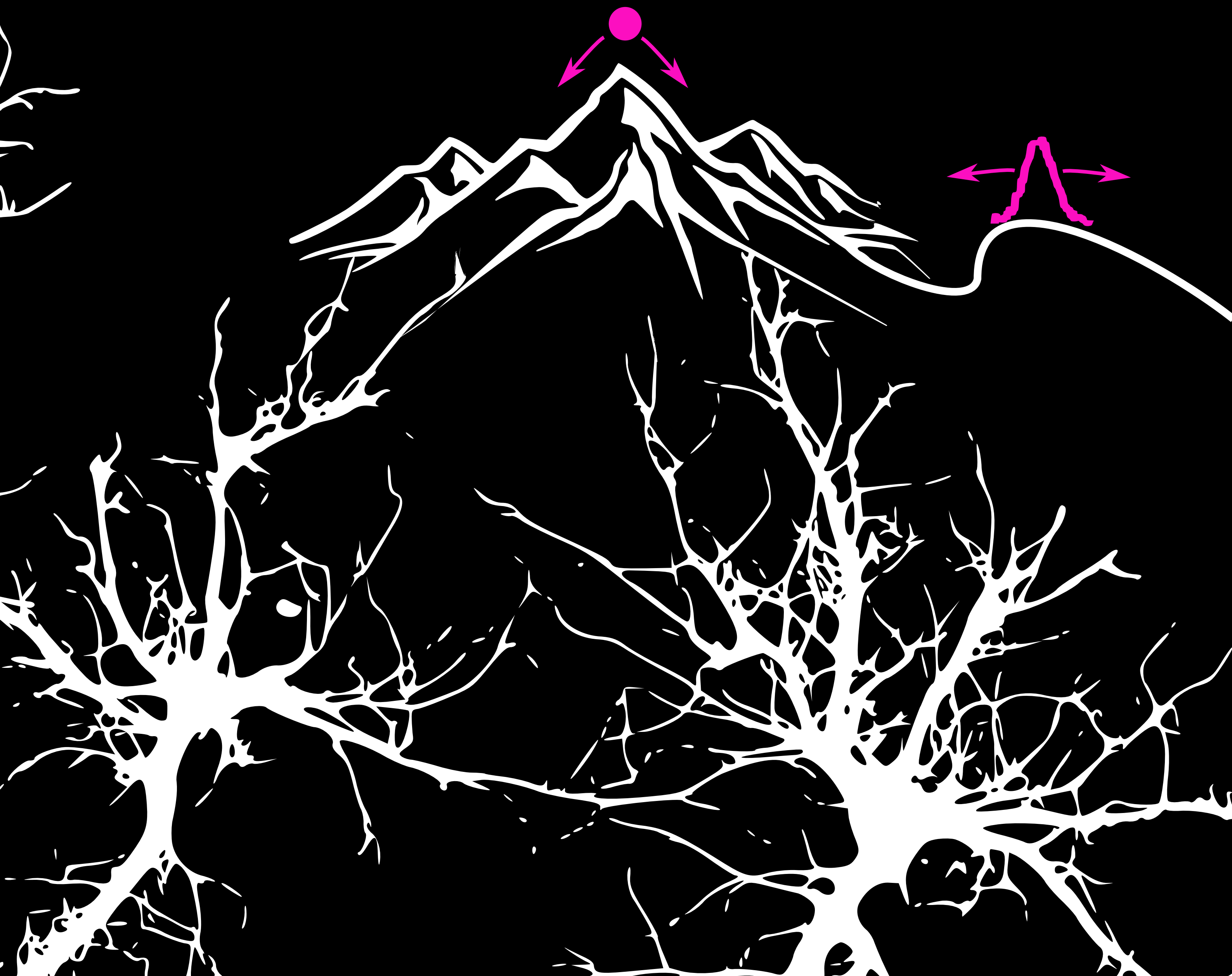
MEMORIES, ATTRACTORS, SPACE AND VOWELS

Francesca Schönsberg - PhD Thesis



MEMORIES, ATTRACTORS, SPACE AND VOWELS

FRANCESCA SCHÖNSBERG

SUPERVISOR: ALESSANDRO TREVES

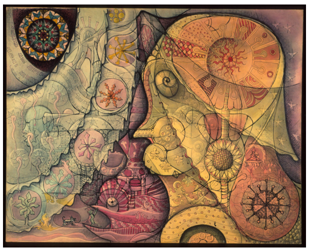
A thesis submitted for the degree of
Doctor of Philosophy

SISSA – SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

TRIESTE, JULY 2021

ma per seguir virtute e canoscenza

SISSA



GATE

MARIE CURIE ACTIONS

Francesca Schönsberg: *Memories, attractors, space and vowels*.
PhD Thesis, July 2021

A Margherita e Alda

ACKNOWLEDGEMENTS

I sincerely thank Alessandro, who has been mentoring me at every step of this thesis with kindness, encouraging me without imposing topics, control or pressure, but being always available to discuss, at any time.

I thank Yasser Roudi, for co-supervising me with enthusiasm in the second part of this thesis, for hosting me in Trondheim and for sharing his notes, from 2007, with the first attempt of applying the Gardner approach to TL units.

I thank Rémi Monasson for hosting me in Paris, for giving perceptive advises on the Gardner calculation and for the availability in the recent remote discussions to make analytical sense of the third part of this thesis.

I thank Beessan Warasna, Ayman Salman and Aya Mousa for hosting me at the Palestinian Neuroscience Initiative, in Al-Quds University, and showing me the joy with which one can do research under occupation, despite the little availability of tools, making it also as an act of resistance.

I thank my Phd office mate Kwang Il Ryom for often helping me with brilliant clarity in reminding the basics of statistical physics and for being available for entertaining discussions.

I thank my Phd colleague Oleksandra Soldatkina, with whom we shared the M-GATE steps since the beginning and who has always been there with glee for any help.

I thank my parents and my sister for the unconditional support, as well as my whole friends and beloved.

CONTENTS

ABSTRACT

Higher cognitive capacities, such as navigating complex environments or learning new languages, rely on the possibility to memorize, in the brain, continuous noisy variables. Memories are generally understood to be realized, e.g. in the cortex and in the hippocampus, as configurations of activity towards which specific populations of neurons are "attracted", i.e towards which they dynamically converge, if properly cued. Distinct memories are thus considered as separate attractors of the dynamics, embedded within the same neuronal connectivity structure. But what if the underlying variables are continuous, such as a position in space or the resonant frequency of a phoneme? If such variables are continuous and the experience to be retained in memory has even a minimal temporal duration, highly correlated, yet imprecisely determined values of those variables will occur at successive time instants. And if memories are idealized as point-like in time, still distinct memories will be highly correlated. How does the brain self-organize to deal with noisy correlated memories? In this thesis, we try to approach the question along three interconnected itineraries.

In Part ii we first ask the opposite: we derive how many uncorrelated memories a network of neurons would be able to precisely store, as discrete attractors, if the neurons were optimally connected. Then, we compare the results with those obtained when memories are allowed to be retrieved imprecisely and connections are based on self-organization. We find that a simple strategy is available in the brain to facilitate the storage of memories: it amounts to making them more sparse, i.e. to silencing those neurons which are not very active in the configuration of activity to be memorized. We observe that the more the distribution of activity in the memory is complex, the more this strategy leads to store a higher number of memories, as compared with the maximal load in networks endowed with the theoretically optimal connection weights.

In part iii we ask, starting from experimental observations of spatially selective cells in quasi-realistic environments, how can the brain store, as a continuous attractor, complex and irregular spatial information. We find indications that while continuous attractors, per se, are too brittle to deal with irregularities, there seem to be other mathematical objects, which we refer to as quasi-attractive continuous manifolds, which may have this function. Such objects, which emerge as soon as a tiny amount of quenched irregularity is introduced in

would-be continuous attractors, seem to persist over a wide range of noise levels and then break up, in a phase transition, when the variability reaches a critical threshold, lying just above that seen in the experimental measurements. Moreover, we find that the operational range is squeezed from behind, as it were, by a third phase, in which the spatially selective units cannot dynamically converge towards a localized state.

Part iv, which is more exploratory, is motivated by the frequency characteristics of vowels. We hypothesize that also phonemes of different languages could be stored as separate fixed points in the brain through a sort of two-dimensional cognitive map. In our preliminary results, we show that a continuous quasi-attractor model, trained with noisy recorded vowels, can effectively learn them through a self-organized procedure and retrieve them separately, as fixed points on a quasi-attractive manifold.

Overall, this thesis attempts to contribute to the search for general principles underlying memory, intended as an emergent collective property of networks in the brain, based on self-organization, imperfections and irregularities.

Some ideas and figures have appeared previously in the following publications and oral presentations:

ARTICLES

- F. Schönsberg, Y. Roudi* and A. Treves*, "Efficiency of local learning rules in threshold-linear associative networks.", Physical Review Letters, 126(1), 018301

BOOK CHAPTERS

- O. Soldatkina*, F. Schönsberg* and A. Treves, "Challenges for place and grid cell models", in "Computational Neuroscience Approaches to Cells and Circuits", M. Giugliano et al, eds. Springer, 2021 (in press).

ORAL CONTRIBUTIONS AND POSTERS

- *Efficiency of local learning rules in threshold-linear associative networks*. Poster. FENS Conference, July 2020, (Glasgow) Online.
- *Gardner storage capacity and the actual cost of Hebbian learning in threshold linear associative networks*. Talk. Theory meetings in Yasser Roudi's group. May 2020, (Trondheim) Online.
- *Memory and attractor dynamics.*. Talk. Palestinian Neuroscience Initiative (PNI), Al Quds University, February 2019, Al Quds, Palestine.
- *Maximal storage capacity and the actual cost of one-shot learning in cortical associative networks*. Talk. MGATE Winter School, Weizmann Institute.
- *Gardner approach for threshold linear units to understand memory in the brain.* Talk. Tinkos Conference, October 2019, Belgrade, Serbia.
- *Progresses in exploring if grid maps with different peak heights can be attractive states*. Talk. Theory meetings in Yasser Roudi's group. March 2019, NTNU, Trondheim, Norway.
- *Grid and Place cells are part of the same network. How?*. Talk. MGATE Winter school "Principles of neural computations", February 2018, Nijmegen, Netherlands.

ARTICLES IN PREPARATION

- F. Schönsberg, Y. Roudi and A. Treves, "The nature of retrieval to no-retrieval transition in threshold-linear networks with Hebbian learning"
- F. Schönsberg, R. Monasson and A. Treves, "Continuous quasi-attractor manifolds for irregular place maps"
- E. Cortesi, F. Schönsberg, I. Lona, A. Treves, "Attractor model for vowels cognitive maps"

Part I

GENERAL INTRODUCTION

1 "PARTICULAR TREASURES OF THE PAST"

To be rooted is perhaps the most important and least recognized need of the human soul. [..] A human being has roots by virtue of his real, active and natural participation in the life of a community which preserves in living shape certain particular treasures of the past and certain particular expectations for the future.
— Simone Weil, *L'Enracinement (1949)*

Before they seize power and establish a world according to their doctrines, totalitarian movements conjure up a lying world of consistency which is more adequate to the needs of the human mind than reality itself in which, through sheer imagination, uprooted masses can feel at home.
— Hannah Arendt, *The Origins of Totalitarianism (1951)*

1.1 MEMORY AND MARY CALKINS

"What one of the numberless images which might [..] follow upon the present percept or image will actually be associated with it?" So Mary Whiton Calkins[1] was introducing her Phd Dissertation on Memory in 1895 [1].

[1] (Hartford 1863, Newton 1930)

She was a pioneer in the study of memory in those days of excitement in the newborn psychology community. In her Thesis, resulting from the experimental studies she carried out from 1892 to 1894, she devoted an initial special emphasis to the differentiation between objects of consciousness and contents of consciousness. As reported in Ref. [2], these terms closely correspond to what is *cue* and *to-be-remembered* item in current memory research terminology. She designed experiments aimed at providing a deeper understanding of *"the nature of associations"* in our brain, testing what we would call today short term memory. In these, among other results, she discovered the method of paired associates: she saw that remembering a number which is presented always with a color is easier than remembering numbers presented each time with different colors, however extravagant they would be. She did not use the term *paired associates* in the original work [2] (she will do it later, in her autobiography) and the method was named in 1908 by Edward Thorndike, who did not cite her.

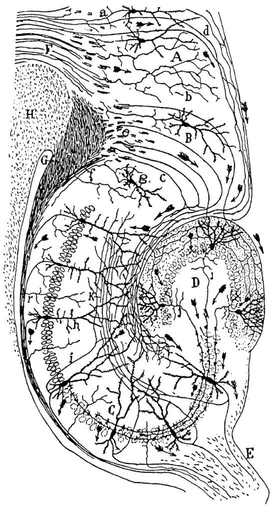
Despite the note sent in 1894 to the President of Harvard College by the professor with whom she was working informally, asking to enroll her in the PhD as she was without *"any doubt [..] superior also to all candidates of the philosophical Ph.D. during the [previous] years"* Mary Whiton Calkins was not accepted because she was a woman. Nor she was granted a PhD in 1895, when an unauthorized committee of Harvard professors, after evaluating her Thesis sent the positive response to the President and Fellows of Harvard College, nor in 1927, when 13 other professors, including Thorndike, sent a petition to Harvard to give her the title, nor nowadays as a *post-mortem* award [1].

She would eventually publish her thesis in Psychological Review [3, 4] and after not being awarded a PhD her last paper on memory would be in 1898 [2], shifting then her focus to the psychology of introspection and to the concept of self, among other interests, pursued with her own research group. Still today, despite the recognition which her ideas and results are gradually receiving, the study of memory is predominantly traced back solely to her contemporary Hermann Ebbinghaus. She died of cancer, leaving to her posterity an autobiography, and to Eleanor Gamble, one of the few other women present in science at that time, the direction of her Laboratory [4]. She is known as the first female president of the American Psychological Association.[2]

Mary W. Calckins was also an active and outspoken femminist, pacifist and socialist, supporting cases such as that of the italian anarchists Sacco and Vanzetti [6].

[2] Changing president every year since 1892 APA has had in 128 president elections a total of 19 women presidents of whom 8 in the past 11 years [5].

1.1.1 *Memory in this Thesis*



About 20 years after Mary Calkins' dissertation, the neuroanatomist Ramon y Cajal sketched in Madrid his illuminating drawing[3] of the hippocampus [7]. 33 years after that, Dr. Scoville, a neurosurgeon in the hospital of Hartford, the city where Mary Calkins was born, removed the hippocampus from his patient Henry Molaison in a tragic attempt to treat his epilepsy. Henry Molaison[4], as a consequence, completely lost his capability to form new episodic memories [8], as understood by the studies of Brenda Milner, who followed him afterwards. From then onward it became clear that somewhere in the hippocampus is embedded our capability to form *short term memories*. Four years before that surgery, in 1949, Edward Tolman,[5] born in the same city where Mary Calkins died, formulated the psychological hypothesis that animals are able to create "cognitive maps" [10]. The connection between cognitive maps, hippocampus and memory will be clarified in the following decades, fostered by the discovery of spatially selective cells.

In Chapter 2 I will briefly introduce part of these studies, which form the general phenomenological context of Part iii of the present thesis.

[3] reproduced here on the left

[4] known as the "H.M. patient"

[5] Edward Tolman (1886 – 1959) was a prominent learning theorist in the 30s and an active opponent of the loyalty oaths, imposed in Berkeley in 1949 as an attempt to take out communists and other disloyal people from academia [9].

1.2 DANIEL AMIT AND THE NON-NEUTRALITY OF SCIENCE

Daniel J. Amit, born in Poland in 1938, was an eminent physicist and one of the pioneers in the field of neural networks and computational neuroscience [11, 12]. In the 80′, attracted to neuroscience as some other physicists at that time, he realized that the tools of statistical mechanics could be applied to the study of memory. This led to his famous calculation published with Gutfreund and Sompolinsky in 1985 [13] and to the ones of Elizabeth Gardner a few years later [14]. From there on, attractor neural networks and, later, continuous attractor neural networks, became fundamental mathematical objects to understand memory, objects which will be introduced in Sect. 3 and Sect. 10.

The rest of this chapter, instead, will not be directly related to the present thesis and is devoted, for the delight of keeping memory alive, to introduce his scientific political involvement .

Daniel J. Amit considered that *"one of the main intellectual duties of a scientist is to apply the rigorous standard of his profession also to his own (research) activity intended as social activity"*[6]. In 2003, as a consequence of the US invasion of Iraq he refused to collaborate with the American Physical Society[7], which he motivated to ArabNews saying:

"There can be absolutely no doubt that science has been directly linked to weapons of mass destruction for at least two centuries. [..] Today this problem is particularly grave because every aspect of existence has become a weapon, not only physics."[8]

[6] Excerpts translated from the essay Daniel Amit published in the magazine Prometeo in 2005 [15] and preliminary presented in a conference in 2003

[7] The letter exchange made public by Daniel Amit between him and the American Physical Society can be read here [16]

[8] From an interview by ArabNews to Daniel Amit on 15.03.2003. The interview has been reproduced with permission on the webpage of Luis A. Florit at IMPA (Instituto de Matemática Pura e Aplicada) Rio de Janeiro and is available here [17]

Daniel Amit was actively involved in those debates on the *non-neutrality* of science developed since the mid twentieth century [18] within and outside academia. His concerns were not only focused on the relation between scientific discoveries and war, but also, more in general, on the enslavement of research to power and technology and on the dangers underlying the idealization of science. In an oral contribution to a conference, in 2003, he was arguing:

"It would be enough to look at what kind of projects are promoted by national and international agencies, to realize that most of the funding is assigned to projects that benefit a wrong or questionable idea of economic development, which promotes the virtual, the superfluous, the military, at the expense of the social and the ecological conservation. [..] The world of research collaborates, roughly, in all this, driven by the ease of obtaining funding, and by the media exposure coupled with today's production-marketing process. Overproduction in fields such as communication, information technology, [..], is identified with science itself, and as such, it is defended in its most prestigious journals. [..] Also contributes to this a perverse relationship with the media, which is selling science as a cover element of the prevailing socio-economic project, offering scientists the temptation of public exposure.[..] The confusion [..] between technology and science, and the authoritarian defense of an ideal concept of science (still a 'miracle') ad exclusion of other ways of knowing, serves as a perfect cover for a system increasingly in crisis, more and more violent."[6]

From then on, a large amount of studies in the field of neural networks started to be focused on machine learning and artificial intelligence. If part of the scientific community finds it extremely exciting, voices from other fields are warning over the social impacts of applying artificial intelligence to people's data[9], while several other applications can be found in weapon technologies [20].

One could wonder what would Daniel Amit think about the current development of the field, and what would be if those debates were again to take place within the scientific community.

Daniel Amit, who was also an active pacifist opposed to the occupation of Palestine [21, 22], took his own life in Jerusalem in 2007.

[9] See for example the meticulous researches of the Harvard professor Shoshana Zuboff summarized in her book "*The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*" [19]

2 PHENOMENOLOGICAL CONTEXT

By 1971, at the time the first place cell was discovered in the rat hippocampus [23] it was already known that the hippocampal formation was crucial for episodic memory (see Sect. 1.1.1). It had also already been hypothesized that animals may create "cognitive maps" [10] to comprehensively represent an environment, and the discovery of spatially selective cells provided a putative neurophysiological expression of both, memory and spatial cognitive maps. In this chapter I will generally introduce spatially selective cells, their features and some of the challenges they pose. In Chapter 9 I will specifically introduce recent experimental results regarding the irregularity of place cells and in Appendix A I give a short anatomical overview over the organization of the hippocampal system.

2.1 SPATIALLY SELECTIVE CELLS

When considering spatial cognition, cells in the hippocampal division have been first characterized by looking at the *firing rate map* of each cell. In such a map, the spike events are plotted in a drawing representing the environment in which the animal is moving, at the position of the head of the animal when each spike occurred. Spikes clustered in a specific region form a *field*. The number of spikes occurring in each spatial bin is typically divided by the time spent in that bin, and the map is then regularized to look smoother. A common trait to the various types of spatially selective cells is that the localization of their fields appears unrelated to the position of each cell in the tissue, and neighboring cells do not necessarily show overlapping firing fields in the environment. Here I will provide a short recap of the main spatially selective cell types as they are usually categorized:

- **Place cells**: Originally discovered half a century ago [23], their activity is peaked at one or a few positions in space in the typical environments in which rodents are made to run in the laboratory, as in Fig. 1a. In one-dimensional environments, such as circular paths, n-arm mazes or linear tracks, place fields are typically directional, i.e. they occur only when the animal is running in one direction, whereas in two-dimensional environments they tend to be, or to become, non-directional. Place cells have been most extensively described in CA3 and in CA1 (see Appendix A), where it is estimated that between a quarter and a half of all pyramidal cells show at least one place field in a typi-

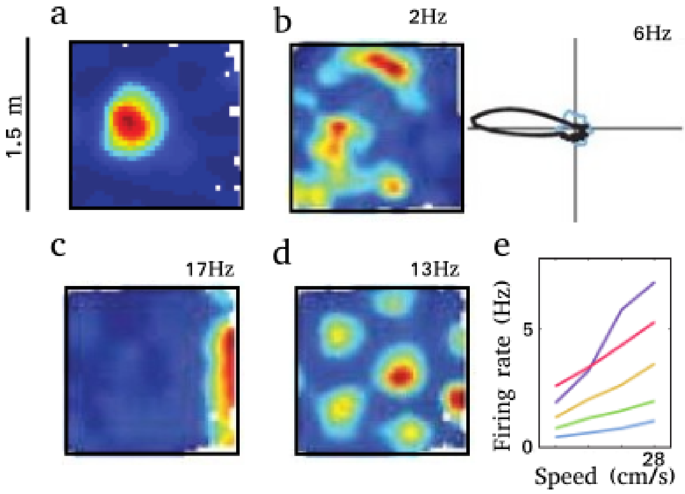


Figure 1: Firing behavior of spatially selective cells. a) Firing rate map of a place cell; b) Firing rate map (left) and head direction map (right) of an head direction cell; c) Firing rate map of a border cell; d) Firing rate map of a grid cell; e) Schematic of characteristic tuning curves of five speed cells. Plot a) adapted from [24], b)c)d) from [25], e) from [26].

cal 1 m^2 box. Place activity, like other types of selective spiking, is typically modulated by the speed of the animal.

- **Head Direction cells**: First reported in 1984 [27], HD cell activity depends on the direction of the head of the animal, which on average tends to coincide with, but is quite distinct from, its direction of motion, as in Fig. 1b. They are found in a variety of areas, especially in the parasubiculum and in the EC (see Appendix A).

- **Grid cells**: A startling discovery [28], their activity is peaked, ideally, at the vertices of a hexagonal lattice, spacing from a few tens of centimiters upwards, giving rise, in a typical two-dimensional box, to several grid fields per cell, see Fig. 1d. The spacing and orientation of the lattice appear to be shared by neighboring cells, but not the position of their fields. Whereas in EC layer II grid activity is characterized as a-directional (but see [29]), in deeper layers of EC the activity of most grid cells is modulated by head direction, and they are called conjunctive (grid) cells [30]. The spacing of the grid lattice increases towards the ventral portion of mEC [31] in what appear to be discrete steps, or *modules*. Grid cells are predominantly found in mEC but are also present in the pre/para-subiculum.

- **Border cells**: Described by [25], the activity of border cells is intense at one or several borders of the environment the animal is exploring, as in Fig. 1c. They are found in the EC and pre/para-subiculum.

- **Speed cells**: Originally found in [26], their firing rates linearly depend on the velocity at which the animal is navigating, as in Fig. 1d. They have been found in the EC, but variants sensitive to angular velocity have been recently reported also in the pre/para-subiculum.
- **Object, object-trace, object-vector, social cells**: A still burgeoning variety of selectivity types is observed when objects (or other animals[32]) are introduced in the same environment, starting with those observed by [33], which fire selectively at positions related to an object and which were found in the lateral enthorinal cortex.

2.1.1 *Stability*

One should note that the selectivity of cells in the hippocampal region tends to be stable: when an animal is exposed a second time to a familiar environment, the activity of each neuron reproduces on average the same map as the previous time, as if the environment is somehow memorized in that maps. The behavior of spatially selective cells across different environments instead generally differs depending on the type of response. Grid cells *align*: when moving a rodent across environments, grid cells within the same module appear to change grid orientation, but coherently, while field size, spacing and relative distance between fields are maintained constant [34]. Place cells, instead, *remap*: they change their firing patterns, in a manner that appears totally unpredictable from knowledge of its place field(s) in the original environment, or from the changes expressed by nearby cells [35, 36]. The same cell may show two place fields in one box, none in another, and be selective for an odor in an olfactory discrimination task [37].

2.1.2 *Irregularities*

The difference between grid-*alignment* and place-*remapping* leads to a rather general hypothesis regarding the putative hippocampus-mEC memory-for-navigation system: grid cells would provide a detailed metric, a universal chart, while place cells would represent the landmarks and additional information to give a more complete notion of a territory.

This hypothesis, however, started to be challenged with those results highlighting place and grid inhomogenities. Several experimental studies, have shown for example that grid regularity can be distorted as soon as the environment becomes more complex, as in the *hairpin* maze [38], in the presence of goals [39] or with non-standard shapes of the walls [40]. On the other hand, the variability in the peak rates of the grid fields of the same cell, not just in their position, has been shown to be reliable, hence possibly carrying some information [41, 42]. Also place fields recorded in large environments turned to be usually multiple and highly irregular (see Chapter 9). All such effects, which are expected to be huge in the natural environments in which the grid and place cell system has presumably evolved, challenge the possibility of reducing spatially selective cells to idealized models. Part iii of this thesis originates from this debate.

3 THEORETICAL FRAMEWORK

How can the hippocampal network store and retrieve memories?

3.1 FROM DAVID MARR TO ATTRACTOR NEURAL NETWORKS

The initial attempts to approach a quantitative answer to this question probably date back to 1971, when the student David Marr published his *theory of archicortex* (i.e., of the hippocampus) [43]. Marr's vision was of a *simple* memory system where representations are abstract entities and he developed a detailed neural network theory for this function. In his model the hippocampus gets inputs, direct or indirect, from all the sensory areas, and "binds" them in a way that later, when cued with partial information, it can retrieve them. Marr's theory was an attempt to structure such narrative into a well-defined mathematical model, aiming to understand the anatomical structure of the hippocampus based on the memory impairment described in patients with hippocampal damage. This general logic is clear, and it has been profoundly inspirational for later work by many researchers. The implementation, however, was rather complicated, additionally hampered by the lack of adequate mathematics – it will be contributed by physicists over 10 years later – and of adequate numerics.

Marr's work did not consider place cells, that were being discovered at the same time by O'Keefe and Dostrovsky in rodents [23]. The discovery would stimulate a computational hypothesis in a different direction: that the location of the animal in space is computed within the hippocampus, and therefore its internal circuitry has to be understood as functional to self-localization, and hence in general to navigation, rather than to memory [44].

16 years later McNaughton and Morris recombined the two hippocampal narratives – the memory function and the spatial function– in a review [45]: they suggested that the hippocampal circuitry *stores* spatial representations within its synapses. The emphasis of the review was on the mechanics of learning, in particular, McNaughton and Morris suggested a set of simple network models, all based on the Hebb [46] idea that "neurons that fire together, wire together". At the core of each network there was a matrix of associatively modifi-

able weights, which was taken to capture the occurrence of conjunctive activity between input patterns on two streams.

3.2 ATTRACTOR NEURAL NETWORKS

The recurrent connectivity of the auto-associative model, one of those proposed in [45] and resembling the connectivity structure of the hippocampal region CA3 (see Appendix A), implies that the neurons, serving as inputs and output to the same synaptic matrix, will tend to reach a stable configuration, or *pattern*, if they can find one in which the activation of each neuron is consistent with that of the neurons that feed its inputs. This consistency is of the same nature as that describing the relaxation dynamics of dissipative physical systems of interacting variables to a steady state, as envisaged by John Hopfield in his seminal paper on content addressable memories [47]. Amit, Gutfreund and Sompolinsky showed how the attractors of such dynamics can be studied with a beautiful nontrivial mathematical formalism derived from the statistical physics of disordered systems [13].

In particular, Amit, Gutfreund and Sompolinsky developed a mean field theory for a fully connected network composed of binary units storing uncorrelated patterns and "*diagnosed*" the retrieval of a memory state as the dynamical persistence of a pattern evoked by external inputs [13]. In this, they derived an overlap parameter between the stored and retrieved memories and saw that above a certain number P_c of patterns to be memorized this overlap abruptly decreases to zero, see Fig. 2a. As a consequence they could define the critical storage capacity for a network of N units as $\alpha_c = \frac{P_c}{N} \simeq 0.138$ at which patterns are retrieved. Beyond this memory load no retrieval is possible, and the thermodynamic average of the overlap parameter drops abruptly from $m = 0.967$ to zero, in the formal analysis of the model. At m=0.967, effectively only 1.5% of the units are in a state different from the one they are expected to take in the memory.

Their sophisticated calculations were later named as "the first triumph of statistical physics" applied to neural networks [11] or a first *tour de force* [48], the second one being the calculations developed three years later by Elizabeth Gardner [14].

Gardner, in particular, formalized for the same binary fully connected network considered by Amit *et al.* the fractional volume of the interaction space and moved with respect to [13], to consider as dynamical variables the connections instead of the activities of each unit. Through a replica approach she could show that increasing the load (i.e. the number of patterns to be memorized) such interaction volume progressively shrinks. As a consequence she could estimate

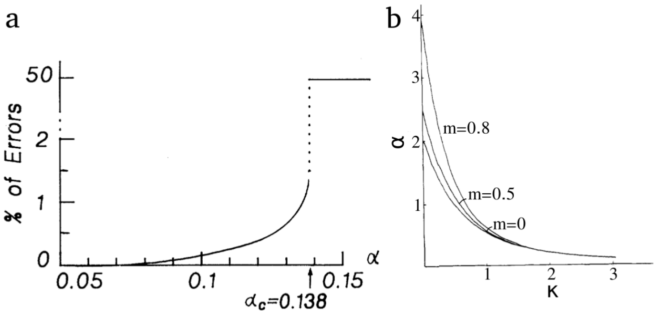


Figure 2: a) Transition of the % of error in the retrieved patterns for a fully connected binary associative network endowed with Hebbian learning [13] as a function of the storage load α. b) Storage capacity for an optimally connected binary associative network as a function of a stability parameter κ for different values of the magnetization m, the result reported in the text is for $\kappa = 0$ and $m = 0$. Figure a) is adapted from [13] figure b) from [14].

the critical storage capacity $\alpha_c = \frac{P_c}{C} = 2$ as the maximal number of patterns P_c which can be perfectly retrieved (i.e. with an overlap=1) in a network having C connections per unit, see Fig. 2b. This value was considered as the upper bound a neural network can reach if the connections are perfectly tuned, as it will be later put into practice with *back-propagation* algorithms.

As will be introduced in Sect. 4.1, the 14 fold increase in the storage capacity between Amit's *et al.* [13] and Gardner's [14] results, together with the success of back-propagation algorithms and their consequent powerful applications in artificial intelligence, consolidated the belief that self-organized Hebbian learning is inefficient in storing memories as compared to long training procedures. The belief was not redeemed with the following analytical studies which, even if still focusing on binary responses, considered more biologically realistic features as high dilution in the connections and sparse coding in the patterns [49–52].

3.2.1 *Comparing the results*

In order to compare the results obtained in those studies it is relevant to note that the 1.5% error bound only applies to the fully connected Hopfield model studied by Amit, Gutfreund and Sompolinsky [13], in which the retrieval/no-retrieval transition when the capacity is reached is a first-order transition. For the highly diluted connectivity Hopfield model studied by Derrida, Garnder and Zippelius [49], however, this transition becomes smooth, and the overlap approaches zero as $m \sim \sqrt{\alpha - \alpha_c}$. The error rate at capacity, which in this case is at $\alpha_c = \frac{P}{C} = \frac{2}{\pi}$ becomes 50%. Still, the Gardner optimal capacity calcu-

lation yields the same optimal capacity of 2 for diluted connectivity, as the approach is equivalent for fully connected or highly diluted systems. In other words, for diluted networks endowed with binary units, which can be argued to be more biologically realistic than the fully connected ones, there is still a large gap between the optimal and the Hebbian capacity, even though the Hebbian error at capacity is 50% and the overlap is zero.
In addition to its dependence on the level of connectivity, this "error", as estimated in Amit *et al.* [13], is only really meaningful when the stored patterns are binary and one can count how many neurons should have been active in the retrieved pattern but are not, and vice versa: this is directly reflected in the so-called overlap. In Part ii of the thesis we derive the Gardner storage capacity for realistic units responses. In this case, with non-binary patterns, defining error in such a way is not feasible, as being active or inactive is not the only information in the pattern. Therefore, as in previous works on associative networks of graded response units [53–55] one can only look at whether the overlap vanishes in the thermodynamics limit, or remains non-zero. In Sect. 6.2 we explicitly evaluate this transition for graded responses and Hebbian learning.

3.2.2 *The shift to Threshold linear unit responses*

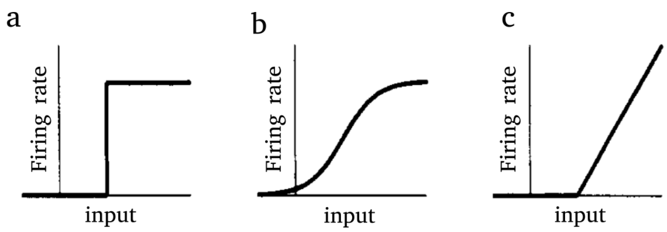


Figure 3: Examples of a) Binary/Heaviside b) Sigmoidal c) Threshold-linear unit responses. Figure adapted from [56]

As introduced by Treves in the early 90's [56, 57] assuming binary units, though useful for analytical calculations, may hamper a comprehensive understanding of the actual phenomena occurring in the brain. Some of the most general features neuronal responses exhibit, indeed, can be summarized as i) inactivity below a certain voltage ii) strong dependence on the input level above the threshold iii) saturation above a certain value of the input, due to each neuron's refractory period. When the binary assumption (Fig. 3a) is adopted one considers that neurons spend most of their time either silent ot at the saturation level, squeezing the intermediate values to a single point. One could thus adopt a sigmoidal transfer function (Fig. 3b), which, from a mathematical point of view would however reduce, in certain limits such as high gain, to the Heaviside step function [56].

The alternative proposed by Treves in 1990 was the Threshold-linear (TL) transfer function, already adopted at least from the 50′s in other fields of neuroscience, such as the studies on the retina [58, 59]. The TL input-current-to-output-frequency function was shown to be simple enough to be mathematically treatable in statistical mechanics approaches, but complex enough to grasp the intermediate dependence on the input and the resting state below the threshold. Neglecting the description of the saturation, however, one should consider long-time collective effects preventing neurons to reach it [56].
With those unit responses a mean field theory was developed and the storage capacity was estimated for the fully connected [57], the highly diluted [53], the directed [54] networks and then generalised to arbitrary dilution in [60].
In Sect. 6.1 a recap of the calculations for the highly diluted limit is reported. In Sect. 5 the Gardner storage capacity for a network endowed with Threshold Linear units is explicitly derived.

3.3 CONTINUOUS ATTRACTOR NEURAL NETWORKS

Applying the formalism, however, and even simply conceptualizing attractor dynamics, is less straightforward when dealing with the representation of spatial, continuous variables instead of simple uncorrelated patterns.

Let us take therefore first look at attractor dynamics in Head direction (HD) cells to understand basic aspects of continuous attractor dynamics in the representation of space. With Head direction cells, introduced in Chapter 2, the striking finding is that the direction that most activates a cell remains the same in every environment, familiar or new. In fact, this is striking because often the information on the basis of which the animal can calculate its head direction is partially misleading, e.g., when an object has been moved. Further, when most of it is not coming through the senses, for example because lights are turned off, and olfactory cues have been washed, HD can be reconstructed from memory, if a system exists that keeps it in memory.

This system can be an attractor network, and in fact such an observation has motivated the development of a simplified version of the theory of continuous attractor neural networks. In 1995, Skaggs *et al.* [61] proposed that a *ring attractor* could interpret sensory cues and keep HD in active (short-term) memory. To understand it intuitively, imagine: one places head direction cells on a ring, each at the angle it is most responsive to (see Fig. 4), and the connections between the neurons are taken to have been strengthened by Hebbian plasticity, resulting in neurons close to each other on the imaginary ring exciting each other. What we can observe then, is a bump of activity or an "activity clump", which would correspond to the animal's head direction, wherever it is pointing, among the 2π directions on the ring. The interactions among the units – producing attractor dynamics – compactify, stabilize and can keep in short-term memory a position on the ring.

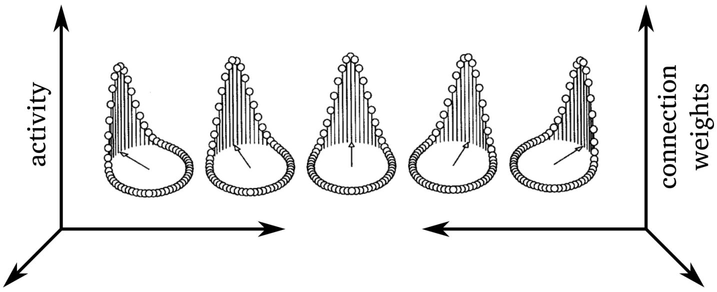


Figure 4: Head direction cell ring. Each plot represents either the connection strength J_{ij} of unit j (pointed with the arrow) towards all other units; or one among the continuous manifold of fixed point configurations of activity towards which the network can evolve due to the underlying connectivity structure. Adapted from [62].

Could this system also include the selection of one among a number of rings? The question becomes very concrete, and easy to visualize, if applied to place cells. In this case each place cell is placed at the position of its field center on a two dimensional abstract-sheet resembling the real environment. The same idea of connecting nearby cells applies then also to two dimension. The continuous manifold, in this case, corresponds to configurations representing the activity of all neurons in a specific position of the environment. The idea, then, is that the movement of the animal translates, through some path-integration mechanism, into a drift of the configuration of activities along the manifold towards the closest fixed configuration. It was determined that such continuous attractors can store multiple distinct *charts* or *maps* within the same connectivity matrix [55, 63], as attractor neural networks can store independent patterns. Each map, in this case, is a continuous manifold of solutions, each representing the ensemble of the population activities at each position of one putative environment. Additional details can be found in Sect. 10.

Let us now take a step back and specify a few aspects which will turn useful in Part iii of this thesis.

3.3.1 *Learning with a Kernel or through the Hebbian rule*

If one considers an open ring attractor, i.e. a line attractor with periodic boundary conditions, one can imagine each unit to be placed at a position on this line. If one thinks about place cells in one dimension, then the position of each neuron would be the center of its field. The idea introduced above is that units which are closed on the line have strong connections, those which are far, instead, have weak or none. This can be obtained in two ways:

1. Defining mathematically an interaction Kernel $\mathcal{K}$ as a function of the distance between the position of each cells x, and thus writing the connectivity matrix as

$$J_{ij} = \mathcal{K}(|x_i - x_j|) \tag{1}$$

2. First creating the activity profiles $\{\vec{\eta}\}$ to be memorized, where $\vec{\eta}(s)$ is the activity of all units in a position s and $\vec{\eta}_i$ is the firing rate map of unit i in the environment at each position s. Then, defining the connections through the Hebbian rule

$$J_{ij} = \int ds \left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\left(\frac{\eta_j(s)}{\langle\eta\rangle} - 1\right) \tag{2}$$

where $\langle\eta\rangle$ is the mean activity. Both approaches lead to "clumped", i.e., localized, activity states for the examples described in this section. However, if one wants to explore the behaviour of irregular systems, in which the activity profiles of different units cannot be simply described in relation to a single place field center, then mainly the second approach is the one to follow.

3.3.2 *The activity-space and the overlap-space*

When one refers to "clumped" activity states, "bumps" or "activity pockets" can be imagined in two ways, corresponding to the spaces in which the bump is visualized

1. In the way we have described it so far the bump is visualized in the activity space: placing each unit at the position of its field center (or its head direction selectivity) one can see that the fixed point configurations are clustered around a certain unit.

2. An alternative visualization, which turns out to be particularly useful when studying irregular systems, is in the overlap space related to point 2. in Sect. 3.3.1. For each dynamical variable $\vec{V}$ describing the configuration of activity at a certain time, or a fixed point configuration, one can define an overlap parameter

$$O\left(\vec{V}, \vec{\eta}(s)\right) = \frac{\sum_i^N V_i \cdot \eta_i(s)}{\sqrt{\sum_i^N (V_i)^2 \cdot \sum_i^N (\eta_i(s))^2}} \tag{3}$$

giving the cosine similarity between the current activity configuration and the vector $\eta\vec{(s)}$ representing a candidate fixed point on the manifold stored with Eq. (2) . Evaluating this overlap between a given a vector $\vec{V}$ and all discretized $\vec{\eta}(s)$ (one per each discretized position s) and plotting this value at the s position corresponding to $\vec{\eta}(s)$ one will obtain, for a continuous attractor, a clump centered at a goven s, with maximal value 1.

The second approach enables to visualize and analyse bumps when these are not as evident in the activity space. Imagine a perfectly regular continuous attractor where all patterns to be memorized $\vec{\eta}(s)$ are translationally invariant. If one evaluates J_{ij} with (2) and runs dynamics starting from a certain $\vec{\eta}(s)$ one will get to a stable configuration still centered at s. This can be seen in the activity space, Fig.5a and in the overlap space Fig.5b. Both are bumps. While in the activity space, due to the tuning of the parameters, one can see that the range of activity values varies from the initial configuration $\vec{\eta}(s = 500)$ (red) to the fixed one (black), in the overlap space this difference is hardly visible. If however one tries to store irregular maps, where each unit has more than one field and/or these are irregular in shape, as in Fig. 5 c and d, then the bump is clearly analyzable only in the overlap space.

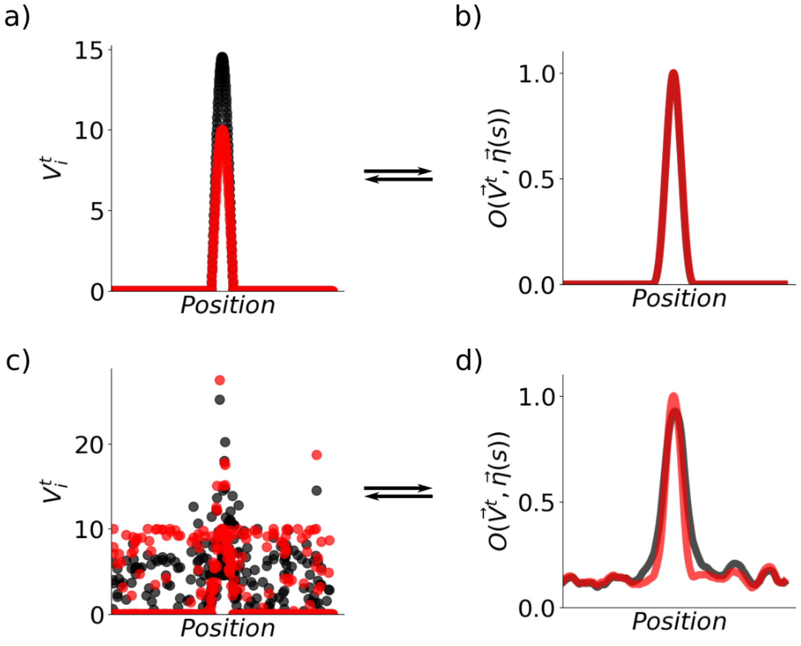


Figure 5: Examples of the visualization of the bump in *activity* space (a,c) or in the *overlap* space (b,d) for a regular continuous attractor (a,b) and an irregular one (c,d). Red corresponds to the initial condition $\vec{\eta}(s)$, black to the fixed points of the dynamics.

