UNIVERSITÀ DEL SALENTO

DIPARTIMENTO DI MATEMATICA E FISICA
'ENNIO DE GIORGI'

PhD Thesis in Nanotechnology

Neural Networks and Learning Machines: from mathematical foundation to applications in biological complexity.

Francesco Alemanno

Advisors:
Dr. Loretta del Mercato
Prof. Adriano Barra

Referees:
Prof. Remi Monasson
Prof. Ido Kanter

XXXIII CICLO

List of publications

Publications produced during the PhD

1. E. Agliari, **F. Alemanno**, M. Aquaro, A. Barra, F. Durante, ***Recurrent neural networks that generalize from examples and optimize by dreaming***, submitted to Neural Networks (2021)
[Theoretical Paper] [Status: Submitted]

2. A. Fachechi, E. Agliari, **F. Alemanno**, A. Barra, ***Dreaming Boltzmann machines outperform standard ones***, submitted to IEEE Tr. NN & LS (2021).
[Theoretical Paper] [Status: Submitted]

3. A Chandra, S Prasad, **F. Alemanno**, A Barra, C. Bucci, G. Gigli, L. Del Mercato, ***A fully automated computational approach for precisely measuring organelle acidification with optical pH sensors***, submitted to Nano (2021).
[Experimental Paper] [Status: Submitted]

4. **F. Alemanno**, M. Cavo, D. Delle Cave, E. D'Amore, A. Fachechi, G. Gigli, E. Lonardo, A. Barra, L. del Mercato, ***Quantifying heterogeneity to drug response in cancer-stroma kinetics***, submitted to PNAS (2021).
[Experimental Paper] [Status: Submitted]

5. E Agliari, **F. Alemanno**, A Barra, G De Marzo, ***The emergence of a concept in shallow neural networks***, arXiv preprint arXiv: 2109 00454, (2021).
[Theoretical Paper] [Status: In Press in Neural Networks]

6. E Agliari, L Albanese, **F. Alemanno**, A Fachechi, ***Pattern recognition in Deep Boltzmann machines***, arXiv preprint arXiv: 2106 08978, (2021).
[Theoretical Paper] [Status: In Press in Journal of Physics A]

7. E Agliari, **F. Alemanno**, A Barra, A Fachechi, L Moretti, ***Analysis of temporal correlation in heart rate variability through maximum entropy principle in a minimal pairwise glassy model***, Nature Sci. Rep. **10** (1), 1-14, (2020).
[Experimental Paper] [Status: Published]

8. A Chandra, S Prasad, **F. Alemanno**, A Barra, E Lonardo, E Parasido, L. Del Mercato, ***Microgel-based in vitro tumoroid platform for real time assessment of drug sensitivity and resistance***, Cancer Res. **80** (16 Supplement), 2967-2967 (2020).
[Experimental Paper] [Status: Published]

9. E Agliari, **F. Alemanno**, A Barra, A Fachechi, ***Generalized Guerra's interpolation schemes for dense associative neural networks***, Neural Networks **128**, 254-267, (2020).
[Theoretical Paper] [Status: Published]

10. MM Cavo, **F. Alemanno** D Delle Cave, E D'Amone, A Barra, E Lonardo, L. Del Mercato, ***Quantifying stroma-tumor cell interactions in three-dimensional cell culture systems***, Cancer Res. **80** (11), 53-54, (2020).
[Experimental Paper] [Status: Published]

11. **F. Alemanno**, M Centonze, A Fachechi, ***Interpolating between Boolean and extremely high noisy patterns through minimal dense associative memories***, Journal of Physics A: Mathematical and Theoretical **53** (7), 074001, (2020).
[Theoretical Paper] [Status: Published]

12. E Agliari, **F. Alemanno**, A Barra, M Centonze, A Fachechi, ***Neural networks with a redundant representation: detecting the undetectable***, Physical Review Letters **124** (2), 028301, (2020).
[Theoretical Paper] [Status: Published]

13. E Agliari, **F. Alemanno**, A Barra, A Fachechi, ***Dreaming neural networks: rigorous results***, Journal of Statistical Mechanics: Theory and Experiment **8**, 083503, (2019).
[Theoretical Paper] [Status: Published]

14. E Agliari, **F. Alemanno**, A Barra, A Fachechi, ***On the Marchenko Pastur law in analog bipartite spin-glasses***, Journal of Physics A: Mathematical and Theoretical **52** (25), 254002, (2019).
[Theoretical Paper] [Status: Published]

Contents

Introduction

This PhD thesis summarizes three years of work in understanding artificial information processing in order to develop new neural networks whose capabilities outperform the state of the art in the fields of *pattern recognition* and *signal detection* in Artificial Intelligence. The ultimate aim is to apply these techniques to high dimensional inferential problems as those typically emerging in any modern design of a biotechnological experiment.
Indeed purposes of this work are twofold. From one side we aim at developing a theoretical framework for Artificial Intelligence, where a comphrension of its processing mechanisms and spontaneously emerging computational skills may find a natural place: this field of Science has been historically grounded on the *statistical mechanics of complex systems*, i.e. the maximum entropy variational extremization a' la Jaynes and Parisi theory of spin glasses, that will thus be the two leitmotif of the whole thesis. From the other side, we want to convert the above theoretical comprehension of information processing by neural network and learning machines in concrete practical applications in the laboratory, at work with problems in Biological Complexity.
As a result the thesis is naturally split into three main Chapters.
The first chapter, the *Theoretical Backbone*, is a conditio sine qua non in order to be sure to share the basic mathematical instruments we need during the thesis. Once quickly streamlined statistical inference and statistical mechanics in the beginning of the chapter, than we move to deepen the two archetypal models we need to rely on along the whole manuscript, namely the Curie-Weiss model (as the harmonic oscillator for simple systems) and the Sherrington-Kirkpatrick model (as the hegemon example of a complex paradigm). Once simple and complex references are provided we can build up the simplest architectures of neural networks and learning machines, namely the Hopfield neural network and its dual representation, the Restricted Boltzmann machine: while historically heuristic approaches were available to describe these systems already in the past century (e.g. the so-called *Replica Trick* in the '80s & '90s), since the groundbraking interpolation techniques introduced by Francesco Guerra in the early 2000, it is finally possible to obtain a rigorous mathematical control of these networks and in this thesis the formalization of the new neural networks I studied is achieved by these novel mathematical tools [1].
Once such a minimal summary is over with the first chapter, the second chapter reports new research in the field of Theoretical Artificial Intelligence. At first in order to make the Hopfield network a precious intrument at work in the Labs, the pivotal generalization that we have to face is to shift its *storing skills* toward *learning skills*: namely, within the standard statistical mechanical formulation of such a network the most important question addressed in the past has been *given N binary neurons to give rise to a fully connected Hebbian network, how many -already defined- patterns P the network is able to cope with*

[1] Indeed we paid attention to a systematic usage of the same techniques along the whole manuscript: all the models (the three ones of the first Chapter and the two generalizations provided in Chapter Two) have all been addressed with the Guerra's interpolation technique and -where required- by the same signal-to-noise analysis with the hope of making the manuscript easier to be understood.

at its best? - this is a question on the *maximal storage capacity* α of the network, whose answer is $\alpha_{\max} = P_{max}/N \sim 0.14$ for the Hopfield neural network. We should generalize this question toward *given the same network, if it is not supplied with already prescribed patterns (that we can call archetypes for obvious reasons) rather it is fed by examples, which are the critical sizes of the datasets containing the examples that we must ensure to the Hopfield network such that it can reconstruct the archetype so to be able to learn and generalize it?* This will be the first research point addressed in this thesis.
Once established these thresholds for learning, still in the second chapter, I will focus on two major generalizations (always biologically inspired) of the Hopfield reference, that is neural networks equipped with *ultra-memory* and with *ultra-detection* skills, as I briefly summarize hereafter:

- The first generalization -*ultra-memory*- raises to enlarge the maximal storage capacity, i.e. $\alpha_c \sim 0.14$ as, from general information theoretic argument we know that the upper bound for such a capacity should be $\alpha_{top} = 1$[1] hence the actual bound $\alpha_{max} \sim 0.14$ seems rather unsatisfactory. The idea that we implemented is to allow the network *to sleep*: namely, by suitably stylizing in mathematical equations the two key mechanisms of dreaming in mammals (Random Eye Movement and Slow Wave Sleep), I prove that the network -if allowed to take some rest- can actually saturate the critical capacity for symmetric networks as prescribed by Information Theory argument reaching $\alpha_c \sim 1.00$. As we will see, beyond constituting a remarkable step forward in Theoretical Artificial Intelligence per se and a new bridge between artificial information processing and biological information processing, this enlarged capacity results in remarkable computational implications, at first the possibility of avoiding (or better minimizing) overfitting while learning from datasets.

