiation with the minimal area. If we increase the amplitude beyond this optimal value, the necessary area under the stimulus curve $I(t)$ starts to increase again.

In order to understand this effect, we look again at the phase-plane, in particular at the dependence of the separatrix on the timescale separation. In the limit in which the amplitude goes to infinity and the duration goes to 0 while the area of the whole plasticity-inducing stimulus stays the same, the stimulus can be described by a Dirac-δ function. In Fig. 3.5D we can see

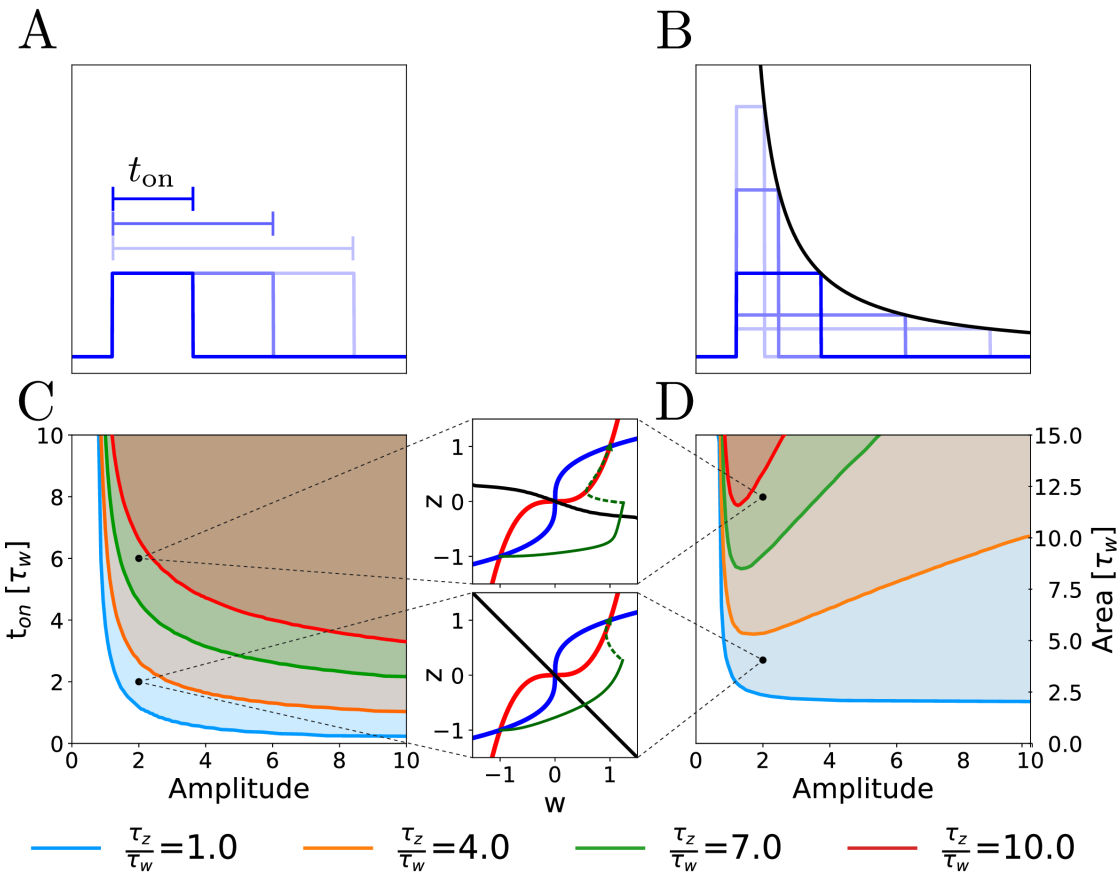


Figure 3.5 – **Potentiation during a single episode.** Different curves correspond to different ratios of the time constant τ_z and τ_w in Eq. 3.2. **A:** Schematic representation of single-episode stimulations, corresponding to different choices of t_{on}. **B:** Schematic representation of different single-episode stimulations with constant area. The black line is proportional to 1/amplitude in order to stress that all pulses have the same area. **C:** Separation curves between regions of successful or unsuccessful potentiation as a function of amplitudes and duration t_{on} of a the plasticity-inducing stimulus. The initial state is always the unpotentiated synapse ($w = z = -1$). The shaded region of the parameter space is the one in which the synapse gets potentiated. **D:** Same as **C** but as a function of amplitude and area of the stimulus during the episode. The two insets show examples of trajectories (green lines) in the phase-plane for two different parameters choices. The solid green lines represent the dynamical evolution of the system during the application of the external stimulus, while the dotted green line shows the relaxation of the system to a stable fixed point after the stimulation. Red: w-nullcline; blue: z-nullcline; black: separatrix. The parameters that are not specified in the figure are: $C_w = C_z = 1$, $I = 0$.

that, if $\tau_z \gg \tau_w$, the separatrix tends to an horizontal line for $w \gg 1$. Since a δ-pulse input is equivalent to an instantaneous horizontal displacement of the momentary system state in the phase-plane, a single δ-pulse cannot bring the system across the separatrix. The δ-pulse stimulation is, of course, a mathematical abstraction. In a real experimental protocol, such a stimulation can be approximated by a short episode of very intense high frequency stimulation. Due to the necessary finite duration of an episode, the system response in the phase-plane will not be a perfectly horizontal displacement. However, achieving potentiation with short pulses can still be considered as difficult, because it would require a disproportionately large stimulation amplitude.

Our findings highlight the fact that changing parameters, such as the ratio of τ_w and τ_z, gives rise to different behaviors of the model in response to changes in the stimulation protocols. We may use this insight to design optimal experimental protocols for single-episode plasticity induction. In particular, a model with timescale separation would predict the existence of an optimal stimulus intensity for which the total stimulus area necessary for potentiation is minimized. We emphasize that any model where consolidation works on a timescale that is slower than that of plasticity induction will exhibit timescale separation and be therefore sensitive to details of the stimulation protocol.

3.4.3 Repeated episodes

As a second case, we consider the potentiation of a synapse induced by repetitions of several stimulation episodes. In an experimental setting, this type of stimulation would correspond to several episodes of high-frequency stimulations, characterized by three parameters: the intensity of stimulation during each episode (amplitude), the duration (t_{on}) of each episode and the inter-episode interval, t_{off}; cf. Fig. 3.6A,B. To keep the analysis transparent we apply a number of repetitions large enough to decide whether potentiation is successful or not given the three parameters. Notice that if $t_{\text{off}} = 0$ we are back to the DC stimulation as defined in the previous section.

The curves in Fig. 3.6C,D show the separation between parameters that lead to successful potentiation (shaded) or not (white) in the amplitude-t_{off} space for fixed values of t_{on} and for different τ_z/τ_w ratios. In Fig. 3.6C we fix $t_{\text{on}} = \tau_w$. We observe that, at least for low timescale ratios, it exists an amplitude above which the synapse gets potentiated independently of t_{off}, which suggests that, for this intensity of the stimulation, the potentiation happens during the first episode. The amplitude necessary to obtain potentiation in one pulse, however, increases rapidly with the τ_z/τ_w ratio (see Section 3.4.2). On the other hand, if the value of t_{off} is small enough (i.e. for high repetition frequency), potentiation can be achieved with smaller amplitudes and the timescale ratio is less important (notice the superimposed lines in the bottom left part of the plot). If we decrease the pulse duration to $t_{\text{on}} = 0.01\tau_w$, we obtain

