Low-dimensional continuous attractors in recurrent neural networks : from statistical physics to computational neuroscience

Aldo Battista

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Low-Dimensional Continuous Attractors in Recurrent Neural Networks

from Statistical Physics to Computational Neuroscience

Soutenue par
Aldo BATTISTA
Le 27 Octobre 2020

École doctorale n° 564
Physique en Île-de-France

Spécialité
Physique

Composition du jury :

Jean-Pierre, NADAL École Normale Supérieure	*Président du jury*
Sandro, ROMANI Janelia Research Campus	*Rapporteur*
K. Y. Michael, WONG HKUST	*Rapporteur*
Elena, AGLIARI La Sapienza	*Examinateur*
Pierfrancesco, URBANI IPhT/CEA	*Examinateur*
Rémi, MONASSON École Normale Supérieure	*Directeur de thèse*

LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

ACKNOWLEDGMENTS

After a very long journey the time for thanks has come, which makes me a little sad because it is a concrete sign that this beautiful adventure has come to an end.

In fact, I spent three wonderful years in the Physics Laboratory of the ENS (and in general in Paris) that have deeply changed me (hopefully for the better) both from a scientific and personal point of view, and all this is mainly due to the colleagues and friends with whom I shared this experience.

So, I take this opportunity to thank them, even if those who know me are aware that I am not good at expressing my feelings in words. I hope to be able to communicate even a small part of the deep sense of gratitude they deserve.

I can only begin by deeply thanking my supervisor Rémi Monasson who opened the doors to this trip. I am truly grateful to him first of all for making me passionate about neuroscience and for showing me how statistical physics can be a really powerful weapon in the study of problems at first sight far from physics. I thank him for the continuous technical and personal support that I have received during these years and that I wish every Ph.D. student to have. It is obvious to say that without him this work would be nothing but a distant mirage and for this I will always be extremely grateful!

A sincere thanks also goes to Simona Cocco and to all the fantastic colleagues with whom I had the pleasure to work in the same research group during these years. All of them were always available to discuss about science and this helped me a lot, but I also wanted to express my gratitude to them for their friendship. All this has made the working group actually become a second family and the laboratory a second home and for this I am extremely thankful to everyone. I would also like to take this occasion to thank Alessandro Treves and his students for the good times spent in Trieste, which were very stimulating from a scientific and human point of view. Moreover, a special thanks goes to my office mates, especially for tolerating me during these three long years, but also for the countless scientific discussions that have been crucial for me.

Thanks also to my mentor and tutor, respectively Francesco Zamponi and Jean-Pierre Nadal, for having followed the evolution of this work and for their advice during this path. Thanks a lot even to the members of the committee for accepting with great enthusiasm to correct my thesis and for the valuable suggestions that have certainly improved the quality of this work.

Furthermore, I would like to take this opportunity to warmly thank all my friends, from the Parisians to the Romans and the ones of my homeland (Ciociaria) without whom I would certainly not be the person I am but above all for having supported me all these years.

Finally, it is my duty and pleasure to thank my whole family for always being supportive in my choices and for the affection I receive from them constantly, especially thanks to my twin brother Pietro with whom I have shared most of my life (and will continue to do so).

ABSTRACT

How sensory information is encoded and processed by neuronal circuits is a central question in computational neuroscience. In many brain areas, the activity of neurons is found to depend strongly on some continuous sensory correlate; examples include simple cells in the V1 area of the visual cortex coding for the orientation of a bar presented to the retina, and head direction cells in the subiculum or place cells in the hippocampus, whose activities depend, respectively, on the orientation of the head and the position of an animal in the physical space.

Over the past decades, continuous attractor neural networks were introduced as an abstract model for the representation of a few continuous variables in a large population of noisy neurons. Through an appropriate set of pairwise interactions between the neurons, the dynamics of the neural network is constrained to span a low-dimensional manifold in the high-dimensional space of activity configurations, and thus codes for a few continuous coordinates on the manifold, corresponding to spatial or sensory information.

While the original model was based on how to build a single continuous manifold in an high-dimensional space, it was soon realized that the same neural network should code for many distinct attractors, *i.e.*, corresponding to different spatial environments or contextual situations. An approximate solution to this harder problem was proposed twenty years ago, and relied on an ad hoc prescription for the pairwise interactions between neurons, summing up the different contributions corresponding to each single attractor taken independently of the others. This solution, however, suffers from two major issues: the interference between maps strongly limit the storage capacity, and the spatial resolution within a map is not controlled.

In the present manuscript, we address these two issues. We show how to achieve optimal storage of continuous attractors and study the optimal trade-off between capacity and spatial resolution, that is, how the requirement of higher spatial resolution affects the maximal number of attractors that can be stored, proving that recurrent neural networks are very efficient memory devices capable of storing many continuous attractors at high resolution.

In order to tackle these problems we used a combination of techniques from statistical physics of disordered systems and random matrix theory. On the one hand we extended Gardner's theory of learning to the case of patterns with strong spatial correlations. On the other hand we introduced and studied the spectral properties of a new ensemble of

random matrices, *i.e.*, the additive superimposition of an extensive number of independent Euclidean random matrices in the high-density regime.

In addition, this approach defines a concrete framework to address many questions, in close connection with ongoing experiments, related in particular to the discussion of the random remapping hypothesis and to the coding of spatial information and the development of brain circuits in young animals.

Finally, we discuss a possible mechanism for the learning of continuous attractors from real images.

RÉSUMÉ

La manière dont l'information sensorielle est codée et traitée par les circuits neuronaux est une question centrale en neurosciences computationnelles. Dans de nombreuses régions du cerveau, on constate que l'activité des neurones dépend fortement de certains corrélats sensoriels continus ; on peut citer comme exemples les cellules simples de la zone V1 du cortex visuel codant pour l'orientation d'une barre présentée à la rétine, et les cellules de direction de la tête dans le subiculum ou les cellules de lieu dans l'hippocampe, dont les activités dépendent, respectivement, de l'orientation de la tête et de la position d'un animal dans l'espace physique.

Au cours des dernières décennies, les réseaux neuronaux à attracteur continu ont été introduits comme un modèle abstrait pour la représentation de quelques variables continues dans une grande population de neurones bruités. Grâce à un ensemble approprié d'interactions par paires entre les neurones, la dynamique du réseau neuronal est contrainte de s'étendre sur une variété de faible dimension dans l'espace de haute dimension des configurations d'activités, et code ainsi quelques coordonnées continues sur la variété, correspondant à des informations spatiales ou sensorielles.

Alors que le modèle original était basé sur la construction d'une variété continue unique dans un espace à haute dimension, on s'est vite rendu compte que le même réseau neuronal pouvait coder pour de nombreux attracteurs distincts, correspondant à différents environnements spatiaux ou situations contextuelles. Une solution approximative à ce problème plus difficile a été proposée il y a vingt ans, et reposait sur une prescription ad hoc pour les interactions par paires entre les neurones, résumant les différentes contributions correspondant à chaque attracteur pris indépendamment des autres. Cette solution souffre cependant de deux problèmes majeurs : l'interférence entre les cartes limitent fortement la capacité de stockage, et la résolution spatiale au sein d'une carte n'est pas contrôlée.

Dans le présent manuscrit, nous abordons ces deux questions. Nous montrons comment parvenir à un stockage optimal des attracteurs continus et étudions le compromis optimal entre capacité et résolution spatiale, c'est-à-dire comment l'exigence d'une résolution spatiale plus élevée affecte le nombre maximal d'attracteurs pouvant être stockés, prouvant que les réseaux neuronaux récurrents sont des dispositifs de mémoire très efficaces capables de stocker de nombreux attracteurs continus à haute résolution.

Afin de résoudre ces problèmes, nous avons utilisé une combinaison de techniques issues de la physique statistique des systèmes désordonnés et de la théorie des matrices

aléatoires. D'une part, nous avons étendu la théorie de l'apprentissage de Gardner au cas des modèles présentant de fortes corrélations spatiales. D'autre part, nous avons introduit et étudié les propriétés spectrales d'un nouvel ensemble de matrices aléatoires, c'est-à-dire la superposition additive d'un grand nombre de matrices aléatoires euclidiennes indépendantes dans le régime de haute densité.

En outre, cette approche définit un cadre concret pour répondre à de nombreuses questions, en lien étroit avec les expériences en cours, liées notamment à la discussion de l'hypothèse du remapping aléatoire et au codage de l'information spatiale et au développement des circuits cérébraux chez les jeunes animaux.

Enfin, nous discutons d'un mécanisme possible pour l'apprentissage des attracteurs continus à partir d'images réelles.

PUBLICATIONS

This manuscript comprises the research work that the author (Aldo Battista) has conducted during the last three years under the supervision of Prof. Rémi Monasson at the Laboratoire de Physique de l'École Normale Supérieure, and includes the following published as well as original results.

Journal articles

— [28] Aldo Battista and Rémi Monasson. « Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks.» In: *Physical Review Letters* 124.4(2020), p.048302.

— [29] Aldo Battista and Rémi Monasson. « Spectrum of multispace Euclidean random matrices. » In: *Physical Review E* 101.5(2020), p.052133.

OVERVIEW OF THE CHAPTERS

This thesis work is organized as follows.

— First of all, in Chapter 1 we discuss the close link between statistical physics and computational neuroscience, thus justifying our approach.

— Afterwards, we start providing motivations for our work by discussing in detail the low-dimensional manifold hypothesis, the cornerstone of the whole thesis. That's the content of Chapter 2.

— Once we have clarified the approach and motivations of the thesis, we are going to introduce its subject in Chapter 3, *i.e.*, recurrent neural networks and attractors. This pedagogical introduction to the subject is always accompanied by links with statistical physics and experimental evidences in the field of neuroscience that justify the development of the models themselves.

— In Chapter 4 we go into the heart of the thesis, where we propose and study in detail a new theory with the aim of solving the problems on continuous attractor neural networks to store multiple manifolds presented in the previous Chapter 3. In particular this has given us the opportunity to generalize Gardner's classic theory for the capacity of the perceptron to the case of patterns with strong spatial correlation.

— The theory presented in Chapter 4 led to the introduction of a new ensemble of random matrices, *i.e.*, the superimposition of an extensive number of independent random Euclidean matrices in the high-density limit. Chapter 5 is devoted to a detailed study of the spectral properties of this new class of matrices.

— In Chapter 6 we report additional details on the theory studied in Chapter 4. We also propose a generalization of the results found so far, in particular by inserting biological constraints into the model. Links with ongoing experiments are made as well as another application of our setting is shown in the context of storing continuous attractors in a recurrent neural network starting from real images.

— Finally, Chapter 7 contains a summary of the results obtained in this work as well as indications for lines of research to follow in the near future.

CONTENTS

1

STATISTICAL PHYSICS MEETS COMPUTATIONAL NEUROSCIENCE

> Since this is a thesis in theoretical (statistical) physics with applications to computational neuroscience, it is worth reminding the reader of the strong link, both from an historical and methodological point of view, between these two apparently very distant fields of research. This is the purpose of this short Chapter.

Statistical physics was born from thermodynamics towards the end of the nineteenth century and later developed in the field of condensed matter physics [106, 129, 192].

The main objective of this research area is to deduce easily measurable macroscopic quantities as a result of microscopic laws and also to explain collective phenomena such as phase transitions [113, 138, 189]. Its spirit can be summed up in the concise and famous P.A. Anderson phrase of 1972: "more is different", or in other words, when we move from an individual level of description of nature, *i.e.*, a single molecule of water, to a collective level, *i.e.*, a bottle full of water, new and non-trivial phenomena appear, *i.e.*, liquid-solid transition.

In the case of thermodynamics and condensed matter physics, there are basically two reasons for the development of statistical physics:

— what is observed in experiments is usually on a macroscopic scale, so one is generally interested in macroscopic quantities;

— it is often not possible to calculate microscopic quantities because of the huge number of parameters.

A statistical approach therefore manages to reconcile well what is needed with what is feasible.

Systems with multiple levels of description are omnipresent in nature, which is why the tools of statistical physics in recent decades have been used in many research fields, even seemingly far from physics, *i.e.*, sociology [52], economics [42, 148], finance [40], biology [33], immunology [6, 10, 34], route planning [263] and neuroscience [14, 145].

In particular, this thesis concerns applications of statistical physics to the field of neuroscience, *i.e.*, the study of the brain.

It is now evident the union between these two fields that seemed at first sight very different: the brain is formed by microscopic units, the neurons, which are interconnected in a network, via synapses, and each of them follows its own local microscopic dynamics and from this we want to deduce the global states.

This problem is very reminiscent of magnetic systems in physics in which the spins (magnetic moments) interact with each other so that long-range ferromagnetic order can arise from the local couplings between them: the spins are schematized as binary variables, *i.e.*, they can be in both up and down state, as well as a simplified model of a neuron can be active or silent. There is therefore a direct mapping between the celebrated Ising model [117], which is the milestone for magnetic system models, and the schematic representations of neural networks [65].

To complete the analogy just think that real neural networks have randomness factors both in the network structure ("quenched noise") and in the response of each neuron ("fast noise", equivalent to a temperature parameter). Therefore, all the mathematical tools developed by statistical physicists can be used to investigate neuroscience problems.

It is therefore no coincidence that many works in the literature have tried to address computational neuroscience problems with statistical physics techniques such as models of neural networks that reproduce observations of brain activity or that are able to perform specific functions [14, 67, 89].

Today, applications of techniques from the 1980s of statistical physics of disordered systems, *i.e.*, spin glasses, to real and artificial neural networks [18, 128] are a very hot topic for two main reasons:

— the first is to understand the success of machine learning algorithms, particularly in the field of deep learning [97]. In fact, in recent years these algorithms have brought to the state of the art performance in several fields such as image [107, 135, 218] and speech recognition [99, 110], natural language processing [61, 159], text translation [22, 137, 226], computational medical diagnosis [72] and artificial image/video generation [98, 199];

— the second comes from the incredible improvement of experimental techniques in neurobiology, such as electrophysiological and fluorescence-based functional recordings of neurons, that allow now to study in a quantitative way what in the 1980s was only speculative [222].

2 THE LOW-DIMENSIONAL MANIFOLD HYPOTHESIS

Before starting this long journey it is essential to spend a few pages for discussing the low-dimensional manifold hypothesis, which encapsulates the key motivations that inspired this thesis. Although the reasons of this work are coming from computational neuroscience, we discuss here the emergence of this hypothesis also in the context of machine learning in order to emphasize its generality and importance.

The low-dimensional manifold hypothesis states that real-world high-dimensional data may lie on low-dimensional manifolds embedded within the high-dimensional space [75]. This definition may seem complicated at first sight but in reality its meaning is very intuitive. In the following we introduce this hypothesis in the field of machine learning and computational neuroscience respectively so as to concretize this abstract statement with concrete examples.

2.1 IN MACHINE LEARNING

Representing and interpreting efficiently noisy high-dimensional data is an issue of growing importance in modern machine learning. A common procedure consists of searching for representations of data in spaces of (much) lower dimensions, an approach known as manifold learning [32, 53, 118, 146].

Manifold learning's approach has as a key assumption the low-dimensional manifold hypothesis which states that although many data are a priori high-dimensional, in reality they are intrinsically of much lower dimension. There are several reasons in favour of this hypothesis, for example:

— this can result from physical laws such as translation, rotation, change of scale and so on. If we consider a real image, whose dimensionality (the number of pixels constituting the photo) is usually very high and we apply the above mentioned transformations, it is clear that all the variants of the starting photo are linked together by a number of parameters much smaller than the number of pixels, *i.e.*, each parameter corresponding to a particular physical law. So in the end we have that the different photos live in a space of dimensionality much lower than the a priori very high number of pixels, see Fig. 1(a);

— moreover, if we consider a classic data-set used in machine learning like MNIST [140], see Fig. 1(b), it is possible to estimate the intrinsic dimension, that is the

number of variables needed in a minimal representation of the data, of its digits by looking at the number of local transformations required to convert a number into another one of its variants, and it turns out that this number is $\simeq 10$ [63], much less than the number of pixels of the images that make up MNIST, equal to 784.

The same goes for a more complicated example, let's consider a data-set consisting of photos of a person's face in different poses, Fig. 1(c); each picture is made of, say, 1000×1000 pixels. It is clear that this data-set is a very small subset of all possible colored pictures, which is defined by a $3 \cdot 10^6$–dimensional vector [1]. The reason is that, for a given face, there are only ~ 50 varying degrees of freedom (the position of all muscles), a very small number compared to 10^6 [139]. Hence, all data points lie in a (non-linear) manifold, of very low dimension compared to the one of the initial pixel space.

We can therefore generally conclude that even though data may be high-dimensional, very often the number of relevant dimensions is much smaller.

The low-dimensional manifold hypothesis explains (heuristically) why machine learning techniques are able to find useful features and produce accurate predictions from data-sets that have a potentially large number of dimensions (variables) bypassing the curse of dimensionality problem. In fact, when the dimension of data increases, the volume of the configuration space grows so fast (exponentially) that the available data become sparse and this sparsity is problematic for any method that requires statistical significance. The fact that the actual data-set of interest really lives in a space of low dimension, means that a given machine learning model only needs to learn to focus on a few key features of the data-set to make decisions. However, these key features may turn out to be complicated functions of the original variables. Many of the algorithms behind machine learning techniques focus on ways to determine these (embedding) functions [194].

A simple example of application of this hypothesis comes from the context of supervised learning [35, 96, 97], that is the fitting of input-output relation from examples, with neural networks for high-dimensional data classification because if the data actually live in manifolds of much smaller dimension, it is necessary to classify the manifolds [58], see Fig. 1(a).

1. The factor 3 comes from the RGB color channels.

Figure 1 – (a) The set of variations associated with the image of a dog form a continuous manifold of dimension much lower than the pixels space. Other object images, such as those corresponding to a cat in various poses, are represented by other manifolds in same space. Figure adapted from [58]; (b) Examples of digits corresponding to the number two in the MNIST data-set. Figure taken from [152]. (c) Pictures of a person with various facial expressions. They lie in a very low dimensional manifold of the vector space of pictures with 1000×1000 pixels. Figure taken from [139].

2.2 IN COMPUTATIONAL NEUROSCIENCE

Low-dimensional representations of high-dimensional data are not restricted to machine learning, and are encountered in other fields, in particular, computational neuroscience [1, 86, 250].

In fact, what is typically done in a neuroscience experiment is to measure with electrodes the activity of neurons in a real neural network, see Fig. 2(a). By looking at the population of measured neurons one can often find that the activity of the network can be explained in terms of relative activation of groups of neurons, called neural modes or cell assemblies, see Fig. 2(b). This means that even if the network activity is a priori high-dimensional, if one looks at its trajectory in the space of neural configurations as a function of time, it will be confined to live in a linear (or even non linear) manifold of much smaller dimension $D \ll N$, being N the number of neurons that make up the network, see Fig. 2(c) and Fig. 2(d).

It is now legitimate to ask what the dimensions of the manifold mean. One of the most reliable hypothesis is that these collective coordinates generated by the neural network activity in the D-dimensional manifold encode sensory correlates, *i.e.*, they encode some external stimulus, as the orientation of a bar presented to the retina [114, 115], and can be used for example from the motor cortex to make decisions and/or produce actions [86].

Also related to this is the fact that low-dimensional continuous attractors provide a paradigm for analog memories, in which the memory item is represented by an extended manifold, *i.e.*, the cognitive map of a place in a certain context, see Section 3.3.

A crucial question to answer is therefore, as we will investigate in detail in the next Chapters, how to engineer the connections in a network of interconnected neurons such as to map a low-dimensional dynamics to the high-dimensional one of the units that make up the net [23].

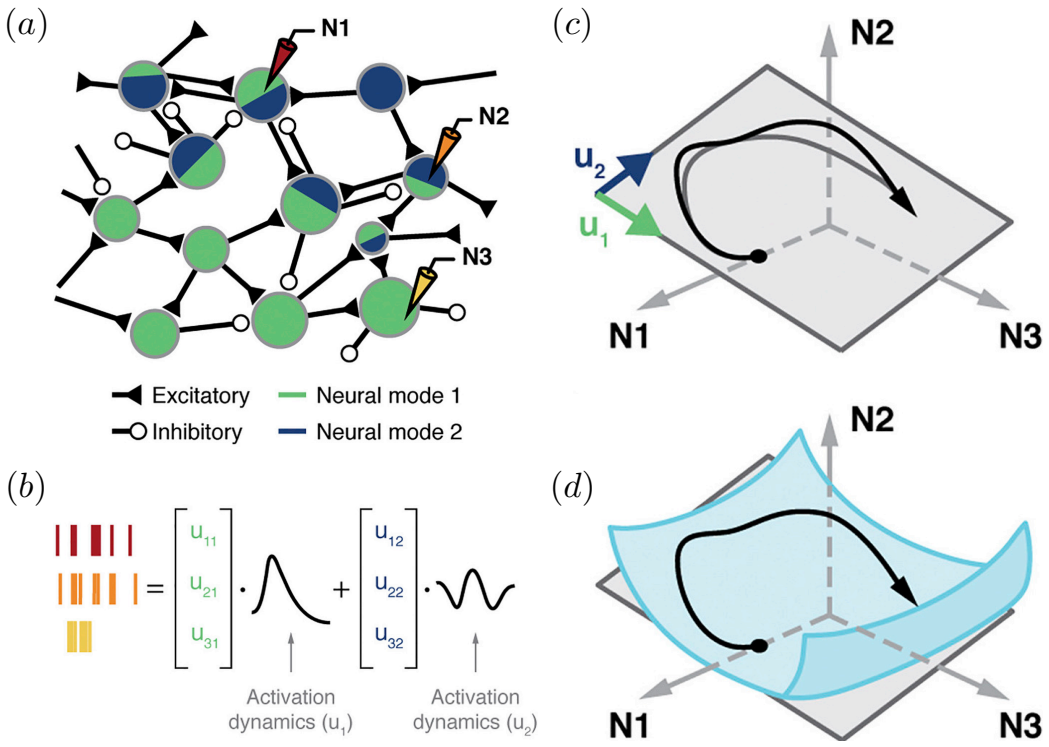

Figure 2 – (a) Neural modes as a generative model for population activity. The relative area of the blue/green regions in each neuron represents the relative magnitude of the contribution of each cell assembly to the neuron's activity; (b) Spikes from three recorded neurons during task execution as a linear combination of two neural modes; (c) Trajectory of time-varying population activity in the neural space of the three recorded neurons (black). The trajectory is mostly confined to the neural manifold, a plane shown in gray and spanned by the neural modes u_1 and u_2; (d) A curved, nonlinear neural manifold, shown in blue. Figure adapted from [86].

3

RECURRENT NEURAL NEURAL NETWORKS AND ATTRACTORS

Once we have presented in Chapter 2 the fundamental hypothesis on which this work is based, we can introduce the subject of this thesis, *i.e.*, recurrent neural networks and attractors, through a pedagogical illustration of a series of models introduced in the field of computational neuroscience. These models are always presented together with the experimental evidence that led to their formulation and the connections with statistical physics are also explained in detail, see Chapter 1 for a discussion on the link between statistical physics and computational neuroscience. The aim of this Chapter is therefore to place the work of this thesis in a very precise context within the literature, stressing the problems of the current theory and therefore the need for it.

3.1 DISCRETE AND CONTINUOUS ATTRACTORS IN RECURRENT NEURAL NETWORKS

Let's start by defining in a pictorial way what a recurrent neural network (RNN) is, what an attractor is, and the difference between discrete and continuous attractors [1].

A RNN is a kind of non-linear dynamical system defined by a set of N activity variables (neurons) σ_i, $i = 1, \dots, N$, interconnected via pairwise synapses $\{W_{ij}\}$ (in the following we will never consider the case of self-connections, *i.e.*, $W_{ii} = 0, \forall\, i$), where, depending on the models, both neurons and synapses can assume binary or continuous values and also respect from time to time different constraints of biological nature that we will discuss later, see Fig. 3(a).

In addition, the activity variables are updated over time, which can also be considered discrete or continuous as the circumstances require, following a non-linear dynamics dictated by the connectivity matrix $\mathbf{W}$ (or even by external fields), whose choice obviously defines the network properties in a crucial way.

The state of the RNN can therefore be represented by a point evolving in a very high-dimensional space, the space of neural configurations of dimension N. In particular we will be interested in studying the trajectory of this point after a long time, especially in the case where the dynamics of the network remains blocked on different fixed points

1. As it will become clearer during the thesis, in this context we will always use the terms manifold, map, environment and continuous attractor with the same meaning.

depending on the initial condition of the neurons activity variables: these fixed points are the celebrated attractors.

According to the choice of the synapses (without considering the presence of any external fields) we can have different scenarios for the structure of these fixed points:

— we can have that the different fixed points (specific configurations of the network activity variables where the dynamics get stucked) are isolated from each other and divided by attraction basins, that define according to the initial condition which will be the fixed point to which the dynamics of the network will converge: this is the case of discrete attractors, also called point attractors or 0-dimensional attractors, see Fig. 3(b);

— moreover, we can also have situations where the attractors instead of being isolated points, are composed of a continuous set of fixed points (manifolds) living in a D-dimensional space, where typically $D \ll N$: this is the case of continuous attractors, see Chapter 2. Also here it is possible to have several continuous attractors as fixed points of the same network and divided by attraction basins, where, however, now depending on the initial condition of the network in an attraction basin, the dynamics of the RNN can remain blocked at any point of the relative attractor, see Fig. 3(c). As we discussed in Section 2.2, the important thing is to understand the physical meaning of the collective coordinate $\mathbf{r}$ which represents the state of the network onto the continuous attractor, that is a D-dimensional vector.

It is important to note that the dynamics really gets stucked [2] to a fixed point if it is noise free (deterministic), *i.e.*, zero temperature Glauber dynamics [91], otherwise there will be fluctuations around the fixed points that will depend on the level of the neural noise. In particular, as we will see later, with the right temperature it is possible to spontaneously generate for the state of the network transitions between one fixed point and another in the case of discrete attractors and the same is true also in the case of continuous attractors where, however, in addition to transitions between different attractors, a diffusive dynamics of the collective coordinate $\mathbf{r}$ on the attractors themselves is present as well.

Questions that we will answer in the following concern how to engineer the choice of synapses in order to build attractors with ad hoc (biological) properties and in particular what is the maximum number of attractors that can be stored in a recurrent network.

2. Note that in general a dynamical system can have other types of attractors such as periodic or chaotic attractors [248] that we will not consider in this thesis, in fact we will always be interested, both in the case of discrete and continuous attractors, in fixed points where the dynamics of the system remains stuck in the absence of small noise or external fields (stable or indifferent equilibrium points).

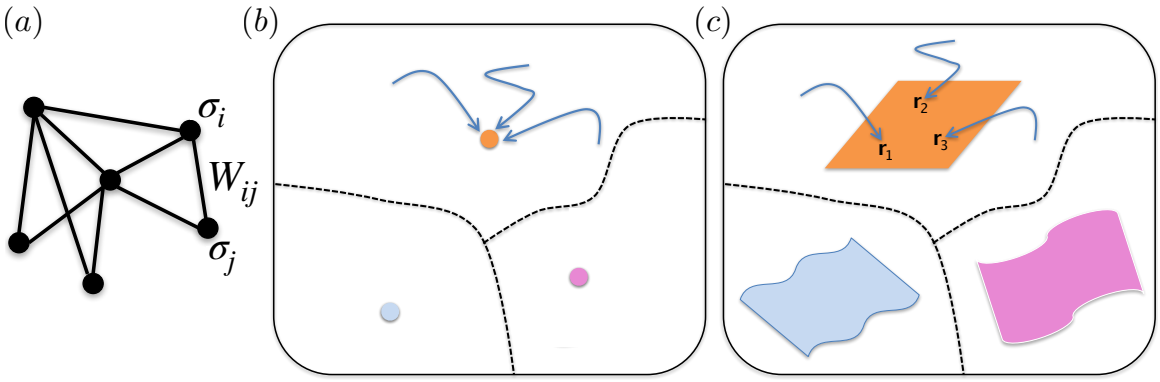

Figure 3 – (a) Sketch of a recurrent neural network (RNN) with $N = 6$ neurons and connectivity matrix $\mathbf{W}$; (b) Schematic diagram of the space of configurations of a RNN (N-dimensional space) in which the connections have been chosen to store multiple discrete attractors. The dashed lines indicate the attraction basins between the different fixed points while the arrows are examples of possible trajectories for a deterministic dynamics (converging to the nearest fixed point depending on the initial condition) of the network, for different initial configurations; (c) Same as (b) but in the case of multiple continuous attractors. $\mathbf{r}$ is a D-dimensional collective coordinate that describe the "position" along one of the manifolds. Figure adapted from [28].

We will also give below a strong emphasis to the experimental evidence (direct and indirect) of these mechanisms in the brain, especially in the context of memory, where memories correspond to the above mentioned attractors.

3.2 HOPFIELD MODEL: MULTIPLE POINT-ATTRACTORS

Undoubtedly the milestone in this field of research is the seminal work of J.J. Hopfield in 1982 [112] where the model named after him was formulated.

He showed that the computational properties used by biological organisms or for the construction of computers may emerge as collective properties of systems that have a large number of simple equivalent components (or neurons).

In practice J.J. Hopfield had proposed a practical way to choose connections in a RNN with many neurons in order to build multiple 0-dimensional attractors, showing that a very simple model of interacting binary neurons was able to have non-trivial collective properties, in particular to build autoassociative memories.

Basically there are two main ways to store information on a device: addressable and autoassociative memory.

— The first way consist in comparing input search data (tag) against a table of stored data, and returns the matching one [188];

— the second is any type of memory that enables one to retrieve a piece of data from only a tiny sample of itself.

The Hopfield model together with all the models we will see in the following are autoassociative memories, and are particularly important to study because they are more biologically plausible than addressable ones.

Moreover, this model has remarkable properties as the robustness to the removal of a certain number of connections, the ability to correct patterns (memories) presented with errors, the ability to store patterns with a time sequence and recall them in the right order, although the single elementary components had independent dynamics without a clock that synchronized them [112].

Before we discuss the Hopfield model in detail, let us recall the fundamental ingredients of biological inspiration that led to its formulation.

3.2.1 *Ingredients of the model*

The model is an extremely simplified schematization of a real neural network, where the basic units are binary spins[3] inspired by the Ising model of ferromagnetism from statistical physics [117], see Chapter 1. These units or better neurons are therefore binary variables, *i.e.*, $\sigma_i = \pm 1$[4], where the state -1 represents a silent neuron while the state $+1$ an active neuron.

In reality, neurons are more complicated because they are cells whose activity is given by the difference in potential between the inner and outer part of their membrane. Typically this potential difference is about -70 mV when the neuron is silent, and when the neuron becomes active a very sharp electrical wave (action potential) localized in time is emitted and propagated along the axon which is the output of the neurons. The duration of this wave is of the order of a millisecond and if we properly discretize the time it is possible to distinguish if the neuron is active or not depending on the presence of this wave, see Fig. 4(a) for a schematic representation of two interconnected neurons and Fig. 4(b) for a schematic plot of an action potential.

3. In the following we will use the terms spins, neurons, activity variables and units indiscriminately.
4. All the following can be trivially formulated with the notation $\sigma_i = 0, 1$ which we will use afterwards. Here we consider the notation $\sigma_i = \pm 1$ because the Hopfield model was historically introduced in this way.

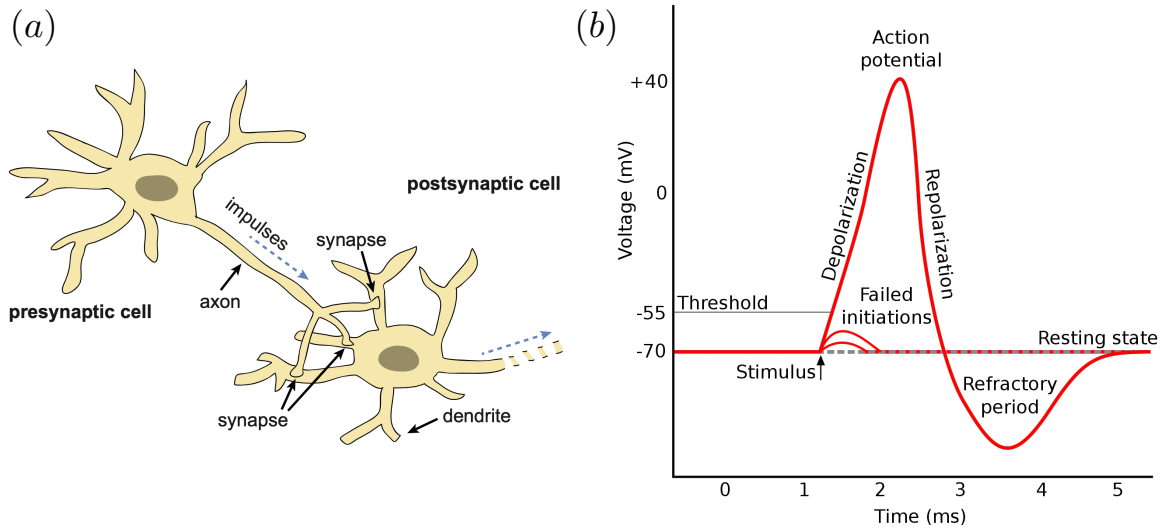

Figure 4 – (a) Schematic representation of two interconnected neurons. The contact areas where the information is transmitted are called synapses. A signal from the presynaptic cell is transmitted through the synapses to the postsynaptic cell. Figure taken from [230]; (b) Approximate plot of a typical action potential shows its various phases as the action potential passes a point on a cell membrane. The membrane potential starts out at approximately -70 mV at time zero. A stimulus is applied at time $\simeq$ 1 ms, which raises the membrane potential above -55 mV (the threshold potential). After the stimulus is applied, the membrane potential rapidly rises to a peak potential of 40 mV at time $\simeq$ 2 ms. Just as quickly, the potential then drops and overshoots to -90 mV at time $\simeq$ 3 ms, and finally the resting potential of -70 mV is reestablished at time $\simeq$ 5 ms. Figure taken from [253].

The different neurons then communicate with each other via these electrical signals that pass through the connections between the neurons called synapses [5]. The important thing about synapses is that when electric waves arrive at the end of the axons and come into contact with them, they may or may not amplify this signal depending on the strength and type of synaptic interaction. There are basically two types of synapses, excitatory (positive) and inhibitory (negative), the former tend to amplify the signal, while the latter tend to weaken it.

Moreover, a neuron is connected to many other neurons and therefore receives from all of them the different electrical signals coming from axons and then modulated by synapses at a fixed time. What happens in the neuron is therefore a weighted sum of these signals which is then compared to a threshold, if this threshold is exceeded the neuron will emit an action potential, so it will be active, otherwise it will remain in its

5. From now on we will also use the terms synapses, weights, connections and interactions indiscriminately.

resting state, so it will be silent. This dynamics can be represented by the following simple equation:

$$\sigma_i(t+1) = \text{sign}\left(\sum_{j\neq i} W_{ij}\sigma_j(t) - \theta_i\right), \tag{3.1}$$

where $\sigma_i(t+1)$ represents whether neuron i at the discrete time step $t+1$ is active or silent, W_{ij} represents the synapse between neuron i and neuron j which can be both excitatory and inhibitory, θ_i represents the threshold associated with neuron i and sign is the sign function defined as:

$$\text{sign}(x) := \begin{cases} -1, & \text{if } x < 0 \\ +1, & \text{otherwise} \end{cases} . \tag{3.2}$$

Once we have defined the dynamics of the model in Eq. (3.1), starting from an initial configuration for the activity of the neurons and a choice for the synaptic coefficients everything is well defined. The next step is then how to determine the choice of synapses. To do this J.J. Hopfield took inspiration from D. Hebb's seminal work "The organization of behavior" of 1949, in particular the idea that "neurons that fire together wire together" [108] or rather, quoting him directly:

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

In practice the idea is that if neurons have a tendency to be synchronously active together, then the connection between them will be strengthened over time. If, on the other hand, the opposite happens, *i.e.*, that there is no synchronicity in the activity of some neurons, the connections between them will be weakened.

It is interesting to note that this mechanism is different from what we typically have in physics where the connections in an interacting system are usually given in the Hamiltonian in a static way, while here they are dynamically modified during dynamics through this mechanism of feedback of neuron activity on synapses. In this case it is as if the Hamiltonian of the system self-modifies itself according to the dynamics it has produced in previous times.

It is also important to mention that since Hebb's original theoretical formulation, many experimental studies, both in vitro and in vivo, have investigated the physiological basis of synaptic potentiation. The long-lasting strengthening of the synaptic connection

between two neurons is called long-term potentiation, or LTP [36, 37, 170, 176]. In the hippocampus, a region of the brain that we will later discuss in Section 3.3, the best-known mechanism that enables LTP is the transduction of electrical signals into chemical ones that activate the potentiation mechanisms in both the pre-synaptic and post-synaptic neurons, mediated by the N-methyl-D-aspartate (NMDA) receptor complex [151]. Note that a similar mechanism occurs for the weakening of synapses, called long-term depression or LTD. It should be noted, however, that synaptic plasticity is still the subject of much research and several plausible mechanisms have been proposed, see as an example [160].

3.2.2 *Model details and properties*

Now that we have described all the ingredients properly we can define the Hopfield model in detail.

Let's suppose we want to store a certain set of neuron activity configurations (patterns), where by store we mean that these configurations must be the fixed points for the dynamics established in Eq. (3.1). These patterns are defined as $\xi_i^\mu = \pm 1$, where i is the neuron index, which goes from 1 to N as usual, and μ is the pattern index, which ranges from 1 to P. The individual elements of the patterns are sampled randomly and independently of each other. $\xi_i^\mu = +1$ corresponds to an active neuron, while $\xi_i^\mu = -1$ corresponds to a silent neuron.

Formalizing therefore in mathematical terms the idea of D. Hebb we can write the following prescription for the connectivity matrix of the RNN that must memorize these patterns obtaining so the famous Hebb rule [6]:

$$W_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu \,, \tag{3.3}$$

where this matrix is of rank P and where the hebbian mechanism is clear, in fact, if two neurons in a pattern are in the same state the connection between them is strengthened, otherwise weakened.

The question to ask now is that if with this ad hoc choice of the weights, Eq. (3.3), the patterns are actually fixed points of the dynamics (3.1) [7].

In the case $P = 1$ it is easy to understand why this works because this situation is equivalent to the Curie-Weiss model of statistical physics [113, 123, 189, 201, 202, 215, 239], which is basically a mean field version of the Ising model where we take all interactions

6. Note that thus the network presents both excitatory and inhibitory synapses.
7. For simplicity we consider the case without external fields, $\theta_i = 0$, $\forall i$.

equal to $W_{ij} = \frac{1}{N}$, $\forall\, i \neq j$. So we know that dynamics (3.1) will take the network to a state with non-zero magnetization, that is:

$$\frac{1}{N}\sum_i \sigma_i = +1 \quad \text{or} \quad \frac{1}{N}\sum_i \sigma_i = -1\,, \tag{3.4}$$

depending on the initial activity configuration. It's interesting to note that in this situation the network will be organized in the configuration with all 1s or all -1s after some time, *i.e.*, we have stored the pattern with all the components equal to $\xi_i^\mu = -1$ and the pattern with all the components equal to $\xi_i^\mu = 1$ (the network state after a long time has maximum overlap with one of the two memorized patterns).

Now that we have seen how to store trivial configurations we can go further and try to understand how to memorize more interesting patterns. It can be done by multiplying the couplings in the Curie-Weiss model by a global gauge: this is the Mattis model [113, 123, 189, 201, 202, 215, 239]. In particular, now we choose interactions like

$$W_{ij} = \frac{1}{N}\xi_i^1\xi_j^1\,, \tag{3.5}$$

where by a simple change of variables we have stored a richer pattern, like the ones we would like to store in the Hopfield model. In fact, also in this case the dynamics (3.1) will bring the network to a magnetized (with maximum overlap) state, for the same reason as in the Curie-Weiss model, but now on the pattern with components ξ_i^1 or the opposite one according to the initial condition of the network.

The idea behind the Hopfield model then is that if the patterns to memorize are orthogonal enough so that they don't interact too much[8], we could put together many Mattis models (one for each pattern to store) and memorize simultaneously in the same network multiple patterns, from this comes the rule in Eq. (3.3).

In addition, the dynamics in Eq. (3.1) is simply a zero temperature updating rule that is equivalent to minimize the following energy[9] [14, 109]:

$$E[\{\sigma_i\}] = -\frac{1}{2}\sum_{i<j} W_{ij}\sigma_i\sigma_j + \sum_i \theta_i\sigma_i\,. \tag{3.6}$$

We immediately recognize that this is the energy of a spin glass (frustrated[10] disordered magnetic systems) with local fields generated by the local thresholds $\{\theta_i\}$. So we know

8. For this reason the patterns in the Hopfield model are chosen randomly.
9. This is true if there are no self-connections and if the connectivity matrix is symmetric, as in this case.
10. In the Hopfield model the frustration comes from the Hebb rule (3.3), which leads both to excitatory and inhibitory couplings.

that this dynamics will converge after some time to the nearest local minimum depending on the initial condition, and the question is whether or not these minima are related to the patterns we want to store, see Fig. 5(a). The answer to this question was found first numerically by Hopfield himself [112] and then analytically by the seminal work of Amit, Gutfreund and Sompolinsky (AGS) [17] using techniques of statistical physics of disordered systems, *i.e.*, the replica method [51, 78, 157]. This answer is affirmative, in the limit of large N (thermodynamic limit), if the number of patterns to be mamorized divided by the number of neurons that make up the network is not too big, more precisely if $\alpha = \frac{P}{N}$, that is called load, is less than the critical capacity [11] $\alpha_c \simeq 0.138$ (without considering the presence of external fields, *i.e.*, $\theta_i = 0, \forall i$), see Fig. 5(b).

Moreover, AGS [17] generalized the dynamics of the system from deterministic, see Eq. (3.1), to stochastic by introducing a temperature T such that the system obeys the detailed balance according to the Hamiltonian (3.6).

In this setting, always in the absence of external fields, they were able to find a phase diagram in the thermodynamic limit that includes different phases. In particular at high temperatures we have a paramagnetic phase (PM) in which the spins are substantially random and uncorrelated, therefore an uninteresting phase. If instead we look at lower temperatures we see the presence of two phases, the ferromagnetic phase (FM) and the spin glass phase (SG). At not too high values of α we find ourselves in the FM phase where the minimums of the energy (3.6) actually correspond to the stored patterns. If instead we consider values of α too big we enter the SG phase in which the local minima of (3.6) have nothing to do with the stored patterns and then our network stops being an associative memory, this is due to the fact that when we want to store to many patterns in the connectivity matrix at a certain point the interference between them becomes so strong to generate this catastrophic loss of memories, see Fig. 5(c).

3.2.3 *Why is it important to study simple models?*

Before moving on to the next topic it is important to stress that the Hopfield model is extremely simplified and far from achieving a good level of biological realism [12], for example the neurons are schematized as simple spins, the connectivity matrix is symmetric and this is not true in general in real neural networks, the network is a priori fully-connected while generally biological networks are quite sparse, there is no constraint on the sign of weights while in neuroscience there is a clear difference between excitatory neurons that are connected by positive synapses and inhibitory neurons connected by

11. This is trivially true in the case of large N and finite P, so $\alpha = 0$ [109].
12. For more sophisticated and biologically plausible models of RNNs that store descrete attractors see for instance [48, 154, 169].

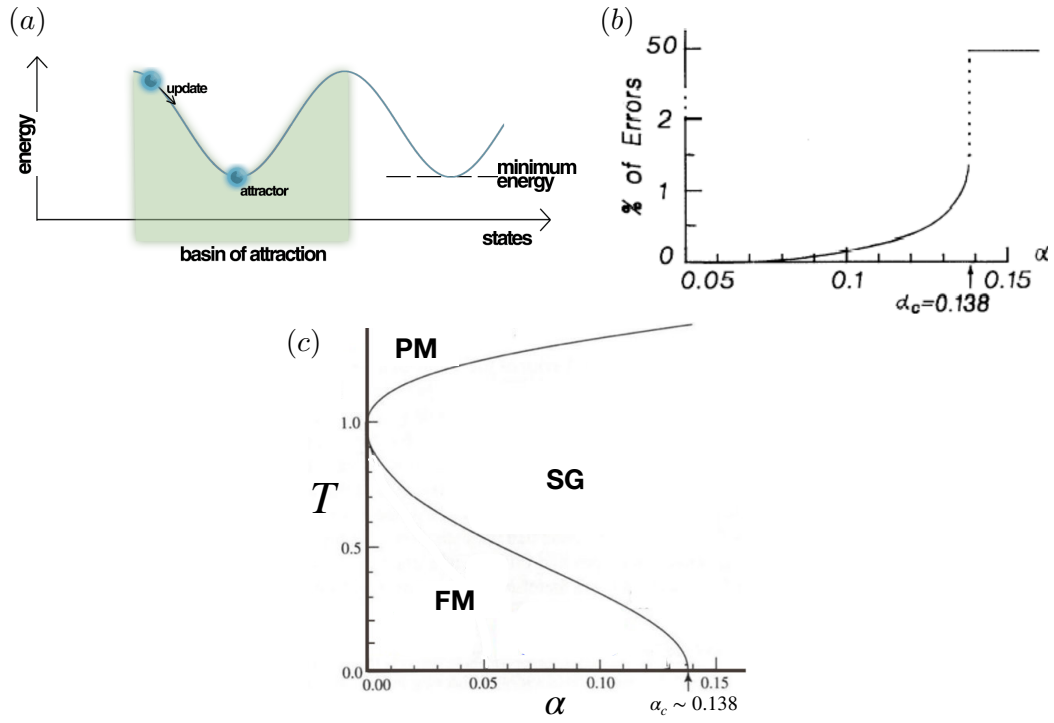

Figure 5 – (a) Energy landscape of a Hopfield network, highlighting the current state of the network (up the hill), an attractor state to which it will eventually converge, a minimum energy level and a basin of attraction shaded in green. Note how the update of the Hopfield network is always going down in energy. Figure taken from [254]; (b) Average percentage of errors in the Hopfield model as a function of $\alpha = \frac{P}{N}$ at zero temperature and with no external fields. Figure taken from [17]. (c) Phase diagram, temperature T vs load $\alpha = \frac{P}{N}$, of the Hopfield model in the absence of external fields. The continuous line represents the transition line between paramagnetic (PM), ferromagnetic (FM) and spin glass (SG) phases. Figure adapted from [17, 155].

negative synapses according to Dale's rule [66] and there is no adaptation mechanism in the dynamics of the network [14, 67, 89, 109].

Nevertheless, this model is recognized as a milestone in the field of theoretical neuroscience because it is one of the few paradigms that combine Hebb's rule with attractors. It is also interesting because with this model it is possible to engineer the connectivity matrix to allow different states of activity for the network. Moreover, after the work of AGS [17] the Hopfield model paved the way for the application of spin glass theory [157] beyond physics (together with simulated annealing [128] in computer science and engineering) starting to create a strong sociological impact in the statistical physics community towards theoretical biology and computer science.

Nonetheless, generalizations of the Hopfield model are nowadays a hot topic of research both in regards to neuroscience problems [73], but also machine learning thanks to the mapping between this model and (Restricted) Boltzmann Machines [3, 4, 238] or even immune networks applications [2, 5].

3.3 REPRESENTATION OF SPACE IN THE BRAIN

So far we have discussed in detail the Hopfield model (Section 3.2) which, as we have seen, is the milestone of recurrent neural network models to store multiple discrete attractors, on the other hand we have introduced motivations for another class of attractors found for example in computational neuroscience, *i.e.*, continuous attractors, see Section 2.2 and Fig. 3(c).

Before going on with the models on continuous attractor neural networks, however, it is necessary to give the reader more specific biological motivations. For this reason in the following we introduce the hippocampus, a very important part of the brain where the presence of the mechanisms discussed in Section 2.2 is hypothesized in particular for the representation of space in the brain [235, 236], where the activity of a recurrent network of noisy neurons (CA3 region of the hippocampus) that is high-dimensional, encodes for a continuous variable of low dimension as the position in an environment of an animal, and then we focus on the specific case of the special neurons that perform this task, *i.e.*, place cells [173].

For other examples in the brain of the presence of the continuous attractor mechanism see Section 3.8.

3.3.1 *Hippocampus*

The hippocampus is a part of the mammalian brain that is located in the medial temporal lobe and is part of the limbic system, see Fig. 6. It is composed of about 30 million neurons in humans and about 0.3 million neurons in rats. All vertebrate species, including reptiles and birds, have an homologous region. This part of the brain is one of the most studied by psychologists and neuroscientists because of its crucial role in spatial navigation and episodic memory.

The first evidence we have of the use of the hippocampus in a memory process was observed in the famous case of the patient H.M., who suffered from severe anterograde amnesia after receiving bilateral ablation of the hippocampus as a treatment for epilepsy [213]. This first evidence was followed by many others that confirmed the strong correlation between hippocampal damage/injury and problems with declarative, *i.e.*, involving conscious recall, memory formation and consolidation by human patients [11, 266].

In rodents, the hippocampus has been highly studied because of its role in spatial navigation and memory, *i.e.*, the process of memorizing and recalling the cognitive map associated with an environment in a certain context. The term "cognitive map", coined by the American psychologist Edward Tolman [232], refers to an allocentric representation of the surroundings embedded in a Euclidean metric, which enables navigation through the cognition of the spatial distances between locations and objects. Tolman himself gave the first strong experimental evidence for a map-based navigation system in rodents by showing the ability of the rats to devise shortcuts to a known prize position. [233]. In general, the role of the hippocampus in the formation and recall of cognitive maps (spatial memories) has been demonstrated many times in experiments that required specific positions to be memorized in an environment. An important example is the Morris water maze where a rat, who is an able swimmer but dislikes being in the water, is trained to swim to an hidden platform in a specific location within a pool of milky water. An healthy animal quickly learns the position of the platform in the sense that the average time to reach the goal decreases rapidly with the number of trials [171]. Morris and co-workers compared the results of healthy rats to those in which the hippocampus had previously been injured, showing that in the latter case the performance was drastically reduced [172].

Now without going into too much detail about the anatomy of the hippocampus, for which we refer to [20], it is important to say that a sub-part of it, called CA3, is practically a network of neurons with many recurrent connections, like those of the models discussed in Section 3.1, so that it is natural to try to formulate a mathematical model for this part of the brain [235, 236].

3.3.2 *Place cells and place fields*

The fundamental connection between the hippocampus, including the CA3 region, and spatial navigation is due to the crucial discovery in 1971 by O'Keefe (who won the Nobel Prize in physiology or medicine for this finding in 2014) and Dostrovsky of a population of pyramidal cells [13] (a type of excitatory cells) which are active only when an animal is in specific positions of an environment [186]. The sharp firing specificity of these neurons granted them the name of "place cells", and their spatial receptive fields were named "place fields", see Fig. 13(a).

If, for example, we consider a rat exploring a square arena there will be a neuron in the hippocampus that will be very active when the rat is at some specific position and practically silent outside, and the same goes for other neurons with other positions in the

13. The name of pyramidal cells comes from their shape.

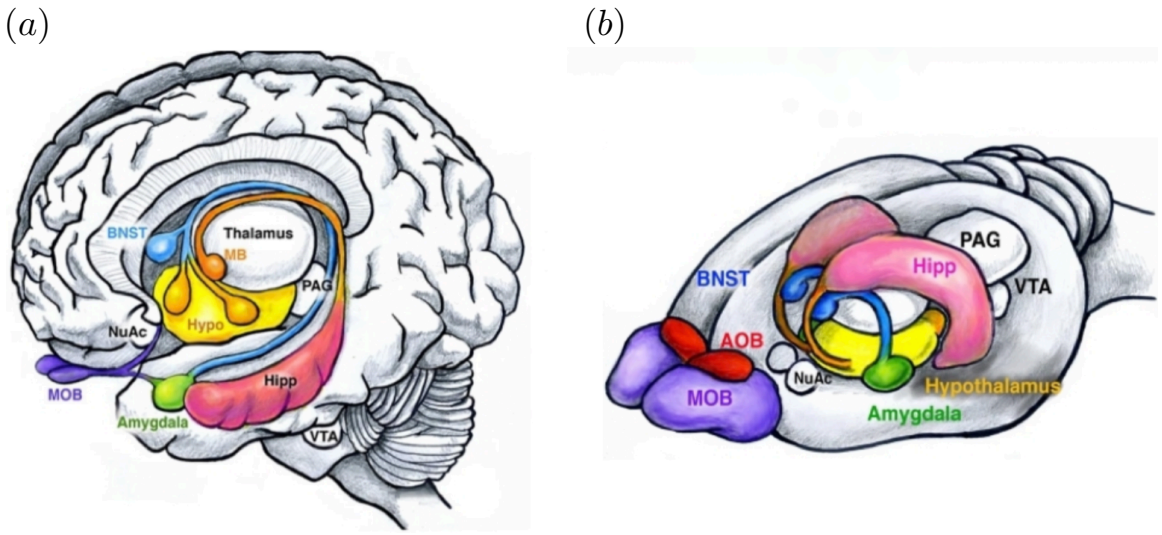

Figure 6 – Main structures of the human and rodent limbic system. (a) Human brain showing the amygdala (green), bed nucleus of stria terminalis (BNST, blue), hypothalamus (yellow), and hippocampus (pink). The hippocampus (pink) attaches to the mamillary bodies (orange) through the fimbria-fornix. Olfactory inputs are received by the olfactory bulbs (MOB, purple). Other structures include the nucleus accumbens (NuAc), ventral tegmental area (VTA), and the periaqueductal gray (PAG). (b) Similar structures are found in rodents. Figure and caption adapted from [221].

environment. The area of space such that the neuron is highly active is exactly the place field [14] associated with that particular neuron (place cell) [178], see Fig. 7 [15].

From the seminal discovery of place cells the number of research works exploded in the following years in order to characterize their different properties. In a given environment the locations of the place fields corresponding to the different place cells are randomly positioned and the whole population of place cells is able to cover the whole surroundings [182, 255]. Another very important property of place cells is the stability of their activity over time, in the sense that after an animal has memorized an environment, *i.e.*, has associated place fields to some place cells (cognitive map), if it returns to visit the same setting after weeks the correspondence between place cells and place fields is almost unchanged [231]. In addition, the place fields are quite robust to small disturbances and transformations of external landmarks [105, 143, 175].

14. Note that in the case of animals like rats the place fields are two-dimensional while in the case for example of bats, which can fly, the latter are three-dimensional as they have access to an additional spatial dimension [262].

15. We recommend the reader to watch the attached video in which the activity of some place cells in a rat is recorded while exploring a track: https://www.youtube.com/watch?v=lfNVv0A8QvI. The measurements were made in the Wilson lab at MIT.

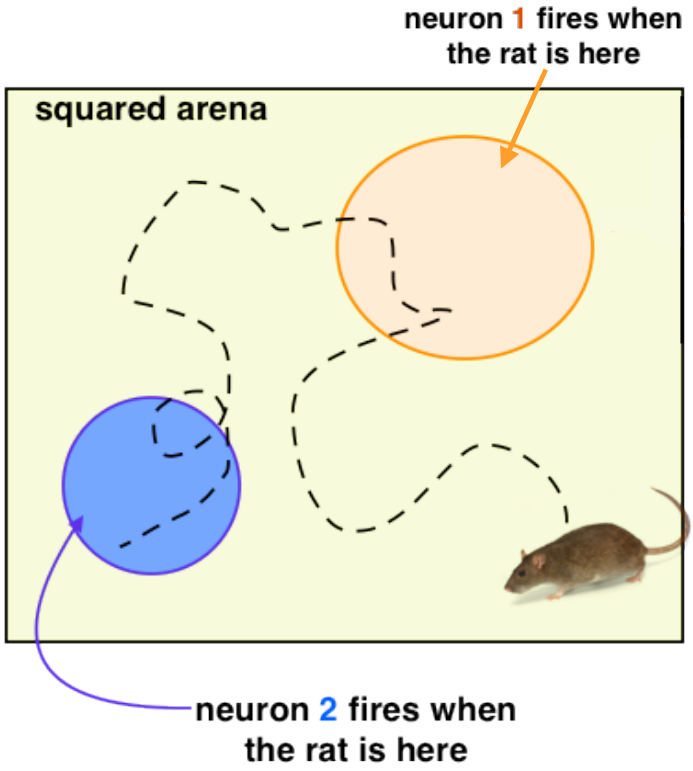

Figure 7 – Schematic of a rat exploring a square arena. The place fields corresponding to two particular place cells of the rat are shown. When the rat is located inside a place field the corresponding place cell will be very active, otherwise it will be almost silent.

Moreover, it is known that external inputs, like visual inputs, are very important for the formation of a place field by a place cell but are not necessary to maintain their activity. This property is clear in the experiment of Quirk and collaborators [198], who showed that the correspondence between place cells and place fields in an environment stored by a rat was almost unchanged if the light was turned off (no visual input) so that the animal was still able to maintain a good correlation between its position and the activity of the place cells measured. This mechanism of the rodent to integrate its linear and angular self-motion in order to maintain correlation between where it is and where the cells should activate is called path integration [153].

Finally, it is worth discussing the experiment of Buzsaki and colleagues [69], who showed that the activation sequences of groups of place cells can be generated internally by the network itself and is not necessarily due to physical inputs, in fact, they found that before performing an action in a known environment a rat activates in order, during a planning phase of the action, the same set of place cells that then will actually activate during the motion (but in a time scale about 20 times less than the physical time).

A similar mechanism occurs if the planned action is successful, with the difference that this time the sequence of place cells activity is reproduced in reverse. This process is called backward replay and is associated with a mechanism of plasticity of the synapses (learning) in the sense that the activity of the neurons is reproduced in order to strengthen/weaken the connections for example with a Hebb mechanism, see Section 3.2, or more sophisticated ones, *i.e.*, bidirectional synaptic plasticity [160]. So it is important to stress that the activity of place cells is not only generated by external physical inputs but can be generated internally by the network.

Thanks to said properties, namely spatial selectivity, stability over time, and so on, the place cells population has been proposed as a suitable candidate for the neurological basis of the cognitive map [187].

We have therefore seen that the CA3 region of the hippocampus can be anatomically schematized by a recurrent neural network in which neurons (place cells) enjoy special properties such as the fact that their activity is linked to the position of an animal in physical space, recalling the mechanisms proposed in Section 2.2 and Fig. 3(c) in the context of continuous attractors [235, 236]. How is it possible to design a model that reproduces this phenomenology? This is the purpose of the next Section 3.4.

3.4 STORING A SINGLE CONTINUOUS ATTRACTOR

We thus begin to present models of recurrent neural networks that qualitatively reproduce what was discussed in Section 3.3, *i.e.*, how to store a continuous attractor in a RNN.

There is certainly no shortage in literature of models of this type called continuous attractor neural networks (CANN) in which a large and noisy population of neurons can reliably encode “positions" in low-dimensional sensory manifolds and continuously update their values over time according to input stimuli [12, 31, 81–83, 237, 265].

The first model ever of this type is the one of Amari [12] of 1977 in which we consider a recurrent neural net composed of N neurons arranged in a ring structure and connected through excitatory couplings whose strength decays with the distance between the neurons (therefore local excitation), together with another neuron connected with all the others that is excited from these but that in turn inhibits all the others (global inhibition). From these two fundamental ingredients of local excitation and global inhibition, given an initial condition of the network through an external input, it is possible to generate a localization of the activity (bump or cell assembly). A mechanism of this type explains well what happens in the head-direction cells system or in the ellipsoid body of the flies, see Section 3.8, in which a ring attractor encodes the angle of the head of an animal (one

dimensional continuous attractor), in which simultaneously active neurons encode for close angles while the others are silent.

Without going too much into the details of Amari's model, for which we refer to the original article [12], in the following we will focus on another CANN model of fundamental importance that contains the same ingredients of the ring attractor one but has been developed specifically to explain the phenomenology of place cells and place fields presented in Section 3.3, namely the model of Tsodyks and Sejnowsky [237].