- The second generalization -*ultra-detection*- constists in equipping the network with higher order interactions w.r.t. the pairwise reference (the so-called *dense network* limit): in particular, inspired by the redundant representation in humans (roughly speaking the presence of two sources providing similar information as e.g. the two eyes we use to see), I studied a generalized Boltzman machine equipped with two identical input layers and prove the latter to be the dual of a dense Hopfield network whose Hebbian kernel contains redundant information: remarkably this redundancy allows the network to tune its signal-to-noise threshold for signal detection. Indeed I proved the (quite counterintuitive at a first glance) result that -while the standard Hopfield network can detect a signal whose intensity is $O(1)$ if it lies in a sea of noise at worst of the same magnitude of the signal -hence $O(1)$- the present dense generalization can detect a signal $O(1)$ even if lost in a sea of noise $O(\sqrt{N})$. Clearly there is a price to pay for this skill: when at work, the network sacrifices extensive memory storage to free neurons and synapses in order to play with the redundancy coming from the two input layers for finding out the information hidden in the noise. This closes my theoretical work in Artificial Intelligence discussed in this thesis.

The last part of the thesis -Chapter Three- reports two applications of high-dimension statistical inference (that the maximum entropy criterion allows to account for) on two biological problems: detecting interactions between (pancreatic) cancerous cells and those cells that surround -or infiltrate- the tumoral mass -namely the stroma- and inferring the

[1] Actually it is possible to reach even $\alpha_{top} = 2$ by removing the symmetry of the synaptic matrix, but this requires working if off-equilibrium settings not yet crystal clear for disordered systems and will not be deepened here.

timescales involved in heart rate variability in healthy and pathogenic patients as I briefly summarize again hereafter.

- Detecting cancer-stroma interactions: once suitably marked for fluorescence (in order to be able to recognize the two cellular lineages) we mix, in vitro, cancerous cells and stroma cells and via time-lapse confocal microscopy we record the dynamics of these cells twice: the first time cells are left alone to dialogue, the second time a chemoterapeutic drug (i.e. gemcitabine) is added in the colture medium. From a physical perspective, the dataset for network's training thus constists in the entire phase space of these cells (i.e., positions and velocities) and the pourpose of the inference is to detect kinetic cellular interactions in order to see how these are affected (or not) by the presence of the chemoterapy. I studied two difference lineages of pancreatic cancer and I obtained different results: on a given tumoral line (the more aggressive), stroma and cancer dialogues were absent with and without the presence of gemcitabine, questioning of the efficacy of this drug for this particular type of cancer (indeed the progression of the clonal expansion of the cancer raised almost unperturbately) while for the other tumoral line, the effect of the gemcitabine is to highly increase interaction between cancer and stroma (consistently with the death count of roughly $O(50\%)$ of cancerous cells) thus highglighting a key mechanism of action of the drug.

- Inferring timescales in heart rate variability: while naively one may point out that the cardiac frequency always stays confined between 40 and 300 beats per minute (hence there are not at all several timescales to cross), the attention in this research is not on the beats per minute, rather on their variation per minute: the latter indeed spans over several scales and and it is known to be power-law distributed (in particular the characteristic $1/f$ noise typical of heart-rate variability stems from the interplay between the parasympathetic and orthosympathetic systems). I studied historical series recorded from standard Holter for healthy patients and those suffering from cardiac decompensation or suffering from atrial fibrillation and -via a suitable adaptation of the maximum entropy inferential criterion- I revealed the existence of huge differences in their related statistics both in the time and frequency domains highlighting novel aspects of non-invasive early diagnosis which could be tomorrow integrated in a Personalized Medicine, where these approaches are becoming pervasive already nowadays.

Finally I'd like to remark that, if we see the skills of the Hopfield networks and Boltzmann machines emerging as a consequence of the minimization of their respective cost functions under the prescription of maximization of their relative entropies, both the research chapters -the former dealing with biologically-inspired neural networks, the latter dealing with high-dimensional biological inference- are tributes to maximum entropy variational scheme that, as stated, is indeed the first conceptual pillar of the whole thesis. Further, if we note that dialogues among cancerous and stroma cells in the first experiment are of complex nature (i.e. both promoting and inhibiting dynamics) and we appreciate the frustration resulting from the interplay between the parasympathetic and orthosympathetic systems in heart rate variability, also the Parisi representation of a complex systems plays naturally as the second pillar over which the whole thesis raises.

Chapter 1

Part 1: Theoretical Backbone

1.1 The mathematical pillars

Statistical mechanics aroused in the last decades of the XIX century thanks to its founding fathers L. Boltzmann, J.C. Maxwell and J.W. Gibbs. Its scope (at that time) was to provide a consistent theoretical background formalizing the already existing empirical thermodynamics, in order to reconcile its noisy and irreversible behaviour with a deterministic and time reversal microscopic dynamics. While trying to get rid of statistical mechanics in just a few words is almost meaningless, its modus operandi may be summarized via toy-examples. Let us start with a very simple system, e.g. a perfect gas, in which molecules obey a Newton-like microscopic dynamics (without friction - as we are at the molecular level - thus time-reversal). Rather than focusing on each particular particle trajectory to characterize the state of the system (that would be computationally prohibitive because we shall integrate numerically an Avogadro number of coupled ODEs), we define *order parameters* (variables describing the system's behaviour from a macroscopic perspective, e.g. the density) in terms of microscopic variables (the particles belonging to the gas). By averaging their evolution over suitable probability measures and simultaneously imposing minimum energy and maximum entropy principles, via this route it is possible to infer the macroscopic behaviour of the system in agreement with thermodynamics, hence linking the microscopic deterministic and time reversal mechanics with the macroscopic strong dictates stemmed by the second principle (i.e. the arrow of time coded by the entropy growth). Despite famous attacks to Boltzmann theorem (e.g. by Zermelo or Poincaré), statistical mechanics was immediately recognized as a deep and powerful bridge between microscopic dynamics of system's constituents and (emergent) macroscopic properties shown by the system itself, as exemplified by the equation of state for perfect gases obtained by considering the Hamiltonian for a single particle accounting for the kinetic contribution only [1, 2].

One step beyond the perfect gas scenario (where no interaction takes places among atoms), J.D. Van der Waals and J.C. Maxwell in their pioneering works focused on real gases, in which particle interactions were finally considered by introducing a non-zero potential in the microscopic Hamiltonian describing the system. This extension required fifty-years of deep changes in the theoretical physics perspective in order to be able to face new classes of questions. The remarkable reward lies in a theory of phase transitions where the focus is no longer on details regarding the system constituents, but rather on the characteristics of their interactions. Indeed, phase transitions, namely abrupt changes in the macroscopic state of the whole system, are not due to the particular system considered, but are primarily due to the ability of its constituents to perceive interactions over the

thermal noise. For instance, when considering a system made of a large number of water molecules, whatever the level of resolution to describe the single molecule (ranging from classical to quantum), by properly varying the external tunable parameters (e.g. the temperature), the system eventually changes its state with a phase transition from liquid to vapor (or solid, depending on parameter values): of course, the same applies generally to liquids (not just to water).

The fact that the macroscopic behaviour of a system may spontaneously show *cooperative, emergent* properties (actually hidden in its microscopic description and not directly deducible when looking at its single components) was definitely appealing in neuroscience. In fact, in the 70s, neuronal dynamics along axons, from dendrites to synapses, was already rather clear (see e.g. the celebrated book by Tuckwell [3]) and not much more intricate than circuits that may arise from basic human creativity. In this context, the aptness of a *thermodynamic formulation* of neural interactions - *revealing* possible emergent capabilities - was immediately pointed out, despite the route was not clear yet. Indeed, we will try to show in this thesis that one of the main rewards in using statistical mechanics to inspect the spontaneous information processing skills neural networks show is the concept of *phase diagram*: we will be able to identify, in the space of the tunable parameters of the network (e.g. the level of noise the network is embedded in or the information load of the network, etc.), regions where some emerging skills are available, regions where other behaviours appear and regions where the network no longer works as an information processing system. This is exactly the opposite perspective w.r.t. the extensive empirical trials that constitute nowadays the main route to Machine Learning, as seen from an engineering-prone perspective.

Along the same lines, while we will largely rely upon statistical mechanics to paint these phase diagrams, we can also adopt a pure statistical inference perspective - in order to match our results with those existing in the Engineering Literature where much of the results have been framed in statistical terms: the bridge will be the Maximum Entropy Principle acting as the Roman *Giano Bifronte* as it can be used to literally ground both statistical mechanics as well as statistical inference, as we will quickly revise in this introductory section.[1]

1.1.1 Statistical mechanics in a teaspoon

This framework requires a probability measure on a given space, that is invariant with respect to the Hamiltonian flow. For a system of N particles this measure can be easily deduced, and it is related to the Hamiltonian function, that we choose to be

$$H_N(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2m} \sum_{i=1}^{N} p_i^2 + \sum_{i \neq j} V(q_i - q_j),$$

where in this generic construction $\boldsymbol{p} = (p_1, \dots, p_N)$ and $\boldsymbol{q} = (q_1, \dots, q_N)$ are the Lagrangian coordinates in the phase space of the system, with p_i and q_i respectively being the momentum and the position of particle i, and V is a potential. Setting these quantities to be in the three-dimensional euclidean space, the state space is $\Omega = \mathbb{R}^{6N}$. When working on spin or neural networks, the state variable are idealized with Boolean vectors $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)$, where each $\sigma_i \in \{-1, +1\}$ represents the spins orientation (up or down) or the neuron's

[1] Of course here we are tacitely assuming the reader to be familiar with these fields of Science as, obviously, nor there is hope to be exhaustive on such broad themes in just a few pages, neither this is the scope of the present manuscript.