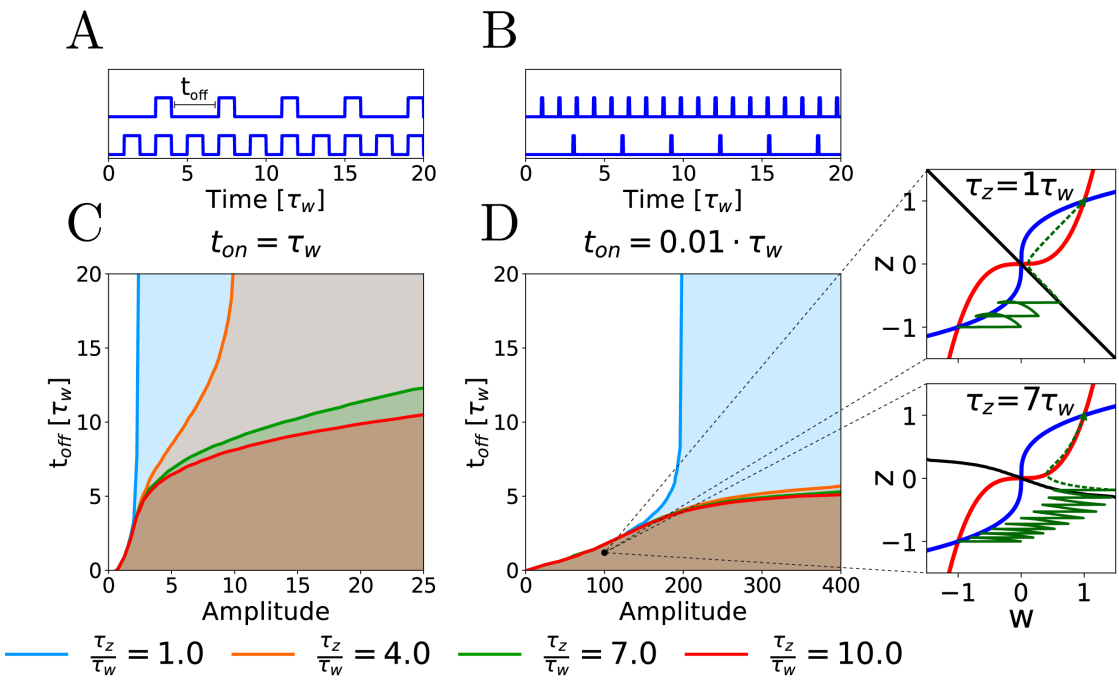


Figure 3.6 – **Potentiation with repeated episodes. A:** Schematic representation of stimulation protocols characterized by different t_{off}, while $t_{\text{on}} = \tau_w$ is fixed. **B:** Schematic representation of stimulation protocols with $t_{\text{on}} = 0.01\tau_w$. **C:** Potentiation region for stimulation with long episodes for fixed $t_{\text{on}} = \tau_w$. The curves for different ratios τ_z/τ_w (see color code) indicate the separation between the region that yields potentiation (shaded) and the region that does not (white) as a function of intensity of stimulation in each episode (amplitude) and inter-episode interval (t_{off}). **D:** Potentiation regions for protocols with shorter episode $t_{\text{on}} = 0.01\tau_w$. The potentiation region is shaded. The two insets show examples of trajectories (green lines) in the phase-plane for the same choice of stimulation parameters but different timescale separation. The solid green lines represent the dynamical evolution of the system during the application of the external stimulus, while the dotted green line shows the relaxation of the system to a stable fixed point after the end of the stimulation protocol. The parameters that are not specified otherwise are: $C_w = C_z = 1$, $I = 0$.

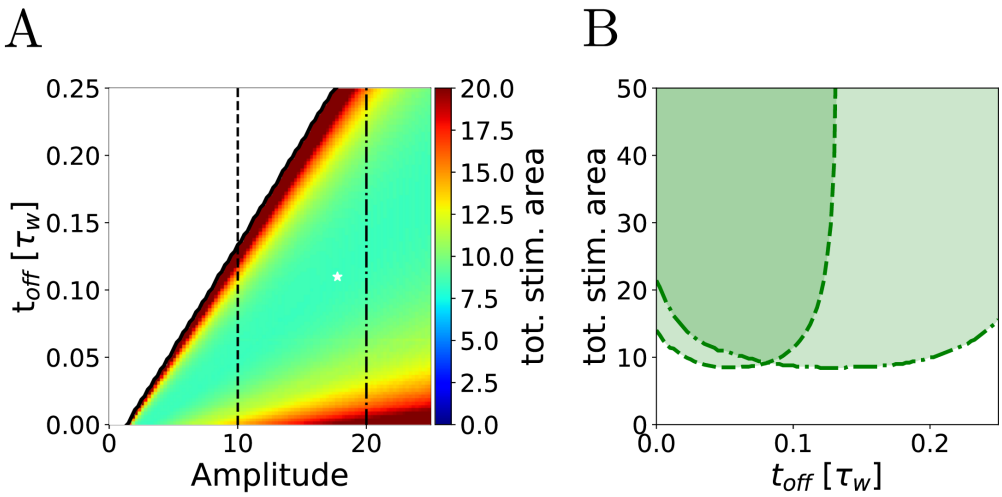


Figure 3.7 – **Stimulation effort needed to achieve potentiation for $\tau_z/\tau_w = 7$ and $t_{\text{on}} = 0.01\tau_w$. A:** The potentiation domain (shaded region in Fig. 3.6) is colored proportionally to the stimulation area needed to achieve potentiation with a repetitive pulse stimulus. The minimum stimulation area is $\simeq 8.34$, it is indicated by the white star and corresponds to the parameters values $t_{\text{off}} = 0.11\tau_w$, amplitude = 17.75 and 47 pulses. **B:** Slices of the diagram in panel A for amplitude = 10 (dashed line) and for amplitude = 20 (dash-dotted line) are plotted against t_{off}. One can notice that for a fixed stimulation amplitude, there is an optimal repetition frequency $f = \frac{1}{t_{\text{on}}+t_{\text{off}}}$ that minimizes the stimulus area required to achieve potentiation. The parameters that are not specified otherwise are: $C_w = C_z = 1$, $I = 0$.

qualitatively similar separation curves, but potentiation now requires much larger values for the amplitude of the plasticity-inducing stimulus (see Fig. 3.6B), than $t_{\text{on}} = \tau_w$ (see Fig. 3.6A). Importantly, in the case of timescale separation (e.g. $\tau_z = 7\tau_w$) several repetitions are needed before the consolidation variable z has sufficiently increased so that the synapse state crosses the separatrix (Fig. 3.6, insert).

In analogy to the analysis performed in section 3.4.2, we search for an optimal stimulation protocol in the case of repeated episodes. In order to allow a direct comparison between single and repetitive episodes, we measured the total area under the stimulation curve $I(t)$ in the repetitive episode scenario, limited to the minimal number of episodes sufficient to achieve potentiation. In Fig. 3.7A, we show the minimum stimulation area (number of episodes times the area of each rectangular plasticity inducing stimulus) required to achieve potentiation, as a function of the amplitude and the frequency of the stimulus for strong timescale separation ($\tau_z/\tau_w = 7$). We notice that the minimum stimulation area (white star) corresponds to $t_{\text{off}} \sim 10\, t_{\text{on}}$, i.e., the waiting time between episodes is ten times long than each episode. In real experimental conditions, however, it might be difficult to control the intensity

of the stimulation. For this reason, we consider a fixed intensity (e.g. amplitude $I = 10$ in Fig. 3.7A) and vary the inter-episode time t_{off}. We find that there exists an optimal stimulation frequency to obtain potentiation with minimal total area (see Fig. 3.7B). These results highlight two main facts: (i) for many stimulation intensities (only two are shown in the graph), one can find an optimal repetition frequency, (ii) there is a broad region in the parameter space (t_{on}, amplitude, and area) where the number of pulses needed to achieve consolidation is constant. Indeed, the broad region around the minima in Fig. 3.7B (fixed amplitude and t_{on}) where the area is approximately constant corresponds a constant number of pulses ($n_{\text{pulses}} = \text{area}/\, t_{\text{on}}$).

3.5 Discussion

We introduced and analyzed a minimal mathematical model of synaptic consolidation, that consists of two ODEs with linear coupling terms and cubic non-linearity. Since it is a two-dimensional model, the system can be studied using phase-plane techniques. While our model can have up to four stable fixed points, we focused on the case of two stable fixed points, to allow the physical interpretation of the fixed points as an unpotentiated or potentiated synapse. The weight variable w should be seen as the bistable building block of complex synapses. While there is evidence that the potentiation of a single synapse is a all-or-none process (Petersen et al., 1998; O'Connor et al., 2005; Bosch et al., 2014), recent results challenge this view in favor of a modular structure of the synapse (Lisman, 2017). Either way, it is possible to identify a bistable basic element of the synapse.