Afterwards we will focus on how to translate these concepts into a statistical physics framework and at the end we will discuss the importance of the concept of continuous attractor.

3.4.1 *Tsodyks and Sejnowsky's model*

Here we consider the CANN model introduced by Tsodyks and Sejnowsky [237] in 1995 as a model of the hippocampus CA3 network that produce place selective activity in an environment, see Section 3.3.

Let's so consider a recurrent neural network defined by the following rate equations:

$$\frac{d\nu_i}{dt} = -\nu_i + g\Big(\sum_{j\neq i} W_{ij}\nu_j + I_i\Big)\,, \tag{3.7}$$

$$W_{ij} = W_0 \exp\Big(-\frac{|\mathbf{r}_i - \mathbf{r}_j|}{\Delta}\Big) - W_1\,, \tag{3.8}$$

where ν_i is the average spiking rate[16] of the neuron i, I_i is it's external sensory input, g is the gain function of the neurons (like a ReLU function), W_{ij} is the strength of the synaptic coupling between the neurons i and j coding for locations (centers of the place fields) $\mathbf{r}_i$ and $\mathbf{r}_j$ respectively in the physical space and Δ is the diameter of the place fields, see Fig. 7. The excitatory connections in the model are mainly local and the coupling matrix W_{ij} is an exponential function of the distance between their place fields centers. The uniform inhibitory inputs W_1 can be considered as a global feedback inhibition.

If in Eq. (3.7) we do not consider the g term, we get a relaxation equation, this means that after some time the firing rate of all the neurons is zero. What makes the system non-trivial is the fact that the neurons are connected via the couplings $\{W_{ij}\}$, and possibly an $\{I_i\}$ external input. If we choose the couplings intelligently so that to have a mapping between the neurons in the network and the position they encode in physical space by

16. Note then that this is a model with continuous and positive neurons, unlike the standard Hopfield model defined with binary units, see Section 3.2.

choosing excitatory couplings if the distances in physical space are small and inhibitory otherwise, we will be able to create a bump of activity that means that neurons that are active at the same time encode for close positions in physical space. Moreover, this bump can move due to the introduction of an eventual noise term in the dynamics (3.7) or to weak inputs $\{I_i\}$, thus defining a way to move along the attractor.

3.4.2 *CANN and statistical physics: Lebowitz and Penrose's model*

Now let's try to put together the ideas just presented and formalize a very simple model with which we can make a mapping with a classic model of statistical physics, the Lebowitz and Penrose's model [141].

Hence, we start as usual with defining a set of N binary neurons, which can take either $\sigma_i = 1$ or $\sigma_i = 0$ depending on whether they are active or silent. Moreover, these neurons are interconnected through an appropriate matrix of connectivity $\{W_{ij}\}$ (without self-connections), see Fig. 3(a).

The weights of the net in this case must be chosen in order to store a continuous variable such as the position of an animal in an environment, as we have seen in Section 3.3. To do this we start by defining a simplified cognitive map model to be memorized, that is, we consider a square environment [17] of unitary area and with periodic boundary conditions [18] where we associate to each neuron a place field [19], *i.e.*, a fixed radius disk [20] located in a random position $\mathbf{r}_i$ of the environment, see for example one of the two environments in Fig. 14(b).

It is important to note that there is no relationship between the arrangement of the different place fields in an environment and the arrangement of the corresponding place cells in the neural network also because, as we will discuss in Section 3.5, the same network must store different environments in which the different place fields associated with the same neuron do not have any correlation (global random remapping).

17. Here we focus on the two-dimensional case in analogy with the place cells and place fields in rodents discussed in Section 3.3, but the following is trivially generalizable to any finite dimension D.

18. The use of periodic boundary conditions is justified only in the modeling of head-direction cells where the attractor to be memorized is actually a ring in $D = 1$, see Section 3.8. Already in the case of place cells this assumption is no longer valid because the presence of edges (walls) in an environment is important, see Section 3.8, nevertheless we will continue to use this assumption for theoretical convenience.

19. In general it is not true that in a given environment each neuron has a place fields, in fact typically about thirty percent of neurons have it while the others do not (silent cells) [9], for now we do not take into account this aspect for simplicity.

20. From experiments it is known that place fields can have different shapes, but for the moment we consider place fields all the same and circular with the same radius for simplicity.

Once an environment model has been defined, it is necessary to associate the possible locations with a corresponding network activity status. If we consider a specific position $\mathbf{r}$ on the map we will consider that the neuron i is active if the distance between the position $\mathbf{r}$ and the center of the corresponding place field $\mathbf{r}_i$ is less than the radius of the place field, *i.e.*, the position is within the place field associated with neuron i, see Fig. 7.

To memorize an environment we mean that the set of patterns associated to the different positions, which are obviously spatially correlated by construction, must be fixed points of the network dynamics and this defines a continuous attractor, see for example one of the manifolds in Fig. 3(c).

Now it is important to understand how to choose the synapses of the network to memorize this continuous attractor. In the simple case of the Mattis model, see Section 3.2, we had that interactions were only a function of the specific pattern to store, see Eq. (3.5), but now we are interested in memorizing all the different positions associated with the spatial map.

The standard way to do this, following what already seen in Tsodyks and Sejnowski's model, is to consider a connectivity matrix in which the synapses between neuron i and neuron j are a function of the distance between the respective place fields in the environment with centers located in $\mathbf{r}_i$ and $\mathbf{r}_j$, that is:

$$W_{ij} = w\big(|\mathbf{r}_i - \mathbf{r}_j|\big) \ , \tag{3.9}$$

where $|\cdot|$ denotes the distance in the map. If w is sufficiently excitatory at short distances and inhibitory at long ones, a bump state spontaneously emerges, in which active neurons tend to code for nearby positions in the map. Classic examples of kernels w are for instance exponential, step function or Gaussian ones.

It is important to note that the mechanism presented in Eq. (3.9) is reminiscent of Hebb's rule as the concept that neurons that fire together wire together remains valid. In fact, neurons that have place fields nearby will often be active together so we need to strengthen the connection between them, while neurons whose place fields are far apart will hardly be active together so it is better to have a weak connection between them.

To start studying the distribution of neural activity configurations associated to this kind of model what is missing is the definition of a dynamics associated to the network. We consider here a scheme of the same type already used in the Hopfield model, see Section 3.2, but with the introduction of a finite temperature, given by the following

formula that explain how neuronal states are updated stochastically according to the probabilities

$$\text{Prob}\left(\sigma_i(t+1)|\{\sigma_j(t)\}\right) = \frac{1}{1+\exp\left[-\frac{1}{T}\left(2\sigma_i(t+1)-1\right)\left(\sum_{j(\neq i)} W_{ij}\,\sigma_j(t)-\theta_i\right)\right]}\,, \quad (3.10)$$

where T is a temperature parameter to be set such that a bump of activity may form and sustain itself [21].

The thresholds $\{\theta_i\}$ in this dynamics are simply effective fields that keep the activity of the network at a fixed level typically giving an inhibitory contribution (they are negative and somehow schematize the effect of the interneurons) and are necessary because the connections between place cells in Eq. (3.9) are instead excitatory (positive).

It is interesting to note that a model of the kind just presented is very similar to a very old model of statistical physics introduced in 1966 by Lebowiz and Penrose to explain the liquid/vapor transition [141, 168].

Consider a D-dimensional lattice, whose N sites $\mathbf{x}_i$ can be occupied by a particle ($\sigma_i = 1$), or left empty ($\sigma_i = 0$) [22]. The energy of a configuration $\{\sigma_i\}$ is given by the Ising-like Hamiltonian

$$E\left[\{\sigma_i\},\{\mathbf{x}_i\}\right] = -\sum_{i<j} w\left(|\mathbf{x}_i-\mathbf{x}_j|\right)\,\sigma_i\,\sigma_j\,, \quad (3.13)$$

where w is a positive and decaying function of its argument, *i.e.*, of the distance between sites. At fixed number of particles and low enough temperature T translation invariance on the lattice is spontaneously broken: particles tend to cluster in the $\mathbf{x}$-space, and form a high density region (liquid drop) surrounded by a low-density vapor. The density profile of this "bump" of particles hardly fluctuates, but its position can freely move on the lattice, and defines a collective coordinate for the microscopic configuration of particles, see Fig. 8.

21. Note that in the null temperature limit, *i.e.*, $T = 0$, we obtain a deterministic dynamics of the type:

$$\sigma_i(t+1) = \Theta\Big(\sum_{j\neq i} W_{ij}\,\sigma_j(t)-\theta_i\Big)\,, \quad (3.11)$$

where Θ is the Heaviside step function defined as:

$$\Theta(x) := \begin{cases} 0, & \text{if } x<0 \\ 1, & \text{otherwise} \end{cases}\,. \quad (3.12)$$

So exactly the same dynamics as the Hopfield model rewritten for spins 0 or 1 instead of ± 1, see Section 3.2.

22. The fact that the particles are on a grid here whereas previously we have seen that place fields have random positions in the environment does not change the phenomenology of the model in the large N limit.

Fixing the number of particles here is equivalent to find in our model of CANN an inhibitory threshold that keeps the number of active neurons in the network constant.

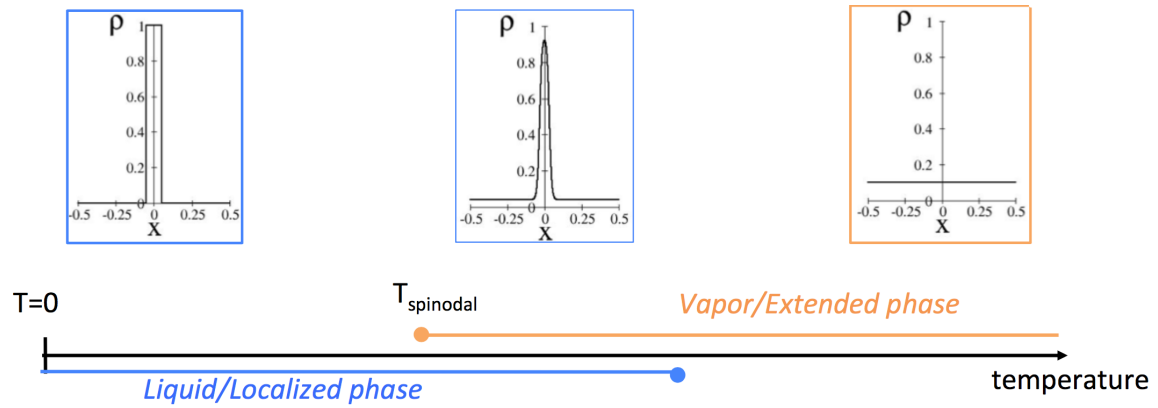

Figure 8 – Phase diagram of Lebowitz and Penrose's theory of the lquid/vapor transition, in dimension $D = 1$ (periodic boundary conditions). Insets show the density of particles $\rho(x)$ as a function of position over space, $x \in [0; 1]$. Note the coexistence between the homogeneous and bump states at intermediate temperatures. The location of the bump is arbitrary. Figure taken from [60].

3.4.3 *Continuous attractors and population coding*

From the neuroscience point of view, the existence of a collective coordinate, weakly sensitive to the high stochasticity of the microscopic units, is central to population coding theory.

It should now be clear the equivalence between the RNN models to store a continuous attractor and the Lebowiz and Penrose model. The key in both cases is in fact to have synapses that are functions of some distance, in the first case the distance between the place fields in the environment and in the second case the distance between the particles. Moreover, the fact of creating a liquid droplet at low temperatures in the second case is equivalent to create a bump of activity in the first case in which the neurons that are active at the same time code for nearby positions in the map.

Since the system is invariant under translation, the bump of activity can spread both because of the temperature in the dynamics and because of a weak external input and explore all the positions associated with the environment which are equivalent to each other (no preferred positions), this means that with this choice of the connectivity matrix we have actually managed to store a continuous attractor.

Moreover, the shape of the bump is determined by the Hamiltonian of the system but its position is totally random. So we got exactly what we wanted, a model of a population of noisy neurons in which we have the emergence of a collective coordinate (the position of an animal in an environment) that is robust to the error of the single neuron, *i.e.*, a robust encoding. Obviously the number of coded coordinates is equal to the dimension of the stored map and it is also important to note that the activity bump may persist in the absence of external input, exactly as we saw for the place cells in Section 3.3.

In addition, we notice that we can have two points of view with this kind of models, either we look at a fixed time to the activity of the neuron population and we see the emergence of this collective variable coding for the position in the map, see Fig. 9, or we look at the activity of a single neuron and we see that this is active when the distance of the collective variable, *i.e.*, the position of the animal in the environment, is close to the center of the place field of the neuron considered, which is instead what we typically see in experimental measurements, see Section 3.3.

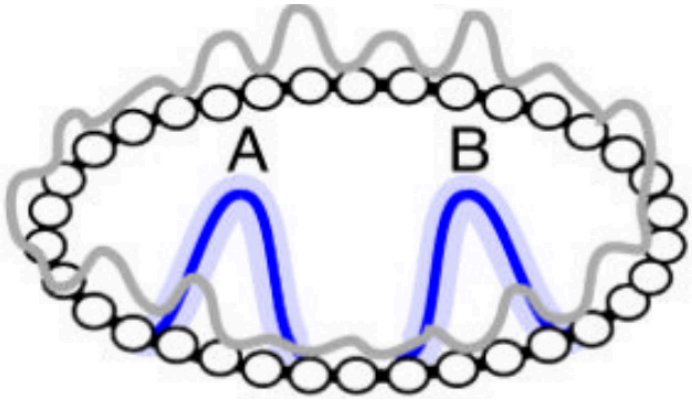

Figure 9 – An example network of N neurons (small circles) with 1D continuous attractor dynamics. Local excitatory and global inhibitory connections (not shown) between all neurons stabilize population states that are local activity bumps (*i.e.*, blue bump A or B; gray: transient/unstable activity profiles). Figure adapted from [264].

In general models of CANNs have many computationally appealing properties, such as efficient population decoding, smooth tracking of moving objects, and implementing parametrical working memory. The computational advantages of CANNs and their successes in modeling brain functions have suggested that CANNs serve as a canonical model for neural information representation [260].

3.5 WHY IS IT NECESSARY TO STORE MULTIPLE CONTINUOUS ATTRACTORS?

We have therefore seen in detail how to store a continuous attractor in a recurrent neural network, see Section 3.4, that could correspond to the memorization of a cognitive map of an environment by a network of place cells, see Section 3.3. In reality, however, the story is more complicated than this because the hippocampus needs to contain simultaneously several cognitive maps corresponding to different environments and also, for the same environment, to different contexts. But first of all, how do the different cognitive maps relate to each other?

For different environments, place fields associated to different place cells can re-position in a supposedly random way, a property called "global remapping" [23] [121, 143, 175, 185], or rather, given a certain place cell this can encode positions (have a place field) in different environments and the location of the different place fields, corresponding to the same cell, seems to be totally random without any correlation between the various maps.

The fact that the different maps are unrelated because of random remapping will be one of the key elements to model a recurrent neural network that stores multiple continuous attractors, see Section 3.6, because in this way the configurations of neurons encoding different positions in different environments are orthogonal to each other with high probability, this recalls the case of the Hopfield model, *i.e.*, multiple discrete attractors, where random (orthogonal) patterns were chosen in order to limit the interference between them, see Section 3.2, here the idea is exactly the same.

An important example of this mechanism is provided by the experiment of Alme and collaborators [9] in which rats are trained to store eleven different rooms. Once these environments have been memorized the animals are able to recall the cognitive maps according to the room they are in, and the different maps present the above mentioned random remapping phenomenon for the place fields in the various rooms associated to the same place cell. This experiment shows how the hippocampus can store several maps at the same time, raising the natural theoretical question, obviously very complicated to study experimentally and of which we do not have biologically plausible estimates, on the maximum number of continuous attractors that a network of this type can store (critical capacity), which we will start to deal with in Section 3.6. It has to be said, however, that despite the difficulty of experimentally measuring the network capacity, we know that, for example, wild rats (and not only laboratory ones) are able to navigate perfectly in

23. Regarding the hippocampal region CA3 we have only global remapping. In reality, however, there is also another type of remapping, called "rate remapping", in which a place fields vary in the frequency of spikes of the respective place cells without changing its position in the map, which is present in another region of the hippocampus, that is CA1 [20] (which we will not consider in the models presented in this thesis).

many environments, so it is reasonable to think that they are able to store a large number of cognitive maps in their hippocampus.

Moreover, the specific positioning of place fields are flexible and might shift or entirely re-arrange upon drastic changes in external landmarks and boundaries [130, 175], odors [8], or even abstract variables such as contextual conditions or the task to be performed [8, 122, 124, 220].

Finally, it is important to note that it is possible to have different cognitive maps also depending on the sensory modalities used by the animal, for example Geva-Sagiv et al. [90] showed in bats that they used different cognitive maps (which differed through global remapping) for the same environment depending on whether the animal used as external inputs the visual one or echolocation. Other examples where global remapping is observed by changing or combining different input modalities, such as visual and path integration, include experiments performed in virtual reality where one has full control over the different input modalities to which the animal is subject [56].

At this point it is therefore obvious that the next step is to try to combine the Hopfield model for multiple discrete attractors, see Section 3.2, with what we have seen in Section 3.4 for single continuous attractors in order to have a model that can reproduce the phenomenology just discussed. This will be the purpose of the next Section 3.6.

3.6 THE CASE OF MULTIPLE CONTINUOUS ATTRACTORS

We discussed in Section 3.4 how to store a single D-dimensional attractor in a recurrent network, but it is clear from Section 3.5 that it may be necessary to have a model of a RNN that stores many continuous attractors in the same connectivity matrix. To solve this problem Samsonovich and McNaugthon presented the following RNN model [208] in 1997 that is kind of a mix between the Hopfield model to store multiple discrete attractors, see Section 3.2, and the model presented in Section 3.4 to store a single continuous attractor [208].

Then we will present a version of this model inspired by statistical physics, that is the model of Rosay and Monasson [166], with which it is possible to study in particular the critical capacity of RNN that store multiple continuous attractors, that is the maximum number of D-dimensional attractors L that can be stored in a network of N neurons.

3.6.1 *Samsonovich and McNaugthon's model*

In the models presented for a single continuous attractor, like the model of Tsodyks and Sejnowsky (Section 3.4), we have seen how to memorize a map it is essential to choose connections between place cells so that they are a function of the distance between the centers of the place fields in the map of the respective neurons, see Eq. (3.8). Now instead of storing only one map, we want to memorize L of them in the same connectivity matrix, where the different maps differ because of the random remapping phenomenon discussed in Section 3.5, *i.e.*, given a place cell, this encodes a position (has a place field) in each map and these positions are totally random and uncorrelated in the different maps. Based on this fact and also the Hebb rule presented in the Hopfield model where the contributions of the different patterns were summed in the connectivity matrix, see Eq. (3.3), Samsonovich and McNaugthon proposed the following rule for the couplings matrix of a RNN that must store multiple continuous attractors [208]:

$$W_{ij} = \sum_{\ell=1}^{L} \exp\left(-\frac{|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|^2}{\Delta^2}\right), \tag{3.14}$$

where $\ell = 1, \dots, L$ and $\mathbf{r}_i^\ell$ is the location of the place-field center of the cell i in the map ℓ. Moreover, using a network dynamics of the type defined in Eq. (3.7) they showed that when one cognitive map is retrieved, the activity is organized as a coherent 2D bump on the corresponding map, while looking scattered and uninformative on all the other maps (how it must be given that at a fixed time an animal is in a specific map position and also because of the random remapping the different maps are orthogonal to each other).

3.6.2 *Rosay and Monasson's model*

From the model discussed above we can therefore understand that to store multiple continuous attractors in one RNN we can in general use the rule defined in Eq. (3.9) for each of the maps separately and then consider as connectivity matrix the sum of the connectivity matrices of the single environments, obtaining in this way, assuming each one of the L maps contributes equally to the learning process, a rule of type

$$W_{ij} = \sum_{\ell=1}^{L} w\big(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\big)\,, \tag{3.15}$$

where $\mathbf{r}_i^\ell$ is the center of the place field of neuron i in environment ℓ.

In particular, what we mean to store more than one map in the same connectivity matrix is precisely the fact that if we consider the dynamics of network [24] (3.10) with rule (3.15) at some time we will have a bump of activity in one of the maps (the map of which we are recalling some specific position) and basically noise in the others or rather that the fixed points associated with dynamics (3.10) at the zero temperature are all the configurations of the neurons built starting from a position $\mathbf{r}$ in one of the maps in which the place cells associated with the place fields close to $\mathbf{r}$ are active and the others silent. Obviously we will have the bump only in one of the maps at a fixed time because the different maps are orthogonal to each other (random remapping), *i.e.*, the activity patterns that encode for the positions of a specific map will substantially give noise in the other maps or better that the patterns encoding for the different positions in the different maps are orthogonal with high probability.

In the following we consider the model of Rosay and Monasson [25] [166], inspired by the model of Lebowitz and Penrose (Section 3.4), in which the issue of storing multiple continuous attractors in a RNN is set as a statistical physics problem.

So we consider a network of binary neurons where the N place cells are modeled by binary units σ_i equal to 0 (silent state) or 1 (active state). These neurons interact together through excitatory couplings $\{W_{ij}\}$. Moreover, they interact with inhibitory interneurons, whose effect is to maintain the total activity of the place cells to a fraction f of active cells (global inhibition). We also assume that there is some stochasticity in the response of the neurons, controlled by a noise parameter T. All these assumptions come down to considering that the network states are distributed according to the Gibbs distribution associated to the Hamiltonian

$$E_W[\{\sigma_i\}] = -\sum_{i<j} W_{ij}\sigma_i\sigma_j\,, \tag{3.16}$$

restricted to configurations of spins $\{\sigma_i\}$ such that

$$\sum_i \sigma_i = f\,N\,. \tag{3.17}$$

We want to store $L+1$ environments in the coupling matrix, indexed by ℓ, each defined as a random permutation π^ℓ of the N neurons' place fields that are initially arranged on a D-dimensional grid (the case of only one map is reminescent of the Lebowitz and

24. From now on we will continue to use only binary neurons, not as in the above mentioned model of Samsonovich and McNaugthon.

25. See as well the Battaglia and Treves's model [27] which has several points in common with the one we are describing.

Penrose's model). This schematize the random remapping of place fields from one map to the other. With this definition, an environment is said to be stored when activity patterns localized in this environment are stable states of the dynamics. In other words, the configurations where active neurons have neighbouring place fields in this environment are equilibrium states. To make this possible, we assume a Hebbian prescription for the couplings $\{W_{ij}\}$ that is a straightforward extension of the Hopfield synaptic matrix, see Section 3.2, to the case of quasi-continuous attractors. This rule is described as follows:

— additivity: $W_{ij} = \sum_{\ell=1}^{L} W_{ij}^{\ell}$ where the sum runs over all the environments.

— potentiation of excitatory couplings between units that may become active together when the animal explores the environment:

$$W_{ij}^{\ell} = \frac{1}{N} \text{ if } d_{ij}^{\ell} \leqslant d_c \ , \quad 0 \text{ if } d_{ij}^{\ell} > d_c \ , \tag{3.18}$$

where d_{ij}^{ℓ} is the distance between the place field centers of i and j in the environment ℓ. d_c represents the distance over which place fields overlap. This distance is chosen in such a way that each neuron i is connected to the same number of other neurons j, regardless of the spatial dimension D. If $\hat{w}N$ is this number: $\hat{w}$ ($\ll 1$) is the fraction of neurons to which each neuron is connected. So for example $d_c = \frac{\hat{w}}{2}N$ in $D = 1$ and $d_c = \sqrt{\frac{\hat{w}N}{\pi}}$ in $D = 2$. The $\frac{1}{N}$ factor in Eq. (3.18) ensures that the total input received by a cell remains finite as N goes to infinity.

Rosay and Monasson [26] were able to show in this model that in the thermodynamic limit it is possible to store an extensive number of maps L (at fixed dimension D) with a recurrent neural network composed of N neurons if the ratio $\alpha = \frac{L}{N}$ is less than a critical capacity $\alpha_c(D)$, where this quantity is of order 1. It's important to note that while in the Hopfield model, see Section 3.2, the load parameter α was defined as the ratio between the number of patterns P to store divided by the number of neurons N, now here we have the number of maps L at the numerator (each of which corresponds a priori to many patterns, one for each position in the map to store).

We show in Fig. 10 the phase diagram associated with this model in the case $D = 1$, $f = .1$ and $\hat{w} = .05$ [27]. It is then interesting to make a comparison of this phase diagram with the one concerning the Hopfield model, see Fig. 5(c). In both phase diagrams there are different phases depending on temperature, in the Rosay and Monasson case in particular the clump phase (CL), the paramagnetic phase (PM) and the spin glass phase (SG). This CL phase is the equivalent in the Hopfield model of the ferromagnetic phase, and is the

26. See [166] for all the details on the replica computation which led to finding the phase diagram of the model. This computation is somehow a generalization of the AGS one for the standard Hopfield model, see Section 3.2.

27. The phase diagram remains qualitatively similar to variations in model parameters, see [166].

region of the phase diagram interesting to consider in order for the neural network to function as associative memory. The neural configurations allowed in the CL region are those in which a bump of activity is able to form in a random position of one of the maps, in the other phases this is not possible. At high noise level (high T), the system is in the PM phase, where no coherent representation is formed and the population activity is comparable to random. Finally, when the number of stored maps exceeds the critical capacity ($\alpha_c(D,T)$) the system falls into the SG phase, a behavior characterized by the presence of many local minima where the effective noise induced by the competition between maps freezes the activity and no spatial or map selectivity is achieved.

In addition, from a qualitative comparison with the Hopfield model it seems here that the system is more robust against noise, in fact, the ferromagnetic region of the Hopfield model has a triangular shape while the CL phase in this case is more squared, suggesting that the storage of correlated patterns is more robust against noise than completely random patterns.

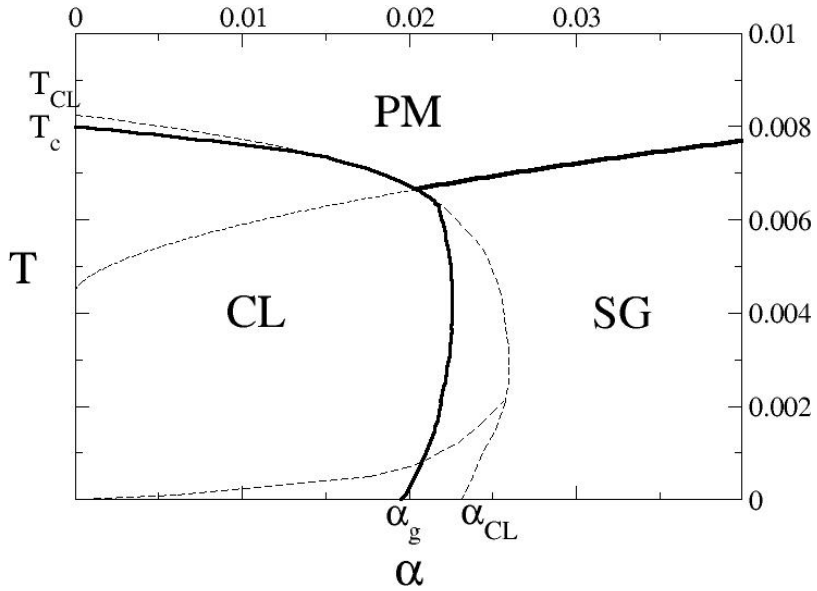

Figure 10 – (a) Phase diagram in the $(T, \alpha = \frac{L}{N})$ plane in $D = 1$, $f = .1$ and $\hat{w} = .05$ for the Rosay and Monasson's model. Thick lines: transitions between phases. Figure taken from [166].

Besides the static properties of the model, Rosay and Monasson have also studied its dynamical features [167] that are much richer than in the standard the Hopfield setting. First of all, in the case of a single environment, *i.e.*, $\alpha = 0$, we have the emergence of a quasi-particle (the bump) with an activated diffusion, expected from what we have discussed in Section 3.4; furthermore, in the case of several maps, *i.e.*, $\alpha > 0$, in addition to

the diffusion of the bump within a map we have as well spontaneous transitions between the different maps [28], see Fig. 11 [29].

In fact, a dynamics at a moderate noise level will most of the time tend to move in the space of the configurations in one of the manifolds (composed of a set of spatially related fixed points) dictated by the initial condition, but occasionally it may jump to another set of fixed points corresponding to another map, see Fig. 3(c). In the case of the Hopfield model instead the situation was much simpler because the different patterns corresponded to different isolated fixed points so there may be only, with a sufficient level of noise, drastic jumps between a fixed point and another, see Fig. 3(b).

3.7 ISSUES WITH CURRENT THEORY

The type of models discussed in Section 3.6 represent the state of the art to date on how to store multiple continuous attractors in a recurrent neural network by using the ad hoc prescription defined in Eq. (3.15) for the pairwise interactions between neurons, despite the presence of the following important problems both from a practical and theoretical point of view.

1. First of all, due to interferences (crosstalk) between the different manifolds, the attractors are not continuous any longer, and are effectively discrete. How to practically learn smooth attractors and ensure a prescribed level of spatial resolution, *i.e.*, in a nutshell the typical distance between nearest neighbour fixed points in every attractor, is an open issue.

   In fact, as soon as we want to store more than one map in the RNN, the activity bump gets stuck in some preferred locations in the retrieved map due to the interferences coming from the other non-retrieved maps [54]. In other words, rule (3.15) does not define truly continuous attractors, as large barriers oppose the motion of the bump along the map [167].

   To better understand this phenomenon, we refer to the case of having to memorize only one continuous attractor. In this case, in fact, as we have seen in Section 3.4, there is invariance under translation, in the sense that if we consider two neurons coding for two positions (so having their place fields centered in that positions) on the map, the interaction between them depend only on the distance between these positions, see Eq. (3.9), consequently, an activity bump can be located in any position of the environment at the same cost, see Eq. (3.13). But as soon

28. A full characterization of the phenomenology of these spontaneous transitions can be found in [168].

29. See also the video at https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.115.098101. Note that the circles of different colors represent the centers of the place fields whose corresponding place cells are active at some time in the different maps.

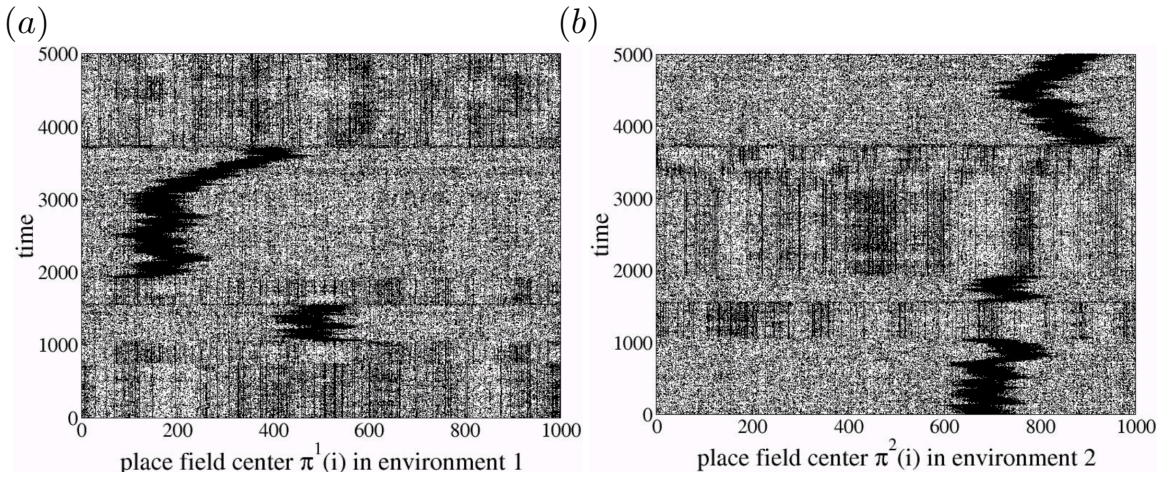

Figure 11 – Time evolution of a 1D network with N = 1000 units and 2 stored environments observed in a Monte Carlo simulation at T = 0.007, illustrating the coexistence of diffusion within one map and transitions between maps. Each black dot represents an active unit. Both panels represent the same data, they only differ by the ordering of the units along the x-axis (remapping of the place fields). (a) the units i are arranged according to their place field centres in environment 1. (b) they are arranged according to their place field centres in environment 2. The y-axis represents time (in Monte Carlo rounds). Between time 0 and time ~ 1000 the activity is localized in map 2 and delocalized in map 1. Then it undergoes several transitions between 1 and 2. Between times ~ 2000 and ~ 3700, the network is in attractor 1 and the bump diffuses within this attractor. Finally, it ends up and diffuses in map 2. Note the abruptness of the transitions between maps dynamics of quasi-continuous attractors is much richer than in the case of more basic models as the Hopfield network, where the only possible evolution is to transit from one attractor to the other, see Fig. 3(b). Figure taken from [167].

as we try to store in the same connectivity matrix more than one map with rule (3.15), and consider a bump of activity that has formed in one of the maps, we will break this invariance under translation because in the connections there will be the contributions of the other maps (that we are not recalling) that will act as a noise term, this can be described with the fact that in the case of a single manifold the bump is travelling in an effective flat free energy landscape while in the case of multiple maps this landscape is rough due to this crosstalk term, see Fig. 12 and [167]. This means that if we are at a low level of neural noise (low temperature) the bump will remain stuck in one of the free energy minima, in other words we have no longer stored a continuous attractor but only a discrete set of positions that correspond to the minima of this free energy.

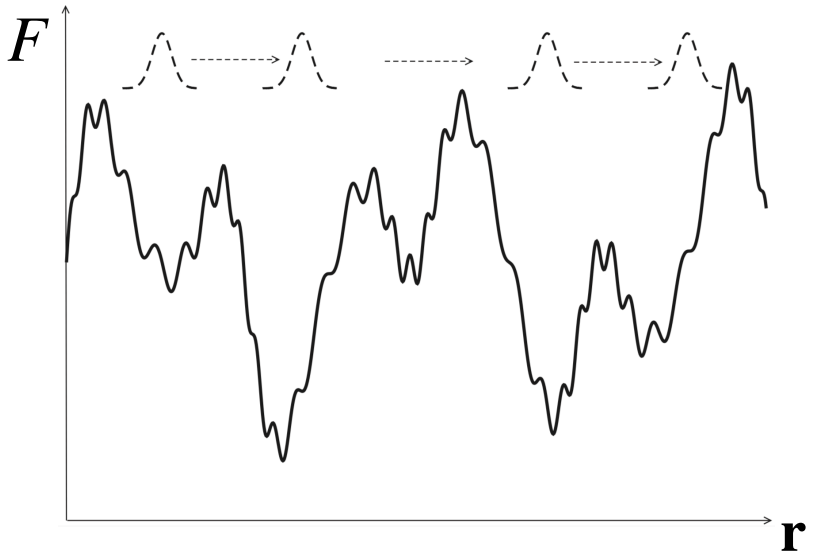

Figure 12 – Sketch of the effective free energy landscape probed by the bump of neural activity (dashed bump) moving through an environment in the case of storing multiple continuous attractor in a RNN. Figure taken from [167].

2. This issue of spatial resolution is also unclear from a theoretical point of view. Capacity calculations [27, 166] require that a bump can form in any of the memorized maps, in at least one position: they offer no guarantee about the existence of other memorized positions, that is, we have no explicit control over which positions we actually have stored in each map (what are precisely the fixed points of the neural dynamics?).

3. Moreover, theoretical studies [27, 166, 168] show that the maximal capacity corresponding to the prescription (3.15) is very low, see Fig. 10. It is reasonable to expect that the optimal storage capacity could be much higher: a $\sim$ 15-fold increase was found from the Hebb-rule critical capacity, $\simeq 0.138$ [17], to the optimal capacity, $\alpha_c = 2$ [87] in the case of 0-dimensional attractors, corresponding to the Hopfield model [112].

   Where by optimal capacity we mean that the weights of the neural net have been chosen to maximize the number of attractors (number of maps at fixed number of stored positions per map) to memorize, rather than making an a priori choice for the couplings and then see how many maps can be memorized with that choice following the Hebb-like rules (3.15) approach.

4. Optimal learning could also provide detailed insights on the statistical structure of the neural couplings $\{W_{ij}\}$, which could be compared to the physiological distribution of synaptic connections [46].

The purpose of this thesis is to solve these problems. In fact, as we shall see in the next Chapters, we will show how to achieve optimal storage of continuous attractors and study the optimal trade-off between capacity and spatial resolution, that is, how the requirement of higher spatial resolution affects the maximal number of attractors that can be sustained in a RNN.

3.8 EXPERIMENTAL EVIDENCES FOR CONTINUOUS ATTRACTORS

Before presenting a solution to the problems posed in Section 3.7, it is important to provide other examples in the brain of neurons whose activity depends strongly on some continuous sensory correlates in addition to the place cells that we have already discussed in detail, see Sections 3.3 and 3.5.

3.8.1 *Head-direction cells*

The place cells we have already described are not the only example of neurons in the brain showing a sharp spatial selectivity in their activation properties. In fact, after their discovery, the head-direction (HD) cells in the septal presubiculum were found [200, 227]. These cells have the surprising property of being active or silent depending on the orientation of a rodent's head with respect to a reference direction of the environment, see Fig. 13(b).

Neurons responsive to the direction of the motion have later been discovered in other regions, such as the entorhinal cortex [209], the anterior and lateral dorsal thalamic nuclei [162], the lateral mammillary nucleus [216], the retrosplenial cortex [57] and the striatum [252], suggesting that the directional signal can be calculated in regions of the brain outside the hippocampal formation [243].

Since HD cells fire allocentrically and depending only on the ongoing direction of the animal, and not on the specific location within the environment, they have been interpreted as the "compass" used for navigation in the cognitive map [227]. HD cells primarily rely on external landmarks to represent the motion direction [228, 229], although they are known to respond to self-motion cues [95] as well as contextual conditions [119] when visual information is unavailable or unreliable [130].

3.8.2 *The fruit fly central complex*

It is also very important to mention experimental evidence in animals simpler than rodents, here in fact we will focus on the fruit fly, in fact even insects are excellent in space navigation but at the same time they are much simpler than rodents and this allows to study the neural circuits really in detail with state-of-the-art experimental techniques such as two-photon calcium imaging in head-fixed walking flies and optogenetics in order to get a mechanistic description of the latter. In particular here we will discuss neurons that are equivalent to HD cells previously discussed for rodents giving the first direct experimental evidence of the mechanism of continuous attractors in the brain.

Specifically, let's consider a part of the brain of the fruit fly, called central complex, common with some variations to a large class of insects, and in particular let's focus on a sub-part called ellipsoid body, where are arranged in a ring a type of neurons named compass cells or E-PG neurons (similar to HD cells).

Looking at the activity of these neurons Seelig et al. [214] have identified that this ring structure presents a single activity bump that acts as an abstract representation of the fly orientation and where the activity of the network is confined to move on this ring in the neural activity space. Moreover, this representation persists in the absence of visual and self-motion cues.

Then, using the fact that in this context we have a compass network with a strong topography, that it is possible to monitor the activity of entire populations of cell types, that it is possible to selectively perturb the activity of specific neurons and that one has access to both structural and functional connectivity, Kim et al. [127] have been able to demonstrate that the neural circuitry of the ellipsoid body of the fly must be implemented by a recurrent network with local excitation and uniform inhibition, *i.e.*, the same type of models discussed in Section 3.4, giving for the first time direct evidence of this mechanism in the brain.

Another important question to be understood from a mechanistic point of view is how the bump formed in the compass network can move in the absence of visual stimuli (in the dark). To do this Turner et al. [241], always using two-photon calcium imaging and electrophysiology in head-fixed walking flies, were able to identify in the central complex another population of neurons, called P-ENs, that simultaneously encodes heading and angular velocity, and is excited selectively by turns in either the clockwise or counter-clockwise direction. They showed how these mirror-symmetric turn responses combine with the neurons' connectivity to the E-PG neurons to create a mechanism for updating the fly's heading representation when the insect turns in absence of visual inputs.

It should now also be clear that the study of animals as simple as drosophila is essential to understand the functioning of neural circuits from a mechanistic point of view (much more complicated thing to do in more complex brains like those of mammals) and in fact this line of research is proving extremely fruitful today [103, 126]. In addition, mechanisms similar to those presented for insects have been successful even in more complex animals such as mammals (bats) [77].

3.8.3 *Grid cells*

Another key example is the one provided by the grid cells discovered by E. and MB. Moser (also Nobel Prize winners together with O'Keefe in 2014 for this finding) in the medial entorhinal cortex (mEC) [85, 105]. These neurons have the particular property of exhibiting triangular, periodic grid-like spatial selectivity, see Fig. 13(c).

The grid cells can be described by the orientation and period of the grid. These cells enjoy a property called "topography" which consists in the fact that neighboring neurons in the mEC have similar period, or grid spacing, while this increases along the dorsoventral axis of the cortex [45]. Moreover, grid cells have the fundamental property of not changing their mutual relationship in different environments, namely that the superimposition of their firing fields is constant regardless of external conditions, since the firing fields translate and rotate coherently in different familiar environments [84].

The independence of spatial encoding from the context of grid cells, as opposed to the complex variability of place cells, has led to their interpretation as a putative substrate for the representation of a universal metric for navigation [105, 153, 184, 264]. A context-independent metric is essential to perform path integration, *i.e.*, the process of updating one's cognition of self-location based on the estimation of linear and angular direction and velocity from proprioception and vestibular information, which allows for navigation in a known environment even in the absence of visual guidance, as we had already mentioned in Section 3.3.

Moreover, since anatomically mEC projects into the hippocampus, it is thought that the weighted sum of grid cells activity can produce localized activity (place field) in the hippocampus and also that changes in these weights may be responsible for the phenomenon of random remapping. The process of formation of place cells from grid cells is an hot topic of study today, in fact, for example, based on the above, it is not clear why in newborn animals are observed place cells before the emergence of grid cells. Experiments in which the activity of neurons is measured simultaneously in mEC and hippocampus could help to better understand this fundamental problem [49].

It is important to note, from the point of view of mathematical models of recurrent neural networks that store continuous attractors, that the key difference between single

activity bump models such as place cells or even head-direction cells and models with multiple bumps (in this case the bumps are located on the triangular lattice), lies on the fact that in the first case we have a global inhibition mechanism, while in the other case we must have an inhibition that depends on the distance, so a local and not global inhibition (just having a local excitation is not enough) [217, 223].

3.8.4 *Prefrontal cortex*

Indirect evidence for the continuous attractor mechanism is present as well in the prefrontal cortex (PFC), a part of the mammalian brain that subserve higher cognitive functions like decision making and working memory, during the delay period in an oculomotor delayed response task.

Indeed, Wimmer at al. [256] analyzed behavioral and electrophysiological data from monkeys performing a spatial working memory task and tested predictions from a continuous attractor hypothesis, see Sections 2.2 and 3.4, for spatial memory maintenance. Their analysis support the idea that PFC activity represents spatial memories in a fixed-shape bump of activity that is used for guiding behavior, but is liable to cumulative encoding errors as a result of random fluctuations.

Thus validating the concept of prefrontal persistent activity as the basis of spatial working memory, supporting the continuous (or finely discretized) nature of spatial memory encoding in PFC, and being consistent with bump attractor dynamics mediating cognitive function in the cortex.

3.8.5 *Other examples*

Other examples for the mechanism of continuous attractors in the brain include:

— simple cells in the V1 area of the visual cortex coding for the orientation of a bar presented to the retina, see Hubel and Wiesel (Nobel prize winners in physiology or medicine in 1981) experiments [114, 115];

— border cells [183] responsive to the walls of an environment were reported in several regions of the hippocampal formation, see Fig. 13(d);

— speed cells [136] that are neurons which firing rates depend on an animal's speed through its environment;

See [131, 260] for nice reviews on the topic.

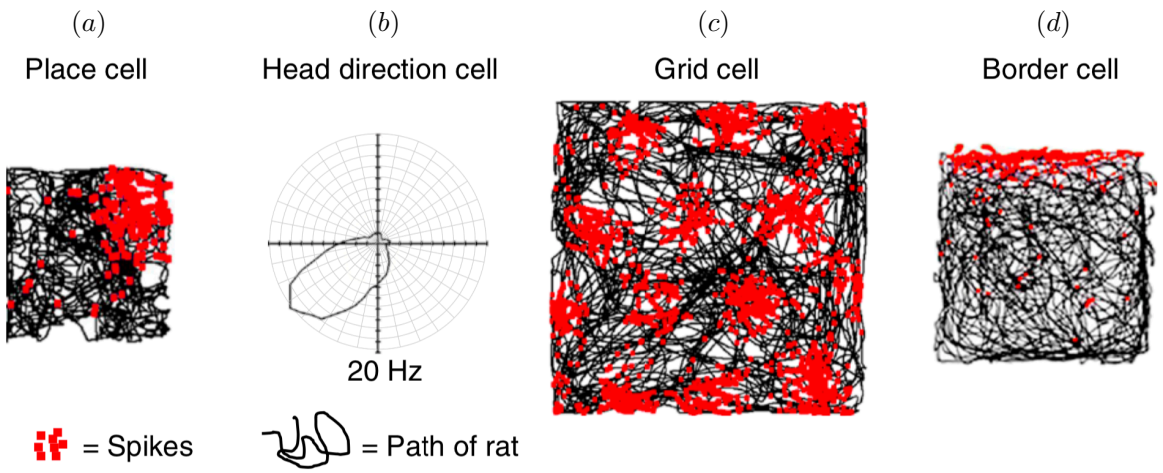

Figure 13 – Spatial cognition requires having a cognitive map, *i.e.*, an internal representation of the outside world [232]. Single neuron recordings have led to the discovery of various neuronal cell assemblies, each with its own specific function, in the hippocampus and the enthorhinal cortex. The different biological features of each serve different specific functions, and have been found to provide the neural basis of spatial navigation. The activity of a place cell (a), a head direction cell (b), a grid cell (c) and a border cell (d) in environments of different sizes are illustrated in the spike plots from left to right. Figure taken from [150].

4

OPTIMAL CAPACITY-RESOLUTION TRADE-OFF IN MEMORIES OF MULTIPLE CONTINUOUS ATTRACTORS

After Chapter 3 it should be clear that recurrent neural networks are powerful tools to explain how attractors may emerge from noisy, high-dimensional dynamics. We study in this Chapter how to learn the $\sim N^2$ pairwise interactions in a recurrent neural network with N neurons to embed L manifolds of dimension $D \ll N$, showing that the capacity, *i.e.*, the maximal ratio L/N, decreases as $|\log \epsilon|^{-D}$, where ϵ is the error on the position encoded by the neural activity along each manifold, or in other words, that recurrent neural networks are flexible memory devices capable of storing a large number of manifolds at high spatial resolution. These results rely on a combination of analytical tools from statistical physics of disordered systems and random matrix theory, extending Gardner's classical theory of learning to the case of patterns with strong spatial correlations.

All the findings presented in this Chapter have been published in [28].

4.1 INTRODUCTION

At this stage it should be evident the objective of the thesis, namely how to store multiple continuous attractors in a recurrent neural network (RNN). Furthermore, after the reading of Chapter 3, the reasons for this kind of research should be understood from the point of view of computational neuroscience, see Section 3.5, as well as the problems with the current theory on the subject should be clear, see Section 3.7. Recall in fact that in the last twenty years no theoretical progress has been made on this fundamental problem despite the presence of several critical aspects of popular models such as:

— due to interferences between the different manifolds stored in the same network, the attractors are not continuous any longer, and are effectively discrete. How to practically learn smooth attractors and ensure a prescribed level of spatial error ϵ, where by spatial error we indicate the typical distance on the maps between two nearby positions where the activity bump can get stuck (actually memorized map positions), is an open issue.

— Moreover, theoretical studies [27, 166, 168] show that the critical capacity (maximal number of maps that can be memorized in a RNN) corresponding to the standard Hebbian-like prescription (3.15), an ad hoc learning rule for the pairwise interactions between neurons summing up the different contributions corresponding to

each single attractor taken independently of the others, see Section 3.6, is very low. In addition, these studies do not convey any information about the spatial error with which the different continuous attractors are encoded in the sense that it is not clear in the different maps which and how many positions are really stored.

Is it possible to define not a priori learning rules for the network connectivity matrix in order to have control over the spatial error with which the different maps are actually stored? Also, once this is done, is it possible to understand the optimal trade-off between capacity and spatial error? Answering these questions is of fundamental importance also because one might be interested in very different scenarios, *i.e.*, to store in a RNN a few maps but with very small spatial error or maybe to store many maps in a coarser way. In the following we present a theory in order to bypass these problems.

4.2 THE MODEL

To answer the questions mentioned in Section 4.1 and Section 3.7 we introduce the following model with which we want to study the optimal storage of multiple quasi-continuous manifolds with prescribed spatial error ϵ in a RNN with N binary neurons ($\sigma_i = 0, 1$) and real-valued, oriented connections W_{ij}, same scheme as Figs. 3(a) and (c).

Inspired by the phenomenology of place cells and place fields, see Section 3.3, we define in this context a map as a D-dimensional cube with unitary volume and periodic boundary conditions [1] in which each neuron $i = 1, \dots, N$ of the RNN has a randomly located input (place) field (PF), *i.e.*, a D-dimensional sphere of volume $\phi_0 < 1$, centered in position $\mathbf{r}_i$, see one of the two maps in Fig. 14(b).

We want than to store in the network $L = \alpha N$ (as usual we define the parameter $\alpha = \mathcal{O}(1)$ as the load, see Section 3.6) maps that differ through random rearrangements of the PF center positions, $\mathbf{r}_i^\ell$, $\ell = 1, \dots, L$ (see Fig. 14(b)). Imposing random remapping between the different maps is essential to make them orthogonal as we already discussed in Section 3.5, where by orthogonal maps we mean that the activity configurations of the neurons coding for the positions of the different environments are orthogonal to each other, and thus avoid as much interference between them as possible.

Now instead of trying to memorize the different maps with a priori learning rules as Hebbian-like ones (3.15), see Section 3.6, where by a priori we mean here that the learning of the maps through the choice of the connectivity matrix is done without having an explicit control on the positions actually stored in the different environments (what are the fixed points of the network dynamics in the different maps?), we approximate each

1. The more realistic case of environments in which we remove periodic boundary conditions is discussed in Section 6.2.

map ℓ through a collection of p random positions $\hat{\mathbf{r}}^{\ell,\mu}$, $\mu = 1,\dots,p$, and we will define in the next Section 4.3 how to store in the connectivity matrix of the RNN explicitly these positions. The spatial error ϵ, which is an index to the accuracy of how each map is stored, is defined as the typical distance between nearest neighbor positions on an environment, see Fig. 14(a), and in this setting where we have approximated the maps with p random positions it scales as $\epsilon \simeq p^{-\frac{1}{D}}$ [55]. In this way we can have control on the accuracy of the coding because increasing p correspond to decrease the spatial error ϵ, and in particular we will be able to store quasi-continuous attractors in the large p limit.

At this stage for every position in every map, $\hat{\mathbf{r}}^{\ell,\mu}$, we extract an N-dimensional pattern of activity to be stored in the network: the neuron i is active ($\sigma_i^{\ell,\mu} = 1$) if the distance between this position and the center of the PF associated with the neuron i in map ℓ, $|\hat{\mathbf{r}}^{\ell,\mu} - \mathbf{r}_i^{\ell}|$, is smaller than the PF radius r_c, and silent ($\sigma_i^{\ell,\mu} = 0$) otherwise, see Fig. 14(b). Also this way of building the patterns to memorize is clearly inspired by the mechanism of place cells and place fields discussed in Section 3.3. We end up with a data-set of $p \times L$, N-dimensional, binary patterns $\{\sigma_i^{\ell,\mu}\}$, where i is the neuron index, ℓ the environment index and μ the index of the position to be stored in map ℓ.

It is essential to note that each pattern built in this way corresponds to a bump state for the network in the sense that precisely by construction, given a position in a certain environment, the related pattern has active neurons corresponding to PFs in that map that are close to the position itself (those that contain it), while the other neurons whose PFs are far away from the position (those that do not contain it), are silent.

Then, by storing a map in a RNN we mean that the patterns associated to the sampled positions on that map must be fixed points of the network dynamics and we will see in the next Section 4.3 how to learn the connections between neurons for this to be possible.

4.3 LEARNING THE OPTIMAL COUPLINGS

How do we store the different maps, that is the patterns generated in the previous Section 4.2, in the RNN?

Before discussing this we need to define a dynamics for the network. In this specific case we consider the standard sequential zero temperature Monte Carlo dynamics [25, 79, 91, 239]:

$$\sigma_i(t+1) = \Theta\Big(\sum_{j\neq i} W_{ij}\sigma_j(t)\Big)\,, \tag{4.1}$$

where Θ is the Heaviside step-function and at every time step $t+1$ we choose uniformly at random the neuron i to update among the N ones, exactly the same kind of dynamics encountered in the previous Chapter 3 with 0 or 1 spins and without considering this time

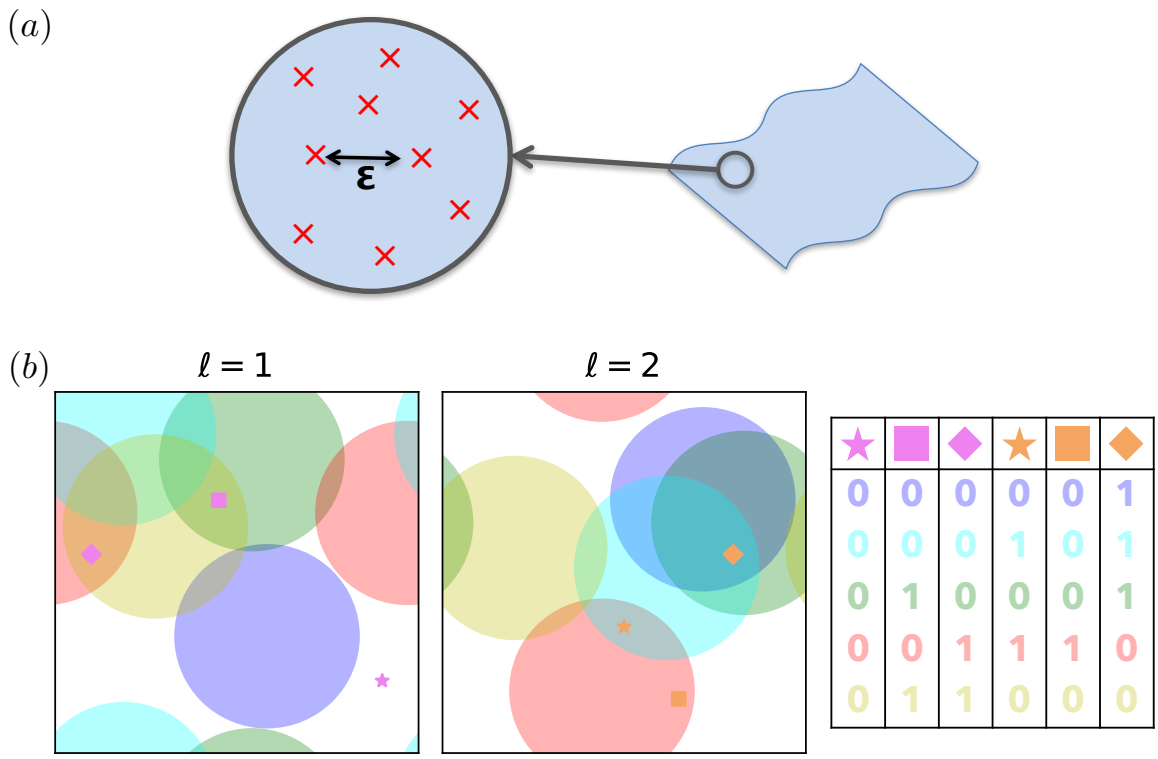

Figure 14 – (a) Sketch of a manifold to store in the RNN, see Fig. 3(c). We require to memorize p points (left, red crosses) on the manifold, whose separation defines the spatial error ϵ. (b) Place fields (PF) of $N = 5$ neurons in two maps. Each color identifies one neuron; the corresponding PF define the regions (with periodic boundary conditions) in the maps in which the neuron is active. The table lists, for each map, $p = 3$ activity patterns corresponding to the marked points.

the local external threshold fields, in this case we don't need any threshold because as we will see the connectivity matrix will have both positive and negative interactions and this is enough to keep the network activity fixed. Moreover it is important to recall, as we saw in Section 3.2, that such a noiseless dynamics converges from any initial configuration of the neurons to its nearest fixed point, *i.e.*, a configuration of the network neurons activity such that the dynamics remain stuck.

At this point we want to choose minimal conditions for the connectivity matrix $\mathbf{W}$ such that the patterns generated in the previous Section 4.2 are fixed points of the neural dynamics (4.1), that is they are stored in the RNN. This property is equivalent to impose for every pattern in the data-set

$$\Delta_i^{\ell,\mu} \equiv (2\sigma_i^{\ell,\mu} - 1) \sum_{j \neq i} W_{ij} \sigma_j^{\ell,\mu} > 0 \,, \quad \forall i, \ell, \mu \,, \tag{4.2}$$

where $\Delta_i^{\ell,\mu}$ is defined as the stability of the neuron i for the pattern associated with the map ℓ in position μ.

If we now look carefully at the definition of the RNN it is easy to realize that since we don't have self-connections ($W_{ii} = 0, \forall i$), if we focus on a single neuron i and wonder what is the connectivity matrix row $\mathbf{W}_i$, an $(N-1)$-dimensional vector, that satisfies (4.2), we immediately realize that this problem is a perceptron problem [206, 207] with neuron i as output and the other $N-1$ neurons as input, see Fig. 15(a). In particular we have N independent perceptron problems to solve, one for each output neuron i, so that to consider all the rows of the connectivity matrix that after learning will be non-symmetric a priori. The important thing is that even though we have N distinct problems to solve, learning is done on the same data-set with the only difference being to change the output and input neurons in the different perceptrons, which after the learning phase will not be uncorrelated and will produce a well defined dynamics for the whole network.

In order for the patterns to be fixed points of the dynamics according to (4.2) we have to solve the N classification problems associated to the different perceptrons. In fact, if we look at the $p \times L$ binary patterns in the $(N-1)$-dimensional input space of one of the perceptrons, we see that the patterns are divided into two classes: the class that includes the patterns in which the output neuron is active and the class in which it is silent. Choosing the row of the connectivity matrix $\mathbf{W}_i$ that satisfies (4.2) for a fixed output neuron i is equivalent to find the hyperplane that linearly separates the two classes. This problem, if solvable, admits infinite solutions (all solutions that can be reached by the perceptron algorithm [71, 109]), and among all these solutions we are interested in finding the hyperplane that maximizes the distance between the two classes, this problem therefore corresponds to a convex optimization problem with an unique solution [43], where this optimal distance is called maximal stability [21, 132] or in machine learning terminology hard margin, because this particular solution to the perceptron problem is equivalent to a support vector machine (SVM) with a linear kernel and hard margin [35, 62, 211], see Fig. 15(b).

This choice of $\mathbf{W}_i$ ensure the biggest basins of attraction in the pattern space, *i.e.*, robustness against thermal noise [16, 132, 133], because the optimal hyperplane ensures that every perceptron is as robust as possible in the classification of the two classes (smallest generalization error).