activity (spiking or not spiking). Here the state space is $\Omega = \{-1, +1\}^N$.
From now on we will only consider systems with a noisy microscopic behaviour. In this stochastic context, we define the *entropy* functional for the system as the following:

$$S[\mathcal{P}] = -\int_\Omega dx\ \mathcal{P}(x) \ln \mathcal{P}(x),$$

with $x = (\boldsymbol{p}, \boldsymbol{q})$, $\mathcal{P}$ being the probability distribution over the state space Ω. Entropy is by definition the measure of the system disorder. In fact, the smaller is the subset of Ω on which the density $\mathcal{P}$ is concentrated and the smaller is the measured entropy. Indeed, if the system is described by a probability distribution that is highly concentrated in a small area of the state space it means that the system is actually not that random but is rather ordered. For example, if we consider the discrete case with N possible states, the entropy function is described by

$$S_N[\mathcal{P}] = -\sum_{i=1}^{N} \mathcal{P}_i \ln \mathcal{P}_i, \tag{1.1}$$

with the closure condition

$$\sum_{i=1}^{N} \mathcal{P}_i = 1.$$

Let's consider the simple case of a uniform distribution

$$\mathcal{P}_i = \begin{cases} \frac{1}{N} & i \leq N, \\ 0 & i > N, \end{cases} \tag{1.2}$$

where $\mathcal{P}_i$ is the probability that state i is occupied. Then, $S = \ln N$, meaning that the number of configurations in which the system can be found with a considerable probability is e^S and thus confirming the meaning of S as a measure of the system disorder.

We now illustrate how expression (1.1) can also be interpreted as the number of system configurations. Let us consider a set of systems (*ensemble*) made of N identical systems and suppose that each one of them can take on K different possible states. A configuration of this system is given by the numbers $N_1, \ldots, N_K$, where N_i is the number of the systems in the ensemble occupying the i-th state. The number of possibilities that satisfy this configuration is given by the multinomial coefficient

$$\frac{N!}{\prod_{i=1}^{N} N_i!} = \mathcal{N},$$

with the condition that $\sum_i N_i = N$. Applying Stirling's formula, the entropy S_N is $S_N = 1/N \ln \mathcal{N}$ following this computation:

$$\frac{1}{N} \ln \mathcal{N} = \frac{1}{N}\Big(N \ln N - \sum_{i=1}^{K} N_i \ln N_i\Big) = \sum_{i=1}^{K} \frac{N_i}{N}\Big(\ln N - \ln N_i\Big) =$$

$$= -\sum_{i=1}^{K} \frac{N_i}{N} \ln \frac{N_i}{N} = -\sum_{i=1}^{K} \mathcal{P}_i \ln \mathcal{P}_i = S_N[\mathcal{P}],$$

in which the probability $\mathcal{P}_i$ has been identified with the frequency N_i/N thanks to the law of large numbers (tacitely highlighting that we will be interested in evaluating quantities in the asymptotic limit $N \to \infty$, called *thermodynamic limit* in Statistical Mechanics. Therefore we obtained an interpretation of an ensemble entropy, and it is the one that we will use throughout this thesis: S is proportional to the logarithm of the number of ways that a given configuration can appear.

Remark 1.1. Thanks to the previous definitions and examples, we can conclude that for a smaller entropy we have a system that is concentrated on a small number of states and thus we have more information about it.

Now, we show how Gibbs measure has the ability of maximizing the entropy functional. To do this, we consider the evolution of a set of N (a large and fixed number) interacting Hamiltonian systems in thermal equilibrium, meaning that the energy of a generic subsystem j presents small fluctuations on the average value fixed at E_j. We can say that the ensemble is in thermal equilibrium if every subsystem gives out and receives an equal quantity of energy from the other subsystems. Assuming that we know E_N, the ensemble's average value of the total energy is given by

$$E_N = \sum_{i=1}^{N} \mathcal{P}_i E_i,$$

where the sum is carried on all the possible values that can be observed in the ensemble. For simplicity, we shall consider a discrete case in which E_j stands in a discrete set $\mathcal{E}_N$ and every subsystem of the ensemble takes average energy levels in $\mathcal{E}_N$. From the second principle of thermodynamics, we know that the entropy of an isolated system grows as the information decreases while system evolves. Hence, it comes naturally to look for the probability distribution $\mathcal{P}_j$ of all the available energy levels that maximize the entropy $S_N[\mathcal{P}]$. This distribution exists and it is called the *Gibbs measure*. The problem can be translated in a mathematical form as

$$\begin{cases} \max_{\mathcal{P}_j} S_N[\mathcal{P}], \\ \sum_{i=1}^{N} \mathcal{P}_i E_i = E_N, \\ \sum_{i=1}^{N} \mathcal{P}_i = 1. \end{cases} \tag{1.3}$$

This is a constrained maximization problem, whose solution is obtained by means of Lagrange multipliers, i.e. finding the maximum of the following function

$$S_{N,\beta,\gamma}[\mathcal{P}] = -\sum_i \mathcal{P}_i \ln \mathcal{P}_i + \beta\Big(\sum_i \mathcal{P}_i E_i - E_N\Big) + \gamma\Big(\sum_i \mathcal{P}_i - 1\Big).$$

The solution is quite nice and simple, and reads

$$\mathcal{P}_i = \frac{e^{-\beta E_i}}{Z_N},$$

where $Z_N = e^{1-\gamma} = \sum_i e^{-\beta E_i}$ is known as the *partition function*. The computed values of $\mathcal{P}_i$ are in fact maximum points for the entropy. The parameter β can be calculated with the following condition:

$$\frac{1}{Z_N} \sum_{i=1}^{N} E_i e^{-\beta E_i} = E_N,$$

from which we can also show why β can be interpreted as the inverse of the temperature. To clarify this point, we introduce the function

$$F_N(\beta, \mathcal{E}_N) = \ln Z_N,$$

whose associated differential is

$$dF_N = \frac{\partial F_N}{\partial \beta} d\beta + \sum_{i=1}^{N} \frac{\partial F_N}{\partial E_i} dE_i = -E_N d\beta - \beta \sum_{i=1}^{N} \frac{N_i}{N} dE_i, \tag{1.4}$$

where we have replaced $\mathcal{P}_i$ with the frequency N_i/N of the event of having the energy level E_i in the ensemble. We can rewrite equation (1.4) as

$$d\big(F_N + E_N\beta\big) = \beta\Big(dE_N - \sum_{i=1}^{N} \frac{N_i}{N} dE_i\Big), \tag{1.5}$$

and give a nice physical interpretation. In fact, if we suppose to work on different ensemble subsystems (e.g. varying their dimension, parameters, etc.), the quantity $\sum_i N_i/N dE_i$ represents the work on the ensemble needed to change the energy levels of the systems and dE_N its internal energy variation. Thus, for the first principle of thermodynamics, $dE_N - \sum_i N_i/N \ dE_i$ is nothing but the amount of exchanged heat dQ_N between the ensemble and the external environment. Hence, the identification of $\beta = 1/T$, where T it the ensemble temperature, is straightforward since it is the only way to make βdQ_N exact.

From the second principle of thermodynamics, we know that $d(F_N + E_N/T)$ must be the system entropy differential, being $dQ_N/T = dS_N$. Hence, taking $\beta = 1/T$, we have the (extensive) free energy of the system:

$$F_N(\beta) \equiv -\frac{1}{\beta} \ln Z_N = E_N - TS_N. \tag{1.6}$$

The free energy $F_N(\beta)$ (or -alternatively- the statistical pressure $alpha_N(\beta) = -\beta F(\beta)$, that conveys the same information, vide infra) is a state function that can be expressed through the system order parameters and control parameters, such that, by extremizing this observable w.r.t. the order parameters -to impose thermodynamic principia- we obtain a set of coupled equations for their evolution in the space of the tunable parameters, whose inspection gives rise to the phase diagrams, that are the highest reward of this approach.

Remark 1.2. The order parameter values that minimize F_N describe the equilibrium states of the system. In fact, minimizing the free energy, they also maximize the system entropy and minimize the system energy and are fulfilled by the most number of (allowed) microscopic states. Thus, they are the most probable values.

An equivalent way to find the values of $\mathcal{P}_i$ that maximize the entropy is based on the search of the free energy F_N minima satisfying the conditions in (1.3). Plugging the definitions of entropy (1.1) and average energy into (1.6) and imposing the minimum conditions, we have

$$\begin{aligned} &F_{N,\mu}(\beta) = \sum_i \mathcal{P}_i E_i + T\sum_i \mathcal{P}_i \ln \mathcal{P}_i + \mu\Big(\sum_i \mathcal{P}_i - 1\Big) = 0, \\ &\frac{\partial F_N}{\partial \mathcal{P}_i}(\beta) = E_i + T\ln \mathcal{P}_i + T + \mu = 0 \quad \Rightarrow \quad \mathcal{P}_i = e^{-E_i/T} \cdot e^{-\mu/T-1}. \end{aligned}$$

Forcing the normalization on $\mathcal{P}_i$, we get μ such that $e^{-\mu/T-1} = 1/\{\sum_i e^{-E_i/T}\} \equiv 1/Z_N$, so that $\mathcal{P}_i = e^{-E_i/T}/Z_N$.