Part II

LEARNING EFFICACY WITH THRESHOLD-LINEAR UNITS

Since the 80s, there has been a general consensus that biologically plausible self-organized learning rules, such as the Hebbian learning rule, are very inefficient as compared to iterative algorithms. An important contributing factor to the formation of this consensus comes from the theoretical analyses performed by Elizabeth Gardner of the best possible learning outcome in binary networks. Here we derive the Gardner storage capacity for associative networks of threshold linear units and show that when attention is shifted towards biologically plausible graded response units the emerging scenario varies drastically: Hebbian learning turns out to be highly efficient to store memories through a sparsification of the retrieved patterns.

4 INTRODUCTION: ELIZABETH GARDNER'S APPROACH

Elizabeth Gardner (Cheshire 1957, Edinburgh 1988) was a scientist who, at age of 30, was regarded among the most profound thinkers in the emerging field of neuronal networks. She died a few months before turning 31, nine months after starting her five-year Advanced Fellowship [64].

She was one of the few women in an academic environment mostly comprised of men. In the memorials written by her closest colleagues, she is remembered for the brilliant creativity, the outstanding intellectual standards [65] and the reserved manners [64].

This part of the thesis is based on the mathematical intuitions and formalism developed by Elizabeth Gardner in the two years before her death [14]. In particular it extends the treatment of the interaction space of neuronal networks, which she originally defined for binary units, to neuronal plausible unit responses (i.e. threshold linear or ReLu), introduced in the 90′ [56] in the analysis of associative networks.

The work presented in this part of the thesis has been done under the co-supervision of Yasser Roudi and is published in [66], except for Sect. 6.2 which will be published in [67].

The coming chapters are organized as follows: Chapter 5 contains the core analytical work, whose main findings and ideas are summarized in Sect. 5.2.1. Chapter 6 is devoted to a comparison between the results obtained with Hebbian learning and those derived in chapter 5. In particular, after summarizing, in Sect. 6.1, the main results derived with Hebbian learning [53, 54, 57], a comparison between theoretical distributions is performed. Chapter 7 extends the comparison to sample experimental data obtained in the 90′s and chapter 8 discusses the main results of the whole Part and relates them with other works. First, the next section is dedicated to a concise introduction and overview.

4.1 LEARNING AND UNIT RESPONSES

Learning in neuronal networks is believed to happen largely through changes in the weights of the synaptic connections between neurons. Local learning rules, those that self-organize through weight changes depending solely on the activity of pre- and post-synaptic neurons, are generally considered to be more biologically plausible than non-local ones [46, 68, 69]. But how effective are local learning rules? Quite ineffective, has been the received wisdom since the 80's, when non-local iterative algorithms came to the fore. However, this wisdom, when it comes to memory storage and retrieval, is largely based on analysing networks of binary neurons [14, 47, 49, 70], while neurons in the brain are not binary.

A better, but still mathematically simple description of neuronal input-output transformation is through threshold-linear (TL) activation function [56, 71], also predominantly adopted in recent deep learning applications (called ReLu in that context) [72–75]. Therefore, one may ask if the results from the 80's highlighting the contrast between the effective, iterative procedures used in machine learning and the self-organized, one-shot, perhaps computationally ineffective local learning rules are valid beyond binary units [76].

The Hopfield model, a most studied model of memory, is a fully connected network of N binary units endowed with a local, *Hebbian* learning rule [47, 70]: the weight between two units increases if they have the same activity in a memory pattern; otherwise it decreases. The network can retrieve only up to $p_{max} \simeq 0.14N$ patterns, while, in comparison, Elisabeth Gardner showed [14] that with C connections per unit, the optimal capacity that such a network can attain is $p_{max} = 2C$, about 14 times higher; the bound can be approached through iterative procedures like back-propagation that progressively reduce the difference between current and desired output. This consolidated the impression that unsupervised, Hebbian plasticity may well be of biological interest, but is rather inefficient for memory storage. In the fully connected Hopfield model, the transition to no-retrieval is discontinuous: right below the storage capacity, $\sim 1.5\%$ of units in a retrieved pattern are misaligned with the stored pattern, but 50%, i.e., chance level, just above the capacity [70]. This rather low error certainly contributes to the low capacity. However, the negative characterization of Hebbian learning in binary networks persisted even when more errors occur: in the more biologically relevant highly diluted networks the error smoothly goes to 50% [50], but the capacity is still a factor of 3 away [49], *approaching* the bound only when the fraction of active unit in each pattern is $f \ll 1$ [51].

What about TL units? Are they more efficient in the unsupervised learning of memory patterns? Here we study the optimal pattern capacity *à la Gardner* in networks of TL units. Past work discussed above [51] had suggested that the distribution of activity (along with the connectivity) may play a role in how efficient Hebbian learning is, but, back then, this only meant changing f. Besides being a better model of neuronal input-output transformation, by allowing non-binary patterns, TL units permit a better understanding of the interplay between the retrieval properties of recurrent networks and the distribution of the activity stored in the network. In fact, we show that while for binary patterns the Gardner bound is larger than the Hebbian capacity no matter how sparse the code, this does not, in general, hold for non-binary stored patterns: the Hebbian capacity can even surpass the bound. This perhaps surprising violation of the bound is because the Gardner calculation imposes an infinite output precision [77], while Hebbian learning exploits its loose precision to *sparsify* the retrieved pattern. In other words, with TL units, Hebbian capacity can get much closer to the optimal capacity or even surpass it, by retrieving a sparser version of the stored pattern. We find that experimentally observed distributions from the Inferior-Temporal (IT) visual cortex [78], which can be taken as patterns to be stored, would be sparsified about 50% by Hebbian learning, and would reach about $50\% - 80\%$ of the Gardner bound.

5 GARDNER STORAGE CAPACITY FOR THRESHOLD LINEAR UNITS

5.1 MODEL DESCRIPTION

We consider a network of N units and p patterns of activity, $\{\eta_i^\mu\}_{i=1,..,N}^{\mu=1,..,p}$ each representing one memory stored in the connection weights via some procedure. Each η_i^μ is drawn independently for each unit i and each memory μ from a common distribution $\Pr(\eta)$. The activity of unit i is denoted by ν_i and is determined by the activity of the C units feeding to it as

$$\nu_i = g[h_i - \vartheta]^+ \tag{4a}$$

$$h_i\{\nu_i\} = \frac{1}{\sqrt{C}}\sum_j J_{ij}\nu_j, \tag{4b}$$

where $[x]^+ = x$ for $x > 0$ and $= 0$ otherwise; and both the gain g and threshold ϑ are fixed parameters. The *storage capacity, or capacity for short,* is defined as $\alpha_c \equiv p_{max}/C$, with p_{max} the maximal number of memories that can be stored and individually retrieved. The synaptic weights J_{ij} are taken to satisfy the spherical normalization condition for all i

$$\sum_{j\neq i} J_{ij}^2 = C. \tag{5}$$

We are interested in finding the set of J_{ij} that satisfy Eq. (5), such that patterns $\{\eta_i^\mu\}_{i=1,..,N}^{\mu=1,..,p}$ are self-consistent solutions of Eqs. (4), namely that for all i and μ we have, $h_i^\mu = \vartheta + \eta_i^\mu/g$ if $\eta_i^\mu > 0$ and $h_i^\mu \leqslant \vartheta$ if $\eta_i^\mu = 0$.

5.2 REPLICA ANALYSIS TO DERIVE THE STORAGE CAPACITY

In this section, we give a detailed mathematical derivation of the Gardner bound, reported instead schematically in the next section.

We start by considering a single threshold-linear unit whose activity is denoted by u. The unit receives C inputs v_j, for $j = 1 \cdots C$, through synaptic weights J_j. The activity of the unit is determined through the threshold-linear activation function as

$$\begin{aligned} u &= g[h_i - \vartheta]^+ \\ h\{v\} &= \frac{1}{\sqrt{C}} \sum_j J_j v_j, \end{aligned} \tag{6}$$

We assume that we have p patterns of activity over the inputs, that we denote by ξ_j^μ, with $\mu = 1 \cdots p$. For each input pattern μ we also consider a desired output activity for each unit that we denote η^μ. We are interested in finding how many patterns can be stored in the synaptic weights, such that the input activity elicits the desired output activity, assuming that the synaptic weights satisfy the spherical constraint

$$\sum_{j \neq i} J_j^2 = C. \tag{7}$$

Following [14], the fractional volume in the space of interactions J that satisfy Eq. (7) and the correct output η^μ given the inputs ξ_j^μ can be written as

$$\begin{aligned} V = {} & \frac{\int \prod_{j,j\neq i} dJ_j \delta\left(\sum_j J_j^2 - C \right) \prod_\mu \left[\left(1 - \delta_{\eta^\mu,0} \right) \delta\left(h^\mu - \vartheta - \frac{\eta^\mu}{g} \right) \right.}{\int \prod_{j,j\neq i} dJ_j \delta\left(\sum_j J_j^2 - C \right)} + \\ & + \frac{\left. \delta_{\eta^\mu,0} \Theta\left(\vartheta - h_i^\mu \right) \right]}{\int \prod_{j,j\neq i} dJ_j \delta\left(\sum_j J_j^2 - C \right)}, \end{aligned} \tag{8}$$

Calculating the optimal capacity essentially boils down to calculating, in the thermodynamic limit $C \to \infty$, the expectation of the logarithm of this fractional volume V over the distribution of η and ξ and finding for what value of p it shrinks to zero. For calculating $\langle \ln V \rangle_{\eta,\xi}$, we use the replica trick $\langle \ln V \rangle = \lim_{n\to 0} \frac{\langle V^n\rangle - 1}{n}$, which turns the problem to that of computing the replica average $\langle V^n \rangle_{\xi,\eta}$, namely

$$\langle V^n \rangle_{\xi,\eta} = \Bigg\langle \prod_{a=1,..,n} \prod_{\mu} \frac{\int \prod_{j,j\neq i} dJ_j^a \delta\Big(\sum_j (J_j^a)^2 - C \Big)}{\int \prod_{j,j\neq i} dJ_j^a \delta\Big(\sum_j (J_j^a)^2 - C \Big)} \cdot \tag{9}$$
$$\cdot \Bigg[\Big(1 - \delta_{\eta^\mu,0} \Big) \delta\Big(h^{a,\mu} - \vartheta - \frac{\eta^\mu}{g} \Big) + \delta_{\eta^\mu,0} \Theta(\vartheta - h^{a,\mu}) \Bigg] \Bigg\rangle_{\xi,\eta} .$$

We first compute the numerator. To compute the averages over ξ and η in the numerator, we note that the delta function can be written as

$$\begin{aligned} \delta(h^{a,\mu} - \vartheta - \frac{\eta^\mu}{g}) &= \int \frac{dx_\mu^a}{2\pi} \exp\left\{ i x_\mu^a \Big(\frac{1}{\sqrt{C}} \sum_j J_j^a \xi^\mu - \vartheta - \frac{\eta^\mu}{g} \Big) \right\} \\ &= \int \frac{dx_\mu^a}{2\pi} \exp\Big[- \frac{i x_\mu^a}{g} \Big(\eta^\mu + g\vartheta \Big) \Big] \exp\Big[\frac{i x_\mu^a \sum_j J_j^a \xi^\mu}{\sqrt{C}} \Big]. \end{aligned} \tag{10}$$

For the average of the Heaviside function, we write

$$\begin{aligned} \Theta(\vartheta - h^{a,\mu}) &= \int_0^\infty d\lambda_\mu^a \delta[\lambda_\mu^a - (\vartheta - h^{a,\mu})] \\ &= \int_0^\infty \frac{d\lambda_\mu^a}{2\pi} \int_{-\infty}^\infty dy_\mu^a \exp[i y_\mu^a (\lambda_\mu^a - (\vartheta - h^{a,\mu}))] \\ &= \int_0^\infty \frac{d\lambda_\mu^a}{2\pi} \int_{-\infty}^\infty dy_\mu^a \exp\Big[i y_\mu^a (\lambda_\mu^a - \vartheta) \Big] \exp\Big[\frac{i y_\mu^a \sum_j J_j^a \xi^\mu}{\sqrt{C}} \Big]. \end{aligned} \tag{11}$$

We now use the above identities in Eqs. (10) and (11) to compute the following quantity that appears in the numerator of Eq. (9), assuming independently drawn ξ and η as

$$\begin{aligned} e^{CM} &\equiv \Bigg\langle \prod_{\mu,a} (1 - \delta_{\eta^\mu,0}) \delta(h^{a,\mu} - \vartheta - \frac{\eta^\mu}{g}) + \delta_{\eta^\mu,0} \Theta(\vartheta - h^{a,\mu}) \Bigg\rangle_{\xi,\eta} \\ &= \prod_\mu \Bigg\langle (1 - \delta_{\eta^\mu,0}) \Big\langle \prod_a \delta(h^{a,\mu} - \vartheta - \frac{\eta^\mu}{g}) \Big\rangle_{\xi^\mu} + \\ &+ \delta_{\eta^\mu,0} \Big\langle \prod_a \Theta(\vartheta - h^{a,\mu}) \Big\rangle_{\xi^\mu} \Bigg\rangle_{\eta^\mu} . \end{aligned} \tag{12}$$

In order to compute the average of the delta functions in Eq.(12), we use the approximation

$$\begin{aligned}\Big\langle \exp(x) \Big\rangle &= \langle 1 + x + \frac{x^2}{2} + \mathcal{O}(x^3) \rangle = 1 + \langle x \rangle + \frac{\langle x^2 \rangle}{2} + \langle \mathcal{O}(x^3) \rangle \\ &\approx \exp\left\{ \langle x \rangle + \frac{\langle x^2 \rangle}{2} - \frac{\langle x \rangle^2}{2} \right\}\end{aligned}$$

to calculate the following average

$$\begin{aligned}&\left\langle \exp\left\{ \frac{i \sum_{a,j} x_\mu^a J_j^a \xi_j^\mu}{\sqrt{C}} \right\} \right\rangle_{\xi^\mu} = \\ &= \exp\left\{ \frac{i}{\sqrt{C}} \sum_{a,j} x_\mu^a J_j^a \langle \xi_j^\mu \rangle - \frac{1}{2C} \sum_{a,b,j,k} x_\mu^a x_\mu^b J_j^a J_k^b \langle \xi_j^\mu \xi_k^\mu \rangle + \right. \\ &\left. - \frac{1}{2} \left(\frac{i}{\sqrt{C}} \sum_{a,j} x_\mu^a J_j^a \langle \xi_j^\mu \rangle \right) \left(\frac{i}{\sqrt{C}} \sum_{b,k} x_\mu^b J_j^b \langle \xi_k^\mu \rangle \right) \right\} \\ &= \exp\left\{ \frac{i}{\sqrt{C}} \sum_{a,j} x_\mu^a J_j^a \langle \xi_j^\mu \rangle - \frac{1}{2C} \Big[\sum_{a,b,j} x_\mu^a x_\mu^b J_j^a J_j^b \langle (\xi_j^\mu)^2 \rangle + \right. \\ &+ \sum_{a,b,j,k!=j} x_\mu^a x_\mu^b J_j^a J_k^b \langle \xi_j^\mu \rangle \langle \xi_k^\mu \rangle \Big] + \frac{1}{2C} \Big[(\sum_{a,b,j} x_\mu^a x_\mu^b J_j^a J_j^b \langle \xi_j^\mu \rangle^2) + \\ &\left. + \sum_{a,b,j,k!=j} x_\mu^a x_\mu^b J_j^a J_k^b \langle \xi_j^\mu \rangle \langle \xi_k^\mu \rangle \Big] \right\} \\ &= \exp\left\{ \frac{i}{\sqrt{C}} \sum_{a,j} x_\mu^a J_j^a \langle \xi_j^\mu \rangle - \frac{1}{2C} \sum_{a,b,j} x_\mu^a x_\mu^b J_j^a J_j^b \langle (\xi_j^\mu)^2 \rangle + \right. \\ &\left. + \frac{1}{2C} \sum_{a,b,j} x_\mu^a x_\mu^b J_j^a J_j^b \langle \xi_j^\mu \rangle^2 \right\}\end{aligned} \tag{13}$$

where in going from the second to the third line in Eq. (13), we have used the fact that $\langle \xi_j^\mu \xi_k^\mu \rangle = \langle \xi_j^\mu \rangle \langle \xi_k^\mu \rangle$. Expanding the second exponential in the second line of Eq. (10), we can write, in the large C limit

$$\begin{aligned}&\Big\langle \prod_a \delta(h^{a,\mu} - \vartheta - \frac{\eta^\mu}{g}) \Big\rangle_{\xi^\mu} = \\ &= \int_{-\infty}^{\infty} \Big[\prod_a \frac{dx_\mu^a}{2\pi} \Big] \exp\Big[- \frac{i}{g} (\eta^\mu + g\vartheta) \sum_a x_\mu^a + i d_1^{inp} \sum_a x_\mu^a m^a + \\ &- \frac{d_3^{inp}}{2} \Big(\sum_a (x_\mu^a)^2 + 2 \sum_{a<b} x_\mu^a x_\mu^b q^{ab} \Big) \Big] \\ &\equiv I_1(q^{ab}, m^a, \eta^\mu)\end{aligned} \tag{14}$$

in which we have assumed symmetric replicas and defined $d_1^{inp} \equiv \langle \xi_j^\mu \rangle$, $d_2^{inp} \equiv \langle (\xi_j^\mu)^2 \rangle$, $d_3^{inp} \equiv d_2^{inp} - (d_1^{inp})^2$ and

$$q^{ab} = \tfrac{1}{C} \textstyle\sum_j J_j^a J_j^b \tag{15a}$$

$$m^a = \tfrac{1}{\sqrt{C}} \textstyle\sum_j J_j^a \tag{15b}$$

Similarly, using the identity in Eq. (11) we have

$$\begin{aligned} &\Big\langle \prod_a \Theta(\vartheta - h^{a,\mu}) \Big\rangle_{\xi^\mu} = \\ &= \int_0^\infty \Big[\prod_a \frac{d\lambda_\mu^a}{2\pi} \Big] \int_{-\infty}^\infty \Big[\prod_a dy_\mu^a \Big] \exp \Big[i \sum_a (\lambda_\mu^a - \vartheta) y_\mu^a + \\ &+ i d_1^{inp} \sum_a y_\mu^a m^a - \frac{d_3^{inp}}{2} \Big(\sum_a (y_\mu^a)^2 + 2 \sum_{a<b} y_\mu^a y_\mu^b q^{ab} \Big) \Big] \\ &\equiv I_2(q^{ab}, m^a). \end{aligned} \tag{16}$$

Using Eq. (14) and (16), the quantity $M(q^{ab}, m^a)$ defined through Eq. (12) can be written as

$$\begin{aligned} &M(q^{ab}, m^a) = \\ &\frac{p}{C} \ln \Bigg[\langle (1 - \delta_{\eta^\mu,0}) I_1(q^{ab}, m^a, \eta^\mu) + \delta_{\eta^\mu,0} I_2(q^{ab}, m^a) \rangle_{\eta^\mu} \Bigg]. \end{aligned} \tag{17}$$

We now insert Eq. (17) back to Eq. (9) and enforce the definitions of m and q in Eq. (15) using the identities

$$\begin{aligned} 1 &= C \int \frac{dq^{ab} d\hat{q}^{ab}}{2i\pi} \exp \Bigg(- C \hat{q}^{ab} q^{ab} + \hat{q}^{ab} \sum_j J_j^a J_j^b \Bigg) \\ 1 &= \sqrt{C} \int \frac{dm^a d\hat{m}^a}{2i\pi} \exp \Bigg(- \sqrt{C} \hat{m}^a m^a + \hat{m}^a \sum_j J_j^a \Bigg) \end{aligned} \tag{18}$$

and the normalization of Eq. (9) using

$$\delta \Bigg(\sum_j {J_j^a}^2 - C \Bigg) = \int \frac{dE^a}{4i\pi} \exp \Big(- \frac{E^a}{2} \sum_{j \neq i} {J_j^a}^2 + \frac{C E^a}{2} \Big) \tag{19}$$

such that the numerator in Eq. (9) can be written as

$$\begin{aligned} A = &\int \Big[\prod_a \frac{dE^a}{4i\pi} \Big] \Big[\prod_a \sqrt{C} \frac{dm^a d\hat{m}^a}{2i\pi} \Big] \Big[\prod_{a<b} C \frac{dq^{ab} d\hat{q}^{ab}}{2i\pi} \Big] \cdot \\ &\cdot e^{C[M(q,m) - \frac{1}{\sqrt{C}} \sum_a \hat{m}^a m^a - \sum_{a<b} \hat{q}^{ab} q^{ab} + \sum_a \frac{E^a}{2}]} \\ &\cdot \int \Big[\prod_{j,a} dJ_{ij}^a \Big] e^{- \sum_{a,j} \frac{E^a}{2} (J_j^a)^2 + \sum_{a,j} \hat{m}^a J_j^a + \sum_{a<b} \hat{q}^{ab} J_{ij}^a J_{ij}^b}. \end{aligned} \tag{20}$$

Defining the function

$$\begin{aligned}W(\hat{q}^{ab}, \hat{m}^{a}, E^{a}) = \ln \int \Big[\prod_{a} dJ^{a}\Big] \exp\Big(-\frac{1}{2}\sum_{a} E^{a}(J^{a})^{2} + \\ + \sum_{a} \hat{m}^{a} J^{a} + \sum_{a<b} \hat{q}^{ab} J^{a} J^{b}\Big)\end{aligned} \tag{21}$$

we can write

$$\begin{aligned}A = \int \Big\{ \Big[\prod_{a} \frac{dE^{a}}{4i\pi}\Big]\Big[\prod_{a} \sqrt{C}\frac{dm^{a} d\hat{m}^{a}}{2i\pi}\Big]\Big[\prod_{a<b} C\frac{dq^{ab} d\hat{q}^{ab}}{2i\pi}\Big] \cdot \\ \cdot\, e^{C[M(q^{ab},m^{a}) + W(\hat{q}^{ab},\hat{m}^{a},E^{a}) - \frac{1}{\sqrt{C}}\sum_{a}\hat{m}^{a}m^{a} - \sum_{a<b}\hat{q}^{ab}q^{ab} + \sum_{a}\frac{E^{a}}{2}]} \Big\}\end{aligned} \tag{22}$$

We can then compute A in Eq. (22) using the saddle point approximation, by maximizing the argument of the exponential, that is maximising

$$\begin{aligned}G(q^{ab}, \hat{q}^{ab}, m^{a}, \hat{m}^{a}, E^{a}) \equiv M(q^{ab}, m^{a}) + W(\hat{q}^{ab}, \hat{m}^{a}, E^{a}) + \\ - \frac{1}{\sqrt{C}}\sum_{a} \hat{m}^{a} m^{a} - \sum_{a<b} \hat{q}^{ab} q^{ab} + \sum_{a} \frac{E^{a}}{2}.\end{aligned} \tag{23}$$

In order to proceed to make this extremisation we assume a replica symmetric ansatz:

$$\begin{aligned} q^{ab} &= q \\ \hat{q}^{ab} &= \hat{q} \\ m^{a} &= m \\ \hat{m}^{a} &= \hat{m} \\ E_{a} &= E \end{aligned} \tag{24}$$

with these assumptions

$$G(q, \hat{q}, m, \hat{m}, E) = M(q, m) + W(\hat{q}, \hat{m}, E) + \frac{n}{2}(-\frac{2\hat{m}m}{\sqrt{C}} + \hat{q}q + E). \tag{25}$$

In the above Eq. (25), W and M are calculated using the limits for $n \to 0$ of the expressions in Eq. (17) and (21), as follows. For W, we use the Gaussian trick (i.e. the one dimensional Hubbard-Stratonovich transformation)

$$\begin{aligned}\int_{-\infty}^{\infty} dt \exp(-at^{2} \pm bt) = \exp\Big[\frac{b^{2}}{4a}\Big]\sqrt{\frac{\pi}{a}} \\ \to e^{-x^{2}/2} = \int_{-\infty}^{\infty} \frac{dt}{\sqrt{2\pi}} e^{-t^{2}/2 \pm tx}\end{aligned} \tag{26}$$

combined with the replica symmetric expression for W to get

$$\begin{aligned} W(\hat{m},\hat{q},E) &= \ln\int\Big[\prod_a dJ^a\Big]\exp\Big(-\frac{E}{2}\sum_a(J^a)^2+\hat{m}\sum_a J^a+ \\ &+\frac{\hat{q}}{2}\Big(\sum_a J^a\Big)^2-\frac{\hat{q}}{2}\sum_a(J^a)^2\Big) \\ &= \ln\int\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}\left[\int dJ\exp\left(-\frac{E+\hat{q}}{2}J^2+(\hat{m}+\sqrt{\hat{q}}t)J\right)\right]^n. \end{aligned} \tag{27}$$

where we have applied the transformation to the third exponent in the first line. Using $a^n\approx 1+n\log a$ and $\log(1+a)\approx a$, we have

$$W(\hat{m},\hat{q},E)=n\int\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}\ln\left[\int dJ\exp\left(-\frac{E+\hat{q}}{2}J^2+(\hat{m}+\sqrt{\hat{q}}t)J\right)\right] \tag{28}$$

In order to perform the Gaussian integrals one can show that for general a, b parameters:

$$\begin{aligned} &\int dxe^{ax^2\pm bx}=\sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}} \\ &\int\frac{dx}{\sqrt{2\pi}}e^{-x^2/2}(a+bx)^2=a^2+b^2 \end{aligned}$$

Therefore, integrating over J in Eq. (28), leads to:

$$W(\hat{m},\hat{q},E)=n\Big(\int\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}\ln\sqrt{\frac{2\pi}{E+\hat{q}}}+\int\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}\frac{(\hat{m}+\sqrt{\hat{q}}t)^2}{2(E+\hat{q})}\Big) \tag{29}$$

and integrating over t, finally leads to:

$$W(\hat{m},\hat{q},E)=\frac{n}{2}\Big[\ln(2\pi)-\ln(E+\hat{q})+\frac{\hat{q}+\hat{m}^2}{E+\hat{q}}\Big] \tag{30}$$

Computing M is a bit more tricky.

$$M(q,m)=\frac{p}{C}\ln\left[\langle(1-\delta_{\eta^\mu,0})I_1(q,m,\eta^\mu)+\delta_{\eta^\mu,0}I_2(q,m)\rangle_{\eta^\mu}\right]. \tag{31}$$

as one has to compute $I_1(q, m, \eta^\mu)$ and $I_2(q, m)$. Using the Gaussian trick in Eq. (26) and assuming replica symmetry we rewrite Eq. (14) as

$$\begin{aligned}
I_1(q, m, \xi) &= \int_{-\infty}^{\infty} \Big[\prod_a \frac{dx_\mu^a}{2\pi}\Big] \exp\Big\{\Big[-\frac{i}{g}(\eta^\mu + g\vartheta) + i d_1^{inp} m\Big] \sum_a x_\mu^a + \\
&\quad - \frac{d_3^{inp}}{2} \sum_a (x_\mu^a)^2 + - d_3^{inp} q \sum_{a<b} x_\mu^a x_\mu^b \Big\} \\
&= \int_{-\infty}^{\infty} \Big[\prod_a \frac{dx_\mu^a}{2\pi}\Big] \exp\Big\{\Big[-\frac{i}{g}(\eta^\mu + g\vartheta) + i d_1^{inp} m\Big] \sum_a x_\mu^a + \\
&\quad - \frac{d_3^{inp}}{2} \sum_a (x_\mu^a)^2 + \frac{d_3^{inp} q}{2} \sum_a (x_\mu^a)^2 - \frac{q d_3^{inp}}{2} \Big(\sum_a x_\mu^a\Big)^2 \Big\} \\
&= \int Dt \Big\{ \int \frac{dx_\mu}{2\pi} \exp\Big[-i\Big(g^{-1}\eta^\mu + \vartheta - d_1^{inp} m + \\
&\quad - t\sqrt{q d_3^{inp}}\Big) x_\mu - \frac{d_3^{inp}}{2}(1-q) x_\mu^2\Big]\Big\}^n
\end{aligned} \tag{32}$$

with $Dt = \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}$. In a very similar way we can write Eq. (16) as

$$\begin{aligned}
I_2(q, m) = \int Dt \Big\{ \int_0^{\infty} \frac{d\lambda_\mu}{2\pi} \int_{-\infty}^{\infty} dy_\mu \exp\Big[i\Big(\lambda_\mu^a - \vartheta + d_1^{inp} m + \\
+ t\sqrt{q d_3^{inp}}\Big) y_\mu - \frac{d_3^{inp}}{2}(1-q)(y_\mu)^2\Big]\Big\}^n.
\end{aligned} \tag{33}$$

We define $P(\eta^\mu > 0) = f$ and rewrite Eq. (17) as

$$\begin{aligned}
M(q, m) &= \frac{p}{C} \ln\{\langle(1 - \delta_{\eta^\mu,0})\rangle_{\eta^\mu} \langle I_1(q, m, \eta^\mu)\rangle_{\eta^\mu} + \langle \delta_{\eta^\mu,0}\rangle_{\eta^\mu} I_2(q, m)\} \\
&= \frac{p}{C} \ln\Big[f \langle I_1(q, m, \eta^\mu)\rangle_{\eta^\mu} + (1-f) I_2(q, m)\Big].
\end{aligned} \tag{34}$$

Simplifying for the sake of visualization Eq. (32) and (33) as

$$\begin{aligned}
I_1(q, m, \eta^\mu) &= \int Dt Y^n \\
I_2(q, m) &= \int Dt K^n
\end{aligned} \tag{35}$$

where

$$
\begin{aligned}
Y \equiv & \int \frac{dx_\mu}{2\pi} \exp\Big[-i\Big(g^{-1}\eta^\mu + \vartheta - d_1^{inp} m - t\sqrt{q d_3^{inp}}\Big)x_\mu + \\
& - \frac{d_3^{inp}}{2}(1-q)x_\mu^2\Big] \\
K \equiv & \int_0^\infty \frac{d\lambda_\mu}{2\pi} \int_{-\infty}^{\infty} dy_\mu \exp\Big[i\Big(\lambda_\mu^a - \vartheta + d_1^{inp} m + t\sqrt{q d_3^{inp}}\Big)y_\mu + \\
& - \frac{d_3^{inp}}{2}(1-q)(y_\mu)^2\Big]
\end{aligned}
\tag{36}
$$

one can use again $a^n \approx 1 + n \ln a$ and $\ln(1+a) \approx a$, which is valid for $n \to 0$, to write $M(q, m)$ as

$$
\begin{aligned}
M(q,m) &= \frac{p}{C} \ln\Big[f\Big\langle \int DtY^n \Big\rangle_{\eta^\mu} + (1-f)\int DtK^n\Big] \\
&= \frac{p}{C} \ln\Big[\int Dt[f\Big\langle 1 + n\ln Y \Big\rangle_{\eta^\mu} + (1-f)(1+n\ln K)\Big] \\
&= \frac{p}{C} \ln\Big[1 + n\Big(f\int Dt\Big\langle \ln Y \Big\rangle_{\eta^\mu} + (1-f)\int Dt\ln K\Big)\Big] \\
&= \frac{p}{C} n\Big(f\int Dt\Big\langle \ln Y \Big\rangle_{\eta^\mu} + (1-f)\int Dt\ln K\Big)
\end{aligned}
\tag{37}
$$

Turning back to the original notation we can further develop the terms composing the above approximation. The first one yields:

$$
\begin{aligned}
& \int Dt\Big\langle \ln Y \Big\rangle_{\eta^\mu} = \int Dt \int \Big\langle \frac{dx_\mu}{2\pi} \cdot \\
& \cdot \exp\Big[-i\Big(g^{-1}\eta^\mu + \vartheta - d_1^{inp} m - t\sqrt{q d_3^{inp}}\Big)x_\mu - \frac{d_3^{inp}}{2}(1-q)x_\mu^2\Big]\Big\rangle_{\eta^\mu} \\
& = \int Dt\Big\langle \ln\Big[\exp\Big\{-\frac{\Big(d_1^{inp} m - g^{-1}\eta^\mu - \vartheta + t\sqrt{q d_3^{inp}}\Big)^2}{2 d_3^{inp}(1-q)}\Big\} \cdot \\
& \cdot \sqrt{\frac{2\pi}{d_3^{inp}(1-q)}}\frac{1}{2\pi}\Big]\Big\rangle_{\eta^\mu} \\
& = \frac{1}{2}\Bigg[- \ln 2\pi - \ln d_3^{inp}(1-q) + \\
& - \frac{\Big\langle\Big(d_1^{inp} m - g^{-1}\eta^\mu - \vartheta\Big)^2\Big\rangle_{\eta^\mu} + q d_3^{inp}}{d_3^{inp}(1-q)}\Bigg]
\end{aligned}
$$

(38)

and the second one yields:

$$\begin{aligned}
&\int Dt \ln K = \int Dt \ln \int_0^\infty \frac{d\lambda_\mu}{2\pi} \int_{-\infty}^\infty dy_\mu \exp\left[i \left(\lambda_\mu^a - \vartheta + d_1^{inp} m + \right.\right. \\
&\left.\left. + t\sqrt{q d_3^{inp}} \right) y_\mu - \frac{d_3^{inp}}{2}(1-q)(y_\mu)^2 \right] \\
&= \int Dt \ln \int_0^\infty \frac{d\lambda_\mu}{2\pi} \exp\left[- \frac{\left(d_1^{inp} m + \lambda_{\mu - \vartheta + t\sqrt{q d_3^{inp}}} \right)^2}{1 d_3^{inp}(1-q)} \right] \cdot \\
&\cdot \sqrt{\frac{2\pi}{d_3^{inp}(1-q)}} \\
&= \int Dt \ln \int_{\frac{d_1^{inp} m - \vartheta + t\sqrt{q d_3^{inp}}}{\sqrt{d_3^{inp}(1-q)}}}^\infty \frac{dz}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}
\end{aligned} \tag{39}$$

where in the last passage we made a simple change of variables. Therefore we can rewrite Eq. (34) as:

$$\begin{gathered}
M(q,m) = \frac{p}{C} n \left\{ \frac{f}{2} \left[- \ln[2\pi d_3^{inp}(1-q)] + \right.\right. \\
\left.\left. - \frac{\left[\left\langle \left(d_1^{inp} m - g^{-1}\eta^\mu - \vartheta \right)^2 \right\rangle_{\eta^\mu} + q d_3^{inp} \right]}{d_3^{inp}(1-q)} \right] + (1-f) \int Dt \ln H(u) \right\} \\
\text{where} \\
u \equiv \frac{d_1^{inp} m - \vartheta + t\sqrt{q d_3^{inp}}}{\sqrt{d_3^{inp}(1-q)}} \\
H(u) \equiv \int_u^\infty \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}.
\end{gathered} \tag{40}$$

Now we can evaluate the derivatives

$$\frac{dG}{d\hat{m}} = \frac{dG}{d\hat{q}} = \frac{dG}{dE} = \frac{dG}{dm} = \frac{dG}{dq} = 0 \tag{41}$$

where $G = G(q, \hat{q}, m, \hat{m}, E)$ given by Eq. (25), and set them to zero to find the maximum of Eq. (25), with $W(\hat{m}, \hat{q}, E)$ given by Eq. (30) and $M(q, m)$ given by Eq. (40).

With the first three derivatives equalized to zero, which are applied only to the second and third term of Eq. (25), and assuming $Cq \gg m^2$ and $|C(1-2q)| \gg m^2$ as $C \to \infty$, we obtain the relations

$$\begin{aligned} \hat{m} &= -\frac{m}{\sqrt{C}(q-1)} \\ \hat{q} &= \frac{q}{(1-q)^2} \\ E &= \frac{1-2q}{(q-1)^2}. \end{aligned} \tag{42}$$

Substituting them into Eq. (25) we have to perform the last two derivatives. $\frac{dG}{dm}$ can be simply evaluated, applying the Leibniz integral rule $\frac{d}{dx}[\int_{a(x)}^{b(x)} f(x,t)dt] = f(x,b(x))\frac{d}{dx}b(x) - f(x,a(x))\frac{d}{dx}a(x) + + \int_{a(x)}^{b(x)} \frac{d}{dx} f(x,t)dt$ based on which $\frac{d}{dt}H(u(m)) = \frac{d}{dm}\int_{u(m)}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} = -\frac{1}{\sqrt{2\pi}} e^{-\frac{u(m)^2}{2}} \frac{d}{dm}u(m)$ yielding:

$$\begin{aligned} \frac{dG}{dm} &= 0 = -fd_1^{inp}(d_1^{inp}m - g^{-1}\langle\eta^{\mu}\rangle - \vartheta) + \\ &- \frac{\sqrt{d_3^{inp}(1-q)(1-f)d_1^{inp}}}{\sqrt{2\pi}} \int DtH(u)^{-1}e^{-u^2/2} \end{aligned} \tag{43}$$

The derivative in q is a bit more tricky:

$$\begin{aligned} \frac{dG}{dq} &= 0 = \frac{dM}{dq} + \frac{nq}{2(1-q)^2} - \frac{nq}{2(1-q)^2} = \\ \frac{p}{C}n\Bigg\{ &- \frac{f}{2}\left[\frac{\langle(d_1^{inp}m - g^{-1}\eta^{\mu} - \vartheta)^2\rangle + qd_3^{inp}}{d_3^{inp}(1-q)^2}\right] + \\ -\frac{(1-f)}{\sqrt{2\pi}} &\int Dte^{-u^2/2}\left[\frac{t\sqrt{d_3^{inp}} + (d_1^{inp}m - \vartheta)\sqrt{q}}{2\sqrt{d_3^{inp}}(1-q)\sqrt{q(1-q)}}\right] H(u)^{-1}\Bigg\} \end{aligned} \tag{44}$$

where we have used as before $\frac{d}{dq}H(u(q)) = \frac{d}{dq}\int_{u(t)}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} = -\frac{1}{\sqrt{2\pi}} e^{-\frac{u(q)^2}{2}} \frac{d}{dq}u(q)$ but as a function of q. Now the term multiplied by $(1-f)$ should be integrated by parts, i.e $\int_a^b u(x)v'(x) = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx$. Remembering that $Dt \equiv \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}$ one indeed can see that

$$\begin{aligned} \frac{d}{dt}\left[e^{-\frac{t^2}{2}}e^{-\frac{u^2}{2}}\right] &= -\left(t + u\sqrt{\frac{q}{1-q}}\right)\left(e^{-\frac{t^2}{2}}e^{-\frac{u^2}{2}}\right) = \\ -&\left(\frac{t\sqrt{d_3^{inp}} + (d_1^{inp}m - \vartheta)\sqrt{q}}{\sqrt{d_3^{inp}(1-q)}}\right)\left(e^{-\frac{t^2}{2}}e^{-\frac{u^2}{2}}\right) \end{aligned} \tag{45}$$

so one can re-write the term multiplied by $(1-f)$ in (44) as

$$\begin{aligned}
&\frac{(1-f)}{\sqrt{2\pi}}\int Dte^{-u^2/2}\left[\frac{t\sqrt{d_3^{inp}}+(d_1^{inp}m-\vartheta)\sqrt{q}}{2\sqrt{d_3^{inp}}(1-q)\sqrt{q(1-q)}}\right]H(u)^{-1}=\\
&=-\frac{(1-f)}{2\sqrt{2\pi}\sqrt{q(1-q)}}\int\frac{dt}{\sqrt{2\pi}}\frac{d}{dt}\left[e^{-\frac{t^2}{2}}e^{-\frac{u^2}{2}}\right]H(u)^{-1}=\\
&=-\frac{(1-f)}{2\sqrt{2\pi}\sqrt{q(1-q)}}\left\{\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}e^{-\frac{u^2}{2}}H(u)^{-1}\Big|_{t=-\infty}^{t=+\infty}+\right.\\
&\left.-\int Dte^{-\frac{u^2}{2}}\frac{d}{dt}H(u)^{-1}\right\}=\\
&=\frac{(1-f)}{2\sqrt{2\pi}\sqrt{q(1-q)}}\left\{-\int Dte^{-\frac{u^2}{2}}(-)H(u)^{-2}(-)\cdot\right.\\
&\left.\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\frac{\sqrt{qd_3^{inp}}}{\sqrt{d_3^{inp}(1-q)}}\right\}
\end{aligned}\tag{46}$$

where in the last passage we used again the Leibniz integral rule with the derivative in t. Substituting back Eq. (46) in the second term of Eq. (44) and canceling out the repeated terms enables to reach right away the simplified solution:

$$\begin{aligned}
\frac{dG}{dq}=0=\frac{\alpha}{q}\Bigg\{&f\Big[\frac{\langle(d_1^{inp}m-g^{-1}\eta^{\mu}-\vartheta)^2\rangle+qd_3^{inp}}{d_3^{inp}}\Big]+\\
&+\frac{(1-f)(1-q)}{2\pi}\int DtH(u)^{-2}e^{-u^2}\Bigg\}
\end{aligned}\tag{47}$$

where $\alpha\equiv p/C$ is the storage load.
As in Gardner [14] we take the limit $q\to1$, where the volume shrinks to a single point and it exist a unique configuration of weights satisfying the equations. In this limit, the storage load α becomes the critical capacity α_c. Note that in this limit:

$$\lim_{q\to1}u=\begin{cases}\infty & \text{if } t>\frac{\vartheta-d_1^{inp}m}{\sqrt{d_3^{inp}}}\\ -\infty & \text{if } t<\frac{\vartheta-d_1^{inp}m}{\sqrt{d_3^{inp}}}.\end{cases}\tag{48}$$

and

$$\begin{aligned}
&\lim_{u\to-\infty}H(u)\approx1\\
&\lim_{u\to\infty}H(u)\approx\frac{1}{\sqrt{2\pi}u}e^{-u^2/2}(1-\frac{1}{u^2})=\frac{1}{\sqrt{2\pi}u}e^{-u^2/2}
\end{aligned}$$

where in the second approximation we have Taylor expanded $H(u)$ around $u=0$.

This enables to further simplify the above equations, as one can define the variable

$$x = \frac{\vartheta - d_1^{inp} m}{\sqrt{d_3^{inp}}} \tag{49}$$

which can be used to divide the integral into two components, i.e.

$$\int DtH(u)^{-\kappa} e^{-\kappa \frac{u^2}{2}} = \cancelto{0}{\int_{-\infty}^{x} DtH(u)^{-\kappa} e^{-\kappa \frac{u^2}{2}}} + \int_{x}^{\infty} DtH(u)^{-\kappa} e^{-\kappa \frac{u^2}{2}} \tag{50}$$

where $\kappa = 1$ in Eq. (43) and $\kappa = 2$ in (47).