To put in formulas these concepts, we are interested in finding the connectivity matrix $\mathbf{W}$ that maximizes the stabilities of the patterns to be stored

$$\kappa = \max_{\mathbf{W}} \min_{\{i=1\dots N, \ell=1\dots L, \mu=1\dots p\}} \left[(2\sigma_i^{\ell,\mu} - 1) \sum_{j(\neq i)} W_{ij}\, \sigma_j^{\ell,\mu} \right], \tag{4.3}$$

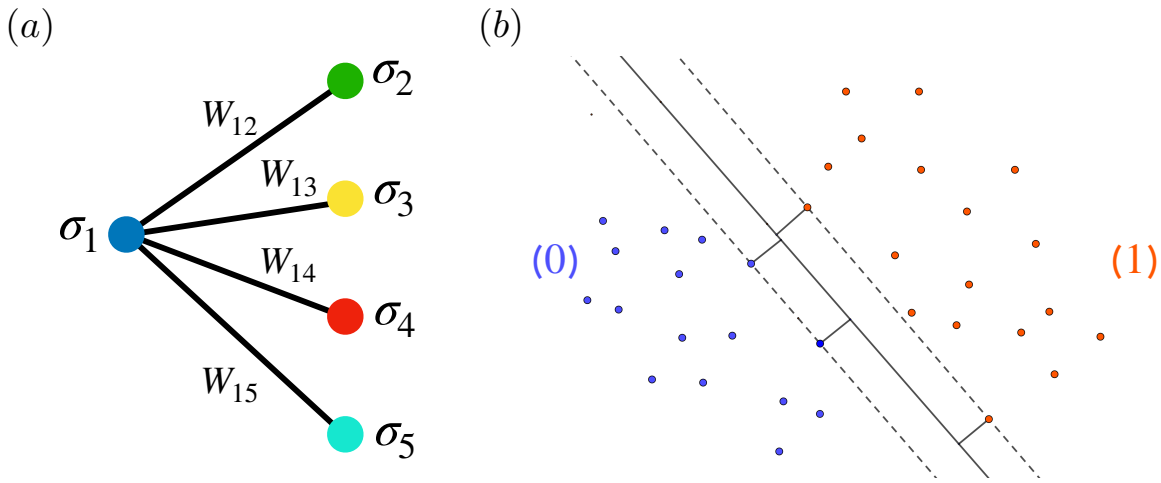

Figure 15 – (a) One of the N perceptrons that make up the RNN with $N = 5$ neurons and where the output neuron is the one correspondent to the index $i = 1$. (b) Example of a learning problem for a perceptron where the aim is to find the hyperplane separating the two classes, class (0) and class (1), while maximizing its distance from them (in the maximum margin setting). It is clear from the figure that the problem of linearly separating the two classes with an hyperplane admits infinite solutions when the classes are linearly separable.

where all the rows of $\mathbf{W}$ are normalized to 1, *i.e.*, $\sum_{j(\neq i)} W_{ij}^2 = 1 \; \forall \; i$, for details on the SVM algorithm see Appendix A.1.

It is interesting to note that usually SVMs are used in supervised learning (fit of input-output relation from examples typically in high dimensions) problems where in general the output is simply a label associated to the input [58] while in this specific setting we are using the same technique in an unsupervised (find statistical features of data for clustering, dimensional reduction and so on) way in the sense that here the output neurons are part of the activity patterns to be stored that are composed of input and output together.

4.4 RESULTS OF NUMERICAL SIMULATIONS

Now that we have well defined the model in Section 4.2 and the dynamics of the network together with the learning rule in Section 4.3 we can start studying the results of simple numerical simulations and see if qualitatively the typical characteristics of RNNs that store continuous attractors are reproduced, see Sections 3.4 and 3.6.

4.4.1 *Couplings obtained by SVM*

First, we report some qualitative features of the couplings obtained by SVMs after the training process (Section 4.3) on the specific data discussed in Section 4.2.

— As shown in Fig. 16(a) and (c) the couplings W_{ij} are correlated with the distances $d_{ij}^{\ell} = |\mathbf{r}_i^{\ell} - \mathbf{r}_j^{\ell}|$ between the PF centers of the neurons i and j in the different maps ℓ. Note that the dependence on distance is less marked in the single environments as the number L of maps increases, due to the interferences between the maps encoded in the same connectivity matrix, see Section 3.6.

— In order to sustain a bump state with average activity ϕ_0 (obviously for construction the average activity of the patterns to memorize is equal to the area of the PFs), couplings are excitatory at short distances, up to roughly the radius r_c of the PF, and inhibitory at longer ones. The sign of the couplings can be intuitively understood. Two neurons at short distances have largely overlapping PFs: their activities are likely to be equal, and having a large coupling helps increasing the stability, see (4.2). If the distance is bigger than r_c, the activities are likely to be different, hence inhibitory (negative) couplings would increase the stability, see Fig. 16(a) and (c).

— Moreover, in Fig. 16(b) and (d) we show the histograms of the couplings and it seems that the amplitudes decay with N. In agreement with [165] we expect the average values and standard deviations to scale, respectively, as $1/N$ and $1/\sqrt{N}$, see Sections 4.6 and 6.7.

4.4.2 *Finite temperature dynamics* ($T > 0$)

Second, once the coupling matrix $\{W_{ij}\}$ has been learned, we may perform Monte Carlo simulations to investigate the behavior of the network. Instead of considering the natural dynamics of the network without noise (4.1), we implement a noisy dynamical scheme, the same introduced in Section 3.4, but now without the threshold terms (see Section 4.3), where neuronal states are updated stochastically according to the probabilities [25, 79, 91, 239]

$$\text{Prob}\left(\sigma_i(t+1)|\{\sigma_j(t)\}\right) = \frac{1}{1+\exp\left[-\frac{1}{T}\left(2\sigma_i(t+1)-1\right)\sum_{j(\neq i)} W_{ij}\,\sigma_j(t)\right]}\,, \tag{4.4}$$

with T as a temperature parameter to be set such that a bump of activity may form and sustain itself. This happen when T is comparable to, or smaller than the stability κ of the network. All that in order to show the diffusion of the activity bump within a map and

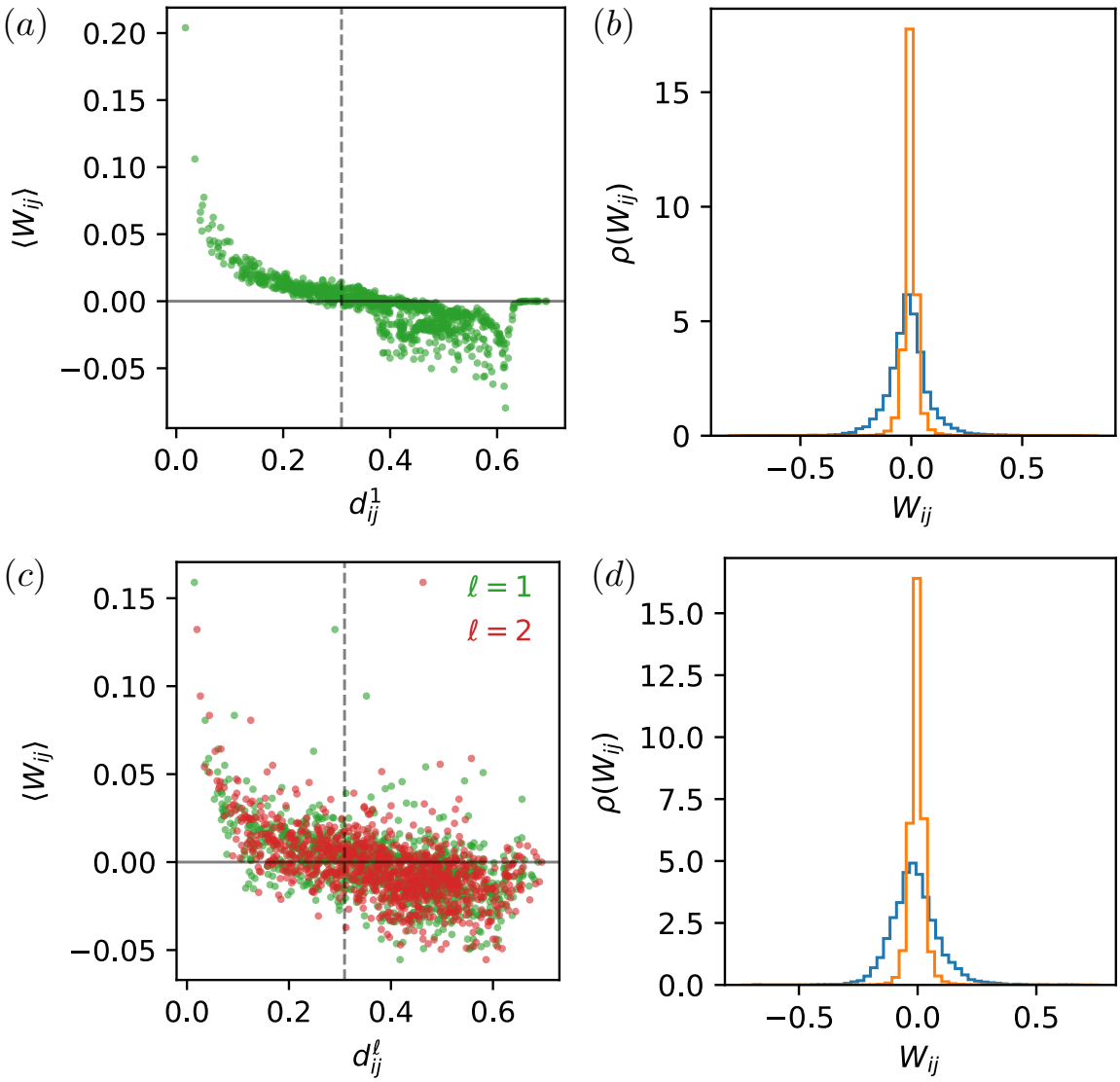

Figure 16 – Couplings obtained after training with SVM for L = 1 (a) and (b) and L = 2 (c) and (d) maps. Left: dependence of the average coupling with the distance between the corresponding neurons; the vertical line locates the radius r_c of the place fields. Averages were computed over 500 samples of the p positions per map at fixed PF centers; N = 1000 neurons. Right: histograms of the couplings, for sizes N = 100 (blue) and N = 1000 (orange) Parameters: D = 2, $\phi_0 = .3$ and p × L = N.

the transitions between maps [167, 168, 208] as already seen previously in Sections 3.1 and 3.6. We illustrate the dynamical properties of the model with two examples [2]:

2. The examples mentioned below can be viewed at: https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.124.048302. In the attached videos the red crosses represent the stored positions of the

— first, we consider a network with $N = 1000$ neurons, in which we store one map ($L = 1$) in dimension $D = 2$ and with average activity $\phi_0 = .3$. We consider then the case in which the learning is done on a small number of points, $p = 30$, resulting in a large value of the stability (as we will see in fact the stability of the network decreases as the constraints that the connectivity matrix must satisfy increase, so L and p), $\kappa = 1.7$. And then the case in which $p = 300$ is higher, and the stability is smaller: $\kappa = .6$. Our noise parameter T is set to .8 to allow the bump to form in both cases. In the large κ case, the bump, which is possible to visualize by looking at which are the PFs that correspond to the active neurons at a given time, gets stuck very quickly in one of the p training positions, depending on the initial configuration, see videos `LargeKappaL1.mp4` and `LargeKappaL1Bis.mp4`. In the small κ case, the bump diffuses on the map, see video `SmallKappaL1.mp4`. For larger p, the bump can easily travel through the environment, with a large diffusion coefficient; in contrast, in the small p case, the stability landscape is very rough (only the stored positions have a very high positive stability value, the others can have a priori negative stabilities, which is why the dynamics remain locked, see Fig. 34(a)) and the bump is stucked close to the stored positions.

— In the second example we consider the case of $L = 2$ maps and $p = 150$ points. The other parameters have the same values as in the first example, *i.e.*, the stability is fixed to $\kappa = .5$. In the video `SmallKappaL2.mp4` we see that, as κ is small, the bump diffuses in one maps and sporadically jumps to the other map, as it should be due to the fluctuations induced by the temperature and a reasonable number of stored positions per map. In fact, since the two maps are orthogonal by construction, if we focus on which are the PFs that correspond to the active neurons at a given time, we observe that in one of the two maps these are clustered and therefore form a bump of activity (the position in one of the maps we are recalling at that time), while in the other map we have that these PFs are basically random, as it should be given that the bump must be located in one of the maps only [3]. Moreover, since the dynamics has a noise level, it is possible to observe the transitions of the bump from one map to another simply by noticing that at a given time the network activity has PFs clustered in one map and random in the other, and when a transition occurs the opposite happens.

different environments in the connectivity matrix of the neural network and the blue circles represent the centres of the PFs whose corresponding neurons are active at a fixed time (the PFs whose neurons are silent at a certain time are not shown). Also note the presence of periodic boundary conditions in the maps that can sometimes make the bump not easy to visualize.

3. The case of having activity bumps located at the same time in multiple maps is possible if they are correlated. Since in our model the PFs are random in the different maps (global remapping) this situation is unlikely to be possible by construction.

Moreover, the fact that the stored attractors can be continuous depends primarily on the number of place cells encoding the map and the size of the PFs, as these parameters define the possible number of distinct patterns and therefore the positions that can be stored in the map itself. Once the map is defined, the dynamical properties of the activity bump are basically fixed by the number of positions memorized per map and if the difference between the points is a fraction of the bump's width, the discretization in practice is no longer perceived. To test if our model is able to do the tracking of moving inputs we made a simulation where, starting from a RNN of $N = 1000$ neurons, we learn $L = 2$ maps in $D = 1$, where the size of the PFs is fixed at $\phi_0 = .2$ and where are stored $p = 250$ positions per map. Then we go to study the dynamics of the network with a temperature $T = .5$ and also add an external input that acts on individual neurons in the following way: $I_i(t) = \exp(-\frac{|x_i^\ell - x(t)^\ell|}{\phi_0})$, where $x(t)^\ell$ represents the center of the input at a given time in one of the two maps. From Fig. 17 it is clear that the center of the activity bump [4] (blue points) follows the input (orange curve) very well, even if the input is abruptly moved from one map to another. From this example we can see that the continuous tracking of moving inputs is achieved in our model.

4.4.3 *Zero temperature dynamics ($T = 0$) and spatial error ϵ*

Now instead, if we are interested in calculating the spatial error ϵ with which the maps have been stored, we consider the natural dynamics of the network (4.1). Starting from an initial activity configuration, we track the system dynamics for at most N^2 Monte Carlo steps (N sweeps), and retain the visited configuration with the minimum number of violated constraints, *i.e.*, with the highest number of non-negative stabilities [5]

$$\Delta_i = (2\sigma_i - 1) \sum_{j(\neq i)} W_{ij}\, \sigma_j \geqslant 0 \ . \tag{4.5}$$

We generate L environments and p points in each of them, learn the coupling matrix corresponding to these $p \times L$ patterns. We then pick at random a position in one of the learned maps, and use that position to construct the initial activity configuration of the dynamics. After the dynamics described above is done we keep the final configuration and use it to decode the final position on that map, as the center of mass of the PFs (on the map) of the active neurons in the final configuration. The distance between this

4. See the next paragraph for details on how to decode the position of the center of the bump from network activity.

5. The choice of N sweeps as a maximal simulation time is empirical: we do not find that significantly better results are obtained by increasing this bound. Actually, the dynamics often ends up in a fixed point with stabilities $\Delta_i \geqslant 0$ for all neurons i in much less sweeps. Nevertheless, it is necessary to set a maximum number of sweeps because the learned connectivity matrix is not a priori symmetrical and therefore the convergence of dynamics is not ensured.

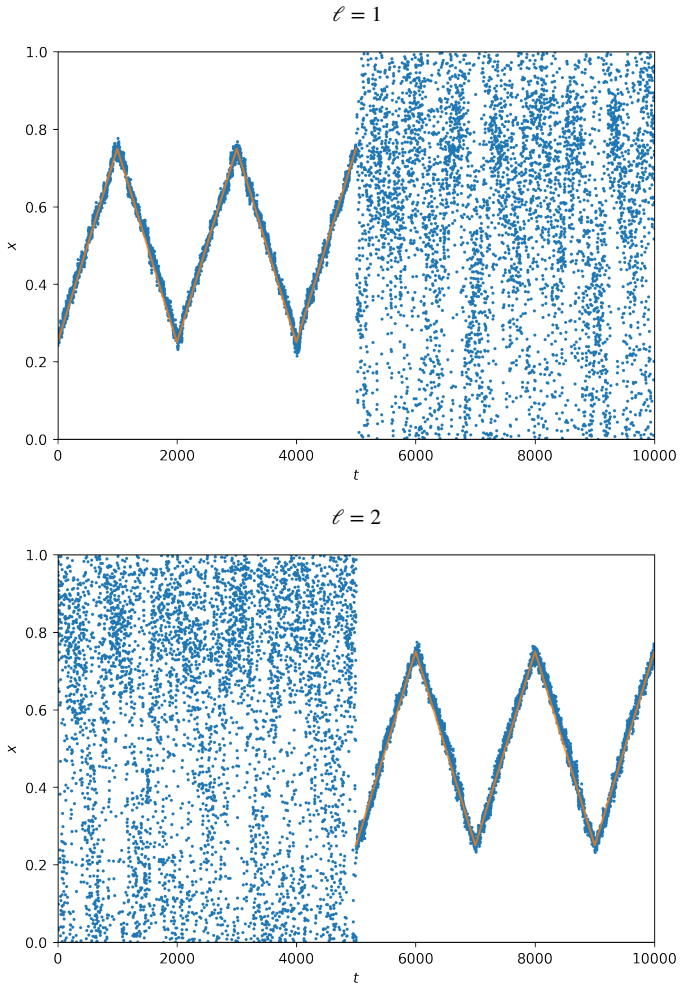

Figure 17 – Tracking a moving input. The $x(t)$ represent the centers of the bump decoded by the activity of the place cells in the two environments (taking into account the periodic boundary conditions). Every temporal step t in the simulation corresponds to a Monte Carlo sweep (that is to the attempt to update N neurons chosen at random following the dynamics (4.4) that includes the external input) and consequently the plot reported corresponds to a single dynamics, and not to the average on different samples of possible trajectories.

estimated position and the initial one (taking care of the periodic boundary conditions), after averaging over many initial positions (100 in the figures showed), defines the spatial error ϵ with which the maps have been stored (the spatial error is roughly the same in all the environments since p is the same in all maps and the positions are random).

Fig. 18(a)[6] shows that ϵ becomes quite large as L increases if we use Hebbian-like learning rules (3.15) for the connectivity matrix. However, with the maximal-stability learning rule (Section 4.3), the spatial error ϵ can be tuned at will by varying p. For a fixed p, ϵ remains remarkably stable as the load increases until its critical value is reached. This is in sharp contradistinction with the Hebb rule case (3.15), for which ϵ quickly increases with the number of maps. The p patterns form a discrete approximation of the map, with average spatial error scaling as $\epsilon = p^{-1/D}$, *i.e.*, as the typical distance between neighboring points [55], see Fig. 18(b).

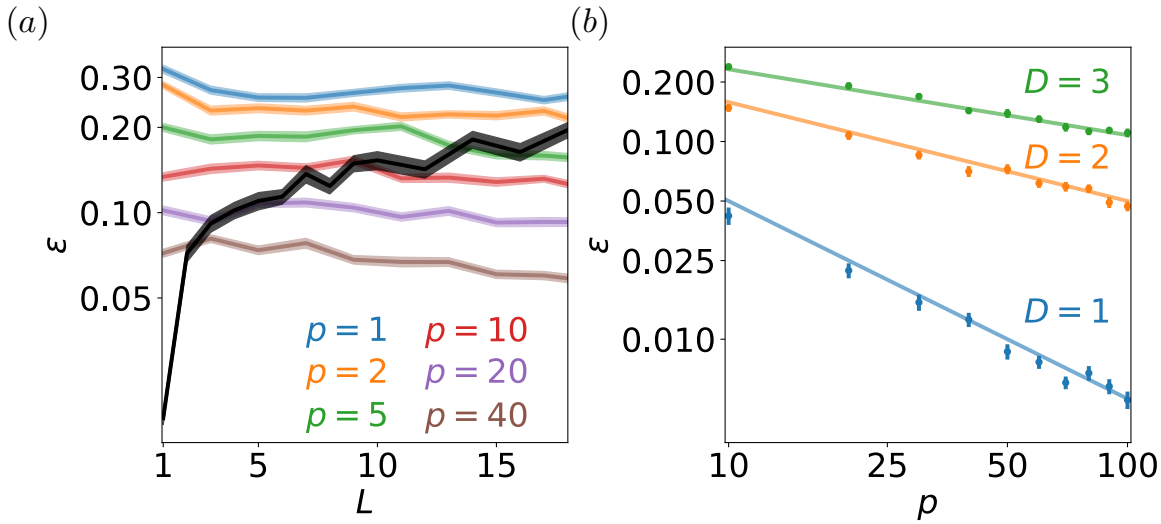

Figure 18 – (a) Spatial error ϵ vs number L of two-dimensional maps in a network of N = 1000 neurons. Black: rule (3.15), with $w(d) = e^{-d/0.01} + w_0$, where $w_0 < 0$ enforces a fraction $\phi_0 = .3$ of active cells. Colors: SVM results for different numbers p of prescribed positions. Line widths show the error bars. (b) spatial error ϵ vs number p of positions in a network of N = 1000 neurons storing L = 5 maps, in dimensions D = 1, 2, 3. Lines show the expected scalings $\epsilon \sim p^{-1/D}$ in log-log scale.

It is therefore clear now that by increasing p we can control the accuracy of the coding by decreasing the spatial error ϵ. It is obvious, however, that requiring small ϵ for the maps has a cost in terms of how many maps can be stored because there are more

6. In this figure we show only the case of an exponential kernel (where the parameters have been chosen such that to have zero spatial error in case of a single map) for the Hebbian rule, but it can be seen from Fig. 19 that for example a Gaussian kernel has similar performance as the exponential one. In general it is true that the specific value of the spatial error depends on the choice of the kernel (and its parameters), however the important point to note is that regardless of the kernel used, a Hebbian type rule has a spatial error that increases with the number of maps while the key to our approach is to have a method that can maintain the same level of spatial error even when storing multiple maps (obviously until the critical capacity is reached).

constraints that must be taken into account by the connectivity matrix. The goal now is therefore to understand this trade-off between capacity and resolution.

4.4.4 *Comparison with Hebb rule*

Before studying how the optimal stability κ defined in Eq. (4.3) and obtained from SVMs (Section 4.3) behaves as a function of $\alpha = \frac{L}{N}$ and p, as a sanity check, we show that it is always much higher than the maximal stability achievable with Hebbian-like rules (3.15), even after the optimization over the interaction kernel w. In fact, this should be true by construction, see Section 4.3.

In order to show that we consider the cases of an exponential kernel,

$$w(d) = a\, e^{-d/b} - 1 \,, \tag{4.6}$$

and of a Gaussian kernel,

$$w(d) = a\, e^{-d^2/b} - 1 \,. \tag{4.7}$$

We then optimize over a and b; the value of the negative offset at large distance is arbitrary, since couplings are normalized row by row. Results for a typical sample are shown in Fig. 19. The stability for the best kernel w is always much lower (and negative in the examples considered here) than the optimal stability κ found with SVM.

4.4.5 *Capacity-Resolution trade-off*

The optimal stability κ (4.3) is shown in Fig. 20(a) as a function of the load α and of the number p of prescribed fixed points. As expected, $\kappa(\alpha, p)$ is a decreasing function of α and p: increasing the number of maps or enforcing finer spatial error reduces the stability.

The value of the load α at which $\kappa(\alpha, p)$ vanishes defines the critical capacity $\alpha_c(p)$, that is, the maximal load sustainable by the network as a function of the required spatial error. Fig. 20(a,inset) shows that $\alpha_c(p)$ decreases proportionally to $1/p$ at low p, and then much more slowly as p grows, see Appendix A.2 for details on how we estimated numerically the critical capacity.

For small p, all $L \times p$ patterns are roughly independent, and we have $\alpha_c(p) \simeq \frac{\alpha_c(1)}{p}$, where $\alpha_c(1)$ is the capacity of the perceptron with independent, biased patterns having a fraction ϕ_0 of active neurons [87]. As p gets large, substantial redundancies between the p patterns within a map appear, as nearby positions define similar patterns, see Fig. 14, and the capacity is expected to decrease less quickly with p. The cross-over takes place at $p_{c.o.} \sim 1/\phi_0$, see Fig. 20(b).

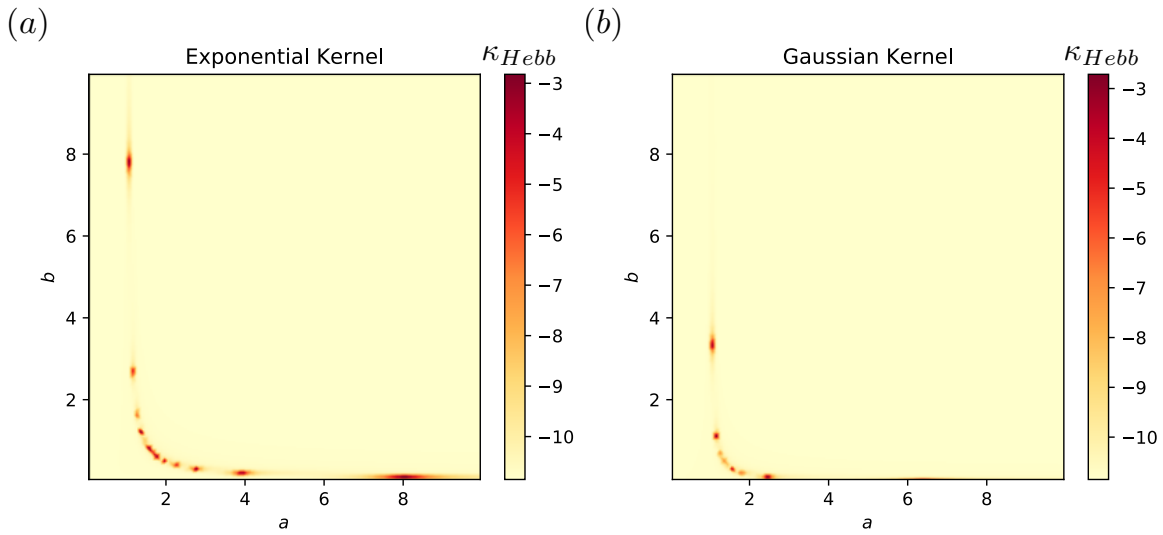

Figure 19 – Stabilities obtained with the Hebb rule with exponential (a) and Gaussian (b) kernels on a given representative sample. The kernel parameters a and b vary from 0 to 10 with a step of .01. Parameter values: $N = 1000$, $D = 2$, $\phi_0 = .3$, $\alpha = .1$ and $p = 5$. The optimal value of the stability on that sample obtained by SVM is $\kappa \simeq .55$.

The nontrivial behavior of $\alpha_c(p)$ when $p \gg p_{c.o.}$ will be characterized in the analytical study in Section 4.5 and Section 4.6.

4.5 GARDNER'S THEORY FOR RNN STORING SPATIALLY CORRELATED PATTERNS

As we have just discussed in Section 4.4, increasing p corresponds to decrease the spatial error ϵ, but we can easily understand that increasing p, and therefore the number of constraints that must be satisfied by the connectivity matrix has a cost in terms of how many maps the RNN can store. In order to try to understand this trade-off between capacity and spatial error analytically we used, as a first attempt, Gardner's approach developed in the 1980s [87, 88].

The idea is to look at the space of all possible connectivity matrices that have our patterns, see Section 4.2, as fixed points of the network dynamics (4.1), *i.e.*, all possible hyperplanes that linearly separate the two classes for all N perceptrons, see Section 4.3. In this feasible set of solutions there will be as well the one that corresponds to the connectivity matrix with maximal stability κ, namely the solution with the maximum margin for every perceptron, see Section 4.3. When we increase $\alpha = \frac{L}{N}$ at fixed p the logarithm of the volume of this set of solutions decrease until it shrinks to a single point, as we will explain in detail below. When this happens we found the critical capacity because adding a new constraint would make the problem impossible to solve.

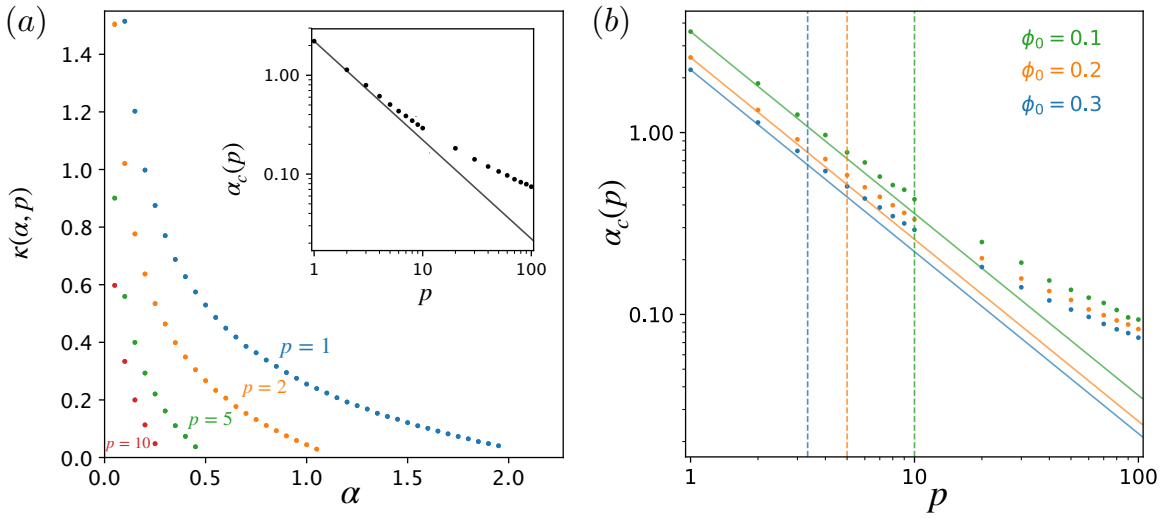

Figure 20 – (a) Optimal stability κ as a function of the load α and the number p of positions obtained from SVM. Parameter values: $D = 2$, $\phi_0 = 0.3$, $N = 1000$ for SVM. Inset: $\alpha_c(p)$ decreases proportionally to $1/p$ (straight line) at low p, and much more slowly for large p. Dots indicate results from SVM ($N = 5000$), averaged over 50 samples; the dot size indicates the maximal error bar. (b) Scaling cross-over of $\alpha_c(p)$ vs p for different values of ϕ_0. The vertical lines correspond to the values of $p_{c.o.} \sim \frac{1}{\phi_0}$. We use for this results $D = 2$, $N = 5000$, and we have averaged over 50 different realization of the environments and different realizations of the p positions. Parameter values: $N = 5000$, $p = 100$, Samples$= 25$.

So here we are going to extend the Gardner theory for the capacity of the perceptron (SVM with linear kernel and hard margin) to the case of continuous attractors.

We shall consider the now usual RNN consisting of N neurons, with $p \times L$ binary (0 or 1) patterns $\{\sigma_i^{\ell\mu}\}$ constructed by drawing randomly p positions in each of the L environments, see Section 4.2, so that the resulting patterns are spatially correlated and to be stored as fixed points of the network dynamics

$$\sigma_i(t+1) = \Theta\Big(\sum_{j\neq i} W_{ij}\sigma_j(t)\Big)\,. \tag{4.8}$$

For the storage of the patterns a strong characterization of fixed points is provided by

$$\Delta_i^{\ell,\mu} \equiv (2\sigma_i^{\ell,\mu} - 1)\sum_{j\neq i} W_{ij}\sigma_j^{\ell,\mu} > \kappa\,, \quad \forall\, i,\, \ell,\, \mu \tag{4.9}$$

which ensures a finite radius of attraction for $\kappa > 0$ [87, 88, 125, 132–134]. Then we can choose to normalize $\mathbf{W}$'s by enforcing the "spherical" constraint

$$\sum_{j\neq i} W_{ij}^2 = 1\,, \quad \forall\, i\,, \tag{4.10}$$

for each row of the matrix of couplings.

The volume in the space of couplings $\{W_{ij}\}$ that corresponds to admissible solutions of the storage problem, *i.e.*, the partition function, can be written, up to a normalization factor, in the form [109]

$$Z = \int_{-\infty}^{\infty} \prod_{i \neq j}^{N} dW_{ij} \prod_{i} \delta\Big(\sum_{j(\neq i)} W_{ij}^2 - 1 \Big) \prod_{i,\ell,\mu} \Theta\Big((2\sigma_i^{\ell\mu} - 1) \sum_{j(\neq i)} W_{ij}\, \sigma_j^{\ell\mu} - \kappa \Big) \tag{4.11}$$

and is equal to the product of the N single-site volumes Z_i, with $i = 1, \ldots, N$. So we may focus for example on the volume associated with $i = 1$:

$$Z_1 = \int_{-\infty}^{\infty} \prod_{j=2}^{N} dW_j\, \delta\Big(\sum_{j \geqslant 2} W_j^2 - 1 \Big) \prod_{\ell,\mu} \Theta\Big((2\sigma_1^{\ell\mu} - 1) \sum_{j \geqslant 2} W_j\, \sigma_j^{\ell\mu} - \kappa \Big), \tag{4.12}$$

where $W_j \equiv W_{1j}$.

Using the replica method [51, 156, 157, 181, 242], we compute the average of $\log Z_1$ over the patterns as

$$\langle \log Z_i \rangle = \lim_{n \to 0} \frac{\langle Z_i^n \rangle - 1}{n}, \tag{4.13}$$

where $\langle . \rangle$ represent the quenched average over the data distribution, so we get

$$\langle Z_1^n \rangle = \int_{-\infty}^{\infty} \prod_{j,a} dW_{ja} \prod_{a} \delta\Big(\sum_{j \geqslant 2} W_{ja}^2 - 1 \Big) \Big\langle \prod_{\ell,\mu,a} \Theta\Big((2\sigma_1^{\ell\mu} - 1) \sum_{j \geqslant 2} W_{ja}\, \sigma_j^{\ell\mu} - \kappa \Big) \Big\rangle, \tag{4.14}$$

where $a = 1, \ldots, n$ is the replica index.

Let us now introduce the integral representation of the Heaviside functions and exploit the statistical independence of the different maps so that we can write

$$\langle Z_1^n \rangle = \int_{-\infty}^{\infty} \prod_{j,a} dW_{ja} \prod_{a} \delta\Big(\sum_{j} W_{ja}^2 - 1 \Big)\, \chi(\mathbf{W})^{\alpha N} \tag{4.15}$$

where $L = \alpha N$ and with $\chi(\mathbf{W})$ equal to

$$\int \prod_{\mu=1}^{p} d\hat{\mathbf{r}}_\mu \int \prod_{j=1}^{N} d\mathbf{r}_j \int_{\kappa}^{\infty} \prod_{\mu,a} dt_{\mu a} \int_{-\infty}^{\infty} \prod_{\mu,a} \frac{d\hat{t}_{\mu a}}{2\pi} e^{\,i \sum_{\mu,a} \hat{t}_{\mu a} t_{\mu a}} \prod_{j} e^{\,-i \sum_{\mu,a} \hat{t}_{\mu a} (2\sigma_1^{\mu} - 1) W_{ja} \sigma_j^{\mu}}, \tag{4.16}$$

where $\hat{\mathbf{r}}_\mu$ denotes the p prescribed locations in the environment, and $\mathbf{r}_j$ the N PF centers of the neurons in the map (these variables must be integrated on the map, *i.e.*, D-dimensional cube with periodic boundary conditions).

Let $\Phi(\mathbf{r})$ be the indicator function of the PF centered in $\mathbf{o}$: $\Phi = 1$ if $|\mathbf{r}| < r_c$, where r_c is the radius of the PF (with $\int d\mathbf{r}\,\Phi(\mathbf{r}) = \phi_0$), and o otherwise. Let $\Gamma(\mathbf{r}) = \int d\mathbf{r}'\Phi(\mathbf{r}')\,\Phi(\mathbf{r}-\mathbf{r}')$ be the correlation function of Φ and $\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) = 2\Phi(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) - 1$. Given p points $\hat{\mathbf{r}}_\mu$, $\mu = 1, \dots, p$ drawn uniformly at random in space, we define the $p \times p$ Euclidean random matrix[7] [158] with entries

$$\mathbf{\Gamma}_{\mu,\nu}\big(\hat{\mathcal{R}} \equiv \{\hat{\mathbf{r}}_\mu\}\big) = \Gamma\big(\hat{\mathbf{r}}_\mu - \hat{\mathbf{r}}_\nu\big) - \phi_0^2 \,. \tag{4.17}$$

We first carry out explicitly the integrals over the PF centers with indices $j = 2, 3, ..., N$, leaving the integrals over $\mathbf{r}_1$ and all $\hat{\mathbf{r}}_\mu$ in $\chi(\mathbf{W})$ so that we obtain

$$\begin{aligned}\chi(\mathbf{W}) = \int \prod_{\mu=1}^{p} d\hat{\mathbf{r}}_\mu \int d\mathbf{r}_1 \int_\kappa^\infty \prod_{\mu,a} dt_{\mu a} \int_{-\infty}^{\infty} \prod_{\mu,a} \frac{d\hat{t}_{\mu a}}{2\pi} e^{i\sum\limits_{\mu,a} \hat{t}_{\mu a} t_{\mu a}} \\ \times e^{-i\phi_0 \sum\limits_{\mu,a,j} W_{ja}\hat{t}_{\mu a}\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu)} e^{-\frac{1}{2}\sum\limits_{\mu,\nu,a,b,j} W_{ja}W_{jb}\mathbf{\Gamma}_{\mu,\nu}(\hat{\mathcal{R}})\hat{t}_{\mu a}\hat{t}_{\nu b}} .\end{aligned} \tag{4.18}$$

We can now introduce the order parameters

$$m^a = \sum_{j\geqslant 2} W_{ja} \,, \ \forall a \tag{4.19}$$

and

$$q^{ab} = \sum_{j\geqslant 2} W_{ja}W_{jb} \,, \quad a < b \tag{4.20}$$

to be added with the standard trick of inserting the integral representation of Dirac delta functions and fix their values through integration and rewrite

$$\begin{aligned}\langle Z_1^n \rangle = \int_{-\infty}^{\infty} \prod_{ja} dW_{ja} \int \prod_a \frac{d\hat{u}^a}{4\pi i} e^{\sum\limits_a \frac{\hat{u}^a}{2}\left(1 - \sum\limits_j W_{ja}^2\right)} \int \prod_a \frac{d\hat{m}^a dm^a}{2\pi i} e^{\sum\limits_a \hat{m}^a\left(m^a - \sum\limits_j W_{ja}\right)} \\ \times \int \prod_{a<b} \frac{d\hat{q}^{ab} dq^{ab}}{2\pi i} e^{\sum\limits_{a<b} \hat{q}^{ab}\left(q^{ab} - \sum\limits_j W_{ja}W_{jb}\right)} \chi(\mathbf{W})^{\alpha N}\end{aligned} \tag{4.21}$$

where $\hat{u}^a$ are the Lagrange multipliers enforcing the spherical constraints, $\hat{m}^a$ the ones that enforce the definition of m^a and $\hat{q}^{ab}$ the ones that enforce the definition of q^{ab}, and they simply come from the integral representation of the Dirac delta functions.

It is possible then to rewrite $\chi(\mathbf{W})$ as

$$\begin{aligned}\chi(\{m^a, q^{ab}\}) = \int \prod_\mu d\hat{\mathbf{r}}_\mu \int d\mathbf{r}_1 \int_\kappa^\infty \prod_{\mu,a} \frac{dt_{\mu a}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \prod_{\mu,a} \frac{d\hat{t}_{\mu a}}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum\limits_{\mu,\nu,a,b} q^{ab}\mathbf{\Gamma}_{\mu,\nu}(\hat{\mathcal{R}})\hat{t}_{\mu a}\hat{t}_{\nu b}} \\ e^{-i\sum\limits_{\mu,a} m^a\phi_0\hat{t}_{\mu a}\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) + i\sum\limits_{\mu,a}\hat{t}_{\mu a}t_{\mu a}} ,\end{aligned} \tag{4.22}$$

7. See Chapter 5 for details on Euclidean random matrices.

where due to translation invariance, the integral over $\mathbf{r}_1$ is irrelevant, and we can set $\mathbf{r}_1 = \mathbf{0}$.

We can also define from (4.21) the quantity $Y(\{\hat{u}^a, \hat{m}^a, \hat{q}^{ab}\})$ as

$$\int_{-\infty}^{\infty} \prod_{ja} dW_{ja} e^{\sum_a \frac{\hat{u}^a}{2}(1-\sum_j W_{ja}^2)} e^{\sum_a \hat{m}^a (m^a - \sum_j W_{ja})} e^{\sum_{a<b} \hat{q}^{ab}(q^{ab} - \sum_j W_{ja} W_{jb})} \tag{4.23}$$

so that we can write (4.21) as

$$\langle Z_1^n \rangle = \int \prod_a \frac{d\hat{u}^a}{4\pi i} \int \prod_a \frac{d\hat{m}^a dm^a}{2\pi i} \int \prod_{a<b} \frac{d\hat{q}^{ab} dq^{ab}}{2\pi i} Y(\{\hat{u}^a, \hat{m}^a, \hat{q}^{ab}\}) \chi(\{m^a, q^{ab}\})^{\alpha N} . \tag{4.24}$$

It is possible now make the Replica Symmetric (RS) Ansatz [51, 157] (expected to be valid since the domain of suitable couplings is convex, see Section 4.3) on the structure of the order parameters so that

$$\hat{m}^a = \hat{m}, \quad \forall a, \tag{4.25}$$

$$\hat{q}^{ab} = \hat{q}, \quad a < b, \tag{4.26}$$

and the same for the conjugate variables and as well the Lagrange multipliers associated to the normalization of the weights.

We can therefore write within the limit of large N and small n:

$$\frac{1}{nN} \log Y = \frac{\hat{u}}{2} - \frac{q\hat{q}}{2} + \lim_{n\to 0} \frac{1}{n} \log \int_{-\infty}^{\infty} \prod_a dW_a e^{-\frac{\hat{u}}{2}\sum_a W_a^2 - \hat{q} \sum_{a<b} W_a W_b - \hat{m} \sum_a W_a} . \tag{4.27}$$

It is now possible to compute the Gaussian integrals over $\{W_a\}$ and take the small n limit so we get:

$$\frac{1}{nN} \log Y = \frac{\hat{u}}{2} - \frac{q\hat{q}}{2} - \frac{1}{2} \log(\hat{u} - \hat{q}) + \frac{\hat{m}^2 - \hat{q}}{2(\hat{u} - \hat{q})} . \tag{4.28}$$

We are now able to calculate the remaining integrals over the order parameters and their conjugated variables with the saddle-point method. So we start by writing the saddle-point equations relative respectively to $\hat{m}$, $\hat{u}$ and $\hat{q}$ from the above equation:

$$\hat{m} = 0, \tag{4.29}$$

$$\frac{1}{\hat{u} - \hat{q}} = 1 - q, \tag{4.30}$$

$$\hat{q} = \frac{-q}{(1-q)^2} . \tag{4.31}$$

By solving this equations we finally find:

$$\hat{m} = 0\,, \tag{4.32}$$

$$\hat{q} = \frac{-q}{(1-q)^2}\,, \tag{4.33}$$

$$\hat{u} = \frac{1-2q}{(1-q)^2}\,. \tag{4.34}$$

By injecting these results in (4.28) we found that when $q \to 1$, so when the volume associated with the admissible solutions is reduced to one point and therefore we are at critical capacity, we have:

$$\frac{1}{nN}\log Y \simeq \frac{1}{2(1-q)}\,. \tag{4.35}$$

As for the term χ this can be rewritten in the RS ansatz and using a Gaussian integral trick as:

$$\chi = \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu\right)}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}} \tag{4.36}$$

$$\times \left\{ \int_\kappa^\infty \prod_\mu dt_\mu \int_{-\infty}^\infty \prod_\mu \frac{d\hat{t}_\mu}{2\pi} e^{-\frac{1}{2}(1-q)\sum_{\mu,\nu}\hat{t}_\mu \Gamma_{\mu,\nu}\hat{t}_\nu} e^{i\sum_\mu \hat{t}_\mu (z_\mu\sqrt{q} + t_\mu - m\phi_0\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu))} \right\}^n .$$

After performing the Gaussian integral in $\{\hat{t}_\mu\}$, taking the small n limit and the $q \to 1$ limit we get:

$$\frac{\log \chi}{n} \simeq -\frac{1}{2(1-q)} \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu\right)}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}}$$

$$\times \min_{\{t_\mu \geqslant \kappa\}} \sum_{\mu,\nu} \left[t_\mu - (z_\mu + m\phi_0\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu)\right] \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} \left[t_\nu - (z_\nu + m\phi_0\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\nu)\right]\,. \tag{4.37}$$

Putting all together we get:

$$\frac{\langle Z_1^n\rangle - 1}{nN} \simeq \frac{1}{2(1-q)} \left\{ 1 - \alpha \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu\right)}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}} \right.$$

$$\left. \times \min_{\{t_\mu \geqslant \kappa\}} \sum_{\mu,\nu} \left[t_\mu - (z_\mu + m\phi_0\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu)\right] \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} \left[t_\nu - (z_\nu + m\phi_0\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\nu)\right] \right\}\,. \tag{4.38}$$

We finally obtain the expression for the critical capacity $\alpha_c(\kappa, p) = \max_m \alpha_c(m; \kappa, p)$, where $\alpha_c(m; \kappa, p)$ is the load α cancelling the terms inside the curly brackets in (4.38).

We can rewrite this final expression for the maximal load at fixed κ and p like

$$\boxed{\alpha_c(\kappa, p) = 1/\min_m \langle E_p(\hat{\mathcal{R}}, \mathcal{Z}, m; \kappa, \phi_0)\rangle_{\hat{\mathcal{R}},\mathcal{Z}}} \tag{4.39}$$

where the minimum is taken over $m = \sum_{j(\neq i)} W_{ij}$. In the formula above, $\langle\cdot\rangle$ denotes the average over the vectors $\hat{\mathcal{R}} = (\hat{\mathbf{r}}_1, ..., \hat{\mathbf{r}}_p)$ of p positions $\hat{\mathbf{r}}_\mu$ drawn uniformly at random in the D-dimensional cube with periodic boundary conditions, and $\mathcal{Z} = (z_1, ...z_p)$ drawn from the multivariate centered Gaussian distribution with $\hat{\mathcal{R}}$-dependent covariance matrix $\mathbf{\Gamma}_{\mu,\nu}(\hat{\mathcal{R}}) = \Gamma\big(\hat{\mathbf{r}}_\mu - \hat{\mathbf{r}}_\nu\big) - \phi_0^2$. Here, $\Gamma(d)$ is the overlapping volume between two PF, whose centers are at distance d from one another, hence, $\Gamma(0) = \phi_0$. The function E_p in Eq. (4.39) is defined through

$$E_p(\hat{\mathcal{R}}, \mathcal{Z}, m; \kappa, \phi_0) = \min_{\{t_\mu \geqslant \kappa\}} \sum_{\mu,\nu} \big[t_\mu - (z_\mu + m\phi_0\sigma(\hat{\mathbf{r}}_\mu)\big]\mathbf{\Gamma}(\hat{\mathcal{R}})^{-1}_{\mu,\nu}\big[t_\nu - (z_\nu + m\phi_0\sigma(\hat{\mathbf{r}}_\nu)\big] \tag{4.40}$$

where $\sigma(d) = 1$ if $d < r_c$, -1 otherwise and we have set $\mathbf{r}_1 = \mathbf{0}$ for traslation invariance. r_c is the radius of the PF, *i.e.*, the smallest number such that $\Gamma(2\,r_c) = 0$.

In practice, computing $\alpha_c(\kappa, p)$ from Eq. (4.39) is quite involved from a numerical point of view, as it requires to solve the p-dimensional semi-definite quadratic optimization problem [43, 68] in Eq. (4.40), as well as to average over the random vectors $\hat{\mathcal{R}}$ and $\mathcal{Z}$. This can be accurately done for small enough p, with results in excellent agreement with the SVM simulations, see Fig. 21.

Notice that as already mentioned in Section 4.4, for $p = 1$, our calculation reproduces Gardner's critical capacity $\alpha_c(\kappa, 1)$ for independent and biased patterns [87]. This is expected as spatial correlations between patterns within a map appear when $p \geqslant 2$, see Appendix A.3.

4.6 QUENCHED INPUT FIELDS THEORY

In order to compute the capacity $\alpha_c(\kappa, p)$ with the Gardner approach in Section 4.5 we have to solve a p-dimensional constrained quadratic optimization problem, depending on p correlated Gaussian random variables, see Eqs. (4.39) and (4.40), and then average over p random positions on the map. This task becomes quickly intractable in practice as p increases.

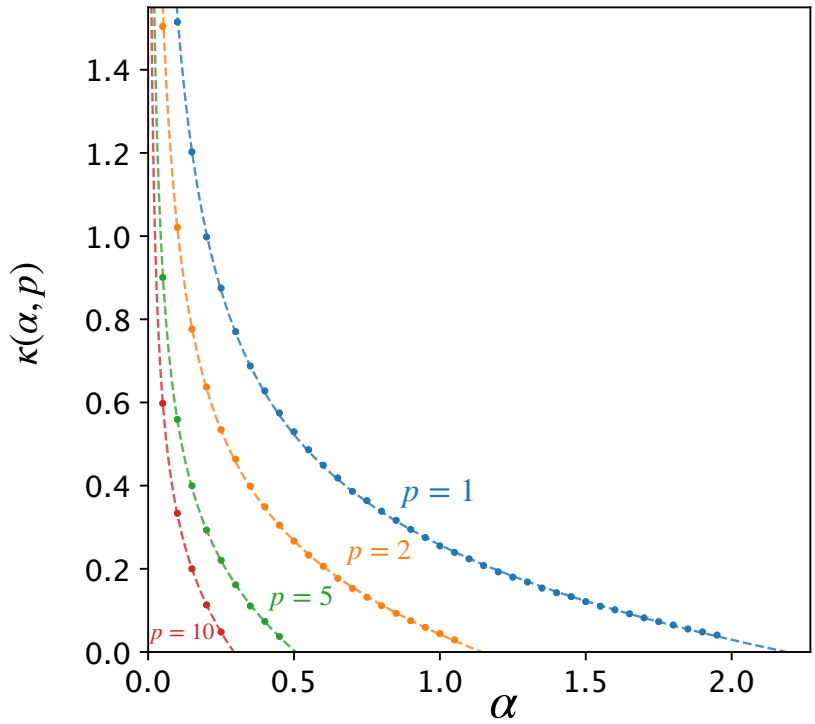

Figure 21 – Optimal stability κ as a function of the load α and the number p of positions. Dots, SVM results; dashed lines, Gardner's theory (4.39). Parameter values: $D = 2$, $\phi_0 = 0.3$, $N = 1000$ for SVM.

In other words Eq. (4.39) seems, unfortunately, intractable for large p. The intricate dependence on p, *i.e.*, showing up through the Gaussian correlations between the p random fields z_μ in Eq. (4.40), stems from the average (in each map ℓ) over the N PF centers $\{\mathbf{r}_i^\ell\}$ at fixed positions $\{\mathbf{r}^{\ell,\mu}\}$.

To avoid introducing these correlations and having an explicit dependence on the parameter p, we consider an alternative calculation scheme, where the p positions in each map are averaged out, while keeping the $L \times N$ centers quenched.

To further simplify the calculation we neglect in the effective action all terms of order $\geqslant 3$ in the couplings $\{W_{ij}\}$ following closely [165] but, as we will discuss later, this Gaussian approximation is expected to be exact in our setting in the large-p limit.

A potentially interesting feature of this approach is that it holds at fixed PFs, instead of averaging over them as in Section 4.5, and could be applied to specific situations, *i.e.*, sets of PFs measured in experiments.

4.6.1 *Replica calculation*

Starting from the replicated volume $\langle Z_1^n \rangle$ in (4.15), we now perform first the average over each one of the p locations in $\chi(\mathbf{W})$ in (4.16) as follows

$$\int d\hat{\mathbf{r}}_\mu \; \exp\left(-i \sum_a \hat{t}_{\mu a}(2\sigma_1^{\ell,\mu} - 1) \sum_{j\geqslant 2} W_{ja}\, \sigma_j^{\ell,\mu}\right) =$$
$$\exp\left(-i \sum_a m_\ell^a\, \hat{t}_{\mu a} - \frac{1}{2} \sum_{a,b} \hat{t}_{\mu a}(q_\ell^{ab} - m_\ell^a m_\ell^b)\hat{t}_{\mu b} + O(\hat{t}^3)\right) \tag{4.41}$$

where we have reintroduced the map index ℓ to underline that the PFs are kept fixed here. The order parameters in the formula above are

$$m_\ell^a = \sum_{j\geqslant 2} W_{ja} \left(2\,\Gamma(|\mathbf{r}_j^\ell - \mathbf{r}_1^\ell|) - \phi_0\right) \tag{4.42}$$

and

$$q_\ell^{ab} = \sum_{j,k\geqslant 2} W_{ja}\, W_{kb}\, \Gamma(|\mathbf{r}_j^\ell - \mathbf{r}_k^\ell|)\ , \tag{4.43}$$

to be introduced as in Section 4.5.

We then simplify the calculation with two approximations:

— We truncate the expansion in powers of $\hat{t}$ in (4.41) to the second order, and omit all higher order terms. This amounts to approximate the distribution of couplings $\{W_{ij}\}$ (at fixed PFs) by a Gaussian. This approximation is valid only if the couplings fluctuate weakly around their means, which is the case in the large-p limit, see Section 4.6.3.

— We also neglect the dependence of the order parameters m_ℓ and q_ℓ above on the map ℓ. The histograms of the overlaps q_ℓ measured by SVM are shown in Fig. 22. As can be seen from the figure, the distribution of overlaps is not concentrated in the large-N limit at fixed p. Therefore, while $m_\ell^a = m^a$ and $q_\ell^{ab} = q^{ab}$ is a valid Ansatz for the saddle-point equations of the log. partition function (due to the statistical equivalence between the maps), we expect Gaussian fluctuations to be relevant even in the infinite-N limit. However, as p increases, these fluctuations are smaller and smaller, and are asymptotically negligible. The order parameters then reduce to, after summation over the maps $\ell = 1, \ldots, L$,

$$m^a \equiv \frac{1}{L} \sum_{\ell=1}^{L} m_\ell^a = \sum_{j\geqslant 2} W_{ja} \left(2\,\mathcal{C}_{1j}(\{\mathbf{r}_j^\ell\}) - \phi_0\right) \tag{4.44}$$

and

$$q^{ab} \equiv \frac{1}{L}\sum_{\ell=1}^{L} q_\ell^{ab} = \sum_{j,k\geqslant 2} W_{ja}\, W_{kb}\, \mathcal{C}_{jk}\big(\{\mathbf{r}_j^\ell\}\big)\ . \tag{4.45}$$

The $N \times N$ multi-space Euclidean random matrix $\mathcal{C}$ appearing in the expressions above is defined by

$$\boxed{\mathcal{C}_{jk}\big(\{\mathbf{r}_i^\ell\}\big) = \frac{1}{L}\sum_{\ell=1}^{L} \Gamma\big(|\mathbf{r}_j^\ell - \mathbf{r}_k^\ell|\big)\ .} \tag{4.46}$$

In the following, we denote by $\rho(\lambda)$ the density of eigenvalues λ of $\mathcal{C}$. This density is self-averaging when the PFs are randomly drawn in the large L, N double limit. Its resolvent, defined as

$$g(U) = \int d\lambda\, \frac{\rho(\lambda)}{\lambda + U}\ , \tag{4.47}$$

where the integral runs over the support of $\rho(\lambda)$, is solution of the implicit equation

$$U = -\frac{1}{g(U)} + \sum_{\mathbf{k}\neq \mathbf{0}} \frac{\alpha\, \hat{\Gamma}(\mathbf{k})}{\alpha + g(U)\, \hat{\Gamma}(\mathbf{k})}\ , \tag{4.48}$$

where the $\hat{\Gamma}(\mathbf{k})$'s are the components of the Fourier transform of Γ on the D-dimensional infinite reciprocal cube, as we will see in detail in Chapter 5 [8].

Within the RS Ansatz, the overlap matrix q^{ab} is fully characterized by its diagonal and off-diagonal elements that we denote by, respectively, s and q:

$$s = \sum_{i,j\geqslant 2} \langle \mathcal{C}_{ij}\, [W_{1i}\, W_{1j}]\rangle \quad , \qquad q = \sum_{i,j\geqslant 2} \langle \mathcal{C}_{ij}\, [W_{1i}]\, [W_{1j}]\rangle\ . \tag{4.49}$$

where, as in Section 4.5, the brackets denotes the average over the random patterns, and the square parenthesis stand for the average over all couplings satisfying the inequalities (4.9).

8. In Chapter 5 we will show (Section 5.2) how to compute the expression for the implicit equation of the resolvent $g(U)$, that is Eq. (4.48), associated to the matrix $\mathcal{C}$ defined in Eq. (4.46), see the result in Eq. (5.21). An alternative derivation of the same result is presented in Section 5.3. It is important to note however that the notations in Chapter 4 and Chapter 5 are slightly different in the sense that $g(U)$ in the former corresponds to $s(z)$ in the latter.

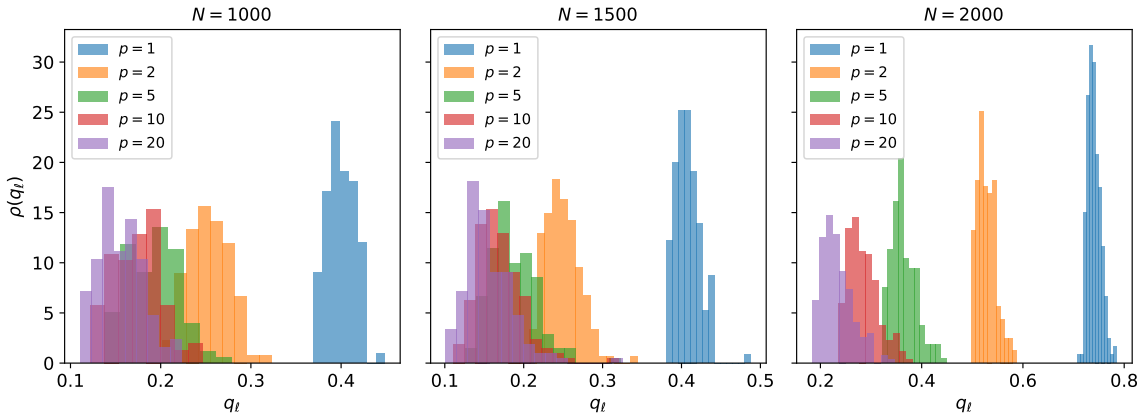

Figure 22 – Distributions of the overlaps q_ℓ for different values of N and p. It is clear that the histograms are roughly Gaussian. We use for this results $D = 2$, $\phi_0 = .3$, $\alpha = .1$, and we have averaged over 500 realization of the p positions at fixed PF.