From the definitions given above, we can learn the following relations:

$$\begin{aligned} F_N &= E_N - TS_N = \\ &= \sum_i \mathcal{P}_i E_i + T\sum_i \mathcal{P}_i \ln \mathcal{P}_i|_{\mathcal{P}_i = {Z_N}^{-1} e^{-E_i/T}} = \\ &= \frac{1}{Z_N}\sum_i E_i e^{-\beta E_i} + \frac{T}{Z_N}\sum_i e^{-\beta E_i} \ln\Big(\frac{1}{Z_N} e^{-\beta E_i}\Big) = -T\ln Z_N, \\ S_N &= \beta^2 \frac{\partial F_N}{\partial \beta}, \quad E_N = F_N + \beta \frac{\partial F_N}{\partial \beta}. \end{aligned}$$

Ultimately, we will be interested in the thermodynamic limit for the intensive (i.e. normalized to the system size) free energy, referred to as $f(\beta)$ (i.e. we drop the index N) and to find its minima. Thus

$$f(\beta) \doteq \lim_{N\to\infty} \frac{1}{N} F_N(\beta) = \lim_{N\to\infty} -\frac{1}{\beta N} \ln Z_N. \tag{1.7}$$

Equivalently, we can study the statistical pressure, referred to as $\alpha_N(\beta)$ when dealing with a finite system of size N, and as $\alpha(\beta)$ when dealing with the thermodynamical limit, that is

$$\alpha(\beta) = \lim_{N\to\infty} \alpha_N(\beta) = -\beta \lim_{N\to\infty} \frac{1}{N} F_N(\beta) = \lim_{N\to\infty} \frac{1}{N} \ln Z_N(\beta). \tag{1.8}$$

The reason behind this equivalent choice between the free energy and the statistical pressure is mainly historical: indeed, in the standard approaches to field theory and statistical mechanics by theoretical physicists, they used to deal with the free energy, yet the mathematical physicists (in particular in the celebrated trilogy of papers describing Quantum Field Theory as a Statistical Mechanical Theory by F. Guerra, L. Rosen and Barry Simon) formulated the whole theory by using $\alpha(\beta)$ rather than $f(\beta)$): as we will tackle the whole thesis with rigorous Guerra's rigorous techniques, often we will naturally prefer to keep speaking about the statistical pressure rather than the free energy.

Once we are able to write explicitly the free energy or the statistical pressure in terms of the system order parameters, we proceed with the calculation of the state equations for these quantities. This procedure consists in deriving the statistical pressure with respect to each order parameter in order to find its critical points where the statistical pressure has a maximum (or a minimum in the case we are dealing with the free energy).

A licit question could be the following: why are we considering the thermodynamic limit when neural networks cannot physically contain infinitely many neurons? Other than obtaining the associative memory characteristic (technically speaking, solely in the thermodynamic limit, neural networks are a form of non-ergodic systems [4], as, along the same reasoning, just in that limit phase transitions do exist [5]), we can also give a merely practical justification: in this limit, most of the probability distributions describing crucial observables (e.g. those pertaining to thermodynamic functions as free energy, energy and entropy) become delta-peaked, thus ultimately allowing a simpler description of the system under study (w.r.t. finite volume expressions).

1.1.2 Statistical inference in a nutshell

Following the same attitude of the previous Section, where we forced a deep and complicated discipline in just a few pages, here we address Statistical Inference. In fact, this is another giant field, but we will cover solely one of its many ramifications. This Section deals with a particular application of information theoretic concepts to problems of statistical inference (typically addressed in Machine Learning), that is density estimation for a random variable X (with values $x \in \Omega$) which is not completely specified, in the sense that the full set of probabilities $\{\mathcal{P}_i, i = 1, \dots, N\}$ or, in the case of a continuous random variable, the probability density function $\mathcal{P}(x)$ are unknown. We assume that information about probabilities is available in terms of averages $\langle f_\alpha(x) \rangle$ for a family $\{f_\alpha\}$ of functions of X (e.g. the moments $\mu_n = \langle x^n \rangle$ of X). The task is once more to estimate $\mathcal{P}$ solely on the basis of available information. Remarkably, the method of choice here is again the Maximum Entropy Principle, for density estimation this time, as we briefly revise.

The solution to the problem formulated above, as proposed by Jaynes [6] in the 1950s, is based on the observation that the (Shannon [7]) entropy associated to a random variable

X, that is

$$S(X) = -k \sum_{x \in A} \mathcal{P}(x) \ln \mathcal{P}(x), \tag{1.9}$$

describes the average uncertainty about actual outcomes of observations of X (with some normalizing factor k whose knowledge is now inessential), therefore measuring our ignorance about X (see also [8, 9, 10]). According to Jaynes, a consequence of that observation is that the best estimate of a set of probabilities $\{\mathcal{P}(x), x \in A\}$, compatible with the available information, is given by an assignment of probabilities maximizing the entropy - that is, our ignorance about X - subject only to constraints coming from experiemental findings. Note that this can be seen as a mathematical formalization of the Occam razor and it is not too far from the old *Principle of Sufficient Reason* by Liebniz.

One thereby expects to prevent inappropriate implicit assumptions about X, involving properties that we have in fact no knowledge of, from sneaking into the probability assignment that is being made. Jaynes prescription thus provides a systematic method of being maximally unbiased in a probability estimate and only using known averages. In order to formulate the solution in detail, we return to the previous convention of making explicit the dependence of the entropy on the distribution $\mathcal{P}$ by using the notation $S[\mathcal{P}]$.

The problem to be solved can now formally be stated as follows. Let X be a random variable, with the set A of possible realizations given. It is assumed that the only information available about the probabilities $\{\mathcal{P}(x), x \in A\}$ is given in terms of a set of averages

$$\langle f_\alpha(x) \rangle = \sum_{x \in A} \mathcal{P}(x) f_\alpha(x) = \bar{f}_\alpha, \qquad f_\alpha \in \mathcal{M},$$

with $\mathcal{M} = \{f_\alpha(x)\}$ denoting a given family of functions. We stress that this family must *always* contain the function $f_0(x) \equiv 1$, whose trivial average

$$\langle f_0(x) \rangle = \sum_{x \in A} \mathcal{P}(x) = 1,$$

ensures that $\mathcal{P}(x)$ is a probability and thus $S[\mathcal{P}]$ a real entropy. Denoting $\mathcal{P}^*$ as the best estimate of the probability distribution compatible with the above constraints, then it is found according to the following prescription

$$S[\mathcal{P}^*] = \max_{\mathcal{P}} \{S[\mathcal{P}]\} \quad \text{such that} \quad \langle f_\alpha(x) \rangle = \bar{f}_\alpha. \tag{1.10}$$

We will now briefly discuss some prototypical examples to get acquainted with entropy maximization by an inferential perspective.

- Worst Example: **Uniform Distribution**

  Let us suppose we know nothing about the system under consideration. Then, the only constraint is that $\mathcal{P}^*$ is a probability distribution, so that Jaynes criterion turns into the maximization of the functional

  $$S_0[\mathcal{P}] = S[\mathcal{P}] + k\alpha_0 \Big(\sum_{x \in A} \mathcal{P}(x) - 1 \Big).$$

  Then, $\mathcal{P}^*$ is obtained with the conditions

  $$\frac{\partial S_0[\mathcal{P}]}{\partial \mathcal{P}(x)} = -k \ln \mathcal{P}(x) - k + k\alpha_0 = 0, \tag{1.11}$$

  $$\frac{\partial S_0[\mathcal{P}]}{\partial \alpha_0} = k \Big(\sum_{x \in A} \mathcal{P}(x) - 1 \Big) = 0, \tag{1.12}$$

and it is trivial to check that the solution is the uniform distribution (as expected since we have no a priori information on the system, see eq. 1.2).

- Crucial Example: **Gaussian Distribution**

  Let us suppose now that - as in the standard experimental settings - we can measure the first empirical momenta regarding the system under study, i.e. the mean and the variance. Again, the functional to maximize can immediately be written in Lagrangian form as

$$\begin{aligned} S_2[\mathcal{P}] = S[\mathcal{P}] + k\alpha_0\Big(\sum_{x\in A}\mathcal{P}(x) - 1\Big) + k\alpha_1\Big(\sum_{x\in A} x\mathcal{P}(x) - \mu_1\Big) \\ + k\alpha_2\Big(\sum_{x\in A}\mathcal{P}(x)(x-\mu_1)^2 - \mu_2\Big). \end{aligned} \tag{1.13}$$

  In a similar fashion as before, $\mathcal{P}^*$ is found by solving

$$\frac{\partial S_0[\mathcal{P}]}{\partial \mathcal{P}(x)} = 0, \tag{1.14}$$

$$\frac{\partial S_0[\mathcal{P}]}{\partial \alpha_s} = 0, \tag{1.15}$$

  for $s = 0, 1, 2$. It is again trivial - but also crucial - to check that the solution is the Gaussian distribution (as expected since we have information on the mean and the variance of the system under consideration), namely

$$\mathcal{P}^*(x) = \frac{1}{Z}e^{\alpha_1 x + \alpha_2 x^2} = \frac{1}{Z}e^{-(x-\hat{\alpha}_1)^2/2\hat{\alpha}_2^2},$$

  with $\hat{\alpha}_1 = \mu_1$ and $\hat{\alpha}_2 \equiv \sigma^2 = \mu_2 - \mu_1^2$.