The simple application of the limit $q \to 1$ with the above approximations, substituting back u as in Eq. (40) and the new variable x as in Eq. (49) leads to the final set of equations for the storage capacity

$$\begin{cases} f(x + \frac{d_1^{out}}{g\sqrt{d_3^{inp}}}) = (1-f)\int_x^\infty Dt(t-x) \\ \frac{1}{\alpha_c} = f\Big[x^2 + \frac{d_2^{out}}{g^2 d_3^{inp}} + \frac{2x d_1^{out}}{g\sqrt{d_3^{inp}}} + 1\Big] + (1-f)\int_x^\infty Dt(t-x)^2. \end{cases} \tag{51}$$

where $d_{1,2,3}^{out}$ are defined in the same way as $d_{1,2,3}^{inp}$ except that the averages are now over the output distribution η.

Going from the calculation reported above for the threshold-linear perceptron it is straightforward to calculate the optimal capacity of a network of threshold linear units. Considering the network defined as in Eq. (4), the corresponding volume we need to calculate can be written as

$$V_T = \frac{\int \Big\{ \prod_{i,j,j\neq i} dJ_{ij} \delta\Big(\sum_{j,j\neq i} J_{ij}^2 - C\Big) \prod_{i,\mu}}{\int \prod_{i,j,j\neq i} dJ_{ij} \prod_i \delta\Big(\sum_{j,j\neq i} J_{ij}^2 - C\Big)} \cdot \tag{52}$$
$$\cdot \Big[\Big(1 - \delta_{\eta^\mu,0}\Big)\delta\Big(h_i^\mu - \vartheta - \frac{\eta^\mu}{g}\Big) + \delta_{\eta^\mu,0}\Theta\Big(\vartheta - h_i^\mu\Big)\Big]\Big\}$$

Since V_T can be written as the product of the individual volumes of the connection weights towards each unit, as $V_T = \prod_i^N V_i$ and thus $\langle \ln V_T \rangle_\eta = N \langle \ln V_i \rangle_\eta$, we will essentially be dealing with individual perceptrons like the one we have just studied. Putting $d_1^{inp} = d_1^{out} = d_1$ and $d_2^{inp} = d_2^{out} = d_2$ and thus $d_3^{inp} = d_3^{out} = d_3$ for $\forall i$, we arrive to Eq. (54).

We evaluate the maximal storage capacity in the limit $g \to \infty$, which

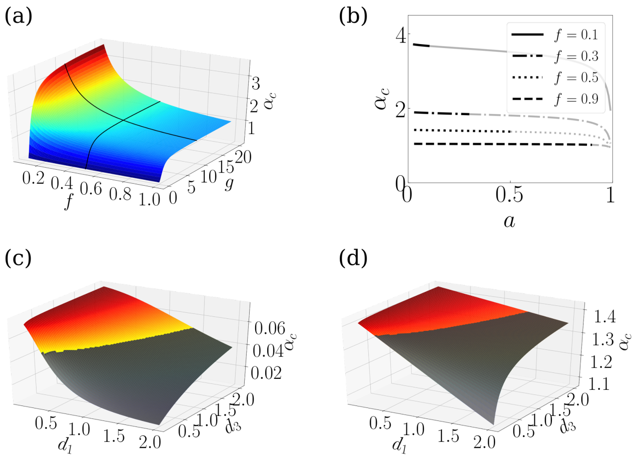


Figure 6: Dependence of the Gardner capacity α_c on different parameters: in (a) as a function of g and f ($d_1 = 1.1, d_2 = 2$), in (b) as a function of $a = d_1^2/d_2$ for different values of f ($g = 10$, $d_1 = 1.1$), in (c) and (d) as a function of d_1 and d_3 for $g = 0.2$ and $g = 10$, respectively ($f = 0.5$). Note that fixing f restricts the available range of a, as a cannot be larger than f; the inaccessible ranges are shadowed in (b-d).

is approached already for moderate values of g. Eq. (54) in the $g \to \infty$ limit reduces to:

$$\begin{cases} 0 = fx - (1-f)\int_x^\infty Dt(t-x) \\ \frac{1}{\alpha_c} = f(x^2+1) + (1-f)\int_x^\infty Dt(t-x)^2, \end{cases} \tag{53}$$

which provides the universal α_c^G bound for errorless retrieval, dependent only through f on the distribution of the patterns.

5.2.1 *Summary of the derivation*

Adapting the procedure introduced in [14] for binary units to our network, we evaluate the fractional volume of the space of the interactions J_{ij} which satisfy Eqs. (4)-(5), using the replica trick and the replica symmetry ansatz, we obtain the standard order parameters $m = \frac{1}{\sqrt{C}}\sum_j J_{ij}$ and $q = \frac{1}{C}\sum_j J_{ij}^a J_{ij}^b$ corresponding, respectively, to the average of the weights within each replica and to their overlap between two replicas a and b. Increasing p, for $C \to \infty$, shrinks the volume of the compatible weights, eventually to a single point, i.e., when there is only a unique solution and the storage capacity is reached. This corresponds to the case where all the replicated weights are equal $q \to 1$, implying that only one configuration satisfying all the equations exists. Adding a further memory pattern would make it impossible, in general, to satisfy them all. At the end, we obtain the following equations for α_c

$$\begin{aligned} 0 &= -f(x + \frac{d_1}{g\sqrt{d_3}}) + (1-f)\int_x^\infty Dt(t-x) \\ \frac{1}{\alpha_c} &= f\left[x^2 + \frac{d_2}{g^2 d_3} + \frac{2xd_1}{g\sqrt{d_3}} + 1\right] + (1-f)\int_x^\infty Dt(t-x)^2, \end{aligned} \tag{54}$$

where we have introduced the averages over $\Pr(\eta)$: $d_1 \equiv \langle \eta_i^\mu \rangle$, $d_2 \equiv \langle (\eta_i^\mu)^2 \rangle$ and $d_3 \equiv d_2 - d_1^2$; $x = (\vartheta - d_1 m)/\sqrt{d_3}$ is the normalized difference between the threshold and the mean input, while $f = Pr(\eta > 0)$ is the fraction of active units and $Dt \equiv dt \exp(-t^2/2)/\sqrt{2\pi}$. The two equations yield x and α_c. Both equations can be understood as averages over units, respectively of the actual input and of the square input, which determine the amount of quenched noise and hence the storage capacity. The capacity, α_c, then depends on the proportion f of active units, but also on the gain g, and on the cumulants d_1 and d_3. Fig. 6a shows that at fixed g, α_c increases as more and more units remain below threshold, ceasing to contribute to the quenched noise. In fact, α_c diverges as $[2f \ln(1/\sqrt{2\pi}f)]^{-1}$, for $f \to 0$; see Appendix B. At fixed f, there is an initially fast increase with g followed by a plateau dependence for larger values of g. One can show that $\alpha_c \to \frac{g^2}{g^2+1}$ as $f \to 1$, i.e., when all the units in the memory patterns are above threshold, it is always $\alpha_c < 1$ for any finite g. At first sight this may seem absurd: a linear system of N^2 independent equations and N^2 variables always has an inverse solution, which would lead to α_c being (at least) one. Similar to what was already noted in [77], however, the inverse solution does not generally satisfy the spherical constraint in Eq. (5); but it does, in our case, in the limit $g \to \infty$ and this can also be understood as the reason why α_c is highest when g is very large. In practice, Fig. 6 indicates that over a broad range of f values, α_c approaches its $g \to \infty$ limit already for moderate values of g; while the dependence on d_1 and d_3 is only noticeable for small

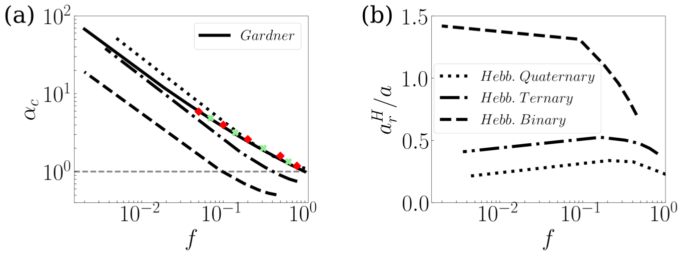


Figure 7: Hebbian vs Gardner capacity. (a) α_c^H vs. f for different sample distribution of stored patterns compared to the analytically calculated universal α_c^G; the red diamonds and green crosses are reached using perceptron training for binary and ternary patterns, respectively. (b) the sparsification of the stored patterns at retrieval, for Hebbian networks at their capacity.

g, as can be seen by comparing Fig. 6c and d. For $g \to \infty$, one sees that Eqs. (6) depend on $\Pr(\eta)$ only through f. Eqs. (54), at $g \to \infty$, have been verified by explicitly training a threshold linear perceptron with binary patterns, evaluating α_c numerically as the maximal load which can be retrieved with no errors; See Appendix C for details. Estimated values of α_c are depicted by red diamonds in Fig. 14, and they follow the profile of the solid line describing the $g \to \infty$ limit of Eq. (54).

6 COMPARISON WITH A HEBBIAN RULE: THEORETICAL ANALYSIS

6.1 RECAP OF THE DERIVATION OF THE HEBBIAN CAPACITY IN THRESHOLD-LINEAR NETWORKS

In this initial section we provide a brief recap of the main ideas, analytical tools and results reported in [53, 54, 57] about the storage capacity of networks endowed with threshold-linear units. In the most general case, one considers that the threshold-linear unit i receives an input

$$h_i = \sum_j J^c_{ij} V_j + b\left(\sum_j V_j/N\right) + \sum_\mu s^\mu \frac{\eta_i^\mu}{\langle\eta\rangle_\eta} \tag{55}$$

where the first term is the standard term coming from the activity of the other units through the synaptic weights J^c. The second term is supposed to provide a general feedback, perhaps through inhibitory neurons that are not explicitly modelled, and it only depends on the mean network activity via a function b. The last term is the strength of the input aligned with one or more stored patterns. To study self-sustained attractors, we set s^μ to zero (which implies also $\delta = 0$ in the notation of [53]). The unit activities V_i are subject to the threshold-linear activation function, and the weight matrix is structured in *one-shot* by Hebbian learning. In a general case, the Hebbian learning rule can be defined in terms of the firing rates η of the units as

$$J_{ij} = c_{ij}\frac{1}{C}\sum_{\mu=1}^{p} F(\eta_i^\mu)G(\eta_j^\mu) \tag{56}$$

where J_{ij} is the synaptic efficacy/connection strengths of the inputs coming from neuron j to neuron i; $F(\eta_i^\mu)$ and $G(\eta_j^\mu)$ are generic functions describing how the learning rule depends on the activity of the postsynaptic neuron i and the presynaptic neuron j respectively, in pattern μ; c_{ij}, instead, defines whether a synapse is present or not.
The presynaptic component can be defined in order to have mean $\langle G(\eta)\rangle_\eta = 0$ (necessary to allow to store a number of patterns which increases with the connectivity C). One can further assume that the postsynaptic component is the same as the presynaptic one [54], as:

$$G(\eta) = F(\eta) = \frac{\eta - \langle\eta\rangle}{\langle\eta\rangle} \tag{57}$$

such that, if $c_{ij} = c_{ij}$ the system can be described in terms of an energy function. Assuming Eq. (57) the mean a_F and the variance c_F

of the postynaptic factors coincide with the ones of the presynaptic component, being $a_F = 0$ and the variance

$$T_0 \equiv c_F \equiv \langle G(\eta)^2 \rangle_\eta - \langle G(\eta) \rangle_\eta^2 = \frac{\langle \eta^2 \rangle - \langle \eta \rangle^2}{\langle \eta \rangle^2} = \frac{1-a}{a} \tag{58}$$

where a is the sparseness of the code defined in Eq. (75).
The postsynaptic function $F(\eta)$ can also have different forms; in [54], for example, the case of NMDA-resembling postynaptic activity is addressed. Here we will report the analytical results derived in [53, 54, 57] for $F(\eta)$ as defined in (57), which leads to a covariance rule for expressing Hebbian learning, leading from Eq. (56) to:

$$J_{ij}^C = c_{ij} \frac{1}{C \langle \eta \rangle_\eta^2} \sum_{\mu=1}^{p} (\eta_i^\mu - \langle \eta \rangle_\eta)(\eta_j^\mu - \langle \eta \rangle_\eta). \tag{59}$$

were C is the number of connections per unit and $c_{ij} = 1$ if there is a connections from unit j to unit i, $c_{ij} = 0$ otherwise.

Calculations for the storage capacity were performed for different types of network: the fully connected [57], the highly diluted [53], the directed [54] ones, and then generalised to arbitrary dilution in [60]. Here we focus on networks with extremely diluted connectivity [53], namely when $\frac{C}{N} \to 0$ and $C, N \to \infty$ such that the synapses J_{ij} and J_{ji} can be considered independent, and also on fully connected networks, useful to grasp a deeper understanding of the error at retrieval and the nature of Hebbian learning.

The calculation of the storage capacity involves the definition of the overlap order parameters

$$\hat{x}^\mu = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\eta_i^\mu}{\langle \eta \rangle_\eta} - 1 \right) \langle V_i \rangle \tag{60}$$

measuring the overlap between the stored patterns η and the activity of units V_i, where $\langle \cdots \rangle$ (without subscripts) denotes thermal average. One assume without loss of generality that one of the patterns, let us say the first pattern, is to be retrieved, and one then assumes the existence of stable states of the system for which $\hat{x}^1$ is non-zero while $\hat{x}^{\mu_s}, \mu \neq 1$ are zero in the thermodynamic limit. In particular, comparing the overlap with the average retrieved activity

$$x(t) = \frac{1}{N} \sum_{i=1}^{N} \langle V_i(t) \rangle \tag{61}$$

once the dynamics reaches a fixed point, one would have $x^1 \gg x$. One should note that the specific quantity which emerges from the calculations is the subtracted overlap $\hat{x}^1 = x^1 - x$ i.e.

$$\hat{x}^1 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\eta_i^1}{\langle \eta \rangle_\eta} - 1 \right) \langle V_i \rangle \tag{62}$$

This is done in mean-field theory, developed either by means of the replica trick or signal-to-noise analyses, which yield self-consistent equations for the overlaps and other order parameters that appear. The mean field theory is based on four main assumptions:

1. the thermodynamic limit $C \to \infty$, with $\frac{C}{N} \to 0$;
2. the storage capacity is an extensive quantity where $p \to \infty$ and $\alpha \equiv \frac{p-1}{C}$ is finite;
3. only one single pattern has non-zero correlation;
4. the evolution of the overlap $\hat{x}^{\mu}(t)$, of the mean activity $x(t)$ and of the mean square activity

$$y(t) = \frac{1}{N}\sum_{i=1}^{N} \langle V_i(t)^2 \rangle \tag{63}$$

   reach the fixed points $\hat{x}^{\mu}$, x and y.

An important order parameter that appears in the mean-field theory of attractor neural networks is the variance of the quenched noise [69]: it comes from the contribution to the field acting on each unit from the correlation of the activity of the network and that of non-retrieved patterns, i.e. those different from the first pattern. This correlation albeit small for each individual non-retrieved pattern gives a significant contribution when p is comparable to C and is what makes retrieval impossible for large p. It thus needs to be included for calculating the storage capacity. Following [53, 57], we denote this parameter as ρ. In the case of threshold-linear units two other order parameters are important that measure the relative magnitude of the signal (the part of the input to units that makes the units have the correct activity for retrieval) to the quenched noise ρ. The first one

$$w = \frac{-\hat{x}^1 - \vartheta}{T_0 \rho} \tag{64}$$

is the signal of the background versus the noise due to memory loading; ϑ is the threshold (Eq. (4)), $b(x)$ is a general function, depending on the average activity and which would contribute as well to (4) and where ρ is the noise due to memory loading, deriving from all other patterns than the retrieved one. The second signal to noise

$$v = \frac{\hat{x}^1 + s^1}{T_0 \rho} \tag{65}$$

is specific to the units that have to be active. Here we consider for simplicity the pattern specific external stimulus $s^1 = 0$.

The self-consistent mean-field equations that emerge from the calculations can be written in terms of the following quantities

$$\begin{aligned}
A_1(w,v) &= \frac{1}{vT_0}\Big\langle\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\int^{+} Dz\Big[w+v(1+F(\eta))-z\Big]\Big\rangle_{\eta}\\
A_2(w,v) &= \frac{1}{vT_0}\Big\langle\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\int^{+} Dz\Big[w+v(1+F(\eta))-z\Big]\Big\rangle_{\eta}\\
A_3(w,v) &= \Big\langle\int^{+} Dz\Big[w+v(1+F(\eta))-z\Big]^2\Big\rangle_{\eta}\\
A_4(w,v) &= \frac{1}{v}\Big\langle\int^{+} Dz\Big[w+v(1+F(\eta))-z\Big]\Big\rangle_{\eta}
\end{aligned} \tag{66}$$

where the average is over the distribution P_η and over a Gaussian variable z up to a threshold, such that:

$$\int^{+} Dz() = \int_{-\infty}^{w+v(1+F(\eta))} \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,() \tag{67}$$

and that, by simple substitution of T_0 can be rewritten as

$$A_1(w,v) = \tfrac{a}{v(1-a)}\Big\langle\Big(\tfrac{\eta}{\langle\eta\rangle}-1\Big)(x\phi(x)+\sigma(x))\Big\rangle_{\eta} - \langle\phi(x)\rangle_{\eta} \tag{68}$$

$$A_2(w,v) = \tfrac{a}{v(1-a)}\Big\langle\Big(\tfrac{\eta}{\langle\eta\rangle}-1\Big)(x\phi(x)+\sigma(x))\Big\rangle_{\eta} \tag{69}$$

$$A_3(w,v) = \Big\langle(x^2+1)\phi(x)+x\sigma(x)\Big\rangle_{\eta} \tag{70}$$

$$A_4(w,v) = \tfrac{1}{v}\Big\langle(x\phi(x)+\sigma(x))\Big\rangle_{\eta} \tag{71}$$

where

$$x \equiv w + v\tfrac{\eta}{\langle\eta\rangle} \tag{72}$$

$$\phi(x) \equiv \frac{[1+\mathrm{erf}(\frac{x}{\sqrt{2}})]}{2} = \frac{\mathrm{erfc}(\frac{-x}{\sqrt{2}})}{2} \tag{73}$$

$$\sigma(x) \equiv \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{74}$$

and where the sparsity parameter, a, defined as

$$a \equiv \frac{\langle\eta^2\rangle}{\langle\eta\rangle_{\eta}^2}, \tag{75}$$

shows up as a crucial quantity.

As far as the calculation of capacity is concerned, for the fully connected network, these equations must satisfy the conditions

$$E_1^{fc}(w,v) = 0 = A_1(w,v)^2 - \alpha A_3(w,v) \tag{76}$$

$$E_2^{fc}(w,v) = 0 = A_1(w,v)\Big(\tfrac{1}{gT_0} - A_2(w,v)\Big) - \alpha A_2(w,v) \tag{77}$$

and for the highly diluted network

$$E_1^{hd}(w,v) = 0 = A_2(w,v)^2 - \lambda\alpha A_3(w,v) \tag{78}$$

$$E_2^{hd}(w,v) = 0 = A_2(w,v) - \tfrac{1}{gT_0}. \tag{79}$$

where $\lambda = \frac{c_F + a_F^2}{T_0} = 1$ given Eq. (57). In other words, the storage capacity α_c can be computed by finding the largest value of α for which equation E_1 can be satisfied, while equation E_2 can be used to extract the optimal value of g.

The value of other order parameters, e.g. ρ, or $\hat{x}^1$ for each value of α and any given choice of the distributions of η can also be calculated. In order to extract the overlap one can find ρ (from Eq.s (28) of [57] in a fully connected net and Eq.s (13) of [53] in an highly diluted one) as

$$\rho = xA_2/vA_4 \tag{80}$$

and

$$\hat{x}^1 = T_0 \rho v \tag{81}$$

inverting Eq. (65).

6.2 RETRIEVAL TO NO-RETRIEVAL TRANSITION IN THRESHOLD-LINEAR HOPFIELD NETWORKS

For the purpose of comparing, in the threshold-linear domain, the *error-less* storage capacity (evaluated in Sect. 5.2) with the *error-full* Hebbian capacity (summarized in Sect. 6.1), and to properly relate this comparison with respective one in the binary domain, one needs to know how the overlap approaches zero in threshold-linear networks endowed with Hebbian learning. For this purpose, in this chapter, we initially take as an example the binary distribution, defined as

$$p(\eta) = (1-a)\delta(\eta) + a\delta(1-\eta) \tag{82}$$

in order to show how the indirect measure of the overlap v varies at the storage capacity with the sparsity and the type of network, then we provide an analytical derivation of the general constraint required to have first or second order phase transition in v at the storage capacity. Finally we take again the binary distribution at two sparsity values and show how the proper overlap $\hat{x}^1$ approaches its value when having a first or a second order phase transition.

In Fig. 8 one can appreciate how the values of (α_c, v_c, w_c) vary as a function of a for the highly diluted and the fully connected networks for a binary distribution of patterns. One can notice that if for the feedback networks the indirect measure of the overlap at the storage capacity v_c never goes to zero, there is instead a sparsity value for which this occurs in the highly diluted limit, after which the antipattern $(v < 0)$ is retrieved .

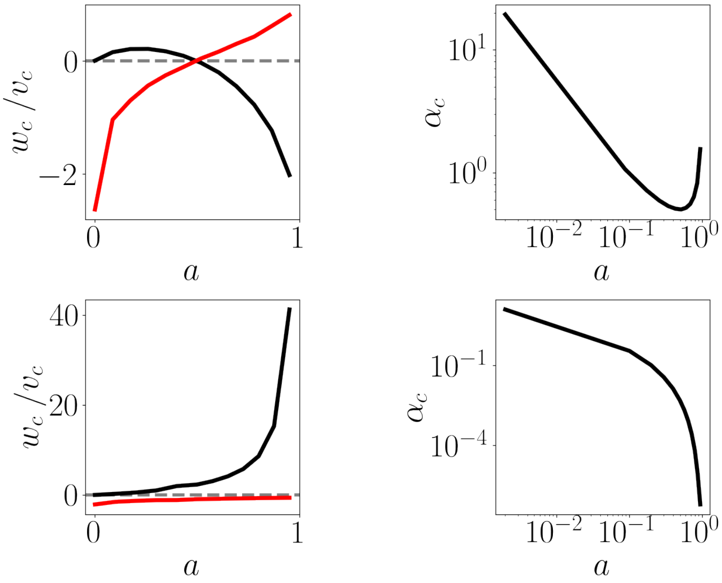


Figure 8: Parameters at the storage capacity satisfying Eq. (76)(77) (fully connected) and (78)(79) (highly diluted) for a binary distribution. First row: Highly diluted networks, second row: fully connected network. First column: red= w_c, black= v_c.

Order of the phase transition for $\nu \to 0$*:*

In the extremely diluted limit one can then calculate what are the conditions on the distribution in order to have a second order phase transition in the overlap, i.e. the limit $\nu \to 0$ of $A_2(w,\nu)$ and $A_3(w,\nu)$.

To do so we expand around $x_0 = w$, given that, as defined in Sect. 6.1, $x = w + \nu\frac{\eta}{\langle\eta\rangle}$ and here $\nu \to 0$. In this expansion, three equations should hold, namely:

$$\begin{cases} E_1(\nu,w,\alpha) = A_2(w,\nu)^2 - \alpha A_3(w,\nu) = 0 \\ \frac{d}{dw}E_1(\nu,w,\alpha) = 2A_2(w,\nu)\frac{d}{dw}A_2(w,\nu) + \frac{d}{dw}A_3(w,\nu) = 0 \\ \frac{d}{d\nu}E_1(\nu,w,\alpha) = 2A_2(w,\nu)\frac{d}{d\nu}A_2(w,\nu) + \frac{d}{d\nu}A_3(w,\nu) = 0 \end{cases} \tag{83}$$

The first one, as presented in Sect. 6.1, corresponds to the equation of the storage capacity, the second and third ones, instead, are related to the geometry of the solution. In the (w,ν) plane, indeed, one finds that the solutions lie on an island which progressively shrinks while increasing the storage load. At the storage capacity the solution is an individual maximum in the (w,ν) plane, thus both derivatives needs to be zero.

We now perform the expansions and derivatives, where we use the following identities which one can show:

$$\frac{d}{dx}\sigma(x) = -x\sigma(x) \tag{84}$$

$$\frac{d}{dx}\varphi(x) = \sigma(x) \tag{85}$$

$$\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big) = 0 \tag{86}$$

$$\int d\eta P(\eta)\frac{\eta}{\langle\eta\rangle}\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big) = T_0 \tag{87}$$

$$\frac{d}{dx}(x\varphi(x) + \sigma(x)) = \varphi(x) \tag{88}$$

$$\frac{d}{dx}[(x^2+1)\varphi(x) + x\sigma(x)] = 2(x\varphi(x) + \sigma(x)) \tag{89}$$

The Taylor expansion of A_2 is

$$A_2(w,v) \approx \frac{1}{vT_0}\cancelto{0}{(w\varphi(w) + \sigma(w))\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)} \quad + \\ + \frac{1}{\cancel{v}T_0}\cancel{v}\varphi(w)\cancelto{T_0}{\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)\frac{\eta}{\langle\eta\rangle}}$$

So:

$$A_2(w,v) = \varphi(w) + \mathcal{O}(v^2) \tag{90}$$

Its derivatives are

$$\frac{d}{dw}A_2(w,v) = \frac{1}{vT_0}\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)\varphi(x) \tag{91}$$

$$\frac{d}{dv}A_2(w,v) = \frac{1}{vT_0}\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)\varphi(x)\frac{\eta}{\langle\eta\rangle} - \frac{1}{v}A_2 \tag{92}$$

which, if expanded around $x_0 = w$ give:

$$\frac{d}{dw}A_2(w,v) \approx \frac{1}{vT_0}\cancelto{0}{\varphi(w)\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)} \quad + \\ + \frac{1}{\cancel{v}T_0}\cancel{v}\sigma(w)\cancelto{T_0}{\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle} - 1\Big)\frac{\eta}{\langle\eta\rangle}}$$

so:

$$\frac{d}{dw}A_2(w,v) = \sigma(w) + \mathcal{O}(v^2) \tag{93}$$

and

$$\begin{aligned}\frac{d}{dv}A_2(w,v) \approx{}& \cancel{\frac{1}{vT_0}\phi(w)\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\frac{\eta}{\langle\eta\rangle}}+\\ &+\frac{1}{\cancel{v}T_0}\cancel{v}\sigma(w)\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\frac{\eta}{\langle\eta\rangle}\frac{\eta}{\langle\eta\rangle}+\\ &-\cancelto{0}{\frac{1}{v^2T_0}(w\phi(w)+\sigma(w))\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)}+\\ &\cancel{-\frac{1}{v^{\cancel{2}}T_0}\cancel{v}\phi(w)\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\frac{\eta}{\langle\eta\rangle}}+\\ &-\frac{1}{2\cancel{v^2}T_0}\cancel{v^2}\sigma(w)\int d\eta P(\eta)\Big(\frac{\eta}{\langle\eta\rangle}-1\Big)\frac{\eta}{\langle\eta\rangle}\frac{\eta}{\langle\eta\rangle}\end{aligned}$$

so:

$$\begin{aligned}\frac{d}{dv}A_2(w,v) &= \sigma(w)T_1+\mathcal{O}(v^3)\\ T_1 &= \frac{1}{2}\frac{\langle\eta^3\rangle-\langle\eta\rangle\langle\eta^2\rangle}{\langle\eta\rangle(\langle\eta^2\rangle-\langle\eta\rangle^2)}\end{aligned} \tag{94}$$

The Taylor expansion of A_3 is

$$\begin{aligned}&A_3(w,v)\approx[(w^2+1)\phi(w)+w\sigma(w)]\cancelto{1}{\int d\eta P(\eta)}+\\ &2(w\phi(w)+\sigma(w))v\cancelto{1}{\int d\eta P(\eta)\frac{\eta}{\langle\eta\rangle}}\end{aligned}$$

so

$$A_3(w,v)=(w^2+1)\phi(w)+w\sigma(w)+\mathcal{O}(v) \tag{95}$$

Its derivatives are

$$\frac{d}{dw}A_3(w,v)=2\int d\eta P(\eta)(x\phi(x)+\sigma(x)) \tag{96}$$

$$\frac{d}{dv}A_3(w,v)=2\int d\eta P(\eta)(x\phi(x)+\sigma(x))\frac{\eta}{\langle\eta\rangle} \tag{97}$$

which, if expanded around $x_0=w$, both result in:

$$\frac{d}{dw}A_3(w,v)=2[w\phi(w)+\sigma(w)]+\mathcal{O}(v) \tag{98}$$

$$\frac{d}{dV}A_3(w,v)=2[w\phi(w)+\sigma(w)]+\mathcal{O}(v) \tag{99}$$

Therefore, the three equations which should be satisfied are:

$$\begin{cases}\phi(w)^2-\alpha[(w^2+1)\phi(w)+w\sigma(w)]=0\\ \phi(w)\sigma(w)-\alpha[w\phi(w)+\sigma(w)]=0\\ \phi(w)\sigma(w)T_1-\alpha[w\phi(w)+\sigma(w)]=0\end{cases} \tag{100}$$

from the last two equations we get

$$T_1 = 1 \tag{101}$$

and, from the first two

$$\frac{\phi(w)^2}{[(w^2+1)\phi(w)+w\sigma(w)]} = \frac{\phi(w)\sigma(w)}{[w\phi(w)+\sigma(w)]} \tag{102}$$

Left and right hand side of Eq. (102) are plotted in Fig. 9. One can see

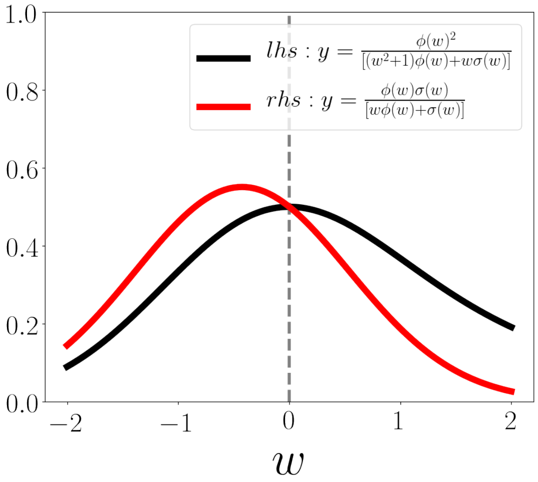


Figure 9: Plot of left and right hand side of Eq. (102)

that there is a unique w value which satisfy the above equations by developing Eq. (102) to

$$w[\phi(w)^2 - \sigma(w)^2 - w\sigma(w)\phi(w)] = 0 \tag{103}$$

where the term within the brackets never goes to zero, as shown in Fig. 10

In conclusion for $\nu \to 0$, the phase transition is second order only for those distributions which enable $T_1 \equiv 1$, and, when this is true, then $w* = 0$ and $\alpha_c = \frac{1}{2}$.

One can plot what is the required relation between the cumulants of a distribution in order to have a second order phase transition by rewriting $T_1 = 1$ as

$$\begin{gathered}\frac{1}{2}\frac{\langle\eta^3\rangle - \langle\eta\rangle\langle\eta^2\rangle}{\langle\eta\rangle(\langle\eta^2\rangle - \langle\eta\rangle^2)} = 1 \\ \langle\eta^3\rangle - 3\langle\eta\rangle\langle\eta^2\rangle + 2\langle\eta\rangle^3 = 0 \\ \frac{\langle\eta^3\rangle}{\langle\eta\rangle^3} - 3\frac{\langle\eta^2\rangle}{\langle\eta\rangle^2} + 2 = 0\end{gathered} \tag{104}$$

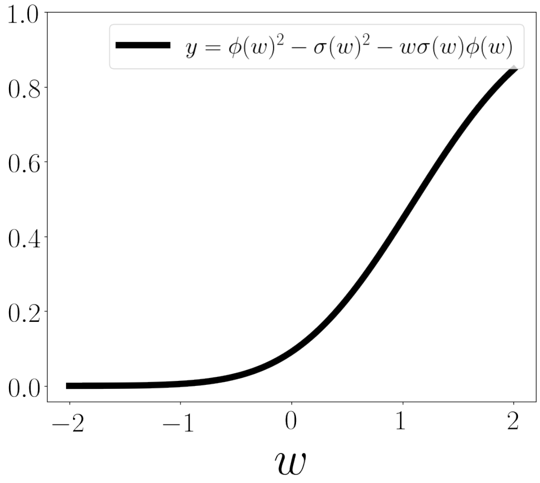


Figure 10: Plot of the term within the brackets of Eq. (103)

where in the second and third passages the condition of existence is $\langle\eta^2\rangle \neq \langle\eta\rangle^2$ and $\langle\eta\rangle \neq 0$, yielding to the requirement that:

$$\frac{\langle\eta\rangle^3}{\langle\eta^3\rangle} = \frac{a}{3-2a} \tag{105}$$

Here one can see that for the binary distribution, having $\frac{\langle\eta\rangle^3}{\langle\eta^3\rangle} = a^2$ the intersection occurs at $a = \frac{1}{2}$. In Fig. 11 one can see such intersection in the $\frac{\langle\eta\rangle^3}{\langle\eta^3\rangle} - a$ plane.

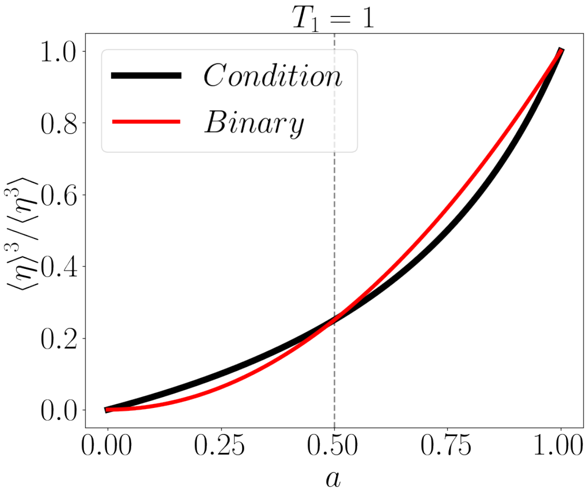


Figure 11: In black we plot the condition, defined as in E. (105) in order to have a second order phase transition. In red we plot the relation between $\frac{\langle\eta\rangle^3}{\langle\eta^3\rangle}$ and a for a generic binary distribution, identifying the intersection point with a gray vertical line.

$\hat{x}^1$ *as a function of* α:

Back to the binary example we can thus show numerically how ν and $\hat{x}^1 = T_0 \rho \nu$, vary while the system approaches the storage capacity. In the binary distribution $T_1 = \frac{1}{2a}$ so, in order to fulfill the requirement $T_1 = 1$ the sparsity has to be $a = \frac{1}{2}$, as mentioned in the previous section. Indeed, as already shown in Fig. 8, at $a = 0.5$ ν_c goes to zero. In Fig. 12 we show how ν and w, both for the stable and unstable intersection points, vary while approaching α_c for $a = 0.5$ and $a = 0.2$ for the highly diluted and the fully connected networks. To do so, we need to fix g at the storage capacity. We found, by numerically solving the equations, that for the fully connected network the required triads (a, α_c, g_c) are $(0.2, 0.105, 0.032)$ and $(0.5, 0.0049, 1.253)$; instead for the highly diluted they are $(0.2, 0.667, 0.525)$ and $(0.5, 0.5, 2)$. The behavior of ν, w, α is plotted in Fig. 12 while the one of $\hat{x}^1$ in Fig. 13. One has to notice that if $\hat{x}^1$ is the subtracted overlap defined as $\hat{x}^1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\eta_i^1}{\langle\eta\rangle_\eta} - 1\right)\langle V_i\rangle$, then $(\hat{x}^1)^{MAX} = 1 - a$ as visible also in the plots. In Fig. 13 one can appreciate that the full-connectivity not only decreases the storage capacity but it makes asymmetric the behavior of the stable vs unstable $\hat{x}$, the latter one approaching 0 for $\alpha \to 0$.

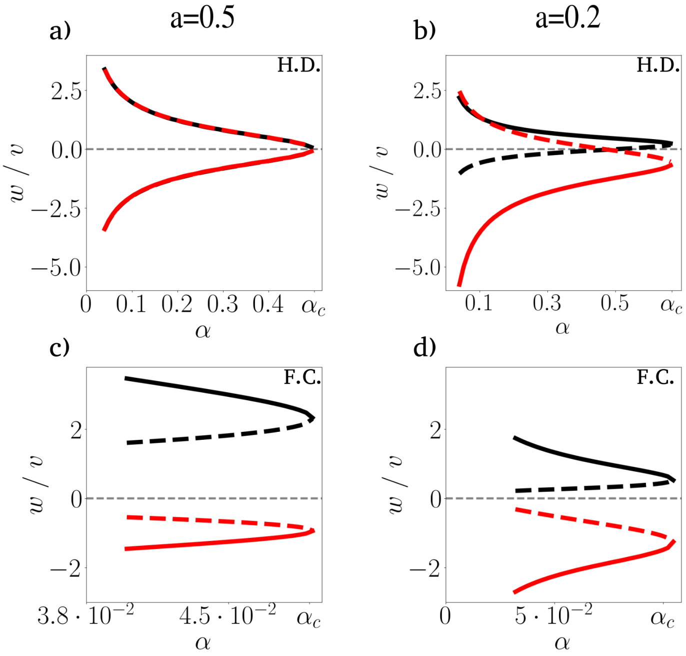


Figure 12: Values of ν (black) and w (red) satisfying a)b) Eq. (78)(79) (highly diluted -H.D.-) and c)d) (76)(77) (fully connected -F.C.-) while approaching the storage capacity α_c for sparsity a)c) $a = 0.5$ and b)d) $a = 0.2$. The stable solutions are depicted with a solid line, the unstable ones with a dashed line.

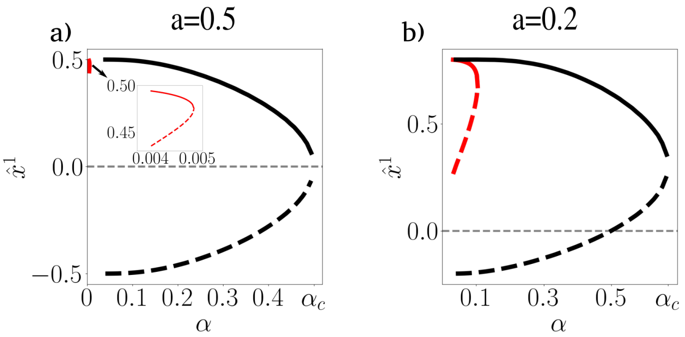


Figure 13: Decay of the overlap with the stored memory, as evaluated in Eq. (81), in a fully connected (red) and highly diluted (black) networks, for two sparsity parameters (a)a=0.5, b)a=0.2)). The inset in a) correspond to a restricted area of the plot. The solid line correspond to the stable solutions for (w, v), the dashed lines for the unstable ones.

6.3 SELF-ORGANIZED HEBBIAN LEARNING IS MORE EFFICIENT THAN OPTIMAL PROCEDURES

With highly diluted connectivity and non-sparse patterns a binary network can get to $1/\pi$ of the bound, even if with vanishing overlaps, much closer than in the fully connected case. This is intuitively because the quenched noise is diminished as J_{ij} and J_{ji} become effectively independent. Besides its biological relevance, with TL units, the fair comparison to the capacity *à la Gardner* is thus that of a Hebbian network with *highly diluted* connectivity. In what follows, we indicate the Gardner capacity as calculated in the previous section and the Hebbian capacity, by α_c^G and α^H, respectively, and use similar superscript notations for other quantities.

The capacity of the TL network with diluted connectivity was evaluated analytically in [53, 54]; see Sect. 6.1 for a recap. Whereas for $g \to \infty$ the Gardner capacity depends on $\Pr(\eta)$ only via f, for Hebbian networks it does depend on the distribution, and most importantly on a, the *sparsity*

$$a = \langle \eta_i^\mu \rangle^2 / \langle (\eta_i^\mu)^2 \rangle \tag{106}$$

whose relation to f depends on the distribution [53, 54].

Fig. 14 shows the results for 3 examples of binary, ternary and quaternary distributions for which f and a are related through $f = a$, $9a/5$ and $9a/4$, respectively, see Appendix D.1; the Hebbian and the Gardner capacities diverge in the sparse coding limit.

When attention is restricted to binary patterns in Fig. 14a, the Gardner capacity, α_c^G, *seems* to provide an upper bound to the capacity reached with Hebbian learning; more structured distributions of activity, however, dispel such a false impression: the quaternary example already shows higher capacity for sufficiently sparse patterns.

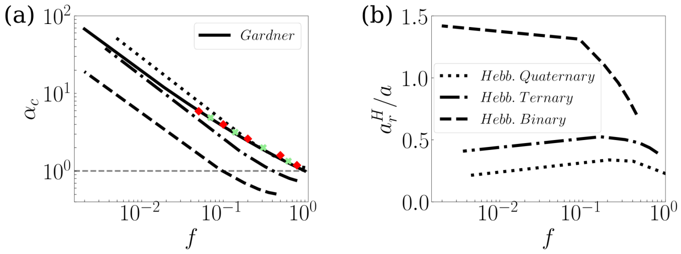


Figure 14: Hebbian vs Gardner capacity. (a) α_c^H vs. f for different sample distribution of stored patterns compared to the analytically calculated universal α_c^G; the red diamonds and green crosses are reached using perceptron training for binary and ternary patterns, respectively. (b) the sparsification of the stored patterns at retrieval, for Hebbian networks at their capacity.

The bound, in fact, would only apply to perfect errorless retrieval, whereas Hebbian learning creates attractors which are, up to the Hebbian capacity limit, correlated but not identical to the stored patterns; in particular, we notice that when considering TL units and Hebbian learning, in order to reach close to the capacity limit, the threshold has to be such as to produce sparser patterns at retrieval, in which only the units with the strongest inputs get activated. Fig. 14b shows the ratio of the sparsity of the retrieved pattern produced by Hebbian learning, $a_r^H = \langle v_i^\mu \rangle^2 / \langle (v_i^\mu)^2 \rangle$ (estimated as described in Appendix D.4) to that of the stored pattern a, vs. f: except for the binary patterns at low f, the retrieved patterns, at the storage capacity, are always sparser than the stored ones. The largest sparsification happens for quaternary patterns, for which the Hebbian capacity overtakes the Gardner bound, at low f. Sparser patterns emerge as, to reach close to α_c^H, ϑ has to be such as to inactivate most of the units with intermediate activity levels in the stored pattern. Of course, the perspective is different if α_c^H is considered as a function of a_r instead of a, in which case the Gardner capacity remains unchanged, as it implies retrieval with $a_r = a$, and is above α_c^H for each of the 3 sample distributions; see Fig. 31 of Appendix D.1.

7 COMPARISON WITH A HEBBIAN RULE: EXPERIMENTAL DATA

Having established that the Hebbian capacity of TL networks can surpass the Gardner bound for some distributions, we ask what would happen with distributions of firing rates naturally occurring in the brain. We considered published distributions of single neurons in IT cortex in response to short naturalistic movies [78]. Such distributions can be taken as examples of patterns elicited by visual stimuli, to be stored with Hebbian learning, given appropriate conditions, and later retrieved using attractor dynamics, triggered by a partial cue [79–84]. How many such patterns can be stored, and with what accompanying sparsification?