Following closely [165], we obtain the expression of the average logarithm of the volume,

$$\frac{\langle \log Z_1 \rangle}{N} = -\frac{1}{2} q\hat{q} + s\hat{s} + m\hat{m} + \hat{u} - \frac{1}{2}\int d\lambda\, \rho(\lambda) \left[\log(2\hat{u} + (2\hat{s} - \hat{q})\lambda) + \frac{\hat{q}\lambda}{2\hat{u} + (2\hat{s} - \hat{q})\lambda} \right] + \frac{\hat{m}^2\, \Xi}{2(2\hat{s} - \hat{q})} + \alpha p \int Dz \log H\left(\frac{z\sqrt{q - m^2} - m + \kappa}{\sqrt{s - q}} \right) \tag{4.50}$$

where Dz denotes the Gaussian measure, $H(x) = \int_x^\infty Dz = \frac{1}{2}\,\mathrm{erfc}(\frac{x}{\sqrt{2}})$, and the $\hat{\cdot}$ Lagrange parameters enforce the definitions of the order parameters ($\hat{u}$ enforces the normalization condition over the rows of the $\mathbf{W}$ matrix). The quantity Ξ is a function of the argument [9]

$$U = \frac{2\hat{u}}{2\hat{s} - \hat{q}}\,, \tag{4.51}$$

and is defined as

$$\Xi(U) = \sum_{j,k \geqslant 2} H_j \left(U\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} H_k \qquad \text{with} \qquad H_j = 2\,\mathcal{C}_{1j} - \phi_0\,, \tag{4.52}$$

9. The physical meaning of U is to represent the eigenvalues of the multi-space Euclidean random matrix associated to the different environments, similar to the pattern correlation matrix in [165]. Always in the same reference it is possible to better understand the relationship between U and κ on the critical line, see for example the Eq. (17) associated, where ν corresponds to our U. Moreover, in Section 3.4 of [164] (a work prior to [165]) it is shown how ν is also found from the modified Hebb rule to take into account the correlations. The presence of ν is not specific of the replicas, but is also found in simpler approaches (Hebb + denoising) and allows to interpolate between $\mathbf{K}$ equal to the identity matrix for very high ν and $\mathbf{K}$ equal to the inverse of the correlation matrix of the patterns for $\nu \to 0$ according to the constraints that one puts on denoising, see Eq. (3.31) of [164].

where **Id** is the identity matrix of size $N-1$. In the above equation, the inverse is intended over the $(N-1)$-dimensional restriction of the matrix U **Id** $+ \mathcal{C}$ to entries $j, k \geqslant 2$. The quantity $\Xi(U)$ can be calculated, we report in Appendix A.4 the details, with the result:

$$\Xi(U) = 1 + 4U - \frac{4}{g(U)} \,. \tag{4.53}$$

4.6.2 *Log. volume and saddle-point equations close to the critical line*

As α reaches its maximal value (at fixed κ and p), the set of couplings satisfying the inequalities (4.9) shrink to a single solution, and we expect s, q to reach the same value according to Eq. (4.49). We therefore look for an asymptotic expression for $\frac{1}{N^2}\langle \log Z \rangle$ in Eq. (4.50) when

$$\epsilon = s - q \,, \tag{4.54}$$

is very small and positive [10]. In this regime, we expect the conjugated Lagrange parameters to diverge as inverse powers of ϵ. More precisely, calling

$$\hat{\epsilon} = 2\hat{s} - \hat{q} \,, \tag{4.55}$$

we assume that

$$\hat{\epsilon} = \frac{V}{\epsilon} \quad \text{and} \qquad \hat{q} = \frac{T}{\epsilon^2} \,, \tag{4.56}$$

as $\epsilon \to 0$. To the leading order, we obtain

$$\frac{1}{N}\langle \log Z_1 \rangle = \frac{F(\alpha)}{2\epsilon} + O\Big(|\log \epsilon|\Big) \,, \tag{4.57}$$

where $F(\alpha)$ is the extremum over m, q, U, V, T of $F(\alpha; m, q, U, V, T)$ that is equal to

$$V\left(q + U - \frac{m^2}{\Xi(U)}\right) + T\left(1 - \frac{1}{V}\int d\lambda\, \rho(\lambda)\frac{\lambda}{\lambda + U}\right) - \alpha p (q - m^2)\int_x^\infty \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} (z - x)^2 \tag{4.58}$$

10. Note that here ϵ is defined as the difference between the order parameters $s - q$, where we restrict our analysis to the critical line $\alpha_c(\kappa, p)$ at fixed p so that $s - q = \epsilon \to 0^+$. Hence, in this case ϵ is different from the spatial error considered in the rest of the manuscript which is defined in the limit of p large as $\epsilon \sim p^{-\frac{1}{D}}$.

with $x = \frac{m-\kappa}{\sqrt{q-m^2}}$ and U defined in (4.51). Note that, in order to obtain (4.58), the saddle point equation over $\hat{m}$ in (4.50) was derived and solved for $\hat{m}$ explicitly. Extremizing over U, T, V, we obtain

$$V = \int d\lambda\, \rho(\lambda) \frac{\lambda}{\lambda + U} \,, \tag{4.59}$$

$$T = -\left(q + U - \frac{m^2}{\Xi(U)}\right) \int d\lambda\, \rho(\lambda) \frac{\lambda}{\lambda + U} \,, \tag{4.60}$$

$$1 + \frac{m^2}{\Xi(U)^2} \frac{d\Xi}{dU} = \left(q + U - \frac{m^2}{\Xi(U)}\right) \int d\lambda\, \rho(\lambda) \frac{\lambda}{(\lambda + U)^2} \,. \tag{4.61}$$

Note that the derivative of Ξ with respect to U can be easily computed from the derivative of g with respect to U according to Eq. (4.53). Following the implicit equation over g in Eq. (4.48), we find

$$\frac{dg}{dU}(U) = \frac{1}{\displaystyle\sum_{\mathbf{k}\neq\mathbf{0}} \frac{\alpha \hat{\Gamma}(\mathbf{k})}{(\alpha + g\,\hat{\Gamma}(\mathbf{k}))^2} - \frac{1}{g^2}} \,. \tag{4.62}$$

We may now write the saddle-point equations over q and m, which give, after some elementary manipulation,

$$\alpha\, p\, H(x) = \int d\lambda\, \rho(\lambda) \frac{\lambda}{\lambda + U} \,, \tag{4.63}$$

$$\frac{m}{m-\kappa}\left(\frac{1}{\Xi(U)} - 1\right) = \frac{1}{\sqrt{2\pi}\, x\, e^{x^2/2}\, H(x)} - 1 \,. \tag{4.64}$$

The three coupled equations (4.61,4.63,4.64) allows one, in principle, to compute q, m, U, and therefore T, V and $F(\alpha)$. In addition, the optimization of $\langle \log Z \rangle$ in (4.57) over ϵ immediately gives $F(\alpha) = 0$, hence, a fourth equation to determine the critical value of α at fixed κ and p. This last equation read, after the simplification according to Eq. (4.64),

$$\frac{U}{\kappa} = m\left(\frac{1}{\Xi(U)} - 1\right) \,. \tag{4.65}$$

Resolution of these equations gives access to $\kappa(\alpha, p)$, in very good agreement with the numerical results obtained with SVM, see Fig. 23(a). Small deviations can, however, be noticed and diminish with increasing p as expected, see Section 4.6.3.

In addition, the order parameters q and m are shown as functions of p in Fig. 24, in good agreement with SVM results for large p ($\gg p_{c.o.}$).

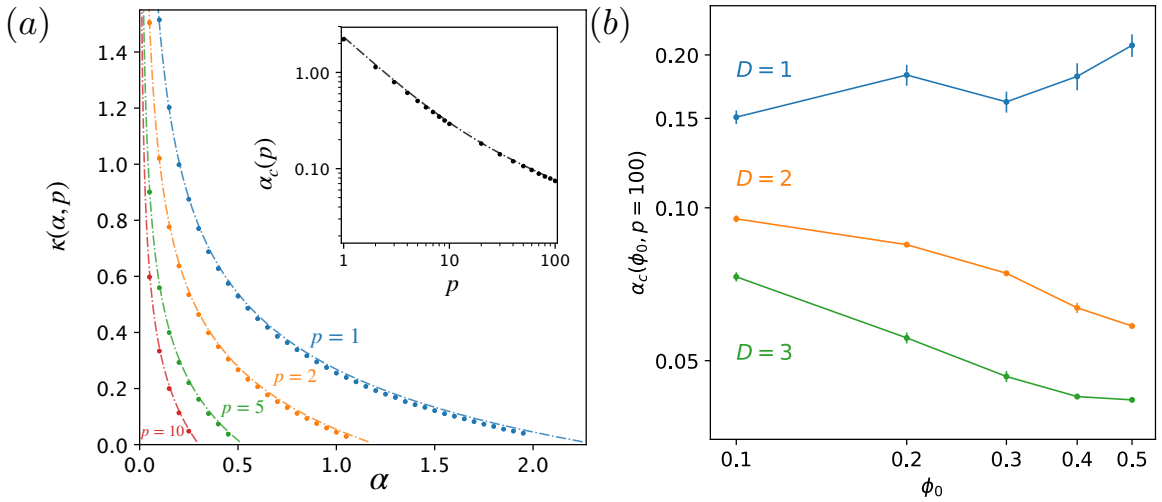

Figure 23 – (a) Optimal stability κ as a function of the load α and the number p of positions. Dots, SVM results; dashed lines dashed-dotted lines, quenched PF theory. Parameter values: $D = 2$, $\phi_0 = 0.3$, $N = 1000$ for SVM. Inset: $\alpha_c(p)$ vs p. Dots indicate results from SVM ($N = 5000$), averaged over 50 samples; the dot size indicates the maximal error bar. The dashed-dotted line shows the predictions from the quenched PF theory. (b) Critical capacity obtained by SVM vs ϕ_0 for different values of D in log-log scale. Parameter values: $N = 5000$, $p = 100$, Samples$= 25$.

Furthermore, the value of p at which the confluence between the results from the quenched theory and SVM takes place is a decreasing function of the PF size ϕ_0 and of the map dimension D, see Appendix A.5.

4.6.3 *Large-p behavior of the critical capacity*

We now focus on the maximal capacity, obtained when $\kappa \to 0$. According to (4.65), U vanishes, and equations (4.63,4.64) as a well as the implicit Eq. (4.48) on the resolvent g give a set of two coupled equations for x and the resolvent g:

$$\frac{1}{g} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}(\mathbf{k})}{1 + g\, p\, H(x)\, \hat{\Gamma}(\mathbf{k})} \,, \tag{4.66}$$

$$1 - \frac{4}{g} = x\sqrt{2\pi}\, H(x)\, e^{x^2/2} \,. \tag{4.67}$$

from which the capacity can be computed as a function of the number p of points,

$$\alpha_c(p) = \frac{1}{p\, H(x)} \,. \tag{4.68}$$

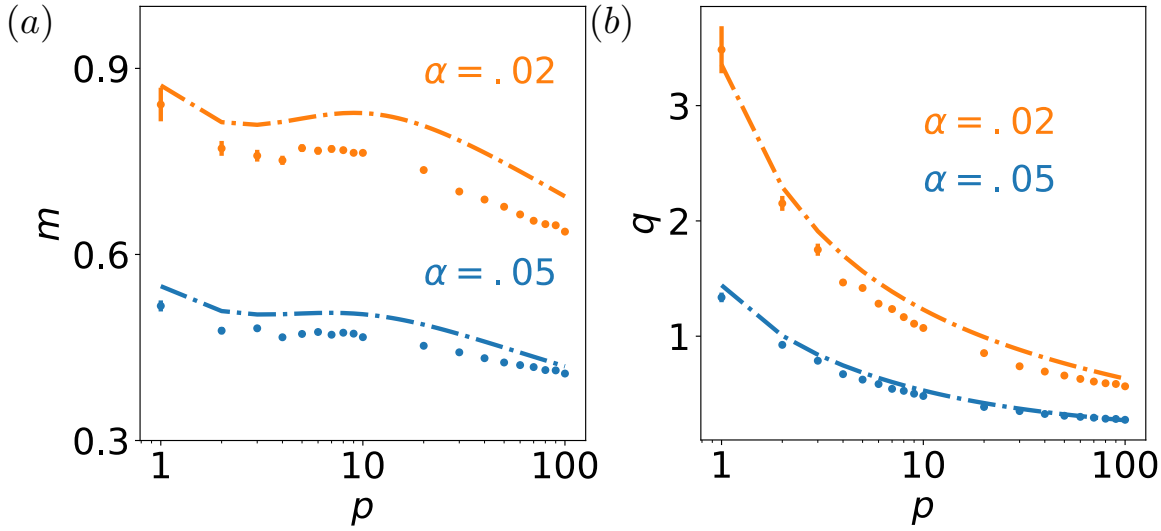

Figure 24 – Order parameters m (a) and q (b) vs p. Dots, SVM results ($N = 2500$), averaged over 50 samples; dashed-dotted lines, quenched PF theory. Parameters: $D = 2$, $\phi_0 = 0.3$, $\alpha = 0.02$ (top) and 0.05 (bottom), for which up to, respectively, $p_c \simeq 2500$ and $p_c \simeq 250$ points can be memorized.

In practice, we can choose x at will, compute g from (4.67), then p from (4.66), and, finally, $\alpha_c(p)$ from (4.68), see Fig. 23(a,inset) for the results of the numerical resolution of these equations.

Remark that equation (4.66) can be rewritten as

$$p\, H(x) = G\Big(g\, p\, H(x)\Big) \qquad \text{with} \qquad G(y) = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{y\, \hat{\Gamma}(\mathbf{k})}{1 + y\, \hat{\Gamma}(\mathbf{k})} \,. \tag{4.69}$$

According to dimensional analysis, the large momentum scaling of the Fourier coefficients is given by

$$\hat{\Gamma}(\mathbf{k}) \sim \frac{\phi_0^2}{\left(k\, \phi_0^{\frac{1}{D}}\right)^{D+1}} = \frac{\phi_0^{1-\frac{1}{D}}}{k^{D+1}} \,, \tag{4.70}$$

where $k = |\mathbf{k}|$ and D is the dimension of the physical space. We deduce that, for large arguments y,

$$G(y) \sim A_1(D)\, \phi_0^{\frac{D-1}{D+1}}\, y^{\frac{D}{D+1}} \qquad \text{with} \qquad A_1(D) = \int \frac{d^D\mathbf{u}}{|\mathbf{u}|^{D+1} + 1} \,. \tag{4.71}$$

In addition, using the asymptotic expansion of the erfc function, we have

$$x\sqrt{2\pi}\,H(x)\,e^{x^2/2} \simeq 1 - \frac{1}{x^2} \tag{4.72}$$

for large x.

Combining these expressions allows us to obtain the asymptotic relation between x and y,

$$y^{\frac{1}{D+1}} = 4\,A_1(D)\,\phi_0^{\frac{D-1}{D+1}}\,x^2 \ . \tag{4.73}$$

and, to the leading order in p,

$$x \simeq \sqrt{2\log p} - \left(D + \frac{1}{2}\right)\frac{\log\log p}{\sqrt{2\log p}} \ . \tag{4.74}$$

We then deduce the asymptotic scaling of the critical capacity

$$\boxed{\alpha_c(p) \sim A(D)\,\frac{\phi_0^{-(D-1)}}{(\log p)^D} \qquad (p \to \infty)\ ,} \tag{4.75}$$

with

$$A(D) = \frac{1}{8^D\,A_1(D)^{D+1}} \ . \tag{4.76}$$

Equation (4.75) is the main result of this thesis. Informally speaking, the very slow decay of the critical capacity with p, see Fig. 23(a,inset), means that recurrent neural nets can efficiently store multiple spatial maps, even at high spatial resolution. More precisely, enforcing a strong reduction of the spatial error, such as $\epsilon \to \epsilon^2$, results in a moderate drop of the maximal sustainable load, $\alpha_c \to \alpha_c/2^D$.

In addition, the capacity is predicted to be a decreasing function of the PF size in dimensions $D = 2, 3$, but not in dimension $D = 1$. This asymptotic statement is qualitatively corroborated by SVM results, even for moderate values of p, see Fig 23(b).

Moreover, the scaling for x in Eq. (4.74) entails the following relation between the order parameters q and m in the large–p regime,

$$\frac{q}{m^2} - 1 \sim \frac{1}{2\,\log p} \ . \tag{4.77}$$

To interpret the consequences of the equation above, we consider a set of replicated couplings, $\{W_{ia}\}$. For any random position $\mathbf{r}$ in map ℓ defining the pattern $\boldsymbol{\sigma}$, we define the rescaled and centered random variable

$$B\left(\mathbf{r}|\{W_{ia}\}\right) = \frac{1}{m_\ell^a}\left((2\sigma_i^\ell - 1)\sum_{j\geqslant 2} W_{ja}\,\sigma_j^\ell - m_\ell^a\right) . \tag{4.78}$$

By definition of the order parameter m, the average value of B vanishes:

$$\left\langle B\left(\mathbf{r}|\{W_{ia}\}\right)\right\rangle_{\mathbf{r}} = 0 . \tag{4.79}$$

Equation (4.77) implies that the variance of B is

$$\left\langle B\left(\mathbf{r}|\{W_{ia}\}\right)^2\right\rangle_{\mathbf{r}} \simeq \frac{1}{2\log p} , \tag{4.80}$$

as p gets large and the load takes its maximal value (critical capacity). In other words, the standard deviation of B scales as $(\log p)^{-1/2}$ for large p. We thus expect that the k^{th} cumulant of B will scale as $(\log p)^{-k/2}$. Under this assumption, the distribution of the stability t has mean value m and fluctuations of the order of $\Delta t = m/\sqrt{\log p}$. These fluctuations are negligible in the large–p limit, since resolution of the saddle-point equation (4.61) shows that

$$m \simeq \frac{D}{4} - \frac{D^2}{256\,(\log p)^3} + o\left(\frac{1}{(\log p)^3}\right) \tag{4.81}$$

at the critical point. Hence, $\Delta t \sim (\log p)^{-1/2}$ is smaller and smaller as p increases, and the distribution of t is well approximated by a Gaussian in the large–p limit. The Gaussian approximation obtained by discarding all powers of $\hat{t}$ of order $\geqslant 3$ in Eq. (4.41) in our quenched PF theory is therefore expected to be exact in this limit.

5

SPECTRUM OF MULTI-SPACE EUCLIDEAN RANDOM MATRICES

This Chapter is dedicated to a detailed investigation of the spectral properties of the random matrix introduced in Section 4.6, namely in Eq. (4.46), given by the superimposition of an extensive number of independent random Euclidean matrices in the high-density limit. More specifically, we calculate analytically its resolvent both with free probability theory techniques and with the replica method from statistical physics of disordered systems. The results for the spectrum and eigenmodes are shown in the particular case of the model presented in Section 4.2, and are corroborated by numerical simulations.

All the findings presented in this Chapter have been published in [29].

5.1 INTRODUCTION

In the twenty years following their introduction, Euclidean Random Matrices (ERM) have been studied in a variety of contexts in physics [44, 158, 219] and mathematics [38, 64, 70]. Examples of applications of ERM include the theoretical description of vibrations in topologically disordered systems [100, 101, 190], wave propagation in random media [92, 219], relaxation in glasses [59], Anderson localization [13] and many more [93]. There is also no lack of ERM applications in the modeling of biological networks [102, 174].

While determining the spectral properties of ERM is generally quite involved due to the existence of correlations between the entries of these matrices, a well-understood limit is the so-called high-density regime [38, 158]. Assume N points $\mathbf{r}_i$ are drawn uniformly at random in a bounded space, *i.e.*, the unit D-dimensional hypercube $\mathcal{H}_D$, and define the N-dimensional ERM $\mathbf{M}^{(1)}$ with entries $M^{(1)}_{ij} = \Gamma(|\mathbf{r}_i - \mathbf{r}_j|)/N$. Here, $|\cdot|$ denotes the Euclidean distance (with periodic boundary conditions over $\mathcal{H}_D$), and Γ is a given function that depends only on the distance $|\cdot|$ and that has a finite range, independent of N. In the large-N limit (for fixed D), the points effectively form a dense, statistically uniform sampling of the hypercube; the eigenmodes of $\mathbf{M}^{(1)}$ are well approximated by Fourier plane waves [93, 158], with eigenvalues

$$\hat{\Gamma}(\mathbf{k}) = \int_{\mathcal{H}_D} d\mathbf{r}\, e^{-i\,2\pi\,\mathbf{k}\cdot\mathbf{r}}\, \Gamma(|\mathbf{r}|)\ , \tag{5.1}$$

where the components of $\mathbf{k} = (k_1, k_2, ..., k_D)$ are integer-valued.

Hereafter, we consider a novel statistical ensemble of ERMs in the high-density regime obtained by mixing multiple spaces. Instead of having a single set of N random points $\mathbf{r}_i$, we consider L such sets, $\mathbf{r}_i^\ell$, with $\ell = 1, ..., L$ (and $i = 1, ..., N$ as usual). Each index ℓ points to a different "space" (hypercube), and for simplicity all points are drawn uniformly at random in the different spaces. We define the superimposition of all the ERM attached to the spaces, with entries

$$C_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \Gamma\left(\left|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell\right|\right) \ . \tag{5.2}$$

The random matrix (5.2) has not yet been considered in statistical physics and is of exactly the same type as the one we encountered in Eq. (4.46) and called Multispace-ERM (MERM).

Let us briefly recall the motivation coming from the field of computational neuroscience that led us to introduce (Section 4.6) and study in detail MERM, more precisely, the need to understand how the hippocampal place-cell network [173] can account for multiple cognitive maps, coding for various environmental and contextual situations. From a model perspective the points $\mathbf{r}_i^\ell$ correspond to the positions of the centers of the place field of place cell i in map ℓ. The resulting statistical ensemble for MERM is sketched in Fig. 25. An important issue is the maximal number L of maps the hippocampal recurrent neural network (with N neurons) can sustain, more precisely, the maximal ratio

$$\alpha = \frac{L}{N} \ , \tag{5.3}$$

called critical capacity. This capacity depends on the dimension of the maps, $D \ll N$, and of their spatial accuracy (the precision with which N-dimensional neural configurations encode D-dimensional positions along the map). In Section 4.6 we have shown how the critical capacity could be determined from the knowledge of the resolvent of $\mathbf{C}$. A non trivial statistical setting is obtained when the number L of spaces is of the order of the matrix size, N. More precisely, we consider herefater the double infinite size limit $L, N \to \infty$ at fixed ratio α. This choice corresponds to the assumption that the hippocampal network activity can code for many different environments [9] or different contexts [122], and operates, as hypothesized for other cortical areas [24, 46, 47], in a regime close to maximal capacity.

This Chapter is organized as follows. The spectrum of MERM is computed using arguments borrowed from free probability theory in 5.2, and re-derived using the replica method in 5.3. Finally we show the results for the spectrum and eigenmodes for the choice of Γ corresponding to Fig. 25 and compare with numerical simulations in 5.4.

The reader unfamiliar with random matrix theory can find in Appendix B a brief summary of the key concepts needed to understand the following.

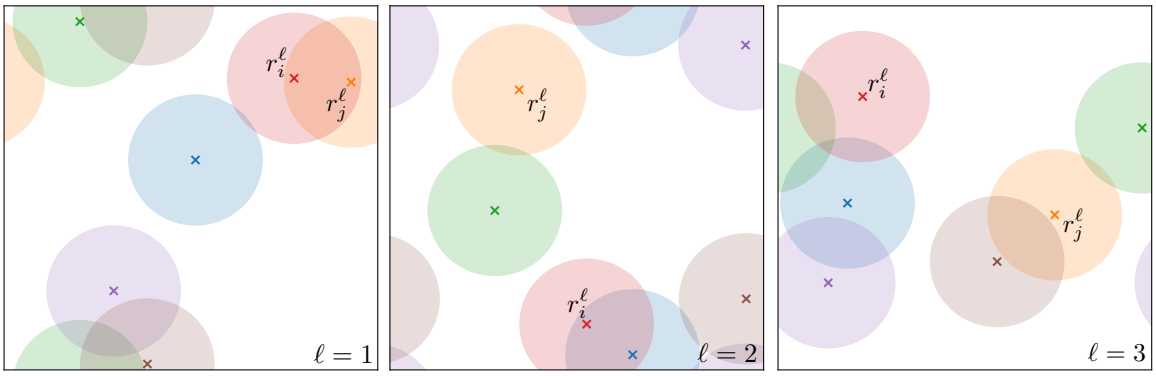

Figure 25 – Basic statistical ensemble of MERM considered in this Chapter, same as the one defined in Eq. (4.46) corresponding to the model defined in Section 4.2. $L = 3$ sets of $N = 6$ points, $\mathbf{r}_i^\ell$, with $\ell = 1, ..., L$ and $i = 1, ..., N$ are drawn uniformly at random in unit squares $\mathcal{H}_2$ (dimension $D = 2$). Points are represented by crosses, whose colors identify their indices i. The MERM is defined through (5.2), where Γ is a generic function of the distance between points. A possible choice for Γ, inspired from the place cells in neuroscience, is the overlap (common area) between pairs of disks (place fields) of surface $\phi_0 < 1$ and having centers $\mathbf{r}_i^\ell$ in each space ℓ.

5.2 SPECTRUM OF MERM: FREE-PROBABILITY-INSPIRED DERIVATION

Let us consider an extensive number L of spaces, see (5.3), with

$$M_{ij}^{(L)} = \frac{1}{N} \sum_{\ell=1}^{L} \Gamma\left(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\right) , \tag{5.4}$$

where the points are independently drawn from one space ℓ to another and where the single elements of the sum are ERM defined from N points $\mathbf{r}_i^\ell$ drawn uniformly at random in the D-dimensional unit hypercube $\mathcal{H}_D$:

$$M_{ij}^{(1)} = \frac{1}{N} \Gamma(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|) . \tag{5.5}$$

Before computing the spectrum of (5.4) we recall how to derive the spectral properties of (5.5) in the high-density regime heuristically, for a mathematically rigorous derivation see [38].

For any ERM $\mathbf{M}^{(1)}$, we can always formally write $\sum_{j=1}^{N} M_{ij}^{(1)} v_j(\mathbf{k}) = \hat{\Gamma}_i(\mathbf{k}) v_i(\mathbf{k})$ with $v_i(\mathbf{k}) = \frac{e^{i\mathbf{k}\cdot\mathbf{r}_i}}{\sqrt{N}}$ and

$$\hat{\Gamma}_i(\mathbf{k}) = \frac{1}{N} \sum_{j=1}^{N} e^{-i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j)} \Gamma(|\mathbf{r}_i - \mathbf{r}_j|). \tag{5.6}$$

In the large N limit the phase $-i\mathbf{k}\cdot(\mathbf{r}_i-\mathbf{r}_j)$ vary weakly between neighboring points $\mathbf{r}_i$ and $\mathbf{r}_j$ so that the sum in Eq. (5.6) can be approximated by an integral, thus $\hat{\Gamma}_i(\mathbf{k})$ does not depend anymore on i, becoming an eigenvalue of $\mathbf{M}^{(1)}$, $\hat{\Gamma}(\mathbf{k}) = N\hat{\Gamma}_0(\mathbf{k})$, where

$$\hat{\Gamma}_0(\mathbf{k}) = \frac{1}{N}\int_{\mathcal{H}_D} d\mathbf{r}\, e^{-i\,2\pi\,\mathbf{k}\cdot\mathbf{r}}\,\Gamma(|\mathbf{r}|) \tag{5.7}$$

is the Fourier transform of $\Gamma(|\mathbf{r}|)$. This eigenvalue is associated with the eigenvector $(\frac{e^{i\mathbf{k}\cdot\mathbf{r}_1}}{\sqrt{N}},\ldots,\frac{e^{i\mathbf{k}\cdot\mathbf{r}_N}}{\sqrt{N}})$. In fact, given a Fourier mode $\mathbf{k}$, the associated Fourier eigenvectors define a two dimensional space (linear combinations of sine and cosine corresponding to $\mathbf{k}$), and we denote by $\mathbf{v}^{\perp}$ the projection of the eigenvector $\mathbf{v}$ of $\mathbf{M}^{(1)}$ orthogonal to this two dimensional-space. We show the squared norm of $\mathbf{v}^{\perp}$ in Fig. 26 (averaged over several random realization and for the specific model defined in Fig. 25 but for $D = 1$ and $L = 1$) vs N in log-log scale. It is clear from Fig. 25 that the squared norm of the orthogonal projection of $\mathbf{v}$ scales as $\frac{1}{N}$; hence, each orthogonal squared component scales as $\frac{1}{N^2}$, and each orthogonal component as $\pm\frac{1}{N}$. This indicates that the eigenvectors of $\mathbf{M}^{(1)}$ are well approximated by Fourier modes (to the order $\frac{1}{\sqrt{N}}$) in the large N limit.

5.2.1 *Case of the extensive eigenvalue* - $\mathbf{k}=\mathbf{0}$

We would like to compute the resolvent (Stieltjes transform) of $\mathbf{M}^{(L)}$ using arguments from free-probability theory [144, 161, 180, 246]. Heuristically, asymptotic freeness between the different ERMs relies on the fact that their eigenvectors basis are mutually incoherent. In the $N\to\infty$ limit, the eigenvalues of $\mathbf{M}^{(1)}$ in space ℓ are given by (5.1) with associated eigenvectors of components $v_i(\mathbf{k}) \simeq e^{i\,2\pi\,\mathbf{k}\cdot\mathbf{r}_i^{\ell}}/\sqrt{N}$ [38, 93, 158]. All ERMs defined in the sum in (5.4) have mutually incoherent eigenbasis only if we restrict the analysis to the subspace orthogonal to the uniform mode attached to $\mathbf{k}=\mathbf{0}$, shared by all the spaces. Though this argument is not rigorous, we expect this restriction to allow us to find all the eigenvalues of $\mathbf{M}^{(L)}$, except the one corresponding to the asymptotically uniform eigenvector.

Furthermore it is easy to determine the leading behavior (when N is large) of the eigenvalue of $\mathbf{M}^{(L)}$ corresponding to $\mathbf{k}=\mathbf{0}$. As the corresponding eigenvector is expected to have all components equal to $N^{-1/2}$, we find that the corresponding eigenvalue is extensive in N and approximately equal to $\Lambda = N\,\alpha\,\hat{\Gamma}(\mathbf{0})$. For the matrix $\mathbf{C}$ the corresponding eigenvalue is $z_{ext} = \frac{\Lambda}{\alpha} = N\,\hat{\Gamma}(\mathbf{0})$.

From now on we concentrate on calculating the spectrum of $\mathbf{M}^{(L)}$ corresponding to $\mathbf{k}\neq\mathbf{0}$; the term "resolvent" will refer to the resolvent in the $\mathbf{k}\neq\mathbf{0}$ subspace.

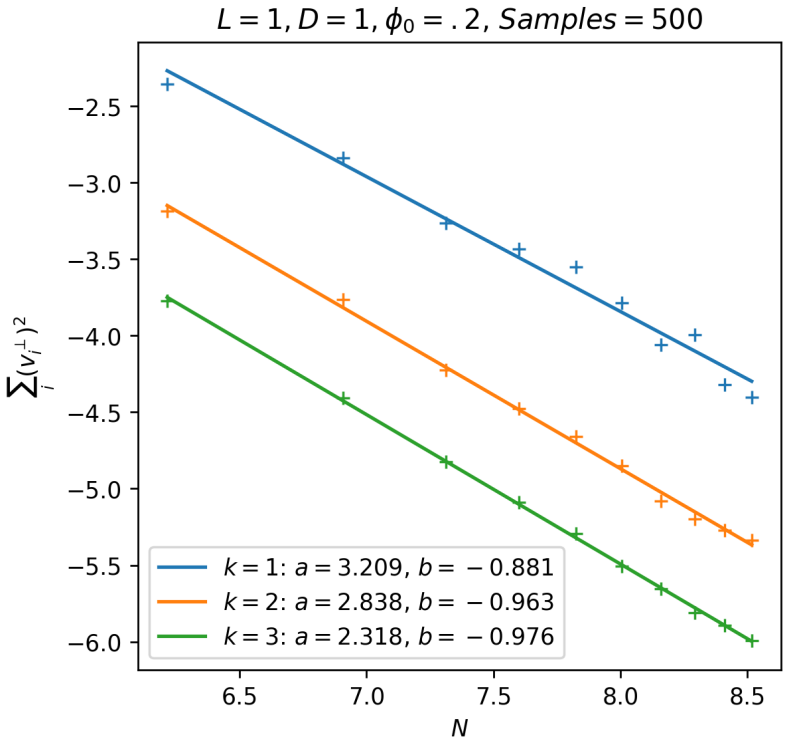

Figure 26 – $\sum_i (v_i^\perp)^2$ vs N in log-log scale for the specific model defined in Fig. 25 but for $D = 1$, $L = 1$ and for the top 3 eigenvectors corresponding to $k = 1, 2, 3$. Each point is averaged over $Samples = 500$ realizations and the errorbars are simply given by the standard deviation divided by $\sqrt{Samples - 1}$. The fits are of the form $\log y = a + b \log x$.

5.2.2 *Case of a single space (L=1)*

The resolvent of $\mathbf{M}^{(1)}$ is defined as [1]

$$s_1(z) = \frac{1}{N}\left\langle \text{Trace}\left(\mathbf{M}^{(1)} - z\,\mathbf{Id}\right)^{-1}\right\rangle_{(1)}, \tag{5.8}$$

1. To avoid ambiguity with the standard results presented in Appendix B.1 we must specify that the resolvent considered in this work $s(z)$ is defined as minus the resolvent used in the common notation in the literature, *i.e.*, $s(z) = -g(z)$, obviously nothing changes if we are consistent with the notation.

where $\langle\cdot\rangle_{(1)}$ stands for the average over the distribution of the matrix (5.5). It is easy to rewrite the resolvent when $N \gg 1$,

$$s_1(z) = -\frac{1}{zN}\left(N + \sum_{\ell=1}^{\infty} \sum_{\substack{\mathbf{k}\neq\mathbf{0} \\ (|\mathbf{k}|\leqslant N)}} \hat{\Gamma}(\mathbf{k})^\ell \, \frac{1}{z^\ell}\right) = -\frac{1}{z} - \frac{1}{N\,z}\,\gamma\left(\frac{1}{z}\right) \tag{5.9}$$

with

$$\gamma(u) = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{u\,\hat{\Gamma}(\mathbf{k})}{1 - u\,\hat{\Gamma}(\mathbf{k})} \tag{5.10}$$

and where the sum runs over $\mathbb{Z}^D$ without the $\mathbf{k} \neq \mathbf{0}$ term.

5.2.3 *Case of multiple spaces* ($L = \alpha N$)

We now consider the case of $\mathbf{M}^{(L)}$. It's resolvent $s_L(z)$ is defined as

$$s_L(z) = \frac{1}{N}\left\langle \text{Trace}\,\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-1}\right\rangle_{(L)}, \tag{5.11}$$

where $\langle\cdot\rangle_{(L)}$ stands for the average over the distribution of the matrix (5.4), can be computed through the following steps, as explained in Appendix B.2:

1. Invert (functionally) the resolvent $s_1(z)$ of $\mathbf{M}^{(1)}$[2]: we first rewrite (5.9) into the following implicit equation for the inverse resolvent:

$$z_1(s) = -\frac{1}{s} - \frac{1}{N\,s}\,\gamma\left(\frac{1}{z_1(s)}\right). \tag{5.12}$$

   We then send N to infinity in the above equation, and obtain that $z_1(s) = -1/s$ in this limit. Using this expression for the argument of the γ function in (5.12) we obtain the $\frac{1}{N}$–correction to the inverse resolvent:

$$z_1(s) = -\frac{1}{s} - \frac{\gamma(-s)}{N\,s}. \tag{5.13}$$

2. Compute the R-transform of $\mathbf{M}^{(1)}$, defined through

$$R_1(s) \equiv z_1(-s) - \frac{1}{s}. \tag{5.14}$$

   Note the unusual presence of a minus sign in the argument of z_1 in the above equation, due to the fact that our resolvent is defined as the opposite of the standard resolvent [144]. Using (5.13), we obtain

$$R_1(s) = \frac{\gamma(s)}{N\,s} + o\left(\frac{1}{N}\right). \tag{5.15}$$

2. That is, calculate the Blue function, see Appendix B.1.

3. Compute the R-transform of $\mathbf{M}^{(L)}$ through[3]

$$R_L(s) = L\, R_1(s)\ . \tag{5.16}$$

Using (5.15), we obtain,

$$R_L(s) = \alpha\frac{\gamma(s)}{s} + o(1)\ , \tag{5.17}$$

where the corrections $o(1)$ vanish when both $N, L \to \infty$ at fixed ratio α.

4. Write the functional inverse resolvent of $\mathbf{M}^{(L)}$ through

$$z_L(s) = R_L(-s) - \frac{1}{s} = -\frac{1+\alpha\,\gamma(-s)}{s}\ . \tag{5.18}$$

5. Compute the resolvent $s_L(z)$ of $\mathbf{M}^{(L)}$. From (5.18) and (5.10) we find the implicit equation satisfied by s_L:

$$z = \alpha\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\hat{\Gamma}(\mathbf{k})}{1+s_L\,\hat{\Gamma}(\mathbf{k})} - \frac{1}{s_L}\ . \tag{5.19}$$

Note that for what is needed in Section 4.6 we are interested in the spectral properties of the matrix $\mathbf{C}$ with entries

$$C_{ij} = \frac{1}{L}\sum_{\ell=1}^{L}\Gamma\left(\left|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell\right|\right) = \frac{1}{\alpha}M_{ij}^{(L)}\ . \tag{5.20}$$

Obviously, the resolvent s of $\mathbf{C}$ is related to the resolvent s_L of $\mathbf{M}^{(L)}$ through the equation $s(z) = \alpha\, s_L(\alpha z)$, see Appendix B.1. Hence we obtain our fundamental implicit equation for the resolvent of $\mathbf{C}$[4]:

$$\boxed{z = \sum_{\mathbf{k}\neq\mathbf{0}}\frac{\alpha\,\hat{\Gamma}(\mathbf{k})}{\alpha + s\,\hat{\Gamma}(\mathbf{k})} - \frac{1}{s}\ .} \tag{5.21}$$

3. Here the use of free-probability is crucial, see Appendix B.2. In fact, the important property to note is that since the different ERM corresponding to single environments in the subspace $\mathbf{k} \neq \mathbf{0}$ have orthogonal eigenbasis with high-probability due to random remapping (they are asymptotically free), the R-transform of the MERM ($R_L(s)$) is simply given by the sum of the R-transforms of the single ERM ($R_1(s)$), which are equal to each other, so $R_L(s) = L\, R_1(s)$.

4. That is the result we announced in Section 4.6. Note that compared to Eq. (4.48) there is only a change of notation, what was previously $g(U)$ has now become $s(z)$.

5.3 SPECTRUM OF MERM: REPLICA-BASED DERIVATION

Here we re-derive the implicit equation (5.19) for the resolvent of $\mathbf{M}^{(L)}$ defined in (5.11) using the replica method coming from statistical physics of disordered systems. We start by rewriting the definition of the resolvent, see Appendix B.1, as

$$s_L(z) = \frac{2}{N}\partial_z \Big\langle \log\det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}} \Big\rangle_{(L)} , \tag{5.22}$$

where $\langle\cdot\rangle_{(L)}$ it's still the average over the distribution of the matrix (5.4). With this representation the determinant $\det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}}$ can be expressed as a canonical partition function:

$$\mathcal{Z}_L(s) = \det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}} = \int \prod_i \frac{d\phi_i}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_i \phi_i^2 - \frac{1}{2}\sum_{ij}\phi_i M_{ij}^{(L)}\phi_j\right) , \tag{5.23}$$

where i, j go from 1 to N. Notice that it is legitimate to adopt a real-valued Gaussian representation for the inverse square root of the determinant. Each ERM $\mathbf{M}^{(l)}$ is a correlation matrix, and have real, non-negative eigenvalues; consequently, $\mathbf{M}^{(L)}$, which is the sum of correlation matrices, also has real and non-negative eigenvalues.

Resolvent (5.22) can be calculated using the replica trick [156, 157, 181]:

$$s_L(z) = \frac{2}{N}\partial_z \big\langle \log \mathcal{Z}_L(s)\big\rangle_{(L)} = \frac{2}{N}\partial_z\Big[\lim_{n\to 0}\frac{1}{n}\log\big\langle \mathcal{Z}_L(s)^n\big\rangle_{(L)}\Big] \tag{5.24}$$

with

$$\big\langle \mathcal{Z}_L(s)^n\big\rangle_{(L)} = \int \prod_{ia} \frac{d\phi_i^a}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_a\sum_i (\phi_i^a)^2\right)\Big\langle \exp\Big(-\frac{1}{2}\sum_a\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\Big)\Big\rangle_{(L)} , \tag{5.25}$$

where we have replicated the system n times, *i.e.*, a goes from 1 to n.

In order to perform the average in (5.25) we rewrite (5.4) by considering the ℓ-th space ERM in its eigenbasis:

$$M_{ij}^{(L)} = \frac{1}{N}\sum_{\ell=1}^{L} \Gamma\left(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\right) = \sum_\ell \sum_{\mathbf{k}\neq\mathbf{o}} \nu_{\mathbf{k}i}^\ell\, \hat{\Gamma}(\mathbf{k})\, \nu_{\mathbf{k}j}^\ell , \tag{5.26}$$

where ℓ goes from 1 to L, and the sum over $\mathbf{k}$ discards the $\mathbf{k} = \mathbf{o}$ extensive mode because as discussed in the previous Section 5.2. The eigenvector components, $\nu_{\mathbf{k}i}^\ell \simeq \frac{1}{\sqrt{N}}\sin\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right)$, $\frac{1}{\sqrt{N}}\cos\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right)$, are real due to the symmetry $\hat{\Gamma}(\mathbf{k}) = \hat{\Gamma}(-\mathbf{k})$. Hence we get

$$\Big\langle \exp\Big(-\frac{1}{2}\sum_a\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\Big)\Big\rangle_{(L)} = \Big\langle \exp\Big(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{o}}\hat{\Gamma}(\mathbf{k})\big(\sum_i \nu_{\mathbf{k}i}^\ell\phi_i^a\big)^2\Big)\Big\rangle_{(L)} . \tag{5.27}$$

We now use the Stratonovich trick to linearize $(\sum_i v^\ell_{\mathbf{k}i}\phi^a_i)^2$:

$$\left\langle \exp\Big(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{0}}\hat{\Gamma}(\mathbf{k})\big(\sum_i v^\ell_{\mathbf{k}i}\phi^a_i\big)^2\Big)\right\rangle_{(L)} = \prod_\ell\int\prod_{a,\mathbf{k}\neq\mathbf{0}}\frac{du^a_{\ell,\mathbf{k}}}{\sqrt{2\pi}}$$
$$\times\exp\Big(-\frac{1}{2}\sum_{a,\mathbf{k}\neq\mathbf{0}}\big(u^a_{\ell,\mathbf{k}}\big)^2\Big)\left\langle\exp\Big(-i\sum_{a,\mathbf{k}\neq\mathbf{0}}\sqrt{\hat{\Gamma}(\mathbf{k})}\,u^a_{\ell\mathbf{k}}\sum_i\phi^a_i v^\ell_{\mathbf{k}i}\Big)\right\rangle_{(L)}. \tag{5.28}$$

Using the fact that $\langle v^\ell_{\mathbf{k}i}\rangle = 0$ and $\langle v^\ell_{\mathbf{k}i}v^\ell_{\mathbf{k}'j}\rangle = \frac{1}{N}\delta_{ij}\delta_{\mathbf{k}\mathbf{k}'}$ it is easy to perform the average in the above equation, with the result

$$\left\langle\exp\Big(-i\sum_{a,\mathbf{k}\neq\mathbf{0}}\sqrt{\hat{\Gamma}(\mathbf{k})}\,u^a_{\ell\mathbf{k}}\sum_i\phi^a_i v^\ell_{\mathbf{k}i}\Big)\right\rangle_{(L)} = \exp\Big(-\frac{1}{2}\sum_{a,b}\sum_{\mathbf{k}\neq\mathbf{0}}\hat{\Gamma}(\mathbf{k})q^{ab}u^a_{\ell\mathbf{k}}u^b_{\ell\mathbf{k}}\Big) \tag{5.29}$$

where we have defined the overlap q^{ab} as

$$q^{ab} = \frac{1}{N}\sum_i\phi^a_i\phi^b_i \tag{5.30}$$

to be fixed through

$$1 = \int\prod_{a\leqslant b}\frac{d\hat{q}^{ab}dq^{ab}}{\frac{2\pi i}{N}}\exp\Big(N\sum_{a\leqslant b}\hat{q}^{ab}q^{ab} - \sum_{a\leqslant b}\hat{q}^{ab}\sum_i\phi^a_i\phi^b_i\Big). \tag{5.31}$$

We can finally write $\langle\mathcal{Z}_L(s)^n\rangle_{(L)}$ as

$$\int\prod_{a\leqslant b}\frac{d\hat{q}^{ab}dq^{ab}}{\frac{2\pi i}{N}}\exp\left\{N\left[\log\int\prod_a\frac{d\phi^a}{\sqrt{2\pi}}\exp\Big(\frac{z}{2}\sum_a(\phi^a)^2 - \sum_{a\leqslant b}\hat{q}^{ab}\phi^a\phi^b\Big)\right.\right.$$
$$\left.\left. + \sum_{a\leqslant b}\hat{q}^{ab}q^{ab} + \alpha\log\int\prod_{\mathbf{k}\neq\mathbf{0},a}\frac{du^a_{\mathbf{k}}}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\sum_{\mathbf{k}\neq\mathbf{0},a}(u^a_{\mathbf{k}})^2 - \frac{1}{2}\sum_{\mathbf{k}\neq\mathbf{0}}\sum_{a\leqslant b}\hat{\Gamma}(\mathbf{k})q^{ab}u^a_{\mathbf{k}}u^b_{\mathbf{k}}\Big)\right]\right\}. \tag{5.32}$$

The Gaussian integrals over ϕ^a and $u^a_{\mathbf{k}}$ can be easily computed. We then make the Replica Symmetric (RS) Ansatz on the structure of the order parameters q^{ab} and their conjugate variables $\hat{q}^{ab}$, so that

$$q^{ab} = r + (q-r)\delta_{ab} \tag{5.33}$$

and

$$\hat{q}^{ab} = \hat{r} + (\hat{q}-\hat{r})\delta_{ab}\,. \tag{5.34}$$

The integrals over q, r, $\hat{q}$ and $\hat{r}$ are then estimated using the saddle-point method valid for large N, and then taking the small n limit. The resulting expression for the resolvent of (5.4) is

$$s_L(z) = 2\partial_z \left[\mathrm{opt}_{q,r,\hat{q},\hat{r}} \lim_{n\to 0} \lim_{N\to\infty} \frac{1}{nN} \log \left\langle \mathcal{Z}_L(s)^n \right\rangle_{(L)} \right] = 2\partial_z \left[\mathrm{opt}_{q,r,\hat{q},\hat{r}} f(q,r,\hat{q},\hat{r}) \right] , \tag{5.35}$$

where f is the free energy density equal to

$$\begin{aligned} f(q,r,\hat{q},\hat{r}) = \hat{q}q - \frac{1}{2}\hat{r}r - \frac{\alpha}{2} \sum_{\mathbf{k}\neq\mathbf{0}} \left[\log\left(1+\hat{\Gamma}(\mathbf{k})(q-r)\right) + \frac{\hat{\Gamma}(\mathbf{k})r}{1+\hat{\Gamma}(\mathbf{k})(q-r)} \right] \\ -\frac{1}{2}\log\left(2\hat{q}-\hat{r}-z\right) - \frac{\hat{r}}{2(2\hat{q}-\hat{r}-z)} \end{aligned} . \tag{5.36}$$

The saddle-point equations obtained by optimizing $f(q,r,\hat{q},\hat{r})$ with respect to $\hat{q}$, $\hat{r}$, q and r read

$$\begin{aligned} q &= -\frac{\hat{r}}{(2\hat{q}-\hat{r}-z)^2} + \frac{1}{2\hat{q}-\hat{r}-z} , \quad r = -\frac{\hat{r}}{(2\hat{q}-\hat{r}-z)^2} , \\ \hat{q} &= \frac{\alpha}{2} \sum_{\mathbf{k}\neq\mathbf{0}} \left(\frac{\hat{\Gamma}(\mathbf{k})}{1+\hat{\Gamma}(\mathbf{k})(q-r)} - \frac{r\,\hat{\Gamma}(\mathbf{k})^2}{\left(1+\hat{\Gamma}(\mathbf{k})(q-r)\right)^2} \right) , \\ \hat{r} &= -\alpha \sum_{\mathbf{k}\neq\mathbf{0}} \frac{r\,\hat{\Gamma}(\mathbf{k})^2}{\left(1+\hat{\Gamma}(\mathbf{k})(q-r)\right)^2} . \end{aligned} \tag{5.37}$$

This system of equations admits $r = \hat{r} = 0$ as a solution, which gives, according to (5.35), the following implicit equation satisfied by $s_L(z)$:

$$z = \alpha \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}(\mathbf{k})}{1+s_L\,\hat{\Gamma}(\mathbf{k})} - \frac{1}{s_L} . \tag{5.38}$$

This equation is identical to (5.19) obtained using free probability theory.

It's interesting to note that since we consider the solution corresponding to $r = \hat{r} = 0$, an annealed computation where we directly bring the average of the different realizations of $\mathbf{M}^{(L)}$ inside the logarithm of the partition function, instead of keeping it out as in our version of the quenched computation that led us to use the replica method, it would have brought to the same result (obviously the quenched computation is a cleaner way to proceed) [157].

Moreover, we should as well observe that in this setting we have not needed to use the asymptotic freeness for the different ERMs $\mathbf{M}^{(1)}$ as we did in the free-probability computation of the previous Section 5.2, so in this sense this approach seems more powerful. In

fact, we can also use this method to study the case of MERM where the statistical features of the L ERM's are non-independent from space to space (in this case we can't use free probability because we lose the asymptotic freeness between the different ERMs because the entries of these are correlated with each other), see Section 6.5.

5.4 APPLICATION AND COMPARISON WITH NUMERICS

5.4.1 *Numerical computation of the spectrum*

We now aim at solving the implicit equation (5.21) satisfied by the resolvent of $\mathbf{C}$. We show in Fig. 27(a) the representative curve of z as a function of s around the pole at the origin ($s = 0$). A set of forbidden disjoint intervals, $z \in [z_-^{(m)}, z_+^{(m)}]$, with $m = 1, ..., M$ is found, which cannot be reached for real-valued s; the number M of these intervals is a decreasing function of the ratio α. When z lies in one of these intervals, we look for a solution to equation (5.21) with

$$s = s_r + i\, s_i \ , \tag{5.39}$$

where the imaginary part s_i is strictly positive. For $z = x + i\,\epsilon$, the density of eigenvalues at x is given by $\rho(x) = \lim_{\epsilon \to 0} s_i(z)/\pi$ by virtue of well-known properties of the Stieljes transform, see Appendix B.1. From now on we will indicate with z the eigenvalue and with $\rho(z)$ the correspondent density, bearing in mind the $\epsilon \to 0$ limit.

The implicit equations fulfilled by s_r and s_i for $z \in [z_-^{(m)}, z_+^{(m)}]$, with $m = 1, ..., M$ read

$$z = \sum_{\mathbf{k} \neq \mathbf{0}} \frac{\alpha^2\, \hat{\Gamma}(\mathbf{k})}{\left(\alpha + s_r\, \hat{\Gamma}(\mathbf{k})\right)^2 + \left(s_i\, \hat{\Gamma}(\mathbf{k})\right)^2} \ , \tag{5.40}$$

$$\frac{1}{s_r^2 + s_i^2} = \sum_{\mathbf{k} \neq \mathbf{0}} \frac{\alpha\, \hat{\Gamma}(\mathbf{k})^2}{\left(\alpha + s_r\, \hat{\Gamma}(\mathbf{k})\right)^2 + \left(s_i\, \hat{\Gamma}(\mathbf{k})\right)^2} \ , \tag{5.41}$$

and can be solved numerically. Figure 28 shows the density of eigenvalues for various values of α. We observe the presence of the disconnected intervals $[z_-^{(m)}; z_+^{(m)}]$ corresponding to non-zero density $\rho(z)$, referred to as "connected components" below. These connected components originate from the discrete spectrum of ERM (with eigenvalues labelled by $\mathbf{k}$) and progressively merge as α increases (Fig. 27(b)). We now discuss the mechanism leading to merging in the large $|\mathbf{k}|$, small α regime.

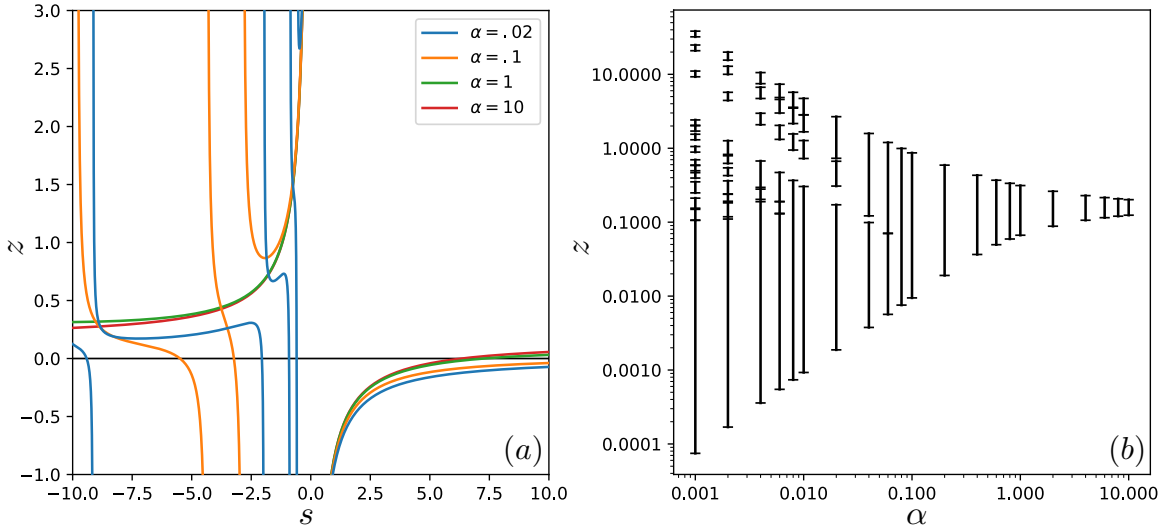

Figure 27 – (a) z vs s, see (5.21), close to the origin ($s = 0$), for different values of α. (b) Support of the spectrum for different values of α: black segments show the interval of eigenvalues z with non-zero density $\rho(z)$. Results obtained by taking for Γ the overlap (common length) between segments of length $\phi_0 = .2$, centered in points $\mathbf{r}_i^\ell$ randomly drawn in the unit interval $\mathcal{H}_1$ ($D = 1$), more precisely $\Gamma(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|) = \phi_0 - |\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|$.

5.4.2 *Merging of density "connected components": behavior of the density at small α*

For small α, we look for a solution of equation (5.21) near the poles, so that to consider only a value $\mathbf{k} \neq \mathbf{o}$ in the sum over the modes:

$$z(\mathbf{k}) = \frac{\alpha\,\hat{\Gamma}(\mathbf{k})}{\alpha + s(\mathbf{k})\,\hat{\Gamma}(\mathbf{k})} - \frac{1}{s(\mathbf{k})}\ . \tag{5.42}$$

We find then $s(\mathbf{k})$ such that $\frac{dz(\mathbf{k})}{ds(\mathbf{k})} = 0$, *i.e.*, where the resolvent has singularities (eigenvalues), obtaining:

$$s_\pm(\mathbf{k}) = -\frac{\alpha}{\hat{\Gamma}(\mathbf{k})}\left(1 \pm \sqrt{\alpha}\right)\ , \tag{5.43}$$

this implies that the spectrum has the edges located at:

$$z_\pm(\mathbf{k}) = \frac{\hat{\Gamma}(\mathbf{k})}{\alpha}\left(1 \pm 2\sqrt{\alpha}\right)\ . \tag{5.44}$$

This means that when α become sufficiently small the spectrum develop a connected component in correspondence of every $\mathbf{k} \neq \mathbf{o}$ centered in $z_\mathbf{k} = \frac{1}{2}(z_-(\mathbf{k}) + z_+(\mathbf{k})) = \frac{\hat{\Gamma}(\mathbf{k})}{\alpha}$ and

of half-width $\frac{1}{2}(z_+(\mathbf{k}) - z_-(\mathbf{k})) = \frac{2\hat{\Gamma}(\mathbf{k})}{\sqrt{\alpha}}$. In order now to understand how the density of eigenvalues behaves inside these connected components we look to a solution of equation (5.42) of the form

$$s(\mathbf{k}) = s_r(\mathbf{k}) + i\, s_i(\mathbf{k}) \,, \tag{5.45}$$

so that to finally obtain the parametric equations for the density $\rho(z)$ of eigenvalues equal to z:

$$\rho(x;\mathbf{k}) = \frac{\alpha^{\frac{3}{2}}}{\pi\hat{\Gamma}(\mathbf{k})}\sqrt{1-x^2} \quad , \qquad z(x;\mathbf{k}) = \frac{\hat{\Gamma}(\mathbf{k})}{\alpha}\left(1 + 2x\sqrt{\alpha}\right) \,, \tag{5.46}$$

where $x \in [-1;1]$. This solution makes sense only for the modes $\mathbf{k}$ and ratios α such that the local semi-circle distributions attached to two contiguous eigenvalues do not overlap. More precisely, the ratio α should be smaller than

$$\alpha_{merging}(\mathbf{k}) \simeq \frac{(\hat{\Gamma}(\mathbf{k}) - \hat{\Gamma}(\mathbf{k}^c))^2}{4(\hat{\Gamma}(\mathbf{k}) + \hat{\Gamma}(\mathbf{k}^c))^2} \,, \tag{5.47}$$

where $\mathbf{k}^c$ is the momentum vector corresponding to the closest eigenvalue to $\hat{\Gamma}(\mathbf{k})$. This formula gives the values of the ratios at which the small connected components of $\rho(z)$ (Figs. 27(b) and 28) successively merge, and is asymptotically correct for large $|\mathbf{k}|$.

When α is sufficiently large, all connected components have merged into a single continuous, semi-circle distribution, as could be expected from the vanishing correlation between the matrix elements of $\mathbf{C}$, centered in $z_1 = \frac{1}{2}(z_- + z_+) = \hat{\Gamma}_1$ and of half-width $\frac{1}{2}(z_+ - z_-) = 2\sqrt{\hat{\Gamma}_2/\alpha}$, with $\hat{\Gamma}_1 = \sum_{\mathbf{k}\neq\mathbf{o}} \hat{\Gamma}(\mathbf{k})$ and $\hat{\Gamma}_2 = \sum_{\mathbf{k}\neq\mathbf{o}} \hat{\Gamma}(\mathbf{k})^2$.

5.4.3 *Eigenvectors of MERM and Fourier modes associated to the ERMs*

We briefly discuss here the properties of the eigenvectors of MERM. We consider a connected component of eigenvalues originated from the same ERM eigenvalue (labelled by $\mathbf{k}$), see previous Subsection 5.4.2. To quantify how much the MERM eigenvectors $\mathbf{v}$ are related to the 2L eigenvectors (Fourier modes) of the L ERMs, we write

$$v_i = \sum_{\ell=1}^{L} \left(\gamma_\ell \frac{1}{\sqrt{N}} \cos\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right) + \delta_\ell \frac{1}{\sqrt{N}} \sin\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right) \right) + R_i \,, \tag{5.48}$$

where γ_ℓ and δ_ℓ are the projection coefficients onto the 2L ERMs eigenvectors and $\mathbf{R}$ is the component of $\mathbf{v}$ orthogonal to this subspace [5].

The distributions of the coefficients γ_ℓ, δ_ℓ and of the norm of $\mathbf{R}$ are shown in Fig. 29 in the case $L = 5$ and for increasing values of N. We observe that

5. $\mathbf{R}$ is equivalent to $\mathbf{v}^\perp$ encountered in Section 5.2 in the case of single ERM ($L = 1$).