We showed that our minimal model responds to stimulation protocols in a non-trivial way: we quantified the total stimulation strength by the stimulus area defined as duration times intensity, where intensity is a combination of intra-episode frequency and current amplitude of extra-cellular pulses. We found that the minimal stimulus area necessary to induce potentiation depends non-monotonically on the choice of stimulus parameters. In particular, we found that, for both single-episode and multiple-episode stimulation, it is possible to choose the stimulation parameters (intensity, duration and inter-episode frequency) optimally, so as to minimize the stimulus area (Fig. 3.5 and 3.7). Fig. 3.7 can be used to compare the minimum stimulation area needed to achieve potentiation in a single episode (corresponding to the choice $t_{\text{off}} = 0$) to the case of repetitive pulses ($t_{\text{off}} \neq 0$). We conclude that, for a fixed stimulation area, stimulation over several episodes is advantageous to achieve potentiation, in agreement with some widely used protocols (Larson and Munkácsy, 2015). The effect is stronger if the consolidation variable z is slow compared to the weight variable ($\tau_z \gg \tau_w$). Note that in experiments it is often impossible to have a fine control on the stimulation amplitude: extra-cellular stimulation of fibers must be strong enough to excite the post-synaptic neuron, but there is no control of the post-synaptic firing rate, which could undergo adaptation or exhibit other time-dependent mechanisms. In summary, the existence of an optimal stimula-

tion frequency is the direct consequence of two very fundamental synaptic properties: (i) the bistability of a synaptic basic element, and (ii) the time scale separation between the internal synaptic mechanisms.

Parameter	**Fig. 2**	**Fig. 4**	**Fig. 5**	**Fig. 6**	**Fig. 7**
C_w	1	1	1	1	1
C_z	variable	variable	1	1	1
w_0	1	1	1	1	1
z_0	1	1	1	1	1
τ_w	1	1	1	1	1
τ_z	1	variable	variable	variable	7
dt	0.01	0.01	0.01	0.01	0.01

Table 3.1 – Parameter values used in Figs. 3.5- 3.7 unless otherwise indicated in figures captures.

The minimum of the total stimulus area is particularly pronounced in the regime of strong separation of timescale ($\tau_z \gg \tau_w$), which is the relevant regime in view of the experimental consolidation literature which suggests multiple consolidation mechanisms with a broad range of timescales (Bliss and Collingridge, 1993; Reymann and Frey, 2007). Assuming that the timescale τ_w is on the order of a few seconds, as suggested by some plasticity induction experiments at the level of single contact points (Petersen et al., 1998), we can interpret a short stimulation episode of duration $0.01 \cdot \tau_w \sim 20$ms as a burst of few pulses at high frequency. For example, one particularly interesting protocol is the theta burst stimulation, where each burst consists of 4 pulses at 100Hz corresponding to a burst of 30 ms duration (Larson and Munkácsy, 2015). Assuming that this stimulation does not correspond to an extremely small amplitude value (a reasonable assumption since the experimentalists want to induce LTP), our model predicts an optimal frequency (see Fig. 3.7) on the order of $t_{\text{off}} = 0.11\tau_w \sim 200$ms, which is in rough agreement with the experimental protocols where theta bursts are repeated every 200ms (Larson and Munkácsy, 2015). When comparing the optimal stimulation frequency obtained by our model to experimental data, we should keep in mind that in experiments timing effect come from different sources. In (Larson and Munkácsy, 2015) the key factor that determines the optimal stimulation protocol is the feed-forward inhibition. Moreover in both (Larson and Munkácsy, 2015) and (Sajikumar et al., 2005; Kumar and Mehta, 2011) the position of the stimulation (apical vs basal) plays a fundamental role, together with priming of NMDA channels. Finally, the fraction of NMDA vs AMPA receptors is another fundamental element. None of these factors is taken into account in our simplified model.

We have described the simplified dynamics of a bistable basic element of synaptic consolidation (which can be interpreted as a single contact point or a synaptic sub-unit). However, in most experiments, we can only observe the collective effect of many such elements together

(Malenka, 1991; Bliss and Collingridge, 1993). Such a collective effect can be interpreted as the average number of potentiated contact points. For a detailed comparison between our model and these experiments, it would be needed to simulate the dynamics of the pre- and post-synaptic neuron groups and of their contact points, in order to obtain an average quantity that can be compared with the continuous change of EPSP observed in experiments. Such approach has been taken in (Ziegler et al., 2015) and it requires several assumptions, among others the specification of the dependence of the plasticity induction current I on the pre- and post-synaptic activity, the parameters of the two populations and possible recurrent interactions (see also 3.4.1). For these reasons, such a comparison goes beyond the aim of this work. Moreover, since the model describes a single synaptic contact, it cannot be applied to more complex experiments that involve cross-tagging, where effects of protein synthesis are shared between several synapses (Sajikumar and Frey, 2004b). On the other hand, our results highlight the fact that our model shares similar response properties with the population-averaged quantities measured in experiments, such as its sensitivity to the stimulation frequency and the preference for multiple repetitions. Altogether, these findings suggest that our model possesses the necessary dynamical repertoire to reproduce some of the experimental results, such as (Malenka, 1991; Bliss and Collingridge, 1993).

Using our model we can only make some qualitative predictions on experimentally measurable quantities. For example, by comparing the panels C and D in Fig. 3.6, we can see that the optimal stimulation parameters change when varying the episode duration t_{on}. More precisely, our model predicts that for shorter t_{on} the optimal stimulus requires a large stimulus intensity during each episode.

The proposed framework is related to a number of previous modeling approaches to synaptic consolidation. In particular, the memory formation in networks of excitatory and inhibitory neurons in (Zenke et al., 2015a) is based on a synaptic plasticity model with a linear weight variable and a slower consolidation variable, corresponding to a choice of $K_w = 0$ in Eq. 3.2 of our model. If we exploit this relation between the two models, the coupling term C_w should depend on the post-synaptic activity. Such a time-dependent coupling coefficient is similar to the gating variable in the write-protected model (Ziegler et al., 2015). The write-protected model (Ziegler et al., 2015) can be considered as a three-dimensional generalization of our framework. In our model the weight variable w is directly coupled to the consolidation variable z whereas in the write-protected model w is coupled to an intermediate tag-related variable which is then coupled to z.

The dynamical understanding of the interplay between stimulation protocol and autonomous dynamics gained here by studying the two-dimensional system can be also applied to a three-(or higher-)dimensional generalization, under the assumption that coupling exists only between pairs of variables and that there is timescale separation. Using such a multi-

dimensional generalization, it would be possible to explain a much larger set of experimental results. In addition, the model presented in (Ziegler et al., 2015) features coupling coefficients that are dynamically adjusted as a function of the induction protocol itself. A change of coupling C makes a model at the same time more expressive and harder to analyze; cf. section 3.3.3.

The cascade model (Fusi et al., 2005) can be related to the model in the present paper by introducing several slow variables $z_1, \ldots, z_n$ with time constants $\tau_1, \ldots, \tau_n$. The coupling from k to $k+1$ is analogous to the coupling of w to z in Eq. 3.2. Even though this extended model and the cascade model share the concept of slower variables, there are some important differences between the two. First, the cascade model (Fusi et al., 2005) is intrinsically stochastic, i.e. the stochasticity due to spiking events is combined with the stochasticity of plasticity itself. Second, the transitions among states in the cascade model are instantaneous (Fusi et al., 2005) . In our framework instead, even though there are discrete *stable* states, the transitions need some time to happen and this is exactly why the frequency of a repetitive stimulus matters in our model.

Similarly to the cascade model (Fusi et al., 2005), the “communicating vessels” model (Benna and Fusi, 2016) relies on multiple hidden variables. However, in contrast to the cascade model (Fusi et al., 2005), the dynamics in the “communicating-vessels” model are determined by continuous variables that obey continuous-time differential equations (Benna and Fusi, 2016). If we truncate the “communication-vessels” model to a single hidden variable, the resulting dynamics fall into our framework, with the simple choice $K_w = K_z = 0$. Extensions to multiple hidden variables with progressively bigger timescales is possible analogously to our discussion above. Indeed experimental results show that the internal bio-chemical mechanisms of the synapse are characterized by different timescales (Reymann and Frey, 2007; Bosch et al., 2014).

Similar to the cascade model, the “state based model" proposed in (Barrett et al., 2009), consist of synapses whose state can shift from e-LTP to l-LTP according to some transition-rates. The model captures two internal mechanisms (tagging and anchor for AMPAR). The probability that a particular synapse is in a specific state is a continuous quantity that depends on the transition probabilities. The main similarity to our model is the presence of a bistable basic synaptic element.

Finally the synaptic plasticity model proposed in (Shouval et al., 2002) proposes a non-linear dynamics for the synaptic weights, similarly to our model. The main goal of (Shouval et al., 2002) is to relate the amount of NMDAR with the Calcium level in the synapse. However, their model is not bistable and no attempt is made to capture the internal synaptic state.

To summarize, our model focuses on a single transition using two variables. If these variables have different intrinsic timescales, the temporal pattern of the stimulation protocol plays a

crucial role. We believe that these insights are applicable beyond our two-variable model in situations where multiple variables covering multiple timescales are pair-wise coupled to each other. This includes well-known consolidation models such as the cascade model (Fusi et al., 2005), the communicating vessels model (Benna and Fusi, 2016) and the write-protected synapse model (Ziegler et al., 2015).