7.1 ESTIMATION OF THE BINNED RETRIEVED DISTRIBUTIONS

We considered sample distributions from [78], where each neuron emits, in time bins of fixed duration (100msec), $0, \dots, n, \dots, n_{max}$ spikes, with relative frequency c_n, such that $\sum_{n=0}^{n_{max}} c_n = 1$. We take these values from Fig. 2 of [78] and they correspond to the blue histograms in Fig. 15 below (and in Fig. 33 in Appendix E). We assume they are the distributions of the patterns to be stored. If the weights are those described by the Gardner calculation, these patterns can be retrieved as they are, and their distribution remains the same. If they are stored with Hebbian weights close to the maximal Hebbian capacity, however, the retrieved distributions look different. In this section we derive the retrieved distribution given a stored one.

The firing rate V of a neuron in retrieving a stored pattern η is assumed proportional to $w + v\eta/\langle\eta\rangle + z$ [54], where the parameters w and v are appropriately rescaled signal-to-noise ratios (general and pattern-specific), such that the normally distributed random variable z, of zero mean and unitary variance, is taken to describe all other non constant (noise) terms, besides η itself. Averaging over z one can write, as in Eq. (167a), that at the maximal capacity

$$\langle V \rangle (\eta) = g \int_{-x(\eta)}^{\infty} Dz \, [x(\eta) + z] = g \, [x_c \phi(x_c) + \sigma(x_c)] \,, \qquad (107)$$

where $x(\eta) \equiv w + v\eta/\langle\eta\rangle$ and at the saddle-point the parameters w and v take the values w_c and v_c that maximize capacity, as explained in [54]. This implies setting an optimal value for the threshold ϑ, which in the analysis is absorbed into the parameter w, and which determines the sparsity of the retrieved distribution. The gain g remains, however, a free parameter, that affects neither sparsity nor capacity. It is a rescaled version of the original gain g in the hypothetical TL transfer function. In other words, the maximal Hebbian capacity determines the shape of the retrieval activity distribution, but not its scale (e.g., in spikes per sec).

To produce a histogram, that details the frequency with which the neuron would produce n spikes at retrieval, e.g. again in bins of 100msec, one has to set this undetermined scale. We set it arbitrarily, with the rough requirement that the frequency of producing n_{max} spikes at retrieval be below what it is in the observed distribution, taken to describe storage, and negligible for $n_{max}+1$ spikes. Having set the scale g, the frequency with which the neuron emits n spikes at retrieval, with $0 < n < n_{max}$ is the probability that $n - 1/2 < V < n + 1/2$, that is, it is a sum over contributions from each η, such that

$$\begin{aligned} n-\frac{1}{2} < g(w_c + v_c\frac{\eta}{\langle\eta\rangle} + z) < n+\frac{1}{2} \\ \frac{n}{g} - \frac{1}{2g} - x_c < z < \frac{n}{g} + \frac{1}{2g} - x_c \end{aligned} \tag{108}$$

i.e.,

$$\Pr(n) = \sum_{\eta=0}^{\eta_{max}} c_\eta \Big[\varphi\Big(\frac{n}{g}+\frac{1}{2g}-x_c\Big) - \varphi\Big(\frac{n}{g}-\frac{1}{2g}-x_c\Big)\Big], \tag{109}$$

with appropriate expressions for the two extreme bins. These are the distributions shown in Fig. 15 in chapter 7, and in Fig.2 below. We took $g = \frac{1}{2}$, as this value satisfies the *a priori* requirements and allows to keep the same number of bins in the retrieved memory as in the stored one (and the coefficients sum up to one, to a very good approximation).

7.2 THE MOST ACTIVE UNITS REMAIN ACTIVE

Fig. 15a-b show the analysis of two sample distributions from [78]. The observed distributions, in blue, labeled "Gardner", are those we assume could be stored and retrieved, exactly as they were, with a suitable training procedure bound by the Gardner capacity. In orange, we plot the distribution that would be retrieved following Hebbian learning operating at its capacity, see Sect. 7.1 for the estimation of the retrieved distribution. Note that the absolute scale of the retrieved firing rate is arbitrary, what is fixed is only the shape of the distribution, which is sparser (as clear already from the higher bar at zero). The pattern in Fig. 15a, which has $a < 0.5$, could also be fitted with an exponential distribution having $f = 2a$ (see Appendix D.2). In that panel we also show the values of $\alpha_c^{H_{exp}}$ and $a_r^{H_{exp}}$, calculated assuming the exponential fit, along with values from the observed discrete distribution ($\alpha_c^{H_{naive}}$ and $a_r^{H_{naive}}$). Fig. 15c shows both α_c^G and $\alpha_c^{H_{exp}}$ versus f; we have indicated by diamonds the Hebbian capacities for the 9 empirical distributions in [78] and by circles the fitted values for those which could be fitted to an exponential. In Appendix D.3 we also discuss the fit to a log-normal, which is better at reproducing experimental distributions with a mode above zero [85], as in Fig. 15b. There are three conclusions that we can draw from these data. First, the Hebbian capacity from the empirical distributions is about 80% of that of the exponential fit, when available. Second, in general for distributions like those of these neurons, the capacity achieved by Hebbian learning is about 50% – 80% of the Gardner capacity, depending on the neuron and whether we take its discrete distribution "as is", or fit it to an exponential (or, e.g., to a log-normal) shape. Third, with Hebbian learning retrieved patterns tend to be 2 – 3 times sparser than the stored ones, again depending on the particular distribution, empirical or exponential fit (as for non-sparse distributions, which

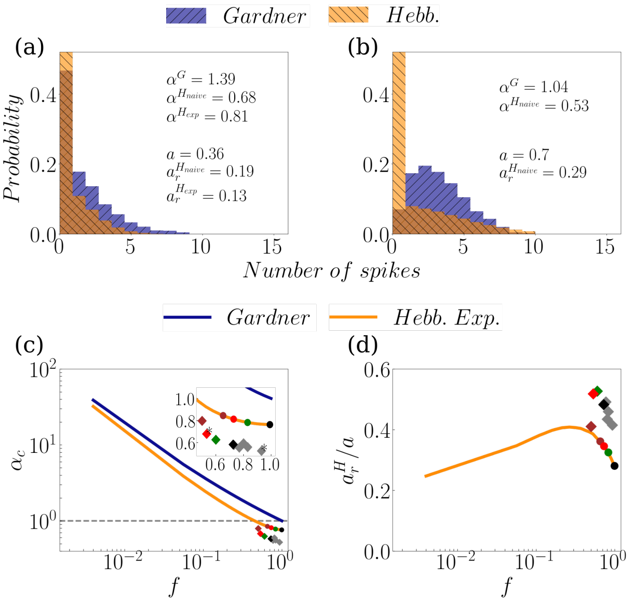


Figure 15: Hebbian vs. Gardner capacity for experimental data. (a,b) histograms of two experimentally recorded spike counts (blue) and the retrieved distributions, if the patterns were stored using Hebbian learning (orange). Note that the retrieved distributions *à la Gardner* would be the same as the stored patterns. (c) Analytically calculated Gardner capacity α_c^G (blue), compared to $\alpha_c^{H_{exp}}$ for the Hebbian learning of an exponential distribution (orange, circles). $\alpha_c^{H_{naive}}$ is shown by diamonds. The asterisks mark the two cells whose distribution is plotted in (a-b). (d) Sparsification of the retrieved patterns, for Hebbian learning. In Appendix E the same analysis in a)b) is performed on all other recorded cells.

could be better fit by a log-normal, see Appendix D.3). As illustrated in Fig. 15d, the empirical distributions achieve a lower capacity than that of their exponential fit, as the latter leads to further sparsification at retrieval.

8 DISCUSSION

While instrumental in conceptualizing memory storage [55, 82, 86], Hebbian learning has been widely considered a poor man's option, relative to more powerful machine learning algorithms that could reach the Gardner bound for binary units and patterns. No binary or quasi-binary pattern of activity has ever been observed in the cerebral cortex, however. A few studies have considered TL units, showing them to be less susceptible to memory mix-up effects [87, 88] or perturbations in the weights and inputs values [89] but, in the framework of *à la Gardner* calculations, they have focused on issues other than associative networks storing sparse representations. For instance, a replica analysis was carried out in [77] with a generic gain function, but then discussed only in a quasi-binary regime. Others considered monotonically increasing activation functions under the constraint of non-negative weights [90]. Here, we report the analytical derivation of the Gardner capacity for TL networks, validate it via perceptron training, and compare it with Hebbian learning. We find that the bound can be reached or even surpassed, and that retrieval leads to sparsification. For sample experimental distributions, we find that one-shot Hebbian learning can utilize $50-80\%$ of the available "errorless" capacity if retrieving sparser activity, compatible with recent observations [84].

In deriving the Gardner bound, we assumed errorless retrieval and it remains to be seen how much allowing errors increases this bound for TL units and neurally plausible distributions. For the binary case of [50], as already mentioned, this errorless bound is still above the Hebbian capacity of the highly diluted regime, with its continuous (second order) transition, i.e., with vanishing overlap at storage capacity [50]. How does the overlap behave in the TL case? For highly diluted TL networks with Hebbian learning, in fact, except for special cases, the transition at capacity is discontinuous: the overlap drops to zero from a non-zero value that depends on the distribution of stored neural activity but can be small [67]. It is worth noting, though, that while in the binary case the natural measure of error is simply the fraction of units misaligned at retrieval, in the TL case error can be quantified in other ways. In the extreme in which only the most active cells remain active at retrieval, those retrieved memories cannot be regarded as the full pattern, with its entire information content, but more as a pointer, effective perhaps as a mechanism only to distinguish between different possible patterns or to address the full memory elsewhere, as posited in *index* theories of 2-stage memory retrieval [91]. Further understanding would also derive from comparing the maximal information content per synapse for TL units, with Hebbian or iterative learning, as previously studied for binary networks [92]. Using non-binary patterns might also afford a solution to the low storage capacity observed in balanced memory networks storing binary patterns [93].

Our focus here has been on memory storage in associative neural networks, with the overarching conclusion that the relative efficiency of Hebbian learning is much higher when units have a similar transfer function to real cortical neurons. The efficiency of local learning rules had also been challenged by their comparatively weaker performance in other (machine learning) settings [94], while results to the contrary are also reported [95, 96]. It may therefore be argued that the efficiency of local learning in these settings might also be fundamentally dependent on both the types of units used and the data, observations consistent with the findings in [96] and [94], respectively. In evaluating a learning rule, it may therefore be crucial to consider whether it is suited to the transfer function and data representation it operates on.

Part III

CONTINUOUS QUASI-ATTRACTORS FOR IRREGULAR MEMORIES

Animals, like humans, navigate in complex territories. It is believed that mammals, at least, create *cognitive maps* of the environments they explore. The neurophysiological discovery of spatially selective cells in the hippocamal formation provided neuronal candidates involved in this capability. As a consequence, it has been hypothesized that the dynamics underlying the retrieval of environmental *cognitive maps* could be driven by *continuous attractors*, brittle mathematical objects which break with irregularities. However, the wilder the environment the more spatially selective cells seem to be activated unevenly. Can a continuous attractor theory contemplate also complex nonuniform activity? In this chapter we argue that it can, relaxing the requirement of a continuous manifold of fixed points to a *quasi-attractive continuous manifold*, intended as a direction of flow. We find that quasi-attractive manifolds *persist* under noise up to a critical value at which they abruptly break up. We show that some remarkably variable experimental recordings lie just at the edge of this transition.

9 PHENOMENOLOGICAL INTRODUCTION: COMPLEX MAPS FOR IRREGULAR ENVIRONMENTS

Each night, fruit bats can fly up to tens of kilometers to forage and then, usually, they return back to to their home colonies. When fruit bat pups become independent fliers, their initial flights are close to home and nights after nights the area explored gets larger. Already from the first night alone, though, fruit bats are able to perform shortcuts. It seems that vision is the sole sensory modality involved in this capability. In particular, it seems that a few landmarks are enough for bats to get the proper direction to home, as if they are triangulating them on a map [97].

Norway rats dig their nests under the soil, forming convoluted structures made by interconnected burrows and halls. Outside their nests, on the surface, Norway rats trace interconnected walking routes in the vegetation forming geometrically closed shapes. Each time they walk outside the nest, they tend to follow such traced routes. When snow comes and all olfactory, visual and tactile references of the surface paths are lost, one can see the rats still following the same routes, as if they have a map-like representation based on trees or other persistent vertical objects [98].

How are these complex maps encoded in the brain? How are they stored and retrieved?

The psychological hypothesis that animals are able to create "cognitive maps" dates back to Tolman's research in the 40s [10], and found, in the spatial domain, its first neurophysiological major support in O'Keef's discovery of place cells [23] and its second one in the discovery of grid cells [28].

When both these neurophysiological discoveries were published, a widely shared interpretation was that place cells set up the map, grid cells, instead, represent the detailed metric on the map. Grid fields, indeed, were observed not to change when the animal is exposed to a new environment -the map is one-, place field's, instead, do: they "remap" from one environment to the other **[fyh+07]**. Under this assumption place cells carry the contextual information, the "you are in this territory", whereas grid cells are more precise in telling "here you are" on the map.

Indeed, the simplicity of the original recordings, performed in small scale, regular and non–complex environments, transferred into the hypothesis that such peculiar, fascinating behavior could be the general norm, even in large, complex, wild environments.

As introduced in the first chapter, however, grid and place cells are embedded in a large variety of other spatially selective cells and the more one studies the behavior of all these cells in irregular environments, the more their sharp categorizations totter.

In this section I will briefly review some specific literature focusing on place cell irregularities, obtained monitoring fruit bats and rats in environments tending to be like those in the wild. These results are the point of departure of our theoretical work, presented in the next chapters.

For general considerations over spatially selective cells, their ununiformities and some open questions, refer to the global introduction in chapter 2.

9.1 MULTIPLE AND IRREGULAR PLACE FIELDS

Place cell irregularities, noted since the original experiments in simple small cages, roughly quantified already in 1993 in the dentate gyrus [99], started being appreciated as a fundamental feature only through these studies contemplating larger recording environments. Indeed, the first strong evidence of the irregular nature of place cells appeared as side effects of experiments devoted to different questions. It was only after a few of such unexpected results that place cells complexity came into the stage as a possible window into the nature of neurons: they are irreducible to simplified categorizations and thus one needs to reformulate the paradigms to conceptualize spatially selective cells in the hippocampus. Here I will review a few fundamental steps in this direction.

9.1.1 *Rats walking in large environments*

In 2008, through an experimental effort [100] aimed at clarifying whether also ventral CA3 cells respond preferentially within specific place fields, the authors recorded neurons in three main CA3 dorso-ventral areas of rats running on an 18 meter long track. Under these conditions they showed that i) all areas (dorsal - intermediate -ventral) have place cells ii) the field size scales up dorso-ventrally iii) running was associated with theta rhythmicity and phase precession at all dorsoventral recording levels. The data reported in this paper, especially those in the supplemental material, show that each place cell can have multiple fields. If indeed one looks at individual cell firing profiles, in 1d, multiple distinct and well recognizable fields are visible, furthermore they have different radii and maximal peak heights (Fig. 16a). Interestingly, when the same 1d long setup was used to record rats grid cells [104], the resulting profiles of activity from ventral mEC could look similar to those recorded in ventral CA3 [100] for a naive eye (Fig. 17).

The same year but a different group, through an experimental effort aimed at providing evidences in support of the so–called "ensemble hypothesis" and against the "dedicated-coding" one [102], i.e. the hypothesis that place cell's coding is distributed through the activity of the whole ensemble of neurons, the authors performed recordings in a chamber six times larger ($1,5 \times 1,4$ meters) and enriched (with stairs) as compared with the standard empty cylinder (68-cm-diameter). Under these conditions the authors showed that rats CA1 neurons, recorded in a free foraging tasks, are more likely to be place cells than in small environments and show multiple, irregularly arranged fields (Fig. 16c). When comparing cylinder with chamber recordings, data showed that cells which where active in both environments increased the number of fields on average 2.5 times and the fields size, on average, 1.8 times (Fig. 16d); the authors also reported that, on average, both in the "small" cylinder and in the "big" chamber, larger fields had greater peak firing rates.

Some years later, following the train of thoughts started in [102], the authors provided quantitative evidences that in 2d large, this time

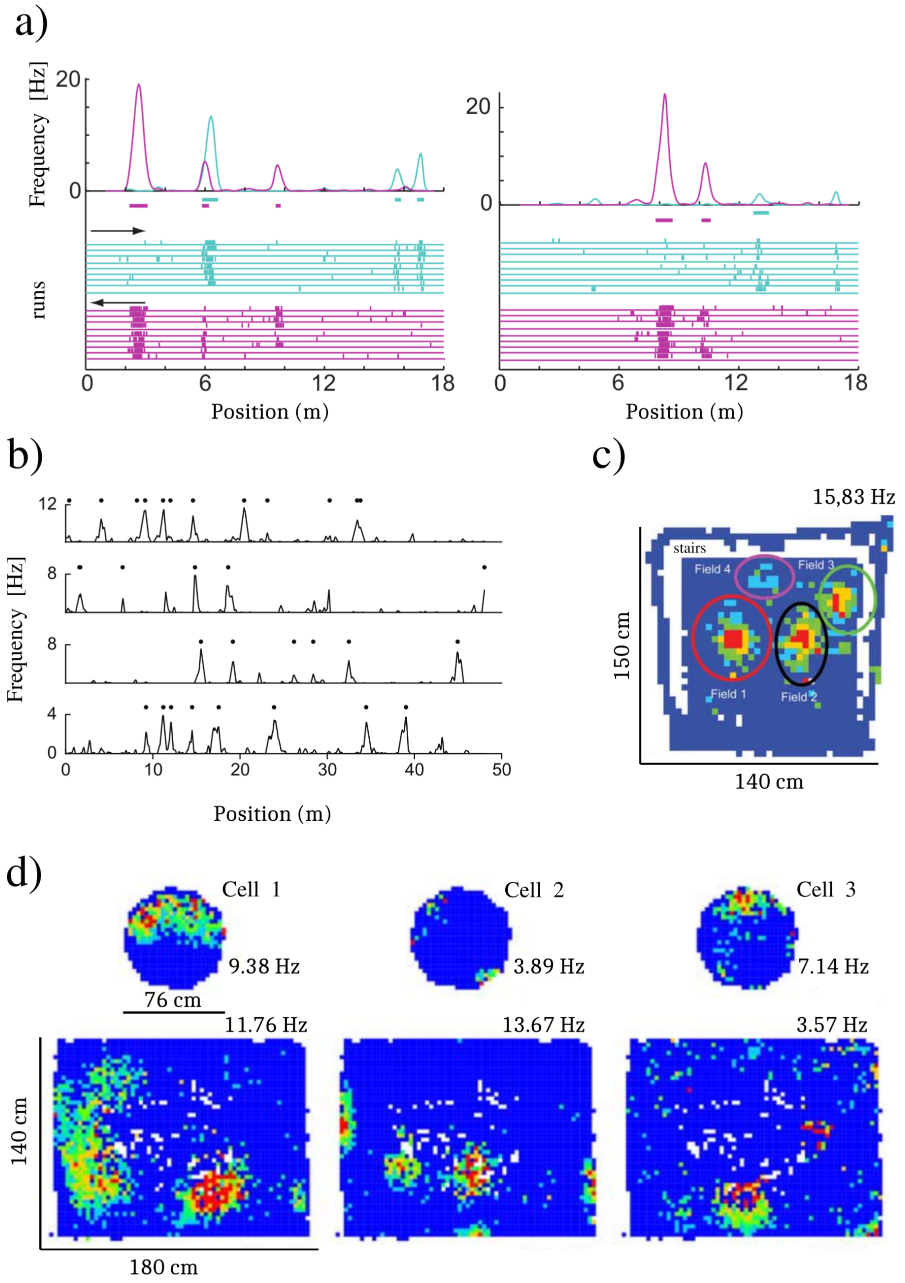


Figure 16: Place fields recorded in large environments are multiple and irregularly distributed in position, firing rates and radius sizes. a) Two examples of dorsal CA3 place cells firing response recorded on an 18 meter long track. b) Four examples of dorsal CA1 place cells firing response recorded on a 45 meter long track. c) Example of CA1 place cell response recorded in an enriched chamber with stairs (white profiles) of large sizes. d) Four examples of dorsal CA1 place cells recorded in a small cylinder and in a large chamber where more than one field becomes visible. Panel a) is adapted from [100] Fig. S5, panel b) is adapted from [101] Fig.2, panel c) is adapted from [102] Fig. 6, panel d) is adapted from [103] fig.3.

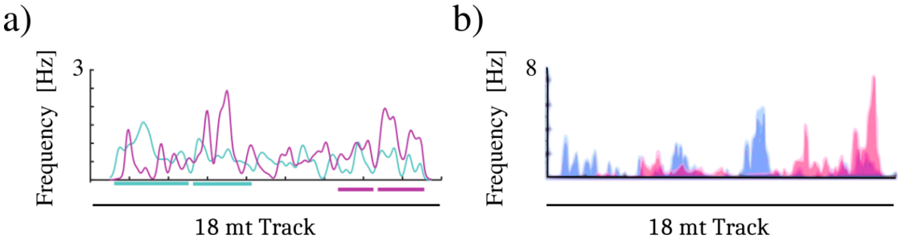


Figure 17: Comparison between recordings on the same 18mt long track of a) one example place cell in ventral CA3 b) one example grid cell in ventral mEC. Blue and pink refer to recordings in forward and backward direction. Panel a) is adapted from [100] fig. S5, panel b) is adapted from [104] Fig. 3G.

non–enriched, enclosures, place fields in rats CA1, CA3 and DG are multiple and irregularly spaced [103] (Fig. 16d). In the same paper, they argued that some geometrical aspects could be the reason why CA3 cells recorded in 1d in [100] where mostly showing one field instead in 2d they would show multiple ones. In particular they argued that directionally tuned cells, such as head direction ones, could be fundamental in modulating the discharge patterns of CA3 place cells. The evidence of fields multiplicity shown in [103] were in agreement with studies published before, such as [105], focusing on other aspects but acknowledging the presence of multiple fields in rat DG recordings or [106], characterizing the firing behavior of rat CA1 cells along the whole proximodistal axis using 2m diametr cylinders, and noticing field multiplicity in distal CA1 cells.
A few years later, in [101] the authors studied an extended linear track of 45 meters in order to test whether a limit exists in the number of place cells which can be recruited for coding a single environment. The authors showed that in novel environments without goals, the number of fields per cell follow a gamma-Poisson distribution [107], i.e each cell's field number is hypothesized to be taken from a Poisson process whose rate is defined by a gamma distribution. The recordings reported in the paper show that fields have irregular sizes and different maximal peak heights (Fig. 16b). A recent follow-up of this study [108], enforced the gamma-Poisson hypothesis as the distribution governing the number of place fields, focusing on mice instead of rats.

9.1.2 *Bats flying over long distances*

Bats navigate in the environment through different sensory modalities and they are the sole mammal which is able to fly. Since 2007, a vast amount of experimental work has been dedicated to study the neural basis of their putative cognitive maps. It was seen that bats walking in cages (hence on a 2-dimensional pavement) have cells in

their hippocampus and mEC exhibiting spatial selectivity like those recorded in rats [109–111]. Also in three dimensions fruit bats were shown to have spherical place cells [112] with equal resolutions in all directions around a center, head direction cells in a toroid reference frame [113] and (somewhat irregular) grid cells [114].
The information coming from bats has been crucial to form intuitions over some theoretical debates: on one side, theoretically, the discovery of place and grid cells in bats, which do not display consistent theta ritmicity [109, 115], discarded the hypothesized oscillatory origin of place and grid coding. On another side, grid fields in bats were characterized as those maintaining a constant distance among each other, leading to the hypothesis that hexagonal lattice symmetry may only appear in 2d.

A recently published work [116] shows that place cells recorded in bats flying in a 200mt long track have multiple irregular fields (Fig. 18). In particular, recorded cells showed up to 20 fields along the track, with sizes spanning from a few meters up to 30m and with peak rates ranging from a few frequencies up to 45 Hertz.

These experimental results, especially those, on bats, have been the origin of our theoretical study aimed at understanding how irregularity may affect continuous attractor dynamics.

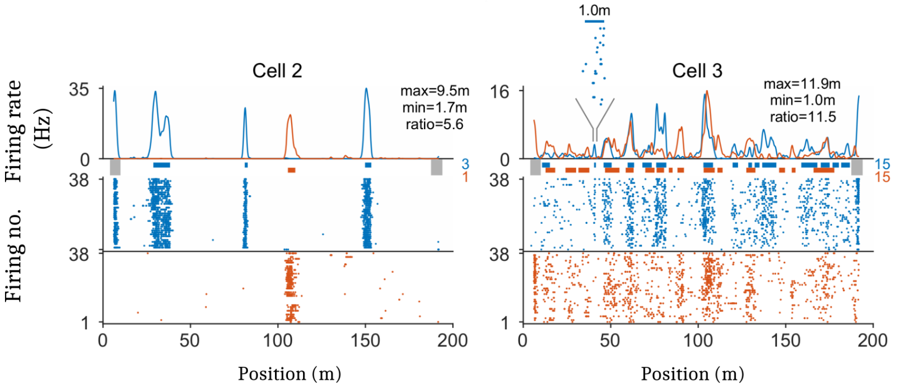
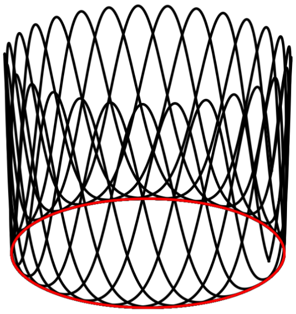


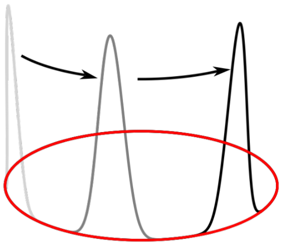
Figure 18: Place fields recorded in a 200mt long track, from the bat hippocampus. Sample firing rate map of two cells. Figure adapted from [116]

10 THEORETICAL INTRODUCTION: CONTINUOUS ATTRACTORS ARE BRITTLE

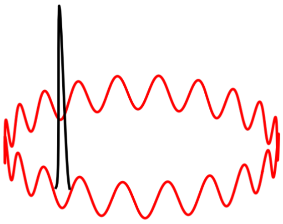


Continuous attractor neural networks, introduced in chapter [-], were brought to neuroscience in 1977 [117] and as models to describe neuronal selectivity to sensory inputs in 1995 [118]. Their opening to the field of memory occurred instead between 1995 – 1998. After the application of continuous attractors to head direction cells [61] and to place fields in 1995 [119], the fundamental concept of *map / chart* was introduced in 1997 [63]: the evidence that multiple representations co-exist in the hippocampus [36] transferred into the hypothesis that each *map* is a continuous attractor representing the coordinates of a putative environment. The recurrent collaterals present in CA3 where hypothesized to be the substrate to store such *multi-map* continuous attractor. Each map would be, in this context, the imaginary arrangement of a population of place cells on an abstract plane where each cell is placed at the position of its field center in the *environment* related to that map. The navigation in each environment would be based on *path integration* exploiting the continuity of the attractor [63]. The storage capacity of such a multi-map continuous attractor, i.e the maximal number of *environments* which can be stored per connection, was finally evaluated in 1998 [55].
Since then, continuous attractor neural networks (cANN) are regarded as cardinal mathematical objects to gain intuition over the mechanisms underlying memory storage and retrieval of continuous variables, and as fundamental concepts in models of place and grid cell networks.

The biological plausibility of continuous attractors, however, keeps being challenged by their brittleness to fast and quenched noise.

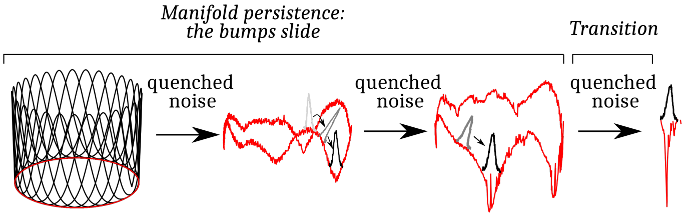


Fast noise is that which refers to the dynamics: if any perturbation outside the *manifold* is easily corrected back, as if the manifold represents a river at the bottom of a valley, the same does not hold for any perturbation along the manifold, i.e. in the direction of the river. Any tiny push in that direction (red on the schematic) is enough to bring the system to another configuration, i.e. to a nearby position. Noise is ubiquitous in the brain, making this issue a fundamental challenge which focused the attention of a vast number of studies. In Sect. 10.1 I will shortly touch on some of the main results obtained.



Quenched noise, instead, is that which refers to the network structure and thus to the shape of the attractor. Since the initial works in the 90′s it was noted that inhomogenities in the coupling parameters lead to the collapse of the continuous attractor into sets of discrete fixed points [119, 120]. The nature of such inhomogenities plays a fundamental role in the effects on the continuous attractor and the putative emergent scenarios are not yet fully explored. In Sect. 10.2 I will briefly review the results which, so far, contributed to unravel the interplay between quenched noise and continuous attractors models of neuronal responses and memory.

As introduced in chapter 9, irregularities are an inherent feature of place fields. In attractor neural networks based on Hebbian learning, thus, *irregular fields* act indirectly as *quenched noise* deforming the shape of the continuous attractor.



In our work, we study the effect of a systematic increase in fields variability acting as *quenched noise* in an attractor neural network. We notice that for sufficiently low variability the discrete fixed points still lie on the continuous manifold, which remains stable, though deformed, in the remaining directions: the bumps of activity monotonically *slide* on the manifold to reach the fixed points. So it does driven by *fast noise* in intact continuous attractors [117, 118, 121–129], or under quenched disorder induced by multiple-maps storage [130–134], by quenched random noise [122, 135–140] or by specific asymmetries in the couplings [60, 93, 141–143]. Crucially, however, we see that above a critical level of irregularities, the quasi-attractive manifold abruptly disappears with a *transition* and fixed points are reached through a trajectory in the phase plane outside of the manifold. We hypothesize that place fields storage and retrieval, even in the situation in which there is only an individual environment, could thus be driven by a *quasi-attractive continuous manifold* contemplating different resolutions of place fields, which would turn into a *continuous attractor* solely in the case of uniform individual fields.

10.1 BRITTLENESS TO FAST NOISE

Continuous attractor neural networks (cANN), introduced as mathematical objects able to track time-varying stimuli in real time, can, as a side effect, be pushed from one state to another by a little amount of noise in the direction of the manifold. This was noted already by the group of Amari, which firstly proposed the application of cANN to Neuroscience [117] and which, around 30 years later, proposed one of the first attempts to systematically study the phenomena [126]. In particular, the authors simplified the dynamics on the tangent of the attractor as an Ornstein-–Uhlenbeck process and saw that for such simplified model, the average error, i.e. the probability that the bump position on the manifold would not ecode the correct position in real space, increases in the dynamical evolution up to a constant value given by a signal to noise ratio. Meanwhile, the same phenomenon had been acknowledged by several studies applying cANN to orientation selective cells in visual cortex and head direction cells [118, 121–125, 131] but was not addressed specifically.

In a follow-up study a few years later, Si Wu and colleagues studied analytically a simplified cANN where each stationary state was assumed to have a gaussian shape. The authors described the dynamics through functions of quantum harmonic oscillators, which enabled to decompose it into different modes (modification of the Gaussian bump in i)peak height ii)position iii)width and iv)skewness) [127]. Considering as the predominant mode the position, i.e the movement of the Gaussian bump along the manifold, the authors developed a time dependent perturbative approach able to track the network dynamics of such simplified model.

In 2012, Burak and Fiete approached the same problem and studied the effects of neuronal noise given by the irregular statistic of the spikes. In particular the authors explored how Poisson spiking neurons influence a random drift along the manifold of the instantaneous attractor state, defined as the state where the network would go without noise. In this way, they analytically derived an information-diffusion inequality setting a lower bound on the diffusion of such state [128].

Between 2013 and 2015 three consequent studies by Monasson and Rosay focused specifically on a *multi-map* continuous attractor model for binary place cells and evaluated analytically its detailed dynamical properties. After estimating the phase diagram [129], the authors analyzed the diffusion of the bump, there called "clump", within one map [134] and the transition towards a different map [144]. The authors showed that the brittleness to *fast noise*, i.e. the diffusion of the bump due to neural noise, is in competition with the *quenched disorder*, i.e. the tendency to transit towards another map. In the single-environment case, i.e. in the absence of *quenched disorder*, the authors derived a description of the dynamical evolution obtaining an analytical expression of the diffusion coefficients, in excellent agreement with numerical simulations.

Since then, other studies have been exploring analytical details of the clump diffusion along the manifold in multi-charts models of place cells, thus considering the underlying quenched disorder, as Zhong *et al*, who evaluated the update speed given by external stimuli [140].

10.2 BRITTLENESS TO QUENCHED NOISE

How the susceptibility of continuous attractors to fragmentation may deal with the storage of continuous variables represents an intriguing riddle lasting since 1995, when this phenomena was firstly pointed out for a place cell cANN storing an individual environment [119]. In particular, in that study Tsodyks and Sejnowski showed that if the stored fields distribution in simulations shifts from being uniform and regular to uniform but random, then the fixed points collapse from being a semi-continuous set to a very low number. The authors emphasized that such positions correspond to those on the manifold characterized by strongest interactions.

The following studies on the issue, between 1996 − 2003, mainly focused on head direction models and on potential mechanisms able to maintain, despite disorder in the connections, the *spatial working memory function*, i.e. the possibility to retrieve a semi-continuous attractor. Zhang in 1996 studied the phenomena by adding Gaussian noise in the connections and approximated the speed of the clustering drift towards a limited number of fixed points [122]. He argued that selective Hebbian learning, activated when active movements are occurring, could be a possible mechanism used by the brain to smooth over irregularities. In two consequent studies in 2002, instead, Stringer *et al* studied the effect of noise in the connections derived from a self-organized learning procedure in 1d [135] and 2d [136]. The authors proposed two biologically inspired mechanisms to stabilize the tendency to drift, either enhancing those cells already firing, as biologically motivated by short term enhancement, or taking advantage of the nonlinear activation of neurons with NMDA receptors, enhancing the activity of those already sufficiently active. Finally, in 2003, Renart *et al.* devoted an entire study to the phenomenon and to a possible mechanism to overcome the drift. In particular the authors incorporated in the disordered cANN an activity dependent scaling of the synaptic weights and showed that this would lead to similar long term average firing rates per each neuron, homogenizing the network and thus recovering a robust working memory, i.e. the ability to durably retrieve the activity bumps [137].

The same year Treves proposed a cANN where quenched disorder took the form of discrete Hopfield patterns, generating a model for the simultaneous retrieval of discrete and continuous information in cortical patches [141]. The model was then analysed in analytical [60, 142] and numerical [93] detail in three consequent papers by Roudi and Treves. The authors, in this context, where not focused on the drift but rather on mechanisms used by the brain to differentiate *where* and *what* information. The *where* information would be associated with the continuity of the attractor, i.e. with a number of fixed points on the *2d* manifold, and a gain modulation was proposed as a possible mechanism able to increment the tendency of a bump to stay still.

The following year, in 2004 Treves proposed a self-organized cANN [130] model for multiple *charts* of place fields in CA3, generated following a Hebbian learning process. The model was further analysed by Pepp and Treves in 2007 who emphasized how the continuous attractor *wrinkles* due to the quenched disorder introduced by the storage of multiple *charts*. The authors showed that on the wrinkled surface the majority of fixed points on the manifold is unstable [131] and thus firstly hypothesized that efficient coding of position could be possible only on timescales between the attraction to the surface and before the bump drifts. In 2013 Cerasti and Treves, in an analogous model, showed that increasing the system size would decrease the fragmentation of the attractor but the tendency to drift would persists [133].

In 2006 Hamaguchi *et al.* proposed a detailed analytical study of a ring cANN endowed with binary units and symmetric quenched disorder taken from a Gaussian distribution. In particular the authors derived the phase and bifurcations diagrams of the model and related *fast* and *quenched* disorder, showing that the latter could be beneficial in reducing the intrinsic drift tendency due to *fast* noise [138].

In 2010, focusing again on a *multi-chart* cANN model for place cells, Hopfield hypothesized that the bump shift due to the storage of other charts could be a mechanism underlying mental exploration [132]. In 2011, instead, Itskov *et al.* studied a ring architecture, highly perturbed by quenched random noise, modelling a generic parametric working memory. The authors showed analytically that through synaptic facilitation the drift of the bumps, induced by the distortion of the attractor, could be slowed down, enabling the retrieval of each memory in biologically plausible timescales [139].

Between $2013 - 2015$, as introduced above, three papers by Rosay and Monasson explored the analytical details of the interplay between *fast* and *quenched* disorder in a binary *multi-chart* model for place cell storage [129, 134, 144]. In particular the authors showed that the tendency to cross-talk between maps is in competition with the tendency to drift within one map.

Finally, a more recent result published by Spalla *et al* [143], studies the effect of an asymmetric component in the connections, a sort of ordered quenched disorder, in a multi-chart cANN model of threshold linear place cells. In particular the authors show that asymmetric connections lead to a bump of activity constantly drifting on the manifold and evaluated the storage capacity, which turns out to be enhanced with respect to the symmetric case [55].

11 MODEL DEFINITION

Large scale variations within the firing maps of place cells have been widely reported (see Sect. 9), a body of observations which challenges previously established theoretical models (see Sect. 10).

An alternative theoretical scenario has been proposed recently in [116], where the authors consider multiple classical continuous attractors, of various scales, interacting with each others, thus enabling fields to span the scale of each attractor.

In this study, instead, we explore the effect of irregularities on the formation of a simple continuous attractor in an associative memory neural network and ask what could one consider as *retrieval*, given that the manifold of fixed points is extremely brittle to any source of noise.

We find that i) while fixed points become few and discrete, yet they lie on the continuous manifold which *persists*, i.e. it remains attractive with respect to the other N-1 dimensions in phase space and that ii) this holds true up to a critical level of noise, after which, the *quasi-attractive continuous manifold* seems to abruptly break up through a *phase transition*.

Applying our analysis to the experimental distributions obtained in [116], we observe that the recordings, notable for their irregularity, lie just at the edge of the transition.

This lead us to hypothesize that memory storage and retrieval of place cells firing patterns can be understood as the establishment of a *persistent continuous quasi-attractive manifold*, intended as a robust direction of flow, which converges to the standard concept of continuous attractor neural network if the quenched patterns are precisely regular.

11.1 MODEL DESCRIPTION

Assume that we have a network of N threshold linear (TL) units, with full recurrent connectivity, which attempts to store in memory the neural representation, in terms of place fields, of a continuous variety parametrized by $\vec{s}$, which we take to have dimensionality d. For simplicity we start with $d = 1$ and drop the vector symbol, although the analysis should be easy to generalize. Later we will consider also the generalization to sparse connectivity, analysing in particular the case of a so-called *highly diluted* network. The continuous representation (taken to be imposed by external inputs to the network, e.g., coming mainly from the Dentate Gyrus if the network is the CA3 one) is expressed by (non-negative) firing rate variables $\{\eta_i(s)\}$, where each unit i is taken to have produced, at the memory encoding stage, a number of place fields of variable peak rate and width. Such variability we assume to be effectively summarized by two (quenched) disorder parameters $\sigma_{dimension}$ (σ_d) and σ_{peak} (σ_p). This has led to a matrix of recurrent weights given by

$$J_{ij} = \frac{1}{N}\int_{\mathcal{S}} \frac{ds}{S}\left[\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right]\left[\frac{\eta_j(s)}{\langle\eta\rangle} - 1\right] \tag{110}$$

with $\bar{\eta}_i \equiv \int_{\mathcal{S}} \frac{ds}{S}\eta_i(s)$.

11.2 ENERGY

At retrieval, we assume the network, once it has been released from external inputs, to evolve driven solely by the recurrent interactions, changing continuously the output of its units according to

$$\frac{dV_i(t)}{dt} = -V_i(t) + g[h_i(t)]^+ \tag{111}$$

where $[\cdot]^+$ sets negative values to zero, g is a fixed gain parameter, and

$$h_i(t) = \sum_{j\neq i} J_{ij}V_j(t) - T\left(\frac{\sum_j V_j(t)}{N}\right) \tag{112}$$

is the input current to each unit, relative to a common threshold value T that is written to incorporate feedback inhibition. We consider $T(\nu) = 4k(\nu - \nu_0)^3$ where we use the shorthand $\nu = \sum_i V_i/N$, and ν_0 is a desired mean value (see Appendix F.6 for its implementation). Define a quantity

$$E(\{V_i\}) = -\frac{1}{2}\sum_i^N\sum_{j\neq i}^N J_{i,j}V_iV_j + NB\left(\frac{\sum_i V_i}{N}\right) + \frac{1}{2g}\sum_i (V_i)^2 \tag{113}$$

with $dB(\nu)/d\nu \equiv T(\nu)$, i.e $B(\nu) = k(\nu - \nu_0)^4$. For any unit i above threshold, $V_i(t) > 0$, we have

$$\frac{dE(t)}{dt} = \sum_i \frac{\delta E(\{V_i\})}{\delta V_i}\frac{dV_i(t)}{dt} = -\frac{1}{g}\sum_i\left[\frac{dV_i(t)}{dt}\right]^2 < 0 \tag{114}$$

showing that E behaves as an energy function in the hyperquadrant $V_i(t) > 0$. Then the Hessian is

$$\frac{\delta^2 E(\{V_i\})}{\delta V_i \delta V_j} = -(1-\delta_{ij})J^R_{ij} + \frac{12}{N}k(\nu - \nu_0)^2 + \delta_{ij}/g. \tag{115}$$

In Eq. (115) J^R_{ij} is the weight matrix J_{ij} restricted to the units above threshold: columns and rows of those below threshold are removed.

11.3 MANIFOLD

When each unit i has a single field in the quenched pattern $\eta(s)$ and the fields of the entire population are identical and regularly placed in N discrete equispaced positions in S, then in the $N \to \infty$ limit Eq. (110) leads to a continuous attractor. When the regularity of the fields in $\{\eta_i(s)\}$ (or the regularity in the network connectivity, which we do not treat yet) is perturbed, the continuous attractor "breaks": only some of the fixed points in the continuous attractor remain; they turn, then, into discrete ones. Such a system can be well described through the energy function (113). One can further define a cosine overlap parameter (a different measure from the overlap $x(s,t)$ which it is convenient to use in the later replica analysis)

$$O(\eta(s), V(t)) = \frac{\sum_i^N \eta_i(s) \cdot V_i(t)}{\sqrt{\sum_i^N (\eta_i(s))^2 \cdot \sum_i^N (V_i(t))^2}} \tag{116}$$

as the cosine similarity between each configuration on the manifold $\{\vec{\eta}(s)\}$ and the state $\vec{V}(t)$ in the time-evolution of the dynamics, as introduced in Sect. 3.3.2 (we refer to this, as the overlap space). If this quantity is evaluated over all $\vec{\eta}(s)$ and it has a bump-like profile in s with high overlap at the center we consider the state to be *localized* on the manifold[1]. We take the s value at which the maximum occurs as the position on the manifold where the activity is localized. If, instead, the overlap evaluated over all $\eta(s)$ is spread out, we consider the state to be non-localized.