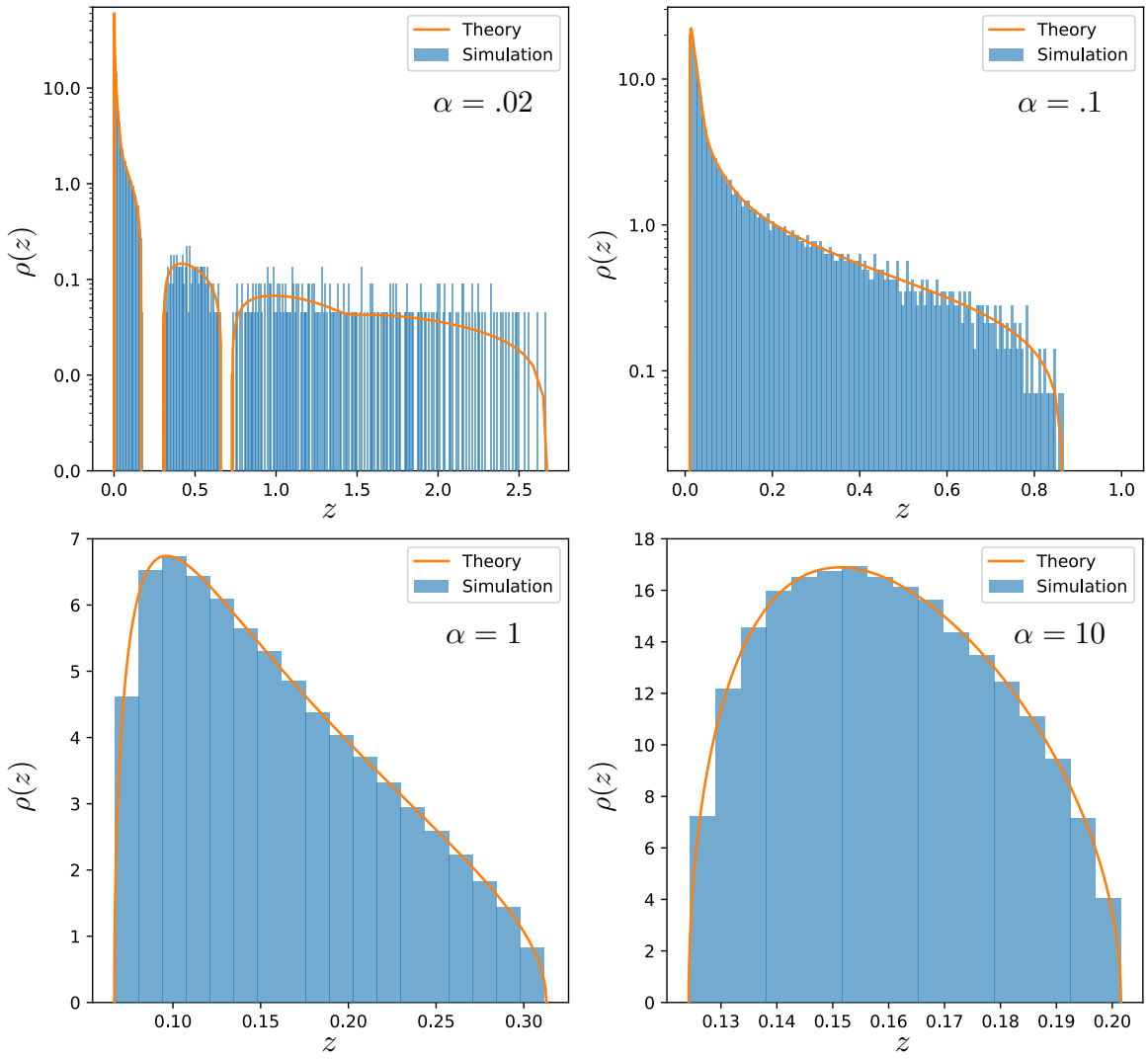

Figure 28 – Density of eigenvalues of $\mathbf{C}$, without the extensive eigenvalue $z_{\mathtt{ext}}$, for various values of α. Orange: results from (5.40). Blue: outputs of numerical diagonalization for $N = 2500$. Same model as in Fig. 27.

— the magnitude of γ_ℓ and δ_ℓ seems to be independent of N (Fig. 29(a)), which implies that these coefficients remain finite as $N \to \infty$. Conversely, the projections of $\mathbf{v}$ on Fourier modes attached to a momentum $\mathbf{k}' \neq \mathbf{k}$ vanishes with increasing N, see Fig. 29(b). Hence, $\mathbf{v}$ retains some coherence with the 2L eigenvectors of the ERMs attached to the connected component even in the infinite size limit (provided L remains finite).

— the norm of $\mathbf{R}$ seems to get peaked as N grows around a non-zero value. Therefore, $\mathbf{v}$ has a substantial component outside the 2L-dimensional subspaces spanned by the ERM eigenmodes.

Notice that the magnitudes of the γ, δ coefficients and of the norm of $\mathbf{R}$ are related to each other through $\langle \gamma^2 \rangle = \langle \delta^2 \rangle = (1 - \langle \mathbf{R}^2 \rangle)/L$ to ensure the normalization of $\mathbf{v}$. The results above were derived for finite L and large N; in the double scaling limit where both L, N are large at fixed ratio α, we find that the coefficients γ, δ of the projections on the Fourier modes attached to the connected component also scale as $N^{-1/2}$, in accordance with the number of those modes.

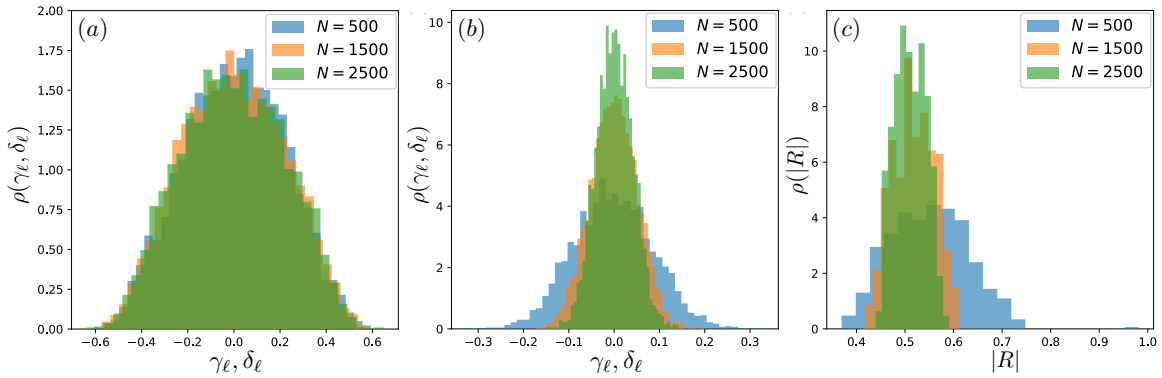

Figure 29 – (a) Histogram of the coefficients γ_ℓ and δ_ℓ for different values of N. Results correspond to the $k = 1$ connected component of eigenvalues in dimension $D = 1$ and for $L = 5$ spaces, averaged over 50 samples. Same model as in Fig. 27. (b) Histograms of the projections of eigenvectors $\mathbf{v}$ to the $k = 2$ Fourier modes of the ERMs. (c) Histograms of the norm of the orthogonal component $\mathbf{R}$, see (5.48).

6 TOWARDS GREATER BIOLOGICAL PLAUSIBILITY

In this Chapter we give additional details of the model presented in Chapter 4 and also generalize the latter to take into account some constraints of biological nature. In particular, we try to understand to what extent the main result of this thesis, the scaling found in Eq. (4.75), is robust to variations of the model. We also try to propose links of our theory with ongoing experiments. Finally, we present another application of our setting to the case of storing continuous attractors in recurrent neural networks starting from real images. Not all the results presented in this Chapter can be considered at a final stage and this will give us the opportunity to present in the Conclusions different lines of research to follow in the near future, see Chapter 7.

Part of the findings presented in this Chapter have been published in [28] and [29].

6.1 INTRODUCTION

In Chapter 4 we found a non-trivial result suggesting that recurrent neural networks are very efficient devices in the storage of continuous attractors, contrary to what it seemed in standard models based on Hebbian-like learning rules of the type (3.15), see Section 3.7. Despite this positive result, there is still work to be done especially in view of the fact that the model presented in Chapter 4 should be a schematization of the recurrent network of place cells in CA3 of the hippocampus, see Sections 3.3 and 3.5, in fact, we have not considered different constraints of biological nature that this type of network must meet, such as:

— take into account the border effects instead of simple periodic boundary conditions as they are crucial in spatial navigation to identify the presence of walls in an environment [26], also remember the presence of cells with this specific function, the border cells, see Section 3.8;

— make explicit the difference between excitatory and inhibitory synapses in the connectivity matrix according to Dale's principle [66], which basically states that a neuron performs the same chemical action at all of its synaptic connections to other cells, regardless of the identity of the target cell. In particular, it is known that the connections between place cells are only excitatory and the network activity is kept fixed thanks to the presence of the interneurons that are purely inhibitory and whose only function is to balance the positive contribution of place cells;

— consider more realistic models of place cells in which place fields can have different sizes [116, 163], take into account the presence of silent cells [231], not all place cells have a place field in a given environment, indeed only about thirty percent do, and also consider that a place cell can have more than one place field in a given environment, which is what usually happens when considering large maps [76, 193, 203];

— take into account that in the context of place cells and fields, it is known that neurons have some individuality, that is, retain some properties in the different environments. In particular it was reported experimentally [142, 203] that each place cell has its own propensity to have a place field per square meter: many neurons have very low propensity values, *i.e.*, have no place field at all in many maps, and few neurons that have very high propensity and therefore tend to code almost all maps even with more than one place field connected component per map;

— consider the sparse nature of synapses [104] and neuron activity observed in CA3 [9];

— consider that the positions to be stored in a map as well as the centers of the input (place) fields do not necessarily have to be distributed uniformly in the different environments. The latter is due to the lack of homogeneity of sensory inputs (visual, auditory, olfactory and so on) and the former due to the trajectories of the animal, which may prefer to spend more time in some areas than in others (reward areas) [41, 111, 147]. Linked to this is also the fact to consider that maps can be stored with different levels of spatial error.

It is therefore important to understand if the results found so far in Chapter 4 remain valid also in the case of the above mentioned variations of the starting model.

Moreover, since the theory presented in Section 4.6 has been developed with quenched place fields perhaps it is possible to propose some links with experiments in which we have access to place fields dispositions in different environments, maybe in order to better understand the phenomenon of random remapping, see Section 3.5.

In addition, another interesting topic to discuss is the learning dynamics of the network. In fact, so far we have always considered offline SVM algorithms (in which the network always has access to the whole data-set during the learning phase), instead of online procedures (in which instead patterns are presented one at a time) that are definitely more biologically plausible. This is a fundamental step in the perspective of wanting to study the dynamics of learning in a recurrent neural network that progressively matures continuous attractors and that could be connected to experiments on cognitive maps formation performed on new born animals [74].

Finally, we show an application of our setting in a different context of the place cells network, in particular to understand how to store continuous attractors in a recurrent neural network starting from real images [267].

6.2 BORDER EFFECTS

Until now, periodic boundary conditions (PBCs) have been considered for the maps in order to simplify the analytical calculations, *i.e.*, translation invariance, see Section 4.2. Here we try to discuss numerically what is the effect of removing PBCs in the model. In particular, now for environment we really consider D-dimensional cubes and not torus, the borders of the maps are rigid and any place field near the edge is a cut sphere according to the proximity from the walls.

In Fig. 30 we compare using SVMs the optimal stability κ defined in Eq. (4.3) as a function of the load $\alpha = \frac{L}{N}$ and of the number p of prescribed fixed points in case the maps do not have PBCs with the case they do.

From simulations it seems that the scaling found in the setting with PBCs is also valid if they are removed. However, at α and p fixed, κ increases slightly in the case PBCs are not present. This can be explained heuristically by noting that when we don't have PBCs the possible configurations of the patterns on the map borders decreases compared to the case of PBCs, so as p increases, the redundancy in the patterns starts to be seen first. Even if the result of the simulations seems quite intuitive, unfortunately an analytical treatment in case of removing PBCs becomes much more complicated.

The main problem is that removing the PBCs breaks the invariance under translation in the model 4.2 and therefore, in the computation with Gardner approach, see Section 4.5, things get more complicated. In theory we could compute numerically the same equations (4.39) but having to explicitly solve the integral on $\mathbf{r}_1$, that is not irrelevant any longer, and also the expression of the matrix (4.17), that is

$$\mathbf{\Gamma}_{\mu,\nu}\left(\{\hat{\mathbf{r}}_\mu\}\right) = \Gamma\left(\hat{\mathbf{r}}_\mu - \hat{\mathbf{r}}_\nu\right) - \phi_0^2 \,, \tag{6.1}$$

becomes more complicated since the latter is no longer a standard ERM because it is no longer a function only of the distance between the positions, but also of their center of mass since the invariance under translations is broken, what happens at the edge of the environment is different from what happens at the bulk.

The computation with quenched place fields in Section 4.6 does not seem possible as well in this case for the same reason. We do not know how to calculate the spectrum of

the matrices of single environment that are no longer ERM because of the presence of the edges, consequently we do not have access to the spectrum of the MERM

$$C_{jk}\left(\{\mathbf{r}_i^\ell\}\right) = \frac{1}{L}\sum_{\ell=1}^{L} \Gamma\left(|\mathbf{r}_j^\ell - \mathbf{r}_k^\ell|\right) , \tag{6.2}$$

see Eq. (4.46) and Chapter 5, and therefore to all the rest.

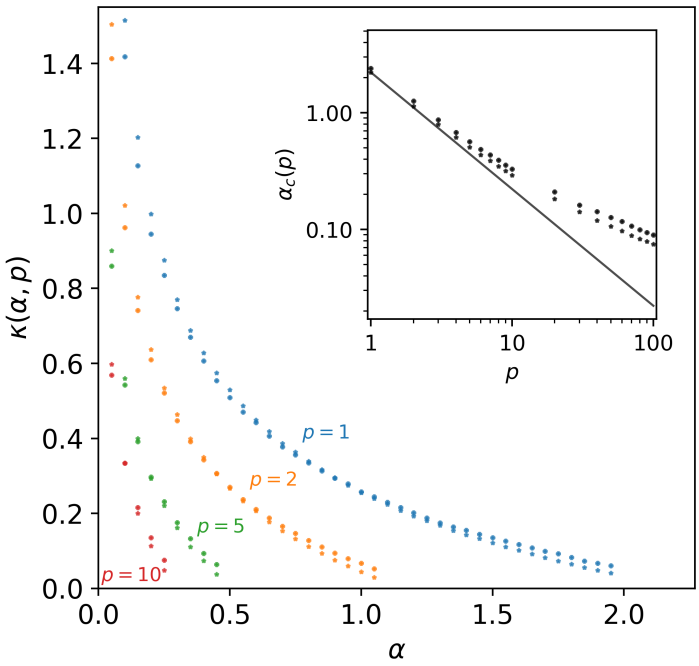

Figure 30 – Optimal stability κ as a function of the load α and the number p of positions obtained with SVMs. Stars, results with PBCs; crosses, results without PBCs. Parameter values: $D = 2$, $\phi_0 = 0.3$, $N = 1000$. Inset: $\alpha_c(p)$ decreases proportionally to $1/p$ (straight line) at low p, and much more slowly for large p. Stars indicate results from SVMs ($N = 5000$), averaged over 50 samples, with PBCs. Crosses indicate the same in the case without PBCs.

6.3 POSITIVE COUPLINGS CONSTRAINT

Here we study the effect of the constraint on the sign of the couplings in the model presented in Chapter 4, in fact, as it is well known in neuroscience the connections between pyramidal cells, place cells in our specific case, are only excitatory.

So in the following we generalize our model in order to take into account this constraint and explicitly considering threshold terms $\{\theta_i\}$ in the RNN that effectively represents the

inhibitory contribution coming from the network of interneurons that serve to keep the network activity fixed.

Somehow, what we are trying to do here is a generalization of the maximal stability perceptron approach with simple random patterns and sign-constrained synapses [24, 46, 47] to the case of patterns with strong spatial correlation.

In the case of random patterns it is known that the connectivity of a recurrent neural network with the constraint of positivity on the weights is quite sparse [46], *i.e.*, large fraction of zero synaptic weights ("potential" synapses), and compatible with the physiological distribution of synaptic connections observed in CA3 [104].

Moreover, always in the case of random patterns it is known that the capacity of a recurrent neural network with the constraint on the sign of the weights is half of the same quantity calculated for a not constrained RNN [15, 19, 50, 177, 245, 259]. What happens in the case of spatially correlated patterns?

Before starting the analytical treatment we try to understand numerically what the introduction of this constraint implies in our case. To do so we use a generalization of the SVM algorithm with the restriction of having only positive weights, the details of which are discussed in Appendix C.1.

6.3.1 *Couplings obtained by SVMs with positive weights constraint*

Hence, we report some qualitative features of the couplings obtained by SVMs with the positive couplings constraint, see Section 4.4 for a comparison with the unconstrained case.

— As shown in Fig. 31(a) and (c) the couplings W_{ij} are, also in this case, correlated with the distances $d^{\ell}_{ij} = |\mathbf{r}^{\ell}_i - \mathbf{r}^{\ell}_j|$ between the PF centers of the neurons i and j in the different maps ℓ. Note that, also here, the dependence on distance is less marked as the number L of maps increases, due to the interferences between the maps.

— In this case the couplings are always positive, so the role of the threshold $\theta_i < 0$ is fundamental to support bump states with an average activity ϕ_0.

— Histograms of the couplings in Fig. 31 (b) and (d) show that the amplitudes decay with N. Also in this case we get the average values and standard deviations to scale, respectively, as $1/N$ and $1/\sqrt{N}$. These results recall the ones already found in [24, 46, 47] for the distribution of synaptic weights in the case of random patterns and comparison with data in the cortex, suggesting also in this setting that synaptic connectivity in a RNN that should memorize continuous attractors (like the CA3 network in the hippocampus) may be optimized to store their number in a robust

fashion. Also here, in fact, it seems that a large fraction of weights is zero suggesting a rather sparse connectivity compatible with what found in [104].

It is worth noting that the weights are of the same order in module both in the standard case of SVM (histograms in Fig. 16 (b) and (d)) and the one where we introduce a constraint on the sign of the couplings (histograms in Fig. 31 (b) and (d)). So the contribution coming from the thresholds must compensate that part of the connections that cannot be negative in the second case. The bias terms however schematize the input coming from a population of inhibitory neurons (the interneurons) whose function is only to keep fixed the activity of the network, so it is reasonable to think that this contribution can compensate the fact that now we consider only excitatory neurons keeping the system biologically feasible, we are simply separating the role of the neurons of interest, that are the place cells (only excitatory), from the population of interneurons schematized by the thresholds.

6.3.2 *Stability obtained by SVMs with positive weights constraint*

In Fig. 32 we compare the optimal stability κ_{pos} defined as

$$\kappa_{pos} = \max_{\{W_{ij}\geqslant 0,\theta_i\}} \min_{\{i=1\dots N,\ell=1\dots L,\mu=1\dots p\}} \left\{ \left(2\sigma_i^{\ell,\mu}-1\right)\left[\sum_{j(\neq i)} W_{ij}\,\sigma_j^{\ell\mu}+\theta_i\right]\right\} \tag{6.3}$$

with the positive weights constraint and where the rows of the connectivity matrix are normalized as usual to unit, with the one in Eq. (4.3) of standard SVMs as a function of the load $\alpha = \frac{L}{N}$ and of the number p of prescribed fixed positions per map (in the case of SVMs with the weights sign constraint α has been multiplied by a factor of 2). It is clear that the numerically calculated curves coincide. This means that the capacity obtained in the constrained case turns out to be half of the unconstrained one.

This result is remarkable since the size of the space of possible couplings is strongly reduced with this restriction and we lose only a finite fraction in term of capacity.

6.3.3 *Adding the positive weights constraint in Gardner's framework*

After the numerical evidences just shown, here we are going to extend Gardner theory for the capacity of the perceptron with sign-constrained synapses [15] to the case of continuous attractors.

The substantial differences with the computation we have presented in Section 4.5 are the fact to consider the integrals over the weights $\{W_{ij}\}$ from 0 to ∞ instead of the whole $\mathbb{R}$ and also the introduction of the threshold terms $\{\theta_i\}$. Nevertheless we report below all the details.

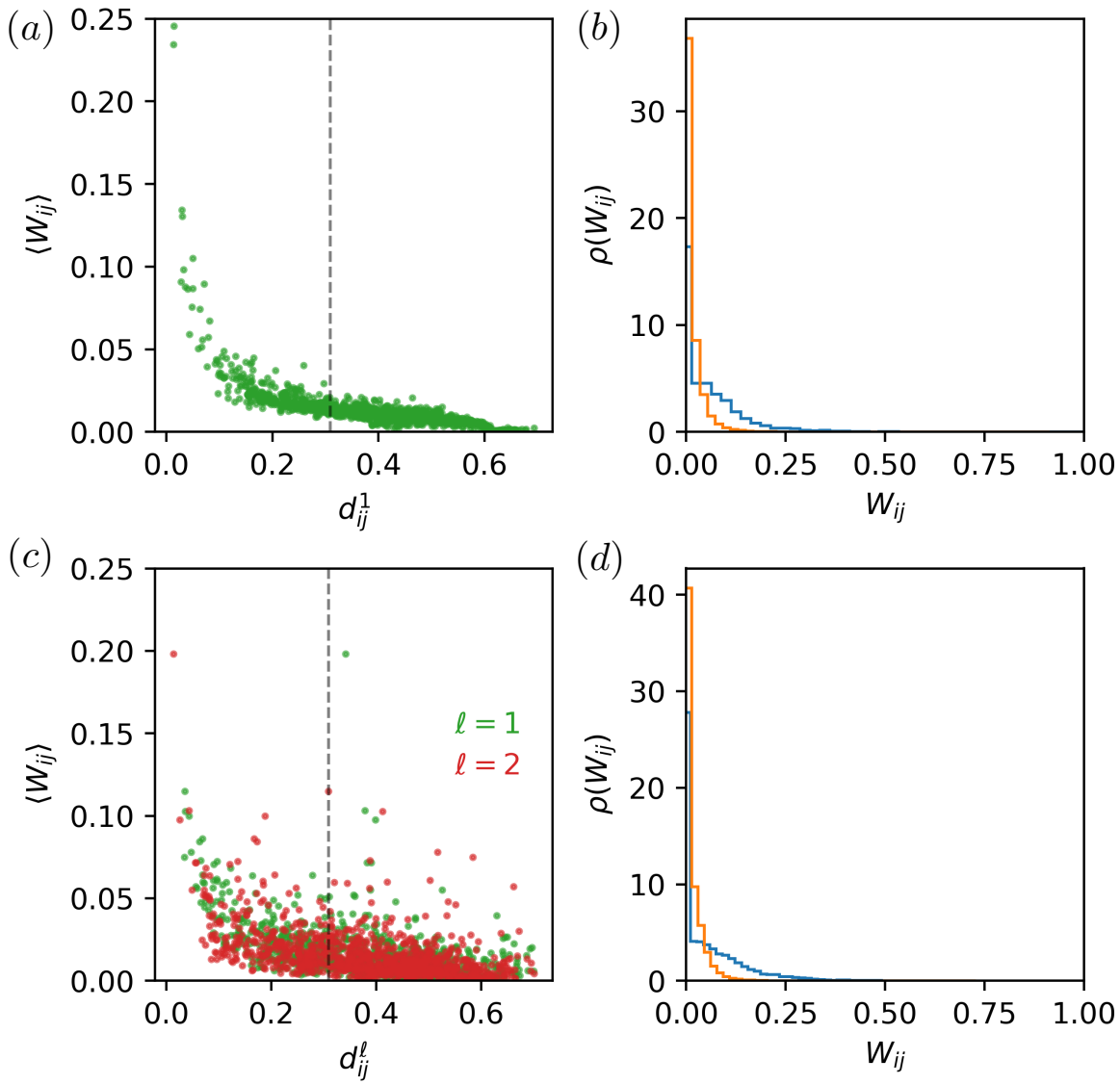

Figure 31 – Couplings obtained after training with SVM with positive weights constraint for L = 1 (a) and (b) and L = 2 (c) and (d) maps. Dependence of the average coupling with the distance between the corresponding neurons; the vertical line locates the radius r_c of the place fields. Averages were computed over 500 samples of the p positions per map at fixed PF centers; N = 1000 neurons. Histograms of the couplings, for sizes N = 100 (blue) and N = 1000 (orange) Parameters: D = 2, $\phi_0 = .3$ and $p \times L = N$.

As in Section 4.5, the training set consist of $p \times L$ binary patterns $\{\sigma_i^{\ell\mu}\}$ constructed by drawing randomly p positions in each of the L environments, defined in Section 4.2, so

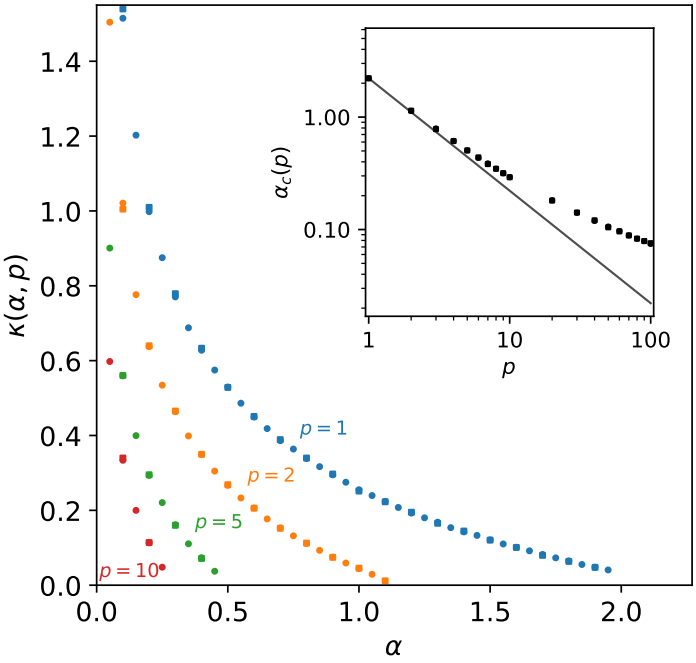

Figure 32 – Optimal stability κ as a function of the load α and the number p of positions obtained with SVMs. Dots, results with standard SVM of Sections 4.3 and A.1; squares, results with positive weights constraint SVM, see Section C.25, in which α has been multiplied by a factor of 2. Parameter values: $D = 2$, $\phi_0 = 0.3$, $N = 1000$, $Samples = 50$. Inset: $\alpha_c(p)$ decreases proportionally to $1/p$ (straight line) at low p, and much more slowly for large p. Here $N = 5000$.

that the resulting patterns are spatially correlated. The stability of the i component of the pattern that correspond to position μ in the environment ℓ is given by [1]

$$\Delta_i^{\ell\mu} = (2\sigma_i^{\ell\mu} - 1)\Big(\sum_{j(\neq i)} \frac{W_{ij}}{\sqrt{N}}\, \sigma_j^{\ell\mu} + \theta_i \Big), \tag{6.4}$$

where $W_{ij} \geqslant 0$ and θ_i is the threshold term. The training set is said to be stored if all the patterns have stabilities larger than some threshold $\kappa \geqslant 0$ [2].

1. To notice that here we normalize the rows of the connectivity matrix to N instead of 1 but we have scaled the weights of $\sqrt{N}$, therefore nothing changes regarding the computation presented in Section 4.5. Moreover, as we have seen in Section A.1, the normalization of weights is irrelevant in our setting and we can choose it arbitrarily.

2. From here until the end of the computation we will call κ what was previously called κ_{pos} to simplify the notation.

The volume in the space of couplings that corresponds to the admissible solutions of the storage problem, is

$$Z = \int_0^\infty \prod_{i\neq j}^{N} dW_{ij} \prod_i \delta\Big(\sum_{j(\neq i)} W_{ij}^2 - N\Big) \prod_{i,\ell,\mu} \Theta\Big((2\sigma_i^{\ell\mu}-1)\Big(\sum_{j(\neq i)} \frac{W_{ij}}{\sqrt{N}}\,\sigma_j^{\ell\mu} + \theta_i\Big) - \kappa\Big) \quad (6.5)$$

and is equal to the product of the N single-site volumes Z_i, with $i = 1\ldots, N$.

So we may focus for example on the volume associated with $i = 1$:

$$Z_1 = \int_0^\infty \prod_{j=2}^{N} dW_j\; \delta\Big(\sum_{j\geqslant 2} W_j^2 - N\Big) \prod_{\ell,\mu} \Theta\Big((2\sigma_1^{\ell\mu}-1)\Big(\sum_{j\geqslant 2} \frac{W_j}{\sqrt{N}}\,\sigma_j^{\ell\mu} + \theta\Big) - \kappa\Big), \quad (6.6)$$

where $W_j \equiv W_{1j}$ and $\theta \equiv \theta_1$. Using the replica method [51, 157, 242], we compute the average of $\log Z_1$ over the patterns. Introducing integral representations of the Heaviside functions and exploiting the statistical independence of the different maps, we write the average of the n^{th} power of the volume,

$$\langle Z_1^n \rangle = \int_0^\infty \prod_{j,a} dW_{ja} \prod_a \delta\Big(\sum_j W_{ja}^2 - N\Big)\, \chi^{\alpha N} \quad (6.7)$$

where $a = 1, \ldots, n$ is the replica index, and

$$\chi = \int \prod_{\mu=1}^{p} d\hat{\mathbf{r}}_\mu \int \prod_{j=1}^{N} d\mathbf{r}_j \int_\kappa^\infty \prod_{\mu,a} dt_{\mu a} \int_{-\infty}^{\infty} \prod_{\mu,a} \frac{d\hat{t}_{\mu a}}{2\pi} e^{i\sum_{\mu,a} \hat{t}_{\mu a}(t_{\mu a} - \theta(2\sigma_1^\mu - 1))} \times \prod_j e^{-\frac{i}{\sqrt{N}}\sum_{\mu,a} \hat{t}_{\mu a}(2\sigma_1^\mu - 1) W_{ja}\sigma_j^\mu}, \quad (6.8)$$

where $\hat{\mathbf{r}}_\mu$ denotes the p prescribed locations in the environment, and $\mathbf{r}_j$ the N PF of the neurons in the map, as in Section 4.5.

We first carry out explicitly the integrals over the PF with indices $j = 2, 3, \ldots, N$, leaving the integrals over $\mathbf{r}_1$ and all $\hat{\mathbf{r}}_\mu$ in χ. We then introduce, as in Section 4.5, the order parameters

$$m^a = \frac{1}{\sqrt{N}} \sum_{j\geqslant 2} W_{ja} \quad (6.9)$$

and

$$q^{ab} = \frac{1}{N} \sum_{j\geqslant 2} W_{ja} W_{jb}\,, \quad (6.10)$$

and rewrite $\langle Z_1^n \rangle$ as

$$\int_0^\infty \prod_{j,a} dW_{j,a} \int \prod_a \frac{d\hat{u}^a}{4\pi i} e^{\sum_a \frac{\hat{u}^a}{2}(N - \sum_j W_{ja}^2)} \int \prod_a \frac{d\hat{m}^a dm^a}{\frac{2\pi i}{\sqrt{N}}} e^{\sum_a \hat{m}^a(\sqrt{N} m^a - \sum_j W_{ja})} \times \int \prod_{a<b} \frac{d\hat{q}^{ab} dq^{ab}}{\frac{2\pi i}{N}} e^{\sum_{a<b} \hat{q}^{ab}(N q^{ab} - \sum_j W_{ja} W_{jb})} \chi^{\alpha N} \tag{6.11}$$

where we have used the integral representation of the Dirac-delta functions.

Still following what seen in Section 4.5, we consider $\Phi(\mathbf{r})$ to be the indicator function of the PF centered in $\mathbf{0}$: $\Phi = 1$ if $|\mathbf{r}| < r_c$, where r_c is the radius of the PF (with $\int d\mathbf{r}\, \Phi(\mathbf{r}) = \phi_0$), and 0 otherwise. Let $\Gamma(\mathbf{r}) = \int d\mathbf{r}' \Phi(\mathbf{r}')\, \Phi(\mathbf{r} - \mathbf{r}')$ be the correlation function of Φ and $\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) = 2\Phi(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) - 1$. Given p points $\hat{\mathbf{r}}_\mu$, $\mu = 1, \dots, p$ drawn uniformly at random in space, we define the $p \times p$ Euclidean random matrix with entries

$$\mathbf{\Gamma}_{\mu,\nu}\big(\hat{\mathcal{R}} \equiv \{\hat{\mathbf{r}}_\mu\}\big) = \Gamma\big(\hat{\mathbf{r}}_\mu - \hat{\mathbf{r}}_\nu\big) - \phi_0^2 \,. \tag{6.12}$$

We can then rewrite χ as

$$\chi = \int \prod_\mu d\hat{\mathbf{r}}_\mu \int d\mathbf{r}_1 \int_\kappa^\infty \prod_{\mu,a} \frac{dt_{\mu a}}{\sqrt{2\pi}} \int_{-\infty}^\infty \prod_{\mu,a} \frac{d\hat{t}_{\mu a}}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{\mu,\nu,a,b} q^{ab} \Gamma_{\mu,\nu}(\hat{\mathcal{R}}) \hat{t}_{\mu a} \hat{t}_{\nu b}} \times e^{-i\sum_{\mu,a}(m^a \phi_0 + \theta)\hat{t}_{\mu a}\, \sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu) + i\sum_{\mu,a} \hat{t}_{\mu a} t_{\mu a}} \,. \tag{6.13}$$

Due to translation invariance, the integral over $\mathbf{r}_1$ is irrelevant, and we can set $\mathbf{r}_1 = \mathbf{0}$.

We also define from (6.11):

$$Y \equiv \int_0^\infty \prod_{j,a} dW_{j,a}\, e^{\sum_a \frac{\hat{u}^a}{2}(N - \sum_j W_{ja}^2)} e^{\sum_a \hat{m}^a(\sqrt{N} m^a - \sum_j W_{ja})} e^{\sum_{a<b} \hat{q}^{ab}(N q^{ab} - \sum_j W_{ja} W_{jb})} \,. \tag{6.14}$$

It is possible now to make the RS Ansatz (expected to be valid also here since the domain of suitable couplings is still convex) on the structure of the order parameters and their conjugate variables.

We can therefore write within the limit of large N and small n:

$$\frac{1}{nN} \log Y = \frac{\hat{u}}{2} + \frac{q\hat{q}}{2} + \lim_{n\to 0} \frac{1}{n} \log \int_0^\infty \prod_a dW_a\, e^{-\frac{\hat{u}}{2}\sum_a W_a^2 + \hat{q}\sum_{a<b} W_a W_b - \hat{m}\sum_a W_a} \tag{6.15}$$

where we have changed $\hat{q}$ in $-\hat{q}$ for simplicity. After having inserted in the exponent of the above expression $(\sum_a W_a)^2$ and used a Gaussian integral trick, is it possible to solve the integral, obtaining in the small n limit:

$$\frac{1}{nN}\log Y = \frac{\hat{u}}{2} + \frac{q\hat{q}}{2} + \frac{1}{2}\log\frac{\pi}{2} - \frac{1}{2}\log(\hat{q}+\hat{u}) + \frac{\hat{m}^2+\hat{q}}{2(\hat{q}+\hat{u})} + \int Dz \log \operatorname{erfc}\frac{\hat{m}-\sqrt{\hat{q}}z}{\sqrt{2}\sqrt{\hat{q}+\hat{u}}}\,, \quad (6.16)$$

where Dz is the Gaussian measure and $\operatorname{erfc} y = \frac{2}{\sqrt{\pi}}\int_y^\infty dx\, e^{-x^2}$.

We shall now concentrate on the saddle-point equations to determine $\hat{m}$, $\hat{u}$ and $\hat{q}$ as a function of q. We shall assume that as $q \to 1$, so when the space of solutions reduces to the optimal coupling matrix, $\frac{\hat{q}}{\hat{q}+\hat{u}}$ diverges and $\frac{\hat{m}}{\sqrt{\hat{q}}} \to 0$, and verify that the solution satisfies this condition. In this limit it is possible to solve $\int Dz \log \operatorname{erfc}\frac{\hat{m}-\sqrt{\hat{q}}z}{\sqrt{2}\sqrt{\hat{q}+\hat{u}}}$ because when $z < 0$ we can use the following asymptotic expansion $\operatorname{erfc} y \simeq \frac{e^{-y^2}}{\sqrt{\pi}y}$, and when $z > 0$ we get trivially a constant. In the end we can write:

$$\frac{1}{nN}\log Y = \frac{\hat{u}}{2} + \frac{q\hat{q}}{2} - \frac{1}{4}\log(\hat{q}+\hat{u}) + \frac{\hat{m}^2+\hat{q}}{4(\hat{q}+\hat{u})} - \frac{1}{4}\log(\hat{q}) + \text{const}\,. \quad (6.17)$$

It is possible then to write the saddle-point equations relative respectively to $\hat{m}$, $\hat{u}$ and $\hat{q}$ from the above equation:

$$\hat{m} = 0\,, \quad (6.18)$$

$$\frac{1}{2} = \frac{\hat{m}^2+2\hat{q}+\hat{u}}{4(\hat{q}+\hat{u})^2}\,, \quad (6.19)$$

$$\frac{q}{2} = \frac{1}{4\hat{q}} + \frac{\hat{m}^2+\hat{q}}{4(\hat{q}+\hat{u})^2}\,. \quad (6.20)$$

By solving this equations we finally find:

$$\hat{m} = 0\,, \quad (6.21)$$

$$\hat{q} = \frac{-2+3q+\sqrt{-4+q(4+q)}}{4(1-q)^2}\,, \quad (6.22)$$

$$\hat{u} = \frac{4+q^2-q(6+\sqrt{-4+q(4+q)})}{4(1-q)^2}\,, \quad (6.23)$$

so that our assumptions are indeed consistent. By injecting these results in (6.17) we find that when $q \to 1$, we have:

$$\frac{1}{nN}\log Y \simeq \frac{1}{4(1-q)}\,. \quad (6.24)$$

As for the term χ this can be rewritten in the RS ansatz and using a Gaussian integral trick as:

$$\chi = \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu\right)}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}} \left\{ \int_\kappa^\infty \prod_\mu dt_\mu \int_{-\infty}^{\infty} \prod_\mu \frac{d\hat{t}_\mu}{2\pi} \right. \tag{6.25}$$
$$\left. \times e^{-\frac{1}{2}(1-q)\sum_{\mu,\nu} \hat{t}_\mu \Gamma_{\mu,\nu}(\hat{\mathcal{R}}) \hat{t}_\nu} e^{i\sum_\mu \hat{t}_\mu (z_\mu \sqrt{q} + t_\mu - (m\phi_0 + \theta)\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu))} \right\}^n .$$

After performing the Gaussian integral in $\{\hat{t}_\mu\}$, taking the small n limit and the $q \to 1$ limit we get:

$$\frac{\log \chi}{n} \simeq -\frac{1}{2(1-q)} \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu}}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}} \min_{\{t_\mu \geqslant \kappa\}} \times \tag{6.26}$$
$$\sum_{\mu,\nu} \left[t_\mu - (z_\mu + (m\phi_0 + \theta)\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu)\right] \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} \left[t_\nu - (z_\nu + (m\phi_0 + \theta)\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\nu)\right] .$$

Putting all together we obtain:

$$\frac{\langle Z_1^n \rangle - 1}{nN} \simeq \frac{1}{2(1-q)} \left\{ \frac{1}{2} - \alpha \int \prod_\mu d\hat{\mathbf{r}}_\mu \int \prod_\mu \frac{dz_\mu}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}\sum_{\mu,\nu} z_\mu \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} z_\nu}}{\sqrt{\det \Gamma(\hat{\mathcal{R}})}} \min_{\{t_\mu \geqslant \kappa\}} \times \right. \tag{6.27}$$
$$\left. \sum_{\mu,\nu} \left[t_\mu - (z_\mu + (m\phi_0 + \theta)\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\mu)\right] \Gamma(\hat{\mathcal{R}})^{-1}_{\mu,\nu} \left[t_\nu - (z_\nu + (m\phi_0 + \theta)\sigma(\mathbf{r}_1 - \hat{\mathbf{r}}_\nu)\right] \right\} .$$

After absorbing into m either ϕ_0, either θ we finally obtain the expression for the critical capacity $\alpha_c(\kappa, p) = \max_m \alpha_c(m; \kappa, p)$, where $\alpha_c(m; \kappa, p)$ is the load α cancelling the terms inside the curly brackets in (6.27).

By comparing the last expression with the one found in Eq. (4.38) it is therefore clear that in this case the capacity is then only half of the one in which we do not consider the restriction on the sign of weights, see Section 4.5.

It is important to note that since the introduction of this constraint brings in the final equations only the difference of a factor 2, the results found in the theory with quenched place fields, see Section 4.6, remains valid also in this setting and therefore the scaling in Eq. (4.75) it is still preserved, despite the introduction of this strict constraint of biological nature. In other words this restriction limits the capacity of a finite factor not causing any problem to the theory developed so far.

6.4 VARIANTS OF THE PLACE CELL MODEL

In order to see how much the scaling of $\alpha_c(p)$ in Eq. (4.75) is robust against the choice of the parametrization $\Phi(\mathbf{r})$ of the manifolds, we are going to show that reducing the number of active neurons in each map, allowing for variations in the sizes of the PFs from neuron to neuron and consider place cells with more than one spatial connected components per map do not affect it.

The function Γ we have considered so far corresponds to the simple model defined in Fig. 25. In a unit cube $\mathcal{H}_D$ in D dimensions with periodic boundary conditions, a set of N positions $\mathbf{r}_i^\ell$ (centres of D-dimensional spheres, place fields, of volume $\phi_0 < 1$) are drawn uniformly and independently at random for each map ℓ. The term $\Gamma\left(\left|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell\right|\right)$ entering in the correlation matrix (5.2) is simply the overlap (common volume) between the two place fields in the same environment, see Fig. 25. We consider below three variants of this model of interest for a more realistic modeling of the CA3 recurrent place cells network.

6.4.1 *Dilution*

Let us first consider single-space ERM in which a fraction ρ_0 of the N place cells (chosen at random among $1, \dots, N$) carry vanishingly small place fields, and the remaining cells have normally place fields of volume ϕ_0 whose centers are randomly located on the map, see Fig. 33(a). All the entries of the ERM $M_{ij}^{(1)} = \Gamma(|\mathbf{r}_i - \mathbf{r}_j|)/N$ such that i or j belongs to the first subset (with point-like PFs) are equal to zero. We are left with a block matrix of dimension $(1-\rho_0)N \times (1-\rho_0)N$, equal to the ERMs considered so far with the model of Fig. 25, see Chapter 5. As a consequence, in the large N limit, the eigenvalues of this block-ERM are equal to $\rho_0\, \hat{\Gamma}(\mathbf{k})$, while the remaining eigenvalues are equal to zero.

The resolvent of this diluted version of ERM in the high-density regime has the same form as (5.9):

$$s_1(z) = -\frac{1}{zN}\left(\rho_0 N + \sum_{\ell=1}^{\infty} \sum_{\substack{\mathbf{k}\neq\mathbf{0} \\ (|\mathbf{k}|\leqslant\rho_0 N)}} \hat{\Gamma}(\mathbf{k})^\ell \, \frac{1}{z^\ell} + (1-\rho_0)N\right) = -\frac{1}{z} - \frac{1}{N\,z}\,\gamma\left(\frac{1}{z}\right) \tag{6.28}$$

where

$$\gamma(u) = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{u\,\rho_0 \hat{\Gamma}(\mathbf{k})}{1 - u\,\rho_0 \hat{\Gamma}(\mathbf{k})}\,. \tag{6.29}$$

The computation of the functional inverse of the resolvent (blue function) of the dilute MERM can be done as in the standard case, see Chapter 5, and we get:

$$z = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\alpha\,\rho_0\hat{\Gamma}(\mathbf{k})}{\alpha + s\,\rho_0\hat{\Gamma}(\mathbf{k})} - \frac{1}{s}\,. \tag{6.30}$$

We can now solve equation (6.30) in order to get the density of eigenvalues. The agreement with the spectrum obtained from numerical simulations is excellent, see Fig. 33(d).

Place cells with vanishingly small place fields serve to model the fact that not all place cells take part in the coding of a given environment (a neuron that presents vanishingly small PF can be seen as a silent cell because it is never active). The hippocampus has about 30M neurons in humans and .3M neurons in rodents, so it is reasonable to think that to effectively store an environment we do not need to use all the cells available, in fact in the experiments it is clear that about 40% of the cells measured in the hippocampus are actually place cells, *i.e.*, they have at least one PF in the environment [231].

6.4.2 *Place fields of different volumes*

We now discuss the case of a multinomial distribution of place fields volumes. We consider first that, in each space, a fraction ρ_1 of the N PFs have volume ϕ_1, while the remaining fraction $\rho_2 = 1 - \rho_1$ have volume ϕ_2, see Fig. 33(b). For every space we build a matrix composed of 4 blocks:

$$\mathbf{M}^{(1)} = \frac{1}{N}\left(\begin{array}{c|c} \Gamma_{11} & \Gamma_{12} \\ \hline \Gamma_{21} & \Gamma_{22} \end{array}\right)\,, \tag{6.31}$$

where the block Γ_{ab} is a $\rho_a N \times \rho_b N$ ERM depending on the overlaps between PFs of volumes ϕ_a and ϕ_b, and with a, b taking values 1 or 2.

We look for eigenvectors of $\mathbf{M}^{(1)}$ of components $v_i(\mathbf{k}) \propto e^{i\,2\pi\,\mathbf{k}\cdot\mathbf{r}_i}$ multiplied by α_a for the site i in the fraction ρ_a, with $a = 1, 2$. We obtain the following eigen-system:

$$\begin{cases} \rho_1\,\hat{\Gamma}_{11}(\mathbf{k})\;\alpha_1 + \rho_2\,\hat{\Gamma}_{12}(\mathbf{k})\;\alpha_2 = \lambda(\mathbf{k})\;\alpha_1 \\ \rho_1\,\hat{\Gamma}_{21}(\mathbf{k})\;\alpha_1 + \rho_2\,\hat{\Gamma}_{22}(\mathbf{k})\;\alpha_2 = \lambda(\mathbf{k})\;\alpha_2 \end{cases}\,. \tag{6.32}$$

In the system above $\hat{\Gamma}_{ab}(\mathbf{k}) = \hat{\gamma}_a(\mathbf{k})\hat{\gamma}_b(\mathbf{k})$ with a, b taking value 1 or 2 and

$$\hat{\gamma}_a(\mathbf{k}) = \int_{\mathcal{H}_D} d\mathbf{r}\,\gamma_a(\mathbf{r})\,e^{-i2\pi\mathbf{k}\cdot\mathbf{r}} \tag{6.33}$$

with $\gamma_a(\mathbf{r})$ being the indicator function of the place field of volume ϕ_a [3]. We find $\alpha_a \propto \hat{\gamma}_a(\mathbf{k})$ and $\lambda(\mathbf{k}) = \rho_1(\hat{\gamma}_1(\mathbf{k}))^2 + \rho_2(\hat{\gamma}_2(\mathbf{k}))^2$.

This result immediately extends to more than two PFs types. If we have K finite (as $N \to \infty$) types of PFs, with associated volumes ϕ_a and fractions ρ_a, with $a = 1, \dots, K$, the eigenvalue of ERM attached to the momemtum $\mathbf{k}$ is given by

$$\lambda(\mathbf{k}) = \sum_{a=1}^{K} \rho_a \, (\hat{\gamma}(\mathbf{k}))^2 \, . \tag{6.34}$$

It is straightforward to write the resulting self-consistent equation for the MERM resolvent by simply changing $\hat{\Gamma}(\mathbf{k}) \to \sum_{a=1}^{K} \rho_a \, (\hat{\gamma}_{aa}(\mathbf{k}))^2$ in Eq. (5.21). In Fig. 33(e) we show the perfect agreement of this theoretical result with numerical simulations.

Place cells with PFs of different volumes are particularly important in the study of heterogeneous environments, *i.e.*, environments in which there are areas of interest (food, water, etc.) and areas of disinterest for the animal that stores it. Typically the areas of interest have a higher density of PFs with a small size, while the areas of disinterest have less PFs but with a larger size [41]. Obviously the size of PFs determines the accuracy with which different parts of the environment are stored, as they determine the width of the activity bump.

6.4.3 *Multiple place fields per cell in each space*

We extend the above setting to the case of multiple place fields per cell in each space. More precisely, we assume that for each place cell $i = 1, \dots, N$, there are c centers $\mathbf{r}^{\ell}_{i,m}$ of PFs, with $m = 1, \dots, c$ in each space ℓ, see Fig. 33(c); we assume that c remains finite as N, L are sent to infinity. The associated MERM is defined as follows

$$C_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \sum_{m,m'=1}^{c} \Gamma\left(\left|\mathbf{r}^{\ell}_{i,m} - \mathbf{r}^{\ell}_{j,m'}\right|\right) \, . \tag{6.35}$$

To better understand what happens in this setting we consider the limit case of a single map:

$$M^{(1)}_{ij} = \frac{1}{N} \sum_{m,m'=1}^{c} \Gamma\left(\left|\mathbf{r}_{i,m} - \mathbf{r}_{j,m'}\right|\right) \, . \tag{6.36}$$

3. $\gamma_a(\mathbf{r})$ is the same as $\Phi(\mathbf{r})$ in Section 4.5 in the case where all the PFs of the different neurons have the same volume in all the maps.

In the high-density regime the eigenvectors of this ERM have components approximately equal to $v_i(\mathbf{k}) \propto \sum_m e^{i\,2\pi\,\mathbf{k}\cdot\mathbf{r}_{i,m}}$, with eigenvalues given by $c\,\hat{\Gamma}(\mathbf{k})$ (for $\mathbf{k} \neq \mathbf{0}$).

The only change to the functional inverse of the correspondent MERM resolvent is $\hat{\Gamma}(\mathbf{k}) \to c\,\hat{\Gamma}(\mathbf{k})$, so that we obtain:

$$z = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\alpha\, c\,\hat{\Gamma}(\mathbf{k})}{\alpha + s\, c\,\hat{\Gamma}(\mathbf{k})} - \frac{1}{s}\,. \tag{6.37}$$

We have solved equation (6.37) in order to get the density of eigenvalues; results are in excellent agreement with numerics, see Fig. 33(f).

Place cells that have multiple PFs serve to model large environments in which this phenomenon is observed [76, 193, 203]. Since the locations in an environment are stored by the entire population of place cells encoding it, having multiple PFs is not harmful as long as there are no problems in decoding the different positions and it may be more convenient to avoid unnecessary cells that are silent in a given map having to take part in encoding it. One is used to think that a place cell can only have one PF in an environment because typically the experiments were done in small rooms where it was not possible to observe this quite common feature. Moreover, the fact that place cells can have more than one PFs in a given environment is not related to grid cells because PFs have centers placed randomly and not on a triangular grid covering the whole environment, see Section 3.8.

6.4.4 *Putting everything together*

As already discussed, in the CA3 region of the hippocampus, neurons may have place fields in some environment and none in other maps, which corresponds to the dilute model presented above. In addition, we have introduced other variants, in which the radius of place fields varies or a place field is made of more than one connected spatial component, as seen in large environments [203]. While the variants of the model considered here lead to different densities of eigenvalues z, the behaviours of these densities for $z \to 0$ and $\alpha \to 0$ seem qualitatively robust, which suggests that the storage capacity of recurrent neural networks is a robust property of the space-to-neural activity encoding.

In fact, the quenched place field theory in these cases extends in a trivial way[4] with respect to the computation presented in Section 4.6, indeed the only irrelevant differences in these situations are the average activity of the network that is different and the spectrum of the associated MERMs which do not change its behaviour for small eigenvalues

4. The new kind of disorders introduced here always enter the correlation matrix that defines the order parameters of the single environments, which also remains in the ERM class. Therefore the picture presented in Fig. 22 remains unchanged in the large p limit where self-averaging properties still apply.

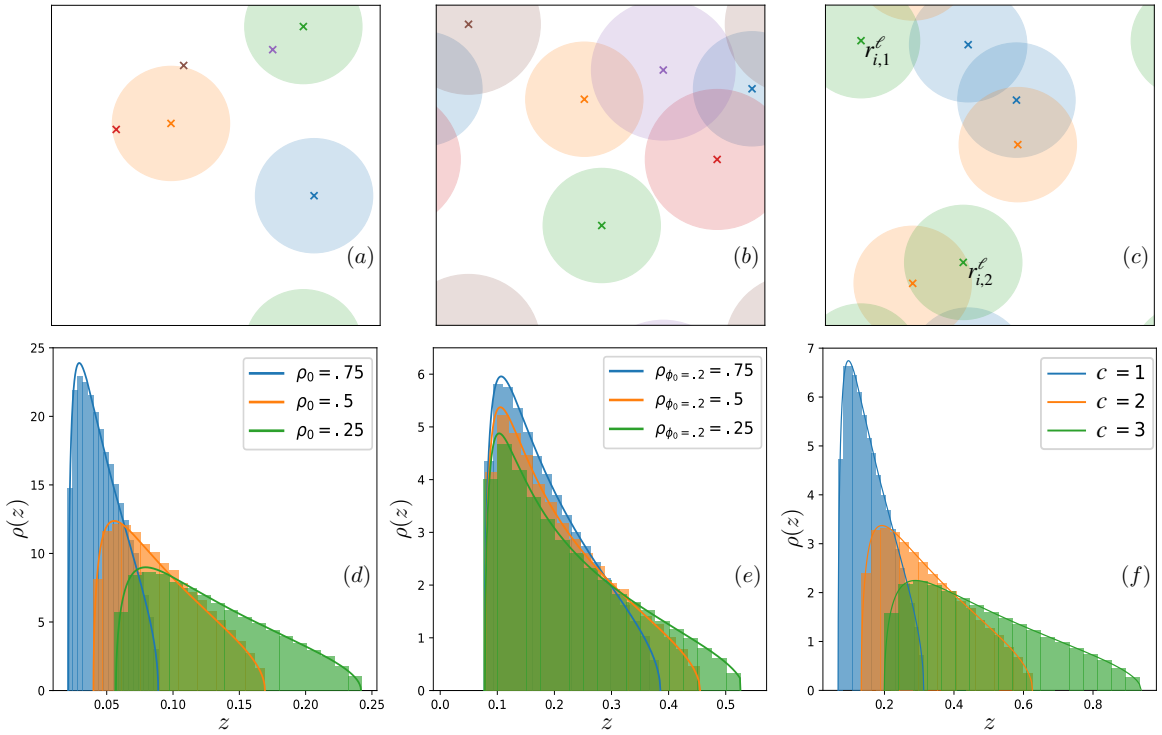

Figure 33 – Top panels: sketches of the model variants, respectively (a) dilution, (b) place fields of different sizes and (c) multiple place fields per cell in a given map. Bottom panels: (d) Density of eigenvalues for the MERM for $\phi_0 = .2$ with different dilution fractions ρ_0. (e) Density of eigenvalues of MERM with different fractions $\rho_{\phi_0=.2}$ of PF with volume $\phi_0 = .2$ and $\rho_{\phi_0=.4}$ of PF with volume $\phi_0 = .4$ in each space. (f) Density of eigenvalues of MERM for $\phi_0 = .2$ with $c = 2$ PF for each index i in each map. Parameters: $N = 2500, D = 1, \alpha = 1$. In all cases we do not show the extensive eigenvalue.

so that the scaling in Eq. (4.75) remains valid even in these contexts of greater biological relevance.

6.5 INDIVIDUALITY OF NEURONS

Until now we have always considered that the different place cells in our recurrent neural network all behave the same way, in the sense that they all have the same propensity to have place fields in different maps. As the experiments in [142, 203] show, however, it doesn't seem to be properly so, in fact each place cells looks to have its own propensity to have a place field per square meter and in particular this propensity seems to be maintained in different environments. In the following we try to extend our model taking this

aspect into consideration together with the variants of the place cells model presented in the previous Section 6.4

In particular, here we want to study the consequences of non-independence between the elementary ERMs composing the MERM on the density of eigenvalues taking into account a biologically plausible statistics for PF propensities that will automatically imply in an elegant way the constraint on the sparsity in the activity of the neurons observed in CA3 [9].

Moreover, since in this case the elements in the same position of the ERMs corresponding to different environments are correlated, we can no longer use the free probability theory approach to derive the spectral properties of the associated MERM (we no longer have the property of asymptotic freeness), but fortunately the calculation can be performed in this case using the replica method, see Sections C.2.1 and C.2.3.

Technically, in the following we generalize the computations presented in Section 4.6 and 5.3 to consider these more biologically relevant settings and also this time try to understand if the scaling found in (4.75) remains preserved even in case of taking into account the individuality of the place cells.

Let's consider therefore a simple schematization of what said above, that is to consider a network of N neurons that are divided in $M = O(1)$ groups with $\beta_\rho N$ neurons each and $\sum_{\rho=1}^{M} \beta_\rho = 1$. Neurons belonging to the same group have c_ρ PFs per map of area $\phi_{\rho,m}$ with $m = 1, \dots, c_\rho$. In this way we explicitly consider that there are neurons that have many place fields (of various areas) in the different maps and others that have few or none.

The MERM that we have to take in consideration in this setting is the following:

$$C_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \sum_{m=1}^{c_i} \sum_{m'=1}^{c_j} \Gamma_{(i,m),(j,m')} \left(\left| \mathbf{r}^{\ell}_{i,m} - \mathbf{r}^{\ell}_{j,m'} \right| \right) , \tag{6.38}$$

where $\Gamma_{(i,m),(j,m')} \left(\left| \mathbf{r}^{\ell}_{i,m} - \mathbf{r}^{\ell}_{j,m'} \right| \right)$ represents the overlap in the map ℓ of the spatial components m and m' associated with place cells i and j with areas that depend on the group of neurons to which they belong.

Before proceeding to the generalization of the quenched place field theory it is necessary to calculate the resolvent of (6.38), the details of the computation can be found in

Appendices C.2.1 and C.2.3 and putting together these results we obtain the following system of equations that allows us to find this resolvent:

$$\begin{cases} g(U) = \sum_{\rho=1}^{M} q_\rho \ , \\ U = -\frac{\beta_\rho}{q_\rho} + \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\alpha \sum_{m=1}^{c_\rho} \hat{\Gamma}(\mathbf{k})_{\rho,m}}{\alpha + \sum_{\nu=1}^{M} q_\nu \sum_{m'=1}^{c_\theta} \hat{\Gamma}(\mathbf{k})_{\theta,m'}}, \ \forall \rho \ , \end{cases} \tag{6.39}$$

where we defined an overlap q_ρ for each group of neurons, $\rho = 1, \dots, M$.

Now using this result and following closely Appendices C.2.2 and C.2.4 we can generalize to this case the theory presented in Section 4.6 and obtain finally the asymptotic scaling of the critical capacity in the large p limit:

$$\alpha_c(p) \sim \frac{A(D)}{(\log p)^D} \Big(\sum_{\rho=1}^{M} \frac{\beta_\rho}{\sum_{m=1}^{c_\rho} \phi_{\rho,m}^{1-\frac{1}{D}}} \Big)^D \qquad (p \to \infty) \ . \tag{6.40}$$

It is interesting to note that in case $D = 1$ the critical capacity does not seem to depend on the size of the PFs and also that we find in the limit cases trivially the results of 4.6, C.2.2 and C.2.4.

6.6 NON UNIFORM DISTRIBUTION OF POSITIONS

Another trivial but at the same time important generalization of our theory is to consider the case of variable spatial error from map to map. In fact, as already discussed in Section 4.1, it is legitimate to think of storing in a RNN different environments with different levels of accuracy depending on the needs and importance (some maps or some areas in a map may be more important to remember for an animal as they may correspond to points of interest such as home, food or a reward in general).

This can be done simply by substituting p with its average value over the maps in Eq. (4.75) fo the model presented in Chapter 4 and suggests that the fraction of maps with finest spatial resolution ϵ should not exceed $\sim \epsilon^D$ when $\epsilon \to 0$, in order not to affect too much the critical capacity, thus keeping unchanged the results found so far.

Moreover, while we have assumed in Chapter 4, for the sake of simplicity, that the distribution of positions was statistically uniform across space, there is no need for this to be so in practice. Experiments have shown that spatial representations of environments are enriched in PFs close to spots of interests (such as water pots [111] or objects [41]) with respect to void regions. We report in the following numerical simulations showing that increasing the density of prescribed positions in regions of the physical space allows us to carve specific attractors in the neural activity space, representing preferentially those

regions. This result is compatible with recent studies establishing the link between PFs distribution and behavioral place preference [147].

More precisely, up to now we have considered for simplicity that the p positions were drawn uniformly at random to produce statistically homogeneous maps, *i.e.*, without preferred positions. It is straightforward to extend this setting to the case of heterogeneous densities of prescribed positions.

Figure 34 shows the spatial distribution of stabilities for homogeneously scattered points (a) and an heterogeneous repartition of points, densely packed in a subregion (diagonal) (b). In the latter case, the strong heterogeneity in the local distribution of stabilities will favor the location of the bump along the zones with a major density of positions. As a consequence, a 1D-attractor is effectively built in the $D = 2$-dimensional space.