3.6 Supplementary information

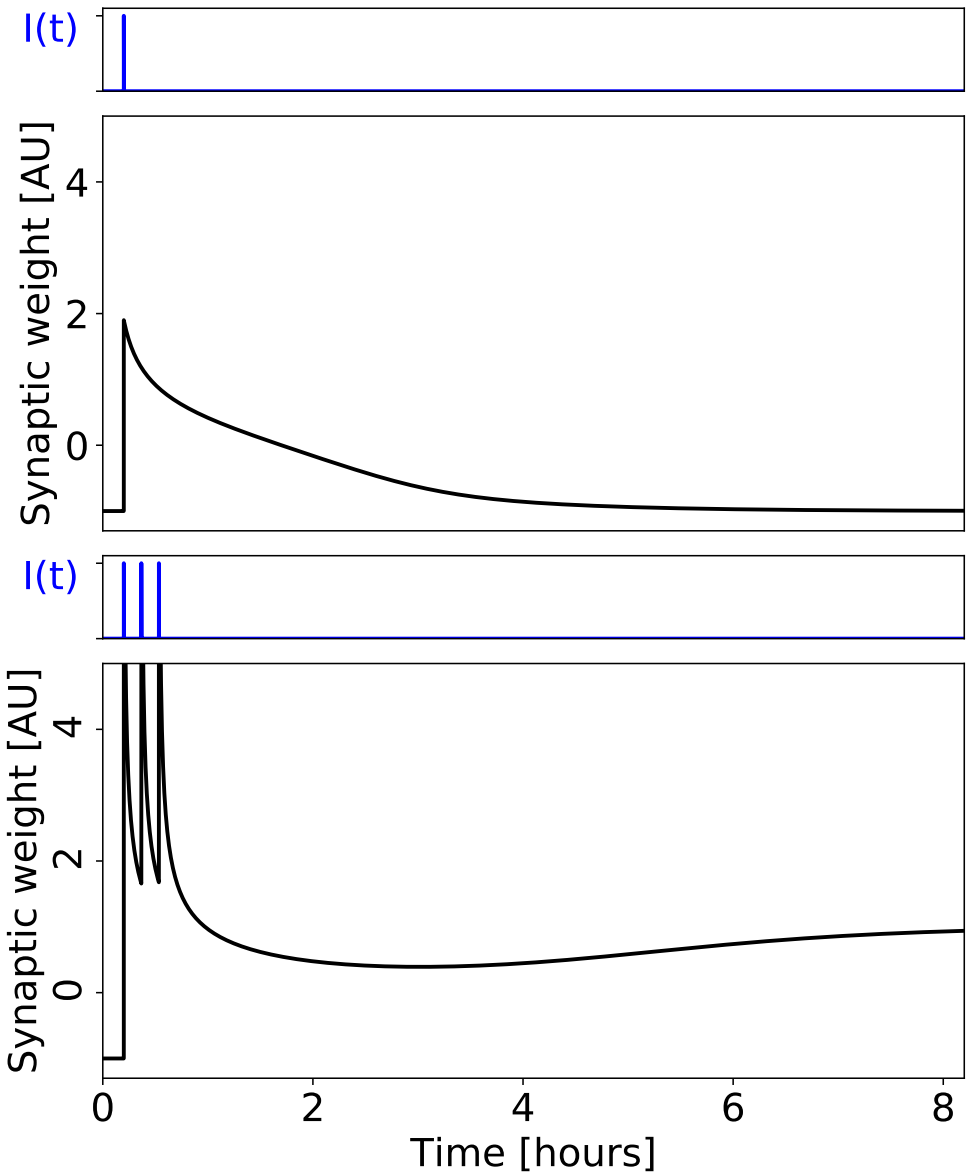


Figure 3.8 – Synaptic weight as a function of time. We try to qualitatively reproduce Fig. 2 of (Frey and Morris, 1997). In blue the trace is of the input current and in black is the response of the synaptic weight which should be proportional to the change in EPSP in the experimental paper. We considered two stimulation protocols. The first one is one single stimulation pulse, or "weak" tetanus , duration 0.2 s. The second in three repeated stimuli, or "strong" tetanus: duration 1 s with 10 min intervals.

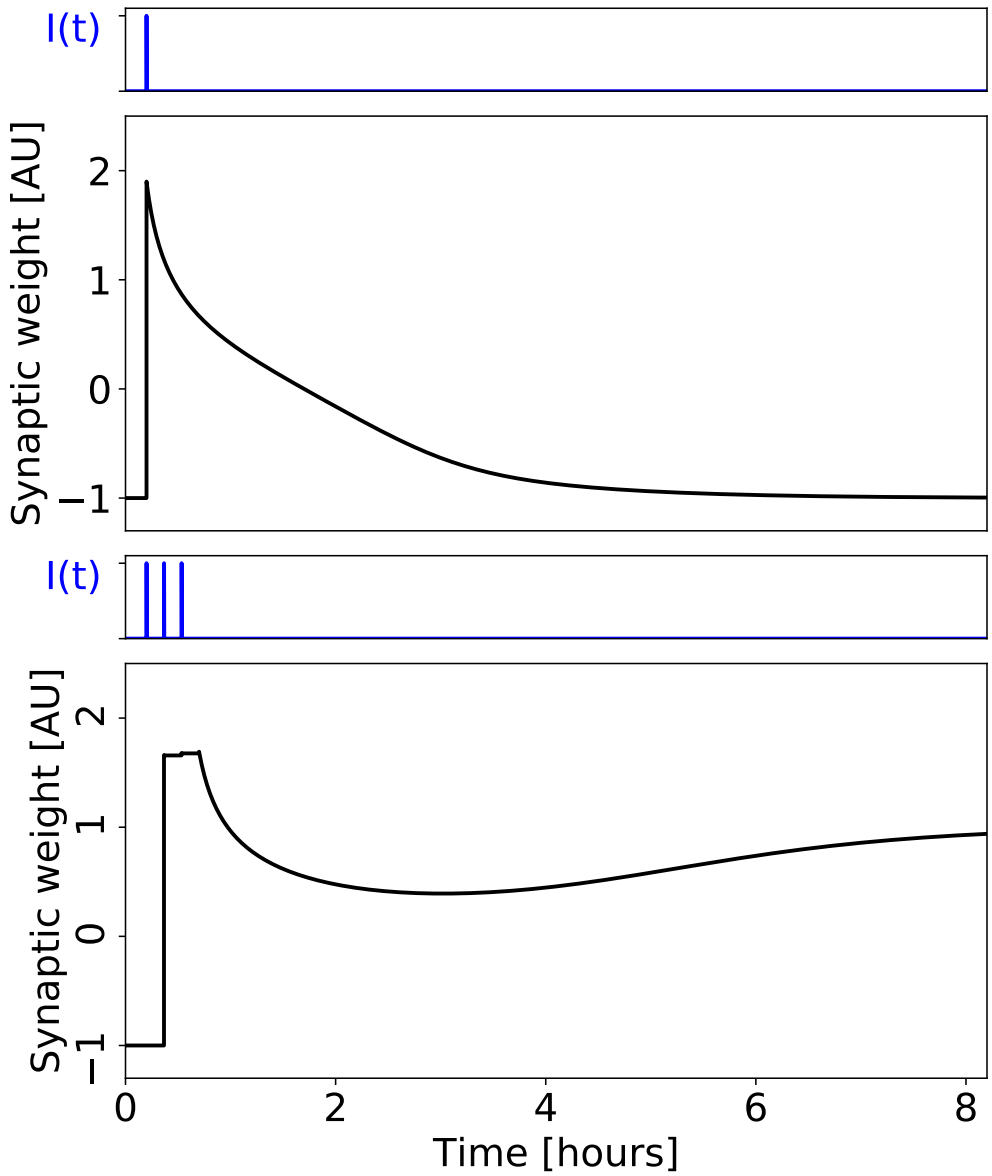


Figure 3.9 – Same as Fig. 3.8, but after applying a low-pass filter on the value of synaptic weights, so to mimic a saturation effect of the synaptic efficacy.

4 Conclusions

4.1 Summary of the main contributions

In the current work, we presented two studies dealing with two fundamental aspects of memory. At the network level, we provide a model that explains how associations between memory engrams can be stored in an auto-associative neural network. Secondly, at the synaptic level, we propose a model of synaptic consolidation, with which we study the consequences of the very essential features of consolidation itself.