During a dynamical evolution towards a fixed point, if the localized state monotonically slides on the manifold, we consider the manifold as *quasi-attractive* or *locally stable*. If, instead, the localized state deforms, spreads out, and perhaps reforms at a different position of the manifold leading to a fixed point, we consider the dynamics to have jumped outside the manifold and the part of the manifold where it should have slid, to have disappeared. In Fig. 38, presented in the following chapter we illustrate, with an example, these concepts.

Remarkably, it should be noted that $\{\vec{\eta}(s)\}$ is not the manifold *per-se*, instead, we would like to define a manifold of configurations $\{\vec{\xi}(s)\}$ where for each s and unit $\vec{\xi}(s) \approx \vec{\eta}(s)$, as the $\{\vec{\eta}(s)\}$ merely represents the quenched patterns which were used to create the $\{\vec{\xi}(s)\}$ through

[1] When the irregularity is low, this can be visualized also as a bump in the activity space (see Sect. 3.3.2 or the supplementary Fig. 37 and Fig.37.) plotting the activity of each cell V_i at the position s on the manifold where $\eta_i(s)$ is maximal.

Hebbian learning, and do not necessarily coincide with them. In fact, the configurations $\{\vec{\eta}(s)\}$ will not in general be stable. Further, only some of the locations s will be represented by reasonably close $\{\vec{\xi}(s)\}$. We would like, however, to refer to a manifold for any location s, not only for those represented by stable configurations (see also Sect. 12 for further details).

It turns out that most parts of the manifold $\{\vec{\xi}(s)\}$ break up almost synchronously at a certain noise level.

Before providing the quantitative measurements and the characterization of the phase space, let us take a step back to introduce how we define the patterns $\{\vec{\eta}(s)\}$ to reproduce real data.

11.4 ALGORITHM FOR DATA GENERATION

We take as paradigmatic experimental results those recently published in Ref. [116], introduced in Sect. 9.1.2, and focus on three main sources of variability: the number of fields, the field size and the field peak rate. Remarkably, our results, presented in the following chapter, do not depend on the details of the distributions and are general also in simpler systems, without the constraints described in the following subsections. The latter, which specify the system we consider in finer detail, are solely required to reproduce satisfactorily the observations in [116].

11.4.1 *Number of fields*

In the experimental results published in [116], the authors show that place cells recorded in large environments can have up to 20 fields (Fig. 19B) spanning from small to large ones (Fig. 19A) with an average of 4.9 fields per cell in each flying direction.

In order to simulate the distribution shown in Fig. 19B, we consider that the number of fields n_F per each unit is randomly drawn from an exponential distribution with probability function $f(n_F, \frac{1}{\zeta}) = \frac{1}{\zeta}\exp(-\frac{n_F}{\zeta})$ under the constraint that zero values, or values above $n_F = 21$, are not accepted. This constraints leads to a distribution of higher average than ζ, and we find (see figure in Appendix F.1) that setting $\zeta = 4.7$ results in $\langle n_F \rangle \approx 4.9$.

11.4.2 *Fields shape and size*

We consider each field κ to be characterized by: i) the position s_κ of its center, ii) its linear dimension, or effective diameter, d_κ and ii) its

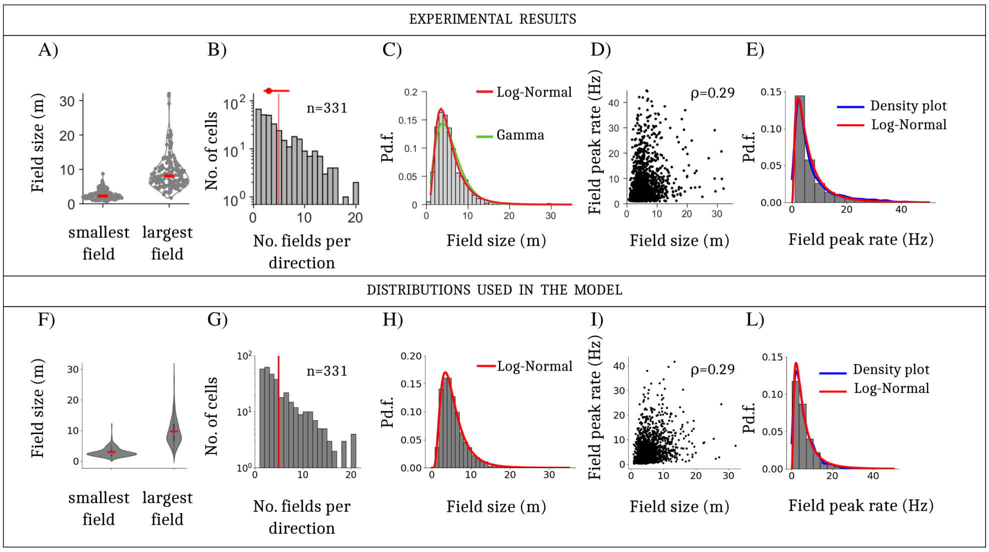


Figure 19: Comparison between experimental results (first row) and the distributions resulting from the algorithm we use (second row). Subplots A)B)C)D) are borrowed from the original article [116] for the sake of comparison. Subplot E) was obtained from the experimental data plotted in D) kindly given us by the authors. A)F) Distribution of smallest and largest field sizes per neuron (A) or per model unit (F) (those shown have at least 2 fields). B)G) Distribution of number of place fields per neuron (B) or per model unit (G) in one direction. In B) the bar at 20 includes all numbers above 20. The average number of fields in B) is 4.9 and coincides to the one obtained with this random realization of our procedure. C) Distribution of fields sizes as obtained in experiments, the log-normal parameters of the fit ($\mu = 1.57m$, $\sigma = 0.575m$) coincides with the fit of the analogous distribution resulting from our procedure (H). D)I) Scatter plot of the field size versus the peak firing rate of each field as obtained in experiments (D) or from our algorithm (I). ρ is the Spearmann correlation coefficient between the two measures in the plot. H) Distribution of the experimental peak firing rates, fitted with a log-normal distribution with parameters ($\mu = 1.549Hz$, $\sigma = 0.884Hz$) L) Distribution of peak firing rates as obtained with the algorithm. The sample distribution in F)G)H)I)L) was obtained with N=331 units.

peak firing rate p_κ.

Please note that we assume periodic boundary conditions, such that fields are effectively lying on a ring of dimension $L = 200\text{m}$, which we depict, throughout the Thesis, as open rather than closed, for ease of visualization. Further, we assume that each field can be modeled as a Gaussian bump, centered at s_κ, with dimension $d_\kappa = 2\sigma$ and maximal height p_κ (Fig. 20).

The activity of a unit i at a certain position s is thus given by:

$$\eta_i(s) = \sum_{\kappa}^{n_F^i} p_\kappa \left[\frac{\exp\left(-\frac{(s-s_\kappa)}{0.5 d_\kappa^2}\right) - \exp(-\frac{1}{2}))}{1 - \exp(-\frac{1}{2})} \right]^+ \tag{117}$$

where n_F^i is the number of fields assigned to unit i and $[\,]^+$ sets all negative values to zero.

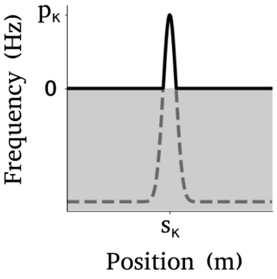


Figure 20: Example of a field κ (black line).

For the sake of reproducing the statistics measured in experiments we do not allow fields to overlap, and we constrain the sum of the dimensions of all fields belonging to a unit to be less than $(L - 3n_F)$ meter, where $L = 200\text{m}$ is the size of the environment and the last subtraction facilitates finding appropriate random positions for the fields on the ring.

We randomly draw the size d_κ of each field from a log-normal distribution $\mathcal{L}(\mu_d, \sigma_d)$ with $\mu_d = 1.57\text{m}$ and $\sigma_d = 0.575\text{m}$, as resulting from the fit to the experimental data, reported in [116] (Fig. 19E,F).

11.4.3 *Peak rates*

The peak firing rate distribution, in the experimental recordings, can be roughly fit with log-normal distribution $\mathcal{L}(\mu_p, \sigma_p)$ with parameters $\mu_p = 1.549\text{Hz}$ and $\sigma_p = 0.884\text{Hz}$, which we estimated from the data presented in ref. [116] (reproduced in Fig. 19D) kindly given us by the authors.

In order to introduce the correlation seen in the experimental recordings (Fig. 19D), given a field with a certain dimension d_κ we define the specific mean

$$\mu_{p_\kappa} = \mu_p + 0.5 \log \left[\frac{d_\kappa}{\exp\left(\mu_d + \frac{\sigma_d^2}{2}\right)} \right] \tag{118}$$

and draw the first guess $\tilde{p}_\kappa$ of the peak firing rate corresponding to that field from a log-normal distribution $\mathcal{L}(\mu_{p_\kappa}, \sigma_p)$. Then, we prevent having fields with peak firing rate much higher than 40Hz by using instead of $\tilde{p}_\kappa$ directly, its non-linear transform p_κ, inspired by those used in phonology to transform Hz into Bark. In particular we obtain the slightly modified peak firing rate of a field κ as:

$$p_\kappa = 30\text{arctan}\left(\frac{\tilde{p}_\kappa}{30}\right) + 6\text{arctan}\left(\left(\frac{\tilde{p}_\kappa}{120}\right)^2\right) \tag{119}$$

These functions, overall, lead to distributions satisfactorily agreeing with the recordings, as shown in Fig. 19 D-I and E-L.

12 CONTINUOUS QUASI-ATTRACTIVE MANIFOLDS

Any tiny source of irregularity in $\{\vec{\eta}(s)\}$ becomes, through Hebbian learning, quenched noise in the connectivity matrix. This breaks the asymptotic continuity of the fixed points on the continuous attractor manifold (see Ch. 10).

We show that what remains of this manifold is a robust direction of flow of the dynamics, which persists to be attractive in the remaining $N-1$ directions, even when irregularities are far more than tiny. We refer to this mathematical object a *quasi-attractive* or *persistent* continuous manifold $\{\vec{\xi}(s)\}$.

Let us characterize it with an example. In Fig. 38A - B we report the energy landscape of two quenched patterns $\{\vec{\eta}(s)\}$, which we use to create the connectivity matrices underlying the dynamical evolution reported, with few representative steps, in Fig. 38C - D respectively –refer to supplementary Figs. 37 and 38 for the plots in the activity space.

One can evaluate whether a dynamical evolution is localized on a manifold by looking at the overlap space (introduced in Sect. 11.3 and Sect. 3.3.2), i.e. at the overlaps of the time dependent variable $\vec{V}(t)$ with all $\vec{\eta}(s)$. If the activity is localized on $\{\vec{\eta}(s)\}$, this can be intended as an indication of the existence of that portion of the quasi-attractive manifold $\{\vec{\xi}(s)\}$.

In Fig. 38C the localized bump in the overlap space monotonically slides on the quasi-attractive manifold towards one of its minima. As a proof of concept of the non-exact identity but high similarity between the quasi-attractive manifold $\{\vec{\xi}(s)\}$ and the patterns $\{\vec{\eta}(s)\}$, introduced in Sect 11.3, one can see that the dynamical evolution in Fig. 38C reaches, roughly, a position lying between the second and third minima of the energy of $\{\vec{\eta}(s)\}$. This is because in $\{\vec{\xi}(s)\}$ the first four minima of $\{\vec{\eta}(s)\}$ become a unique global minimum (see Supplementary Figure 40).

When the noise exceeds a certain threshold, instead, the scenario changes. We observe that during the dynamical evolution the localized bumps deform, spreads out and, under certain conditions, reform on the manifold in a different position (Fig. 38 B-D).

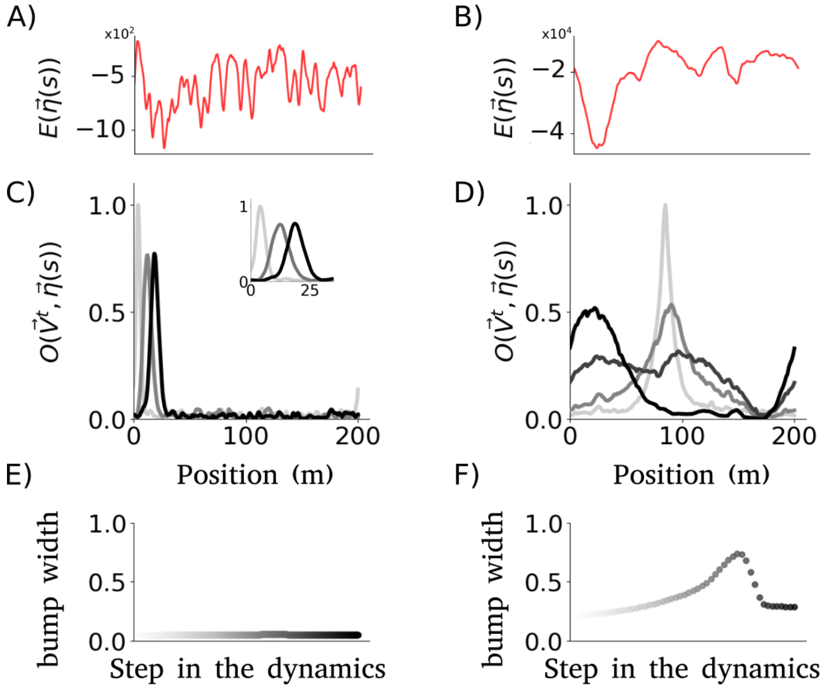


Figure 21: Dynamical evolution which *slides* or *jumps*. A)B) Energy landscape in s of two realizations of quenched patterns $\eta(\vec{s})$ for different noise levels. C)D) Overlap of a few $V(t)$ configurations while the dynamics reach the fixed points, starting from $\vec{\eta}(s = 4)$(C) and $\vec{\eta}(s = 85)$(D). Colors indicate the timestep scaled from light gray (initial condition) to black (fixed point). The dynamics presented in C) slides on the manifold (the inset represents a zoom-in) while the one presented in D) jumps outside and reenters. E)F) Estimation of the standard deviation around the center of mass of the overlap profiles, removed of all ripples below $O(\eta, V) = 0.1$ (see Ch. 13 and supplementary Fig. 39). A)C)E) refer to $\{\vec{\eta}(s)\}$ characterized by $\sigma_d = 0.4$, $\sigma_p = 0.4$, $\zeta = 1$. A)C)D) refer to $\{\vec{\eta}(s)\}$ characterized by $\sigma_d = 0.9$, $\sigma_p = 0.9$, $\zeta = 1$. The networks are composed of $N = 8000$ units, s is discretized into 1000 equally spaced positions every 0.2m and the dynamics is considered to converge when $\sum_i (V_i^t - V_i^{t-1})^2 < 10^{-5}$.

We characterize this behaviour as a *jump* outside the quasi-attractive manifold, and interpret it as its loss of attractivity in the remaining $N-1$ directions, i.e. as its local disappearance. In the supplementary Fig. 36 we report a few more examples of dynamics which we regard as *jumping* outside the former quasi-attractive manifold.

12.1 MEASURE

One can think of different approaches to quantify the loss of attraction, in the remaining $N-1$ directions, of the quasi-attractive continuous manifold. Here, we use three measurements, and refer to Sect. F.3 for the specific details.

1. **Proportion of dynamics which jump**: Given a large number of dynamics starting at different $\vec{\eta}(s)$, with s spanning the whole lengths, we take the proportion of dynamics which exit the manifold, i.e. those in which the localized bump does not slide continuously, as a measurement of the percentage of the manifold which has vanished.

2. $\langle \mathbf{O_{tang}} \rangle$: Given the complete set of residual fixed points on the manifold we evaluate, for each, the eigenvector corresponding to the smallest eigenvalue of the Hessian matrix (Eq. (115)) and estimate its cosine similarity with the direction of the manifold. This quantity, which we name O_{tang}, equals 1 only if the eigenvector closest to instability is exactly aligned with the manifold.

3. **Bump width:** Given a configuration of activity, whether this a dynamical one or a fixed point, we can estimate its localization on the manifold as the standard deviation of the center of mass of its overlap profile with $\{\eta(s)\}$. The closer this value, once normalized, is to 1 the more the configuration of activity is spread on the manifold and the less it is localized. To make this quantity informative we remove all sources of noise in the overlap profile by thresholding it to an arbitrarily set value of 0.1. Fig. 21E-F shows the estimated bump width during the dynamical evolution reported in 21B-C.

13 PHASE TRANSITIONS AND MANIFOLD PERSISTENCE

We study identical realizations of the quenched place field centres of the patterns $\{\vec{\eta}_i\}$, setting the irregularity of the peak firing rates and field numbers at the experimental level, and vary progressively the parameter σ_d (which regulates the fluctuations in the field size).

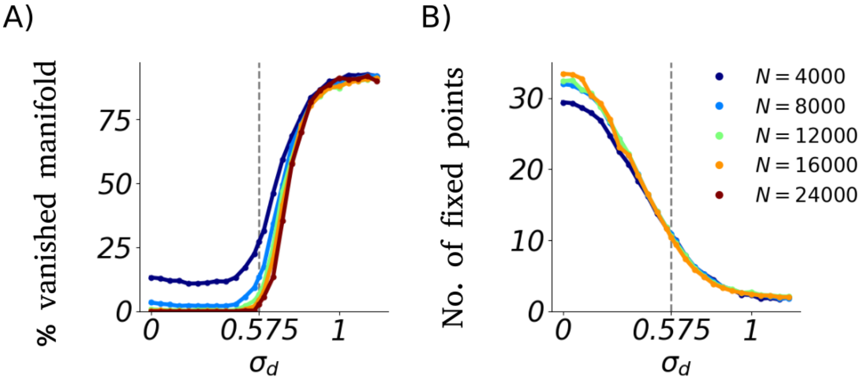


Figure 22: A) Phase transition in quasi-attractive manifold persistence - average proportion of jumps in the dynamics. We consider the proportion of dynamics which jump outside the quasi-attractive manifold as an estimation of its deterioration (see Sect.12.1); the transition gets steeper with the size of the system. B) Average number of fixed points versus σ_d. The number of fixed points do not depend on the size of the system. The vertical line at $\sigma_d = 0.575$ corresponds to random realizations of the distribution modeling experiments (see Fig. 19). Parameters: Data are produced following the algorithm described in Sect. 11.4 and all parameters except σ_d are set to model experimental results ($\zeta = 4.7$ to guarantee $\langle n_F \rangle \approx 4.9$, $\mu_d = 1.57m$, $\mu_p = 1.549Hz$, $\sigma_p = 0.884Hz$). Each point on each curve is obtained averaging over 25-90 quenched realizations of the network (fewer when the system size is larger), simulating 50 dynamics each initialized with a different $\eta(\vec{s})$, with s spanning homogeneously the whole length (one every 4m). s is discretized into 1000 equally spaced positions every 0.2m and dynamics are considered to have converged when $\sum_i |(V_i^t - V_i^{t-1})| < 10^{-10}$. The step size varies from $\gamma = 0.08$ to $\gamma = 0.04$, $g = 17$, $k = 300$.

We find that the quasi-attractive manifold, as we have characterized it in the previous chapter, breaks up with a phase transition at a critical σ_d, which roughly coincides with that fitted from experimental data ($\sigma_d = 0.575$).

While the number of fixed points decreases with σ_d (Fig. 22B), and not, interestingly, with the size of the system, the quasi-attractive manifold persists intact up to the critical σ_d at which it undergoes a transition with a steepness that does depend on the size of the system (Fig. 22A and Fig. 23), as typical in phase transitions.

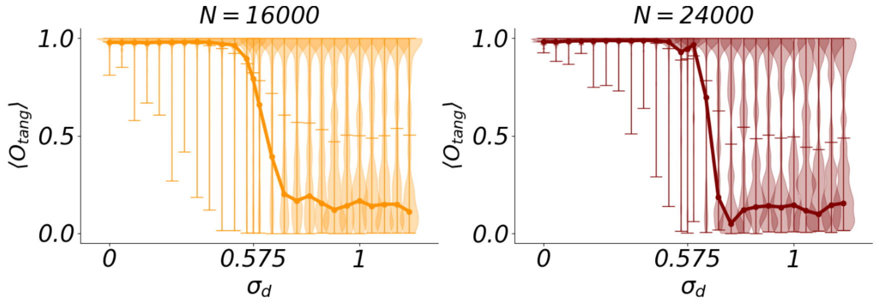


Figure 23: Phase transition in quasi-attractive manifold persistence - direction of the eigenvectors. $\langle O_{tang} \rangle$ quantifies in the range $0-1$ the alignment of the most unstable eigenvectors with the direction of the manifold (see Sect.12.1); the transition from alignment to non-alignment gets steeper with the system size. The solid line represents the 0.25 quantile (75% of the overall data lie above the line), horizontal lines correspond to the average, violin plots indicate the complete distributions. Refer to the parameters of Fig. 22.

The average direction of the most unstable eigenvectors, corresponding to any fixed point (either those which disappear or further stabilize when increasing σ_d), transition similarly, as shown in Fig. 23. If the unstable eigenvectors remain aligned to the manifold up to the critical σ_d, they do not seem to show any preferred direction after the transition.

13.1 A PHASE DIAGRAM WITH 3 DISTINCT REGIONS

We explored numerically the surrounding phase space by systematically varying the other sources of irregularity in the fields. As a short recap of Sect. 11.4, the overall irregularity of the experimental place cell maps, in a neural network model as the one there introduced, can be fully described by three variables:

1. ζ: the rate of the exponential distribution used to draw the number of fields (Sect. 11.4.1)

2. σ_d: the standard deviation of the log-normal distribution used to draw the field dimensions (Sect. 11.4.2)

3. σ_p: the standard deviation of the log-normal distribution used to draw the peak firing rates (Sect. 11.4.3)

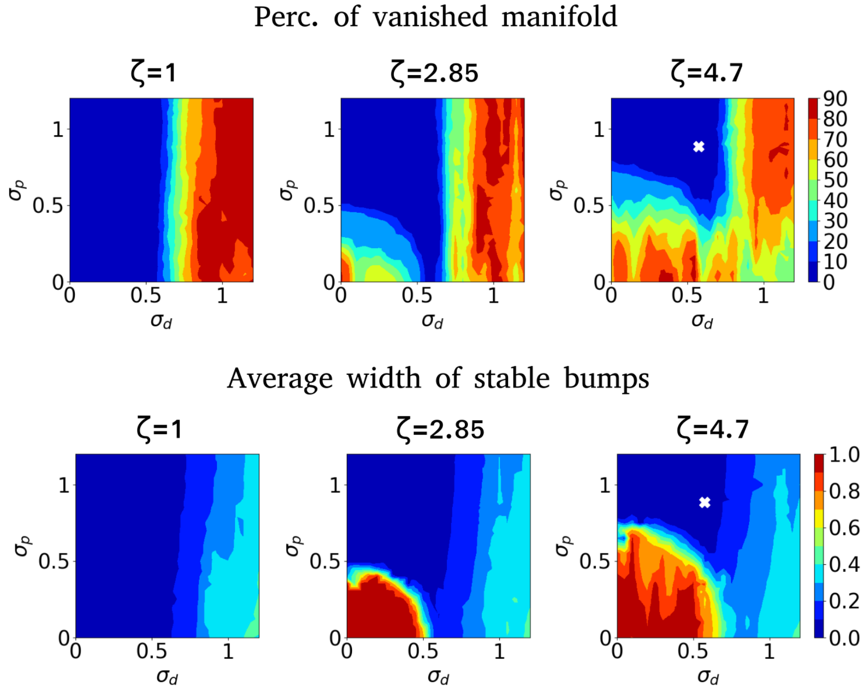


Figure 24: Phase diagrams depicting, in the σ_p-σ_d plane, for increasing average number of fields ζ, the average percentage of the quasi-attractive manifold which breaks up and the average bump width (from 0 to 1) of the fixed points (see Sect.12.1 for details over the measurements). Refer to supplementary figure 41 for the phase diagram reporting the average number of fixed points and their average sparsity. White crosses indicate the position corresponding to the distribution of the experimental results. Parameters: Data are produced following the algorithm described in Sect. 11.4 ($\mu_d = 1.57m$ and $\mu_p = 1.549Hz$ are set to model experimental results). Each plot is comprised of 26x26 data points interpolated. Each data point is averaged over $2-3$ different quenched realizations of the network. Each realization is studied simulating 50 runs of the dynamics, initialized with a different $\eta(\vec{s})$, s spanning homogeneously the whole length (one every $4m$). s is discretized into 1000 equally spaced positions every $0.2m$ and the dynamics are considered to have converged when $\sum_i |(V_i^t - V_i^{t-1})| < 10^{-10}$. The size of the system is $N = 16000$. The step size varies from $\gamma = 0.08$ to $\gamma = 0.04$, $g = 17$, $k = 300$.

The value σ_d of the transition marking the break-up of the quasi-attractive manifold seems to be independent of the other sources of variability (in the number and peak rates of the fields; and also in the connectivity if sparse). In contrast, for low σ_d and σ_p the network seems to be in a third phase, where the dynamics evolve to a non-localized configuration, lying outside the manifold (Fig. 24).

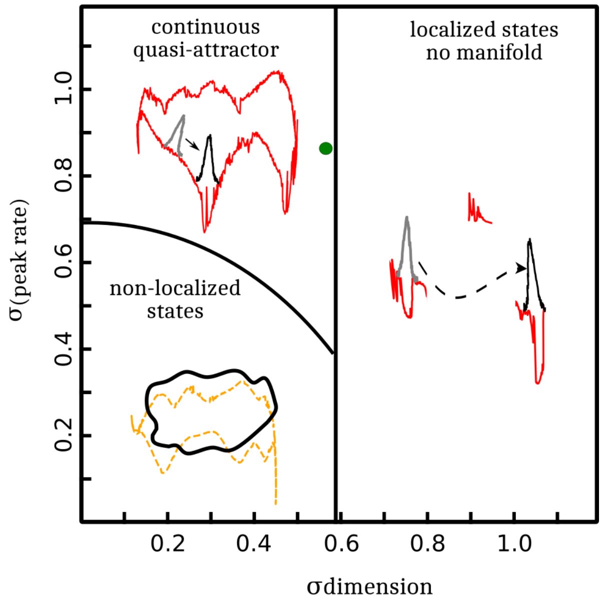


Figure 25: Rough sketch of a phase diagram, schematizing the results we obtain from the simulations illustrated in Fig. 24 for quenched random patterns modelling experimental distributions (refer to Sect. 11.4) which correspond to the green point. The red curves are intended to sketch the energy of the quasi-attractive continuous manifold $\{\vec{\xi}(s)\}$. Gray - black bumps indicate the overlap profile which one can calculate with the $\vec{\eta}(s)$ at each step of the dynamics. The orange dashed manifold, instead, represent a putative manifold $\{\vec{\xi}(s)\}$ which may exist but is not reached by the dynamics.

This can be interpreted as a somewhat counterintuitive effect of limiting the irregularity in both distributions, of field sizes and peak rates: when both are relatively regular, and each unit has several place fields, no place field prevails over the others, and the network is torn, as it were, among the different places where its active units have place fields, and fails to localize its activity. Above a region which appears to be a quarter of a circle in a $\sigma_d - \sigma_p$ plot, with radius growing with ζ, the network does localize its activity, except that if it finds itself to the right of the critical σ_d value mentioned above, it jumps to its attractor state. The delocalized phase indeed disappears when each unit has only one field on average.

The particular semi-circular geometry of this region is outlined also by the sparsity of the fixed points (see Fig. 41 in Appendix F.5) which abruptly increases within the circular region. This seems to occur, though, also for $\zeta = 1$, where states are still localized, even if abruptly more active. The boundaries of the different phases, which we observe in simulations, should be confirmed by an analytical derivation, which is still in progress and which would possibly indicate the role of the sparsity in the transitions (see Appendix H for a preliminary attempt).

The transition towards a non localized state might be due to reaching a sort of storage capacity, or more likely to a specific instability that spreads activity out from the position on the manifold indicated by the external inputs. Note, though, that the question remains, whether numerically we have explored all the appropriate configurations of parameters enabling the network to evolve towards a localized state (see Appendix F.7 for further details about parameter selection).

The overall phase diagram, for ζ set at the experimental level, is sketched in the rough schematic in Fig. 25, which we hope to derive soon analytically.

The main transition, from the presence to the absence of the persistent continuous quasi-attractive manifold seems to be general. Beyond being closely related to the characteristics of the distributions observed in one experiment [116], indeed, it holds also under more general constraints in the quenched patterns, i.e., when units are assigned individual fields of variable size (see Appendix G.1), or with diluted connectivity (see Appendix G.2).

14 DISCUSSION

While the storage of multiple regular place maps –as distinct continuous attractor manifolds in the same connectivity matrix– induces quenched disorder, fragmenting the attractor and competing with fast noise [55, 93, 129, 131, 133, 134, 144]; the storage of a unique irregular map is already enough to break the continuity of the fixed points. We see that the remaining stable states still lie on the manifold and that this turns into an effective direction of flow in the dynamical evolution. We call this mathematical object a continuous quasi-attractor. We find that there is a critical level of noise at which the quasi-attractive manifold seems to abruptly break up, through a phase transition. Applying our analysis to experimental data [116] we observe that the recordings lie at the edges of the transition.

We also find another phase, in which the network does not evolve dynamically towards a localized state, but rather activity spreads out so that it cannot be associated with a location on the manifold. This third phase appears to be present only when each unit has multiple fields, on average, and to be delimited by a circular boundary in the σ_d, σ_p plane.

This result lead us to hypothesize that place cell maps, like those observed in these bat recordings, may be effectively memorized in a quasi-attractive continuous manifold. If the maps are precisely regular than the quasi-attractive manifold becomes a standard continuous attractor. If, instead, they are irregular, the memory function is still preserved and persists up to a critical level of noise σ_d.

In this hypothesis individual memories would be unstable bumps and efficient coding of position could be possible either on a timescale between the attraction and the drift, as previously hypothesized [131], or if some biologically plausible mechanisms, such as those already explored [60, 131, 135–137, 139], could increase the tendency of the bump to stay still.

The three phases of the diagram which we find through numerical simulations, should be derivable through analytical calculations. Our preliminary work in progress in this direction can be found in Appendix H.

Part IV

DO WE "REMAP" BETWEEN LANGUAGES?

Vowels can be approximately described by the first two formants of their sound spectrum. May the brain learn vowels through a two-dimensional cognitive map, much as an animal acquires a cognitive map of an environment, expressed by place cells? If so, are the standard vowels of different languages stored in different vowel maps and are multilingual people remapping when changing language? In this ongoing work, we model the perception of vowels as the convergence towards fixed points on a two dimensional manifold.

15 INTRODUCTION

Understanding a different language, means, at first, differentiating its *speech sounds* or *phones*. While this can be difficult even among native speakers used to different dialects, it becomes hard when one has never heard the other language before. In this case, it seems as if the mental representation people have of their own languages hampers the detection of a new sound by driving it towards an already known *phoneme*. How does this translate neurally? Do we create *phone* cognitive maps, as those rodents, for example, are thought to create for physical space? Do we create cognitive maps for any relevant quantity that can be described as varying in a continuous two-dimensional space?

Vowels offer the perfect temptation to approach such questions theoretically.

15.1 VOWELS IN TWO DIMENSIONS

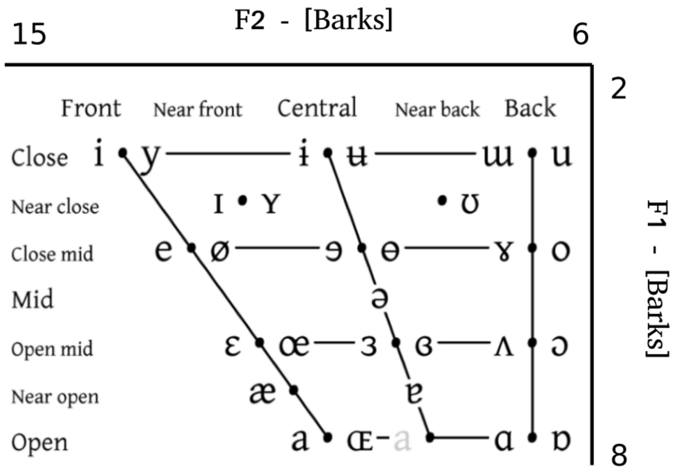


Figure 26: International phonetics alphabet vowel charts. F1 and F2 can be measured in Barks, which are non-linearly related with Hertz through a formula which can be found in Ref. [145, 146]. Barks values are merely indicative, however; each person's vowels can be thought of as shifting and slightly deforming the trapezoid in one way or another, see for example Fig. 27.

Phonemes are the "*smallest identifiable units found in a stream of speech that can be transcribed*" [147] and are usually subdivided into vowels and consonants. One could argue that *phonemes*, therefore, correspond to the smallest identifiable units which we are able to consciously categorize. How does our brain perceive *phones* and, consequently, categorize them into *phonemes*?

A few studies indicate that vowels and consonants are independently coded in the brain [148, 149]. The impairment of one system does not seem to determine an impairment in the other. One can thus focus either on vowels either on consonants in order to gain some specific hints over the mental representation of phonemes.

Here, we will focus on vowels, as they can be differentiated with notable simplicity through two physical, continuous, quantities.

If one records vowels, decomposes their sound spectrum into frequencies, takes the first two resonant components above the fundamental frequency (F1, F2), generally called formants, and uses them as coordinates to represent the vowels in a 2-dimensional plane, one sees that each standard vowel occupies, as a small cloud, a specific position on such (F1, F2) Euclidean space [150]. This is known as a *vowel chart*, and, remarkably, the two dimensions are simply related to two physical quantities [151, 152]. In particular, F1 roughly measures "how open is" and F2 "where is the main occlusion of" the vocal tract while pronouncing the vowel. This is not true for consonants, whose categorization seems to be more complex [146, 153, 154]. Fig. 26 shows the international phonetics alphabet vowel chart while Fig. 27 shows three vowel charts we estimated from recordings.

15.1.1 *Vowels as place fields?*

Given the intriguing possibility of categorizing vowels in two dimensions through an acoustically well defined procedure, it is straightforward to wonder whether our brain uses a similar code. The experimental results in the domain of spatial cognition (Sect. 2), especially those on place cells and their *remapping*, indicate a possible exploratory direction. Indeed, beyond the two dimensional space another indication of potential correspondence seems to exist between place fields and vowels: auditory *place coding* through *tonotopy* [155]. The perception of a sound initially occurs through specific hairs cells in the inner ear which move in response to specific frequencies. In particular, the whole cochlea can be described as comprising a one-dimensional frequency map, with a gradual progression of hair cells sensitive to increasing frequencies [156, 157]. Hair cells do not fire action potentials on their own, they generate a receptor potential through a cascade of events which induces a spike in the connected neuron [158]. The signal then navigates from one synapse to the other

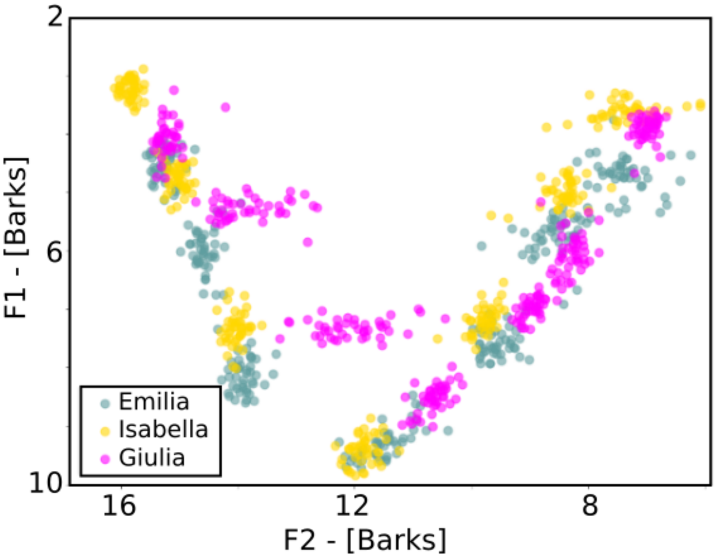


Figure 27: Italian vowel charts of three different young women (color coded) based on the formants extracted from a number of repetitions of each vowel. One can appreciate that each chart is different. Differences can perhaps be understood as related to the bilingualism of two subjects, as will be mentioned later. F1 and F2 are measured in Barks.

eventually reaching the cortex, where the situation is less easy to describe. However, at least at the first stages of the auditory pathway, there is good evidence that specific neurons encode specific frequencies.

One could thus imagine an abstract two dimensional (F1, F2) Euclidean sheet, where each neuron has a field at one (x,y) position. What if each familiar vowel can be thought of as a fixed point on such a plane learned through a Hebbian rule, and then reached via recurrent dynamics? In the following chapter we carry out some preliminary studies on this hypothesis.

In this part of the thesis the main results were obtained with Emilia Cortesi, a Master student I co-supervised from May to December 2020.

16 A CONTINUOUS QUASI-ATTRACTOR STORING VOWELS

16.1 MODEL DESCRIPTION

We consider a fully connected network of $N = 2500$ threshold linear units, each selective to one or more fields in a putative $(F1, F2)$ square chart of 11x11 Barks^2. We discretize this Euclidean $(F1, F2)$ space into $S = 2500$ equidistant positions lying at the vertices of a two-dimensional grid with coordinates $\vec{s} = (F1, F2)$. We consider each unit to be selective to one or more specific locations on the plane according to a Poisson distribution with average value $\lambda = 4$, arbitrarily selected as a hypothesis of realistic variability. The selectivity is taken to realize a place-like code, thus, the firing map of each unit has a number of Gaussian bumps (of standard deviation $\sigma \approx 0.25$ Barks) centered at random positions. Based on the field selectivity we create a template matrix $\{\vec{\eta}\}_{\vec{s}}$, where $\vec{\eta}_{\vec{s}}$ is the population activity of all units at position $\vec{s}$, while $\eta_i(\vec{s})$ is the firing rate map of unit i sampled at all discretized positions $\vec{s}$.
We then construct the connectivity matrix through Hebbian learning as

$$J_{ij} = \frac{1}{S}\sum_{\vec{s}}(\eta_i(\vec{s}) - \langle\eta\rangle)(\eta_j(\vec{s} - \langle\eta\rangle) \qquad (120)$$

and obtain a semi-continuous attractor.
Remarkably, the attractor is *semi*-continuous, not only because the plane is discretized but also because units have multiple selectivity. As shown for the 1d case in Part iii, indeed, when patterns are irregular, be it in the activity profile of different units or in the density of fields at each position, the irregularities act as quenched noise in the connectivity matrix. The retrieved bumps slide on the manifold –here 2-dimensional– which persists, in the sense of remaining attractive as a manifold, up to a certain noise toleration, after which it disappears. While we did not carry out a systematic study in two-dimensions for the exploratory model we just described, we evaluated the residual number of fixed point configurations of activity. We consider the system to realize a semi-continuous attractor, if its activity evolves towards one of several bump profiles, in terms of overlaps with the templates $\eta_i(\vec{s}$'s, which satisfactorily tile the (F1,F2) plane.

We consider this initial connectivity matrix as what one individual would have before learning any language.

16.1.1 *Training procedure*

We mimic phoneme learning as follows:

- we record a vowel pronounced by a native speaker (refer to Appendix I.1 for details over the method), extract the formants $(F1, F2)$, estimate what would be the template in that exact position $\vec{\eta}_{(f1,f2)}$, given the field selectivity of the neurons, and set it as the initial condition of a dynamical evolution;

- we evolve the dynamics synchronously as

  $$\vec{V}^{t+1} = g\left[\bar{J}\vec{V}^{t} - Th\right]^{+} \tag{121}$$

  where $\Theta(\cdot)$ is the Heaviside step function. The average activity is kept fixed through the gain g, and the sparsity through the threshold Th, equal for all units.

- when the dynamics reaches a fixed point we use the final configuration $\vec{V}^{fp}$ to update the connectivity matrix as

  $$J_{ij}^{new} = J_{ij}^{old} + \gamma(V_i^{fp} - \langle V^{fp}\rangle)(V_j^{fp} - \langle V^{fp}\rangle) \tag{122}$$

  with γ being the learning rate, set at 10^{-4}.

- we do the same for all other standard vowels and repeat the procedure about 700 times, repeatedly using all (roughly 25) realizations of each vowel provided by the same subject.

16.1.2 *Probing learning*

We do not define *a priori* where the learned vowel should lie in the $(F1, F2)$ plane. We probe learning through the following procedure, illustrated for a paradigmatic example, shown in Fig. 28:

- As introduced in Sect. 16.1.1, after creating the basic-connectivity matrix, we provide vowels from a native speaker to the learning algorithm (Fig. 28a).

- We run S simulations of the dynamics, with S the number of discretized coordinates $\vec{s}$, each having as initial condition one specific $\vec{\eta}_{\vec{s}}$. We count the number of different fixed points towards which the dynamics evolve and for each we estimate the overlap with all the $\vec{\eta}_{\vec{s}}$s as defined in Eq. (3) to check it satisfactorily tile the (F1,F2) plane. The positions with maximal overlap for each fixed point vector is plotted as a black stars in Fig. 28b.

- We run a number of simulations having as initial stimulus (i.e. as initial condition) one recorded vowel from the same native speaker used for the training and evaluate the fixed configuration towards which the dynamics converge. This is represented in Fig. 28b as the arrows starting from the color coded vowels.

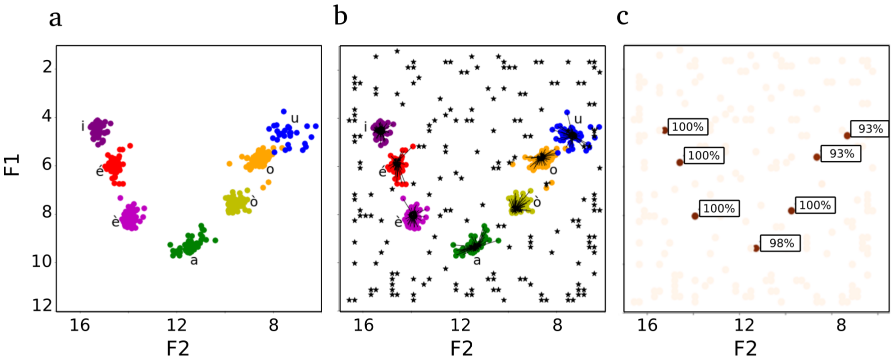


Figure 28: Example of a network learning Italian vowels. a) Formants of vowels recorded from a native speaker. b) fixed points plotted as stars over the positions corresponding to the templates with which the overlap is maximal. Arrows indicate the fixed configurations towards which the dynamics evolve if initialized from recorded vowels. c) Maximal percentage of a specific category of vowels falling into each attractor. F1 and F2 are measured in Barks.

- We evaluate the percentage of vowels which fall into one or another attractive configuration of activity (Fig. 28c).

16.2 THE VOWEL CHARTS OF BILINGUALS CAN BE STORED WITHIN ONE MAP

Having established that a semi-continuous attractor neural network can learn vowels, from a theoretical point of view, storing them as separate fixed points, one can wonder

- whether something similar is occurring in the brain
- if so, whether we *remap* from one space to another when switching languages.