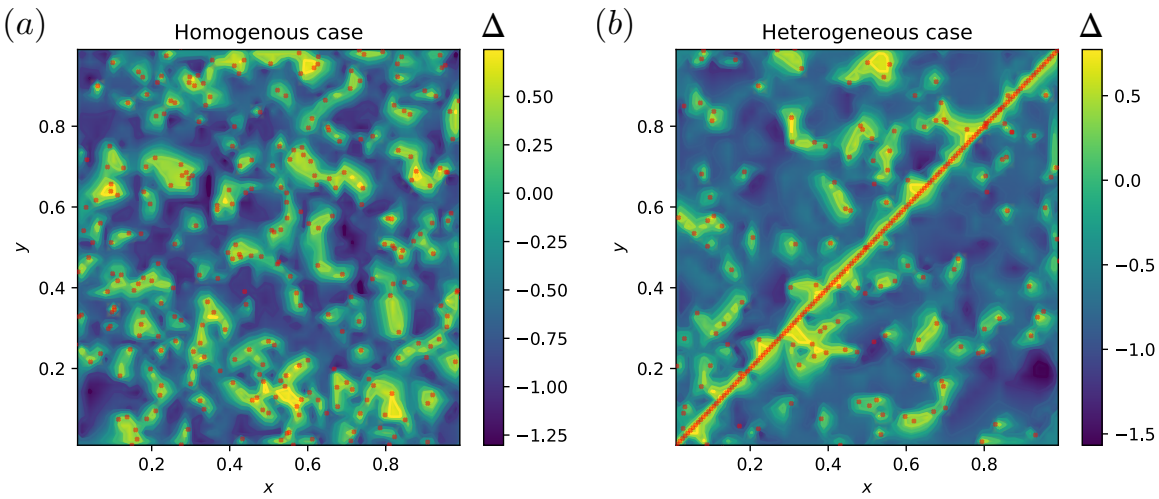

Figure 34 – Distribution of local stabilities after the learning of a map with SVMs. (a) Homogeneous case: the p positions of the data-set are drawn randomly. (b) Heterogeneous case: here the data-set has 150 positions on the diagonal of the maps and the other 150 positions are drawn at random. Here, $D = 2$, $\phi_0 = .3$, $N = 1000$, $p = 300$ and $L = 1$. We show the contour map made from 2500 realization of random positions, for which we evaluate the stabilities of the corresponding patterns. The overall network stabilities (minimal pattern stabilities) in the homogeneous and heterogeneous cases are, respectively, $\kappa \simeq .44$ and $\kappa \simeq .53$ for the samples considered here.

In order to better understand this mechanism we show an example of Monte Carlo simulation `SmallKappaL1Hetero.mkv` [5] that corresponds to the heterogeneous distribution

5. The mentioned example can be viewed at: https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.124.048302. As in the videos presented in Section 4.4, also here the red crosses represent the

of positions shown in Fig. 34(b) and that was obtained with the same parameters as in `SmallKappaL1.mp4`, see Section 4.4, but with 150 out of the $p = 300$ positions drawn on the diagonal of the map; the stability of the network was $\kappa = .5$.

Actually in this case it is expected that the dynamics of the network is favored to move along the diagonal of the map, *i.e.*, the favorite patterns are those that correspond to have active place cells associated with PFs near the diagonal of the map, or in other words PFs such that if we calculate their center of mass (center of the bump) that corresponds to the position decoded at a given time, this is near the diagonal. This happens because the positions to be stored in the map are concentrated on the diagonal (heterogeneous distribution of positions) and after the learning process they have typically high stability values compared to the a priori not stored positions, see Fig. 34(b). Since the dynamics of the network is favored towards patterns that have a high stability value, the bump is confined to move mostly on the diagonal, in this sense we say that we have learned a 1D attractor from a 2D map. Moreover, considering that patterns coding for positions near the diagonal are favoured in dynamics, the activity of place cells whose PFs are near the diagonal is higher [6].

Although the results of numerical simulations seem reasonable, also here, as in Section 6.2 for the periodic boundary conditions, an analytical treatment in which we consider that the positions inside the maps are not homogeneously distributed becomes much more complicated for the same reason, *i.e.*, lack of invariance under translations that technically makes hard the calculation of the ERM spectrum of single environment [191].

6.7 COMPARISON BETWEEN SVM AND THEORETICAL COUPLINGS

Interestingly, the quenched PF theory developed in Section 4.6 can be applied to any particular set of PFs, not necessarily homogeneously distributed over space; knowledge of the PF characteristics, *i.e.*, from experimental measurements, allows us to determine the multispace correlation matrix $\mathcal{C}$ defined in Eq. (4.46) and to make specific predictions. We show here a proof of principle of this approach where we compare the couplings found with SVMs and the ones given by our quenched PF theory on synthetic data, see Section 4.2.

Figure 35 shows how the couplings $\{W_{ij}\}$ depend on the size of the network, N, and of the distances between the PF centers of neurons i, j in the maps. We generally find that the couplings W_{ij} obtained by SVMs and the "thermal" averages $[W_{ij}]$ predicted by the

stored positions of the different environments in the connectivity matrix of the neural network and the blue circles represent the centres of the PFs whose corresponding neurons are active at a fixed time (the PFs whose neurons are silent at a certain time are not shown).

6. Maybe from the video it doesn't seem so because PFs cover 30% of the area of the environment so even place cells with PFs quite far from the diagonal can be active.

quenched PF theory, see Section 4.6, for a fixed set of PF centers are in excellent agreement, see equations (23) and (24) in [165] for details on the calculation of the average couplings and associated standard deviations.

— Both sets of couplings have mean values scaling as $\frac{1}{N}$ and standard deviations scaling as $\frac{1}{\sqrt{N}}$.

— The sign of couplings depend on the distance between their PF centers in the maps. We get excitatory couplings for distances up to the radius of the PFs ($r_c = \sqrt{\frac{\Phi_0}{\pi}}$ in $D = 2$), and inhibitory interactions for larger distances.

Furthermore, a direct application of our quenched PF theory could be to investigate in detail the phenomenon of global random remapping. During all this work we have always assumed random remapping between cognitive maps associated to different environments, *i.e.*, orthogonal maps in which a given place cell presents place fields in different environments in totally random positions without any correlation with other maps. It is difficult to think that this is completely true especially in the case of two maps associated to very similar environments, see for example [204]. It might therefore be interesting to compare the theoretical MERM spectrum assuming random remapping with a numerical spectrum built from real data for example from the Alme et al. [9] experiment and see if there are really differences in the spectrum compared to the assumption of totally random place fields. A technical problem in this sense would be that the numbers presented for example in the above mentioned experiment are very far from the thermodynamic limit with which the theory was developed, in fact, only about $N = 30$ place cells are measured simultaneously in the different rats that have memorized only $L = 11$ environments, so it will be crucial to understand the finite size effects while waiting for better data. In fact, thanks to new generation electrodes (neuropixels) or calcium imaging techniques it is now possible to simultaneously record the activity of thousands of neurons for long periods of time and therefore it is in principle possible to estimate the correlation matrix from PFs measured in different environments and reach L and N values in accordance with the thermodynamic limit with which the theory was developed.

6.8 DYNAMICS OF LEARNING

Another extremely interesting topic in which here we limit ourselves to providing a first step, is the study of the learning dynamics of the network model presented in Chapter 4. More precisely the dynamics of the weights of a recurrent neural network such that starting from discrete attractors, *i.e.*, maps in which few positions have been stored, progressively we get to store continuous attractors, *i.e.*, increasing the number of memorized

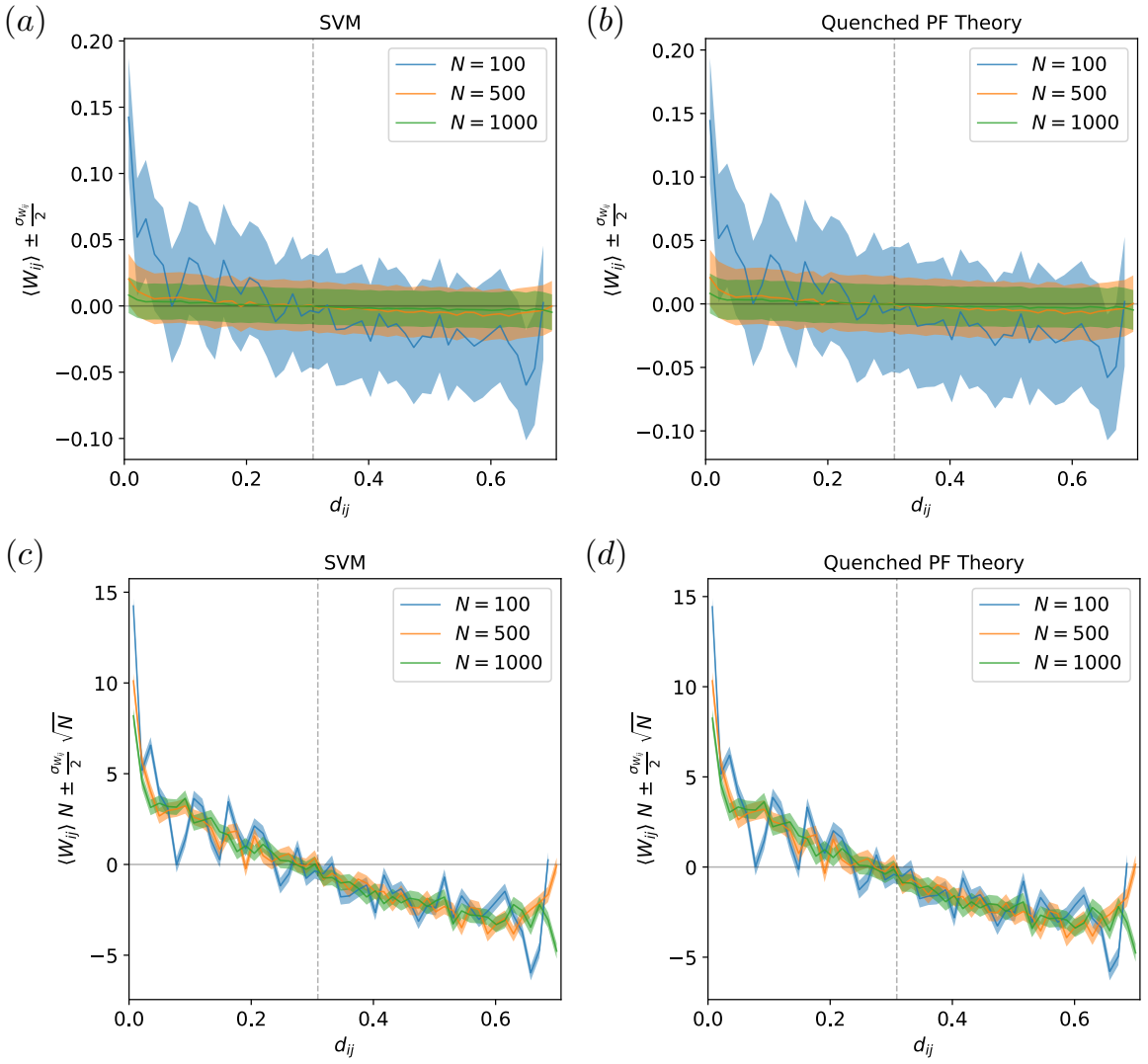

Figure 35 – Comparison of couplings obtained with SVM (a) and (c) and with the Quenched PF Theory (b) and (d). Top: Dependence of couplings on N. Bottom: Dependence of the couplings on distance; The vertical line locates the radius r_c of the PF. These results were obtained for $D = 2$, $\phi_0 = .3$, $\alpha = .1$, $p = 5$; we have averaged over 100 different realizations of the p positions at fixed PF centers for the SVM results. Space was divided in 50 bins with values ranging from 0 to $\sqrt{2}/2$ (the maximal distance achievable in unit square with periodic boundary conditions). Couplings were then put in the corresponding bins for all maps, and the averages and standard deviations were plotted as functions of the bin centers. Average couplings and associated standard deviations with quenched PF theory were computed with (23) and (24) of [165], with the substitution $\alpha_c \to p\,\alpha_c$ as the number of patterns is here $p \times L$.

positions per map thus decreasing sufficiently the spatial error with which the maps are encoded.

In particular here we show that the number of presentations of the patterns needed to stabilize a map is approximately proportional to p.

In fact, the SVM algorithm exposed in Sections 4.3 and A.1 is offline: all the patterns are available to the learning procedure at all times, which is not biologically realistic. As a preliminary attempt to understand how maps are learned, we have implemented an online learning scheme, which is an adaptive version of the perceptron algorithm (adatron) [21], see Appendix C.3 for details. In this procedure, patterns are presented one after the other. We may choose the order of presentation, as well as the learning rate η. In order to study the time needed for the algorithm to store a map, we have run the online learning scheme in the simplest case of a single environment ($L = 1$). We have monitored the stability of the network during this learning phase as a function of the number of training rounds, a round corresponding here to the presentation of all the patterns in the data-set. In Fig. 36, we show that the number of rounds needed to stabilize a map is roughly proportional to p/η (for a fixed ratio p/N).

It would be interesting to relate this finding to biological results. Let us remark that the presentation of repeated rounds considered here could be realistic for an animal exploring the same environment several times, in particular a 1D corridor in which the sequence of visual inputs remains roughly unchanged from one exploration to another. Experiments show that changes in the environment (insertion of one object) lead to the production of a new representation, which is stabilized over 4-5 explorations, see [41].

Moreover, it would be extremely important to study plausible learning rules that could ultimately elucidate how the network progressively maturates to account for more and more fixed points and eventually defines a quasicontinuous attractor, as seems to be the case during the first weeks of development in rodents [74]. In fact, the authors showed that when a newborn rat need to learn a new environment, initially it seems that its representation of the map is more similar to a discrete attractor (a map where few positions are stored), but as the days pass it looks like that the attractors become progressively more continuous (this is due to the fact of starting to store more positions in the same map so as to reduce the spatial error with which the map is stored).

Finally, another important study with practical application would be the analysis of the diffusion coefficient of the bump of activity (which can be measured experimentally, see [225]) to tie it with the number of positions stored in a map and also understand the time it takes to store these positions online. From measurements of the activity bump during sleep of a rodent that has previously explored an environment we may be able to link these quantities and have access to a deeper mechanism of learning the maps themselves.

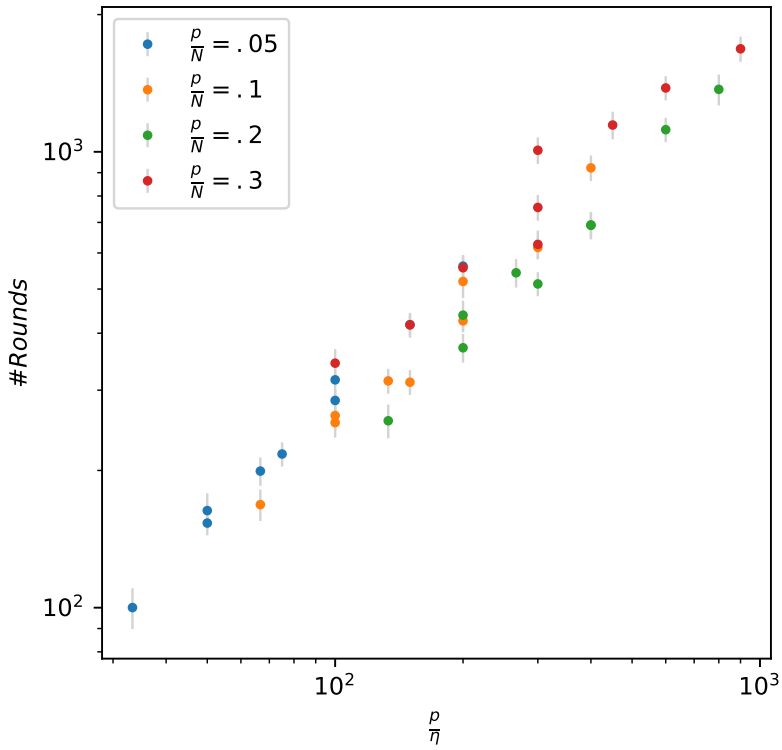

Figure 36 – Number of rounds required for adatron to reach the same stability value obtained with the standard SVM algorithm vs. the ratio $\frac{p}{\eta}$ at fixed $\frac{p}{N}$. The values of N range between 1000, 1500 and 2000, while the ones of η between .5, 1 and 1.5. The other parameters are set to $L = 1$, $D = 1$, $\phi_0 = .3$, $Samples = 100$.

6.9 LEARNING CONTINUOUS ATTRACTORS IN RNN FROM REAL IMAGES

In this Section we present an application of the model presented in Chapter 4 in a different context than seen so far, that is how to memorize continuous attractors in a recurrent neural network starting from real images, inspired by the work of Zou et al. [267]. In particular the authors of [267] have proposed a biologically plausible scheme of a neural system that stores continuous attractors based on two fundamental steps.

— The first is how to generate high-level representations of objects such that the correlations between the neural representations reflect the semantic representations between them, using for this purpose a deep neural network trained on a large number of real images [218].

— The second is how to learn correlated patterns in an RNN, using an orthogonal Hebb rule [224], where as usual the representations between the neural patterns are encoded in the weights of the network.

In the end they were able to show with this procedure that if the images presented to the network are linked by a continuous feature a continuous attractor is efficiently stored in a RNN.

In the following we present preliminary results in which instead of the orthogonal Hebb rule we use SVM learning, see Sections 4.3 and A.1, to memorize correlated patterns.

Moreover, besides showing results for a single continuous attractor stored in a RNN, as in [267], we will try also to understand if it is possible to store more than one in the same connectivity matrix (in this case two).

Let's start by discussing what we have done in detail. So far we have always learned continuous attractors starting from ad-hoc patterns, in the sense that by construction, see Section 4.2, near positions in an environment had similar neural configurations (with a large overlap), while positions corresponding to different environments have neural configurations with almost no overlap (this is due to the random remapping between the different maps).

If we consider now the case of real images this situation is not so immediate, consider for example the data shown in Fig. 37 where there are photos of objects (a guitar and a sax) on a black background and rotated by different angles, in particular we have an image per angle, then $p = 360$ data per object[7].

As already noted by Zou et al. [267] however, representations of images of the same category in different positions, maybe rotated by a certain angle, can be more different (low overlap between the pixel vectors associated with the corresponding images) than representations of images belonging to different categories. This is due to the fact that the initial representations do not have a semantic meaning but only a spatial one, so for example two guitars rotated by 90 degrees can have almost no overlap. It is therefore important to extract from the data high-level representations so that data belonging to the same category always have very correlated representations while data belonging to different categories always have quite orthogonal representations, as it should be.

Following closely [267], in order to extract these representations with a semantic meaning we use the convolutional deep neural network VGG16 [218] that has been trained on a very large number of real images mimicking somehow the dorsal visual pathway

7. Actually the images that we will use in the following have a lower resolution that is $224 \times 224 \times 3$ pixels, the factor 3 represents the RGB color scale, and where the values of the different pixels are normalized between 0 and 1 so that to fit the input of the deep neural network that we are going to use [218].

Figure 37 – Examples of the images that we are going to use for the learning phase. (a) Images of a guitar rotated by different angles on a black background, in particular we will have a figure for each angle from 1 to 360 degrees. (b) The same of (a) but for images of a sax.

[261]. In particular the representations of the images obtained on the deepest layer before the classifier contain exactly what we are looking for. In the case of the specific network VGG16 these high-level representations have size $N = 4096$.

So now we present the images of the data-set in Fig. 37 to VGG16 extracting all the high-level representations. Using our standard notation we have therefore patterns of size $N = 4096$ corresponding to two maps $L = 2$, one for the guitar and one for the sax, where we have a certain level of spatial resolution given by the number of data per map, $p = 360$, one image per angle, and we would like to store these data in a recurrent neural network with the procedure used so far with SVMs, see Sections 4.3 and A.1 [8].

8. A technical problem is that the representations extracted from VGG16 have real non negative components while in our procedure we need binary patterns, so we had to binarize the data by hand hoping not to lose too much information with this operation.

Once built the data-set we can perform the learning with SVMs and then try to study the dynamics of the associated recurrent neural network with the finite temperature scheme defined in Section 4.4.

First of all we show in Fig. 38 the results of the network dynamics after learning a single map, the one associated with the guitar, in which we display the findings in the bi-dimensional space corresponding to the first two principal components extracted from the training data with PCA [257].

It is interesting to note that although we projected the data in a two-dimensional space these seem to be arranged on a circle, so a one-dimensional manifold, this makes sense since the different guitars differ from each other only by one variable (an angle).

We start the dynamics from an initial configuration corresponding to one of the initial images and set a temperature of the order of the network stability. If we look at the evolution of the network state in the space where we projected the training data, it appears that this remains confined to the manifold defined by the data themselves, as it should be, see Fig. 38. Moreover, it seems as well that close points in the manifold correspond to guitars rotated more or less of the same angle.

Even more interesting is the case of storing two manifolds in an RNN at the same time, considering the data-set with both guitar and sax images. Also here we show after the learning phase with SVMs the results of Monte Carlo simulations at finite temperature. The first important thing to note in the examples that will follow is that after projecting the data of the two objects on the first two principal components obtained from PCA carried out on all the data together, the data corresponding to different classes are clustered and this shows that therefore the high-level representations obtained from VGG16 are able to distinguish the different categories going to capture the semantic meaning of the different objects.

In addition, in the videos `Tsmall.mp4` and `Thigh.mp4`[9] we show two examples of the dynamics initialized from the same image but starting from different temperatures. In the

9. The video of the simulation at lower temperature can be seen at: https://www.dropbox.com/s/e14acuw075k3ggi/T_small.mp4, while the one corresponding to higher temperature can be seen at: https://www.dropbox.com/s/7kds8vs4g1lhx7d/T_high.mp4. In the above videos we present in the left panel the data projections in the two-dimensional space of the first two principal components extracted from the entire data-set. The blue dots correspond to the images of the guitars while the orange dots correspond to the images of the sax, the red star instead corresponds to the state of the network projected on the same space that evolves over time, a frame of the videos corresponds to a Monte Carlo sweep. While the central and right panels show the normalized overlap of the network state at a certain time with respect to the images of the data-set as a function of the angle of rotation of the guitar and the sax respectively. The red star present at every time in one between the central and right panels corresponds to the data-set image with greater overlap with respect to the current network state, identifying the location of the bump in one of the two maps.

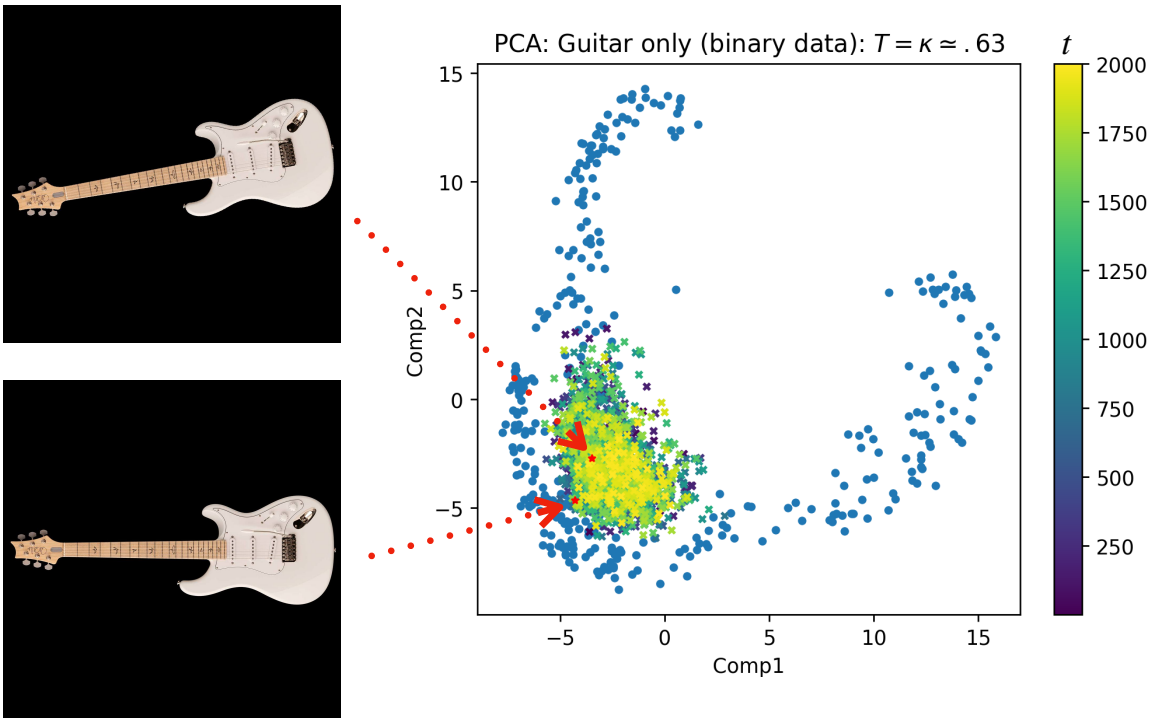

Figure 38 – Study of the dynamics of a RNN after learning through SVMs the high-level data representations extracted from VGG16 and associated with the different images of guitars rotated by different angles. The blue dots represent the data on which the learning was done projected on the first two principal components extracted from the data with PCA. The photos associated with the red stars correspond to 2 images of the data-set, in particular it seems that close points in this space correspond to guitars rotated almost of the same angle. In addition, the crosses represent the dynamics of the network starting from one of the images of the data-set and in which we have a temperature of the order of the stability of the network. The colors of the crosses go from dark purple to yellow with the passing of time of the dynamics, in particular we have a cross every Monte Carlo sweep, see Section 4.4.

case of the simulation at lower temperature the dynamics remains confined to diffuse on the starting manifold, while in the example with a little higher temperature in addition to seeing diffusion on the individual manifolds we see as well transitions between them as we already saw in Section 4.4.

So in the end it seems, at least from this preliminary analysis, that our method is able to store continuous attractors in RNN even in more generic situations instead of the artificial data studied in the rest of the thesis. Moreover, we believe that this method allows also here to overcome the issues presented in Section 3.7, provided that we have extracted good representations of the data to store.

7

CONCLUSIONS

In this Chapter we sum up the work that has been done in this manuscript by summarizing the main results obtained and by giving a cue for future lines of research.

Let us begin by summarizing the results of this thesis.

Certainly the main finding of this work has been to show how to achieve optimal storage of multiple continuous attractors in recurrent neural networks together with the study of the optimal trade-off between capacity and spatial resolution, that is, how the requirement of higher spatial resolution affects the maximal number of attractors that can be stored, thus overcoming the theoretical problems left unsolved in the last twenty years, see Section 3.7.

Using a combination of state-of-the-art statistical physics and random matrix theory tools, we showed that the capacity decreases very slowly with the spatial error, more precisely as the inverse D^{th} power of the logarithm of the spatial error, where D is the dimension of the manifolds, see Section 4.6.

This non trivial scaling proves that recurrent neural networks are very efficient memory devices capable of storing many continuous attractors at high resolution.

Moreover, also if the motivations of this work come from the field of computational neuroscience, in particular how to model the CA3 recurrent network in the hippocampus, this setting gave us the opportunity to generalize Gardner's theory, a milestone in statistical physics of disordered systems, to the case of patterns with strong spatial correlation. The study of data with structure is nowadays a very hot topic instead of considering in theoretical approaches simple random patterns, not only in the field of neuroscience but also the one of machine learning [94].

In addition, in this work we have introduced a novel statistical ensemble for Euclidean random matrices (ERM), see Chapter 5, where the element i, j of the matrix depend on the distances between representative points of i and j in more than one space. Using a combination of heuristic assumptions, analytical and numerical calculation, we have shown that the high-density limit is non trivial when the number L of spaces and the size N of the matrix are sent to infinity, with a fixed ratio $\alpha = L/N$. We have analytically studied the density of eigenvalues of this Multiple-space–ERM (MERM) ensemble, based on free-probability identities and on the replica method.

Even though we have introduced MERM in a specific setting we hope that this ensemble of random matrices will find applications and be of interest in other fields, *i.e.*, in applied mathematics or in information theory. In particular, our results could be used for functions Γ with a dependence on the pairwise distances different from the ones considered in this work.

Finally, as we began to discuss in Chapter 6, there are several lines of research to pursue in the near future based on these results.

— An important aspect that we have not considered in this work is the extension to recurrent neural networks with continuous neurons, for example threshold linear [212, 234], instead of simple binary units. A generalization in this sense would not only be useful in view of more biologically plausible models, that would also allow to investigate another type of remapping seen in other regions of the hippocampus as CA1, namely rate remapping [84], but also for example in the application we have shown in Section 6.9 about learning continuous attractors from real images where high-level representations extracted from deep neural networks have usually positive continuous components.

— Moreover, from a theoretical point of view, it would be very interesting to try to understand how to solve the problem of breaking the invariance under translation in the calculation of ERM spectrum in the high-density limit, with consequent applications in our model to the study analytically of maps without periodic boundary conditions (Section 6.2), heterogeneous distributions of positions (Section 6.6) and non-uniform distributions of the PFs in the different maps.

— Taking into account the fact that the different maps are not independent but can have correlations both in terms of PFs centers and positions is certainly an important problem to study, see [204]. Numerically it is easy to implement, just generate maps that are properly correlated with each other, for example assuming that only a fraction of PFs remap or that the different maps are generated from a common one in which the PFs centers are slightly shifted with a displacement extracted from a Gaussian distribution centered in zero and with a variance smaller than the diameter of the PFs. It is also possible to generate positions in different maps on similar trajectories, but this adds correlation between the patterns in different maps only if they have correlated PFs. In general it makes sense to assume maps with correlated PFs in order to store similar environments (with similar external inputs such as visual, auditory, etc.) and also to assume similar trajectories in similar maps as there may be common points of interest where an animal is interested to go (such as food or water). Analytically we are now working on this problem for a better understanding of the phenomenon of random remapping that is not clear is completely random especially in the case of very similar environments, see Section

6.7. Our theory is developed at quenched PFs and this allows us in principle to consider more complicated cases of random PFs provided that we can analytically calculate the spectral properties of the associated MERM (which is not trivial when taking into account correlations). Nevertheless, introducing correlations in general means to have patterns even more similar to each other compared to the case studied in the thesis and this can only increase the capacity of the network (with patterns that contain less information content due to these redundant representations).

— Furthermore, always making reference to how much seen in Section 6.7, it could be interesting to qualitatively compare the theoretical weights with the synapses measured experimentally in CA3 [104], in particular to try to understand if the real connectivity matrix is more compatible with the fact to have the connections proportional to the correlation matrix between the patterns (Hebb rule) or, as the quenched PF theory predicts, see in particular the Eqs. (23) and (24) in [165], as the inverse of the correlation matrix (optimal connectivity).

— A crucial aspect that we have started to discuss in Section 6.8 is the development of a biologically plausible learning rule for the study of the learning dynamics of a RNN, that could be compared with experimental data of newborn animals that progressively mature continuous attractors starting from discrete ones [74]. In Section 6.8 we started to study an online algorithm but still far from being compared with experimental data. This kind of study could be useful in order to answer questions such as how much time is needed to mature a continuous attractor, what is a reasonable value to choose for the learning rate, what is a reasonable initial configuration of the weights to choose, how should the environment be explored from an animal to memorize it more efficiently, how much does previous knowledge of other environments help to memorize a new one and so on. Moreover, in this context we can try to answer a question about the fact that when an animal is facing the learning of a new environment that can be similar to an old environment, it should decide whether to learn the new environment as a new map, or learn it as a reinforcement of an old map. A key ingredient in this direction can be found in [7] where the authors show how an attention modulator (coming from the dentate gyrus which is an area external to the hippocampus) allows to discriminate between similar maps depending on whether or not it is behaviorally important for the animal.

— It would also be extremely interesting to study models in which the propensities of place cells are stochastic and not deterministic as in Section 6.5, in particular referring to the salt and pepper distribution introduced in [142] by Lee et al..

— More generally, our quenched PF theory (Section 4.6) relies essentially on the activity-activity correlation matrix:

$$\mathcal{C}_{ij} = \frac{1}{L}\sum_{\ell=1}^{L} \langle \Phi_i(\mathbf{r})\, \Phi_j(\mathbf{r}) \rangle_{\mathbf{r}}^{(\ell)} \,, \tag{7.1}$$

where the average $\langle\cdot\rangle_{\mathbf{r}}^{(\ell)}$ is taken over the random positions $\mathbf{r}$ of points in the manifold corresponding to map ℓ. What matters in the parametrization of the manifolds and determines the asymtotics of the capacity is the behaviour of the density of eigenvalues of $\mathcal{C}$, $\rho(\lambda)$, for eigenvalues $\lambda \to 0$. It would be very interesting to classify the possible behaviours of ρ and obtain the corresponding scalings of $\alpha_c(p)$ for example, considering more complicated geometries for the environments than simple cubes, considering in the maps additional dimensions beyond spatial ones, such as variables corresponding to olfactory or auditory stimuli, or considering more general contexts for continuous attractors and not restrict ourselves to cognitive maps as in [58].

— We could also extend the model to study the grid cells, see Section 3.8. To do this we should assume that each neuron has fields that are arranged on a triangular grid instead of having one or a few fields per environment, so a position can activate a specific neuron whenever it is inside one of the fields associated with the grid cell. Following [223] we could also study the critical capacity of a model of this type in which the weights are not chosen in Hebbian way and try to understand the optimal trade-off between capacity and spatial resolution. While in the case of place cells this analysis is more motivated because it is known that animals store many maps associated with different environments and contexts, in the case of grid cells this is not so obvious as it seems at the moment that they define a single map (a single continuous attractor) and then it may not be necessary to study RNN that stores an extensive number of continuous attractors based on grid cells, but certainly an interesting theoretical problem to study.

— A very important aspect that we have not discussed in this thesis is the effect of the θ rhythm in the hippocampus which is considered essential for memory formation and navigation [149, 205, 251]. In biologically plausible models of RNN this rhythm is usually modeled with the introduction of an oscillating external field that acts on the dynamics of the neurons, it would be interesting to see how the dynamical properties of our model depend on the introduction of external fields like the latter.

— Finally it would be crucial as well to study in our model both the effect of a possible noise during the learning phase [258], as we have developed both simulations and theory at zero temperature and from a biological point of view it is a great simplification, and to consider at the same time network dynamics with learning dynamics

(*i.e*, to consider dynamical synapses) [80, 249], since we have always studied the dynamical properties of RNNs after the learning of the connectivity matrix (which remains quenched during the network dynamics).

From this long to-do list it is clear that there is still a lot of research work to be done, particularly from the point of view of the biological plausibility of our model. Nevertheless, our results allow to give possible theoretical explanations to the fact that animals in real life (and not in the laboratory) are very able to navigate many environments with great accuracy, since they can in principle store in the CA3 network of their hippocampus a large number of cognitive maps with a good level of spatial resolution. This statement, although clear from an experimental point of view, was difficult to show in terms of standard theories on multiple continuous attractor neural networks, so we hope that our results have clarified this fundamental point.

To conclude, it is important to note that throughout the work we have focused on studying a model of the CA3 network without considering its interaction with other relevant areas of the brain. For example, as discussed in Sections 3.3 and 3.8, in the brain it is present the path integrator (PI), capable of integrating proprioceptive, vestibular and visual flow inputs and possibly supported by the grid-cell network in the medial-enthorinal cortex (mEC), that allows the animal to update the neural representation during navigation. It lacks therefore a generalization of our model in order to understand how contextual and PI inputs are combined by the hippocampal network to produce cognitive maps and accurate positional encoding [196].

APPENDIX-CHAPTER 4

The purpose of this Appendix is to facilitate the reading of Chapter 4. Indeed, we report here some details about the support vector machine algorithm, numerical simulations and also the calculations, both for Gardner's approach and for the quenched input (place) field theory. In addition, we propose again some results already presented in Chapter 4 but changing the model parameters in order to show the robustness of our findings.

A.1 SUPPORT VECTOR MACHINE LEARNING

As seen in Section 4.3 the problem of learning the connectivity matrix in our RNN can be decomposed into N independent problems of support vector machine (SVM) with linear kernel and hard margin [39, 62, 210, 211, 244]. So let's focus now on one of the SVMs, say the one with output neuron i, and discuss the algorithm in detail.

We begin with the two-class classification problem, see Fig. 15(b), of the form

$$h_i = \sum_{j \neq i} W_{ij} \sigma_j \,, \tag{A.1}$$

where $\{\sigma_j\}$ are the input components of the SVM, that is a (N-1)-dimensional vector considering that $j \neq i$, taking binary values 0 or 1. While $\{W_{ij}\}$ are all connections that arrive on neuron i and are a priori real numbers (either positive, negative or null), see Fig. 15(a).

The training data-set, see Section 4.2, comprises $p \times L$ input binary vectors of components $\{\sigma_j^{\ell,\mu}\} \in \{0,1\}$, $\ell = 1,\dots,L$, $\mu = 1,\dots,p$ and $j \neq i$, with corresponding target values $2\{\sigma_i^{\ell,\mu}\} - 1 \in \{1,-1\}$. A new data point is classified according to the sign of h_i and we shall assume for the moment that the training data-set is linearly separable in the input space, so that by definition there exists at least one choice of the $(N-1)$-dimensional vector $\mathbf{W}_i$ (decision boundary) such that a function of the form (A.1) satisfies $h_i^{\ell,\mu} = \sum_{j \neq i} W_{ij} \sigma_j^{\ell,\mu} > 0$ for points having $2\,\sigma_i^{\ell,\mu} - 1 = 1$ and $h_i^{\ell,\mu} < 0$ for points having $2\,\sigma_i^{\ell,\mu} - 1 = -1$, so that the stabilities $\Delta_i^{\ell,\mu} \equiv (2\,\sigma_i^{\ell,\mu} - 1) h_i^{\ell,\mu} > 0$ are positive for all the training data points.

There may of course exist many such solutions that separate the classes exactly, like the ones that can be found from the perceptron algorithm, see Section 4.3. Moreover, as we have already discussed (Section 4.3), among all the possible solutions to this problem that

allow us to correctly classify the training data-set, a good idea is to choose the one with the smallest generalization error (the hyperplane that maximizes the distance between classes). The SVM approaches this problem through the concept of the margin, which is defined to be the smallest distance between the decision boundary and any of the samples. In SVMs the decision boundary is then chosen to be the one for which the margin is maximized, as illustrated in Fig. 15(b).

The perpendicular distance of an input pattern from an hyperplane defined by $h_i = 0$, where h_i takes the form (A.1), is given by $\frac{|h_i^{\ell,\mu}|}{\sqrt{\sum_{j\neq i} W_{ij}^2}}$, see Fig. 15(b). Furthermore, we are only concerned in solutions for which all input patterns are correctly classified, so that $\Delta_i^{\ell,\mu} > 0 \; \forall \, \ell, \mu$. Thus the distance of an input pattern to the hyperplane is given by

$$\frac{(2\sigma_i^{\ell,\mu} - 1)h_i^{\ell,\mu}}{\sqrt{\sum_{j\neq i} W_{ij}^2}} = \frac{(2\sigma_i^{\ell,\mu} - 1)\sum_{j\neq i} W_{ij}\sigma_j^{\ell,\mu}}{\sqrt{\sum_{j\neq i} W_{ij}^2}} \, . \tag{A.2}$$

The margin we are interested in is determined from the distance to the closest input pattern of the data-set, and we wish to optimize the vector $\mathbf{W}_i$ in order to maximize this distance. Hence the maximum margin solution is found by solving

$$\max_{\mathbf{W}_i} \left\{ \frac{1}{\sqrt{\sum_{j\neq i} W_{ij}^2}} \min_{\ell,\mu} \left[(2\sigma_i^{\ell,\mu} - 1) \sum_{j\neq i} W_{ij}\sigma_j^{\ell,\mu} \right] \right\}, \tag{A.3}$$

where we have taken the factor $\frac{1}{\sqrt{\sum_{j\neq i} W_{ij}^2}}$ outside the optimization over ℓ, μ because $\mathbf{W}_i$ does not depend on them.

Direct solution of this optimization problem would be very complex, and so we shall convert it into an equivalent problem that is much easier to solve. To do this we note that if we make the rescaling $\mathbf{W}_i \to \kappa_i \mathbf{W}_i$, then the distance from any input point to the decision surface, given by $\frac{(2\sigma_i^{\ell,\mu}-1)h_i^{\ell,\mu}}{\sqrt{\sum_{j\neq i} W_{ij}^2}}$, is unchanged. We can use this freedom to set

$$(2\sigma_i^{\ell,\mu} - 1) \sum_{j\neq i} W_{ij}\sigma_j^{\ell,\mu} = 1 \tag{A.4}$$

for the input data that is closest to the surface. In this case, all data points satisfy the constraints

$$(2\sigma_i^{\ell,\mu} - 1) \sum_{j\neq i} W_{ij}\sigma_j^{\ell,\mu} \geqslant 1, \quad \forall \, \ell, \mu \, . \tag{A.5}$$

This is known as the canonical representation of the decision hyperplane. In the case of input data points for which the equality holds, the constraints are said to be active, whereas for the remainder they are said to be inactive. By definition, there is always at least one active constraint, because there is always a closest point, and once the margin has been maximized there are at least two active constraints. The optimization problem then simply requires that we maximize $\frac{1}{\sqrt{\sum_{j\neq i} W_{ij}^2}}$, which is equivalent to minimize $\sum_{j\neq i} W_{ij}^2$, and in the end we have to solve the optimization problem

$$\min_{\mathbf{W}_i} \frac{1}{2} \sum_{j\neq i} W_{ij}^2 \tag{A.6}$$

subject to the constraints given by (A.5). The factor of $\frac{1}{2}$ in (A.6) is included for later convenience. This is an example of a quadratic programming problem in which we are trying to minimize a quadratic function subject to a set of linear inequality constraints [30, 43, 179].

In order to solve this constrained optimization problem, we introduce Lagrange multipliers $\{\lambda_{\ell\mu}\}$, with one multiplier for each of the constraints in (A.5), giving the Lagrangian function

$$L(\mathbf{W}_i, \{\lambda_{\ell\mu}\}) = \frac{1}{2} \sum_{j\neq i} W_{ij}^2 - \sum_{\ell,\mu} \lambda_{\ell,\mu} \left\{ (2\sigma_i^{\ell,\mu} - 1) \sum_{j\neq i} W_{ij} \sigma_j^{\ell,\mu} - 1 \right\}. \tag{A.7}$$

Note the minus sign in front of the Lagrange multiplier term because we are minimizing with respect to $\mathbf{W}_i$ and maximizing with respect to $\{\lambda_{\ell\mu}\}$. Setting the derivatives of $L(\mathbf{W}_i, \{\lambda_{\ell\mu}\})$ with respect to $\mathbf{W}_i$ equal to zero, we obtain the following conditions

$$W_{ij} = \sum_{\ell=1}^{L} \sum_{\mu=1}^{p} \lambda_{\ell\mu} \, (2\sigma_i^{\ell\mu} - 1) \, \sigma_j^{\ell\mu} \, . \tag{A.8}$$

Eliminating $\mathbf{W}_i$ from $L(\mathbf{W}_i, \{\lambda_{\ell\mu}\})$ using these conditions then gives the dual representation of the maximum margin problem in which we maximize

$$\tilde{L}(\{\lambda_{\ell\mu}\}) = \sum_{\ell=1}^{L} \sum_{\mu=1}^{p} \lambda_{\ell\mu} - \frac{1}{2} \sum_{\ell,m=1}^{L} \sum_{\mu,\nu=1}^{p} (2\sigma_i^{\ell\mu} - 1)(2\sigma_i^{m\nu} - 1) \lambda_{\ell\mu} \lambda_{m\nu} \sum_{j(\neq i)}^{N} \sigma_j^{\ell\mu} \sigma_j^{m\nu} \tag{A.9}$$

subject to the constraints

$$\lambda_{\ell,\mu} \geqslant 0, \, \forall \, \ell, \mu \, . \tag{A.10}$$

This optimization problem can be solved using available numerical routines [68, 195].

In the end, the only relevant data points for the learning process are the one corresponding to $\lambda_{\ell\mu} > 0$ as they define the weights of the network according to the Eq. (A.8), they infact correspond to points that lie on the maximum margin hyperplanes in the input space, see Fig. 15(b), and are the so called support vectors.

Once we have obtained $\{\lambda_{\ell\mu}\}$ we can then finally find the maximum margin as

$$\kappa_i = \frac{1}{\sqrt{\sum_{j\neq i} W_{ij}^2}} \tag{A.11}$$

using (A.8).

As an illustration of the learning procedure, we show in Fig. 39 how the number of stored patterns (with positive stabilities) grows as a function of the number of iterations of the quadratic optimization algorithm solving (C.3), until all p prescribed patterns are stabilized.

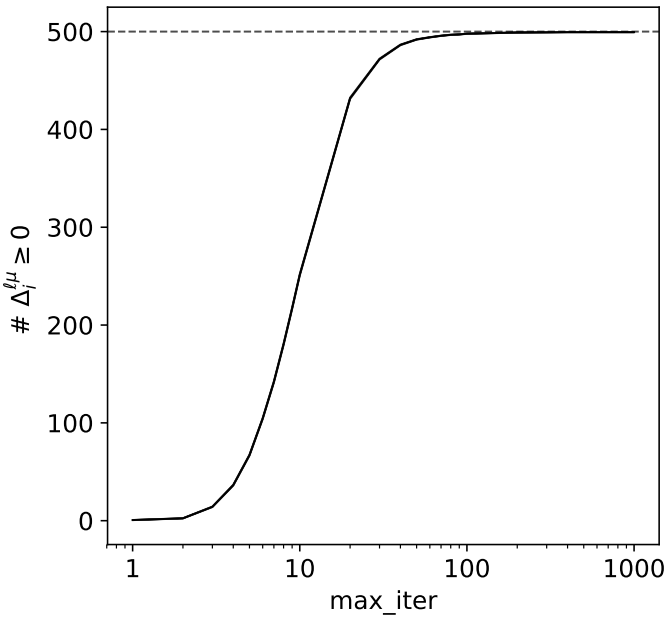

Figure 39 – Number of patterns with positive stabilities (y-axis) vs number of iterations of the quadratic optimization solver (x-axis) for one map ($L = 1$) with $p = 500$ points stored by a network with $N = 1000$ neurons, see Section 4.2 for details about the model and how the patterns were built. Parameter values: $D = 2$, $\phi_0 = .3$.

We are thus now able to calculate the optimal couplings for an SVM and to find the respective maximal margin κ_i. We can then use this algorithm to extract all the rows of

the RNN connectivity matrix, one for each SVM, which therefore is not a priori symmetric. We can as well derive all the margins $\kappa_i \ \forall\, i$ and at the end the margin associated to the RNN is given by the minimum among all the margins of the single SVMs

$$\kappa = \min_i \kappa_i \,, \tag{A.12}$$

see Section 4.3.

Note that formula (A.8) does not give normalized couplings to the unit and so, after calculating the optimal margin with Eq. (A.11), we should eventually normalize them by hand.

A.2 ESTIMATION OF CRITICAL CAPACITY FROM SVMS RESULTS

The critical capacity $\alpha_c(p)$, that is the value of the load $\alpha = \frac{L}{N}$ at which the stability $\kappa(\alpha, p)$ vanishes or, in other words, the maximal load sustainable by the network as a function of the required spatial error, was estimated as follows from SVM results in all the thesis figures in which it appears.

We computed the optimal stabilities κ (at fixed p) for $\tilde{M}$ different values of the load α (obviously the values of α are chosen in a range such that the classes are linearly separable, *i.e.*, $\kappa > 0$), with $\tilde{M}$ generally equal to 20. Then we fitted these points with the empirical function (depending on the parameters a, b, c)

$$\kappa = \frac{a}{\sqrt{\alpha}} + b\,\alpha + c \,, \tag{A.13}$$

and extrapolated from the fit the value of the load at which the fitted function vanished; this defined our estimate for $\alpha_c(p)$. Note that the small α behaviour in equation (A.13) above can be justified analytically from Gardner's calculation [87], see Section 4.5.

A.3 RECOVERING GARDNER RESULTS IN CASE OF ONE POSITION PER MAP

Here we show that, when only a single pattern is considered in each map ($p = 1$), the equation (4.39) is equivalent to the celebrated Gardner critical capacity in the case of biased patterns [87], where the bias comes from the PF area ϕ_0, if for example this area is half the area of the environments, *i.e.*, $\phi_0 = .5$, then the corresponding patterns would be unbiased.

For $p = 1$, the Euclidean random matrix $\Gamma(\hat{\mathcal{R}})$ defined in equation (4.17) reduces to the scalar

$$\Gamma(\hat{\mathcal{R}})_{1,1} = \phi_0(1-\phi_0) \equiv \frac{1-M^2}{4} \,, \tag{A.14}$$

where M is the average activity of the binary pattern in ± 1 notation, *i.e.*, under the change of variable $\sigma(\hat{\mathbf{r}}_1) = \{0,1\} \to \xi = 2\sigma(\hat{\mathbf{r}}_1) - 1 = \{-1,+1\}$.

The convex optimization problem to be solved in (4.38) thus amounts to compute

$$F(z_1, \nu, \kappa) = \min_{\{t_1 \geqslant \kappa\}} \left[\frac{4}{1-M^2} \big(t - (z_1 + \nu\,\xi)\big)^2 \right], \tag{A.15}$$

where $\nu = m\,\phi_0$ and the Gaussian variable z_1 in (4.38) has zero mean and variance $\Gamma(\hat{\mathcal{R}})_{1,1}$.

The minimum over t_1 in (A.15) can easily be determined, with the result

$$F(z_1, \nu, \kappa) = \begin{cases} \dfrac{4}{1-M^2} \big(\kappa - (z_1 + \nu\,\xi)\big)^2 & \text{if } \kappa \geqslant z_1 + \nu\,\xi \\ 0 & \text{otherwise.} \end{cases} \tag{A.16}$$

As $\hat{\mathbf{r}}_1$ is drawn uniformly at random, ξ is a random binary variable:

$$\xi = \begin{cases} +1 & \text{with probability } \dfrac{1+M}{2}, \\ -1 & \text{with probability } \dfrac{1-M}{2}. \end{cases} \tag{A.17}$$

We get, with the normalized Gaussian variable $z = z_1 \times 2/\sqrt{1-M^2}$ and the measure $Dz = dz/\sqrt{2\pi}\,\exp(-z^2/2)$,

$$\frac{1}{\alpha_c(\nu; \kappa, p=1)} = \tag{A.18}$$

$$\frac{1+M}{2} \int_{\frac{2\nu M - 2\kappa}{\sqrt{1-M^2}}}^{\infty} Dz \left(\frac{2\kappa - 2\nu M}{\sqrt{1-M^2}} + z \right)^2 + \frac{1-M}{2} \int_{\frac{-2\kappa - 2\nu M}{\sqrt{1-M^2}}}^{\infty} Dz \left(\frac{2\kappa + 2\nu M}{\sqrt{1-M^2}} + z \right)^2,$$

where ν is chosen in order to maximize $\alpha_c(\nu; \kappa, p=1)$:

$$\frac{1+M}{2} \int_{\frac{2\nu M - 2\kappa}{\sqrt{1-M^2}}}^{\infty} Dz \left(\frac{2\kappa - 2\nu M}{\sqrt{1-M^2}} + z \right) = \frac{1-M}{2} \int_{\frac{-2\kappa - 2\nu M}{\sqrt{1-M^2}}}^{\infty} Dz \left(\frac{2\kappa + 2\nu M}{\sqrt{1-M^2}} + z \right). \tag{A.19}$$

These equations coincide with the results of [87] up to the change $\kappa \to 2\kappa$ due to the fact that the neuron activities take here values 0,1 and not ± 1.

A.4 COMPUTATION OF $\Xi(u)$

Here we explain in detail the calculation of $\Xi(u)$ defined in Eq. (4.52).

Expanding the terms in $\Xi(u)$, we write $\Xi(u) = \Xi_1(u) + \Xi_2(u) + \Xi_3(u)$ with

$$\Xi_1(u) = 4 \sum_{j,k\geqslant 2} c_{1j} \Big(u\,\mathbf{Id} + c\Big)^{-1}_{jk} c_{1k}\,, \tag{A.20}$$

$$\Xi_2(u) = \phi_0^2 \sum_{j,k\geqslant 2} \Big(u\,\mathbf{Id} + c\Big)^{-1}_{jk}\,, \tag{A.21}$$

$$\Xi_3(u) = -4\,\phi_0 \sum_{j,k\geqslant 2} c_{1j} \Big(u\,\mathbf{Id} + c\Big)^{-1}_{jk}\,. \tag{A.22}$$

Computation of Ξ_1: Consider the $N \times N$ matrix $c^{(N)}$, with entries c_{ij} for i, j comprised between 1 and N. Let us also define $\mathbf{Id}^{(N)}$ as the identity matrix in dimension N, while $\mathbf{Id}$ above referred to the identity matrix in dimension $N-1$. Using block-matrix inversion formulas, we write that

$$\Big(u\,\mathbf{Id}^{(N)} + c^{(N)}\Big)^{-1}_{11} = \frac{1}{u + c_{11} - \sum_{j,k\geqslant 2} c_{1j} \Big(u\,\mathbf{Id} + c\Big)^{-1}_{jk} c_{1k}}\,. \tag{A.23}$$

The left hand side of the equation above is equal, in the large–N limit, to the resolvent $g(u)$ of c defined in (4.47). Using $c_{11} = \Gamma(0) = \phi_0$ and the definition of $\Xi_1(u)$, we obtain

$$\Xi_1(u) = 4\left(u + \phi_0 - \frac{1}{g(u)}\right)\,. \tag{A.24}$$

Computation of Ξ_2: Let $|v_+\rangle$ be the normalized vector with N identical components, $(v_+)_i = \frac{1}{\sqrt{N}}$. We have

$$\Xi_2(u) = N\,\phi_0^2 \left\langle v_+ \right| \Big(u\,\mathbf{Id} + c\Big)^{-1} \left| v_+ \right\rangle\,. \tag{A.25}$$

For large N, $|v_+\rangle$ is the top eigenvector of c, with (extensive) eigenvalue $\lambda_+ = N \int d\mathbf{r}\,\Gamma(\mathbf{r}) = N\,\phi_0^2$. Hence,

$$\Xi_2(u) = N\,\phi_0^2 \times \frac{1}{u + N\phi_0^2} \to 1\,, \tag{A.26}$$

in the large-N limit (since u remains bounded, see Section 4.6).

Computation of Ξ_3: As $\mathcal{C}_{jk}$ with $j,k \geqslant 2$ does not depend on the locations $\mathbf{r}_i^\ell$ of the place fields associated to neuron $i = 1$ in the different maps ℓ, we may substitute $\mathcal{C}_{1j}$ in Eq. (A.22) with its average over those positions, equal to ϕ_0^2. We obtain

$$\Xi_3(U) = -4\,\phi_0^3 \sum_{j,k\geqslant 2} \Big(U\,\mathbf{Id} + \mathcal{C}\Big)^{-1}_{jk} = -4\,\phi_0\,, \tag{A.27}$$

in the large-N limit, see the calculation of $\Xi_2(U)$.

Expression of Ξ: Gathering the three terms above, we obtain

$$\Xi(U) = 1 + 4\,U - \frac{4}{g(U)}\,. \tag{A.28}$$

A.5 DEPENDENCE ON ϕ_0 AND D.

Here we discuss the dependence of the results found in Section 4.6 on the size of the place fields ϕ_0 and the dimension of the maps D (as already stated in Section 4.2 the dimension of the place fields is always the same as that of the environments).

In Fig. 40 we show that the value of p such that the results obtained with the quenched PF theory and SVMs match increases as ϕ_0 decrease.

In fact, an analysis of equations (4.66, 4.67), valid in the small ϕ_0 limit, indicate that this minimal value of p scales as

$$p_{match}(\phi_0) \sim \frac{e^{1/(8\phi_0)}}{\phi_0^{3/2}}\,, \tag{A.29}$$

and becomes very large as ϕ_0 becomes small. Realistic values for the place fields area ϕ_0 are reported in the experimental literature [163] and [116] to range between .2 and .3. We stress that place fields are, however, continuous-valued rate fields for real neurons, while, in our model (Section 4.2), they represent binary on/off values. The correspondence between our model and experimental studies relies therefore on the introduction of a cut-off value for the minimal firing rate of place cells; values of ϕ_0 ranging between .2 and .3 seem to be reasonable in view of [163] and [116].

Moreover, Fig. 41 shows the spatial error of trained recurrent neural networks as a function of the number of maps L, the optimal stability κ as a function of the load $\alpha = \frac{L}{N}$ and the critical capacity $\alpha_c(p)$ as a function of the number of stored positions per map, everything for different values of p and for patterns generated starting from maps in dimensions $D = 1$ and 3, completing the results shown for $D = 2$ in Chapter 4. We observe the faster decay of the critical capacity predicted by equation (4.75) with increasing values of D.

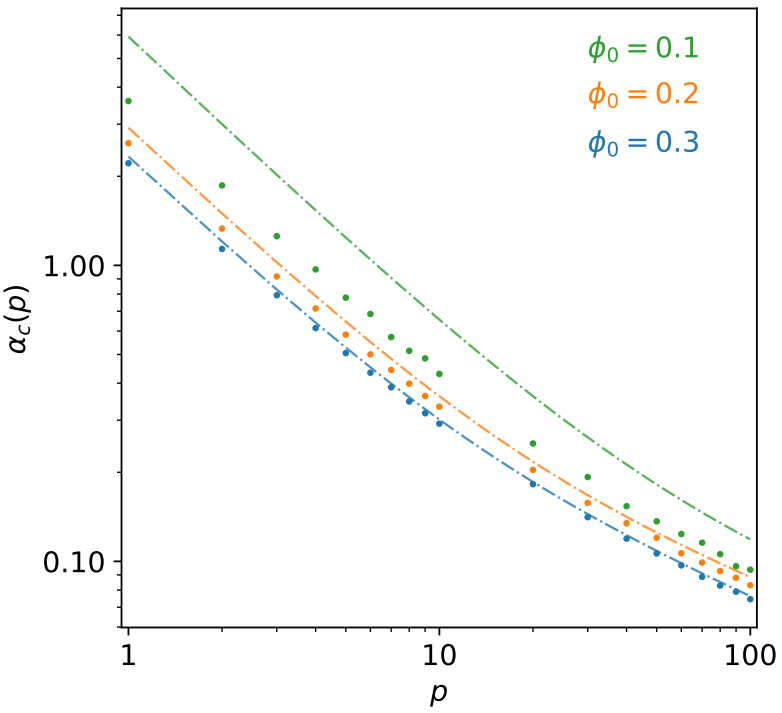

Figure 40 – Scaling cross-over of $\alpha_c(p)$ vs p for different values of ϕ_0. Quenched PF Theory (dashed-dotted lines) gets closer to SVM (scatter plots) as p increase, the value of p for which quenched PF theory and SVM matches increase as ϕ_0 decrease. We use for this results $D = 2$, $N = 5000$, and we have averaged over 50 different realization of the environments and different realizations of the p positions.