4.1.1 Concept association in the human MTL

In Chapter 2 we extend the classical attractor neural networks framework to explain how associations between concepts are stored in the human MTL, a fundamental brain area for episodic memory. Attractor neural networks are the most commonly used tool for modeling associative memory. However, most literature only considers independent memory engrams with relatively high mean-activity, while experimental studies in the MTL (*a*) measured a low mean-activity and (*b*) strongly suggest that neurons shared between engrams encode the association between concepts. This makes the introduction of correlation between low-activity memory patterns a key missing piece of the theory. By including such correlation in a classical low-activity attractor neural network, we can explain how associations might be encoded. We characterize the recall dynamics of low activity memory patterns correlated above chance.

We demonstrated that correlation, even if small, deeply modifies the single pattern retrieval dynamics, and is, therefore, a robust way to encode association. Our results are compatible with the number of shared neurons between memory assemblies experimentally observed, even though we would need a more precise estimation of single-neuron parameters for a better comparison. Moreover, using dynamical mean-field theory, we showed that the free

recall of a chain of concepts is possible in low-activity attractor networks only if patterns are correlated. We quantify the range of correlation that allows the free recall of concepts. In conclusion, our model provides a plausible explanation for how associations are stored in low-activity networks (such as the MTL) and how it influences the retrieval dynamics of memory patterns. The correlation was a key missing piece of the theory of attractor neural network when it is used to model episodic memory in the hippocampus.

Finally, experimental results provide us with two fundamental parameters, which are the sparseness of the representation (or the fraction of active units when the pattern is retrieved) and the fraction of shared units (which is uniquely related to the pair-wise correlation) between two associated patterns. However, correlated memory patterns can be built in several ways while keeping these two parameters fixed. We propose three algorithms to build correlated patterns, which (*i*) keep the measured parameters fixed and (*ii*) are based on different hypotheses. We tested the model prediction with each of these constructions against the experimental dataset from (De Falco et al., 2016) and see which pattern construction (if any) can predict the real distribution of the number of concepts a neuron responds to. We found that the non-hierarchical pattern construction provides a good fit to the data, unlike hierarchical ones which fail in predicting the correct number of highly selective neurons. The results confirm the common opinion that information in the hippocampus is non-hierarchically organised.

4.1.2 Synaptic consolidation

In Chapter 3 we use a simple synaptic model in order to show that the complex response of the synapse to the stimulation protocol can be explained as a consequence of two assumptions about the properties of the synapse. The first is the assumption that synapses are bistable, i.e. they should have at least two stable states, corresponding to depotentiated and potentiated synapses. Secondly, we assume the existence of an internal synaptic mechanism whose dynamics is much slower than the one of the synaptic efficacy, in agreement with experimental evidence that synaptic consolidation is the result of multiple protein mechanisms that are characterized by different timescales. We model the synaptic weights and the internal synaptic mechanism as two dynamical variables that evolve with a fast and slow time scale respectively. The external stimulation input affects directly the synaptic weights and the effect might or might not be transferred to the consolidation variable. We show that it is exactly the time scale separation between the dynamics of the two variables that determines which stimuli lead to synaptic consolidation and which not. Thanks to our analysis we were able to find which combination of stimulation amplitude and frequency are suitable for synaptic consolidation as a function of the time scale separation.

4.2 Possible future developments

In what follows we will propose some possible extensions of the current work in order of scientific relevance.

4.2.1 Synaptic plasticity

Both presented studies leave a few open questions. There is one, in particular, that bridges the gap between the two studies: what is an on-line plasticity rule that allows the creation of correlated memory patterns in an attractor neural network?

While there are plenty of synaptic plasticity models in the computational neuroscience literature, on-line encoding of memory patterns in an attractor neural network remains an open issue. Typically in literature, as well as in the study proposed in Chapter 2, we consider attractor networks in which the memory patterns $\vec{\xi}^{\mu}$ have already been encoded into the connection weights w_{ij} via Hopfield-type connectivity for low activity patterns (Tsodyks and Feigel'man, 1988), that can be regarded as a result of Hebbian learning rule in which patterns are encoded in one shot during the first presentation (Pereira and Brunel, 2018). However, it is important to stress that this interpretation is formally correct only for *independent* patterns.

It is not clear which could be an appropriate Hebbian learning rule to encode correlated patterns in an attractor neural network. In (Zenke et al., 2015b) the authors propose a spiking neuron version of a traditional attractor network. In their model, they show that they can store four correlated stimuli with online synaptic weights updates. The learning rule, which is the result of a fast LTP rule and a slower homeostatic mechanism, is always active, both during the encoding and the retrieval of the memories. The four patterns get correctly encoded, during a stimulation period in which four stimuli are presented to the network one after the other in random order. The stimulation period exceeds in time the time-scale of the homeostatic mechanism and four corresponding attractors (or memory patterns) are created. However, the network fails in storing one more pattern during a second learning session. Instead of creating novel basins of attraction, one of the old ones gets enlarged until it responds to two concepts. In this case, the reconstructed stimulus from the network activity is a combination of the actual pattern and one of the old patterns with which the basin of attraction merged. One possible solution to the issue might be the introduction of a specific homeostatic mechanism to prevent basins of attraction from merging.

On the other hand, in (Mongillo et al., 2003), they propose a plasticity model for learning pairs of associations. The model successfully reproduce the delayed pair-associate tasks, in which two stimuli are presented to a monkey one after the other, with a delay period between the two presentations. The network model is intended to reproduce, not only the behavior, but

also the structure of a "cortical module": while it can be of inspiration, it is not fitted to be directly applied to the human hippocampus.

A hypothetical new Hebbian plasticity rule for learning correlated patterns would depend on how we choose the stimulation to be like. The two main scenarios can be identified as follows:

- the network receives already correlated stimuli and therefore stores them as correlated memory patterns
- the network receives independent stimuli, but sometimes stimuli are presented together. It is the simultaneous stimulation that induces the correlation in the memory representation of the stimuli.

The spiking network from (Zenke et al., 2015b) was failing in learning new concepts in the first scenario. However, in the context of Chapter 2, we are more interested in the second scenario In fact, experiments in which memory associations are induced seem to confirm that the second scenario would be better suited to describe memory engram formation in the human hippocampus (Ison et al., 2015; Rey et al., 2018). In conclusion, a possible extension of the model would be the development of a Hebbian learning rule that is able to learn both a) novel independent stimuli and b) the association between familiar stimuli. In this case, the time correlation between stimuli should be translated into memory pattern correlations. Finally, the fact that the correlation between the memory engrams should not exceed the maximal value $c_{\max}$, should be guaranteed by some homeostatic or feedback mechanism.

4.2.2 Increasing the bio-plausibility of the model

In Chapter 2 we elucidate how recent versions of attractor neural networks are closer to biology compared to the original Hopfield model. First of all we consider low activity, $\{0,1\}$ memory patterns, in contrast with the traditional $\{-1,1\}$ with 50% active neurons. This allows patterns to be interpreted as firing rate units. We have also dropped the assumption that connection weights must be symmetric and we have checked the robustness of the model to a dilution of the weight matrix. However, there is plenty of room for increasing further the bio-plausibility of the proposed model.

First of all, the current rate model violates Dale's rule. The principle is based on the empirical observation that biological neurons fall into two distinct categories: excitatory or inhibitory. Each category can only form a type of synaptic connection with the post-synaptic neuron, independently from the type of the target cell, so that excitatory neurons can only contribute to increase the activity of the post-synaptic neuron and vice-versa, inhibitory neurons can only contribute to decrease the activity of the post-synaptic neuron. In our current implementation

of the synaptic connection, excitatory connections correspond to positive synaptic weights and inhibitory connections correspond to negative synaptic weights. The same neuron might form positive connections with some of the postsynaptic neurons and negative connections with others. While rate based models of attractor neural networks that satisfy Dale's law have been proposed (Burkitt, 1995), this problem is more commonly overcame in spiking implementations.

Secondly, real neurons have thousands of inputs and only one output, on the contrary, to the neural network model implemented in Chapter 2, where each neuron has a similar number of inputs and outputs.

A spiking implementation of the model could be a solution to both problems. There are several ways to implement a spiking version of an attractor neural network. A possible implementation of a spiking version of an attractor memory network is described in (Gerstner et al., 2014). A similar, but more complex implementation is found in (Zenke et al., 2015b). Both of them respect Dale's principle, by introducing separate groups of excitatory and inhibitory neurons, which together reproduce the behavior of the rate model. On the other hand, the main downside of a spiking implementation is that it reacquires a much greater number of parameters compared to the rate model and, as a consequence of this, the need to fine-tune them.