Thus, the Gaussian probability density - apart from its key role in the Central Limit Theorem - enjoys a privileged role also as a maximally unbiased estimator of a probability density function with the only constraints of given first and second moments (or equivalently of given mean and variance).

A final note stressing the overall *harmony* among the two approaches hereafter summarized, is a tribute to *reductionism* (leaving criticism to the Conclusions): in Physics, as long as forces are linear,[1] the Hamiltonian (or energies) are quadratic forms in the microscopic variable (for instance, for a spring whose law is $F = -kx$, as $F = -\partial_x E(x)$ the associated energy is $E(x) = kx^2/2$) and, as a sharp consequence of this, the Boltzmann-Gibbs distribution $\propto \exp(-\beta E)$ is a Gaussian (in the microscopic variables x).

Note that the maximum entropy principle has tacitely been the unique guide in this streamlined and biased summary of the two disciplines we rely along this thesis.
In the next sections we address a celebrated example of a simple system (the Curie-Weiss model) and a celebrated example of a complex system (the Sherrington-Kirkpatrick model): these are the two fundamental ingredients whose generalities we need to know in order to

[1] The assumption of *linearity in the forces* is a natural definition of a "reductionistic description" as, thanks to linearity, a sum of two forces translates in the linear sum of the consequences they imply: it is trivial to visualize this by taking for example a vertical spring in a gravitational field and adding to its lower extremum one or two masses and than checking the relative equilibrium elongation of the spring itself, in formulae: $F_1 = -kx_1,\ F_2 = -kx_2,\ \rightarrow F_{tot} = F_1 + F_2 = -kx_1 - kx_2 = -k(x_1 + x_2) = -kx_{tot}$.

construct a statistical mechanical theory of neural networks and learning machines: in the present thesis, will focus mainly on the Hopfield neural network (as the archetype of an associative memory able to perform pattern recognition) and its dual representation, the (restricted) Boltzmann machine, harmonic oscillator of a learning architecture: note that these are both *pairwise Hamiltonian*, hence we will work at the complex boundaries of the statistical reductionism (that we control mathematically and conceptually in a satisfactory way even from Hard Science perspective), leaving deep and dense generalization to future investigations.

1.2 Simple Systems: The Curie-Weiss paradigm

The Curie-Weiss (CW) model is often introduced during the study of standard statistical mechanics, in particular in relation with the Ising model (1920), originally developed to investigate magnetic properties of matter [1, 2]. Briefly, in the one-dimensional Ising model, each of the N nuclei (labelled with i) is schematically represented by a spin σ_i assuming only two values ($\sigma_i = -1$, spin down and $\sigma_i = +1$, spin up). Only nearest neighbour spins interact reciprocally with positive (i.e. ferromagnetic) interactions $J_{i,i+1} > 0$, hence the Hamiltonian of this system can be written as $H_N(\sigma) \propto -\sum_i^N J_{i,i+1}\sigma_i\sigma_{i+1} - h\sum_i^N \sigma_i$, where h tunes the external magnetic field and the minus signs ensure that spins try to align with the external field and to get parallel each other in order to fulfil the minimum energy principle. Clearly, this model can trivially be extended to higher dimensions. However, due to prohibitive difficulties in facing the metric (rather than topological) constraints of considering nearest neighbour interactions only, soon shortcuts were properly implemented to turn around this path. A (actually crucial for Artificial Intelligence) effective simplification in the treatment of the Ising model is the so called "mean field approximation", whose simplified model is termed the *Curie-Weiss* (CW) model.

The CW model occupies an important place in statistical mechanics literature and its application to information theory. Indeed, it is a paradigm for *simple systems*, whose definition (one out of many) is the requirement that their related amount of free energy minima does not scale with the system size N: in particular, the CW free energy presents only two minima, whatever volume of spins N (even if $N \to \infty$).[1]

1.2.1 The mean field ferromagnetic model

Let us start the analysis of the CW model: in this mean field approximation, where each spin interacts with all the other spins in the network (regardless any definition of distance), the finite volume case is defined on a fully connected graph whose nodes host N Ising spins $\sigma_i \in \{-1, 1\}$ $\forall i = 1, \dots, N$. The interactions are specified with a coupling matrix $\{J_{ij}\}$ (i.e. the weighted adjacency matrix in a graph theoretical jargon) such that $J_{ij} = J > 0$ $\forall i, j = 1, \dots, N$ and $i \neq j$, while the diagonal terms are null. Without loss of generality, we shall assume $J = 1$. For simplicity, we will also require that there is no external field acting on the system (as one body terms $\propto \sum_i h_i\sigma_i$ are always mathematically trivial to handle with since their joint probability distribution factorizes over the sites [8]). Therefore, we can give the following

[1] Moreover, the model can also be interpreted as a neural network in which now neurons replace what were originally called spins, and the values that they acquire are now indicating whether the cell is spiking (+1) or quiescent (−1) [11].

Definition 1.1. The Hamiltonian function $H_N(\boldsymbol{\sigma})$ of the mean field ferromagnetic model (CW) is:

$$H_N(\boldsymbol{\sigma}) = -\frac{1}{N}\sum_{1\le i<j\le N}\sigma_i\sigma_j = -\frac{1}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + \frac{1}{2}. \tag{1.16}$$

Remark 1.3. In the last definition, the last term $1/2$ can be ignored since it is irrelevant in the thermodynamic limit.

Remark 1.4. From now on, through the whole thesis, we write $\sum_{\boldsymbol{\sigma}}$ intending that the sum is carried over all the possible values that $\boldsymbol{\sigma}$ can take in the configuration space $\Omega = \{-1,+1\}^N$.

Definition 1.2. The order parameter for the CW model is the (global) magnetization m defined as

$$m(\boldsymbol{\sigma}) := m = \frac{1}{N}\sum_{i=1}^{N}\sigma_i \in [-1,1]. \tag{1.17}$$

Using this definition, we can also rewrite the Hamiltonian (1.16) as

$$H_N(m) = -\frac{N}{2}m^2,$$

that is clearly minimized for $m^2 = 1$, or equally for $m = \pm 1$. Note further that, as it should, the intensive energy H_N/N does not scale with N, since $H_N(m) \propto N \cdot \mathrm{const}(N)$, where $\mathrm{const}(N)$ means that the quantity is bounded in N.

Definition 1.3. For a given inverse temperature $\beta = 1/T$, the partition function $Z_N(\beta)$ is defined as

$$Z_N(\beta) := \sum_{\boldsymbol{\sigma}} B_N(\beta) = \sum_{\boldsymbol{\sigma}} e^{-\beta H_N(\boldsymbol{\sigma})} = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N}\sum_{ij}\sigma_i\sigma_j}, \tag{1.18}$$

where $B_N := e^{-\beta H_N}$ is the Boltzmann factor.

Definition 1.4. The Gibbs measure $\omega_N(\cdot)$ for a generic function F depending on $\boldsymbol{\sigma}$ is

$$\omega_N(F) := \frac{\sum_{\boldsymbol{\sigma}} F(\boldsymbol{\sigma})B_N(\beta)}{\sum_{\boldsymbol{\sigma}} B_N(\beta)}. \tag{1.19}$$

Definition 1.5. The statistical pressure $\alpha(\beta) = -\beta f(\beta)$ is defined as

$$\alpha(\beta) := \lim_{N\to\infty}\alpha_N(\beta) = \lim_{N\to\infty}\frac{1}{N}\ln Z_N(\beta),$$

where, as standard, $f(\beta) = N^{-1}(E - TS)$ is the (intensive) free energy, namely the difference - at given noise level T - between the energy and the entropy related to the system (normalized to the volume).

Following the statistical mechanics approach, we are interested in obtaining an explicit expression for the thermodynamic limit of the (intensive) free energy (or, equivalently, of the pressure function) in terms of the order parameter: by extremizing such an expression w.r.t. the latter, we will access the equation of state of CW model. This equation allows to inspect phase transitions and painting a phase diagram for the model.

We will solve the problem of writing explicitly the thermodynamic pressure function in three ways: the first is the standard determination of an upper and lower bound for the

finite volume pressure; the second follows a Guerra's (one-parameter) interpolating procedure; finally, the third method that is achieved through another Guerra's (two parameters) interpolating scheme, i.e. the Hamilton-Jacobi formalism. Although the latter method is way more elaborated than necessary for the CW, we present also this method of resolution as a preparatory step to its application to the mean field spin-glass and to the mean field neural network. Furthermore, the latter will act as a guide - once facing an AI rationale in the final chapters of this thesis - to suggest us how to overcome the actual state of the art in this formalization of AI.

Overall this chapter is dedicated more to the techniques (at work on the elementary CW model where every stage of calculation is trivial) than to the Physics (that is rather poor and well-known), so to get the reader acquainted with the underlying mathematical methodologies the thesis has been built on.

In general, as a first step (when possible), it is always mandatory to check the existence of the thermodynamic limit for the free energy. Although it is obvious that it would be rather embarrassing speaking about not-existing quantities, we will see that - in general - for neural networks this knowledge is not yet available: let us start addressing this calculation for the CW model.

1.2.2 The thermodynamic limit

As stated above, the first problem one should face is to prove the existence (and possibly the uniqueness) of the limit of the free energy per site when the size of the system goes to infinity. Indeed, in principle this limit could depend on the particular sequence of system sizes chosen to reach the thermodynamic limit, or, even worst, it could oscillate or simply diverge.