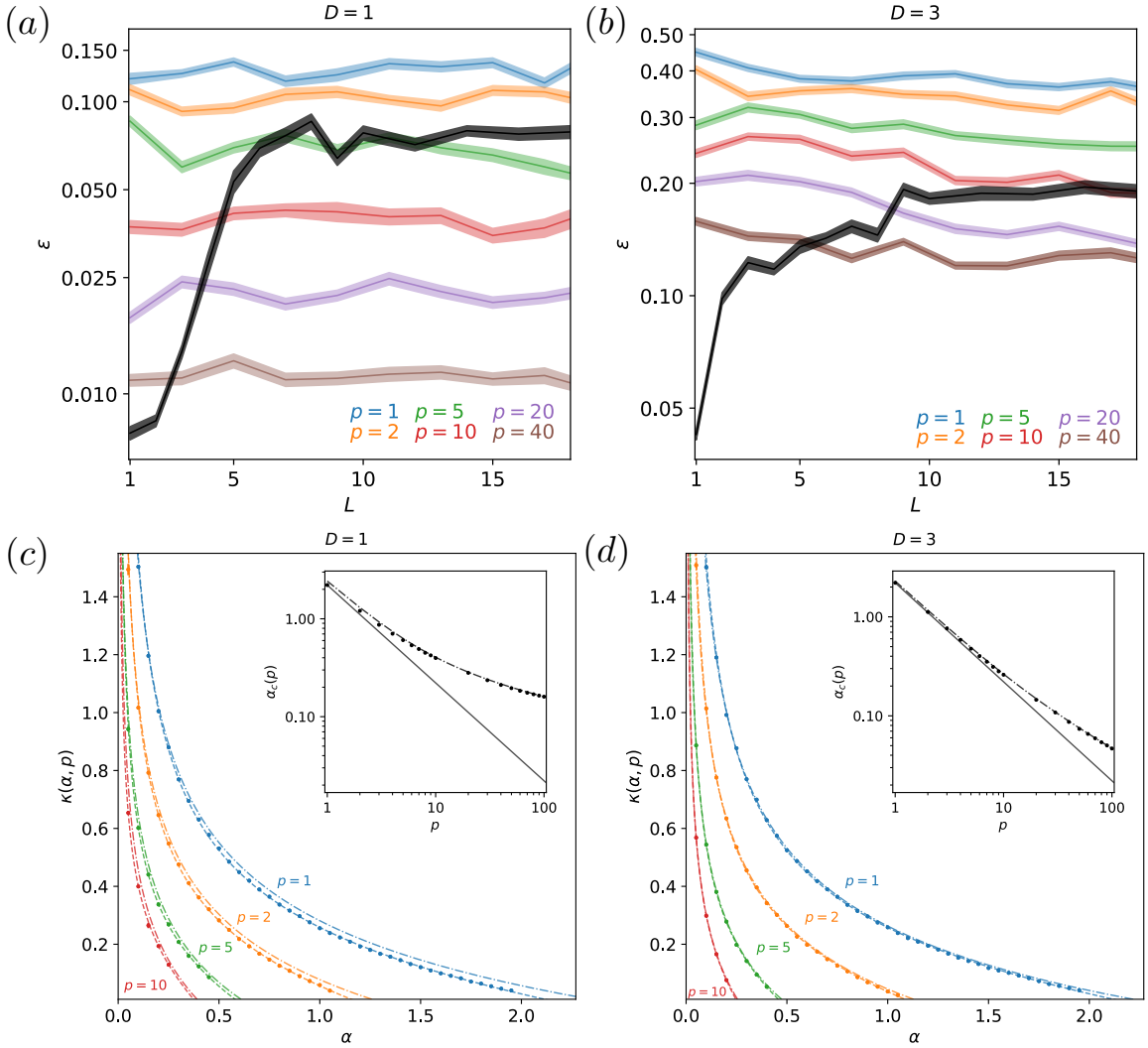

Figure 41 – Same results for spatial error and capacity presented in Chapter 4 but here for D = 1, (a) and (c), and D = 3, (b) and (d). The other parameters are identical to the ones used for the equivalent figures in D = 2, see Figs. 18(a), 21 and 23(a).

B

APPENDIX-CHAPTER 5

This short Appendix is intended to remind the reader of the fundamental notions and quantities of random matrix theory [1] needed to understand Chapter 5, *i.e.*, the study of MERM's spectral properties.

B.1 RESOLVENT, BLUE FUNCTION AND R-TRANSFORM

The eigenvalues Λ_i of a $N \times N$ Hermitian matrix $\mathbf{A}$, random or not, are real. Their density,

$$p(\Lambda) = \frac{1}{N} \left\langle \sum_{i=1}^{N} \delta(\Lambda - \Lambda_i) \right\rangle , \tag{B.1}$$

where $\langle \cdot \rangle$ stands for the average over the distribution of the matrix $\mathbf{A}$, can be obtained from the (one-point) resolvent

$$g(z) = \frac{1}{N} \left\langle \text{Trace} \frac{1}{z\,\mathbf{Id} - \mathbf{A}} \right\rangle = \frac{1}{N} \left\langle \sum_{i=1}^{N} \frac{1}{z - \Lambda_i} \right\rangle , \tag{B.2}$$

where $\mathbf{Id}$ is the identity matrix of size N.

Using the standard relation $\lim_{\epsilon \to 0^+} 1/(\Lambda + i\epsilon) = P\, 1/\Lambda - i\pi\delta(\Lambda)$ (P denotes the Cauchy principal value), we can rewrite Eq. (B.2) as

$$g(\Lambda + i\epsilon) = P \int_{-\infty}^{\infty} d\Lambda' \frac{p(\Lambda')}{\Lambda - \Lambda'} - i\pi p(\Lambda) , \tag{B.3}$$

so that $p(\Lambda)$ may be reconstructed from either the imaginary or the real part of $g(\Lambda + i\epsilon)$:

$$p(\Lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \text{Im}\, g(\Lambda + i\epsilon) , \tag{B.4}$$

$$P \int_{-\infty}^{\infty} d\Lambda' \frac{p(\Lambda')}{\Lambda - \Lambda'} = \text{Re}\, g(\Lambda + i\epsilon) . \tag{B.5}$$

1. For a detailed analysis see for example [144, 197].

In order to compute $g(z)$, we can rewrite Eq. (B.2) in various forms. First, we note that

$$\sum_{i=1}^{N} \frac{1}{z-\Lambda_i} = \partial_z \ln \left[\prod_{i=1}^{N} (z-\Lambda_i) \right] , \tag{B.6}$$

and express the resolvent as

$$g(z) = \frac{1}{N} \partial_z \left\langle \ln \det(z\,\mathbf{Id} - \mathbf{A}) \right\rangle . \tag{B.7}$$

This expression will be used in the replica-based computation presented in Section 5.3.

Another interesting expression for $g(z)$ is a decomposition in terms of the moments of $p(\Lambda)$,

$$\langle \Lambda^m \rangle = \int_{-\infty}^{\infty} d\Lambda\, p(\Lambda) \Lambda^m = \frac{1}{N} \left\langle \text{Trace}\, A^m \right\rangle . \tag{B.8}$$

For Hermitian matrices, $g(z)$ is an holomorphic function of $z \in \mathbb{C}$ except for some cuts along the real axis where eigenvalues of $\mathbf{A}$ are concentrated. Therefore, we can reconstruct $g(z)$ for all z by analytic continuation of its series expansion

$$g(z) = \sum_{m=0}^{\infty} \frac{\langle \Lambda^m \rangle}{z^{m+1}} , \tag{B.9}$$

which, in general, converges only in the vicinity of $|z| \to \infty$.

Other important objects for us are the functional inverse of $g(z)$, also called the Blue function [2], and the R-transform:

$$B(z) \equiv g^{-1}(z) , \tag{B.10}$$

$$R(z) \equiv B(z) - \frac{1}{z} . \tag{B.11}$$

Both of them are fundamental objects of the free random variable theory, see Section B.2. In particular, $R(z)$ is the generating function of the "free cumulants".

Let us now mention a couple of properties useful for the analysis in Chapter 5. The functions $g(z)$, $B(z)$, and $R(z)$ obey the following scaling relations:

$$\begin{aligned} g_{c\,\mathbf{A}}(z) &= \frac{1}{c} g_{\mathbf{A}}(z/c), \\ B_{c\,\mathbf{A}}(z) &= c B_{\mathbf{A}}(c\,z) , \\ R_{c\,\mathbf{A}}(z) &= c R_{\mathbf{A}}(c\,z) , \end{aligned} \tag{B.12}$$

where $c \in \mathbb{C}^*$.

2. So named because it is defined as the functional inverse of the resolvent, the Green function.

B.2 FREE PROBABILITY THEORY IN A NUTSHELL

Free probability theory is a research field in mathematics started by Voiculescu in 1983 [246, 247]. The initial goal of this theory was to say something about the spectral properties of the sum of two matrices $\mathbf{X}_1 + \mathbf{X}_2$ from the knowledge of the spectral properties of the individual ones $\mathbf{X}_1$ and $\mathbf{X}_2$. Unless the two matrices commute, knowing the spectrum of the individual ones is not sufficient to find the spectrum of the sum. In any case free probability theory identifies a sufficient condition, the so-called asymptotic freeness with which this problem can be faced without resorting to the eigenvectors of the matrices. This notion of asymptotic freeness is a generalization of the concept of statistical independence for random variables when these variables, in our particular case the matrices, do not commute.

Let us briefly recall the basic properties of independent variables. We denote by p_x the probability density of the variable x, by $g_x(z) \equiv \langle e^{z\,x} \rangle = \sum_{n \geqslant 0} \langle x^n \rangle \, z^n / n!$ its characteristic function, and by $r_x(z) \equiv \ln g_x(z) = \sum_{n \geqslant 0} c_{x,n} \, z^n$ its cumulant generating function. For two independent real random variables x_1 and x_2, the following relations hold:

$$\langle x_1 x_2 \rangle = \langle x_1 \rangle \, \langle x_2 \rangle \, , \tag{B.13}$$

$$p_{x_1+x_2} = p_{x_1} * p_{x_2} \, , \tag{B.14}$$

$$r_{x_1+x_2} = r_{x_1} + r_{x_2} \, , \tag{B.15}$$

where "$*$" denotes the convolution operation. We will see that these relations find their equivalents for asymptotically free matrices.

By definition, two Hermitian matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are asymptotically free if for all $l \in \mathbb{N}$ and for all polynomials p_i and q_i ($1 \leqslant i \leqslant l$), we have [240]

$$\begin{aligned} &\langle p_i(\mathbf{X}_1) \rangle_\Lambda = \langle q_i(\mathbf{X}_2) \rangle_\Lambda = 0 \\ &\Rightarrow \quad \langle p_1(\mathbf{X}_1) q_1(\mathbf{X}_2) \dots p_l(\mathbf{X}_1) q_l(\mathbf{X}_2) \rangle_\Lambda = 0 \, , \end{aligned} \tag{B.16}$$

where the expectation value $\langle \cdot \rangle_\Lambda$ is defined as

$$\langle X \rangle_\Lambda = \frac{1}{N} \langle \mathrm{Trace}\, X \rangle \, . \tag{B.17}$$

The interpretation of the formal definition (B.16) is the following: two matrices are asymptotically free if their eigenbases are related to one another by a random rotation, or said differently, if their eigenvectors are almost surely orthogonal.

From the definition (B.16), it is easy to compute various mixed moments of $\mathbf{X}_1$ and $\mathbf{X}_2$. Considering $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \langle \mathbf{X}_i \rangle_\Lambda$ that obey $\langle \tilde{\mathbf{X}}_1 \rangle_\Lambda = \langle \tilde{\mathbf{X}}_2 \rangle_\Lambda = 0$, we obtain from Eq. (B.16):

$$\langle \mathbf{X}_1 \mathbf{X}_2 \rangle_\Lambda = \langle \mathbf{X}_1 \rangle_\Lambda \, \langle \mathbf{X}_2 \rangle_\Lambda \, . \tag{B.18}$$

Note that this last condition is not enough to define asymptotic freeness, since matrices do not commute. For example, from Eq. (B.16), the forth moments read

$$\begin{aligned}\langle \mathbf{X}_1\mathbf{X}_1\mathbf{X}_2\mathbf{X}_2\rangle_\Lambda &= \langle \mathbf{X}_1^2\rangle_\Lambda \langle \mathbf{X}_2^2\rangle_\Lambda \,,\\ \langle \mathbf{X}_1\mathbf{X}_2\mathbf{X}_1\mathbf{X}_2\rangle_\Lambda &= \langle \mathbf{X}_1^2\rangle_\Lambda \langle \mathbf{X}_2\rangle_\Lambda^2 + \langle \mathbf{X}_1\rangle_\Lambda^2 \langle \mathbf{X}_2^2\rangle_\Lambda \\ &\quad - \langle \mathbf{X}_1\rangle_\Lambda^2 \langle \mathbf{X}_2\rangle_\Lambda^2 \,. \end{aligned} \tag{B.19}$$

The free cumulants are defined such that the sum property (B.15) is preserved for the generating function of the free cumulants, the so-called R-transform [120, 240]. Interestingly, the R-transform is simply related to the Blue function (B.10), *i.e.*, the functional inverse of the resolvent $g(z)$, by Eq. (B.11).[3] The R-transform of the sum of two asymptotically free matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ obeys:

$$R_{\mathbf{X}_1+\mathbf{X}_2}(z) = R_{\mathbf{X}_1}(z) + R_{\mathbf{X}_2}(z)\,. \tag{B.20}$$

Hence, the problem of finding the eigenvalue distribution of the sum of two free random matrices is straightforward. Applying successively Eqs. (B.10), (B.11), and (B.20), one readily infers $g_{\mathbf{X}_1+\mathbf{X}_2}$ from $g_{\mathbf{X}_1}$ and $g_{\mathbf{X}_2}$. The steps of the algorithm are as follows:

$$\begin{aligned} g_{\mathbf{X}_1},\, g_{\mathbf{X}_2} &\to B_{\mathbf{X}_1}, B_{\mathbf{X}_2} \to R_{\mathbf{X}_1},\, R_{\mathbf{X}_2} \to R_{\mathbf{X}_1+\mathbf{X}_2} \\ &\to B_{\mathbf{X}_1+\mathbf{X}_2} \to g_{\mathbf{X}_1+\mathbf{X}_2}\,. \end{aligned} \tag{B.21}$$

Moreover, the generalization to the sum of an arbitrary number of matrices is trivial.

3. Note that $g(z)$ plays the role of a free characteristic function, see Eqs. (B.8) and (B.9).

C

APPENDIX-CHAPTER 6

In this Chapter we report some details on the numerical, algorithmic and analytical part of the results presented in Chapter 6 in order not to make the latter too heavy to read.

C.1 SVM ALGORITHM WITH SIGN-CONSTRAINED SYNAPSES

Here we generalize the algorithm of support vector machines presented in Sections 4.3 and A.1 to take into account the constraint on the sign of the synapses. In fact, the basic algorithm does not consider this aspect admitting a priori connections of any sign.

After we have generated a data-set of activity patterns, as in Section 4.2, we want to learn the connections $\{W_{ij}\}$ of the recurrent neural network that maximize the stability

$$\kappa_{pos} = \max_{\{W_{ij}\geqslant 0,\theta_i\}} \min_{\{i=1\dots N,\ell=1\dots L,\mu=1\dots p\}} \left\{ (2\sigma_i^{\ell,\mu}-1)\Big[\sum_{j(\neq i)} W_{ij}\,\sigma_j^{\ell\mu} + \theta_i \Big] \right\} \tag{C.1}$$

defined already in Eq. (6.3) at fixed α and p. Indeed, as we have previously seen in Sections 4.3 and A.1, this choice of the weights ensure the biggest basins of attraction in the pattern space, *i.e.*, robustness against thermal noise. In order to do that we implement SVM learning [35, 211], but adding this time the positivity constraint on $\{W_{ij}\}$ and the threshold terms $\{\theta_i\}$ (fundamental now that the weights are only positive to make the problem linearly separable).

In practice, for each neuron i, we want to compute the connections $\{W_{ij}\}$ from the other neurons j (with $W_{ii}=0$, $\forall\, i$, no self-connections) and $\{\theta_i\}$, which are solution of the following primal constrained convex optimization problem

$$\begin{aligned} &\underset{\{W_{ij},\theta_i\}}{\text{minimize}} \quad \frac{1}{2}\sum_{j(\neq i)} W_{ij}^2\,, \\ &\text{subject to} \quad (2\sigma_i^{\ell\mu}-1)\Big(\sum_{j(\neq i)} W_{ij}\,\sigma_j^{\ell\mu}+\theta_i\Big)\geqslant 1, \quad \forall\,\ell,\mu\,, \\ &\qquad\qquad W_{ij}\geqslant 0, \forall\, j(\neq i)\,. \end{aligned} \tag{C.2}$$

We have to solve N such problems to extract all the rows of the coupling matrix and all the threshold terms. The dual form of this problem is

$$\begin{aligned}\underset{\{\lambda_{\ell,\mu},c_{ij}\}}{\text{maximize}} \sum_{\ell=1}^{L}\sum_{\mu=1}^{p}\lambda_{\ell\mu} - \frac{1}{2}\sum_{\ell,m=1}^{L}\sum_{\mu,\nu=1}^{p}(2\sigma_i^{\ell\mu}-1)(2\sigma_i^{m\nu}-1)\lambda_{\ell\mu}\lambda_{m\nu}\sum_{j(\neq i)}^{N}\sigma_j^{\ell\mu}\sigma_j^{m\nu} \\ -\frac{1}{2}\sum_{j(\neq i)}^{N}c_{ij}^2 - \sum_{\ell=1}^{L}\sum_{\mu=1}^{p}(2\sigma_i^{\ell\mu}-1)\lambda_{\ell\mu}\sum_{j(\neq i)}^{N}c_{ij}\sigma_j^{\ell\mu}\,, \\ \text{subject to } \lambda_{\ell,\mu}\geqslant 0,\ \forall\,\ell,\mu\,, \\ c_{ij}\geqslant 0,\ \forall\, j(\neq i)\,, \\ \sum_{\ell=1}^{L}\sum_{\mu=1}^{p}(2\sigma_i^{\ell\mu}-1)\lambda_{\ell\mu}=0\,. \end{aligned} \tag{C.3}$$

where the $\{\lambda_{\ell\mu}\}$'s and $\{c_{ij}\}$'s are Lagrange multipliers enforcing respectively the first and the second sets of constraints in defined in Eq. (C.2). This optimization problem can be solved using available numerical routines [68], as the standard SVM, see Sections 4.3 and A.1.

Once we obtain the $\{\lambda_{\ell\mu}\}$'s and $\{c_{ij}\}$'s we can compute the connections through

$$W_{ij} = c_{ij} + \sum_{\ell=1}^{L}\sum_{\mu=1}^{p}\lambda_{\ell\mu}\,(2\sigma_i^{\ell\mu}-1)\,\sigma_j^{\ell\mu}\,. \tag{C.4}$$

Using the fact that any support vector (data points that lies on the maximal hyperplanes) satisfies

$$(2\sigma_i^{\ell^*,\mu^*}-1)\sum_{j\neq i}W_{ij}\sigma_j^{\ell^*,\mu^*}=1\,, \tag{C.5}$$

we can determine the value of the threshold parameters $\{\theta_i\}$'s thanks to

$$\theta_i = \frac{1}{N_S}\sum_{(\ell^*,\mu^*)\in S}\left[(2\sigma_i^{\ell^*\mu^*}-1) - \sum_{(m^*,\nu^*)\in S}\lambda_{m^*\nu^*}(2\sigma_i^{m^*\nu^*}-1)\sum_{j(\neq i)}^{N}\sigma_j^{\ell^*\mu^*}\sigma_j^{m^*\nu^*}\right], \tag{C.6}$$

where S is the set of support vectors and N_S is their number.

We then normalize the rows of the couplings matrix to unity, *i.e.*, $\sum_{j(\neq i)}W_{ij}^2 = 1$, and divide by the same number also the thresholds $\{\theta_i\}$'s. Finally, the stability κ_{pos} is computed through the formula (6.3).

C.2 ANALYTICAL DETAILS ON THE INDIVIDUALITY OF NEURONS

Here we show a series of generalizations of the computations presented in Sections 5.3 and 4.6 that will allow us to include in the model introduced in Chapter 4 the individuality of neurons discussed in Section 6.5 together with the variants on the place cell models presented in Section 6.4.

C.2.1 *Spectrum of MERM: multi-populations of neurons*

As a first step in the above direction we consider the following MERM:

$$M_{ij}^{(L)} = \frac{1}{N}\sum_{\ell=1}^{L} \Gamma_{ij}\left(\left|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell\right|\right) \ , \tag{C.7}$$

in which the Γ function depend explicitly on the indices i and j, *i.e.*, the N neurons are divided in M finite groups with fractions of neurons β_μ with $\mu = 1, \dots, M$. Every group of neurons has a specific property, *i.e.*, a fixed area for the place fields ϕ_μ.

We are going to compute the resolvent of $\mathbf{M}^{(L)}$ using the replica method.

As in Section 5.3 we start by rewriting the definition of the resolvent as

$$s_L(z) = \frac{1}{N}\Big\langle \text{Trace}\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-1}\Big\rangle = \frac{2}{N}\partial_z\Big\langle \log\det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}}\Big\rangle \ , \tag{C.8}$$

where $\langle\cdot\rangle$ stands for the average over the distribution of the matrix (C.7). With this representation the determinant $\det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}}$ can be expressed as a canonical partition function:

$$\mathcal{Z}_L(s) = \det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}} = \int \prod_i \frac{d\phi_i}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_i \phi_i^2 - \frac{1}{2}\sum_{ij}\phi_i M_{ij}^{(L)}\phi_j\right) \ , \tag{C.9}$$

where i, j go from 1 to N.

The resolvent (C.8) can be calculated using the replica trick [157]:

$$s_L(z) = \frac{2}{N}\partial_z\big\langle \log \mathcal{Z}_L(s)\big\rangle = \frac{2}{N}\partial_z\Big[\lim_{n\to 0}\frac{1}{n}\log\big\langle \mathcal{Z}_L(s)^n\big\rangle\Big] \tag{C.10}$$

with

$$\big\langle \mathcal{Z}_L(s)^n\big\rangle = \int \prod_{ia}\frac{d\phi_i^a}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_a\sum_i(\phi_i^a)^2\right)\Big\langle \exp\left(-\frac{1}{2}\sum_a\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\right)\Big\rangle \ , \tag{C.11}$$

where we have replicated the system n times, *i.e.*, a goes from 1 to n.

In order to perform the average in (C.11) we rewrite (C.7) by considering the ℓ-th space ERM in its eigenbasis:

$$M_{ij}^{(L)} = \frac{1}{N}\sum_{\ell=1}^{L}\Gamma_{ij}\left(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\right) = \sum_{\ell}\sum_{\mathbf{k}\neq\mathbf{0}} \nu_{\mathbf{k}i}^\ell\,\hat{\Gamma}_{ij}(\mathbf{k})\,\nu_{\mathbf{k}j}^\ell\,, \tag{C.12}$$

where ℓ goes from 1 to L, and the sum over $\mathbf{k}$ discards the $\mathbf{k}=\mathbf{0}$ extensive mode as discussed in Chapter 5, $\hat{\Gamma}_{ij}(\mathbf{k}) = \hat{\gamma}_i(\mathbf{k})\hat{\gamma}_j(\mathbf{k})$ with $\hat{\gamma}_i(\mathbf{k})$ being the Fourier transform of the indicator function of the place field of area ϕ_μ and the eigenvector components, $\nu_{\mathbf{k}i}^\ell \simeq \frac{1}{\sqrt{N}}\sin\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right)$, $\frac{1}{\sqrt{N}}\cos\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_i^\ell\right)$, are real due to the symmetry $\hat{\Gamma}(\mathbf{k}) = \hat{\Gamma}(-\mathbf{k})$. Hence we get

$$\left\langle \exp\left(-\frac{1}{2}\sum_{a}\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\right)\right\rangle = \left\langle \exp\left(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{0}}\left(\sum_i \nu_{\mathbf{k}i}^\ell\phi_i^a\hat{\gamma}_i(\mathbf{k})\right)^2\right)\right\rangle. \tag{C.13}$$

We now use the Stratonovich trick to linearize $\left(\sum_i \nu_{\mathbf{k}i}^\ell\phi_i^a\hat{\gamma}_i(\mathbf{k})\right)^2$:

$$\left\langle \exp\left(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{0}}\left(\sum_i \nu_{\mathbf{k}i}^\ell\phi_i^a\hat{\gamma}_i(\mathbf{k})\right)^2\right)\right\rangle = \prod_\ell\int\prod_{a,\mathbf{k}\neq\mathbf{0}}\frac{du_{\ell,\mathbf{k}}^a}{\sqrt{2\pi}} \tag{C.14}$$

$$\times\exp\left(-\frac{1}{2}\sum_{a,\mathbf{k}\neq\mathbf{0}}\left(u_{\ell,\mathbf{k}}^a\right)^2\right)\left\langle \exp\left(-i\sum_{a,\mathbf{k}\neq\mathbf{0}}u_{\ell\mathbf{k}}^a\sum_i\phi_i^a\nu_{\mathbf{k}i}^\ell\hat{\gamma}_i(\mathbf{k})\right)\right\rangle.$$

Using the fact that $\langle\nu_{\mathbf{k}i}^\ell\rangle = 0$ and $\langle\nu_{\mathbf{k}i}^\ell\nu_{\mathbf{k}'j}^\ell\rangle = \frac{1}{N}\delta_{ij}\delta_{\mathbf{k}\mathbf{k}'}$ it is easy to perform the average in the above equation, with the result

$$\left\langle \exp\left(-i\sum_{a,\mathbf{k}\neq\mathbf{0}}u_{\ell\mathbf{k}}^a\sum_i\phi_i^a\nu_{\mathbf{k}i}^\ell\hat{\gamma}_i(\mathbf{k})\right)\right\rangle = \exp\left(-\frac{1}{2}\sum_{a,b}\sum_{\mathbf{k}\neq\mathbf{0}}u_{\ell\mathbf{k}}^a u_{\ell\mathbf{k}}^b\sum_{\mu=1}^{M}q_\mu^{ab}\hat{\gamma}_\mu(\mathbf{k})^2\right) \tag{C.15}$$

where we have used the fact that the N neurons are divided in M groups and we have defined an overlap q_μ^{ab} per group as

$$q_\mu^{ab} = \frac{1}{N}\sum_{i\in\beta_\mu N}\phi_i^a\phi_i^b,\ \forall\mu \tag{C.16}$$

to be fixed through

$$1 = \int\prod_{a\leqslant b}\frac{d\hat{q}_\mu^{ab}dq_\mu^{ab}}{\frac{2\pi i}{N}}\exp\left(N\sum_{a\leqslant b}\hat{q}_\mu^{ab}q_\mu^{ab} - \sum_{a\leqslant b}\hat{q}_\mu^{ab}\sum_{i\in\beta_\mu N}\phi_i^a\phi_i^b\right),\ \forall\mu\,. \tag{C.17}$$

We can finally write $\langle \mathcal{Z}_L(s)^n \rangle$ as

$$\int \prod_\mu \prod_{a \leqslant b} \frac{d\hat{q}_\mu^{ab} dq_\mu^{ab}}{\frac{2\pi i}{N}} \exp\left\{ N \left[\sum_\mu \beta_\mu \log \int \prod_a \frac{d\phi^a}{\sqrt{2\pi}} \right.\right. \\ \left. \exp\left(\frac{z}{2} \sum_a (\phi^a)^2 - \sum_{a \leqslant b} \hat{q}_\mu^{ab} \phi^a \phi^b \right) + \sum_\mu \sum_{a \leqslant b} \hat{q}_\mu^{ab} q_\mu^{ab} \right. \\ \left.\left. + \alpha \log \int \prod_{\mathbf{k} \neq \mathbf{0}, a} \frac{du_{\mathbf{k}}^a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_{\mathbf{k} \neq \mathbf{0}, a} (u_{\mathbf{k}}^a)^2 - \frac{1}{2} \sum_{\mathbf{k} \neq \mathbf{0}} \sum_{a \leqslant b} u_{\mathbf{k}}^a u_{\mathbf{k}}^b \sum_\mu q_\mu^{ab} \hat{\gamma}_\mu(\mathbf{k})^2 \right) \right] \right\} . \tag{C.18}$$

The Gaussian integrals over ϕ^a and $u_{\mathbf{k}}^a$ can be easily computed. We then make the Replica Symmetric (RS) Ansatz on the structure of the order parameters q_μ^{ab} and their conjugate variables $\hat{q}_\mu^{ab}$, so that

$$q_\mu^{ab} = r_\mu + (q_\mu - r_\mu)\delta_{ab}, \ \forall \mu \tag{C.19}$$

and

$$\hat{q}_\mu^{ab} = \hat{r}_\mu + (\hat{q}_\mu - \hat{r}_\mu)\delta_{ab}, \ \forall \mu. \tag{C.20}$$

The integrals over q_μ, r_μ, $\hat{q}_\mu$ and $\hat{r}_\mu$ are then estimated using the saddle-point method valid for large N, and then taking the small n limit. The resulting expression for the resolvent $s_L(z)$ of (C.7) is

$$2\partial_z \left[\mathrm{opt}_{\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}} \lim_{n \to 0} \lim_{N \to \infty} \frac{1}{nN} \log \langle \mathcal{Z}_L(s)^n \rangle \right] = 2\partial_z \left[\mathrm{opt}_{\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}} f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}) \right], \tag{C.21}$$

where $f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\})$ is the free energy density equal to

$$\sum_\mu \hat{q}_\mu q_\mu - \frac{1}{2} \sum_\mu \hat{r}_\mu r_\mu - \frac{\alpha}{2} \sum_{\mathbf{k} \neq \mathbf{0}} \left[\log\left(1 + \sum_\mu \hat{\Gamma}_\mu(\mathbf{k})(q_\mu - r_\mu) \right) + \frac{\sum_\mu \hat{\Gamma}_\mu(\mathbf{k}) r_\mu}{1 + \sum_\mu \hat{\Gamma}_\mu(\mathbf{k})(q_\mu - r_\mu)} \right] \\ - \frac{1}{2} \sum_\mu \beta_\mu \left[\log\left(2\hat{q}_\mu - \hat{r}_\mu - z \right) + \frac{\hat{r}_\mu}{2\hat{q}_\mu - \hat{r}_\mu - z} \right], \tag{C.22}$$

where $\hat{\Gamma}_\mu(\mathbf{k}) = \hat{\gamma}_\mu(\mathbf{k})^2$. The saddle-point equations obtained by optimizing $f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\})$ with respect to $\hat{q}_\mu$, $\hat{r}_\mu$, q_μ and r_μ read $\forall \mu$

$$\begin{aligned}
\frac{q_\mu}{\beta_\mu} &= -\frac{\hat{r}_\mu}{\left(2\hat{q}_\mu - \hat{r}_\mu - z\right)^2} + \frac{1}{2\hat{q}_\mu - \hat{r}_\mu - z}\,, \quad r_\mu = -\frac{\beta_\mu \hat{r}_\mu}{\left(2\hat{q}_\mu - \hat{r}_\mu - z\right)^2}\,, \\
\hat{q}_\mu &= \frac{\alpha}{2}\sum_{\mathbf{k}\neq\mathbf{0}}\left(\frac{\hat{\Gamma}_\mu(\mathbf{k})}{1+\sum_\nu \hat{\Gamma}_\nu(\mathbf{k})(q_\nu - r_\nu)} - \frac{\hat{\Gamma}_\mu(\mathbf{k})\sum_\nu r_\nu\,\hat{\Gamma}_\nu(\mathbf{k})}{\left(1+\sum_\nu \hat{\Gamma}_\nu(\mathbf{k})(q_\nu - r_\nu)\right)^2}\right), \\
\hat{r}_\mu &= -\alpha\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\hat{\Gamma}_\mu(\mathbf{k})\sum_\nu r_\nu\,\hat{\Gamma}_\nu(\mathbf{k})}{\left(1+\sum_\nu \hat{\Gamma}_\nu(\mathbf{k})(q_\nu - r_\nu)\right)^2}\,.
\end{aligned} \tag{C.23}$$

This system of equations admits $r_\mu = \hat{r}_\mu = 0$, $\forall \mu$ as a solution, which gives, according to (C.21), the following system of equations satisfied by $s_L(z)$:

$$\begin{cases} s_L(z) = \sum_\mu q_\mu\,, \\ z = -\frac{\beta_\mu}{q_\mu} + \alpha\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\hat{\Gamma}_\mu(\mathbf{k})}{1+\sum_\nu \hat{\Gamma}_\nu(\mathbf{k})q_\nu}, \ \forall \mu\,. \end{cases} \tag{C.24}$$

Note that we are eventually interested in the spectral properties of the matrix $\mathbf{C}$ with entries

$$C_{ij} = \frac{1}{L}\sum_{\ell=1}^{L}\Gamma_{ij}\left(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\right) = \frac{1}{\alpha}M_{ij}^{(L)}\,. \tag{C.25}$$

Obviously, the resolvent s of $\mathbf{C}$ is related to the resolvent s_L of $\mathbf{M}^{(L)}$ through the equation $s(z) = \alpha\, s_L(\alpha z)$. Hence we obtain our fundamental system of equations for the resolvent of $\mathbf{C}$:

$$\begin{cases} s(z) = \sum_\mu q_\mu\,, \\ z = -\frac{\beta_\mu}{q_\mu} + \sum_{\mathbf{k}\neq\mathbf{0}}\frac{\alpha\hat{\Gamma}_\mu(\mathbf{k})}{\alpha+\sum_\nu \hat{\Gamma}_\nu(\mathbf{k})q_\nu}, \ \forall \mu\,. \end{cases} \tag{C.26}$$

C.2.2 *Quenched PF theory: multi-populations of neurons*

Here we are going to extend the Gaussian theory with quenched PF presented in Section 4.6 to the case of multi-populations of neurons. By multi-populations of neurons we mean that the N neurons are divided in M finite groups with fractions of neurons β_ρ with $\rho = 1, \ldots, M$ and every group of neurons with a specific property, *i.e.*, a fixed area for the place fields $\phi_\rho < 1$.

The computation follows exactly what seen in Section 4.6 until the definition of the order parameters that are now

$$m_\ell^a = \sum_{j\geqslant 2} W_{ja}\left(2\,\Gamma_{1j}\left(|\mathbf{r}_1^\ell - \mathbf{r}_j^\ell|\right) - \phi_j\right) \tag{C.27}$$

and

$$q_\ell^{ab} = \sum_{j,k\geqslant 2} W_{ja}\,W_{kb}\,\Gamma_{jk}\left(|\mathbf{r}_j^\ell - \mathbf{r}_k^\ell|\right)\,, \tag{C.28}$$

notice, in fact, that here the function Γ depend on the indices j and k of the neurons not only for indicating the centers of the PFs in the different maps but also to explicitly distinguish the properties of the different neurons and also ϕ_j varies from neuron to neuron.

Now we make the same approximations as Section 4.6 and so we can write the final expression of the order parameters after adding up all the maps:

$$m^a \equiv \frac{1}{L}\sum_{\ell=1}^{L} m_\ell^a = \sum_{j\geqslant 2} W_{ja}\left(2\,\mathcal{C}_{1j}\left(\{\mathbf{r}_j^\ell\}\right) - \phi_j\right) \tag{C.29}$$

and

$$q^{ab} \equiv \frac{1}{L}\sum_{\ell=1}^{L} q_\ell^{ab} = \sum_{j,k\geqslant 2} W_{ja}\,W_{kb}\,\mathcal{C}_{jk}\left(\{\mathbf{r}_j^\ell\}\right)\,. \tag{C.30}$$

The $N\times N$ multi-space Euclidean random matrix $\mathcal{C}$ appearing in the expressions above is defined as

$$C_{ij} = \frac{1}{L}\sum_{\ell=1}^{L}\Gamma_{ij}\left(\left|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell\right|\right)\,. \tag{C.31}$$

In the following, we denote by $\rho(\lambda)$ the density of eigenvalues λ of $\mathcal{C}$. This density is self-averaging when the PFs are randomly drawn in the large L, N double limit. Its resolvent, defined as

$$g(U) = \int d\lambda\,\frac{\rho(\lambda)}{\lambda + U}\,, \tag{C.32}$$

where the integral runs over the support of ρ, is solution of the following system of equations, see Section C.2.1:

$$\begin{cases} g(U) = \sum_{\rho=1}^{M} q_\rho\,, \\ U = -\frac{\beta_\rho}{q_\rho} + \sum_{\mathbf{k}\neq\mathbf{o}} \frac{\alpha\hat{\Gamma}_\rho(\mathbf{k})}{\alpha + \sum_{\theta=1}^{M}\hat{\Gamma}_\theta(\mathbf{k})\,q_\theta}, \quad \forall\rho\,. \end{cases} \tag{C.33}$$

From here on we obtain the same equations of what we saw in Section 4.6 except for the quantity Ξ defined as

$$\Xi(u) = \sum_{j,k\geqslant 2} H_j \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} H_k \qquad \text{with} \qquad H_j = 2\,\mathcal{C}_{1j} - \phi_j \;, \tag{C.34}$$

and $\mathbf{Id}$ is the identity matrix. In the above equation, the inverse is intended over the $N-1$-dimensional restriction of the matrix $u\,\mathbf{Id} + \mathcal{C}$ to entries $j,k \geqslant 2$.

C.2.2.1 *Computation of* Ξ

Expanding the terms in $\Xi(u)$ in Eq. (C.34) above, we write $\Xi(u) = \Xi_1(u) + \Xi_2(u) + \Xi_3(u)$ with

$$\Xi_1(u) = 4 \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \mathcal{C}_{1k} \;, \tag{C.35}$$

$$\Xi_2(u) = \sum_{j,k\geqslant 2} \phi_j \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \phi_k \;, \tag{C.36}$$

$$\Xi_3(u) = -2 \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \phi_k \; - 2 \sum_{j,k\geqslant 2} \phi_j \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \mathcal{C}_{1k} \;. \tag{C.37}$$

Computation of Ξ_1: Consider the $N \times N$ matrix $\mathcal{C}^{(N)}$, with entries $\mathcal{C}_{ij}$ for i,j comprised between 1 and N. Let us also define $\mathbf{Id}^{(N)}$ the identity matrix in dimension N, while $\mathbf{Id}$ above referred to the identity matrix in dimension $N-1$. Using block-matrix inversion formulas, we write that

$$\left(u\,\mathbf{Id}^{(N)} + \mathcal{C}^{(N)}\right)^{-1}_{11} = \frac{1}{u + \mathcal{C}_{11} - \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \mathcal{C}_{1k}} \tag{C.38}$$

The left hand side of the equation above is equal, in the large–N limit, to the resolvent $g(u)$ of $\mathcal{C}$ defined in (C.32). Using $\mathcal{C}_{11} = \phi_1$ and the definition of $\Xi_1(u)$, we obtain

$$\Xi_1(u) = 4\left(u + \phi_1 - \frac{1}{g(u)}\right) \;. \tag{C.39}$$

Computation of Ξ_2: Let $|v_+\rangle$ be the normalized vector with N identical components, $(v_+)_i = \frac{1}{\sqrt{N}}$. For large N, $|v_+\rangle$ is the top eigenvector of $\mathcal{C}$, with (extensive) eigenvalue $\lambda_+ = N\sum_{\rho=1}^{M} \beta_\rho \phi_\rho^2$. After the computation of the braket of this vector with the matrix of components $\phi_j \left(u\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} \phi_k$, and taking the large N limit we get:

$$\Xi_2(u) = \frac{N\sum_{\rho=1}^{M} \beta_\rho \phi_\rho^2}{u + N\sum_{\rho=1}^{M} \beta_\rho \phi_\rho^2} \to 1 \;. \tag{C.40}$$

Computation of Ξ_3: As $\mathcal{C}_{jk}$ with $j,k \geqslant 2$ does not depend on the locations $\mathbf{r}_i^\ell$ of the place fields associated to neuron $i = 1$ in the different maps ℓ, we may substitute $\mathcal{C}_{1j}$ and $\mathcal{C}_{1k}$ in Eq. (C.37) with their average over those positions, respectively equal to $\phi_1\phi_j$ and $\phi_1\phi_k$. We obtain

$$\Xi_3(U) = -4\,\phi_1\,\Xi_2(U) = -4\phi_1\ , \tag{C.41}$$

in the large-N limit, see calculation of $\Xi_2(U)$ above.

Expression of Ξ: Gathering the three terms above, we obtain

$$\Xi(U) = 1 + 4\,U - \frac{4}{g(U)}\ . \tag{C.42}$$

C.2.2.2 *Large-p behavior of the critical capacity*

The rest of the steps are the same presented in Section 4.6, in fact now we can directly write the set of coupled equations for x and the resolvent g:

$$g = \sum_{\rho=1}^{M} q_\rho\ , \tag{C.43}$$

$$\frac{\beta_\rho}{q_\rho} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}_\rho(\mathbf{k})}{1 + pH(x)\sum_{\theta=1}^{M}\hat{\Gamma}_\theta(\mathbf{k})q_\theta},\ \forall\rho\ , \tag{C.44}$$

$$1 - \frac{4}{g} = x\sqrt{2\pi}\,H(x)\,e^{x^2/2}\ . \tag{C.45}$$

from which the capacity can be computed as a function of the number p of points,

$$\alpha_c(p) = \frac{1}{p\,H(x)}\ . \tag{C.46}$$

In practice, we can choose x at will, compute g from (C.45), then p from (C.43) and (C.44), and, finally, α_c from (C.46).

Now we make the hypothesis that $q_\rho = g\,r_\rho$ with $r_\rho = O(1)\,\forall\rho$ so that we can write (C.43) and (C.44) as:

$$1 = \sum_{\rho=1}^{M} r_\rho\ , \tag{C.47}$$

$$\frac{\beta_\rho}{g\,r_\rho} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}_\rho(\mathbf{k})}{1 + pH(x)g\sum_{\theta=1}^{M}\hat{\Gamma}_\theta(\mathbf{k})r_\theta},\ \forall\rho\ . \tag{C.48}$$

Summing all the equations (C.48) we get:

$$\frac{1}{g} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\sum_{\rho=1}^{M} \hat{\Gamma}_\rho(\mathbf{k}) r_\rho}{1 + pH(x) g \sum_{\theta=1}^{M} \hat{\Gamma}_\theta(\mathbf{k}) r_\theta} \,. \tag{C.49}$$

According to dimensional analysis, the large momentum scaling of the Fourier coefficients is given by

$$\hat{\Gamma}_\rho(\mathbf{k}) \sim \frac{\phi_\rho^2}{\left(k\,\phi_\rho^{\frac{1}{D}}\right)^{D+1}} = \frac{\phi_\rho^{1-\frac{1}{D}}}{k^{D+1}} \,,\ \forall \rho \,, \tag{C.50}$$

where $k = |\mathbf{k}|$ and D is the dimension of the physical space. Using (C.50), equation (C.49) can be rewritten as

$$p\,H(x) = G\Big(g\,p\,H(x) \sum_\rho \phi_\rho^{1-\frac{1}{D}} r_\rho\Big) \qquad \text{with} \qquad G(y) = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{y}{k^{D+1} + y} \,. \tag{C.51}$$

We deduce that, for large arguments y,

$$G(y) \sim A_1(D)\, y^{\frac{D}{D+1}} \quad \text{with} \qquad A_1(D) = \int \frac{d^D\mathbf{u}}{|\mathbf{u}|^{D+1} + 1} \,. \tag{C.52}$$

In addition, using the asymptotic expansion of the erfc function, we have

$$x\sqrt{2\pi}\,H(x)\, e^{x^2/2} \simeq 1 - \frac{1}{x^2} \tag{C.53}$$

for large x. Combining these expressions allows us to obtain the asymptotic relation between x and y,

$$y^{\frac{1}{D+1}} = 4\,A_1(D) \sum_\rho \phi_\rho^{1-\frac{1}{D}} r_\rho\, x^2 \,. \tag{C.54}$$

and, to the leading order in p,

$$x \simeq \sqrt{2\log p} - \big(D + \frac{1}{2}\big) \frac{\log\log p}{\sqrt{2\log p}} \,. \tag{C.55}$$

We then deduce the asymptotic scaling of the critical capacity given by

$$\alpha_c(p) \sim \frac{A(D)}{(\sum_\rho \phi_\rho^{1-\frac{1}{D}} r_\rho)^D (\log p)^D} \qquad (p \to \infty) \,, \tag{C.56}$$

with

$$A(D) = \frac{1}{8^D \, A_1(D)^{D+1}} \, . \tag{C.57}$$

Now in order to show that the hypothesis of $r_\rho = O(1) \, \forall \rho$ is consistent we notice using (C.50) that (C.48) can be rewritten in the following way:

$$\beta_\rho \simeq C \, r_\rho \, \phi_\rho^{1-\frac{1}{D}}, \ \forall \rho \, , \tag{C.58}$$

where C is fixed using (C.47) so that in the end we get:

$$r_\rho \simeq \frac{\beta_\rho \, \phi_\rho^{\frac{1}{D}-1}}{\sum_\theta \beta_\theta \, \phi_\theta^{\frac{1}{D}-1}}, \forall \rho \, , \tag{C.59}$$

and the hypothesis it's verified. Moreover, once we know (C.59) we can rewrite the asymptotic scaling of the critical capacity as

$$\alpha_c(p) \sim \frac{A(D)(\sum_\rho \phi_\rho^{\frac{1}{D}-1} \beta_\rho)^D}{(\log p)^D} \qquad (p \to \infty) \, , \tag{C.60}$$

where we recover trivially the result in Section 4.6 for $M = 1$. Notice also that in $D = 1$ the critical capacity does not depend on the distribution of the sizes of the PFs.

C.2.3 *Spectrum of MERM: multi-populations of neurons (multiple PFs per neuron on a map)*

Let's consider the following MERM:

$$M_{ij}^{(L)} = \frac{1}{N} \sum_{\ell=1}^{L} \sum_{m=1}^{c_i} \sum_{m'=1}^{c_j} \Gamma\left(\left| \mathbf{r}_{i,m}^{\ell} - \mathbf{r}_{j,m'}^{\ell} \right| \right) \, , \tag{C.61}$$

in which the N neurons are divided in M finite groups with fractions of neurons β_μ with $\mu = 1, \ldots, M$. Every group of neurons has a specific property, *i.e.*, c_μ PFs per map. The area of the PFs is fixed to $\phi_0 < 1$ for all the neurons for simplicity.

We are going to compute the resolvent of $\mathbf{M}^{(L)}$ using the replica method coming from statistical physics of disordered systems.

We start by rewriting the definition of the resolvent as

$$s_L(z) = \frac{1}{N} \left\langle \text{Trace} \left(\mathbf{M}^{(L)} - z \, \mathbf{Id} \right)^{-1} \right\rangle = \frac{2}{N} \partial_z \left\langle \log \det \left(\mathbf{M}^{(L)} - z \, \mathbf{Id} \right)^{-\frac{1}{2}} \right\rangle \, , \tag{C.62}$$

where $\langle\cdot\rangle$ stands for the average over the distribution of the matrix (C.61). With this representation the determinant $\det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}}$ can be expressed as a canonical partition function:

$$\mathcal{Z}_L(s) = \det\left(\mathbf{M}^{(L)} - z\,\mathbf{Id}\right)^{-\frac{1}{2}} = \int \prod_i \frac{d\phi_i}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_i \phi_i^2 - \frac{1}{2}\sum_{ij}\phi_i M_{ij}^{(L)}\phi_j\right), \quad \text{(C.63)}$$

where i, j go from 1 to N. The resolvent (C.62) can be calculated using the replica trick [157]:

$$s_L(z) = \frac{2}{N}\partial_z\left\langle \log \mathcal{Z}_L(s)\right\rangle = \frac{2}{N}\partial_z\left[\lim_{n\to 0}\frac{1}{n}\log\left\langle \mathcal{Z}_L(s)^n\right\rangle\right] \quad \text{(C.64)}$$

with

$$\left\langle \mathcal{Z}_L(s)^n\right\rangle = \int \prod_{ia}\frac{d\phi_i^a}{\sqrt{2\pi}} \exp\left(\frac{z}{2}\sum_a\sum_i(\phi_i^a)^2\right)\left\langle \exp\left(-\frac{1}{2}\sum_a\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\right)\right\rangle, \quad \text{(C.65)}$$

where we have replicated the system n times, *i.e.*, a goes from 1 to n.

In order to perform the average in (C.65) we rewrite (C.61) by considering the ℓ-th space matrix in its eigenbasis:

$$M_{ij}^{(L)} = \sum_\ell \sum_{\mathbf{k}\neq\mathbf{o}} v_{\mathbf{k}i}^\ell\, \hat{\Gamma}(\mathbf{k})\, v_{\mathbf{k}j}^\ell\,, \quad \text{(C.66)}$$

where ℓ goes from 1 to L, and the sum over $\mathbf{k}$ discards the $\mathbf{k} = \mathbf{o}$ extensive mode as discussed in Chapter 5. The eigenvector components, $v_{\mathbf{k}i}^\ell \simeq \frac{1}{\sqrt{N}}\sum_{m=1}^{c_i}\sin\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_{i,m}^\ell\right)$, $\frac{1}{\sqrt{N}}\sum_{m=1}^{c_i}\cos\left(2\pi\,\mathbf{k}\cdot\mathbf{r}_{i,m}^\ell\right)$, are real due to the symmetry $\hat{\Gamma}(\mathbf{k}) = \hat{\Gamma}(-\mathbf{k})$. Hence we get

$$\left\langle \exp\left(-\frac{1}{2}\sum_a\sum_{ij}\phi_i^a M_{ij}^{(L)}\phi_j^a\right)\right\rangle = \left\langle \exp\left(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{o}}\hat{\Gamma}(\mathbf{k})\Big(\sum_i v_{\mathbf{k}i}^\ell\phi_i^a\Big)^2\right)\right\rangle. \quad \text{(C.67)}$$

We now use the Stratonovich trick to linearize $\left(\sum_i v_{\mathbf{k}i}^\ell\phi_i^a\right)^2$:

$$\left\langle \exp\left(-\frac{1}{2}\sum_{a,\ell,\mathbf{k}\neq\mathbf{o}}\hat{\Gamma}(\mathbf{k})\Big(\sum_i v_{\mathbf{k}i}^\ell\phi_i^a\Big)^2\right)\right\rangle = \prod_\ell\int\prod_{a,\mathbf{k}\neq\mathbf{o}}\frac{du_{\ell,\mathbf{k}}^a}{\sqrt{2\pi}} \quad \text{(C.68)}$$

$$\times\exp\left(-\frac{1}{2}\sum_{a,\mathbf{k}\neq\mathbf{o}}\left(u_{\ell,\mathbf{k}}^a\right)^2\right)\left\langle \exp\left(-i\sum_{a,\mathbf{k}\neq\mathbf{o}}\sqrt{\hat{\Gamma}(\mathbf{k})}u_{\ell\mathbf{k}}^a\sum_i\phi_i^a v_{\mathbf{k}i}^\ell\right)\right\rangle.$$

Using the fact that $\langle v_{\mathbf{k}i}^\ell\rangle = 0$ and $\langle v_{\mathbf{k}i}^\ell v_{\mathbf{k}'j}^\ell\rangle = \frac{1}{N}c_i\delta_{ij}\delta_{\mathbf{k}\mathbf{k}'}$ it is easy to perform the average in the above equation, with the result

$$\left\langle \exp\left(-i\sum_{a,\mathbf{k}\neq\mathbf{o}}\sqrt{\hat{\Gamma}(\mathbf{k})}u_{\ell\mathbf{k}}^a\sum_i\phi_i^a v_{\mathbf{k}i}^\ell\right)\right\rangle = \exp\left(-\frac{1}{2}\sum_{a,b}\sum_{\mathbf{k}\neq\mathbf{o}}\hat{\Gamma}(\mathbf{k})u_{\ell\mathbf{k}}^a u_{\ell\mathbf{k}}^b\sum_{\mu=1}^{M}q_\mu^{ab}c_\mu\right)$$

(C.69)

where we have used the fact that the N neurons are divided in M groups and we have defined an overlap q_μ^{ab} per group as

$$q_\mu^{ab} = \frac{1}{N} \sum_{i \in \beta_\mu N} \phi_i^a \phi_i^b, \ \forall \mu \tag{C.70}$$

to be fixed through

$$1 = \int \prod_{a \leqslant b} \frac{d\hat{q}_\mu^{ab} dq_\mu^{ab}}{\frac{2\pi i}{N}} \exp\left(N \sum_{a \leqslant b} \hat{q}_\mu^{ab} q_\mu^{ab} - \sum_{a \leqslant b} \hat{q}_\mu^{ab} \sum_{i \in \beta_\mu N} \phi_i^a \phi_i^b \right), \forall \mu \, . \tag{C.71}$$

We can finally write $\left\langle \mathcal{Z}_L(s)^n \right\rangle$ as

$$\begin{aligned} \int \prod_\mu \prod_{a \leqslant b} \frac{d\hat{q}_\mu^{ab} dq_\mu^{ab}}{\frac{2\pi i}{N}} \exp \Bigg\{ N \Bigg[\sum_\mu \beta_\mu \log \int \prod_a \frac{d\phi^a}{\sqrt{2\pi}} \\ \exp\left(\frac{z}{2} \sum_a (\phi^a)^2 - \sum_{a \leqslant b} \hat{q}_\mu^{ab} \phi^a \phi^b \right) + \sum_\mu \sum_{a \leqslant b} \hat{q}_\mu^{ab} q_\mu^{ab} \\ + \alpha \log \int \prod_{\mathbf{k} \neq \mathbf{0}, a} \frac{du_{\mathbf{k}}^a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \sum_{\mathbf{k} \neq \mathbf{0}, a} (u_{\mathbf{k}}^a)^2 - \frac{1}{2} \sum_{\mathbf{k} \neq \mathbf{0}} \hat{\Gamma}(\mathbf{k}) \sum_{a \leqslant b} u_{\mathbf{k}}^a u_{\mathbf{k}}^b \sum_\mu q_\mu^{ab} c_\mu \right) \Bigg] \Bigg\} \end{aligned} \tag{C.72}$$

The Gaussian integrals over ϕ^a and $u_{\mathbf{k}}^a$ can be easily computed. We then make the Replica Symmetric (RS) Ansatz on the structure of the order parameters q_μ^{ab} and their conjugate variables $\hat{q}_\mu^{ab}$, so that

$$q_\mu^{ab} = r_\mu + (q_\mu - r_\mu)\delta_{ab}, \ \forall \mu \tag{C.73}$$

and

$$\hat{q}_\mu^{ab} = \hat{r}_\mu + (\hat{q}_\mu - \hat{r}_\mu)\delta_{ab}, \ \forall \mu. \tag{C.74}$$

The integrals over q_μ, r_μ, $\hat{q}_\mu$ and $\hat{r}_\mu$ are then estimated using the saddle-point method valid for large N, and then taking the small n limit. The resulting expression for the resolvent $s_L(z)$ of (C.61) is

$$2\partial_z \left[\mathrm{opt}_{\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}} \lim_{n \to 0} \lim_{N \to \infty} \frac{1}{nN} \log \left\langle \mathcal{Z}_L(s)^n \right\rangle \right] = 2\partial_z \left[\mathrm{opt}_{\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}} f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\}) \right], \tag{C.75}$$

where $f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\})$ is the free energy density equal to

$$\sum_\mu \hat{q}_\mu q_\mu - \frac{1}{2}\sum_\mu \hat{r}_\mu r_\mu - \frac{\alpha}{2}\sum_{\mathbf{k}\neq\mathbf{0}}\left[\log\left(1+\hat{\Gamma}(\mathbf{k})\sum_\mu c_\mu(q_\mu - r_\mu)\right) + \frac{\hat{\Gamma}(\mathbf{k})\sum_\mu c_\mu r_\mu}{1+\hat{\Gamma}(\mathbf{k})\sum_\mu c_\mu(q_\mu - r_\mu)}\right] \\ -\frac{1}{2}\sum_\mu \beta_\mu\left[\log\left(2\hat{q}_\mu - \hat{r}_\mu - z\right) + \frac{\hat{r}_\mu}{2\hat{q}_\mu - \hat{r}_\mu - z}\right] \tag{C.76}$$

The saddle-point equations obtained by optimizing $f(\{q_\mu, r_\mu, \hat{q}_\mu, \hat{r}_\mu\})$ with respect to $\hat{q}_\mu$, $\hat{r}_\mu$, q_μ and r_μ read $\forall\mu$

$$\frac{q_\mu}{\beta_\mu} = -\frac{\hat{r}_\mu}{(2\hat{q}_\mu - \hat{r}_\mu - z)^2} + \frac{1}{2\hat{q}_\mu - \hat{r}_\mu - z}\,, \quad r_\mu = -\frac{\beta_\mu \hat{r}_\mu}{(2\hat{q}_\mu - \hat{r}_\mu - z)^2}\,, \\ \hat{q}_\mu = \frac{\alpha}{2}\sum_{\mathbf{k}\neq\mathbf{0}}\left(\frac{\hat{\Gamma}(\mathbf{k})c_\mu}{1+\hat{\Gamma}(\mathbf{k})\sum_\nu c_\nu(q_\nu - r_\nu)} - \frac{\hat{\Gamma}(\mathbf{k})^2 c_\mu \sum_\nu r_\nu c_\nu}{(1+\hat{\Gamma}(\mathbf{k})\sum_\nu c_\nu(q_\nu - r_\nu))^2}\right), \\ \hat{r}_\mu = -\alpha\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\hat{\Gamma}(\mathbf{k})^2 c_\mu \sum_\nu r_\nu\, c_\nu}{(1+\hat{\Gamma}(\mathbf{k})\sum_\nu c_\nu(q_\nu - r_\nu))^2}\,. \tag{C.77}$$

This system of equations admits $r_\mu = \hat{r}_\mu = 0$, $\forall\mu$ as a solution, which gives, according to (C.75), the following system of equations satisfied by $s_L(z)$:

$$\begin{cases} s_L(z) = \sum_\mu q_\mu\,, \\ z = -\frac{\beta_\mu}{q_\mu} + \alpha\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\hat{\Gamma}(\mathbf{k})c_\mu}{1+\hat{\Gamma}(\mathbf{k})\sum_\nu q_\nu c_\nu}, \ \forall\mu\,. \end{cases} \tag{C.78}$$

Note that we are eventually interested in the spectral properties of the matrix $\mathbf{C}$ with entries

$$C_{ij} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{c_i}\sum_{m'=1}^{c_j}\Gamma\left(\left|\mathbf{r}_{i,m}^{\ell} - \mathbf{r}_{j,m'}^{\ell}\right|\right) = \frac{1}{\alpha}M_{ij}^{(L)}\,. \tag{C.79}$$

Obviously, the resolvent s of $\mathbf{C}$ is related to the resolvent s_L of $\mathbf{M}^{(L)}$ through the equation $s(z) = \alpha\, s_L(\alpha z)$. Hence we obtain our fundamental system of equations for the resolvent of $\mathbf{C}$:

$$\begin{cases} s(z) = \sum_\mu q_\mu\,, \\ z = -\frac{\beta_\mu}{q_\mu} + \sum_{\mathbf{k}\neq\mathbf{0}}\frac{\alpha\hat{\Gamma}(\mathbf{k})c_\mu}{\alpha+\hat{\Gamma}(\mathbf{k})\sum_\nu q_\nu c_\nu}, \ \forall\mu\,. \end{cases} \tag{C.80}$$

C.2.4 *Quenched PF theory: multi-populations of neurons (multiple PFs per neuron on a map)*

Here we are going to extend the Gaussian theory with quenched PF developed in Section 4.6 to the case of multi-populations of neurons and multiple PFs per neuron in the same map. By multi-populations of neurons we mean that the N neurons are divided in M finite groups with fractions of neurons β_ρ with $\rho = 1, \dots, M$ and every group of neurons with a specific property, *i.e.*, c_ρ PFs with area $\phi_0 < 1$.