4.2.3 Large scale network simulation with diluted weights

In parallel to the previous improvements to the bio-plausibility proposed in the previous section, it would be interesting to check the behavior of the attractor neural network proposed in Chapter 2 in the case of a very diluted synaptic weight matrix. We claimed that the model is particularly suitable to describe area CA3 of the hippocampus, which is known to be a very recurrent network. In fact, there are about 2.8 millions neurons in this area and each neuron is estimated to receive inputs from about 30'000 other neurons (Andersen et al., 2006). However, it is also estimated that there are about 2.8 million neurons in CA3, which means that a bio-plausible dilution would be of about $d = 0.01$, implying to drop of about 99% of connections starting from a fully connected scenario. Implementing such a diluted network posed a theoretical and a computational issue. First, since the highest we set the dilution, the more we are away from the mean-field regime in which the theory is valid. This is particularly true because the weight dilution adds up to the high sparseness of the memory patterns. Second, in order to reproduce such a regime in full network simulations, we would need to scale up the system and increase the number of neurons in the network. However, in trying to do so, we have encountered a computational problem: the running memory usage of full network simulation scales with the square of the number of neurons. The current version of the code (written in Julialang) uses about 16 GB of RAM for a full network simulation counting

$N = 50'000$ neurons, a million neurons network would require about 16^{20} GB of RAM which is of course not feasible. One option we have tested was to use Julialang's tool for sparse matrix computations. Such a tool decreases dramatically the request of memory, for the simulation of $N = 50'000$ less than 3 GB are needed, but the price to pay is an increase of the computation time, which scales roughly with the square of the number of simulated neurons. Also in this case the computational time needed to simulate a million neurons is prohibited. The optimization of the simulation of a very diluted attractor neural network is beyond the goals of the current work, but it would be an interesting future development.

4.2.4 Several time-scales

In this section, we propose possible developments for the results of Chapter 3. Using an essential model with only two dynamical quantities was useful to clarify the direct consequences of time scale separation between the biochemical processes that contribute to synaptic consolidation. In particular, we find that time scale separation induces a filtering effect on the type of stimuli that can induce the consolidation. Only some combination of stimuli amplitude and repetition frequency are suitable to consolidate the synapse. On the other hand, in the Supplementary Information of Chapter 3, it becomes clear that a two-dimensional model is too simple to reproduce the a bio-plausible evolution of the synaptic weights. An improvement was achieved by introducing by low-pass filtering the weight dynamics, but it is clear that two dynamical equations are not enough to capture the complexity of the biochemical cascade involved in synaptic consolidation. A possible future extension of the model proposed in Chapter 3 is to introduce more dynamical quantities that evolve over a range of different time scales, in order to capture the more realistic evolution of the synaptic weights. Several models have been proposed in the past that try to reproduce the synaptic weight dynamics using several dynamical quantities operating over different time scales. Among the others, we would like to cite (Benna and Fusi, 2016).

Bibliography

Akrami, A., Russo, E., and Treves, A. (2012). Lateral thinking, from the hopfield model to cortical dynamics. *Brain research*, 1434:4–16.

Allen, P. A., Kaut, K. P., and Lord, R. R. (2008). Emotion and episodic memory. *Handbook of behavioral neuroscience*, 18:115–132.

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

Amit, D. J. and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 7(3):237–252.

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

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985b). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1987). Information storage in neural networks with low levels of activity. *Physical Review A*, 35(5):2293.

Andersen, P., Morris, R., Amaral, D., Bliss, T., and O'Keefe, J. (2006). *The hippocampus book*. Oxford university press.

Anderson, J. R. and Bower, G. H. (2014). *Human associative memory*. Psychology press.

Artola, A., Bröcher, S., and Singer, W. (1990). Different voltage dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, 347:69–72.

Barrett, A. B., Billings, G. O., Morris, R. G. M., and van Rossum, M. C. W. (2009). State based model of long-term potentiation and synaptic tagging and capture. *PLoS Comput. Biol.*, 5:e1000259.

Bashir, Z. I. and Collingridge, G. L. (1994). An investigation of depotentiation of long-term potentiation in the CA1 region of the hippocampus. *Exp Brain Res.*, 79(2):437–443.

Benna, M. K. and Fusi, S. (2016). Computational principles of synaptic memory consolidation. *Nat. Neurosci.*, 19.

Bhalla, U. S. and Iyengar., R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48.

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

Bliss, T. V. P. and Lømo, T. (1973). Long-lasting potentation of synaptic transmission in the dendate area of anaesthetized rabbit following stimulation of the perforant path. *J. Physiol.*, 232:351–356.

Boboeva, V., Brasselet, R., and Treves, A. (2018). The capacity for correlated semantic memories in the cortex. *Entropy*, 20(11):824.

Bös, S., Kühn, R., and van Hemmen, J. (1988). Martingale approach to neural networks with hierarchically structured information. *Zeitschrift für Physik B Condensed Matter*, 71(2):261–271.

Bosch, M., Castro, J., Saneyoshi, T., Matsuno, H., Sur, M., and Hayashi, Y. (2014). Structural and molecular remodeling of dendritic spine substructures during long-term potentiation. *Neuron*, 82(2):444–459.

Brader, J. M., Senn, W., and Fusi, S. (2007). Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Comput.*, 19:2881–2912.

Brandon, M. P., Bogaard, A. R., Libby, C. P., Connerney, M. A., Gupta, K., and Hasselmo, M. E. (2011). Reduction of theta rhythm dissociates grid cell spatial periodicity from directional tuning. *Science*, 332:595–599.

Brown, T. H., Ganong, A. H., Kairiss, E. W., Keenan, C. L., and Kelso, S. R. (1989). Long-term potentation in two synaptic systems of the hippocampal brain slice. In *Neural models of plasticity*. Academic Press.

Burkitt, A. N. (1995). Study of analogue neural networks that obey dale's law using mean-field theory. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 5, pages 2583–2586. IEEE.

Cabeza, R., Dolcos, F., Graham, R., and Nyberg, L. (2002). Similarities and differences in the neural correlates of episodic memory retrieval and working memory. *Neuroimage*, 16(2):317–330.

Caroni, P., Donato, F., and Muller, D. (2012). Structural plasticity upon learning: regulation and functions. *Nature Rev. Neurosci.*, 13(7):478–490.

Clopath, C., Busing, L., Vasilaki, E., and Gerstner, W. (2010). Connectivity reflects coding: A model of voltage-based spike-timing-dependent-plasticity with homeostasis. *Nat. Neurosci.*, 13:344–352.

Clopath, C., Ziegler, L., Vasilaki, E., Busing, L., and Gerstner, W. (2008). Tag-trigger-consolidation: A model of early and late long-term-potentiation and depression. *PLoS Comput. Biol.*, 4:e1000248.

Cohen, M. A. and Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Trans. Syst. Man. Cybern. Syst.*, 13:815–823.

De Falco, E., Ison, M. J., Fried, I., and Quiroga, R. Q. (2016). Long-term coding of personal and universal associations underlying the memory web in the human brain. *Nature communications*, 7:13408.

Dudai, Y. and Morris, R. G. M. (2000). To consolidate or not to consolidate: what are the questions. In *Brain, perception, memory. Advances in cognitive sciences*, pages 149–162.

Eichenbaum, H., Otto, T., and Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences*, 17(3):449–472.

Ermentrout, B. (1996). Type i membranes, phase resetting curves, and synchrony. *Neural computation*, 8(5):979–1001.

Ermentrout, B. (2002). *Simulating, analyzing, and animating dynamical systems: a guide to XPPAUT for researchers and students*, volume 14. Siam.

Felix-Ortiz, A. C. and Tye, K. M. (2014). Amygdala inputs to the ventral hippocampus bidirectionally modulate social behavior. *Journal of Neuroscience*, 34(2):586–595.

Frey, U. and Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385:533–536.

Frühholz, S., Trost, W., and Grandjean, D. (2014). The role of the medial temporal limbic system in processing emotions in voice and music. *Progress in neurobiology*, 123:1–17.

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

Gerstner, W., Kempter, R., van Hemmen, J. L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76–78.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics. From single neurons to networks and cognition*. Cambridge Univ. Press.

Gjorgjieva, J., Clopath, C., Audet, J., and Pfister, J. P. (2011). A triplet spike-timing–dependent plasticity model generalizes the Bienenstock–Cooper–Munro rule to higher-order spatiotemporal correlations. *Proc. Natl. Acad. Sci. USA*, 108(48):19383–19388.