As it is well-known, for translational invariant systems with short range interactions the existence and uniqueness are proven by dividing the system into large subsystems: the interaction energy among them is a surface effect, negligible with respect to the bulk energy, so that the free energy per site does not change essentially when the system size is increased [5]. When the model is disordered and finite-dimensional with short range interactions, if the disorder distribution is translational invariant, this approach still works: the subsystems interact weakly, due to the short range character of the potential, and the free energy of the blocks can be approximated as independent identically distributed random variables. Then, the existence of the large N limit of the free energy per site follows from the strong law of large numbers.

When dealing with mean field models, surface terms are actually of the same order as the bulk terms, and the approach outlined above does not work. In this case, the proof of the existence of the thermodynamic limit has been provided by Guerra and Toninelli and it is based on a smooth interpolation between a large system, made of N spin sites, and two similar but independent subsystems, made of N_1 and N_2 sites respectively, with $N_1 + N_2 = N$.

We start by considering the trivial inequality

$$2mM - M^2 \leq m^2,$$

holding for any $M \in \mathbb{R}$, which shall be meant as a trial magnetization. Plugging it into the partition function (1.18), we get

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta N}{2} m^2} \geq \sum_{\boldsymbol{\sigma}} e^{\beta m M N} e^{-\frac{1}{2}\beta M^2 N}.$$

The sum is easy to compute, since the magnetization appears linearly and therefore the sum factorizes over each spin. Physically speaking, we replaced the two-body interaction, which is generally difficult to deal with, with a one-body coupling. Then, we try to compensate this replacement by modulating the field acting on each spin with the help of a trial fixed magnetization M and a correction term quadratic in the latter. The result is the following bound:

$$\begin{aligned}\frac{1}{N}\ln Z_N(\beta) \geq & \frac{1}{N}\ln\sum_{\boldsymbol{\sigma}} e^{\beta M\sum_i \sigma_i} + \frac{1}{N}\ln e^{-\frac{1}{2}\beta M^2 N} \geq \\ \geq & \frac{1}{N}\ln\Big(\prod_{i=1}^{N}\sum_{\boldsymbol{\sigma}} e^{\beta M\sigma_i}\Big) - \frac{1}{2}\beta M^2 \geq \\ \geq & \sup_{M\in[-1,1]}\Big\{\ln 2 + \ln\cosh(\beta M) - \frac{1}{2}\beta M^2\Big\},\end{aligned} \tag{1.20}$$

holding for any size of the system N.

The opposite bound needs a few more steps. Firstly, let us notice that the magnetization m can take only $N+1$ distinct values. Using the trivial identity $\sum_M \delta_{mM} = 1$, we can therefore split the partition function into sums over configurations with constant magnetization in the following way:

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}}\sum_{M}\delta_{mM}e^{\frac{1}{2}\beta N m^2}, \tag{1.21}$$

where the sum over M is performed over the values $-1, -\frac{N-1}{N}, \ldots, \frac{N-1}{N}, 1$. Now, inside the sum the relation $m = M$ holds, also implying that $m^2 = 2mM - M^2$. Plugging the last equality into equation (1.21) and using the trivial inequality $\delta_{mM} \leq 1$, we get

$$Z_N(\beta) \leq \sum_{M}\sum_{\boldsymbol{\sigma}} e^{\beta N m M} e^{-\frac{1}{2}\beta N M^2}.$$

With the same calculations performed in (1.20), we have the resulting upper bound:

$$\frac{1}{N}\ln Z_N(\beta) \leq \ln\frac{N+1}{N} + \sup_{M\in[-1,1]}\Big\{\ln 2 + \ln\cosh(\beta M) - \frac{1}{2}\beta M^2\Big\}. \tag{1.22}$$

The upper (1.22) and lower (1.20) bounds converge to the same value of the pressure per site in the thermodynamic limit.

Let us now move to illustrate the idea behind the (much more general) Guerra and Toninelli interpolative approach to prove the existence of this limit [12]. To do this, we start by dividing the N spin system into two subsystems of N_1 and N_2 spins each, with $N = N_1 + N_2$. Denoting by $m_1(\boldsymbol{\sigma})$ and $m_2(\boldsymbol{\sigma})$ the corresponding magnetizations in the two subsytems, trivially defined as

$$m_1(\boldsymbol{\sigma}) = \frac{1}{N_1}\sum_{i=1}^{N_1}\sigma_i, \tag{1.23}$$

$$m_2(\boldsymbol{\sigma}) = \frac{1}{N_2}\sum_{i=N_1+1}^{N}\sigma_i, \tag{1.24}$$

we can easily the global magnetization $m(\boldsymbol{\sigma})$ as a convex linear combination of the two:

$$m(\boldsymbol{\sigma}) = \frac{N_1}{N}m_1(\boldsymbol{\sigma}) + \frac{N_2}{N}m_2(\boldsymbol{\sigma}). \tag{1.25}$$

Since the function $x \to x^2$ is convex, we have

$$Z_N(\beta) \leq \sum_{\boldsymbol{\sigma}} \exp \beta \left(N_1 m_1^2(\boldsymbol{\sigma}) + N_2 m_2^2(\boldsymbol{\sigma}) \right) = Z_{N_1}(\beta) Z_{N_2}(\beta), \tag{1.26}$$

hence

$$N f_N(\beta) = -\frac{1}{\beta} \log Z_N(\beta) \geq N_1 f_{N_1}(\beta) + N_2 f_{N_2}(\beta). \tag{1.27}$$

This is the well known property of superadditivity of the free energy in the system size that guarantees the validity of the Fekete lemma, ensuring convergence of $f_N(\beta) \to f(\beta)$ as $N \to \infty$. The existence of the limit then follows from standard methods: the only other ingredient for the proof, in a nutshell, is that the free energy is bounded from above uniformly in N, which can be easily seen by setting $M = 0$ in Eq. (1.22), to get $f_N(\beta) \leq -\beta^{-1} \log 2$. The property of superadditivity is not only fundamental in proving that the limit exists, but it also implies that the limit equals the $\sup_N f_N(\beta)$.

Operationally, the strategy is to interpolate between the original system of N spins and the two non-interacting subsystems with respectively N_1 and N_2 units, comparing their free energies. To this task we introduce an interpolating parameter $t \in [0, 1]$ and an auxiliary partition function

$$Z_N(\beta, t) = \sum_{\boldsymbol{\sigma}} \exp \beta \left(N t J m^2(\boldsymbol{\sigma}) + N_1 (1-t) J m_1^2(\boldsymbol{\sigma}) + N_2 (1-t) J m_2^2(\boldsymbol{\sigma}) \right). \tag{1.28}$$

For the boundary values $t = 0, 1$, we have

$$-\frac{1}{N\beta} \log Z_N(1) = f_N(\beta), \tag{1.29}$$

$$-\frac{1}{N\beta} \log Z_N(0) = \frac{N_1}{N} f_{N_1}(\beta) + \frac{N_2}{N} f_{N_2}(\beta). \tag{1.30}$$

Taking the derivative respect to t, we obtain

$$-\frac{d}{dt} \frac{1}{N\beta} \log Z_N(\beta, t) = -\omega_t \Big(m^2(\boldsymbol{\sigma}) - \frac{N_1}{N} m_1^2(\boldsymbol{\sigma}) - \frac{N_2}{N} m_2^2(\boldsymbol{\sigma}) \Big) \geq 0, \tag{1.31}$$

where $\omega_t(\cdot)$ denotes the Boltzmann-Gibbs thermal average corresponding to the t-dependent partition function (1.28). Then, integrating over t between 0 and 1 and recalling the boundary conditions (1.29, 1.30), one finds again the superadditivity property (1.27).

1.2.3 Guerra's Interpolating scheme

Now that we know we are speaking about well defined quantities, in this Section we obtain the pressure density function through a celebrated Guerra's interpolation technique: this exploits the real essence of the mean-field nature of these models as we are interpolating between the original system under consideration (i.e. the CW in the present case) and a one-body model. The terms appearing in the latter will be suggested by the model itself and by the mathematical experience collected in making the calculations tractable.