The computation follows exactly what seen in Section 4.6 until the definition of the order parameters that are now

$$m_\ell^a = \sum_{j \geqslant 2} W_{ja} \left(2 \sum_{m=1}^{c_1} \sum_{m'=1}^{c_j} \Gamma\big(|\mathbf{r}_{1,m}^\ell - \mathbf{r}_{j,m'}^\ell|\big) - c_j \phi_0 \right) \tag{C.81}$$

and

$$q_\ell^{ab} = \sum_{j,k \geqslant 2} W_{ja} W_{kb} \sum_{m=1}^{c_j} \sum_{m'=1}^{c_k} \Gamma\big(|\mathbf{r}_{j,m}^\ell - \mathbf{r}_{k,m'}^\ell|\big) \, . \tag{C.82}$$

Now we make the same approximations as Section 4.6 and so we can write the final expression of the order parameters after adding up all the maps:

$$m^a \equiv \frac{1}{L} \sum_{\ell=1}^{L} m_\ell^a = \sum_{j \geqslant 2} W_{ja} \left(2\, \mathcal{C}_{1j}\big(\{\mathbf{r}_{j,m}^\ell\}\big) - c_j \phi_0 \right) \tag{C.83}$$

and

$$q^{ab} \equiv \frac{1}{L} \sum_{\ell=1}^{L} q_\ell^{ab} = \sum_{j,k \geqslant 2} W_{ja} W_{kb} \, \mathcal{C}_{jk}\big(\{\mathbf{r}_{j,m}^\ell\}\big) \, . \tag{C.84}$$

The $N \times N$ multi-space Euclidean random matrix $\mathcal{C}$ appearing in the expressions above is defined as

$$C_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \sum_{m=1}^{c_i} \sum_{m'=1}^{c_j} \Gamma\left(\left| \mathbf{r}_{i,m}^\ell - \mathbf{r}_{j,m'}^\ell \right| \right) \, . \tag{C.85}$$

In the following, we denote by $\rho(\lambda)$ the density of eigenvalues λ of $\mathcal{C}$. This density is self-averaging when the PFs are randomly drawn in the large L, N double limit. Its resolvent, defined as

$$g(U) = \int d\lambda \, \frac{\rho(\lambda)}{\lambda + U} \, , \tag{C.86}$$

where the integral runs over the support of ρ, is solution of the following system of equations, see C.2.3:

$$\begin{cases} g(u) = \sum_{\rho=1}^{M} q_\rho \ , \\ u = -\frac{\beta_\mu}{q_\mu} + \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\alpha \hat{\Gamma}(\mathbf{k}) c_\mu}{\alpha + \hat{\Gamma}(\mathbf{k}) \sum_\nu q_\nu c_\nu}, \ \forall \rho \ . \end{cases} \tag{C.87}$$

From here on we obtain the same equations of what we saw in Section 4.6 except for the quantity Ξ defined as

$$\Xi(u) = \sum_{j,k\geqslant 2} H_j \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} H_k \qquad \text{with} \qquad H_j = 2\,\mathcal{C}_{1j} - c_j \phi_0 \ , \tag{C.88}$$

and $\mathbf{Id}$ is the identity matrix. In the above equation, the inverse is intended over the $N-1$-dimensional restriction of the matrix $u\,\mathbf{Id} + \mathcal{C}$ to entries $j, k \geqslant 2$.

C.2.4.1 *Computation of* Ξ

Expanding the terms in $\Xi(u)$ in Eq. (C.88) above, we write $\Xi(u) = \Xi_1(u) + \Xi_2(u) + \Xi_3(u)$ with

$$\Xi_1(u) = 4 \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} \mathcal{C}_{1k} \ , \tag{C.89}$$

$$\Xi_2(u) = \phi_0^2 \sum_{j,k\geqslant 2} c_j \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} c_k \ , \tag{C.90}$$

$$\Xi_3(u) = -2\phi_0 \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} c_k \ - 2\phi_0 \sum_{j,k\geqslant 2} c_j \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} \mathcal{C}_{1k} \ . \tag{C.91}$$

Computation of Ξ_1: Consider the $N \times N$ matrix $\mathcal{C}^{(N)}$, with entries $\mathcal{C}_{ij}$ for i, j comprised between 1 and N. Let us also define $\mathbf{Id}^{(N)}$ the identity matrix in dimension N, while $\mathbf{Id}$ above referred to the identity matrix in dimension $N-1$. Using block-matrix inversion formulas, we write that

$$\left(u\, \mathbf{Id}^{(N)} + \mathcal{C}^{(N)} \right)^{-1}_{11} = \frac{1}{u + \mathcal{C}_{11} - \sum_{j,k\geqslant 2} \mathcal{C}_{1j} \left(u\, \mathbf{Id} + \mathcal{C} \right)^{-1}_{jk} \mathcal{C}_{1k}} \tag{C.92}$$

The left hand side of the equation above is equal, in the large–N limit, to the resolvent $g(u)$ of $\mathcal{C}$ defined in (C.86). Using $\mathcal{C}_{11} = c_1 \phi_0$ and the definition of $\Xi_1(u)$, we obtain

$$\Xi_1(u) = 4 \left(u + c_1 \phi_0 - \frac{1}{g(u)} \right) \ . \tag{C.93}$$

Computation of Ξ_2: Let $|v_+\rangle$ be the normalized vector with N identical components, $(v_+)_i = \frac{1}{\sqrt{N}}$. For large N, $|v_+\rangle$ is the top eigenvector of $\mathcal{C}$, with (extensive) eigenvalue $\lambda_+ = N\phi_0^2 \sum_{\rho=1}^{M} \beta_\rho c_\rho^2$. After the computation of the braket of this vector with the matrix of components $\phi_0^2 c_j \left(U\,\mathbf{Id} + \mathcal{C}\right)^{-1}_{jk} c_k$, and taking the large N limit we get:

$$\Xi_2(U) = \frac{N\phi_0^2 \sum_{\rho=1}^{M} \beta_\rho c_\rho^2}{U + N\phi_0^2 \sum_{\rho=1}^{M} \beta_\rho c_\rho^2} \to 1 \,. \tag{C.94}$$

Computation of Ξ_3: As $\mathcal{C}_{jk}$ with $j,k \geqslant 2$ does not depend on the locations $\mathbf{r}^{\ell}_{1,m}$ of the place fields associated to neuron $i = 1$ in the different maps ℓ, we may substitute $\mathcal{C}_{1j}$ and $\mathcal{C}_{1k}$ in Eq. (C.91) with their average over those positions, respectively equal to $\phi_0^2 c_1 c_j$ and $\phi_0^2 c_1 c_k$. We obtain

$$\Xi_3(U) = -4\,\phi_0 c_1\,\Xi_2(U) = -4\,\phi_0 c_1 \,, \tag{C.95}$$

in the large-N limit, see calculation of $\Xi_2(U)$ above.

Expression of Ξ: Gathering the three terms above, we obtain

$$\Xi(U) = 1 + 4\,U - \frac{4}{g(U)} \,. \tag{C.96}$$

C.2.4.2 *Large-p behavior of the critical capacity*

The rest of the steps are the same presented in Section 4.6, in fact now we can write the set of equations coupled for x and the resolvent g:

$$g = \sum_{\rho=1}^{M} q_\rho \,, \tag{C.97}$$

$$\frac{\beta_\rho}{q_\rho} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}(\mathbf{k}) c_\rho}{1 + pH(x)\hat{\Gamma}(\mathbf{k}) \sum_{\theta=1}^{M} c_\theta q_\theta}, \ \forall \rho \,, \tag{C.98}$$

$$1 - \frac{4}{g} = x\sqrt{2\pi}\,H(x)\,e^{x^2/2} \,. \tag{C.99}$$

from which the capacity can be computed as a function of the number p of points,

$$\alpha_c(p) = \frac{1}{p\,H(x)} \,. \tag{C.100}$$

In practice, we can choose x at will, compute g from (C.99), then p from (C.97) and (C.98), and, finally, α_c from (C.100).

Now we make the hypothesis that $q_\rho = g\, r_\rho$ with $r_\rho = O(1)\, \forall \rho$ so that we can write (C.97) and (C.98) as:

$$1 = \sum_{\rho=1}^{M} r_\rho \ , \tag{C.101}$$

$$\frac{\beta_\rho}{g\, r_\rho} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}(\mathbf{k}) c_\rho}{1 + p H(x) g \hat{\Gamma}(\mathbf{k}) \sum_{\theta=1}^{M} c_\theta r_\theta}, \ \forall \rho \ . \tag{C.102}$$

Summing all the equations (C.102) we get:

$$\frac{1}{g} = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{\hat{\Gamma}(\mathbf{k}) \sum_{\rho=1}^{M} c_\rho r_\rho}{1 + p H(x) g \hat{\Gamma}(\mathbf{k}) \sum_{\theta=1}^{M} c_\theta r_\theta} \ . \tag{C.103}$$

According to dimensional analysis, the large momentum scaling of the Fourier coefficients is given by

$$\hat{\Gamma}(\mathbf{k}) \sim \frac{\phi_0^2}{\left(k\, \phi_0^{\frac{1}{D}}\right)^{D+1}} = \frac{\phi_0^{1-\frac{1}{D}}}{k^{D+1}} \ , \tag{C.104}$$

where $k = |\mathbf{k}|$ and D is the dimension of the physical space. Using (C.104), equation (C.103) can be rewritten as

$$p\, H(x) = G\Big(g\, p\, H(x) \phi_0^{1-\frac{1}{D}} \sum_\rho c_\rho r_\rho\Big) \qquad \text{with} \qquad G(y) = \sum_{\mathbf{k}\neq\mathbf{0}} \frac{y}{k^{D+1} + y} \ . \tag{C.105}$$

We deduce that, for large arguments y,

$$G(y) \sim A_1(D)\, y^{\frac{D}{D+1}} \quad \text{with} \qquad A_1(D) = \int \frac{d^D \mathbf{u}}{|\mathbf{u}|^{D+1} + 1} \ . \tag{C.106}$$

In addition, using the asymptotic expansion of the erfc function, we have

$$x\sqrt{2\pi}\, H(x)\, e^{x^2/2} \simeq 1 - \frac{1}{x^2} \tag{C.107}$$

for large x. Combining these expressions allows us to obtain the asymptotic relation between x and y,

$$y^{\frac{1}{D+1}} = 4\, A_1(D)\, \phi_0^{1-\frac{1}{D}} \sum_\rho c_\rho r_\rho\, x^2 \ . \tag{C.108}$$

and, to the leading order in p,

$$x \simeq \sqrt{2 \log p} - \Big(D + \frac{1}{2}\Big) \frac{\log\log p}{\sqrt{2 \log p}} \ . \tag{C.109}$$

We then deduce the asymptotic scaling of the critical capacity given by

$$\alpha_c(p) \sim \frac{A(D)\phi_0^{-(D-1)}}{(\sum_\rho c_\rho r_\rho)^D (\log p)^D} \qquad (p \to \infty) \,, \tag{C.110}$$

with

$$A(D) = \frac{1}{8^D \, A_1(D)^{D+1}} \,. \tag{C.111}$$

Now in order to show that the hypothesis of $r_\rho = O(1) \, \forall \rho$ is consistent we notice using (C.104) that (C.102) can be rewritten in the following way:

$$\beta_\rho \simeq C \, r_\rho \, c_\rho, \ \forall \rho \,, \tag{C.112}$$

where C is fixed using (C.101) so that in the end we get:

$$r_\rho \simeq \frac{\beta_\rho}{c_\rho} \frac{1}{\sum_\theta \frac{\beta_\rho}{c_\rho}}, \forall \rho \,, \tag{C.113}$$

and the hypothesis it's verified. Moreover, once we know (C.113) we can rewrite the asymptotic scaling of the critical capacity as

$$\alpha_c(p) \sim \frac{A(D)\phi_0^{-(D-1)} (\sum_\rho \frac{\beta_\rho}{c_\rho})^D}{(\log p)^D} \qquad (p \to \infty) \,, \tag{C.114}$$

where we recover trivially the result in Section 4.6 for $M = 1$ and $c_1 = 1$.

C.3 ADATRON ALGORITHM

Until now we have always used offline algorithms to solve SVM problems where the patterns to be stored were presented all together to the network in an unrealistic way, using standard packages to solve convex optimization problems [68, 195], see Section A.1. Here we show a version of the SVM algorithm of online nature in which the patterns are presented one at a time (more biologically plausible) and that would allow us to study questions like what is the best way to present the patterns in order to stabilize them as soon as possible, or even to present the patterns in such a way to represent the realistic trajectories of an animal that explores one after the other the different environments. So this is a first step for the study of learning dynamics in our model. Fortunately there are already in literature algorithms of this type [21, 71, 132] and in particular in the following we will consider the adatron algorithm of which we report here the details.

The adatron algorithm is equivalent to the support vector machine in the sense that it converges to the same solution of the perceptron with optimal stability (maximum

margin), but with the big difference that it is an online scheme inspired by the perceptron algorithm itself, therefore more biologically plausible.

Below we report the algorithm, referring to [21, 71] for details on the algorithm convergence demonstrations and links with other schemes such as minover or adaline.

Let's consider the case of one of the RNN perceptrons for which we need to find the optimal weights (as usual the different perceptrons are independent so the N problems can be solved separately), see Section 4.3, say the one corresponding to neuron $i = 1$.

Let's start by writing the net weights in the following way

$$W_{1j} = \frac{1}{N} x_1^{\ell,\mu} (2\xi_1^{\ell,\mu} - 1)\xi_j^{\ell,\mu} \tag{C.115}$$

where $x_1^{\ell,\mu}$ are called the embedding strengths.

We can choose any value for $x_1^{\ell,\mu}$ as long as they are non-negative, including the tabula rasa case where they are all null.

Once the problem is initialized we start by presenting to the network one pattern at a time in a sequential way (the algorithm can also be implemented in parallel) and we update the $x_1^{\ell,\mu}$ according to the following rule

$$\delta x_1^{\ell,\mu} = \max\{-x_1^{\ell,\mu}, \eta(1 - \Delta_1^{\ell,\mu})\} \tag{C.116}$$

where the η (learning rate) range must be between 0 and 2 to ensure convergence of the algorithm and $\Delta_1^{\ell,\mu}$ is defined as usual as

$$\Delta_1^{\ell,\mu} = (2\xi_1^{\ell,\mu} - 1) \sum_{j \neq i} W_{ij} \xi_j^{\ell,\mu}. \tag{C.117}$$

Presenting several times the patterns to the net following this rule we will arrive to convergence when the following conditions (Kuhn-Tucker conditions, see [43]) are satisfied, that is

$$\text{either} \quad (x_1^{\ell,\mu} > 0 \quad \text{and} \quad \Delta_1^{\ell,\mu} = 1) \quad \text{or} \quad (x_1^{\ell,\mu} = 0 \quad \text{and} \quad \Delta_1^{\ell,\mu} \geqslant 1)\,, \tag{C.118}$$

for all the patterns.

These conditions are equivalent as those satisfied by the SVM algorithm, in fact once we have reached convergence and normalized to one the weights, we find exactly the same results. In addition, the speed of convergence to the perceptron with maximum stability of this algorithm is exponential.

BIBLIOGRAPHY

[1] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. « Statistical mechanics of complex neural systems and high dimensional data. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.03 (2013), P03014.

[2] Elena Agliari, Alessia Annibale, Adriano Barra, ACC Coolen, and Daniele Tantari. « Immune networks: multitasking capabilities near saturation. » In: *Journal of Physics A: Mathematical and Theoretical* 46.41 (2013), p. 415003.

[3] Elena Agliari, Adriano Barra, Andrea De Antoni, and Andrea Galluzzi. « Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines. » In: *Neural Networks* 38 (2013), pp. 52–63.

[4] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, and Francesco Moauro. « Multitasking associative networks. » In: *Physical review letters* 109.26 (2012), p. 268101.

[5] Elena Agliari, Adriano Barra, Francesco Guerra, and Francesco Moauro. « A thermodynamic perspective of immune capabilities. » In: *Journal of theoretical biology* 287 (2011), pp. 48–63.

[6] Elena Agliari, Elena Biselli, Adele De Ninno, Giovanna Schiavoni, Lucia Gabriele, Anna Gerardino, Fabrizio Mattei, Adriano Barra, and Luca Businaro. « Cancer-driven dynamics of immune cells in a microfluidic environment. » In: *Scientific reports* 4.1 (2014), pp. 1–15.

[7] Manuela Allegra, Lorenzo Posani, Ruy Gómez-Ocádiz, and Christoph Schmidt-Hieber. « Differential Relation between Neuronal and Behavioral Discrimination during Hippocampal Memory Encoding. » In: *Neuron* (2020).

[8] Timothy A Allen, Daniel M Salz, Sam McKenzie, and Norbert J Fortin. « Nonspatial sequence coding in CA1 neurons. » In: *Journal of Neuroscience* 36.5 (2016), pp. 1547–1563.

[9] Charlotte B Alme, Chenglin Miao, Karel Jezek, Alessandro Treves, Edvard I Moser, and May-Britt Moser. « Place cells in the hippocampus: eleven maps for eleven rooms. » In: *Proceedings of the National Academy of Sciences* 111.52 (2014), pp. 18428–18435.

[10] Grégoire Altan-Bonnet, Thierry Mora, and Aleksandra M Walczak. « Quantitative immunology for physicists. » In: *Physics Reports* 849 (2020), pp. 1–83.

[11] David Amaral, Per Andersen, John O'Keefe, Richard Morris, et al. *The hippocampus book*. Oxford University Press, 2007.

[12] Shun-ichi Amari. « Dynamics of pattern formation in lateral-inhibition type neural fields. » In: *Biological cybernetics* 27.2 (1977), pp. 77–87.

[13] Ariel Amir, Yuval Oreg, and Yoseph Imry. « Localization, anomalous diffusion, and slow relaxations: A random distance matrix approach. » In: *Physical review letters* 105.7 (2010), p. 070601.

[14] Daniel J Amit and Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1992.

[15] Daniel J Amit, C Campbell, and KYM Wong. « The interaction space of neural networks with sign-constrained synapses. » In: *Journal of Physics A: Mathematical and General* 22.21 (1989), p. 4687.

[16] Daniel J Amit, MR Evans, H Horner, and KYM Wong. « Retrieval phase diagrams for attractor neural networks with optimal interactions. » In: *Journal of Physics A: Mathematical and General* 23.14 (1990), p. 3361.

[17] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. « Storing infinite numbers of patterns in a spin-glass model of neural networks. » In: *Physical Review Letters* 55.14 (1985), p. 1530.

[18] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. « Statistical mechanics of neural networks near saturation. » In: *Annals of physics* 173.1 (1987), pp. 30–67.

[19] Daniel J Amit, KY Michael Wong, and Colin Campbell. « Perceptron learning with sign-constrained weights. » In: *Journal of Physics A: Mathematical and General* 22.12 (1989), p. 2039.

[20] Per Andersen, Richard Morris, David Amaral, Tim Bliss, and John O'Keefe. *The hippocampus book*. Oxford university press, 2006.

[21] JK Anlauf and M Biehl. « The adatron: an adaptive perceptron algorithm. » In: *EPL (Europhysics Letters)* 10.7 (1989), p. 687.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. « Neural machine translation by jointly learning to align and translate. » In: *arXiv preprint arXiv:1409.0473* (2014).

[23] Omri Barak and Sandro Romani. « Mapping low-dimensional dynamics to high-dimensional neural activity: A derivation of the ring model from the neural engineering framework. » In: *arXiv preprint arXiv:2002.03420* (2020).

[24] Boris Barbour, Nicolas Brunel, Vincent Hakim, and Jean-Pierre Nadal. « What can we learn from synaptic weight distributions? » In: *TRENDS in Neurosciences* 30.12 (2007), pp. 622–629.

[25] Luciano Maria Barone and Enzo Marinari. *Scientific programming: C-language, algorithms and models in science*. World Scientific, 2014.

[26] Caswell Barry, Colin Lever, Robin Hayman, Tom Hartley, Stephen Burton, John O'Keefe, Kate Jeffery, and Neil Burgess. « The boundary vector cell model of place cell firing and spatial memory. » In: *Reviews in the Neurosciences* 17.1-2 (2006), p. 71.

[27] Francesco P Battaglia and Alessandro Treves. « Attractor neural networks storing multiple space representations: a model for hippocampal place fields. » In: *Physical Review E* 58.6 (1998), p. 7738.

[28] Aldo Battista and Rémi Monasson. « Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks. » In: *Physical Review Letters* 124.4 (2020), p. 048302.

[29] Aldo Battista and Rémi Monasson. « Spectrum of multispace Euclidean random matrices. » In: *Physical Review E* 101.5 (2020), p. 052133.

[30] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[31] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. « Theory of orientation tuning in visual cortex. » In: *Proceedings of the National Academy of Sciences* 92.9 (1995), pp. 3844–3848.

[32] Yoshua Bengio, Aaron Courville, and Pascal Vincent. « Representation learning: A review and new perspectives. » In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[33] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012.

[34] Elena Biselli, Elena Agliari, Adriano Barra, Francesca Romana Bertani, Annamaria Gerardino, Adele De Ninno, Arianna Mencattini, Davide Di Giuseppe, Fabrizio Mattei, Giovanna Schiavoni, et al. « Organs on chip approach: a tool to evaluate cancer-immune cells interactions. » In: *Scientific reports* 7.1 (2017), pp. 1–12.

[35] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[36] Tim VP Bliss and Graham L Collingridge. « A synaptic model of memory: long-term potentiation in the hippocampus. » In: *Nature* 361.6407 (1993), pp. 31–39.

[37] Tim VP Bliss and Terje Lømo. « Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. » In: *The Journal of physiology* 232.2 (1973), pp. 331–356.

[38] Charles Bordenave. « Eigenvalues of Euclidean random matrices. » In: *Random Structures & Algorithms* 33.4 (2008), pp. 515–532.

[39] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. « A training algorithm for optimal margin classifiers. » In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[40] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge university press, 2003.

[41] Romain Bourboulou, Geoffrey Marti, François-Xavier Michon, Elissa El Feghaly, Morgane Nouguier, David Robbe, Julie Koenig, and Jerome Epsztein. « Dynamic control of hippocampal spatial coding resolution by local visual cues. » In: *eLife* 8 (2019), e44487.

[42] Paul Bourgine and Jean-Pierre Nadal. *Cognitive economics: an interdisciplinary approach*. Springer Science & Business Media, 2013.

[43] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[44] Édouard Brézin, Vladimir Kazakov, Didina Serban, Paul Wiegmann, and Anton Zabrodin. *Applications of random matrices in physics*. Vol. 221. Springer Science & Business Media, 2006.

[45] Vegard Heimly Brun, Trygve Solstad, Kirsten Brun Kjelstrup, Marianne Fyhn, Menno P Witter, Edvard I Moser, and May-Britt Moser. « Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. » In: *Hippocampus* 18.12 (2008), pp. 1200–1212.

[46] Nicolas Brunel. « Is cortical connectivity optimized for storing information? » In: *Nature neuroscience* 19.5 (2016), p. 749.

[47] Nicolas Brunel, Vincent Hakim, Philippe Isope, Jean-Pierre Nadal, and Boris Barbour. « Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. » In: *Neuron* 43.5 (2004), pp. 745–757.

[48] Nicolas Brunel and Jean-Pierre Nadal. « Modeling memory: what do we learn from attractor neural networks? » In: *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie* 321.2-3 (1998), pp. 249–252.

[49] Daniel Bush, Caswell Barry, and Neil Burgess. « What do grid cells contribute to place cell firing? » In: *Trends in neurosciences* 37.3 (2014), pp. 136–145.

[50] C Campbell and A Robinson. « On the storage capacity of neural networks with sign-constrained weights. » In: *Journal of Physics A: Mathematical and General* 24.2 (1991), p. L93.

[51] Tommaso Castellani and Andrea Cavagna. « Spin-glass theory for pedestrians. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.05 (2005), P05012.

[52] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. « Statistical physics of social dynamics. » In: *Reviews of modern physics* 81.2 (2009), p. 591.

[53] Lawrence Cayton. « Algorithms for manifold learning. » In: *Univ. of California at San Diego Tech. Rep* 12.1-17 (2005), p. 1.

[54] Erika Cerasti and Alessandro Treves. « The spatial representations acquired in CA3 by self-organizing recurrent connections. » In: *Frontiers in cellular neuroscience* 7 (2013), p. 112.

[55] Subrahmanyan Chandrasekhar. « Stochastic problems in physics and astronomy. » In: *Reviews of modern physics* 15.1 (1943), p. 1.

[56] Guifen Chen, John A King, Neil Burgess, and John O'Keefe. « How vision and movement combine in the hippocampal place code. » In: *Proceedings of the National Academy of Sciences* 110.1 (2013), pp. 378–383.

[57] Longtang L Chen, Lie-Huey Lin, Edward J Green, Carol A Barnes, and Bruce L McNaughton. « Head-direction cells in the rat posterior cortex. » In: *Experimental brain research* 101.1 (1994), pp. 8–23.

[58] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. « Classification and geometry of general perceptual manifolds. » In: *Physical Review X* 8.3 (2018), p. 031003.

[59] S Ciliberti, TS Grigera, Víctor Martín-Mayor, G Parisi, and P Verrocchio. « Anderson localization in Euclidean random matrices. » In: *Physical Review B* 71.15 (2005), p. 153104.

[60] Simona Cocco, Rémi Monasson, Lorenzo Posani, Sophie Rosay, and Jérôme Tubiana. « Statistical physics and representations in real and artificial neural networks. » In: *Physica A: Statistical Mechanics and its Applications* 504 (2018), pp. 45–76.

[61] Ronan Collobert and Jason Weston. « A unified architecture for natural language processing: Deep neural networks with multitask learning. » In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 160–167.

[62] Corinna Cortes and Vladimir Vapnik. « Support-vector networks. » In: *Machine learning* 20.3 (1995), pp. 273–297.

[63] Jose A Costa and Alfred O Hero. « Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. » In: *2004 12th European Signal Processing Conference*. IEEE. 2004, pp. 369–372.

[64] Romain Couillet, Florent Benaych-Georges, et al. « Kernel spectral clustering of large dimensional data. » In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1393–1454.

[65] Brian G Cragg and H Nevill V Temperley. « The organisation of neurones: a cooperative analogy. » In: *Electroencephalography and clinical neurophysiology* 6 (1954), pp. 85–92.

[66] HH Dale. *Pharmacology and nerve endings*. 1934.

[67] Peter Dayan and Laurence F Abbott. « Theoretical neuroscience: computational and mathematical modeling of neural systems. » In: (2001).

[68] Steven Diamond and Stephen Boyd. « CVXPY: A Python-embedded modeling language for convex optimization. » In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2909–2913.

[69] Kamran Diba and György Buzsáki. « Forward and reverse hippocampal place-cell sequences during ripples. » In: *Nature neuroscience* 10.10 (2007), pp. 1241–1242.

[70] Noureddine El Karoui et al. « The spectrum of kernel random matrices. » In: *The Annals of Statistics* 38.1 (2010), pp. 1–50.

[71] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[72] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. « Dermatologist-level classification of skin cancer with deep neural networks. » In: *nature* 542.7639 (2017), pp. 115–118.

[73] Alberto Fachechi, Elena Agliari, and Adriano Barra. « Dreaming neural networks: forgetting spurious memories and reinforcing pure ones. » In: *Neural Networks* 112 (2019), pp. 24–40.

[74] U Farooq and G Dragoi. « Emergence of preconfigured and plastic time-compressed sequences in early postnatal development. » In: *Science* 363.6423 (2019), pp. 168–173.

[75] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. « Testing the manifold hypothesis. » In: *Journal of the American Mathematical Society* 29.4 (2016), pp. 983–1049.

[76] André A Fenton, Hsin-Yi Kao, Samuel A Neymotin, Andrey Olypher, Yevgeniy Vayntrub, William W Lytton, and Nandor 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. » In: *Journal of Neuroscience* 28.44 (2008), pp. 11250–11262.

[77] Arseny Finkelstein, Hervé Rouault, Sandro Romani, and Nachum Ulanovsky. « Dynamic control of cortical head-direction signal by angular velocity. » In: *bioRxiv* (2019), p. 730374.

[78] Konrad H Fischer and John A Hertz. *Spin glasses*. Vol. 1. Cambridge university press, 1993.

[79] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[80] CC Alan Fung, KY Michael Wong, He Wang, and Si Wu. « Dynamical synapses enhance neural information processing: gracefulness, accuracy, and mobility. » In: *Neural computation* 24.5 (2012), pp. 1147–1185.

[81] CC Alan Fung, KY Michael Wong, and Si Wu. « Dynamics of neural networks with continuous attractors. » In: *EPL (Europhysics Letters)* 84.1 (2008), p. 18002.

[82] CC Alan Fung, KY Michael Wong, and Si Wu. « Tracking dynamics of two-dimensional continuous attractor neural networks. » In: *J. Phys. Conf. Ser*. Vol. 197. 012017. 2009, p. 6596.

[83] CC Alan Fung, KY Michael Wong, and Si Wu. « A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks. » In: *Neural Computation* 22.3 (2010), pp. 752–792.

[84] Marianne Fyhn, Torkel Hafting, Alessandro Treves, May-Britt Moser, and Edvard I Moser. « Hippocampal remapping and grid realignment in entorhinal cortex. » In: *Nature* 446.7132 (2007), pp. 190–194.

[85] Marianne Fyhn, Sturla Molden, Menno P Witter, Edvard I Moser, and May-Britt Moser. « Spatial representation in the entorhinal cortex. » In: *Science* 305.5688 (2004), pp. 1258–1264.

[86] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. « Neural manifolds for the control of movement. » In: *Neuron* 94.5 (2017), pp. 978–984.

[87] Elizabeth Gardner. « The space of interactions in neural network models. » In: *Journal of physics A: Mathematical and general* 21.1 (1988), p. 257.

[88] Elizabeth Gardner and Bernard Derrida. « Optimal storage properties of neural network models. » In: *Journal of Physics A: Mathematical and general* 21.1 (1988), p. 271.

[89] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[90] Maya Geva-Sagiv, Sandro Romani, Liora Las, and Nachum Ulanovsky. « Hippocampal global remapping for different sensory modalities in flying bats. » In: *Nature neuroscience* 19.7 (2016), pp. 952–958.

[91] Roy J Glauber. « Time-dependent statistics of the Ising model. » In: *Journal of mathematical physics* 4.2 (1963), pp. 294–307.

[92] A Goetschy and SE Skipetrov. « Non-Hermitian Euclidean random matrix theory. » In: *Physical Review E* 84.1 (2011), p. 011150.

[93] A Goetschy and SE Skipetrov. « Euclidean random matrices and their applications in physics. » In: *arXiv preprint arXiv:1303.2880* (2013).

[94] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. « Modelling the influence of data structure on learning in neural networks. » In: *arXiv preprint arXiv:1909.11500* (2019).

[95] Edward J Golob and Jeffrey S Taube. « Head direction cells in rats with hippocampal or overlying neocortical lesions: evidence for impaired angular path integration. » In: *Journal of Neuroscience* 19.16 (1999), pp. 7198–7211.

[96] Ian Goodfellow, Y Bengio, and A Courville. « Machine learning basics. » In: *Deep learning* 1 (2016), pp. 98–164.

[97] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[98] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. « Generative adversarial nets. » In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[99] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. « Speech recognition with deep recurrent neural networks. » In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 6645–6649.

[100] TS Grigera, Víctor Martin-Mayor, G Parisi, P Urbani, and P Verrocchio. « On the high-density expansion for Euclidean random matrices. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2011.02 (2011), P02015.

[101] TS Grigera, Víctor Martín-Mayor, G Parisi, and P Verrocchio. « Vibrations in glasses and Euclidean random matrix theory. » In: *Journal of Physics: Condensed Matter* 14.9 (2002), p. 2167.

[102] Jacopo Grilli, György Barabás, and Stefano Allesina. « Metapopulation persistence in random fragmented landscapes. » In: *PLoS Comput Biol* 11.5 (2015), e1004251.

[103] Eyal Gruntman, Sandro Romani, and Michael B Reiser. « The computation of directional selectivity in the Drosophila OFF motion pathway. » In: *Elife* 8 (2019), e50706.

[104] Segundo Jose Guzman, Alois Schlögl, Michael Frotscher, and Peter Jonas. « Synaptic mechanisms of pattern completion in the hippocampal CA3 network. » In: *Science* 353.6304 (2016), pp. 1117–1123.

[105] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. « Microstructure of a spatial map in the entorhinal cortex. » In: *Nature* 436.7052 (2005), pp. 801–806.

[106] Jean-Pierre Hansen and Ian R McDonald. *Theory of simple liquids*. Elsevier, 1990.

[107] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. « Deep residual learning for image recognition. » In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[108] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[109] John A Hertz. *Introduction to the theory of neural computation*. CRC Press, 2018.

[110] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. « Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. » In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[111] Stig A Hollup, Sturla Molden, James G Donnett, May-Britt Moser, and Edvard I Moser. « Accumulation of hippocampal place fields at the goal location in an annular watermaze task. » In: *Journal of Neuroscience* 21.5 (2001), pp. 1635–1644.

[112] John J Hopfield. « Neural networks and physical systems with emergent collective computational abilities. » In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[113] Kerson Huang. *Statistical Mechanics*. 2nd ed. John Wiley & Sons, 1987.

[114] David H Hubel and Torsten N Wiesel. « Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. » In: *The Journal of physiology* 160.1 (1962), p. 106.

[115] David H Hubel and Torsten N Wiesel. « Receptive fields and functional architecture of monkey striate cortex. » In: *The Journal of physiology* 195.1 (1968), pp. 215–243.

[116] Syed A Hussaini, Kimberly A Kempadoo, Sébastien J Thuault, Steven A Siegelbaum, and Eric R Kandel. « Increased size and stability of CA1 and CA3 place fields in HCN1 knockout mice. » In: *Neuron* 72.4 (2011), pp. 643–653.

[117] Ernst Ising. « Beitrag zur theorie des ferromagnetismus. » In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258.

[118] Alan Julian Izenman. « Introduction to manifold learning. » In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.5 (2012), pp. 439–446.

[119] Pierre-Yves Jacob, Giulio Casali, Laure Spieser, Hector Page, Dorothy Overington, and Kate Jeffery. « An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. » In: *Nature neuroscience* 20.2 (2017), pp. 173–175.

[120] Andrzej Jarosz and Maciej A Nowak. « Random Hermitian versus random non-Hermitian operators—unexpected links. » In: *Journal of Physics A: Mathematical and General* 39.32 (2006), p. 10107.

[121] Kathryn J Jeffery and Michael I Anderson. « Dissociation of the geometric and contextual influences on place cells. » In: *Hippocampus* (2003).

[122] Karel Jezek, Espen J Henriksen, Alessandro Treves, Edvard I Moser, and May-Britt Moser. « Theta-paced flickering between place-cell maps in the hippocampus. » In: *Nature* 478.7368 (2011), pp. 246–249.

[123] Mehran Kardar. « Statistical mechanics of fields. » In: *lecture notes) ocw. mit. edu* (2007).

[124] Eduard Kelemen and André A Fenton. « Dynamic grouping of hippocampal neural activity during cognitive control of two spatial frames. » In: *PLoS biology* 8.6 (2010).

[125] Thomas B Kepler and Laurence F Abbott. « Domains of attraction in neural networks. » In: *Journal de Physique* 49.10 (1988), pp. 1657–1662.

[126] Sung Soo Kim, Ann M Hermundstad, Sandro Romani, LF Abbott, and Vivek Jayaraman. « Generation of stable heading representations in diverse visual scenes. » In: *Nature* 576.7785 (2019), pp. 126–131.

[127] Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. « Ring attractor dynamics in the Drosophila central brain. » In: *Science* 356.6340 (2017), pp. 849–853.

[128] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. « Optimization by simulated annealing. » In: *science* 220.4598 (1983), pp. 671–680.

[129] Charles Kittel, Paul McEuen, and Paul McEuen. *Introduction to solid state physics*. Vol. 8. Wiley New York, 1996.

[130] James J Knierim, Hemant S Kudrimoti, and Bruce L McNaughton. « Place cells, head direction cells, and the learning of landmark stability. » In: *Journal of Neuroscience* 15.3 (1995), pp. 1648–1659.

[131] James J Knierim and Kechen Zhang. « Attractor dynamics of spatially correlated neural activity in the limbic system. » In: *Annual review of neuroscience* 35 (2012), pp. 267–285.

[132] Werner Krauth and Marc Mézard. « Learning algorithms with optimal stability in neural networks. » In: *Journal of Physics A: Mathematical and General* 20.11 (1987), p. L745.

[133] Werner Krauth, Marc Mézard, and Jean-Pierre Nadal. « Basins of attraction in a perceptron-like neural network. » In: *Complex Systems* 2.4 (1988), pp. 387–408.

[134] Werner Krauth, J-P Nadal, and Marc Mezard. « The roles of stability and symmetry in the dynamics of neural networks. » In: *Journal of Physics A: Mathematical and General* 21.13 (1988), p. 2995.

[135] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. « Imagenet classification with deep convolutional neural networks. » In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[136] Emilio Kropff, James E Carmichael, May-Britt Moser, and Edvard I Moser. « Speed cells in the medial entorhinal cortex. » In: *Nature* 523.7561 (2015), pp. 419–424.

[137] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. « Phrase-based & neural unsupervised machine translation. » In: *arXiv preprint arXiv:1804.07755* (2018).

[138] Lev Davidovich Landau, Evgenij Mihajlovič Lifšic, Evgenii Mikhailovich Lifshitz, and LP Pitaevskii. *Statistical physics: theory of the condensed state*. Vol. 9. Butterworth-Heinemann, 1980.

[139] Yann LeCun. « Learning Hierarchies Of Invariant Features. » In: *Slideshare [Online]. Available:* https://www.slideshare.net/yandex/yann-le-cun (2013).

[140] Yann LeCun, Corinna Cortes, and CJ Burges. « MNIST handwritten digit database. » In: *ATT Labs [Online].*
Available: http://yann.lecun.com/exdb/mnist 2 (2010).

[141] Joel L Lebowitz and Oliver Penrose. « Rigorous treatment of the Van Der Waals-Maxwell theory of the liquid-vapor transition. » In: *Journal of Mathematical Physics* 7.1 (1966), pp. 98–113.

[142] Jae Sung Lee, John Briguglio, Sandro Romani, and Albert K Lee. « The statistical structure of the hippocampal code for space as a function of time, context, and value. » In: *bioRxiv* (2019), p. 615203.

[143] Jill K Leutgeb, Stefan Leutgeb, Alessandro Treves, Retsina Meyer, Carol A Barnes, Bruce L McNaughton, May-Britt Moser, and Edvard I Moser. « Progressive transformation of hippocampal neuronal representations in "morphed" environments. » In: *Neuron* 48.2 (2005), pp. 345–358.

[144] Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. *Introduction to random matrices: theory and practice*. Vol. 26. Springer, 2018.

[145] Christopher W Lynn and Danielle S Bassett. « The physics of brain network structure, function and control. » In: *Nature Reviews Physics* 1.5 (2019), p. 318.

[146] Yunqian Ma and Yun Fu. *Manifold learning theory and applications*. CRC press, 2011.

[147] Omar Mamad, Lars Stumpp, Harold M McNamara, Charu Ramakrishnan, Karl Deisseroth, Richard B Reilly, and Marian Tsanov. « Place field assembly distribution encodes preferred locations. » In: *PLoS biology* 15.9 (2017), e2002365.

[148] Rosario N Mantegna and H Eugene Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

[149] Shirley Mark, Sandro Romani, Karel Jezek, and Misha Tsodyks. « Theta-paced flickering between place-cell maps in the hippocampus: A model based on short-term synaptic plasticity. » In: *Hippocampus* 27.9 (2017), pp. 959–970.

[150] Elizabeth Marozzi and Kathryn J Jeffery. « Place, space and memory cells. » In: *Current Biology* 22.22 (2012), R939–R942.

[151] Stephen J Martin, Paul D Grimwood, and Richard GM Morris. « Synaptic plasticity and memory: an evaluation of the hypothesis. » In: *Annual review of neuroscience* 23.1 (2000), pp. 649–711.

[152] Guy Mayraz and Geoffrey E Hinton. « Recognizing hand-written digits using hierarchical products of experts. » In: *Advances in neural information processing systems*. 2001, pp. 953–959.

[153] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. « Path integration and the neural basis of the'cognitive map'. » In: *Nature Reviews Neuroscience* 7.8 (2006), pp. 663–678.

[154] M Mézard, JP Nadal, and G Toulouse. « Solvable models of working memories. » In: *Journal de physique* 47.9 (1986), pp. 1457–1462.

[155] Marc Mézard. « Mean-field message-passing equations in the Hopfield model and its generalizations. » In: *Physical Review E* 95.2 (2017), p. 022117.

[156] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[157] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Company, 1987.

[158] Marc Mézard, Giorgio Parisi, and Anthony Zee. « Spectra of Euclidean random matrices. » In: *Nuclear Physics B* 559.3 (1999), pp. 689–701.

[159] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. « Distributed representations of words and phrases and their compositionality. » In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[160] Aaron D Milstein, Yiding Li, Katie C Bittner, Christine Grienberger, Ivan Soltesz, Jeffrey C Magee, and Sandro Romani. « Bidirectional synaptic plasticity rapidly modifies hippocampal representations independent of correlated activity. » In: *BioRxiv* (2020).

[161] James A Mingo and Roland Speicher. *Free probability and random matrices*. Vol. 35. Springer, 2017.

[162] SJ Mizumori and John D Williams. « Directionally selective mnemonic properties of neurons in the lateral dorsal nucleus of the thalamus of rats. » In: *Journal of Neuroscience* 13.9 (1993), pp. 4015–4028.

[163] Kenji Mizuseki, Sebastien Royer, Kamran Diba, and György Buzsáki. « Activity dynamics and behavioral correlates of CA3 and CA1 hippocampal pyramidal neurons. » In: *Hippocampus* 22.8 (2012), pp. 1659–1680.

[164] Rémi Monasson. « Properties of neural networks storing spatially correlated patterns. » In: *Journal of Physics A: Mathematical and General* 25.13 (1992), p. 3701.

[165] Rémi Monasson. « Storage of spatially correlated patterns in autoassociative memories. » In: *Journal de Physique I* 3.5 (1993), pp. 1141–1152.

[166] Rémi Monasson and Sophie Rosay. « Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram. » In: *Physical review E* 87.6 (2013), p. 062813.

[167] Rémi Monasson and Sophie Rosay. « Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity. » In: *Physical review E* 89.3 (2014), p. 032803.

[168] Rémi Monasson and Sophie Rosay. « Transitions between spatial attractors in place-cell models. » In: *Physical review letters* 115.9 (2015), p. 098101.

[169] Gianluigi Mongillo, Emanuele Curti, Sandro Romani, and Daniel J Amit. « Learning in realistic networks of spiking neurons and spike-driven plastic synapses. » In: *European Journal of Neuroscience* 21.11 (2005), pp. 3143–3160.

[170] RG Morris. « Synaptic plasticity and learning: selective impairment of learning rats and blockade of long-term potentiation in vivo by the N-methyl-D-aspartate receptor antagonist AP5. » In: *Journal of Neuroscience* 9.9 (1989), pp. 3040–3057.

[171] Richard GM Morris. « Spatial localization does not require the presence of local cues. » In: *Learning and motivation* 12.2 (1981), pp. 239–260.

[172] Richard GM Morris, Paul Garrud, JNP al Rawlins, and John O'Keefe. « Place navigation impaired in rats with hippocampal lesions. » In: *Nature* 297.5868 (1982), pp. 681–683.

[173] Edvard I Moser, Emilio Kropff, and May-Britt Moser. « Place cells, grid cells, and the brain's spatial representation system. » In: *Annu. Rev. Neurosci.* 31 (2008), pp. 69–89.

[174] Dylan R Muir and Thomas Mrsic-Flogel. « Eigenspectrum bounds for semirandom matrices with modular and spatial structure for neural networks. » In: *Physical Review E* 91.4 (2015), p. 042808.

[175] Robert U Muller and John L Kubie. « The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. » In: *Journal of Neuroscience* 7.7 (1987), pp. 1951–1968.

[176] Sadegh Nabavi, Rocky Fox, Christophe D Proulx, John Y Lin, Roger Y Tsien, and Roberto Malinow. « Engineering a memory with LTD and LTP. » In: *Nature* 511.7509 (2014), pp. 348–352.

[177] Jean-Pierre Nadal. « On the storage capacity with sign-constrained synaptic couplings. » In: *Network: Computation in Neural Systems* 1.4 (1990), pp. 463–466.

[178] Kazu Nakazawa, Thomas J McHugh, Matthew A Wilson, and Susumu Tonegawa. « NMDA receptors, place cells and hippocampal spatial memory. » In: *Nature Reviews Neuroscience* 5.5 (2004), pp. 361–372.

[179] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[180] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*. Vol. 13. Cambridge University Press, 2006.

[181] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. 111. Clarendon Press, 2001.

[182] John O'Keefe. « Place units in the hippocampus of the freely moving rat. » In: *Experimental neurology* 51.1 (1976), pp. 78–109.

[183] John O'Keefe and Neil Burgess. « Geometric determinants of the place fields of hippocampal neurons. » In: *Nature* 381.6581 (1996), pp. 425–428.

[184] John O'Keefe and Neil Burgess. « Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. » In: *Hippocampus* 15.7 (2005), pp. 853–866.

[185] John O'Keefe and DH Conway. « Hippocampal place units in the freely moving rat: why they fire where they fire. » In: *Experimental brain research* 31.4 (1978), pp. 573–590.

[186] John O'Keefe and Jonathan Dostrovsky. « The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. » In: *Brain research* (1971).

[187] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.

[188] Kostas Pagiamtzis and Ali Sheikholeslami. « Content-addressable memory (CAM) circuits and architectures: A tutorial and survey. » In: *IEEE journal of solid-state circuits* 41.3 (2006), pp. 712–727.

[189] Giorgio Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[190] Giorgio Parisi. « Euclidean random matrices, the glass transition and the boson peak. » In: *The European Physical Journal E* 9.3 (2002), pp. 213–218.

[191] Giorgio Parisi. « EUCLIDEAN RANDOMMATRICES: SOLVEDAND OPEN PROBLEMS. » In: *Applications of Random Matrices in Physics*. Springer, 2006, pp. 219–260.

[192] Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. *Theory of simple glasses: exact solutions in infinite dimensions*. Cambridge University Press, 2020.

[193] EunHye Park, Dino Dvorak, and André A Fenton. « Ensemble place codes in hippocampus: CA1, CA3, and dentate gyrus place cells have multiple place fields in large environments. » In: *PloS one* 6.7 (2011), e22349.

[194] F. Pedregosa et al. « Scikit-learn: Machine Learning in Python. » In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[195] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. « Scikit-learn: Machine learning in Python. » In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[196] Lorenzo Posani, Simona Cocco, and Rémi Monasson. « Integration and multiplexing of positional and contextual information by the hippocampal network. » In: *PLoS computational biology* 14.8 (2018), e1006320.

[197] Marc Potters and Jean-Philippe Bouchaud. « A First Course in Random Matrix Theory. » In: (2019).

[198] Gregory J Quirk, Robert U Muller, and John L Kubie. « The firing of hippocampal place cells in the dark depends on the rat's recent experience. » In: *Journal of Neuroscience* 10.6 (1990), pp. 2008–2017.

[199] Alec Radford, Luke Metz, and Soumith Chintala. « Unsupervised representation learning with deep convolutional generative adversarial networks. » In: *arXiv preprint arXiv:1511.06434* (2015).

[200] JB 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.

[201] Linda E Reichl. *A modern course in statistical physics*. 1999.

[202] Frederick Reif. *Fundamentals of statistical and thermal physics*. Waveland Press, 2009.

[203] P Dylan Rich, Hua-Peng Liaw, and Albert K Lee. « Large environments reveal the statistical structure governing hippocampal representations. » In: *Science* 345.6198 (2014), pp. 814–817.

[204] Sandro Romani and Misha Tsodyks. « Continuous attractors with morphed/correlated maps. » In: *PLoS Comput Biol* 6.8 (2010), e1000869.

[205] Sandro Romani and Misha Tsodyks. « Short-term plasticity based network model of place cells dynamics. » In: *Hippocampus* 25.1 (2015), pp. 94–105.

[206] Frank Rosenblatt. « The perceptron: a probabilistic model for information storage and organization in the brain. » In: *Psychological review* 65.6 (1958), p. 386.

[207] Frank Rosenblatt. « Principles of neurodynamics: Perceptions and the theory of brain mechanisms. » In: (1962).

[208] Alexei Samsonovich and Bruce L McNaughton. « Path integration and cognitive mapping in a continuous attractor neural network model. » In: *Journal of Neuroscience* 17.15 (1997), pp. 5900–5920.

[209] Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L McNaughton, Menno P Witter, May-Britt Moser, and Edvard I Moser. « Conjunctive representation of position, direction, and velocity in entorhinal cortex. » In: *Science* 312.5774 (2006), pp. 758–762.

[210] Craig Saunders, Mark O Stitson, Jason Weston, Leon Bottou, A Smola, et al. « Support vector machine-reference manual. » In: (1998).

[211] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[212] Francesca Schönsberg, Yasser Roudi, and Alessandro Treves. « Efficiency of local learning rules in threshold-linear associative networks. » In: *arXiv:2007.12584* (2020).

[213] William Beecher Scoville and Brenda Milner. « Loss of recent memory after bilateral hippocampal lesions. » In: *Journal of neurology, neurosurgery, and psychiatry* 20.1 (1957), p. 11.

[214] Johannes D Seelig and Vivek Jayaraman. « Neural dynamics for landmark orientation and angular path integration. » In: *Nature* 521.7551 (2015), pp. 186–191.

[215] James Sethna et al. *Statistical mechanics: entropy, order parameters, and complexity*. Vol. 14. Oxford University Press, 2006.

[216] Patricia E Sharp and Kate Koester. « Lesions of the mammillary body region severely disrupt the cortical head direction, but not place cell signal. » In: *Hippocampus* 18.8 (2008), pp. 766–784.

[217] Bailu Si, Sandro Romani, and Misha Tsodyks. « Continuous attractor network model for conjunctive position-by-velocity tuning of grid cells. » In: *PLoS Comput Biol* 10.4 (2014), e1003558.

[218] Karen Simonyan and Andrew Zisserman. « Very deep convolutional networks for large-scale image recognition. » In: *arXiv preprint arXiv:1409.1556* (2014).

[219] SE Skipetrov and A Goetschy. « Eigenvalue distributions of large Euclidean random matrices for waves in random media. » In: *arXiv preprint arXiv:1007.1379* (2010).

[220] David M Smith and Sheri JY Mizumori. « Hippocampal place cells, context, and episodic memory. » In: *Hippocampus* 16.9 (2006), pp. 716–729.

[221] Katie Sokolowski and Joshua G Corbin. « Wired for behaviors: from development to function of innate limbic system circuitry. » In: *Frontiers in molecular neuroscience* 5 (2012), p. 55.

[222] Haim Sompolinsky. « Computational neuroscience: beyond the local circuit. » In: *Current opinion in neurobiology* 25 (2014), pp. xiii–xviii.

[223] Davide Spalla, Alexis Dubreuil, Sophie Rosay, Remi Monasson, and Alessandro Treves. « Can grid cell ensembles represent multiple spaces? » In: *Neural Computation* 31.12 (2019), pp. 2324–2347.

[224] Vipin Srivastava, Suchitra Sampath, and David J Parker. « Overcoming catastrophic interference in connectionist networks using Gram-Schmidt orthogonalization. » In: *PloS one* 9.9 (2014), e105619.

[225] Federico Stella, Peter Baracskay, Joseph O'Neill, and Jozsef Csicsvari. « Hippocampal reactivation of random trajectories resembling Brownian diffusion. » In: *Neuron* 102.2 (2019), pp. 450–461.

[226] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. « Sequence to sequence learning with neural networks. » In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[227] Jeffrey S Taube. « The head direction signal: origins and sensory-motor integration. » In: *Annu. Rev. Neurosci.* 30 (2007), pp. 181–207.

[228] Jeffrey S Taube, Robert U Muller, and James B Ranck. « Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. » In: *Journal of Neuroscience* 10.2 (1990), pp. 420–435.

[229] Jeffrey S Taube, Robert U Muller, and James B Ranck. « Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. » In: *Journal of Neuroscience* 10.2 (1990), pp. 436–447.

[230] Andy Thomas. « Memristor-based neural networks. » In: *Journal of Physics D: Applied Physics* 46.9 (2013), p. 093001.

[231] LT Thompson and PJ Best. « Place cells and silent cells in the hippocampus of freely-behaving rats. » In: *Journal of Neuroscience* 9.7 (1989), pp. 2382–2390.

[232] Edward C Tolman. « Cognitive maps in rats and men. » In: *Psychological review* 55.4 (1948), p. 189.

[233] Edward C Tolman, Benbow F Ritchie, and Donald Kalish. « Studies in spatial learning. I. Orientation and the short-cut. » In: *Journal of experimental psychology* 36.1 (1946), p. 13.

[234] Alessandro Treves. « Threshold-linear formal neurons in auto-associative nets. » In: *Journal of Physics A: Mathematical and General* 23.12 (1990), p. 2631.

[235] Alessandro Treves and Edmund T Rolls. « Computational analysis of the role of the hippocampus in memory. » In: *Hippocampus* 4.3 (1994), pp. 374–391.

[236] Misha Tsodyks. « Attractor neural network models of spatial maps in hippocampus. » In: *Hippocampus* 9.4 (1999), pp. 481–489.

[237] Misha Tsodyks and Terrence Sejnowski. « Associative memory and hippocampal place cells. » In: *International journal of neural systems* 6 (1995), pp. 81–86.

[238] Jérôme Tubiana and Rémi Monasson. « Emergence of compositional representations in restricted Boltzmann machines. » In: *Physical review letters* 118.13 (2017), p. 138301.

[239] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[240] Antonia M Tulino, Sergio Verdú, and Sergio Verdu. *Random matrix theory and wireless communications*. Now Publishers Inc, 2004.

[241] Daniel Turner-Evans, Stephanie Wegener, Herve Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. « Angular velocity integration in a fly heading circuit. » In: *Elife* 6 (2017), e23496.

[242] Pierfrancesco Urbani. *Statistical Physics of glassy systems: tools and applications*. 2018.

[243] NM Van Strien, NLM Cappaert, and MP Witter. « The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. » In: *Nature reviews neuroscience* 10.4 (2009), pp. 272–282.

[244] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[245] R Raju Viswanathan. « Sign-constrained synapses and biased patterns in neural networks. » In: *Journal of Physics A: Mathematical and General* 26.22 (1993), p. 6195.

[246] Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*. 1. American Mathematical Soc., 1992.

[247] Dan Voiculescu. « Asymptotically commuting finite rank unitary operators without commuting approximants. » In: *Acta Sci. Math.(Szeged)* 45.1-4 (1983), pp. 429–431.

[248] Angelo Vulpiani. *Chaos: from simple models to complex systems*. Vol. 17. World Scientific, 2010.

[249] He Wang, Kin Lam, CC Alan Fung, KY Michael Wong, and Si Wu. « How Short-Term Synaptic Depression Reshapes Dynamics of Continuous Attractor Neural Networks. » In: *Proceedings: 2015 International Symposium on Nonlinear Theory and its Applications*. 2015, p. 333.

[250] Xiao-Jing Wang. « Attractor network models. » In: *Encyclopedia of neuroscience*. Elsevier Ltd, 2010, pp. 667–679.

[251] Yingxue Wang, Sandro Romani, Brian Lustig, Anthony Leonardo, and Eva Pastalkova. « Theta sequences are essential for internally generated hippocampal firing fields. » In: *Nature neuroscience* 18.2 (2015), pp. 282–288.

[252] Sidney I Wiener. « Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. » In: *Journal of Neuroscience* 13.9 (1993), pp. 3802–3817.

[253] Wikipedia contributors. *Action Potential — Wikipedia, The Free Encyclopedia*. 2020. URL: https://en.wikipedia.org/wiki/Action_potential.

[254] Wikipedia contributors. *Hopfield network — Wikipedia, The Free Encyclopedia*. 2020. URL: https://en.wikipedia.org/wiki/Hopfield_network.

[255] Matthew A Wilson and Bruce L McNaughton. « Dynamics of the hippocampal ensemble code for space. » In: *Science* 261.5124 (1993), pp. 1055–1058.

[256] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. « Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. » In: *Nature neuroscience* 17.3 (2014), pp. 431–439.

[257] Svante Wold, Kim Esbensen, and Paul Geladi. « Principal component analysis. » In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[258] KY Michael Wong and David Sherrington. « Neural networks optimally trained with noisy data. » In: *Physical Review E* 47.6 (1993), p. 4465.

[259] KYM Wong and C Campbell. « Competitive attraction in neural networks with sign-constrained weights. » In: *Journal of Physics A: mathematical and General* 25.8 (1992), p. 2227.

[260] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. « Continuous attractor neural networks: candidate of a canonical model for neural information representation. » In: *F1000Research* 5 (2016).

[261] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. « Performance-optimized hierarchical models predict neural responses in higher visual cortex. » In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8619–8624.

[262] Michael M Yartsev and Nachum Ulanovsky. « Representation of three-dimensional space in the hippocampus of flying bats. » In: *Science* 340.6130 (2013), pp. 367–372.

[263] Chi Ho Yeung, David Saad, and KY Michael Wong. « From the physics of interacting polymers to optimizing routes on the London Underground. » In: *Proceedings of the National Academy of Sciences* 110.34 (2013), pp. 13717–13722.

[264] KiJung Yoon, Michael A Buice, Caswell Barry, Robin Hayman, Neil Burgess, and Ila R Fiete. « Specific evidence of low-dimensional continuous attractor dynamics in grid cells. » In: *Nature neuroscience* 16.8 (2013), p. 1077.

[265] Weishun Zhong, Zhiyue Lu, David J Schwab, and Arvind Murugan. « Nonequilibrium Statistical Mechanics of Continuous Attractors. » In: *Neural Computation* 32.6 (2020), pp. 1033–1068.

[266] Stuart Zola-Morgan, Larry R Squire, and David G Amaral. « Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. » In: *Journal of Neuroscience* 6.10 (1986), pp. 2950–2967.

[267] Xiaolong Zou, Zilong Ji, Xiao Liu, Yuanyuan Mi, KY Michael Wong, and Si Wu. « Learning a continuous attractor neural network from real images. » In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 622–631.

RÉSUMÉ

La manière dont l'information sensorielle est codée et traitée par les circuits neuronaux est une question centrale en neurosciences computationnelles. Dans de nombreuses régions du cerveau, on constate que l'activité des neurones dépend fortement de certains corrélats sensoriels continus. Au cours des dernières décennies, les réseaux neuronaux à attracteur continu ont été introduits comme un modèle abstrait pour la représentation de quelques variables continues dans une grande population de neurones bruités. Alors que le modèle original était basé sur la construction d'une variété continue unique dans un espace à haute dimension, on s'est vite rendu compte que le même réseau neuronal pouvait coder pour de nombreux attracteurs distincts. Une solution approximative à ce problème plus difficile a été proposée il y a vingt ans, et reposait sur une prescription ad hoc pour les interactions par paires entre les neurones. Cette solution souffre cependant de deux problèmes majeurs: l'interférence entre les cartes limitent fortement la capacité de stockage, et la résolution spatiale au sein d'une carte n'est pas contrôlée. Dans le présent manuscrit, nous abordons ces deux questions en utilisant une combinaison de techniques issues de la physique statistique des systèmes désordonnés et de la théorie des matrices aléatoires. Nous montrons comment parvenir à un stockage optimal des attracteurs continus et étudions le compromis optimal entre capacité et résolution spatiale.

MOTS CLÉS

Réseau Neuronal à Attracteur Continu, Physique Statistique, Théorie des Matrices Aléatoires, Neurosciences Computationnelles.

ABSTRACT

How sensory information is encoded and processed by neuronal circuits is a central question in computational neuroscience. In many brain areas, the activity of neurons is found to depend strongly on some continuous sensory correlate. Over the past decades, continuous attractor neural networks were introduced as an abstract model for the representation of a few continuous variables in a large population of noisy neurons. While the original model was based on how to build a single continuous manifold in an high-dimensional space, it was soon realized that the same neural network should code for many distinct attractors. An approximate solution to this harder problem was proposed twenty years ago, and relied on an ad-hoc prescription for the pairwise interactions between neurons. This solution, however, suffers from two major issues: the interference between maps strongly limit the storage capacity, and the spatial resolution within a map is not controlled. In the present manuscript, we address these two issues using a combination of techniques from statistical physics of disordered systems and random matrix theory. We show how to achieve optimal storage of continuous attractors and study the optimal trade-off between capacity and spatial resolution.

KEYWORDS

Continuous Attractor Neural Network, Statistical Physics, Random Matrix Theory, Computational Neuroscience.