As a first theoretical exploration, we have considered three case studies involving bilingual young women of similar ages: a Hungarian native speaker who has proficiently learnt Japanese and two bilinguals from birth, a native Italian and Swedish speaker (It-Sv) and a native Italian and Chinese speaker (It-Zh); from here onward they will be called simply Judit, Isabella and Giulia, respectively.
Fig.29 shows the conjunctive vowel chart of the subjects. Two preliminary observations can be tentatively made:

1. The vowel charts of one language may be influenced by the other one. The Italian vowel chart in b) (Isabella) is wider than the one in c) (Giulia), as if the Chinese vowels, mostly lying at low F2, drive the fixed points towards the right.
2. Some vowels, whether pronounced within one or the other language are indistinguishable, as for "a" or "o" in c).

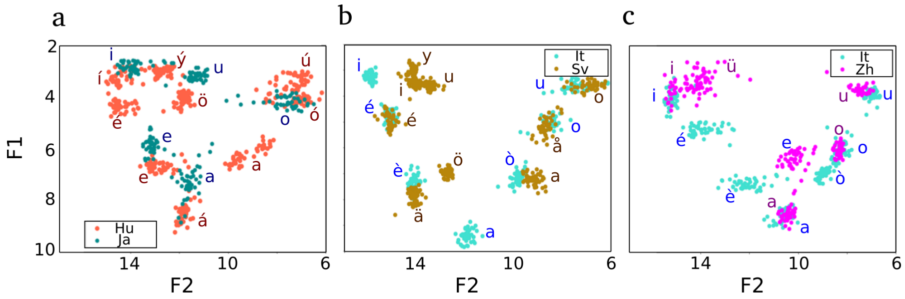


Figure 29: Conjunctive vowel charts of three bilingual subjects. a) Judit: Hungarian-Japanese (Hu-Ja) b) Isabella: Italian-Swedish (It-Sv) c) Giulia: Italian-Chinese (It-Zh). The vowel corresponding to each cloud of points is written nearby. For an better visual distinction between the different realizations of the same vowel, we refer to the separate color-coded plots in the Appendix I.2. F1 and F2 are measured in Barks.

If cognitive vowel charts provide a useful model, these observations could lead to hypothesize that the chart is unique for the two languages of these bilingual subjects, in agreement with previous linguistic results on English-Italian bilinguals [159]. Independent charts, indeed, would not be expected to induce movement in the position of the vowels of either language, as unit selectivity would be effectively orthogonal from one language to the other.

From a computational point of view, instead, we have tentatively confirmed that, in principle, with a network model as the one introduced in the previous section, all vowels presented in each subplot of Fig. 29 can be learned as separate fixed points within one map through Hebbian learning, and retrieved with a proper stimulus.

16.3 EXPERIMENTAL PARADIGM AND CONCLUSION

The preliminary explorations mentioned in this part of the thesis indicate that what remains of a continuous attractor neural network when asymmetries are introduced can model vowel learning, memorization and retrieval within a putative individual cognitive map. Experimental evidence indicating whether any analogous phenomenon is occurring in the real brain is, however, generally lacking. An experimental paradigm defined by Kaya et al [160] indicates a possible testing procedure. In the study, the authors consider bilingual subjects and employ a vowel-confusion paradigm [161]: subjects are induced to place themselves in a mental language environment, as it were, by listening to a story narrated in one of the two languages. Then, together with some binary (yes/no) questions about the story, subjects are asked to answer if pairs of artificial sounds –corresponding to specific $(F1, F2)$ positions– are different. How well, in terms of correct responses, the subjects perform in differentiating two *phones*, is then translated through an algorithm in how perceptually distant the phones are in the subject putative $(\mathsf{F1}, \mathsf{F2})$ cognitive map. Whether such perceptual map differs when subjects are induced to be placed in the other language environment will give crucial information regarding the hypothesis that people "remap" when switching languages. The experimental study is currently being carried out and its forthcoming results, combined with the computational approach here illustrated, will possibly lead to some preliminary conclusions about the mechanisms underlying phoneme cognition.

Beyond vowels, these results could also contribute to understanding whether quasi-attractive manifolds, i.e. what remains of continuous attractor neural networks when irregularity is introduced (see Part iii), could be collective neural behaviours generally emerging in the brain to encode (or place-code) continuous irregular variables.

GENERAL DISCUSSION

The studies presented in this thesis wish to contribute, through two interconnected itineraries and one exploratory application, to the search of a theoretical understanding embracing irregularity and dishomogenities in memory storage.

Memories are generally conceptualized as attractors of the neural dynamics. In particular, two broad classes of attractors, i.e. discrete and continuous ones, are particularly adopted as useful notions in Neuroscience. The first ones are intended to represent uncorrelated and separate memories. The second ones, instead, are useful to describe spatial memory, in particular spatially selective cells.

In our first study, described in Part ii, we consider discrete attractors, and show, through analytical calculations, that the difference between the maximal number of patterns which can be stored in an optimally connected network, in contrast to one connected through a biologically plausible learning strategy, decreases (or even changes sign) when one considers biologically plausible units as compared to binary ones. Optimal connections can be approached through modern machine learning strategies, which iteratively optimize the weights through computationally intense back-propagation algorithms. Biologically plausible connections, instead, are those emerging in an active network following the simple Hebbian learning rule.

Specifically, we see that the storage capacity evaluated *à la Gardner* (i.e of a perceptron) is reached and even surpassed by that obtained with *Hebbian* learning [57] in networks aiming at resembling brain responses, i.e. when the transfer function is *Threshold linear* (or ReLu) instead of step-like. We see that this increase seems to be made possible by an intrinsic feature of non-binarity: a code instantiated by threshold linear units η_i can vary its sparsity $a = \frac{\langle\eta\rangle^2}{\langle\eta^2\rangle}$. When one imposes optimality in the connections one usually analytically considers perfectly retrieved memories, while when one considers Hebbian learning, patterns can be retrieved with errors (the quenched variables are indeed the patterns). The non-optimal but biologically plausible Hebbian learning strategy leads to an emergent re-organization of the activity code which retrieves sparser patterns, composed by those units which were more active in the memory. We find that such trend, which can not occur in the binary reduction as all units are equally active, is the fundamental source of the apparent violation of the storage capacity. We also explore the nature of the retrieval to non-retrieval transition in Hebbian learning and see that the transition shifts from first to second order depending on the detailed distributions expressed by the threshold linear units.

Threshold linear units are essential also in our second study. Neurophysiological recordings of spatially selective cells in rodents and other species, in the past decades and especially in the past few years, particularly when carried out in quasi-ecological settings, have highlighted an intrinsic irregular and chaotic nature of such neuronal responses, irreducible to simple geometrical principles. These irregularities take the form of fields spanning different sizes and numbers but also different peak firing rates. In Part iii of the thesis we study how irregular and complex spatial memory may be accounted for in a continuous attractor theory. We show that while continuous attractors break when considering irregularities, i.e. the number of fixed points on the manifold decreases from an infinite set to a very small number, the few remaining fixed points still lie on the manifold, which becomes a direction of flow, persisting up to a transition. In the standard continuous attractor neural network perspective the number of fixed points on the manifold somehow relates to and quantifies the *memory function* quality: each fixed point is considered as the memory of a specific spatial location and the fewer the fixed points the less an environment is considered to be memorized. While small noise causes limited "wrinkles" in the continuous attractors and the resulting mathematical objects were previously studied as a good approximation of a well behaving continuous attractor, irregularities such as those observed in recent experiments (see Sect. 9) are way to large to fit this characterization.

We show, however, that one may consider as memory, instead of the fixed points, the manifold itself, even if comprised mostly of asymptotically non-fixed points, and one can consider retrieval as occurring at a specific time scale, or the bump to be kept on the manifold by some external input. We call this object a *continuous quasi-attractive manifold* and show that it persists as stable even when considering large scale dishomogenities, thus retrieving irregular place maps up to a critical level of noise, at which it breaks up. We obtain numerically a phase diagram, considering the variation of the continuous quasi-attractor while making the distributions of field size and peak rate increasingly variable. We see that real irregular recordings [116] lie just before the described transition. Further we see that another phase appears, where the activity is not localized. We are currently studying how to derive the boundaries analytically.

While in Part ii of this thesis we show, considering discrete attractors, that Hebbian learning can surpass the Gardner bound at the cost of sparsifying the retrieved patterns, in Part iii we see numerically that the more irregular are the maps to be memorized, in a continuous quasi-attractor, the more only a few units remain active, in the discrete fixed points. One can infer from this comparison that the more maps are irregular, the more the storage of each position in space resembles one of the discrete attractors, thus enhancing the storage capacity of the network.

In the last part of the thesis, presented in Part iv, we explore the neural bases of phonology and hypothesize that vowel charts may be stored, in the brain, through continuous quasi-attractors. We define the problem, run some preliminary simulations and experiments a testing procedure.

The calculation of the Gardner bound for Threshold linear units and the analysis of continuous quasi-attractive manifolds highlight the role of sparsity in regulating memory storage in the brain, in different systems, possibly including phonology.

Part V

APPENDIX

A MAIN ANATOMICAL TRAITS OF THE HIPPOCAMPAL SYSTEM

The hippocampal system, a brain region situated in the medial temporal lobe, can be subdivided in several areas, and first in two main regions, the hippocampal formation and the parahippocampal region [162], which can be differentiated by their gross cytoarchitectonic organization. The hippocampal system is highly similar in different mammalian species and here we give a short overview focusing on rodents.

A.1 THE HIPPOCAMPAL FORMATION – WITH PLACE CELLS

The hippocampus proper, or cornu ammonis (CA) has pyramidal principal cells in one layer – a cortical structure called allocortex – and is further subdivided in a sequence of three areas, CA1, CA2 and CA3, with remarkably distinct connectivity between them. It is flanked on the input end by the dentate gyrus, or DG, which evolves out of the same type of cortex but with small granule cells instead of pyramidal cells, and on the output end by the subicular complex, which, in as many as 5 internal subdivisions [163], links the hippocampus to the adjacent multi–layer cortex. Place fields have been found throughout the hippocampal formation and have been studied especially in CA1 and CA3. For a long time, in fact, it was puzzling how place cells in the two subfields looked so similar, apart from minor statistical differences, when, instead the circuitry is so different: CA3 is dominated by recurrent connections, unlike CA1, and the main afferent connections to CA3 are from the DG granule cells and from Entorhinal Cortex layer II, unlike those to CA1 which are from EC layer III and from CA3 itself (see Fig. 30).

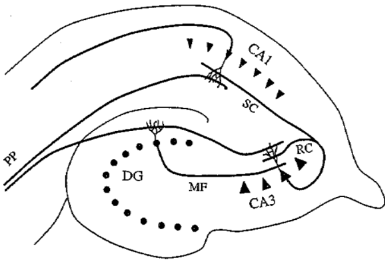


Figure 30: Schematic representation of the connectivity between three main regions of the hippocampus: DG, CA3 and CA1.

A.2 THE PARAHIPPOCAMPAL REGION – WITH ALSO GRID CELLS

The parahippocampal region is characterized in part as periallocortex, to emphasize its transitional nature to fully neocortical structure with multiple layers of principal cells. It is formed by the entorhinal, perirhinal, postrhinal cortices and by the components of the subicular complex, that some prefer to view separately from the subiculum proper. The medial subdivision of the entorhinal cortex (mEC) has risen to particular prominence after the discovery of grid cells, somewhat obscuring the fact that most of its principal cells do not conform to the grid cell stereotype even in standard laboratory settings, nor do those of the other parahippocampal areas. At the system level, perirhinal cortex makes afferent connections to lateral EC that do not appear to convey fine spatial information, unlike the connections from postrhinal cortex to mEC. Grid cells emerge, in this perspective, as one form of refinement of spatial information before it is merged with nonspatial information in the hippocampus, where both lEC and mEC project, and largely transformed into a place cell code, at least in rodents.

A.3 THE ENTORHINO-HIPPOCAMPAL CIRCUITRY

Principal cells from EC layer II reach DG and CA3, while principal cells from EC layer III reach CA1. Internally in the hippocampus, activation propagates in a sort of one-directional loop, with recurrence (in CA3) and shortcuts. DG granule cells project their so-called mossy fibers to CA3, where they make sparse but powerful synapses on the apical dendrites close to the cell body of CA3 pyramidal cells. Since the same CA3 cells receive many more (but weaker) synapses on their distal apical dendrites from the same fibers originating in EC layer II that, *en passant*, connect to the granule cells, a major riddle has been to understand this apparent duplication of the information arriving to CA3, directly and, as it were, translated by the DG. A more recent question involves CA2, which had long been regarded merely as a small transition region between CA3 and CA1; recent evidence on a potentially important role in social cognition [164] has been accompanied by the observation of CA3-like anatomical features in CA2, such as prominent recurrent collaterals [165] and the formation, perhaps in pathological conditions, of mossy synapses [166]. Feedforward connections from CA3 to CA1 (the Schaffer collaterals) and from CA1 to subiculum are also intriguingly combined, in what may be called a heteroassociative architecture, with EC layer III inputs to these two regions. Fibers then project back from CA1 and subiculum to EC layers V and VI.

B

DERIVATION OF THE LIMITS OF THE GARDNER CAPACITY

From Eq. 54 of the main text it is possible to evaluate the two limits of very sparse and non-sparse coding. First, a simple substitution at $f = 1$ leads to

$$x = -\frac{d_1}{g\sqrt{d_3}} \tag{123}$$

$$\alpha_c^{-1} = 1 + \frac{1}{g^2}. \tag{124}$$

The limit $f \to 0$ is a bit trickier. We first rearrange the first equation in Eq. (3) as

$$\frac{f}{1-f} = \frac{1}{(x + \frac{d_1}{g\sqrt{d_3}})} \int_x^\infty Dt(t-x) = \frac{1}{(x + \frac{d_1}{g\sqrt{d_3}})} \Big(\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} - x \int_x^\infty Dt \Big) \tag{125}$$

As f goes to zero, for the left hand side to be equal to the right hand side, we should have $x \to \infty$. We therefore use the expansion

$$\int_x^\infty Dt = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \Big[\frac{1}{x} - \frac{1}{x^3} + \mathcal{O}\Big(\frac{1}{x^5}\Big) \Big]$$

to write the right hand side of Eq. (125) as

$$\frac{f}{1-f} \approx \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi} x^3}. \tag{126}$$

We find a solution to Eq. (126) through the following iterative procedure. We first solve the leading term for $f \to 0$ in $x \to \infty$ namely

$$f \approx \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.$$

yielding

$$x \approx \sqrt{2 \ln \Big(\frac{1}{\sqrt{2\pi} f} \Big)} \tag{127}$$

We then insert x from Eq. (127) into $\exp(-x^2/2) = \sqrt{2\pi}fx^3$ to obtain the logarithmic correction

$$
\begin{aligned}
e^{-\frac{x^2}{2}} &\approx \sqrt{2\pi}fx^3 \\
x &\approx \sqrt{2\ln\left(\frac{1}{\sqrt{2\pi}fx^3}\right)} \\
x &\approx \sqrt{2\ln\left(\frac{1}{\sqrt{2\pi}f}\right)\left(1-\frac{\ln x^3}{\ln\frac{1}{\sqrt{2\pi}f}}\right)} \\
&\approx \sqrt{2\ln\left(\frac{1}{\sqrt{2\pi}f}\right)}\left(1-\frac{3}{4}\frac{\ln\left(2\ln(\frac{1}{\sqrt{2\pi}f})\right)}{\ln\frac{1}{\sqrt{2\pi}f}}\right).
\end{aligned}
\tag{128}
$$

where in the last passage we have used the Taylor expansion of the square $\sqrt{1-y} = 1 - \frac{y}{2} + \mathcal{O}(y^2)$ around $y = 0$ as for $f \to 0$, $\frac{\ln x^3}{\ln\frac{1}{\sqrt{2\pi}f}} \to 0$.

We have tested numerically that the above expression Eq. (128) for x is indeed a solution to Eq. (125) for $f \to 0$.

We now proceed to evaluate α_c and we apply the same Taylor expansion as before

$$
\begin{aligned}
\alpha_c &= \left\{f[\langle((x+\frac{\xi_i}{g\sqrt{d_3}}))\rangle^2+1]+(1-f)\int_x^\infty Dt(t-x)^2\right\}^{-1} \\
&= \left\{f[\langle((x+\frac{\xi_i}{g\sqrt{d_3}}))\rangle^2+1]+\right. \\
&\left.+(1-f)\left(-\frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}}+(1+x^2)\int_x^\infty Dt\right)\right\}^{-1} \\
&\approx \left\{fx^2-\frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}}+\frac{(1+x^2)}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\left(\frac{1}{x}-\frac{1}{x^3}+\frac{3}{x^5}\right)\right\}^{-1} \\
&\approx \left\{fx^2+\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\left(-x+\frac{(1+x^2)(x^4-x^2+3)}{x^5}\right)\right\}^{-1} \\
&\approx \left\{fx^2+\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\left(\frac{2x^2+3}{x^5}\right)\right\}^{-1} = \left\{fx^2+\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{x^2}{2}}}{x^3}\right\}^{-1}.
\end{aligned}
$$

To summarize, in the limit $f \to 0$ we obtain

$$
\begin{cases}
x \approx \sqrt{2\ln\left(\frac{1}{\sqrt{2\pi}f}\right)}\left(1-\frac{3}{4}\frac{\ln\left(2\ln(\frac{1}{\sqrt{2\pi}f})\right)}{\ln\frac{1}{\sqrt{2\pi}f}}\right) \\
\alpha_c \approx \left\{fx^2+\sqrt{\frac{2}{\pi}}\frac{e^{-\frac{x^2}{2}}}{x^3}\right\}^{-1}.
\end{cases}
\tag{129}
$$

Substituting x in α_c to the leading order leads to the limit reported in the main text at page 41.

C DETAILS OF THE TL PERCEPTRON TRAINING ALGORITHM

For the purpose of assessing whether the Gardner capacity for errorless retrieval can be reached with explicit training, we can decompose a network of, say, $N+1 = 10001$ units into $N+1$ independent threshold linear perceptrons. A threshold linear perceptron is just a 1-layer feedforward neural network with N inputs and one output, the activity of which is given by a threshold-linear activation function.

$$[h]^+ = \max(0, h) \tag{130}$$

The network is trained with p patterns. One can then think of the input as a matrix $\bar{\xi}$ of dimension $[N \times p]$ and of the output as a vector $\vec{\eta}$ of dimension $[1 \times p]$.

The aim of the algorithm is to tune the weights such that all p patterns can be memorized. In order to tune the weights we start from an initial connectivity vector $\vec{J}_0$ of dimension $[1 \times N]$ and estimate the output $\vec{\hat{\eta}}$ as:

$$\begin{aligned} \vec{h} &= \vec{J}\bar{\xi} \\ \vec{\hat{\eta}} &= g[\vec{h}]^+ \end{aligned} \tag{131}$$

where g is the gain parameter. We then compare the output $\vec{\hat{\eta}}$ with the desired output $\vec{\eta}$ through the loss function

$$L(\vec{\hat{\eta}}) = \sum_{\mu=1}^{p} \frac{1}{2}(\hat{\eta}^{\mu} - \eta^{\mu})^2. \tag{132}$$

The TL perceptron algorithm can be seen as simply a stripped down version of *back-propagation*, for a 1-layer network: the weights $\vec{J}$ are modified by gradient descent to minimize the loss during the steps $k = 1..k^{MAX}$ where k^{MAX} is the number of steps needed for the gradient descent in order to reach the minima $\frac{dL(\vec{J}_k)}{d\vec{J}_k} = 0$. If at the minima $L(\vec{J}_{k^{MAX}}) = 0$ at least a set of weights exists for errorless retrieval at that p value. The storage capacity $\alpha_c = \frac{p^{max}}{N}$ is evaluated by estimating p^{max} as the highest p value enabling to reach $L(\vec{J}_{k^{MAX}}) = 0$.
Initializing the weights around zero facilitates reaching the minima. The chain derivative that in general implements gradient descent in back-propagation, in this case reduces to

$$\vec{J}_{k+1} = \vec{J}_k + \gamma \frac{g}{p}(\vec{\eta} - \vec{\hat{\eta}})\Theta(\vec{\hat{\eta}})\bar{\xi}^{T} \tag{133}$$

where $\Theta(\vec{\eta})$ is the Heaviside step function applied to all N elements of $\vec{\eta}$ and where γ is a learning rate. Note that the gain g, appearing as a multiplicative factor both in Eq. (133) and Eq. (131) is performing a similar role as the learning rate, with which it can be tuned.

In the simulations presented in Fig. 14 of the main text, in order to obtain the results shown by red diamonds we have used $N = 100$ units, $g = 1$ and binary patterns. For each value f, we increase p and check whether when the connectivity matrix stops changing, we have $L(\vec{\eta}) = 0$. We take p_{max} as the largest value of p for which this is possible for at least a set of random initial weights. As for the learning rate, we use a decreasing scheduling, initially set to $\gamma = 0.2$. As the minimization progresses, some weights stop changing while the others keep changing and therefore when there are only a maximum of 5 weights changing, we decrease the learning rate to $\gamma = 0.02$, and finally in later iterations when the number of still varying weights reduces to 2 we use $\gamma = 0.002$. The initial condition of the weights are drawn from a Gaussian distribution of mean $\mu = 0$ and $\sigma = 10^{-2}$ and each (f, p) combination is tested at least from 20 random initial weights, each for a random data sample. In Fig. 14 of the main text we restrict our analysis to $N = 100$ and $f \geqslant 0.05$ for numerical limitations. Decreasing f implies on one side increasing the number of patterns, thus making the process slower; on the other side it reduces the possibility of finding non-zero values for finite N. Increasing N in order to find non-zero values requires increasing p accordingly, making the process even slower.

According to the analytical calculations, the same f dependence of the capacity found for binary patterns should also hold for other distributions. In Fig. 14 we therefore also show the numerical experiments for input patterns taken at random from the ternary distribution $P(\eta) = (1-f)\delta(\eta) + \frac{f}{2}\delta(1-\eta) + \frac{f}{2}\delta(2-\eta)$ (green crosses in Fig. 1); the same distribution is also used for the outputs. The numerical results are consistent with the analytical results.

D

FURTHER EXPLICIT DERIVATIONS WITH HEBBIAN LEARNING

D.1 THE MATHEMATICAL FORMS OF THE BINARY, TERNARY, QUATERNARY AND EXPONENTIAL DISTRIBUTIONS USED IN THE MAIN TEXT

In chapter 6, we have compared capacity values using a binary, ternary, quaternary and an exponential distribution:

$$p(\eta) = (1-a)\delta(\eta) + a\delta(1-\eta) \tag{134}$$

$$p(\eta) = \left(1 - \frac{9a}{5}\right)\delta(\eta) + \frac{3a}{2}\delta\left(\eta - \frac{1}{3}\right) + \frac{3a}{10}\delta\left(\eta - \frac{5}{3}\right) \tag{135}$$

$$p(\eta) = \left(1 - \frac{9a}{4}\right)\delta(\eta) + \frac{3a}{2}\delta\left(\eta - \frac{2}{9}\right) + \frac{3a}{5}\delta\left(\eta - \frac{5}{9}\right) + \tag{136}$$

$$+ \frac{3a}{20}\left(\eta - \frac{20}{9}\right) \tag{137}$$

$$P(\eta) = (1-2a)\delta(\eta) + 4a\exp(-2\eta) \tag{138}$$

One can see that all distributions are such that $\langle\eta\rangle = \int_0^\infty d\eta P(\eta)\eta = a$ and $\langle\eta^2\rangle = \int_0^\infty d\eta P(\eta)\eta^2 = a$, so that a coincides with the sparsity $\langle\eta\rangle^2/\langle\eta^2\rangle$ of the network. The fraction of active units is thus related to a as $f = a, 9a/5, 9a/4, 2a$ respectively.
One can also easily see that

$$\begin{aligned} A_2^{\text{binary}}(w,v) &= \frac{a}{v}\left[-w\phi(w) - \sigma(w) + \left(w + \frac{v}{a}\right)\phi\left(w + \frac{v}{a}\right) + \right. \\ &\left. + \sigma\left(w + \frac{v}{a}\right)\right] \\ A_3^{\text{binary}}(w,v) &= (1-a)[(w^2+1)\phi(w) + w\sigma(w)] + \\ &+ a\left\{\left[\left(w + \frac{v}{a}\right)^2 + 1\right]\phi\left(w + \frac{v}{a}\right) + \left(w + \frac{v}{a}\right)\sigma\left(w + \frac{v}{a}\right)\right\} \end{aligned} \tag{139}$$

and the same can be explicitly defined also for the ternary and quaternary distributions. For the exponential one, instead, we derive it analytically in the following section.

As a supplement to Fig. 14 in Chapter 6, reproduced here in the 3 separate panels in the upper row in Fig. 31, we show a comparison between the Hebbian capacity and the Gardner one when plotted as a function of the output sparsity (in the bottom row of Fig. 31). The Gardner storage capacity is now in each of these 3 cases above the Hebbian capacity, taken as a function of the output sparsity instead of the input one.

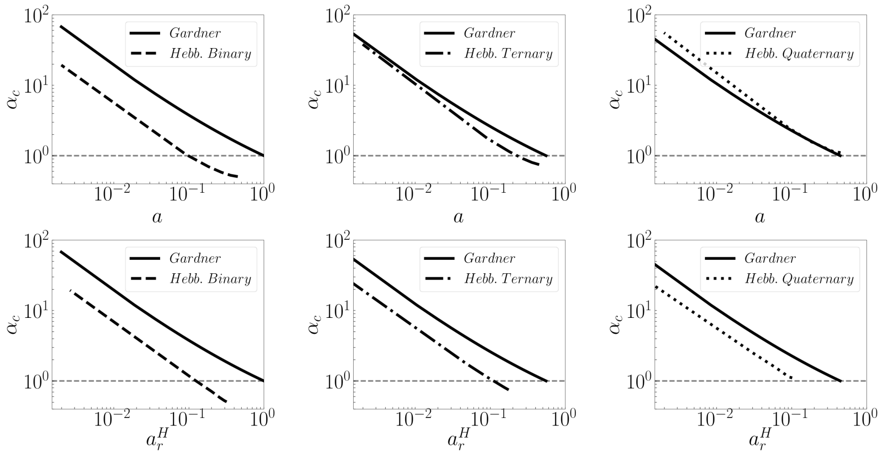


Figure 31: Supplementary to Fig. 14 in Chapter 6. Comparison between the Hebbian and Gardner storage capacity for 3 discrete distributions. The upper row considers as sparsity parameter the one of the input pattern, the lower row the one of the retrieved pattern. The Garner capacity is that given by Eq. (53) in Sect. 5.2

D.2 ANALYTICAL DERIVATION OF THE EXACT HEBBIAN STORAGE LIMIT FOR THE EXPONENTIAL DISTRIBUTION

Here we derive the explicit form of the expression for A_2 and A_3 in Eqs. (69) and (70), introduced of Sect. 6.1 for an exponential distribution of the patterns. In general, for A_2 we write

$$\begin{aligned} A_2 &= \frac{a}{v(1-a)} \int_0^\infty d\eta P(\eta)(\frac{\eta}{\langle\eta\rangle}-1) \int_{-\infty}^{x(\eta)} Dz(x(\eta)-z) \\ &= \frac{a}{v(1-a)} \Big\{ \int_w^\infty Dz \int_{\frac{(z-w)\langle\eta\rangle}{v}}^\infty d\eta P(\eta)(\frac{\eta}{\langle\eta\rangle}-1)(x(\eta)-z) + \\ &+ \int_{-\infty}^{w} Dz \int_0^\infty d\eta P(\eta)(\frac{\eta}{\langle\eta\rangle}-1)(x(\eta)-z) \Big\} \end{aligned} \qquad (140)$$

with $x(\eta) \equiv w + v\eta/\langle\eta\rangle$. Substituting Eq. (138) we obtain

$$\begin{aligned} A_2^{\exp} &= \frac{a}{v(1-a)}(A_{2.1}+A_{2.2}+A_{2.3}) \\ A_{2.1} &= \int_{-\infty}^{w} Dz \int_0^\infty d\eta 4a\exp(-2\eta)(\frac{\eta}{a}-1)(w+\frac{v\eta}{a}-z) \\ A_{2.2} &= \int_{-\infty}^{w} Dz(1-2a)(z-w) \\ A_{2.3} &= \int_w^\infty Dz \int_{\frac{(z-w)a}{v}}^\infty d\eta 4a\exp(-2\eta)(\frac{\eta}{a}-1)(w+\frac{v\eta}{a}-z). \end{aligned} \qquad (141)$$

Solving the equations leads to

$$\begin{aligned}
A_{2.1} &= (1-2a)\sigma(w) + \Big[\frac{v}{a} + w - v - 2wa\Big]\phi(w)\\
A_{2.2} &= (2a-1)(\sigma(w) + w\phi(w))\\
A_{2.3} &= \exp\Big(\frac{2aw}{v}\Big)\exp\Big(\frac{2a^2}{v^2}\Big)\Big[\frac{v(1-a)}{a}\phi\Big(-w-\frac{2a}{v}\Big)+\\
&+\sigma\Big(w+\frac{2a}{v}\Big) - \Big(w+\frac{2a}{v}\Big)\phi\Big(-w-\frac{2a}{v}\Big)\Big].
\end{aligned}\tag{142}$$

Thus

$$\begin{aligned}
A_2^{\exp} &= \phi(w) + \exp\Big(\frac{2aw}{v}+\frac{2a^2}{v^2}\Big)\Big\{\phi\Big(-w-\frac{2a}{v}\Big)+\\
&\frac{a}{v(1-a)}\Big[\sigma\Big(w+\frac{2a}{v}\Big) - \Big(w+\frac{2a}{v}\Big)\phi\Big(-w-\frac{2a}{v}\Big)\Big]\Big\}
\end{aligned}\tag{143}$$

For A_3 we have

$$\begin{aligned}
A_3^{\exp} &= A_{3.1} + A_{3.2} + A_{3.3}\\
A_{3.1} &= \int_{-\infty}^{w} Dz \int_0^{\infty} d\eta 4a\exp(-2\eta)(w+\frac{v\eta}{a}-z)^2\\
A_{3.2} &= \int_{-\infty}^{w} Dz(1-2a)(w-z)^2\\
A_{3.3} &= \int_{w}^{\infty} Dz \int_{\frac{(z-w)a}{v}}^{\infty} d\eta 4a\exp(-2\eta)(w+\frac{v\eta}{a}-z)^2
\end{aligned}\tag{144}$$

Substituting Eq. (138) we obtain

$$\begin{aligned}
A_{3.1} &= (1-2a)\sigma(w) + \Big[\frac{v}{a} + w - v - 2wa\Big]\phi(w)\\
A_{3.2} &= (2a-1)(\sigma(w) + w\phi(w))\\
A_{3.3} &= \exp\Big(\frac{2aw}{v}\Big)\exp\Big(\frac{2a^2}{v^2}\Big)\Big[\frac{v(1-a)}{a}\phi\Big(-w-\frac{2a}{v}\Big)+\\
&+\sigma\Big(w+\frac{2a}{v}\Big) - \Big(w+\frac{2a}{v}\Big)\phi\Big(-w-\frac{2a}{v}\Big)\Big]
\end{aligned}\tag{145}$$

and solving the equations leads to

$$\begin{aligned}
A_{3.1} &= 2a\Big[\sigma(w)(w+\frac{v}{a}) + \phi(w)(1+w^2+\frac{vw}{a}+\frac{v^2}{2a^2}\Big]\\
A_{3.2} &= (1-2a)[w\sigma(w) + (1+w^2)\phi(w))]\\
A_{3.3} &= \frac{v^2}{a}\exp\Big(\frac{2aw}{v}\Big)\exp\Big(\frac{2a^2}{v^2}\Big)\phi\Big(-w-\frac{2a}{v}\Big).
\end{aligned}\tag{146}$$

Thus

$$\begin{aligned}
&A_3^{\exp} = 2v(\sigma(w)+\phi(w)) + w\sigma(w) + (1+w^2)\phi(w) + \frac{v^2}{a}\phi(w) + \exp\Big(\frac{2aw}{v}+\\
&+\frac{2a^2}{v^2}\Big)\phi(-w-\frac{2a}{v}).
\end{aligned}\tag{147}$$

D.3 HEBBIAN CAPACITY OF TL NETWORKS STORING LOG-NORMAL DISTRIBUTED PATTERNS

In chapter 6, we studied the storage capacity of TL networks when the neural activity of the stored patterns are drawn from a number of distributions: binary, ternary, quaternary and exponential. In particular, we analysed the experimental data in relation to the exponential distribution. Several authors, e.g. Buzsáki and Mizuseki, in [85], have observed that often neural activity distribution resemble a log-normal distribution of suitable mean and variance. While not claiming to perform a comprehensive model selection, their study makes the important point that neural activity in many instances has a heavier tail than Gaussian, and is better modelled by a log-normal distribution. In this section, we therefore analyze the storage capacity of TL networks with Hebbian learning also for patterns whose activity follows a log-normal distribution. To be concrete, we assume that the patterns η are drawn from the following distribution

$$P(\eta)d\eta = \frac{1}{\eta}\frac{e^{-\frac{(\ln(\eta)-\mu)^2}{2\kappa^2}}}{\kappa\sqrt{2\pi}}d\eta \tag{148}$$

for which we have

$$\langle\eta\rangle_\eta = e^{\mu+\frac{\kappa^2}{2}} \tag{149a}$$

$$\langle\eta^2\rangle_\eta = e^{2\kappa^2+2\mu}, \tag{149b}$$

where here and in what follows $\langle\cdots\rangle_\eta$ represents averaging with respect to the log-normal distribution in Eq. 148. The sparsity, as defined in Eq. 4 of the main text, then reads

$$a = \frac{\langle\eta\rangle^2_\mu}{\langle\eta^2\rangle_\mu} = e^{-\kappa^2} \tag{150}$$

and it only depends on κ and not on μ.

If we substitute $z = \frac{\ln(\eta)-\mu}{k}$, such that $\eta = e^{zk+\mu}$ we obtain

$$P(e^{\kappa z+\mu})dz = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz. \tag{151}$$

Using this, and the fact that

$$\frac{\eta}{\langle\eta\rangle} = e^{\kappa z+\mu-\mu-\frac{\kappa^2}{2}} = e^{\kappa z-\frac{\kappa^2}{2}}, \tag{152}$$

we can evaluate the quantities A_2 and A_3 defined in Eqs. (69) and (70) of Sect. 6.1 as

$$\begin{aligned} A_2^{\ln-n}(w,v) = \frac{a}{v(1-a)}\int_{-\infty}^{\infty} Dz\Big(e^{\kappa z-\frac{\kappa^2}{2}}-1\Big)\cdot \\ \cdot\Big[(w+ve^{\kappa z-\frac{\kappa^2}{2}})\phi(w+ve^{\kappa z-\frac{\kappa^2}{2}})+\sigma(w+ve^{\kappa z-\frac{\kappa^2}{2}})\Big] \end{aligned} \tag{153}$$

$$A_3^{\ln-n}(w,v) = \int_{-\infty}^{\infty} Dz[(w+ve^{\kappa z-\frac{\kappa^2}{2}})^2+1]\phi(w+ve^{\kappa z-\frac{\kappa^2}{2}})+ \\ +(w+ve^{\kappa z-\frac{\kappa^2}{2}})\sigma(w+ve^{\kappa z-\frac{\kappa^2}{2}}) \tag{154}$$

which can then be used to find the storage capacity, α_c, as the value of α above which Eq. (78) in Sect. D cannot be satisfied.

Fig. 32a shows α_c for the log-normal distribution as a function of the sparsity a. For comparison, we have also included the results for the exponential distribution. One can see that, when plotted as a function of the sparsity of the stored patterns, the capacity of the log-normal distributed patterns is higher than the exponential one. But this high capacity is obtained because of the much sparser retrieved pattern compared to the stored one, as in Fig. 32c. When plotted versus the sparsity of the retrieved pattern, as can be seen Fig. 32b, the capacity of the log-normal distribution is always lower than the one of the exponential distribution. It is also possible to analytically find the limit of $a \to 1$ of the capacity of the log-normal distribution. In or-

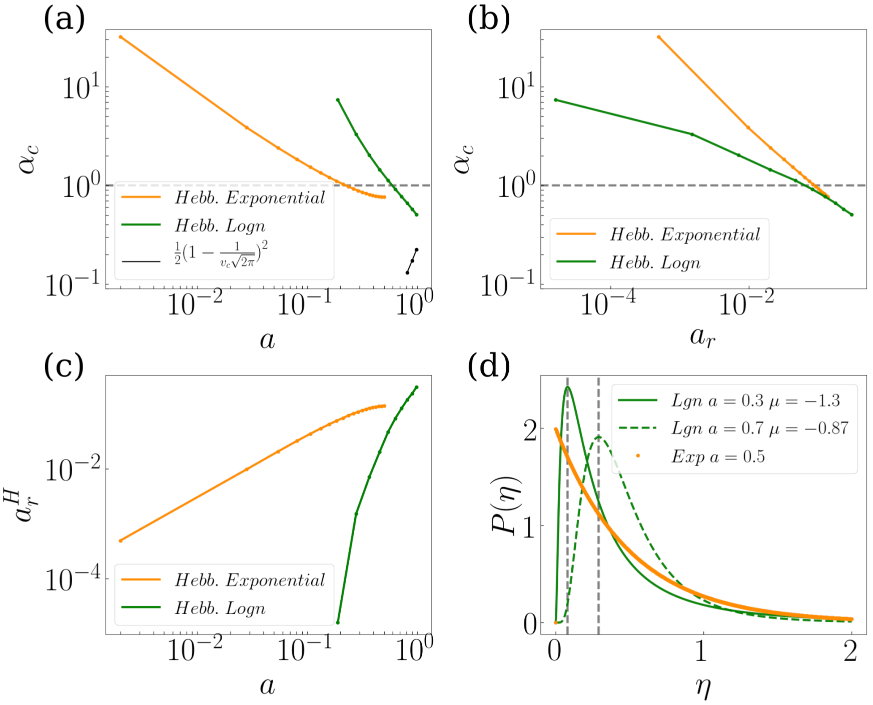


Figure 32: a) α_c vs a, the black line corresponds to Eq. (162) estimating the limit; b) α_c vs a_r; c) a vs a_r. d) Examples of log-normal and Exponential distributions, with parameters (see the legend) such that $<$ eta $>=$ 0.5 in all 3 cases. Note while the log-normal distributions have modes above zero, they have thinner tails than the exponential (but thicker than an ordinary normal). The vertical lines correspond to the maximum evaluated as in Eq. (166)

der to evaluate $\alpha_c = \frac{A_2^2}{A_3}$ as $a \to 1$ we estimate A_2 and A_3, by writing $b \equiv 1 - a$, so that $b \to 0$. In this way:

$$\begin{aligned}
k = \sqrt{-\ln(1-b)} \rightarrow \lim_{b \to 0} k = \sqrt{b} + \mathcal{O}(b^{3/2}) \\
\lim_{b \to 0} A_2 = \frac{1-b}{\nu b} \int_{-\infty}^{\infty} Dz \Big(e^{\kappa z - \frac{\kappa^2}{2}} - 1 \Big) f(k) \\
f(k) = \Big[(w + \nu e^{\kappa z - \frac{\kappa^2}{2}}) \phi(w + \nu e^{\kappa z - \frac{\kappa^2}{2}}) + \sigma(w + \nu e^{\kappa z - \frac{\kappa^2}{2}}) \Big] \\
\lim_{b \to 0} e^{\kappa z - \frac{\kappa^2}{2}} = 1 + \kappa z - \frac{\kappa^2}{2} + \mathcal{O}(k^3)
\end{aligned} \tag{155}$$

so we have

$$\begin{aligned}
f(k) = f(k=0) + k \frac{f(k)}{dk} |_{k=0} + \mathcal{O}(k^3) \\
\frac{df(k)}{dk} = \nu \exp^{\kappa z - \frac{\kappa^2}{2}} (z - k) \phi(w + \nu \exp^{\kappa z - \frac{\kappa^2}{2}}) \\
f(k) \approx (w + \nu)\phi(w + \nu) + \sigma(w + \nu) + k \nu z \phi(w + \nu)
\end{aligned} \tag{156}$$

For $A_2(w, \nu)$ we have:

$$\begin{aligned}
\lim_{b \to 0} A_2 &= \frac{1-b}{\nu b} \int_{-\infty}^{\infty} Dz \Big(\kappa z - \frac{\kappa^2}{2} \Big) \Big[f(k=0) + k \nu z \phi(w + \nu) \Big] \\
&= \frac{1-b}{\nu b} \Big\{ k^2 \nu \phi(w + \nu) \int_{-\infty}^{\infty} z^2 Dz - \frac{k^2}{2} f(k=0) \int_{-\infty}^{\infty} Dz \Big\} \\
&= \frac{1-b}{\nu b} k^2 \Big(\nu \phi(w + \nu) - \frac{f(k=0)}{2} \Big)
\end{aligned} \tag{157}$$

given that

$$k^2 = -\ln(1-b) \approx b + \mathcal{O}(b^2) \tag{158}$$

then

$$\lim_{b \to 0} A_2(w, \nu) = \lim_{a \to 1} A_2(w, \nu) = \phi(w, \nu) - \frac{(w + \nu)\phi(w + \nu) + \sigma(w + \nu)}{2\nu} \tag{159}$$

For $A_3(w, \nu)$ instead we simply have

$$\lim_{a \to 1} A_3(w, \nu) = [(w + \nu)^2 + 1]\phi(w + \nu) + (w + \nu)\sigma(w) \tag{160}$$

Then

$$\lim_{a \to 1} \alpha_c = \frac{\Big(\phi(w, \nu) - \frac{(w+\nu)\phi(w+\nu) + \sigma(w+\nu)}{2\nu} \Big)^2}{[(w + \nu)^2 + 1]\phi(w + \nu) + (w + \nu)\sigma(w)} \tag{161}$$

Plotting $w_c + \nu_c$ as a function of a one can see that $\lim_{a \to 1} w_c + \nu_c \approx 0$. If we substitute that we get

$$\lim_{a \to 1} \alpha_c \approx \frac{1}{2} \Big(1 - \frac{1}{\nu_c \sqrt{2\pi}} \Big)^2. \tag{162}$$

This equation was solved numerically up to the value of $a = 0.99$, obtaining the black line in Fig. 32.