Given the CW Hamiltonian (1.16) and the related partition function (1.18) we introduce the following generalized partition function

$$\begin{aligned} Z_N(\beta, t) &\doteq \sum_{\boldsymbol{\sigma}} \exp \Big\{ \frac{\beta t}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j + (1-t)\psi \sum_{i=1}^{N} \sigma_i \Big\} = \\ &= \sum_{\boldsymbol{\sigma}} \exp \Big\{ \frac{\beta t}{2} N m^2 + (1-t)\psi N m \Big\}, \end{aligned} \tag{1.32}$$

with m defined in (1.17), $t \in [0,1]$ and $\psi \in \mathbb{R}$ is a tunable parameter that we will determine later on. This new generalized partition function is an interpolation between the two-body interaction, once evaluated at $t = 1$, and the much simpler one-body problem, described by $t = 0$[1]. We can then define a generalized pressure $\alpha_N(\beta, t)$ as

$$\alpha_N(\beta, t) \doteq \frac{1}{N} \ln Z_N(\beta, t),$$

the Boltzmann factor $B_N(t)$ such that $Z_N(\beta, t) = \sum_{\boldsymbol{\sigma}} B_N(t)$, and the related generalized Gibbs measure $\omega_t(\cdot)$ following the analogous definition (1.19). The key observation is enclosed in the next

Proposition 1.1. *The statistical pressure for a finite volume N can be written in the following way thanks to the fundamental theorem of calculus:*

$$\alpha_N(\beta) \equiv \alpha_N(\beta, t = 1) = \alpha_N(\beta, t = 0) + \int_0^1 ds \, \Big[\partial_t \alpha_N(\beta, t)\Big]_{t=s}. \tag{1.33}$$

The computation of each term is quite simple. For the one-body (i.e. $t = 0$) term we have

$$\begin{aligned} \alpha_N(\beta, t = 0) =& \frac{1}{N} \ln Z_N(\beta, t = 0) = \frac{1}{N} \ln \Big(\sum_{\boldsymbol{\sigma}} e^{\psi \sum_i \sigma_i} \Big) = \\ =& \frac{1}{N} \ln \Big(\prod_{i=1}^{N} \sum_{\boldsymbol{\sigma}} e^{\psi \sigma_i} \Big) = \ln 2 + \ln \cosh(\psi), \end{aligned} \tag{1.34}$$

while the derivative in (1.33) is

$$\begin{aligned} \frac{\partial}{\partial t} \alpha_N(t) =& \frac{1}{N} \frac{\partial_t Z_N(\beta, t)}{Z_N(\beta, t)} = \frac{1}{N Z_N(\beta, t)} \Big[\sum_{\boldsymbol{\sigma}} \Big(\frac{\beta N}{2} m^2 - \psi N m \Big) B_N(t) \Big] = \\ =& \frac{\beta}{2} \omega_t(m^2) - \psi \omega_t(m). \end{aligned} \tag{1.35}$$

Now, let us go through the following considerations. We know that the average value of the magnetization exists in the thermodynamic limit -let us call this value $M \in [-1, +1]$- and, in that limit, $\mathcal{P}(m) = \delta(m - M)$. Then, trivially we have

$$\omega_t\big((m - M)^2\big) = \omega_t\big(m^2\big) + M^2 - 2M\omega_t(m). \tag{1.36}$$

Looking back at the final result of equation (1.35), we notice that we can manipulate the expression as follows:

$$\frac{\beta}{2} \omega_t\big(m^2\big) - \psi \omega_t(m) = \frac{\beta}{2} \Big(\omega_t\big(m^2\big) - \frac{2\psi}{\beta} \omega_t(m) \Big).$$

Therefore, setting $\psi = \beta M$ and using equation (1.36), we can write a convenient expression for the pressure derivative as

$$\frac{\partial}{\partial t} \alpha_N(\beta, t) = \frac{\beta}{2} \Big(\omega_t\big(m^2\big) - \frac{2\psi}{\beta} \omega_t(m) \Big) = \frac{\beta}{2} \omega_t\Big((m - M)^2 \Big) - \frac{1}{2} \beta M^2. \tag{1.37}$$

Finally, plugging the results of equations (1.34) and (1.37) into (1.33), we can state that the pressure function is defined by the next

[1] The presence of the parameter ψ -rather than ψ_i- is due to the fact that we are working in a mean field approximation, meaning that each spin is equally influenced by a uniform presence of the others.

Theorem 1.1. *The infinite volume limit of the the Curie-Weiss statistical pressure $\alpha(\beta)$ can be written in terms of the magnetization as*

$$\alpha_N(\beta) = \sup_{M\in[-1,1]} \Big\{ \ln 2 + \ln\cosh(\beta M) - \frac{1}{2}\beta M^2 + \frac{\beta}{2}\omega\Big((m-M)^2\Big)\Big\}, \tag{1.38}$$

where the last term at the r.h.s. of the above expression converges to 0 *in the thermodynamic limit (since the order parameter is a self-averaging quantity). The free energy extremization w.r.t. M ensures the requirement of maximum entropy and minimum energy principles, and returns the celebrated self-consistency relation*

$$M = \tanh(\beta M), \tag{1.39}$$

by which the phase diagram of the CW model becomes accessible.

1.3 Complex Systems: The Sherrington-Kirkpatrick paradigm

Spin glasses, besides constituting "a challenge for mathematician" [13], are among the paradigmatic models in complex systems theory, whose distinctive feature is that the number of free energy minima sensibly grows with the system size N. Their fields of applications include optimization theory, computer science, biology, economics etc. [14, 15, 16] and, last but not least, Artificial Intelligence [11, 17].
The expression *spin glass* was originally coined to designate some magnetic alloys with a very peculiar behavior, in particular characterized by lack of long-range order and very slow relaxational dynamics at low temperatures. Experimentally, in such alloys one can observe, for example, a non-periodic arrangement of magnetic moments below a critical temperature, and memory effects in susceptibility and residual magnetization. To understand some of these phenomena, Edwards and Anderson (EA) proposed in 1975 an extension of the Ising model in which the interactions between couple of spins are random variables assuming both positive and negative values. The next step was the introduction of a simpler model by Sherrington and Kirkpatrick (SK), i.e. the mean field version [18] of the EA model. Curiously, the title of the paper was "Solvable Model of a Spin-Glass" but, even if the authors - using the famous replica trick in the replica symmetric approximation - found an explicit form for the free energy, they realized that their solution was only valid above a certain temperature. The correct answer to the problem was found in the '80s with the seminal works by Parisi [19]. There, the author proposed a formula for the free energy per site in the thermodynamic limit and a description of the pure states of the system. However, a rigorous proof of the validity of Parisi formula was carried out only some years ago, and it is splitted across two works by Guerra [20, 21] and Talagrand [22, 23]. Apart from a few exceptions [4, 24, 25], most important rigorous results are quite recent. The existence of the thermodynamic limit for the free energy, for example, was proven by Guerra and Toninelli after more than 20 years, in 2002 [12]. The techniques used for these recent breakthroughs, which are mainly based on interpolation, found fruitful applications also in neighboring fields, such as for example optimization problems and diluted spin glasses, finite-range spin glasses, and neural networks [26, 27, 28, 29, 30], as we will extensively deepen in this thesis.

Spin glasses can be simply defined as magnetic systems with a non-periodic freezing of the spins at low temperatures. The first experiments which drew some attention to these characteristics were performed on dilute solutions of magnetic transition metal impurities in noble metal hosts. In these systems, the impurity moments produce a magnetic

polarization of the host metal conduction electrons, which is positive at some distances and negative at others. Beneath a characteristic temperature, a Mössbauer line-splitting in zero applied field was observed, indicating a local hyperfine field due to local freezing of the magnetic moments. Moreover, the absence of any corresponding magnetic Bragg peak in neutron diffraction demonstrated that the freezing was not periodic. Another sign of this non-ferromagnetic freezing came from earlier measurements of the susceptibility, showing a peak at a similar temperature and therefore highlighting the presence of a phase transition. Other remarkable features, such as preparation-dependence effects and a considerable slowing-down of response to external perturbations, demonstrated the presence of many metastable states in this new low-temperature phase, with significant free energy barriers separating these states. The first historical attempt to produce a theory of the described transition is due to Edwards and Anderson (see e.g. [19, 31, 32]), who proposed a Ising-like Hamiltonian, with the magnetic moments placed on the N sites of a hypercubic lattice, and keeping only a single spin component $\sigma_i = (\vec{\sigma}_i)_z = \pm 1$:

$$H_N(\boldsymbol{\sigma}|\boldsymbol{J}) = -\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j, \tag{1.40}$$

where the nearest neighbors interactions J_{ij} are random independent and identically distributed variables (Gaussians, for example), with random signs. It is then clear that a key ingredient is *disorder*: the Hamiltonian depends not only on the configuration of the system, which we denote by $\boldsymbol{\sigma}$, and possibly on the strength of the external (magnetic) applied fields, but also on some random parameters (usually, the couplings among the elementary degrees of freedom), whose probability distribution is supposed to be known. The random parameters are collectively denoted as "quenched" or "frozen" disorder. From a physical point of view, the word "frozen" means that we are dealing with a disordered system whose impurities have a dynamics which is many orders of magnitude slower than the evolution of the spin degrees of freedom. Therefore, the disorder does not reach thermal equilibrium on the time scales of the spin relaxation and can be considered as fixed (this is somewhat similar to the Born-Oppenheimer adiabatic approximation for dealing with electron and nuclei dynamics in molecular systems). This fact has deep consequences on the way we have to perform the averages over the couplings, compared to the configurations $\boldsymbol{\sigma}$. The second key ingredient, strongly related with the disordered nature of such systems, is *frustration*, i.e., competition between different terms in the Hamiltonian, so that they can not all be minimizied simultaneously. More precisely, a system is said to be frustrated if there exist a loop on which the product of the couplings is negative (see Fig. 1.1). We have seen before (see Sec. 1.2) how in the Curie-Weiss model each spin-spin interaction is minimized when the two spin are parallel, i.e., $\sigma_i\sigma_j = +1$ for all couples $\langle i, j\rangle$. In that case, there are only two such configurations, one with all the spins equal to $+1$, the other with spins -1, and they are connected by the global spin-flip symmetry $\sigma_i \to -\sigma_i \ \ \forall\, i$. If the couplings J_{ij} have random sign (and possibly modulus), the ground state has a high degeneracy and they are not connected to one another by elementary symmetry transformations.[1]

1.3.1 The mean-field spin glass model

The Edwards-Anderson (EA) model is already somewhat simplified with respect to the actual physical situation: a more realistic model could consider, for instance, interactions $\boldsymbol{J} = \{J_{ij}\}$ decaying with distance, instead of nearest-neighbors couplings, or Heisenberg

[1] Notice that frustration disappears when considering the system on graphs without loops, for example a tree.