Graupner, M. and Brunel, N. (2007). STDP in a bistable synapse model based on CaMKII and associate signaling pathways. *PLoS Comput. Biol.*, 3:e 221.

Harding, A., Halliday, G., and Kril, J. (1998). Variation in hippocampal neuron number with age and brain volume. *Cerebral cortex (New York, NY: 1991)*, 8(8):710–718.

Hasselmo, M. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends Cogn. Sci.*, 3:351–359.

Hayashi-Takagi, A., Yagishita, S., M. Nakamura, F. S., Wu, Y. I., Loshbaugh, A. L., Kuhlman, B., Hahn, K. M., and Kasai, H. (2015). Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, 525:333–338.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.

Hemmer, P. and Steyvers, M. (2009). Integrating episodic and semantic information in memory for natural scenes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Holtmaat, A. and Caroni, P. (2016). Functional and structural underpinnings of neuronal assembly formation in learning. *Nat. Neurosci.*, 19:1553–1562.

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. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092.

Ison, M. J. and Quiroga, R. Q. (2008). Selectivity and invariance for visual object perception. *Front Biosci*, 13:4889–4903.

Ison, M. J., Quiroga, R. Q., and Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. *Neuron*, 87(1):220–230.

Jacobson, L. and Sapolsky, R. (1991). The role of the hippocampus in feedback regulation of the hypothalamic-pituitary-adrenocortical axis. *Endocrine reviews*, 12(2):118–134.

Josselyn, S. A. and Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473).

Kanter, I. and Sompolinsky, H. (1987). Associative recall of memory without errors. *Physical Review A*, 35(1):380.

Kastner, D. B., Schwalger, T., Ziegler, L., and Gerstner, W. (2016). A model of synaptic reconsolidation. *Front. Neurosci.*, 10:206.

Katkov, M., Romani, S., and Tsodyks, M. (2015). Effects of long-term representations on free recall of unrelated words. *Learning & Memory*, 22(2):101–108.

Knierim, J. J. (2015). The hippocampus. *Current Biology*, 25(23):R1116 – R1121.

Koelsch, S., Jacobs, A. M., Menninghaus, W., Liebal, K., Klann-Delius, G., Von Scheve, C., and Gebauer, G. (2015). The quartet theory of human emotions: an integrative and neurofunctional model. *Physics of life reviews*, 13:1–27.

Kumar, A. and Mehta, M. R. (2011). Frequency-dependent changes in nmdar-dependent synaptic plasticity. *Frontiers in computational neuroscience*, 5:38.

Larson, J. and Munkácsy, E. (2015). Theta-burst LTP. *Brain Res.*, 1621:38–50.

Lavenex, P. and Amaral, D. G. (2000). Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus*, 10(4):420–430.

Levy, W. B. and Stewart, D. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in hippocampus. *Neurosci.*, 8:791–797.

Lisman, J. (2017). Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes: long-term potentiation, long-term depression, short-term potentiation and scaling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1715):20160260.

Lisman, J. E. and Zhabotinsky, A. M. (2001). A model of synaptic memory: A CaMKII/PP1 switch that potentiates transmission by organizing an AMPA receptor anchoring assembly. *Neuron*, 31:191–201.

Malenka, R. C. (1991). Postsynaptic factors control the duration of synaptic enhancement in area CA1 of the hippocampus. *Neuron*, 6(1):53–60.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215.

Markram, H., Wu, Y., and Tosdyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl. Acad. Sci. USA*, 95:5323–5328.

Marr, D. and Thach, W. T. (1969). A theory of cerebellar cortex. In *From the Retina to the Neocortex*, pages 11–50. Springer.

Marr, D., Willshaw, D., and McNaughton, B. (1971). Simple memory: a theory for archicortex. In *From the Retina to the Neocortex*, pages 59–128. Springer.

Martin, S. J., Grimwood, P. D., and Morris, R. G. M. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Ann. Rev. Neurosci.*, 23:649–711.

Mendis, L. (2016). *Distribution of Lipids in the Human Brain and their Differential Expression in Alzheimer's Disease: A Matrix-Assisted Laser Desorption/Ionisation-Imaging Mass Spectrometry (MALDI-IMS) Study*. PhD thesis, ResearchSpace@ Auckland.

Mézard, M. (2017). Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117.

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

Miller, E. K., Freedman, D. J., and Wallis, J. D. (2002). The prefrontal cortex: categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):1123–1136.

Mizumori, S. J. (2013). Context prediction analysis and episodic memory. *Frontiers in behavioral neuroscience*, 7:132.

Mongillo, G., Amit, D. J., and Brunel, N. (2003). Retrospective and prospective persistent activity induced by hebbian learning in a recurrent cortical network. *European Journal of Neuroscience*, 18(7):2011–2024.

Murray, B. D., Holland, A. C., and Kensinger, E. A. (2013). Episodic memory and emotion.

Nabavi, S., Fox, R., Proulx, C. D., Lin, J. Y., Tsien, R. Y., and Malinow, R. (2014). Engineering a memory with LTD and LTP. *Nature*, 511.

Nader, K., Schafe, G. E., and LeDoux, J. E. (2000). Reply-reconsolidation: The labile nature of consolidation theory. *Nature Rev. Neurosci.*, 1:216–219.

Naim, M., Katkov, M., Romani, S., and Tsodyks, M. (2020). Fundamental law of memory recall. *Physical Review Letters*, 124(1):018101.

Nicolas, F. and Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural Circuits*, 9:85.

O'Connor, D. H., Wittenberg, G. M., and Wang, S. S. H. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proc. Natl. Acad. Sci. USA*, 102(26):9679–9684.

Pereira, U. and Brunel, N. (2018). Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 99(1):227 – 238.e4.

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

Pfister, J. P. and Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *J. Neurosci.*, 26:9673–9682.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597.

Quiroga, R. Q. (2019a). Neural representations across species. *Science*, 363(6434):1388–1389.

Quiroga, R. Q. (2019b). Plugging in to human memory: advantages, challenges, and insights from human single-neuron recordings. *Cell*, 179(5):1015–1032.

Quiroga, R. Q. (2020). No pattern separation in the human hippocampus. *Trends in Cognitive Sciences*.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102.

Recanatesi, S., Katkov, M., Romani, S., and Tsodyks, M. (2015). Neural network model of memory retrieval. *Frontiers in computational neuroscience*, 9:149.

Recanatesi, S., Katkov, M., and Tsodyks, M. (2017). Memory states and transitions between them in attractor neural networks. *Neural computation*, 29(10):2684–2711.

Reddy, L. and Thorpe, S. J. (2014). Concept cells through associative learning of high-level representations. *Neuron*, 84(2):248–251.

Redondo, R. L. and Morris, R. G. M. (2011). Making memories last: the synaptic tagging and capture hypothesis. *Nature Rev. Neurosci.*, 12:17–30.

Rennó-Costa, C., Lisman, J. E., and Verschure, P. F. (2014). A signature of attractor dynamics in the ca3 region of the hippocampus. *PLoS Comput Biol*, 10(5):e1003641.

Rey, H. G., De Falco, E., Ison, M. J., Valentin, A., Alarcon, G., Selway, R., Richardson, M. P., and Quiroga, R. Q. (2018). Encoding of long-term associations through neural unitization in the human medial temporal lobe. *Nature communications*, 9(1):1–13.

Rey, H. G., Gori, B., Chaure, F. J., Collavini, S., Blenkmann, A. O., Seoane, P., Seoane, E., Kochen, S., and Quian Quiroga, R. (2020a). Single neuron coding of identity in the human hippocampal formation. *Current Biology*, 30(6):1152 – 1159.e3.

Rey, H. G., Gori, B., Chaure, F. J., Collavini, S., Blenkmann, A. O., Seoane, P., Seoane, E., Kochen, S., and Quiroga, R. Q. (2020b). Single neuron coding of identity in the human hippocampal formation. *Current Biology*.

Reymann, K. G. and Frey, J. U. (2007). The late maintenance of hippocampal LTP: requirements, phases,synaptic tagging, late associativity and implications. *Neuropharmacology*, 52:24–40.

Rinzel, J. and Ermentrout, G. B. (1998). Analysis of neural excitability and oscillations. *Methods in neuronal modeling*, 2:251–292.

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.*, 29:203–227.

Rolls, E. T. (2007). An attractor network in the hippocampus: theory and neurophysiology. *Learning & memory*, 14(11):714–731.