In order to estimate the exact value at $a \to 1$ we can also require that the two derivatives vanish, i.e. $2A_2A_{2w} - \alpha A_{3,w} = 0$ e $2A_2A_{2v} - \alpha A_{3,v} = 0$. To do so we define $x = w + v$ for simplicity of visualization and write

$$\begin{cases} \left(\phi(x) - \frac{x\phi(x)+\sigma(x)}{2v}\right)^2 - \alpha[x^2+1]\phi(x) + (x)\sigma(x) \\ 2\left(\phi(x) - \frac{x\phi(x)+\sigma(x)}{2v}\right)\left(\sigma(x) + \frac{\phi(x)}{2v}\right) - 2\alpha[x(\phi(x)+\sigma(x)] = 0 \\ 2\left(\phi(x) - \frac{x\phi(x)+\sigma(x)}{2v}\right)\left(\sigma(x) + \frac{\phi(x)}{2v} - \frac{x\phi(x)+\sigma(x)}{2v^2}\right) - 2\alpha[x(\phi(x)+\sigma(x)] = 0 \end{cases} \tag{163}$$

By subtracting the last two equations we obtain

$$2\left(\phi(x) - \frac{x\phi(x)+\sigma(x)}{2v}\right)\frac{x\phi(x)+\sigma(x)}{2v^2} = 0 \tag{164}$$

which can be satisfied only if $v \to \infty$. One can then show that the rest of the equations hold for $x = 0$, thus $\alpha_c^{logn}(a = 1) = 0.5$.
The maximum of the log-normal distribution defined in Eq. (148) is given by:

$$\frac{dP(\eta)}{d\eta} = \frac{P(\eta)}{\eta}\left(-1 - \frac{\ln(\eta) - \mu}{k^2}\right) = 0 \tag{165}$$

where we get the condition $k^2 - \ln(\eta) + \mu = 0$ which is satisfied when:

$$\eta_{max} = ae^{\mu} \tag{166}$$

In conclusion, activity distributions which are well fit by a log-normal result in associative networks that can operate in two somewhat distinct, but continuous regimes: if the distribution is tightly clustered around its mean, i.e. not sparse, $0.5 < a < 1$, the retrieved distribution is not sparse either, and the Hebbian capacity is between $\simeq 1$ and 0.5, comparable but lower than the Gardner capacity. Note that for such values of a no alternative exponential distribution is available, as it would imply $f = 2a > 1$, and indeed for the log-normal $f \equiv 1$ always (implying that a comparison with the Gardner bound would only be limited to its value for $f = 1$, i.e. $\alpha^G(1) \equiv 1$). If instead the distribution is sparse, i.e. k is larger such that $a < 0.5$, the Hebbian capacity is above unity (but below that of the exponential fit, which has a fatter tail), but the retrieved distribution rapidly becomes so much sparser as to make retrieval unfeasible for any reasonably sized network.

D.4 CALCULATING THE SPARSITY OF THE RETRIEVED PATTERNS

Following [54], the average of the activity and the average of the square activity in the patterns retrieved with Hebbian weights are calculated considering that the field, i.e. the input received by a cell with activity η in the memory, is normally distributed around a mean field proportional to x. If we call z a random variable normally distributed with mean zero and variance one, x is already the mean field properly normalized. With the threshold-linear transfer function, the output will be $g(x+z)$ for $x+z>0$ and 0 with probability $\phi(-x)$. Therefore the average activity $\langle V \rangle$ (denoted as x in [53, 54, 57]) and the average square activity $\langle V^2 \rangle$ (denoted as y_0 in [53, 54, 57]) are,

$$\langle V \rangle = g \left\langle \int_{-x(\eta)}^{\infty} Dz \, [x(\eta)+z] \right\rangle_{\eta} = g \left\langle [x_c \phi(x_c) + \sigma(x_c)] \right\rangle_{\eta} \tag{167a}$$

$$\left\langle V^2 \right\rangle = g^2 \left\langle \int_{-x(\eta)}^{\infty} Dz \, [x(\eta)+z]^2 \right\rangle_{\eta} = \tag{167b}$$

$$g^2 \left\langle \left[(1+x_c^2)\phi(x_c) + x_c \sigma(x_c) \right] \right\rangle_{\eta},$$

where

$$x_c \equiv w_c + v_c \frac{\eta}{\langle \eta \rangle}. \tag{168}$$

The sparsity of the retrieved memory is thus $a_r^H = \langle V \rangle^2 / \langle V^2 \rangle$.

E

COMPARISON HEBBIAN-OPTIMAL LEARNING: ADDITIONAL RECORDED CELLS

Supplementary to Fig. 15 in chapter 7, we report in Fig. 33 the same analysis for all 9 single cells reported (using 100ms bins) in [78]. In each panel we write the capacity *à la Gardner* and the Hebbian one (calculated without fitting an exponential) for the 9 empirical distributions, as well as the sparsity of the original distribution and the sparsity of the one that would be retrieved with Hebbian weights.

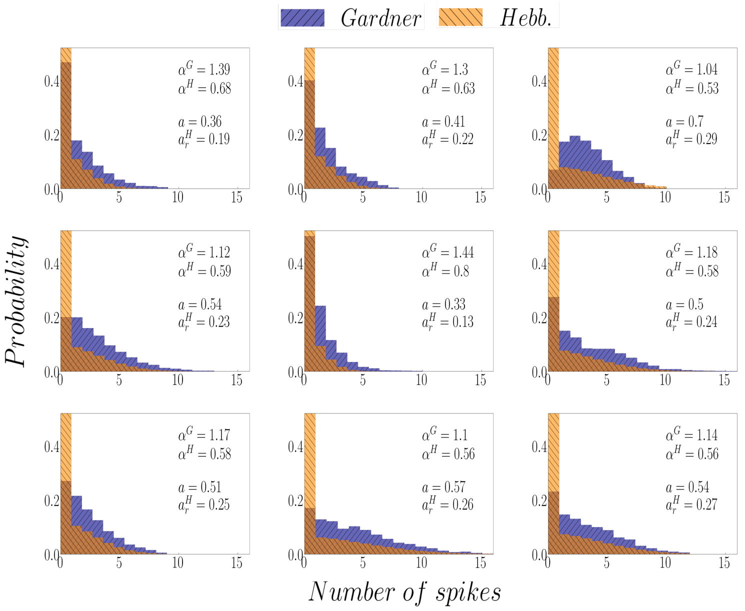


Figure 33: Suplementary to Fig. 15 in the main text.

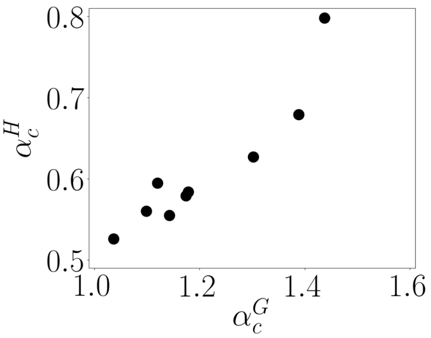


Figure 34: Comparison between the values of the storage capacity *à la Gardner* and Hebbian, for the 9 empirical distributions extracted from [78].

For simplicity of visualization we also show the storage capacity values against each other, calculated *à la Gardner* and *à la Hebb* (again, without fitting an exponential), as a single scatterplot for the 9 distributions, in Fig. 34.

F

FURTHER DETAILS OF THE CONTINUOUS QUASI-ATTRACTOR MODEL

F.1 EFFECTIVE MEAN OF THE MODIFIED EXPONENTIAL DISTRIBUTION

In Fig. 35 we report the numerical relation between the mean ζ of the probability density function $f(n_F, \frac{1}{\zeta}) = \frac{1}{\zeta}\exp(-\frac{n_F}{\zeta})$ and the effective average number of fields $\langle n_F \rangle$ given that values of 0 or > 21 are not accepted.

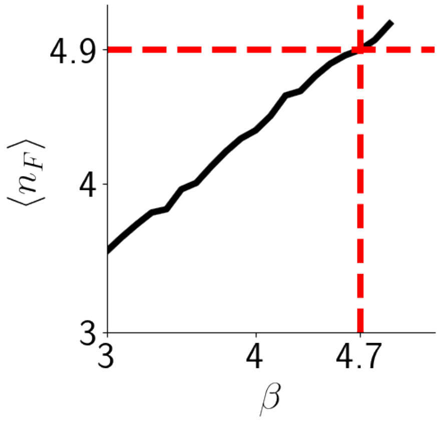


Figure 35: Numerical estimation of $\langle n_F \rangle$ drawing 15000 random numbers from the exponential distribution with p.d.f. $f(n_F, \frac{1}{\zeta})$ (see text) under the described constraints. The horizontal red line sets the experimental average value while the vertical red line sets our arbitrary choice of ζ to obtain, generally, the desired average.

F.2 SUPPLEMENTARY FIGURES – CONTINUOUS QUASI-ATTRACTIVE MANIFOLDS

F.2.1 *Further examples of dynamics which exit the manifold*

Here we report, as a supplement to Fig.21 and Sect. 12, two additional examples of dynamics which we regard as *jumping* and, thus, as indicative of the quasi-attractive manifold break-up.

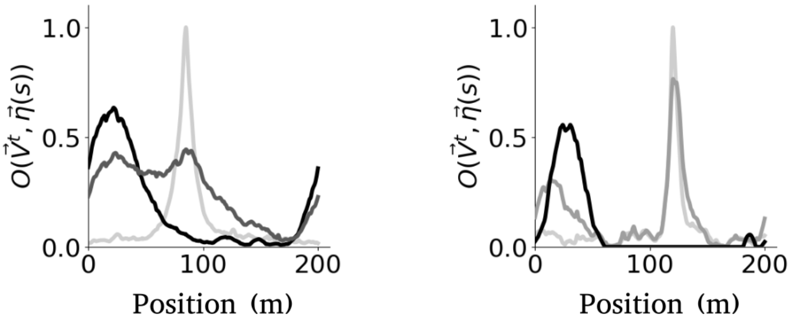


Figure 36: Dynamics which *jump*. Supplementary to Fig.21 we report the plot, in the overlap space, of a few distinctive configurations occurring in two dynamical evolutions (gray to black) which we regard as exiting the manifold, due to their loss of localization. Left) refers to $\{\vec{\eta}(s)\}$ characterized by $\sigma_d = 0.9$, $\sigma_p = 0.9$, $,\zeta = 1$,$N = 8000$. Dynamical step size $\gamma = 0.1$, $g = 1$, $k_B = 300$. The plotted dynamics is initialized from $\vec{\eta}(s = 90)$ and reaches the fixed point in 94 steps, gradually gray to black lines correspond to the configurations at $t = 0, 24, 40$ ($V^{t=40}$ is visually identical to $V^{t=94}$). Right) refers to $\{\vec{\eta}(s)\}$ characterized by $\sigma_d = 0.7$, $\sigma_p = 0.7$, $N = 4000$. Dynamical step size $\gamma = 0.1$, $g = 5$, $k_B = 500$. The dynamics is initialized from $\vec{\eta}(s = 120)$ and reaches the fixed points in 108 steps, gradually gray to black lines correspond to the configurations at $t = 0, 4, 108$ (all steps > 20 were here set black). In both $\{\vec{\eta}(s)\}$ s is discretized into 1000 equally spaced positions every 0.2m and the dynamics is considered to have converged when $\sum_i (V_i^t - V_i^{t-1})^2 < 10^{-5}$.

F.2.2 *Additional comparison between activity and overlap spaces*

In Fig. 37 and 38 we report the plots of the activity space for the sampled configurations in the two dynamical evolutions used as an example in Fig.21 C and D, respectively. As introduced in the main text, the activity space, commonly used to visualize continuous attractors by placing the activity of each unit at the position of its center, in our case of irregular patterns is obtained by plotting the activity of each unit i at the position in s where $\eta_i(s)$ is maximal, i.e. selecting the field center of the field with highest peak rate. If, when the irregularity in $\{\vec{\eta}(s)\}$ is sufficiently low, a noisy bump is visible (Fig. 37) throughout the dynamics, this is not true when the irregularity in $\{\vec{\eta}(s)\}$ is above a certain threshold (Fig. 38).

F.3 SUPPLEMENTARY MEASURE DETAILS

As a supplement to Sect.12.1 here we provide with greater details the procedure we implement to quantify loss of attraction by the manifold .

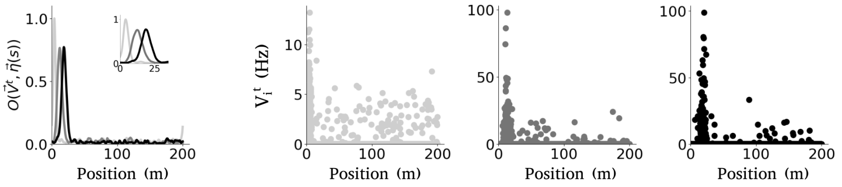


Figure 37: Supplement to Fig. 21C, which is reproduced on the left. The three panels on the right represent the localized configurations in the activity space corresponding to those plotted on the left in the overlap space, analogously color coded. Refer to the caption of Fig. 21C for the parameters.

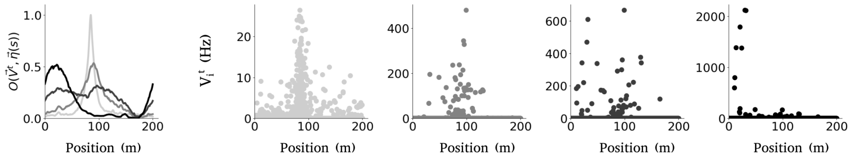


Figure 38: Supplement to Fig. 21D., which is reproduced on the left. The four panels on the right represent the configurations in the activity space corresponding to those plotted on the left in the overlap space, analogously color coded. Refer to the caption of Fig. 21D for the parameters.

F.3.1 *Proportion of dynamics which jump*

Given a system we run a number $\geqslant 50$ of dynamics each starting from a certain configuration $\vec{\eta}(s)$, such that the overall length of the manifold is uniformly sampled. 50 dynamics are enough, for the variable systems we are studying, to reach each fixed point configuration, typically several times.

At each dynamical step, we estimate the position of the center of the bump. If at step $t+1$ the center moved farther than a certain, arbitrarily set, distance ($20cm$) from the center estimated at step t, we check that the overlap values progressively decrease from the position of the new center to the position of the old one. If this does not occur for more than 5 discrete positions (corresponding to $1m$ in the model) we regard this dynamics to have jumped outside the manifold. If instead it occurs for less than $1m$ we consider this effect as a "physiological" ripple of the bump.

F.3.2 $\langle O_{tang} \rangle$

In order to calculate the cosine similarity between the eigenvector closest to instability and the direction of the manifold, one needs to estimate the direction of the manifold around each fixed point. We do this providing as external stimulus (i.e initializing a dynamics) with a template $\vec{\eta}(s)$ corresponding to the activity at a position $1.4\mathfrak{m}t$ (7 discrete positions in s) away from the fixed point position s^* (by position we mean, as elsewhere, the center of its bump in the overlap space). This distance, generally, is short enough that the bump slides towards the fixed configuration (even when the manifold has disappeared elsewhere) and long enough to have reached $\{\vec{\xi}\}$ at s^*-1. We estimate $\vec{\xi}(s^*-1)$ as the average of the configurations of activity which are centered at $\vec{\eta}(s^*-1)$ in the dynamical evolution. We then estimate the direction of the manifold as $\vec{\xi}(s^*-1)-\vec{\xi}(s^*)$.

F.3.3 *Bump width*

As introduced in the main text, one can estimate the standard deviation of the center of mass of a configuration of activity in the overlap space. Given a vector $\vec{O}$, where each entry is the overlap $O_s = O(\eta(s), V)$ (Eq. (116)) with the respective quenched pattern, the center of mass is given by

$$\text{c.m.} = \frac{L}{2\pi}\text{arctan2}\left(-\frac{\sum_s O_s \cos\left(\frac{2\pi}{L}s\right)}{\sum_s O_s}, -\frac{\sum_s O_s \sin\left(\frac{2\pi}{L}s\right)}{\sum_s O_s}\right) \quad (169)$$

where $s^{max} = L$ is the total length and the trigonometry is used to implement periodic boundary conditions. The standard deviation around the center of mass, instead, can be calculated by first estimating the vector $\vec{d}$ of the minimal distances between each s and the center of mass (keeping in mind periodic conditions) and then as

$$\text{st.d.} = \sqrt{\frac{\sum_s O_s d_s^2}{\frac{L^2}{12}\sum_s O_s}} \quad (170)$$

As the irregularities outside the manifold substantially increase the standard deviation, thus reducing the information regarding how localized is the bump, we remove all irregularities below 0.1 (we simply subtract 0.1 by all overlap values and send to zero those which become negative). The difference in the standard deviation of the subtracted vs the non-subtracted overlap can be seen in Fig. 39 and is reflected in the two dynamical evolutions reported in this appendix in Figs. 37 and 38.

F.4 ENERGY ESTIMATION OF $\{\xi(\vec{s})\}$

In Fig. 40 we report, in red, the energy (Eq. (113)) of the same $\{\eta(\vec{s})\}$ as in Fig. 21A, whereas in black we report the energy of the rough estimation of $\{\xi(\vec{s})\}$. Estimating exactly $\{\xi(\vec{s})\}$ is not trivial. Here we do it, for each s, by averaging over all the dynamical variables which had maximal overlap with a certain $\vec{\eta}(s)$ from the third dynamical

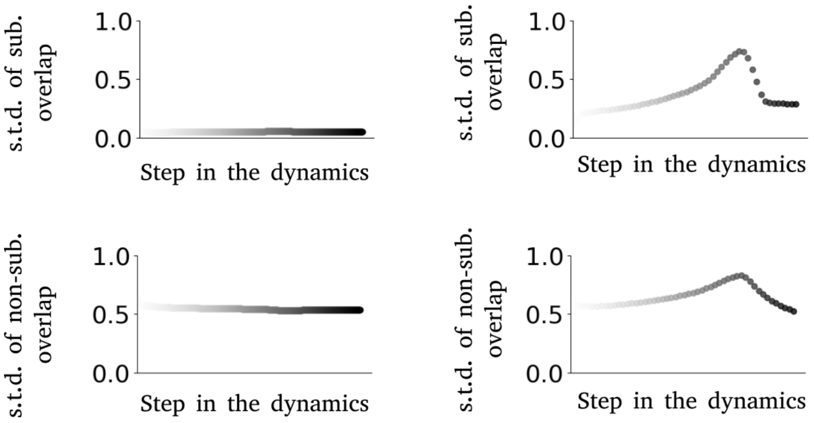


Figure 39: Left column: Supplement to Fig. 21E; Right column: Supplement to Fig. 21F. In this plot we show the difference in measuring the standard deviation of the intact overlap profile versus the reduced one. While the general trend is maintained the differentiation between localized vs non-localized states is more evident when the profile is subtracted.

step in 50 runs, each starting from a different $\eta(\vec{s})$ (such that the whole length is spanned). What should be principally appreciated is the localization of the minima, not the great disparity in the absolute values. The latter depends, also, on i) how one fixes the average and ii) how far $\{\eta(\vec{s})\}$ is from regular patterns: the farther it is the more the disparity, and in the configuration $\{\xi(\vec{s})\}$ only some units tend to be more active whereas the majority tend to be shut down.

F.5 phase diagram - number of fixed points

Here we report as a supplement to Fig. 24 the phase diagrams showing the variation in the number of fixed points and in their sparsity with σ_d, σ_p and ζ. High variability in the peak rate seems to act positively on the continuity of fixed points. This can be possibly intended as high σ_p enabling few outsized peaks, which may promote individual fields as the only relevant ones for most units, thus fostering order. One should consider that for $\zeta = 2.85$ and $\zeta = 4.7$ the lower left semi-circular region (outlined as in Fig. 24 second line) corresponds to few non-localized fixed points, whereas in the rest of the phase space fixed points are localized.
Whereas for $\zeta = 2.85$ and $\zeta = 4.7$ the non-localized region could be intended as related to the abrupt increase in sparsity of the fixed points (Fig. 41 second line) this does not seem to be true for $\zeta = 1.0$ where, while the sparsity abruptly increases, fixed points persist lo-

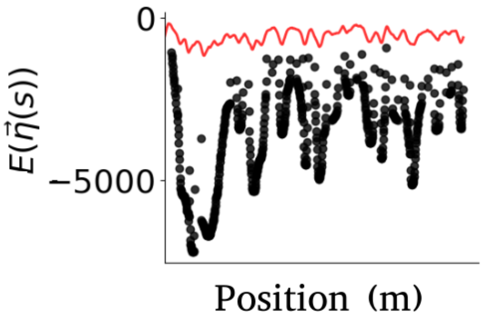


Figure 40: Supplementary to Fig. 21A we report the energy of $\{\eta(\vec{s})\}$ in red (as in Fig. 21A) while in black we report the energy of the estimated $\{\xi(\vec{s})\}$. Each $\vec{\xi}(s)$ is obtained averaging all configurations $V(t)$ having a bump in the overlap space centered at s, for 50 dynamics initialized at discrete $\vec{\eta}(s)$ spanning the entire length. Parameters are the same as in Fig. 21A-C-E

calized (relate to Fig. 24). Further intuitions over the role of sparsity in the transition may come from an analytical derivation, currently in progress.

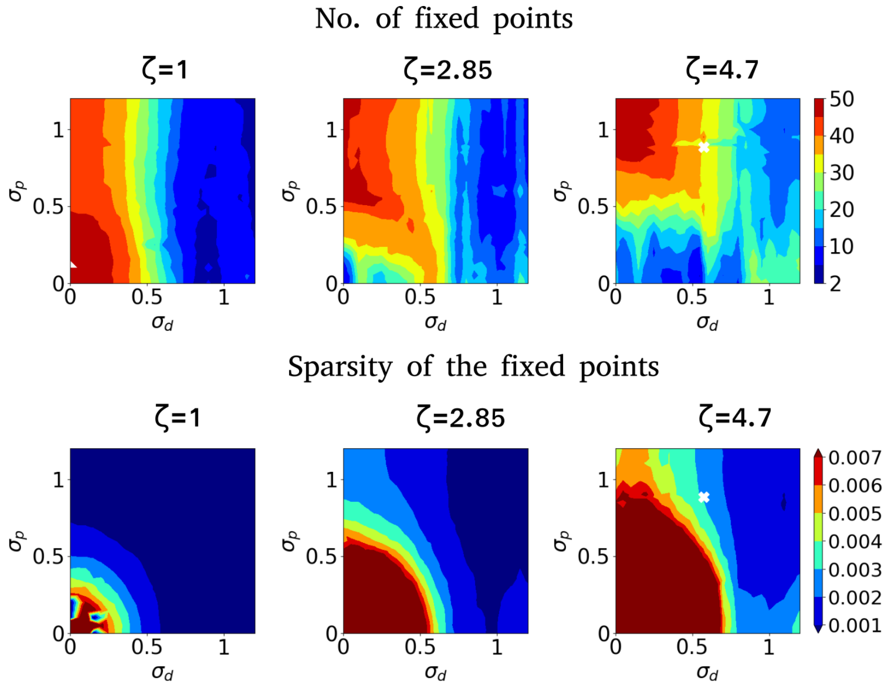


Figure 41: Phase diagram of the number of fixed points and their sparsity as a supplement to Fig. 24. For the second row the dark red includes all values > 0.007 and dark blue all those < 0.001. Refer to Fig.24 for the parameters.

F.6 SIMULATION DETAILS

Let us give a few details concerning the implementations. The dynamics introduced in Eq. 111, in the simulations, updated synchronously with a step size γ as:

$$\vec{V}^{t+1} = (1-\gamma)\vec{V}^t + \gamma g\Big[\bar{J}\vec{V}^t - 4k(\frac{\sum_i^N V_i^t}{N} - \nu_0)^3\Big]^+ \tag{171}$$

We normally set the desired average ν_0 as the one of the specific realization of the quenched patterns $\langle\{\vec{\eta}(s)\}\rangle$, though setting other values did not seem to have any major effect. A small step size is required to prevent the dynamics from entering in bi-stable states. Depending on the parameter, on the specific realization of the quenched patterns $\langle\{\vec{\eta}(s)\}\rangle$ and the initial conditions, the dynamics may require higher or lower γ, which on average was set around $\gamma = 0.04$. The dynamical evolution generally requires a balance between sufficiently high k (to fix, roughly, the mean) and sufficiently high N (especially when the irregularity is high) which often implies, however, a decrease in γ, increasing the computational cost.

While fixing the mean does not seem to play a crucial role in the overall dynamics, and fixed points reached with any value simply scale up or down by a certain factor and are characterize by almost equal smallest eigenvalues, the same is not true for the gain, which is the most fundamental parameter.

F.7 SELECTION OF THE APPROPRIATE GAIN

As foreseen from the initial calculations of the storage capacity in associative networks endowed with threshold linear units [54, 56, 57], recapitulated in Sect. 6.1, the gain is a crucial parameter for this type of transfer functions. Indeed, in the analytical solutions for the storage capacity, the gain roughly defines the slope of a line, which has to intersect a closed curve in a two dimensional plane in order to obtain a solution. There is, therefore, a unique set of gain values which enables the retrieval of a memorized pattern. This range gets tighter and tighter the closer the load of memories is to the storage capacity, and, at the storage limit, when this closed curve shrinks to a point, the gain has a critical value.

There is not, however, an automated procedure (other than the analytical solution for uncorrelated patterns) which identifies the required g range, assuming that the behaviour somehow remains analogous also when patterns are highly irregular and correlated, such as those in the model we study.

Generally, in simulations well below the storage capacity the selection of the gain do not cause any inconvenience, as any g value in a broad range works fine.

When simulating dynamics where the connectivity matrix results through Hebbian learning of highly irregular quenched patterns, instead, we noticed that the gain needs to be sufficiently high to enable the retrieval of a localized attractor state, else the dynamics tend to reach a unique non-localized fixed point. Generally we see the following:

1. The more the quenched patterns (such as those described in the third part of the thesis) are irregular, the more insufficiently high gain values lead the dynamics towards a unique non-localized state.

2. When this gain value increases and passes a threshold (for the simulations we run this threshold can be roughly between $0.5-5$), which is more or less the same for all realizations of the same quenched disorder, the dynamics retrieve a semi-localized state. This, however, may not be so stable, and the dynamics may jump. The range of g values at which this occurs appears to be normally very small.

3. Increasing the gain slightly more leads, generally, to stable localized states, each with high overlap with one of the memories (i.e one of the patterns on the manifold $\{\vec{\eta}(s)\}$.

4. Increasing the gain more and more after this level does not seem to have any specific effect other than progressively decreasing the value of the maximal overlap at the center of mass in the overlap space, still maintaining the state localized, and progressively increasing the activity level of the most active units, which become fewer and fewer. At the "optimal" gain value, the maximal overlap of the localized states in the overlap space may range from around 1 when the connectivity is based on regular quenched patterns to about ≈ 0.5 when the patterns $\{\vec{\eta}(s)\}$ are highly irregular.

In the simulations we set manually the gain to be high enough.

One may argue that the whole arguments and numerical evidences presented in the third part of this thesis are a mere effect of the gain, given that for a unique realization of quenched random variables one can drive the same system in either of the three phases by simply tuning g. In reality, this seems to be true only in the well behaving region corresponding to the quasi-attractive continuous manifold: for the regions of the phase space corresponding to the jumps outside the manifold and to the non-localized fixed states, which given the above characterization would require higher g, we could not find any arbitrarily high gain value enabling a different behaviour.

The precision of the transition value in phase spaces even when the quenched patterns have lower irregularity in the number of fields (smaller ζ), as well as the smooth decrease in the number of fixed points, not transitioning at the critical noise value (see Fig.22) lead us to be confident that the transition is not a mere effect of the gain but a robust complex emergent phenomenon possibly related to the storage capacity.

However, until we will be able to derive an analytical solution there is no exact argument which we can provide else than the ones provided in the above description.

G QUASI-ATTRACTOR PHASE DIAGRAM IN MORE GENERAL SYSTEMS

The transition between the existence of the continuous quasi-attractor manifold and its disappearance, which we have characterized in fully connected networks for distributions resembling the experimental ones (Ch. 13), does not seem to require, to occur, all the details used in the model as specified above.

G.1 REGULAR FIELDS

We also studied a network where each unit in the quenched random patterns is characterized by a unique field regularly placed such that the whole 1d length is equidistantly covered by non overlapping fields centers. Further, we did not impose any correlation between field sizes and peak firing rates and we accepted any outcome of the lognormal distributions (without limiting to values below a certain threshold). For zero variability in field sizes and peak rates (i.e $\sigma_d = \sigma_p = 0$) this quenched variables lead to a standard continuous attractor neural network (or semi-continuous in finite-size numerical simulations): any pattern $\{\eta(s)\}$ corresponds to a fixed point on the continuous "non-quasi" attractor manifold $\{\xi(s)\}$. As soon as any tiny source of irregularity is introduced (i.e. if, for examlple σ_d or σ_p $= 0.05$ $(m)/(Hz)$) then the number of fixed points drastically drops: only some pattern $\{\eta(s)\}$ correspond to a fixed point on the continuous quasi attractor manifold $\{\xi(s)\}$. This is a known phenomenon and it has been already discussed in this thesis.

What we want to emphasize here is that increasing systematically the irregularity in the peak firing rate and field sizes, following lognormal distributions with average values set as those observed in one experiment [116], one reaches a phase transition between the existence and the disappearance of the quasi-attractive manifold, which, for low σ_p, seems to occur at a critical σ_d which coincides with that observed in experiments (Fig. 42 a).

The number of fixed points, which decreases from an infinite number a few tens (e.g., in a specific instance, 28), smoothly varies as well (Fig. 42c). The fixed points in all regions of the phase diagram are localized on the manifold (Fig. 42b) and their sparsity smoothly decreases for increasing variability in both directions (Fig. 42d).

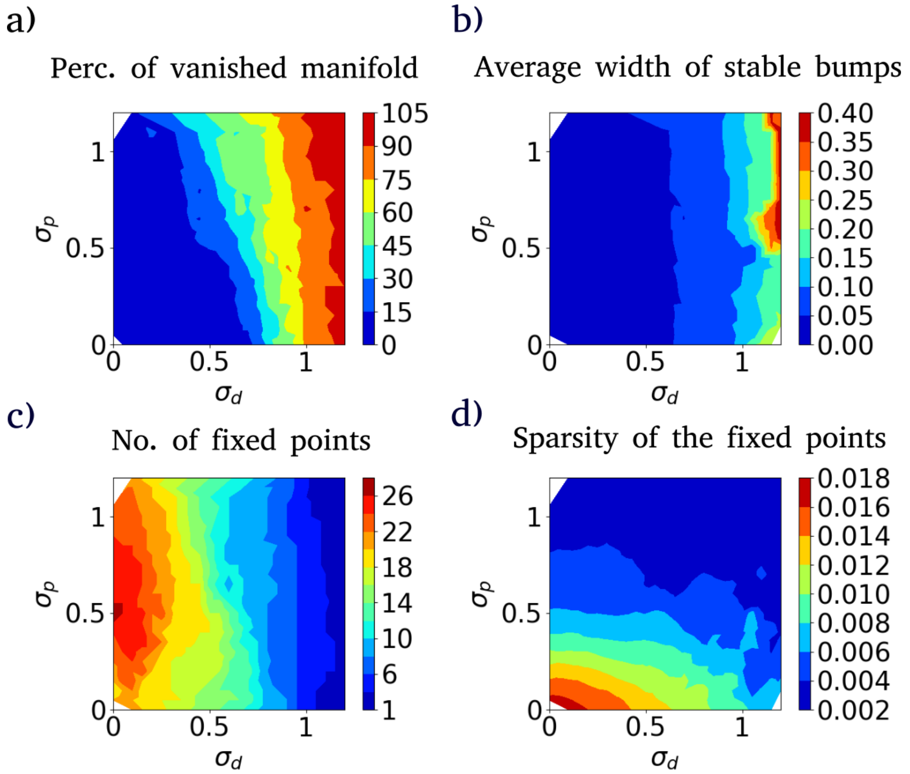


Figure 42: Phase diagram showing 4 indicative measures for a network storing regular quenched random patterns (refer to Sect. 12.1 for details over the measurements). The position at $(0, 0)$ corresponds to a standard continuous attractor. All fixed points are localized on the manifold (b) and they decrease in number already for extremely small σ_d and σ_p (c). A transition occurs to the disappearance of the quasi-attractive continuous manifold (a). Note that for low variability, the sparsity of the localized states is higher (d). Parameters: $\{\vec{\eta}(s)\}$ is produced discretizing the whole length $L = 200m$ into $N = s = 1000$ discrete positions each corresponding to the field center of a unit i (one every $20cm$). The field size and field peak rate of each unit is drawn from two independent lognormal distributions with respective mean $\mu_d = 1.57m$ and $\mu_p = 1.549Hz$ and standard deviation σ_d and σ_p. Each plot includes 26x26 data points corresponding to a unique realization of quenched random patterns. Each data point was produced by 50 simulations each initialized with a different $\eta(\vec{s})$, s spanning homogeneously the whole length (one every $4m$). A simulation is considered to have converged when $\sum_i |(V_i^t - V_i^{t-1})| < 10^{-10}$. The step size varies from $\gamma = 0.5$ to $\gamma = 0.01$, $g = 4$, $k = 300$.

These preliminary observations, performed on a system size 16 times smaller than those presented in Fig. 24, indicate that i) the horizontal transition does not require long range interactions between multiple fields and occurs also when there are only single fields for each unit. ii) That this transition is not occurring at the same critical σ_d value for different levels of irregularities in the peak firing rate. In simulations based on quenched patterns characterized by multiple fields, instead, it seems to occur at the same critical value (Fig. 24). The latter comparison, may lead to two different conjectures: either when quenched patterns are highly irregular in peak firing rates, the presence of more fields somehow benefits, instead of limiting, order; or the correlations between field sizes and field peak rates observed in experiments (and reproduced in the quenched patterns leading to Fig. 24) but erased from the model presented in this section, counterbalance this intrinsic tendency of a regular network to undergo the transition inducing a precise alignment of the critical dimension at its maximal value.

G.2 DILUTED CONNECTIVITY

Connections in the brain are sparse. While we studied systematically a fully connected network, representing the case in which noise reverberation in loops is maximal, we observe that the same transition occurs when networks are sparse, as closer to reality. The proportion of jumps, which we regard as a measurment of the integrity of the quasi-attractive continuous manifold, abruptly increases at a critical value of noise, similar to that seen in fully connected networks (see Fig. 43).

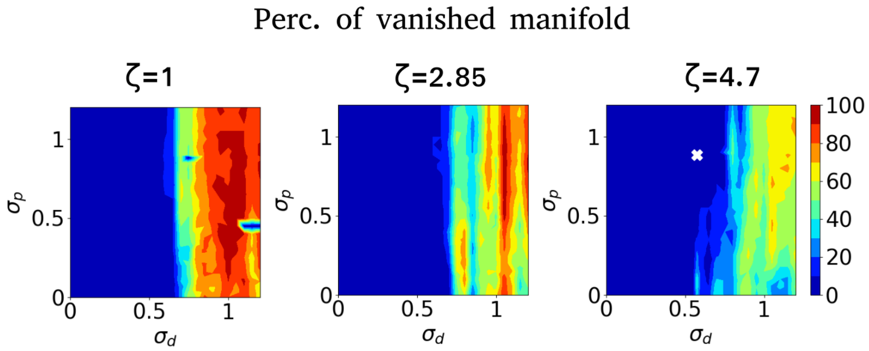


Figure 43: Phase diagram showing the proportion of simulations which jump for diluted networks: 80% of the connections are randomly and symmetrically set at zero. The transition seems to again occur at a value close to that seen in fully connected networks (compare with Fig. 24). Parameters: all parameters, except the dilution, are those of Fig. 24. Interpolated data points, here, are not averaged over three realizations of quenched random patterns but correspond to a single one, hence the more irregular appearance of these plots.

H

ANALYTICAL APPROACH TO THE CONTINUOUS QUASI-ATTRACTOR MODEL

We study a network whose matrix of recurrent weights is given by :

$$J_{ij} = \frac{1}{N}\int_S \frac{ds}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\left(\frac{\eta_j(s)}{\langle\eta\rangle} - 1\right) \tag{172}$$

There are N units, the activity $\eta_i(s)$ is defined at each position s on a ring S of length L defined in terms of "quenched" place fields, indexed by m, where

$$P(m_i) \approx e^{-\frac{m_i}{\langle m\rangle}} \tag{173}$$

The diameters and peaks of each place field are log-normal distributions. The field received by each unit is

$$h_i = \sum_{i,i\neq j} J_{ij}V_j + b\left(\frac{\sum_i^N V_j}{N}\right) \tag{174}$$

and the overall Hamiltonian

$$H = -\frac{1}{2}\sum_i\sum_{j\neq i} J_{ij}V_iV_j - NB\left(\frac{\sum_j V_j}{N}\right) \tag{175}$$

with

$$B(x) = \int^x b(y)dy \tag{176}$$

We define the order parameters

$$x(s) = \frac{1}{N}\sum_i^N \left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\langle V_i\rangle \tag{177}$$

$$x = \frac{1}{N}\sum_i \langle V_i\rangle \tag{178}$$

$$y_0 = \frac{1}{N}\sum_i \langle V_i^2\rangle \tag{179}$$

$$y_1 = \frac{1}{N}\sum_i^N \langle V_i\rangle^2 \tag{180}$$

The calculation of the free energy proceeds by moving the overlaps $x(s)$ to an orthonormal basis of wavelets on the circle X_{kR}, where k is the level or scale. and R is the shift (or phase on the circle).
Writing the partition function as in [57] (refer to the formula for (A3) with s=0) one has

$$Z^n = \Big(\frac{N\beta}{2\pi}\Big)^{pn} \int dt^{kl\gamma} dx^{kl\gamma} \Bigg\{ T_r\{V_i\} \exp\Bigg[\beta \sum_{k,l,\gamma,i} \Bigg(- i t^{kl\gamma} \Big(\frac{\eta_i^{kl}}{\langle \eta_i \rangle} - 1\Big) V_i^\gamma \Bigg)\Bigg] +$$

$$+ \beta N \sum_\gamma \Bigg[i \sum_{k,l} t^{kl\gamma} x^{kl\gamma} + \frac{1}{2} \sum_{kl} (x^{kl})^2 + B(\frac{1}{N} \sum_i V_i^\gamma) \Bigg]\Bigg\} \tag{181}$$

Now we assume that only overlaps with shift $l = 0$ (relative to the fixed point) condense and average over all the others. Keeping only the Gaussian terms:

$$\Bigg\langle \exp\Bigg[- i\beta \sum_\gamma t^{kl\gamma} V_i^\gamma \Bigg(\frac{\eta_i^{kl}}{\langle \eta_i \rangle} - 1 \Bigg)\Bigg]\Bigg\rangle$$

$$\approx \Bigg\langle 1 - i\beta \sum_\gamma t^{kl\gamma} V_i^\gamma \Bigg(\frac{\eta_i^{kl}}{\langle \eta_i \rangle} - 1 \Bigg) - \frac{\beta^2}{2} \Big(\sum_\gamma t^{kl\gamma} V_i^\gamma \Big)^2 \Bigg(\frac{\eta_i^{kl}}{\langle \eta_i \rangle} - 1 \Bigg)^2 \Bigg\rangle$$

$$\approx e^{-\frac{1}{2}\beta^2 \Big(\sum_\gamma t^{kl\gamma} V_i^\gamma \Big)^2 T_0^{kl}} \tag{182}$$

with

$$\begin{aligned} T_0^{kl} &\equiv \Bigg\langle \Bigg(\frac{\eta_i^{kl}}{\langle \eta_i \rangle} - 1 \Bigg)^2 \Bigg\rangle \\ \overline{T_0} &\equiv \int \frac{ds}{S} \Bigg\langle \Bigg(\frac{\eta_i(s)}{\langle \eta_i \rangle} - 1 \Bigg)^2 \Bigg\rangle \end{aligned} \tag{183}$$

and, doing the Gaussian in terms over $t^{kl\gamma}$, one gets to the equation of (A5):

$$\begin{aligned}
\left\langle Z^n \right\rangle = & \Big(\frac{N\beta}{2\pi}\Big)^{p_0 n + \frac{(p-p_0)}{2} n + \frac{(n+3)}{2} n} \int dx^{k\gamma} dt^{k\gamma} dx^{\gamma} dt^{\gamma} dy^{\gamma\delta} dr^{\gamma\delta} \Bigg\{ \\
& \exp \beta N \Bigg[\sum_{\gamma} \Big[\frac{1}{2} \sum_{kl} (x^{kl\gamma})^2 + B(x^{\gamma}) + i \sum_{k} t^{k\gamma} x^{k\gamma} + i t^{\gamma} x^{\gamma} \Big] + \\
& + i \sum_{r\gamma} r^{\gamma\delta} y^{\gamma\delta} - \frac{N\alpha}{2} 2\mathrm{Tr}_{\gamma} \log(T_0^{kl} \beta \hat{y} + \frac{1}{2} \sum_{k,\gamma,\delta,l\neq 0} x^{kl\gamma} (T_0^{kl} \beta y)^{-1}_{\gamma\delta} x^{kl\delta} \Bigg] \cdot \\
& \cdot \mathrm{Tr}\{V_i^{\gamma}\} \Bigg\langle \exp \Big[- i\beta \sum_{k,\gamma,i} t^{k\gamma} \Big(\frac{\eta_i^k}{\langle \eta_i \rangle} - 1 \Big) V_i^{\gamma} \Big] \Bigg\rangle \cdot \\
& \cdot \exp \Big[- i\beta \Big(\sum_{\gamma,i} t^{\gamma} V_i^{r} + \sum_{(\gamma,\delta),i} V^{\gamma\delta} V_i^{\gamma} V_j^{\gamma} \Big) \Big] \Bigg\}
\end{aligned} \tag{184}$$

Proceeding through Eq. (A6) one gets to the equivalent of (19) and (20):

$$\begin{aligned}
f = & - T \Bigg\langle \int Dz \log[T_r(h_1, h_2)] \Bigg\rangle - \frac{1}{2} \sum_{k} (x^k)^2 - B(k) - \sum_{k} t^k x^k - tx + \\
& - r_0 y_0 + r_1 y_1 + \frac{\alpha}{2\beta} \Bigg\{ \log \Big[1 - T_0^{kl} \beta (y_0 - y_1) \Big] - \frac{T_0^{kl} \beta y_1}{1 - T_0^{kl} \beta (y_0 - y_1)} \Bigg\}
\end{aligned} \tag{185}$$

with

$$\alpha = \frac{1}{N \overline{T_0}^2} \sum_{k,l\neq 0} (T_0^{k,l})^2 \tag{186}$$

We continue by "undoing", for clarity, the mean field approach, and writing

$$\begin{aligned}
f = & -T \Bigg\langle \frac{1}{N} \sum_{i} \int Dz \log \mathrm{Tr}_{V_i}(h_i, h_2) \Bigg\rangle - \frac{1}{2} \sum_{k} (x^{k,0})^2 - B(x) + \\
& - \sum_{k} t^{k,0} x^{k,0} - tx - r_0 y_0 + r_1 y_1 + \\
& + \frac{1}{2\beta N} \sum_{k,l\neq 0} \Big\{ \log[1 - T_0^{kl} \beta (y_0 - y_1)] - \frac{T_0^{kl} \beta y}{1 - T_0^{kl} \beta (y_0 - y_1)} \Big\}
\end{aligned} \tag{187}$$

where

$$\begin{aligned}
h_1 & = -t - \sum_{k} t^{k,0} \Bigg(\frac{\eta_i^{k,0}}{\langle \eta_i \rangle} - 1 \Bigg) - z(-2\mathrm{Tr}_i)^{\frac{1}{2}} \\
h_2 & = -r_0 + r_1 \\
\frac{1}{g'} & = \frac{1}{g} - 2h_2
\end{aligned} \tag{188}$$

with saddle point equations on conjugate parameters:

$$\begin{aligned} t &= -b(x) \\ t^{k,0} &= -x^{k,0} \\ r_0 &= \frac{1}{2N}\sum_{k,l\neq 0} T_0^{kl}\frac{1-T_0^{kl}[\psi-\beta y_1]}{(T_0^{kl}\psi)^2} \\ r_1 &= -\frac{1}{2N}\sum_{k,l\neq 0}\frac{(T_0^{kl})^2\beta y_1}{(1-T_0^{k,l}\psi)^2} \end{aligned} \tag{189}$$

with

$$\psi = \beta(y_0 - y_1) \tag{190}$$

different from [57].

$$-2Tr_1 \to \frac{1}{N}\sum_{k,l\neq 0}\frac{(T_0^{k,l})^2 y_0}{(1-T_0^{kl}\psi)^2} \tag{191}$$

$$\psi \to T\left\langle \frac{1}{N}\sum_i \int Dz \frac{d^2}{dh^2}\log \mathrm{Tr}_{\{V_i\}}(h_i, h_2)\right\rangle \tag{192}$$

with the other saddle point equations:

$$x = T\left\langle \frac{1}{N}\sum_i \int Dz \frac{d}{dh}\log \mathrm{Tr}_{\{V_i\}}(h_i, h_2)\right\rangle \tag{193}$$

$$y_0 = T\left\langle \frac{1}{N}\sum_i \int Dz \frac{d}{dh_2}\log \mathrm{Tr}_{\{V_i\}}(h_i, h_2)\right\rangle \tag{194}$$

$$x^{k,0} = T\left\langle \frac{1}{N}\sum_i \left(\frac{\eta_i^{k,0}}{\langle\eta_0\rangle}-1\right)\int Dz \frac{d}{dh}\log \mathrm{Tr}_{\{V_i\}}(h_i, h_2)\right\rangle \tag{195}$$

and, in the $T \to 0$ limit:

$$\psi = g'\left\langle \frac{1}{N}\sum_i \int_{h_i>\vartheta} Dz\right\rangle \tag{196}$$

$$y_0 = (g')^2\left\langle \frac{1}{N}\sum_i \int_{h_i>\vartheta} (h_i-\vartheta)^2\right\rangle \tag{197}$$

$$(-2Tr_1) = (g\overline{T_0})^2 = \frac{1}{2N}\sum_{k,l\neq 0}\frac{(T_0^{kl})^2}{(1-T_0^{kl}\psi)^2}y_0 \tag{198}$$

$$x = g'\left\langle \frac{1}{N}\sum_i \int_{h_i>\vartheta} Dz(h_i-\vartheta)\right\rangle \tag{199}$$

$$x^{k,0} = g'\left\langle \frac{1}{N}\sum_i \left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)\int_{h_i>\vartheta} Dz(h_i-\vartheta)\right\rangle \tag{200}$$

$$h_2 = \frac{1}{2N}\sum_{k,l\neq 0}\frac{T_0^{kl}}{1-T_0^{kl}\psi} \tag{201}$$

with

$$h_i = b(x)+\sum_k\left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)x^{k,0}-z(\rho\overline{T_0}) \tag{202}$$

$$\begin{aligned} f = &-\frac{g'}{2}\left\langle \frac{\sum_i}{N}\int_{h_i>\vartheta} Dz(h_i-\vartheta)^2\right\rangle+\frac{1}{2}\sum_k (x^{k,0})^2+\\ &+xb(x)-B(x)+\frac{(g\overline{T_0})^2}{2}\psi \end{aligned} \tag{203}$$

If

$$v^{k,0} = \frac{x^{k,0}}{\overline{T_0}\rho} \tag{204}$$

$$w = \frac{(b(x)-\vartheta)}{\overline{T_0}\rho} \tag{205}$$

then

$$h_i-\vartheta = (\rho\overline{T_0})\left[w+\sum_k\left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)v^{k,0}-z\right] \tag{206}$$

$$v^{k,0} = g'\left\langle \frac{1}{N}\sum_i\left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)\int^{+} Dz\left[w+\sum_k\left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)v^{k,0}-z\right]\right\rangle \tag{207}$$

$$\frac{x}{\rho\overline{T_0}} = g'\left\langle \frac{1}{N}\sum_i\int^{+} Dz\left[w+\sum_k\left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle}-1\right)v^{k,0}-z\right]\right\rangle \tag{208}$$

$$(\rho_0\overline{T_0})^2 = \frac{1}{2N}\sum_{k,l\neq 0}\frac{(T_0^{kl})^2}{(1-T_0^{kl}\psi)^2} \tag{209}$$

$$\begin{aligned} y_0 =&(g')^2\left\langle \frac{1}{N}\sum_i \int^{+} Dz \left[w + \sum_k \left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle} - 1\right)\nu^{k,0} + \right.\right. \\ &\left.\left. - z\right]^2 \right\rangle \frac{(\rho_0 \overline{T_0})^2}{N} \sum_{k,l\neq 0} \frac{(T_0^{kl})^2}{(1-T_0^{kl}\psi)^2} \end{aligned} \tag{210}$$

Finally, since

$$\frac{1}{g'} = \frac{1}{g} - 2h_2 = \frac{1}{g} - \frac{1}{N}\sum_{k,l\neq 0}\frac{T_0^{kl}}{(1-T_0^{kl}\psi)} \tag{211}$$

$$\frac{\nu^{k,0}}{g'} = \frac{1}{N}\sum_i \left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle} - 1\right)\mathcal{N}(\hbar_i) \tag{212}$$

$$\left(\frac{1}{g'}\right)^2 = \frac{1}{N}\sum_i \mathcal{M}(\hbar_i)\frac{\sum_{k,l\neq 0}}{N}\frac{(T_0^{kl})^2}{(1-T_0^{kl}\psi)^2} \tag{213}$$

$$\frac{\psi}{g'} = \frac{1}{N}\sum_i \Phi(\hbar_i) \tag{214}$$

$$\mathcal{N}(x) = x\phi(x) + \sigma(x) \tag{215}$$

$$\mathcal{M}(x) = (1+x^2)\phi(x) + x\sigma(x) \tag{216}$$

$$\hbar_i = w + \sum_k \left(\frac{\eta_i^{k,0}}{\langle\eta_i\rangle} - 1\right)\nu^{k,0} \tag{217}$$

H.1 STABILITY OF THE SADDLE POINTS WITH SPARSE CONNECTIVITY

Written as "equations of motion" and defining

$$\alpha = \frac{1}{N\overline{T_0}^2}\sum_{k,l\neq 0}(T_0^{k,l})^2 \tag{218}$$

for the "uncondensed patterns". We have

$$x(t+1) = g'\overline{T_0}\rho(t)\frac{1}{N}\sum_i \mathcal{N}(\hbar_i(t)) \tag{219}$$

$$x(s,t+1) = g'\overline{T_0}\rho(t)\frac{1}{N}\sum_i \left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\mathcal{N}(\hbar_i(t)). \tag{220}$$

For the fully connected model:

$$\overline{T_0}^2 g^2(t+1) = (g'\overline{T_0})^2 g^2(t)\frac{\sum_{k,l\neq 0}}{N}\frac{(T_0^{kl})^2}{(1-T_0^{kl}\psi)^2}\frac{1}{N}\sum_i \mathcal{M}(\hbar_i(t)) \tag{221}$$

with

$$\psi(t+1) = g'\frac{1}{N}\sum_i \varphi(\hbar_i) \tag{222}$$

$$\frac{1}{g'} = \frac{1}{g} - \frac{1}{N}\sum_{k,l\neq 0}\frac{(T_0^{kl})}{(1-T_0^{kl}\psi)} \tag{223}$$

$$\hbar_i(t) = \frac{b(x)-\vartheta}{\overline{T_0}\rho(t)} + \int ds\left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\frac{x(s,t)}{\overline{T_0}\rho(t)} \tag{224}$$

while, for the highly diluted model

$$\begin{aligned} g' &= g \\ \rho^2(t+1) &= g^2\rho^2(t)\alpha\frac{\overline{T_0}^2}{N}\sum_i \mathcal{M}(\hbar_i(t)) \end{aligned} \tag{225}$$

To find the stability of a fixed point, we focus on the highly diluted case and linearize the previous equations with $x(t)$ =constant.

$$\begin{aligned} &x(s,t+1) = x_0(s) + \lambda\delta x(s) \approx g\overline{T_0}\rho_0\frac{1}{N}\sum_i \left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\mathcal{N}(\hbar_i(t)) + \\ &+ g\overline{T_0}\frac{1}{N}\sum_i \left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\Bigg\{\varphi(\hbar_{i_0})\int\frac{ds'}{S}\left(\frac{\eta_i(s')}{\langle\eta\rangle} - 1\right)\frac{\delta x(s')}{\overline{T_0}} + \\ &\delta\rho\mathcal{N}(\hbar_0) - \delta\rho\varphi(\hbar_{i_0})\hbar_{i_0}\Bigg\} \end{aligned} \tag{226}$$

$$\begin{aligned} &\rho_0^2 + \lambda 2\rho_0\delta\rho \approx g^2\alpha\rho_0^2\frac{(\overline{T_0})^2}{N}\sum_i \mathcal{M}(\hbar_{i_0}) + \alpha g^2\rho_0\frac{(\overline{T_0})^2}{N}\sum_i \Bigg\{ \\ &2\delta\rho\mathcal{M}(\hbar_{i_0}) - 2\mathcal{N}(\hbar_{i_0})\hbar_{i_0}\delta\rho + 2\mathcal{N}(h_{i_0})\int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle} - 1\right)\frac{\delta x(s')}{\overline{T_0}}\Bigg\} \end{aligned} \tag{227}$$

with the condition

$$0=\delta x=\frac{1}{N}\sum_i\left\{\phi(\hbar_{i0})\int\frac{ds'}{S}\left(\frac{\eta_i(s')}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}}+\delta\rho\sigma(\hbar_{i0})\right\}. \tag{228}$$

$$\delta\rho=-\frac{\sum_i\phi(\hbar_{i0})\int\frac{ds'}{S}\left(\frac{\eta_i(s')}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}}}{\sum_i\sigma(\hbar_{i0})} \tag{229}$$

$$\begin{aligned}\lambda\delta x(s)=&g\overline{T_0}\frac{1}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\left\{\phi(\hbar_{i0})\int\frac{ds'}{S}\left(\frac{\eta_i(s')}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}}+\right.\\&\left.+\delta\rho\sigma(\hbar_{i0})\right\}\end{aligned} \tag{230}$$

$$\lambda\delta\rho=\alpha g^2\frac{(\overline{T_0})^2}{N}\sum\left\{\phi(\hbar_{i0})\delta\rho+\mathcal{N}(\hbar_{i0})\int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}}\right\} \tag{231}$$

and the fixed point conditions

$$1=\alpha g^2\frac{\overline{T_0}^2}{N}\sum_i\mathcal{M}(\hbar_{i_0}) \tag{232}$$

$$x(s)=g\overline{T_0}\rho\frac{1}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\mathcal{N}(\hbar_i) \tag{233}$$

and

$$\delta x(s)=0=\frac{1}{N}\sum_i\left\{\phi(\hbar_{i0})\int\frac{ds'}{S}\left(\frac{\eta_i(s')}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}}+\delta\rho\sigma(\hbar_{i0})\right\} \tag{234}$$

Let us make the following ansatz

$$\begin{cases}\delta x(s)=\epsilon\left\{-x(s)+\gamma\frac{1}{N}\sum_j\left(\frac{\eta_j(s)}{\langle\eta\rangle}-1\right)\phi(\hbar_{j0})+\zeta\right\}\\ \delta\rho=\epsilon\Delta\end{cases} \tag{235}$$

Note that

$$\begin{aligned}
&\int \frac{ds'}{S}\left(\frac{\eta_j(s)}{\langle\eta\rangle}-1\right)\frac{\delta x(s')}{\overline{T_0}} = -\epsilon\int \frac{ds'}{S}\left(\frac{\eta_j(s)}{\langle\eta\rangle}-1\right)\frac{x(s)}{\overline{T_0}}+ \\
&+\epsilon\frac{\gamma}{N}\sum_j\int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\left(\frac{\eta_j(s)}{\langle\eta\rangle}-1\right)\phi(\hbar_{j0})+ \\
&+\epsilon\cancelto{0 \text{ or maybe not?}}{\frac{\zeta}{\overline{T_0}}\int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)} \qquad\qquad = \\
&-\epsilon\left[\hbar_{i0}\rho_0-\frac{b(x)-\vartheta}{T_0}\right]+\frac{\epsilon\gamma}{N}(T_{ii}\phi(\hbar_{i0})+\sum_{i\neq j}T_{ij}\phi(\hbar_{j0}))
\end{aligned} \tag{236}$$

where

$$\begin{aligned}
T_{ii} &= \int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)^2 \\
T_{ij} &= \int\frac{ds'}{S}\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\left(\frac{\eta_j(s)}{\langle\eta\rangle}-1\right)
\end{aligned} \tag{237}$$

So we can re-write Eq. (230) as

$$\begin{aligned}
&-\epsilon\lambda x(s)+\epsilon\lambda\frac{\gamma}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\phi(\hbar_{i0})+\epsilon\lambda\zeta = \\
&g\overline{T_0}\frac{1}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\left\{\epsilon\Delta\sigma(\hbar_{i0})+\frac{\epsilon\gamma}{N}\left[\phi(\hbar_{i0})^2T_{ii}+\sum_{i\neq j}T_{ij}\phi(\hbar_{j0})\phi(\hbar_{i0})\right]+\right. \\
&\left.-\epsilon\phi(\hbar_{i0})\left[\hbar_{i0}\rho_0-\frac{b(x)-\vartheta}{T_0}\right]\right.
\end{aligned} \tag{238}$$

Dividing by ϵ and replacing $\phi(\hbar_{i0})$ with $\mathcal{N}(\hbar_{i0})-\sigma(\hbar_{i0})$

$$\begin{aligned}
&-\ \lambda x(s)+\ \lambda\frac{\gamma}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\phi(\hbar_{i0})+\ \lambda\zeta = \\
&-x(s)+g\overline{T_0}\frac{1}{N}\sum_i\left(\frac{\eta_i(s)}{\langle\eta\rangle}-1\right)\left\{(\Delta+\rho_0)\sigma(\hbar_{i0})+\phi(\hbar_{i0})\left\{\frac{b(x)-\vartheta}{T_0}+\right.\right. \\
&\left.\left.+\frac{\gamma}{N}\left[\phi(\hbar_{i0})T_{ii}+\sum_{i\neq j}T_{ij}\phi(\hbar_{j0})\right]\right\}\right.
\end{aligned} \tag{239}$$

And we can rewrite Eq. (231) as:

$$\begin{aligned}
&\lambda\Delta = \alpha g^2\frac{T_0^2}{N}\sum_i\mathcal{N}(\hbar_{i0})\left\{\frac{\gamma}{N}\langle\phi(\hbar_{i0})T_{ii}+\sum_{i\neq j}T_{ij}\phi(\hbar_{i0})(\hbar_{i0})\rangle+\frac{b(x)-\vartheta}{\overline{T_0}}\right\}+ \\
&+\alpha g^2T_0^2(\Delta+\rho_0)-\rho_0
\end{aligned} \tag{240}$$

I VOWEL CHARTS: METHODS AND ADDITIONAL FIGURES

I.1 RECORDING METHOD AND SOFTWARE

Vowels were recorded from participants using RecordPad, through ordinary headphones microphones and each vowel was recorded in a separate file. In our pilot studies, presented in this thesis, the length of each file was not standardized and depended on the time needed by the subject to stop the recording of each vowel.
Sounds were analyzed using Praat software [167]. In particular, formants were extracted in Python through Parselmouth [168], an open-source library easing the use of Praat in a general python script.

In Fig. 44 (second line) we show the first four formants time traces, as found by praat, highlighted in red and superimposed to the relative spectograms, for seven italian vowels pronounced by a subject. In order to define the first two formants of each vowel as an individual value, each specific time trace was averaged in time. Formants were then converted from Hertz to Barks using a standard formula which can be found in Ref. [145, 146].

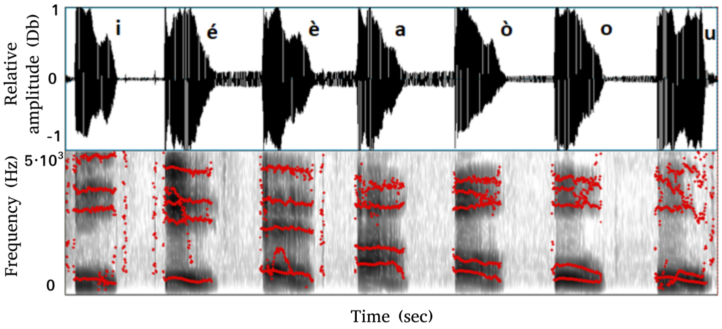


Figure 44: Sample plot obtained with the use of Parselmouth library (Praat): first row) oscillograms of seven sample italian vowels. Second row) Relative spectograms, red traces indicate formants.

I.2 SEPARATED BILINGUAL VOWEL CHARTS

Here I report separately the conjunctive vowels charts shown in Fig. 29.

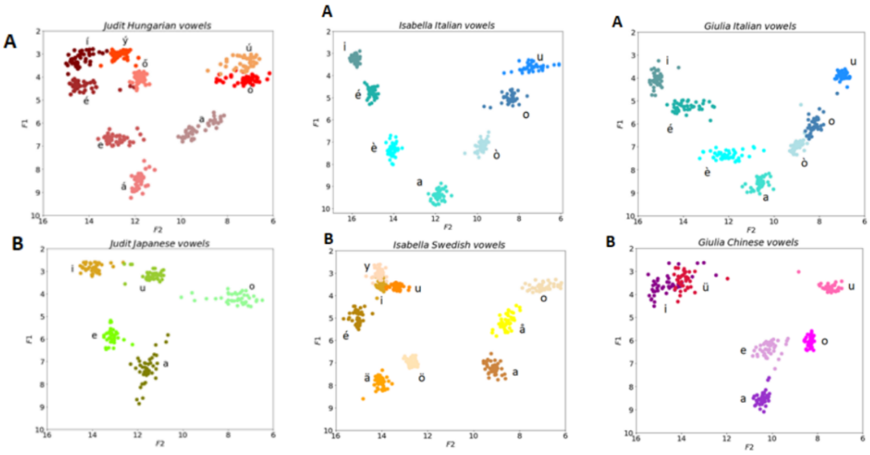


Figure 45: Supplement to Fig. 29, separate vowel charts for the three subjects described in the mai text.

BIBLIOGRAPHY

[1] L. Furumoto, “Mary whiton calkins (1863–1930),” Psychology of Women Quarterly **5**, 55–68 (1980).

[2] S. Madigan and R. O’Hara, “Short-term memory at the turn of the century: mary whiton calkins’s memory research.,” American Psychologist **47**, 170 (1992).

[3] M. W. Calkins, “Association. an essay analytic and experimental.,” The Psychological Review: Monograph Supplements **1**, i (1896).

[4] C. E. Murchison, *A history of psychology in autobiography vol. i.* (Russell & Russell/Atheneum Publishers, visible at psychclassics.yorku.ca/Calkins/murchison.htm, 1930).

[5] “www.apa.org/about/governance/president/former-presidents,” Former APA Presidents.

[6] S. Smith, “Calkins, mary whiton (1863–1930),” Women in World History: A Biographical Encyclopedia. Encyclopedia.com (2021).

[7] S. Ramon y Cajal, “Histologie du système nerveux de l’homme et des vertébrés,” Maloine, Paris **2**, 153–173 (1911).

[8] W. B. Scoville and B. Milner, “Loss of recent memory after bilateral hippocampal lesions,” Journal of neurology, neurosurgery, and psychiatry **20**, 11 (1957).

[9] J. D. Hogan and N. Frishberg, “The general psychologist | april 2015,”

[10] E. C. Tolman, “Cognitive maps in rats and men.,” Psychological review **55**, 189 (1948).

[11] N. Brunel, “Daniel amit (1938–2007),” Network: Computation in Neural Systems **19**, 3–8 (2008).

[12] “Numero speciale in memoria del prof daniel j. amit,” Accastampato **7** (2003).

[13] D. J. Amit, H. Gutfreund, and H. Sompolinsky, “Storing infinite numbers of patterns in a spin-glass model of neural networks,” Physical Review Letters **55**, 1530 (1985).

[14] E. Gardner, “The space of interactions in neural network models,” Journal of physics A: Mathematical and general **21**, 257 (1988).

[15] D. J. Amit, "La scienza nei tempi del neoliberismo - occorrono valide regole di comportamento e coerenza metodologica-," Prometeo - Rivista trimestrale di scienza e storia, n.92 (2005). A copy of the article is available at anticitera.org/archivio/sulla-critica-della-scienza-jean-bricmont-e-daniel-amit/.

[16] "http://luis.impa.br/politics/guerra/carta.html," Letters between Daniel Amit and Daniel Blume (2003).

[17] "luis.impa.br/politics/guerra/amit.html," Arab News.

[18] See for example the Italian scientific magazines SE/scienza esperienza and Sapere, or the works of Lewis Mumford, Ivan Illich, Günther Anders, Pierre Thuillier among others.

[19] S. Zuboff, *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (BBS PubblicAffairs, 2018).

[20] J. M. Fossaceca and S. H. Young, "Artificial intelligence and machine learning for future army applications," in Ground/air multisensor interoperability, integration, and networking for persistent isr ix, Vol. 10635 (International Society for Optics and Photonics, 2018), p. 1063507.

[21] A. Treves, "Non pervenuti," HA KEILLAH, "Foglio bimestrale del gruppo di studi ebraici di Torino", visible at www.hakeillah.com/3_18_11.htm (2018).

[22] A. Capocci, "A trieste una giornata per daniel amit," Il Manifesto (2018).

[23] J. O'Keefe and J. Dostrovsky, "The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat.," Brain research (1971).

[24] M.-B. Moser, D. C. Rowland, and E. I. Moser, "Place cells, grid cells, and memory," Cold Spring Harbor perspectives in biology **7**, a021808 (2015).

[25] T. Solstad, C. N. Boccara, E. Kropff, M.-B. Moser, and E. I. Moser, "Representation of geometric borders in the entorhinal cortex," Science **322**, 1865–1868 (2008).

[26] E. Kropff, J. E. Carmichael, M.-B. Moser, and E. I. Moser, "Speed cells in the medial entorhinal cortex," Nature **523**, 419–424 (2015).

[27] J. Ranck Jr, "Head direction cells in the deep layer of dorsal presubiculum in freely moving rats," in Society of neuroscience abstract, Vol. 10 (1984), p. 599.

[28] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," Nature **436**, 801 (2005).

[29] K. Gerlei, J. Passlack, I. Hawes, B. Vandrey, H. Stevens, I. Papastathopoulos, and M. Nolan, "Grid cells encode local head direction," BioRxiv, 681312 (2020).

[30] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser, "Conjunctive representation of position, direction, and velocity in entorhinal cortex," Science **312**, 758–762 (2006).

[31] H. Stensola, T. Stensola, T. Solstad, K. Frøland, M.-B. Moser, and E. I. Moser, "The entorhinal grid map is discretized," Nature **492**, 72–78 (2012).

[32] D. B. Omer, S. R. Maimon, L. Las, and N. Ulanovsky, "Social place-cells in the bat hippocampus," Science **359**, 218–224 (2018).

[33] S. S. Deshmukh and J. J. Knierim, "Representation of non-spatial and spatial information in the lateral entorhinal cortex," Frontiers in behavioral neuroscience **5**, 69 (2011).

[34] M. Fyhn, T. Hafting, A. Treves, M.-B. Moser, and E. I. Moser, "Hippocampal remapping and grid realignment in entorhinal cortex," Nature **446**, 190–194 (2007).

[35] R. U. Mulner and J. L. Kubie, "The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells," The Journal of Neuroscience **7**, 1951–1968 (1987).

[36] J. L. Kubie and R. U. Mulner, "Multiple representations in the hippocampus.," Hippocampus **1**, 240–242 (1991).

[37] H Eichenbaum, M Kuperstein, A Fagan, and J Nagode, "Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task," Journal of Neuroscience **7**, 716–732 (1987).

[38] D. Derdikman, J. R. Whitlock, A. Tsao, M. Fyhn, T. Hafting, M.-B. Moser, and E. I. Moser, "Fragmentation of grid cell maps in a multicompartment environment," Nature neuroscience **12**, 1325–1332 (2009).

[39] C. N. Boccara, M. Nardin, F. Stella, J. O'Neill, and J. Csicsvari, "The entorhinal cognitive map is attracted to goals," Science **363**, 1443–1447 (2019).

[40] J. Krupic, M. Bauza, S. Burton, C. Barry, and J. O'Keefe, "Grid cell symmetry is shaped by environmental geometry," Nature **518**, 232–235 (2015).

[41] B. Dunn, D. Wennberg, Z. Huang, and Y. Roudi, "Grid cells show field-to-field variability and this explains the aperiodic response of inhibitory interneurons," arXiv preprint arXiv:1701.04893 (2017).

[42] B. R. Kanter, C. M. Lykken, D. Avesar, A. Weible, J. Dickinson, B. Dunn, N. Z. Borgesius, Y. Roudi, and C. G. Kentros, "A novel mechanism for the grid-to-place cell transformation revealed by transgenic depolarization of medial entorhinal cortex layer ii," Neuron **93**, 1480–1492 (2017).

[43] D. Marr, "Simple memory: a theory for archicortex," Philosophical transactions of the Royal Society of London. Series B, Biological sciences **262**, 23–81 (1971).

[44] J. O'Keefe and L. Nadel, *The hippocampus as a cognitive map* (Oxford: Clarendon Press, 1978).

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

[46] D. O. Hebb, *The organization of behavior: a neuropsychological theory* (J. Wiley; Chapman & Hall, 1949).

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

[48] J. Hertz, A. Krogh, R. G. Palmer, and H. Horner, "Introduction to the theory of neural computation," Physics Today **44**, 70 (1991).

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

[50] E Gardner, S. Mertens, and A. Zippelius, "Retrieval properties of a neural network with an asymmetric learning rule," Journal of Physics A: Mathematical and General **22**, 2009 (1989).

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

[52] J.-P. Nadal, "Associative memory: on the (puzzling) sparse coding limit," Journal of Physics A: Mathematical and General **24**, 1093 (1991).

[53] A. Treves, "Dilution and sparse coding in threshold-linear nets," Journal of Physics A: Mathematical and General **24**, 327 (1991).

[54] A. Treves and E. T. Rolls, "What determines the capacity of autoassociative memories in the brain?" Network: Computation in Neural Systems **2**, 371–397 (1991).

[55] F. P. Battaglia and A. Treves, "Attractor neural networks storing multiple space representations: a model for hippocampal place fields," Physical Review E **58**, 7738 (1998).

[56] A. Treves, "Threshold-linear formal neurons in auto-associative nets," Journal of Physics A: Mathematical and General **23**, 2631 (1990).

[57] A. Treves, "Graded-response neurons and information encodings in autoassociative memories," Physical Review A **42**, 2418 (1990).

[58] H. K. Hartline and F. Ratliff, "Spatial summation of inhibitory influences in the eye of limulus, and the mutual interaction of receptor units," The Journal of general physiology **41**, 1049–1066 (1958).

[59] F. Ratliff, "Mach bands: quantitative studies on neural networks," Retina. San Francisco, CA: Holden-Day (1965).

[60] Y. Roudi and A. Treves, "Localized activity profiles and storage capacity of rate-based autoassociative networks," Physical Review E **73**, 061904 (2006).

[61] W. E. Skaggs, J. J. Knierim, H. S. Kudrimoti, and B. L. McNaughton, "A model of the neural basis of the rat's sense of direction.," Advances in Neural Information Processing Systems **7**, 173–180 (1995).

[62] M. B. Zugaro, A. Arleo, A. Berthoz, and S. I. Wiener, "Rapid spatial reorientation and head direction cells," Journal of Neuroscience **23**, 3478–3482 (2003).

[63] A. Samsonovich and B. L. McNaughton, "Path integration and cognitive mapping in a continuous attractor neural network model.," The Journal of Neuroscience **17**, 5900–5920 (1997).

[64] D Sherrington, "Special issue in memory of elizabeth gardner 1957-1988," Journal of Physics A - Mathematical and General, volume=22, number=12, year=1989.

[65] E. Domany, J. L. van Hemmen, and K. Schulten, *Models of neural networks i* (Springer Science & Business Media, 2012).

[66] F. Schönsberg, Y. Roudi, and A. Treves, "Efficiency of local learning rules in threshold-linear associative networks," Physical Review Letters **126**, 018301 (2021).

[67] F. Schönsberg, Y. Roudi, and A. Treves, (forthcoming).

[68] T. H. Brown, E. W. Kairiss, and C. L. Keenan, "Hebbian synapses: biophysical mechanisms and algorithms," Annual review of neuroscience **13**, 475–511 (1990).

[69] D. J. Amit, *Modeling brain function: the world of attractor neural networks* (Cambridge university press, 1992).

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

[71] H. K. Hartline and F. Ratliff, "Inhibitory interaction of receptor units in the eye of limulus," The Journal of general physiology **40**, 357–376 (1957).

[72] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning (icml-10) (2010), pp. 807–814.

[73] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in Proc. icml, 1 (2013), p. 3.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in Proceedings of the ieee international conference on computer vision (2015), pp. 1026–1034.

[75] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[76] U. Pereira and N. Brunel, "Attractor dynamics in networks with learning rules inferred from in vivo data," Neuron **99**, 227–238 (2018).

[77] D. Bollé, R Kuhn, and J Van Mourik, "Optimal capacity of graded-response perceptrons," Journal of Physics A: Mathematical and General **26**, 3149 (1993).

[78] A. Treves, S. Panzeri, E. T. Rolls, M. Booth, and E. A. Wakeman, "Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli," Neural Computation **11**, 601–631 (1999).

[79] J. M. Fuster and J. P. Jervey, "Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli," Science **212**, 952–955 (1981).

[80] Y. Miyashita, "Neuronal correlate of visual associative longterm memory in the primate temporal cortex," Nature **335**, 817–820 (1988).

[81] K. Nakamura and K. Kubota, "Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task," Journal of neurophysiology **74**, 162–178 (1995).

[82] D. J. Amit, S. Fusi, and V. Yakovlev, "Paradigmatic working memory (attractor) cell in it cortex," Neural computation **9**, 1071–1092 (1997).

[83] E. T. Rolls and A. Treves, *Neural networks and brain function* (Oxford university press, Oxford, 1998).

[84] S. Lim, J. L. McKee, L. Woloszyn, Y. Amit, D. J. Freedman, D. L. Sheinberg, and N. Brunel, "Inferring learning rules from distributions of firing rates in cortical neurons," Nature neuroscience **18**, 1804–1810 (2015).

[85] G. Buzsáki and K. Mizuseki, "The log-dynamic brain: how skewed distributions affect network operations," Nature Reviews Neuroscience **15**, 264–278 (2014).

[86] K. Yoon, M. A. Buice, C. Barry, R. Hayman, N. Burgess, and I. R. Fiete, "Specific evidence of low-dimensional continuous attractor dynamics in grid cells," Nature neuroscience **16**, 1077 (2013).

[87] A. Treves, "Are spin-glass effects relevant to understanding realistic auto-associative networks?" Journal of Physics A: Mathematical and General **24**, 2645 (1991).

[88] Y. Roudi and A. Treves, "Disappearance of spurious states in analog associative memories," Physical Review E **67**, 041906 (2003).

[89] C. Baldassi, E. M. Malatesta, and R. Zecchina, "Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations," Physical review letters **123**, 170602 (2019).

[90] C. Clopath and N. Brunel, "Optimal properties of analog perceptrons with excitatory weights," PLoS computational biology **9** (2013).

[91] T. J. Teyler and P. DiScenna, "The hippocampal memory indexing theory.," Behavioral neuroscience **100**, 147 (1986).

[92] J.-P. Nadal and G. Toulouse, "Information storage in sparsely coded memory nets," Network: Computation in Neural Systems **1**, 61–74 (1990).

[93] Y. Roudi and P. E. Latham, "A balanced memory network," PLoS Comput Biol **3**, e141 (2007).

[94] S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton, and T. Lillicrap, "Assessing the scalability of biologically-motivated deep learning algorithms and architectures," in *Advances in neural information processing systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), pp. 9368–9378.

[95] Y. Amit, "Deep learning with asymmetric connections and hebbian updates," Frontiers in computational neuroscience **13**, 18 (2019).

[96] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," Proceedings of the National Academy of Sciences **116**, 7723–7731 (2019).

[97] L. Harten, A. Katz, A. Goldshtein, M. Handel, and Y. Yovel, "The ontogeny of a mammalian cognitive map in the real world," Science **369**, 194–197 (2020).

[98] J. B. Calhoun, *The ecology and sociology of the norway rat* (US Department of Health, Education, and Welfare, Public Health Service, 1963).

[99] M. Jung and B. McNaughton, "Spatial selectivity of unit activity in the hippocampal granular layer," Hippocampus **3**, 165–182 (1993).

[100] K. B. Kjelstrup, T. Solstad, V. H. Brun, T. Hafting, S. Leutgeb, M. P. Witter, E. I. Moser, and M.-B. Moser, "Finite scale of spatial representation in the hippocampus," Science **321**, 140–143 (2008).

[101] P. D. Rich, H.-P. Liaw, and A. K. Lee, "Large environments reveal the statistical structure governing hippocampal representations," Science **345**, 814–817 (2014).

[102] A. A. Fenton, H.-Y. Kao, S. A. Neymotin, A. Olypher, Y. Vayntrub, W. W. Lytton, and N. Ludvig, "Unmasking the ca1 ensemble place code by exposures to small and large environments: more place cells and multiple, irregularly arranged, and expanded place fields in the larger space," Journal of Neuroscience **28**, 11250–11262 (2008).

[103] E. Park, D. Dvorak, and A. A. Fenton, "Ensemble place codes in hippocampus: ca1, ca3, and dentate gyrus place cells have multiple place fields in large environments," PloS one **6**, e22349 (2011).

[104] V. H. Brun, T. Solstad, K. B. Kjelstrup, M. Fyhn, M. P. Witter, E. I. Moser, and M.-B. Moser, "Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex," Hippocampus **18**, 1200–1212 (2008).

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

[106] E. J. Henriksen, L. L. Colgin, C. A. Barnes, M. P. Witter, M.-B. Moser, and E. I. Moser, "Spatial representation along the proximodistal axis of ca1," Neuron **68**, 127–137 (2010).

[107] M. Greenwood and G. U. Yule, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," Journal of the Royal statistical society **83**, 255–279 (1920).

[108] J. S. Lee, J. J. Briguglio, J. D. Cohen, S. Romani, and A. K. Lee, "The statistical structure of the hippocampal code for space as a function of time, context, and value," Cell **183**, 620–635 (2020).

[109] N. Ulanovsky and C. F. Moss, "Hippocampal cellular and network activity in freely moving echolocating bats," Nature neuroscience **10**, 224–233 (2007).

[110] M. M. Yartsev, M. P. Witter, and N. Ulanovsky, "Grid cells without theta oscillations in the entorhinal cortex of bats," Nature **479**, 103–107 (2011).

[111] M. Geva-Sagiv, L. Las, Y. Yovel, and N. Ulanovsky, "Spatial cognition in bats and rats: from sensory acquisition to multiscale maps and navigation," Nature Reviews Neuroscience **16**, 94–108 (2015).

[112] M. M. Yartsev and N. Ulanovsky, "Representation of three-dimensional space in the hippocampus of flying bats," Science **340**, 367–372 (2013).

[113] A. Finkelstein, D. Derdikman, A. Rubin, J. N. Foerster, L. Las, and N. Ulanovsky, "Three-dimensional head-direction coding in the bat brain," Nature **517**, 159–164 (2015).

[114] G. Ginosar, "3d spatial representation: coding of 3d space by 3d grid cells, 3d border cells, 3d head direction cells," in (2018), Poster CogNav.

[115] T. Eliav, M. Geva-Sagiv, M. M. Yartsev, A. Finkelstein, A. Rubin, L. Las, and N. Ulanovsky, "Nonoscillatory phase coding and synchronization in the bat hippocampal formation," Cell **175**, 1119–1130 (2018).

[116] T. Eliav, S. R. Maimon, J. Aljadeff, M. Tsodyks, G. Ginosar, L. Las, and N. Ulanovsky, "Multiscale representation of very large environments in the hippocampus of flying bats," Science **372** (2021).

[117] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," Biological cybernetics **27**, 77–87 (1977).

[118] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky, "Theory of orientation tuning in visual cortex.," Proceedings of the National Academy of Sciences **92**, 3844–3848 (1995).

[119] M. Tsodyks and T. Sejnowski, "Associative memory and hippocampal place cells," International journal of neural systems **6**, 81–86 (1995).

[120] M. Tsodyks, "Attractor neural network models of spatial maps in hippocampus," Hippocampus **9**, 481–489 (1999).

[121] H. S. Seung, "How the brain keeps the eyes still," Proceedings of the National Academy of Sciences **93**, 13339–13344 (1996).

[122] K. Zhang, "Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory," Journal of Neuroscience **16**, 2112–2126 (1996).

[123] D. Lee, B. Reis, H. Seung, and D. Tank, "Nonlinear network models of the oculomotor integrator," in *Computational neuroscience* (Springer, 1997), pp. 371–377.

[124] M. Camperi and X.-J. Wang, "A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability," Journal of computational neuroscience **5**, 383–405 (1998).

[125] A. Compte, N. Brunel, P. S. Goldman-Rakic, and X.-J. Wang, "Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model," Cerebral cortex **10**, 910–923 (2000).

[126] S. Wu, K. Hamaguchi, and S.-i. Amari, "Dynamics and computation of continuous attractors," Neural computation **20**, 994–1025 (2008).

[127] C. A. Fung, K. M. Wong, and S. Wu, "A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks," Neural Computation **22**, 752–792 (2010).

[128] Y. Burak and I. R. Fiete, "Fundamental limits on persistent activity in networks of noisy neurons," Proceedings of the National Academy of Sciences **109**, 17645–17650 (2012).

[129] R. Monasson and S. Rosay, "Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: phase diagram," Physical review E **87**, 062813 (2013).

[130] A. Treves, "Computational constraints between retrieving the past and predicting the future, and the CA3-CA1 differentiation.," Hippocampus **14**, 539–556 (2004).

[131] G. Papp, M. P. Witter, and A. Treves, "The ca3 network as a memory store for spatial representations," Learning & memory **14**, 732–744 (2007).

[132] J. J. Hopfield, "Neurodynamics of mental exploration," Proceedings of the National Academy of Sciences **107**, 1648–1653 (2010).

[133] E. Cerasti and A. Treves, "The spatial representations acquired in CA3 by self-organizing recurrent connections," Frontiers in cellular neuroscience **7**, 112 (2013).

[134] R. Monasson and S. Rosay, "Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: collective motion of the activity," Physical review E **89**, 032803 (2014).

[135] S. Stringer, T. Trappenberg, E. Rolls, and I. Araujo, "Self-organizing continuous attractor networks and path integration: one-dimensional models of head direction cells," Network: Computation in Neural Systems **13**, 217–242 (2002).

[136] S. Stringer, E. Rolls, T. Trappenberg, and I. De Araujo, "Self-organizing continuous attractor networks and path integration: two-dimensional models of place cells," Network: Computation in Neural Systems **13**, 429–446 (2002).

[137] A. Renart, P. Song, and X.-J. Wang, "Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks," Neuron **38**, 473–485 (2003).

[138] K. Hamaguchi, J. Hatchett, and M. Okada, "Analytic solution of neural network with disordered lateral inhibition," Physical Review E **73**, 051104 (2006).

[139] V. Itskov, D. Hansel, and M. Tsodyks, "Short-term facilitation may stabilize parametric working memory trace," Frontiers in computational neuroscience **5**, 40 (2011).

[140] W. Zhong, Z. Lu, D. J. Schwab, and A. Murugan, "Nonequilibrium statistical mechanics of continuous attractors," Neural computation **32**, 1033–1068 (2020).

[141] A. Treves, "Computational constraints that may have favoured the lamination of sensory cortex," Journal of Computational Neuroscience **14**, 271–282 (2003).

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

[143] D. Spalla, I. M. Cornacchia, and A. Treves, "Continuous attractors for dynamic memories," bioRxiv (2020).

[144] R. Monasson and S. Rosay, "Transitions between spatial attractors in place-cell models," Physical review letters **115**, 098101 (2015).

[145] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," The Journal of the Acoustical Society of America **33**, 248–248 (1961).

[146] Z. G. Kaya et al., "Cross-linguistic exploration of phonemic representations," (2018).

[147] E. E. Loos, *Glossary of linguistic terms* (SIL International, 2004).

[148] A. Caramazza, D. Chialant, R. Capasso, and G. Miceli, "Separable processing of consonants and vowels," Nature **403**, 428–430 (2000).

[149] D Boatman, C Hall, M. H. Goldstein, R Lesser, and B. Gordon, "Neuroperceptual differences in consonant and vowel discrimination: as revealed by direct cortical electrical interference," Cortex **33**, 83–98 (1997).

[150] D. Jones, *An english pronouncing dictionary:(on strictly phonetic principles)* (JM Dent, 1917).

[151] A. M. Bell, "Visible speech or self-interpreting physiological letters for the writing of all languages in one alphabet," Simpkin and Marshall, London (1867).

[152] M. Lindau, "Vowel features," Language **54**, 541–563 (1978).

[153] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," The Journal of the Acoustical Society of America **27**, 338–352 (1955).

[154] I. R. Macpherson, *Spanish phonology: descriptive and historical* (Manchester University Press, 1975).

[155] A. D. Manca and M. Grimaldi, "Vowels and consonants in the brain: evidence from magnetoencephalographic studies on the n1m in normal-hearing listeners," Frontiers in Psychology **7**, 1413 (2016).

[156] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," The Journal of the Acoustical Society of America **33**, 1344–1356 (1961).

[157] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," The Journal of the Acoustical Society of America **87**, 2592–2605 (1990).

[158] T. Moser, A. Brandt, and A. Lysakowski, "Hair cell ribbon synapses," Cell and tissue research **326**, 347–359 (2006).

[159] J. E. Flege, C. Schirru, and I. R. MacKay, "Interaction between the native and second language phonetic subsystems," Speech communication **40**, 467–491 (2003).

[160] Z. Kaya, J. Collins, and A. Treves, (forthcoming).

[161] P. Iverson and P. K. Kuhl, "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," The Journal of the Acoustical Society of America **97**, 553–562 (1995).

[162] R. Burwell and K. Agster, "Anatomy of the hippocampus and the declarative memory system, chapter 3.03," Learning and memory-A comprehensive reference. Volume **3** (2008).

[163] S.-L. Ding, "Comparative anatomy of the prosubiculum, subiculum, presubiculum, postsubiculum, and parasubiculum in human, monkey, and rodent," Journal of Comparative Neurology **521**, 4145–4162 (2013).

[164] E. L. Stevenson and H. K. Caldwell, "Lesions to the ca 2 region of the hippocampus impair social memory in mice," European Journal of Neuroscience **40**, 3294–3301 (2014).

[165] K. Okamoto and Y. Ikegaya, "Recurrent connections between CA2 pyramidal cells," Hippocampus **29**, 305–312 (2019).

[166] U. Häussler, K. Rinas, A. Kilias, U. Egert, and C. A. Haas, "Mossy fiber sprouting and pyramidal cell dispersion in the hippocampal CA2 region in a mouse model of temporal lobe epilepsy," Hippocampus **26**, 577–588 (2016).

[167] P Boersma and D Weenink, *Praat: doing phonetics by computer [computer program]. version 6.1. 26 (2020).*

[168] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: a python interface to praat," Journal of Phonetics **71**, 1–15 (2018).

The neuronal silhouettes on the cover are
modified from *Basket and Pyramidals print*
of *Greg Dunn Neuro Art.*

ma per seguir virtute e canoscenza

SISSA