Rolls, E. T. (2018). The storage and recall of memories in the hippocampo-cortical system. *Cell and tissue research*, 373(3):577–604.

Romani, S., Pinkoviezky, I., Rubin, A., and Tsodyks, M. (2013). Scaling laws of associative memory retrieval. *Neural computation*, 25(10):2523–2544.

Rubin, J. E., Gerkin, R. C., Bi, G. Q., and Chow, C. C. (2005). Calcium time course as a signal for spike-timing-dependent plasticity. *J. Neurophysiol.*, 93:2600–2613.

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

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

Sajikumar, S. and Frey, J. U. (2004a). Late-associativity, synaptic tagging, and the role of dopamine during LTP and LTD. *Neurobiol. Learn. Mem.*, 82:12–25.

Sajikumar, S. and Frey, J. U. (2004b). Resetting of synaptic tags is time- and activity dependent in rat hippocampal CA1 in vitro. *Neuroscience*, 129:503–507.

Sajikumar, S., Navakkode, S., Sacktor, T. C., and Frey, J. U. (2005). Synaptic tagging and cross-tagging: the role of protein kinase Mζ in maintaining long-term potentiation but not long-term depression. *J. Neurosci.*, 25:5750–5756.

Senn, W., Tsodyks, M., and Markram, H. (2001). An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing. *Neural Comput.*, 13:35–67.

Shallice, T., Fletcher, P., Frith, C. D., Grasby, P., Frackowiak, R. S., and Dolan, R. J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature*, 368(6472):633–635.

Shamir, M. and Sompolinsky, H. (2000). Thouless-anderson-palmer equations for neural networks. *Physical Review E*, 61(2):1839.

Shiino, M. and Fukai, T. (1992). Self-consistent signal-to-noise analysis and its application to analogue neural networks with asymmetric connections. *Journal of Physics A: Mathematical and General*, 25(7):L375.

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S.-C., and Lindenberger, U. (2010). Episodic memory across the lifespan: The contributions of associative and strategic components. *Neuroscience & Biobehavioral Reviews*, 34(7):1080–1091.

Shouval, H. Z., Bear, M. F., and Cooper, L. N. (2002). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proc. Natl. Acad. Sci. USA*, 99(16):10831.

Sjöström, P. J., Turrigiano, G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32:1149–1164.

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive Hebbian learning through spike-time-dependent synaptic plasticity. *Nat. Neurosci.*, 3:919–926.

Squire, L. R., Genzel, L., Wixted, J. T., and Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766.

Strogatz, S. H. (2014). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Hachette UK.

Tonegawa, S., Pignatelli, M., Roy, D. S., and Ryan, T. J. (2015). Memory engram storage and retrieval. *Current opinion in neurobiology*, 35:101–109.

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

Tulving, E. (1993). What is episodic memory? *Current directions in psychological science*, 2(3):67–70.

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

Turrigiano, G. G., Marder, E., and Abbott, L. F. (1996). Cellular short-term memory from a slow potassium conductance. *J. Neurophysiol.*, 75(2):963–966.

Van Rossum, M. C. W., Bi, G. Q., and Turrigiano, G. G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *J. Neurosci.*, 20(23):8812–8821.

Waydo, S., Kraskov, A., Quiroga, R. Q., Fried, I., and Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, 26(40):10232–10234.

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

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24.

Zeithamova, D. and Preston, A. R. (2010). Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience*, 30(44):14676–14684.

Zenke, F., Agnes, E. J., and Gerstner, W. (2015a). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nat. Commun.*, 6:6922.

Zenke, F., Agnes, E. J., and Gerstner, W. (2015b). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature communications*, 6(1):1–13.

Ziegler, L., Zenke, F., Kastner, D. B., and Gerstner, W. (2015). Synaptic consolidation: from synapses to behavioral modeling. *J. Neurosci.*, 35:1319–1334.

Personal information

Surname(s) / First name(s)	**Gastaldi Chiara**
Address(es)	Batiment AI, EPFL, 1015 Lausanne, Switzerland
Telephone(s)	+41 21 69 35280
Email(s)	chiara.gastaldi@epfl.ch
Nationality(-ies)	Italian
Date of birth	10.04.1991

Education

1.03.2016 until 15.03.2021	**École Polytechnique Fédérale de Lausanne**, Switzerland, **Ph.D. candidate** LCN: Laboratory of computational neuroscience Thesis title: "A dynamical systems approach to synaptic consolidation and associations of concepts in Hippocampus"
09.2013 until 06.2015	**École Polytechnique Fédérale de Lausanne**, Switzerland , Master of Physics and Minor in Computational Science and Engineering
10.2010 until 09.2013	**Universitá degli Studi di Pavia**, Italy, Bachelor of Physics
10.2010 until 09.2013	**Collegio Ghislieri**, Pavia, Italy
10.2010 until 09.2013	**IUSS: Institute for Advanced Study** Pavia, Italy

Publications

In preparation	Chiara Gastaldi, Tilo Schwalger, Emanuela De Falco, Rodrigo Q. Quiroga and Wulfram Gerstner, *When shared concept cells support associations: theory of overlapping memory engrams*
November 2019	Chiara Gastaldi, Samuel P. Muscinelli and Wulfram Gerstner, *Optimal stimulation protocol in a bistable synaptic consolidation model*, Front. Comput. Neurosci. 13:78.
March 2016	Yannis Burnier and Chiara Gastaldi, *Contribution of next-to-leading order and Landau-Pomeranchuk-Migdal corrections to thermal dilepton emission in heavy-ion collisions*, Physical Review C 93.4 (2016): 044902.

Conferences and Seminars

09.03.2021	**Invited speaker**: seminar presentation at Exter University on-line
23.02.2021 until 26.02,2021	**Cosyne**: poster presentation, on-line
06.07.2020 until 07.07. 2020	**ICMNS**: poster presentation, on-line
17.09.2019 until 20.10.2019	**Bernstein Conference**: poster presentation, Berlin, Germany
28.02.2019 until 05.02.2019	**Cosyne**: poster presentation, Lisbon, Portugal
13.07.2018 until 18.07. 2018	**CNS annual meeting**: poster presentation, Seattle, USA
30.05.2017 until 2.06. 2017	**ICMNS**: International Conference on Mathematical Neuroscience, Boulder, USA I gave a talk on "Systematic study of synaptic consolidation models using phase-plane analysis" .
23.05.2017 until 30.05. 2017	**Visiting S. Fusi's lab at Columbia University**, New York, USA
07.2016	**CCN Summer School**: Computational and Cognitive Neuroscience Organized by NYU Shanghai, this summer school combines theoretical lectures with coding sections. Final project: "A memory model for free recall".

Teaching Experience

02.2019 until 07.2019	**Teaching assistant**: "Artificial neural networks" by prof. Gerstner

02.2018 until 07. 2018	**edX**: Computational Neuroscience: Neuronal Dynamics of Cognition I contributed to the course by setting up the web page, writing the on-line exercises and answering to students' questions in the forum. (Link to the course)
02.2017 until 07.2019	**Teaching assistant**: Biological modeling of neural networks During three Spring semesters, I have been a teaching assistant for the course "Biological modeling of neural networks" hold by prof. Gerstner (Link to the course).
09.2017 until 03.2018	**Master project supervision**: Neural implementation of learning by surprise Supervision of the a master student during the development of his final master project and during the following internship.

Other projects

11.2015 until 29.02.2016	**Internship in EPFL**: LCN: Laboratory of computational neuroscience Integrating and optimizing different synaptic plasticity models to form and retrieve memories in neural networks
08.2015 until 10.2015	**Internship in EPFL**: LPPC: Laboratory of particle physics and cosmology Studying the electric conductivity of the quark-gluon plasma.
02.2015 until 07.2015	**Master Project**: Hydrodynamics for the quark gluon plasma and di-lepton emission.
03.2013 until 07.2013	**Bachelor Project**: Spontaneous collapse models for the wave function: theoretical prediction and experimental tests.

Honours and awards

2010 until 2013	Annual fellowship from **IUSS** (Istituto Universitario di Studi Superiori)
2010 to 2011	Fellowship from Regione Lombardia, for scientific subjects

Personal skills and competences

Languages	English, level: Fluent French, level: B1-B2, Intermediate German, level: A1, Elementary
Mother tongue	Italian
Computer skills	**Operative systems:** Linux systems, Windows, MacOS. **Programming languages:** C++, C, Python, Julia, Wolfram Mathematica.
Other skills and personal interests	Mountain lover, basic piano player, passion for travelling.