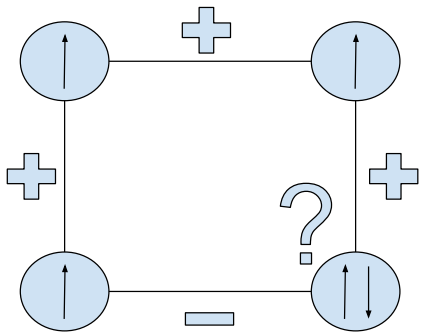

Figure 1.1: **A very simple example of a frustrated system**. The spins tend to be parallel when they interact with a positive coupling and anti-parallel when the interaction is negative. Obviously, not all the conditions can be met simultaneously, meaning that interaction is frustrated.

spins $\vec{\sigma}_i$, with more than one component attached on each site. However, despite its intrinsic limitation, it was already too difficult to be attacked analytically, and suitable approximation schemes were developed. In particular, the most important one (and also the richest in surprises) was the mean-field approximation. In this case, while maintaining the fundamental features of disorder and frustration, the geometrical structure of the lattice is disregarded (as we already discussed for Ising and Curie-Weiss models), allowing for every magnetic moment to interact with all the others, irrespective of the distance. The first model with such requirements was introduced by Sherrington and Kirkpatrick (SK) (see e.g. [19, 31]), whose Hamiltonian is given by the next

Definition 1.6. The mean field spin glass is introduced by the following Sherrington-Kirkpatrick Hamiltonian

$$H_N(\boldsymbol{\sigma}|h;\boldsymbol{J}) = -\frac{1}{\sqrt{N}} \sum_{1\le i<j\le N} J_{ij}\sigma_i\sigma_j - h \sum_{1\le i\le N} \sigma_i. \tag{1.41}$$

where the first term at the r.h.s. is a long range random two-body interaction, while the second one represents the interaction of the spins with an homogeneous magnetic field h. In the following, we will often consider the zero external field case, denoting the Hamiltonian simply with $H_N(\boldsymbol{\sigma}|\boldsymbol{J})$. The $N(N-1)/2$ couplings J_{ij} are assumed to be i.i.d. centered unit Gaussians, so that, denoting with $\mathbb{E}$ the average on disorder, we have

$$\mathbb{E}J_{ij} = 0 \quad \text{and} \quad \mathbb{E}J_{ij}^2 = 1.$$

Note that this choice of the coupling is a matter of convenience: in fact spin glasses share the *universality* property [33], that guarantees that any other symmetric probability distribution with finite moments could be chosen for J_{ij} without modifying the free energy of the system, apart from error terms vanishing in the thermodynamic limit.
The case $J_{ij} = \pm 1$ with equal probability $1/2$, for instance, is often considered in the literature. The normalization factor $1/\sqrt{N}$ guarantees that (intensive) energy, (intensive) entropy and (intensive) free energy density do not scale with N in the thermodynamic limit, as they should. One may point out that, in the Curie-Weiss model, the normalizing

factor is stronger (namely $1/N$, to be compared with $1/N^{1/2}$), but - in the SK case - the random signs of the couplings J_{ij} produce cancellations among the many terms of the Hamiltonian H_N. The correctness of this choice can be easily understood by checking the *linear* extensivity of the (extensive) expectation value for the internal energy of the model: this can be done elementary by considering a duplicated system with configurations $\boldsymbol{\sigma}^1$ and $\boldsymbol{\sigma}^2$, but with the same disorder (i.e. *identical couplings*), and computing

$$\begin{aligned}\mathbb{E}(H_N(\boldsymbol{\sigma}^{(1)}|\boldsymbol{J})H_N(\boldsymbol{\sigma}^{(2)}|\boldsymbol{J})) &= \frac{1}{N}\sum_{i<j}^{1,N}\sum_{k<l}^{1,N}\mathbb{E}(J_{ij}J_{kl})\sigma_i^{(1)}\sigma_j^{(1)}\sigma_k^{(2)}\sigma_l^{(2)} \\ &= \frac{1}{N}\sum_{1\le i<j\le N}\sigma_i^{(1)}\sigma_j^{(1)}\sigma_i^{(2)}\sigma_j^{(2)} \\ &= \frac{N}{2}\Big(\frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)}\Big)^2 - \frac{1}{2}. \end{aligned} \tag{1.42}$$

The quantity

$$q_{12} = q(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)}, \tag{1.43}$$

occurring in the previous equation is fundamental, since it is the order parameter for the model (as we will see in the following), and it is called *overlap*. It measures the resemblance between the configurations of the two copies (or *replicas*, as we will soon better specify) $\boldsymbol{\sigma}^{(1)}$ and $\boldsymbol{\sigma}^{(2)}$, ranging from -1, when each spin of a replica is opposed to the corresponding one of the other copy, to $+1$, when they are perfectly aligned. The fact that the overlap is a resemblance measure is confirmed by its relation with the Hamming distance $d(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})$, which counts the number of non-aligned spins:

$$d(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}) = \frac{1}{2}(1-q_{12}).$$

Then, taking two identical copies $\boldsymbol{\sigma}^{(1)} = \boldsymbol{\sigma}^{(2)}$, we note that

$$\mathbb{E}\left(H_N(\boldsymbol{\sigma}|\boldsymbol{J})\right)^2 = \frac{N}{2} - \frac{1}{2}, \tag{1.44}$$

showing that the normalization factor is correct.

1.3.2 Quenched and annealed free energies

We now start with formalizing the thermodynamic observables for disordered systems. First of all, for a given inverse temperature $\beta = 1/T$, we introduce the following

Definition 1.7. The disorder-dependent partition function $Z_N(\beta,h;\boldsymbol{J})$, the *quenched* average of the free energy per site $f_N(\beta,h)$, and the disorder dependent Boltzmann-Gibbs state $\omega_{\boldsymbol{J}}$ read as

$$Z_N(\beta|h;\boldsymbol{J}) = \sum_{\boldsymbol{\sigma}}\exp(-\beta H_N(\boldsymbol{\sigma}|h;\boldsymbol{J})), \tag{1.45}$$

$$f_N(\beta|h) = -\frac{1}{\beta N}\mathbb{E}\log Z_N(\beta|h;\boldsymbol{J}), \tag{1.46}$$

$$\omega_{\boldsymbol{J}}(A) = Z_N(\beta,h;\boldsymbol{J})^{-1}\sum_{\boldsymbol{\sigma}}A(\boldsymbol{\sigma})\exp(-\beta H_N(\boldsymbol{\sigma}|h;\boldsymbol{J})), \tag{1.47}$$

where $A = A(\boldsymbol{\sigma})$ is a generic observable (for example the energy H_N), depending on the spin configuration $\boldsymbol{\sigma}$.

In some cases it will be more practical to deal, rather than with $f_N(\beta|h)$, with

$$\alpha_N(\beta|h) = \frac{1}{N}\mathbb{E}\log Z_N(\beta|h;\boldsymbol{J}) = -\beta f_N(\beta|h), \tag{1.48}$$

namely the statistical pressure, as already seen for the CW model. As for the Hamiltonian, in the following we will shorten the notation in $Z_N(\beta|\boldsymbol{J})$, $f_N(\beta)$, $\alpha_N(\beta)$ etc. when considering the case of zero external field ($h = 0$). The quenched free energy is the correct average if one looks for the free energy of a system where the disorder is frozen (i.e. its dynamics is many orders of magnitude slower than the dynamics of the spin degrees of freedom), like in real spin glasses.

Remark 1.5. A remark is in order here: it is mandatory to notice that - when mimicking neural networks with statistical mechanical models - we will have to take into account that, in the analogy, while the neurons will be modeled by the spins, while couplings play the role of synapses. Since the latter can be both excitatory as well as inhibitory and they must be accounted by the couplings J_{ij} (or *synaptic matrix* in neural network jargon), it is then clear that the correct reference framework must be a spin-glass and not the simplest ferromagnet. Furthermore, the frustration that these random couplings introduce in the network is the responsible for the proliferation of the free energy minima that is, in turn, something that we will need in order to develop an extensive memory storage (we will deepen these concepts in the following Chapters).

Moreover, the free energy per spin for a given realization of disorder

$$-\frac{1}{\beta N}\log Z_N,$$

is *self-averaging* [34], meaning that its deviations from the quenched value vanish in the thermodynamic limit with probability one.

Definition 1.8. One can also consider the so-called *annealed* free energy

$$f_N^{\mathrm{A}}(\beta|h) = -\frac{1}{\beta N}\log \mathbb{E}Z_N(\beta|h;\boldsymbol{J}), \tag{1.49}$$

where the disorder averages is performed directly on the partition function.

From a physical point of view, this corresponds to the assumption that the couplings relaxation characteristic timescales are on the same level of those relative to spins thermalization (in the landscape produced by the synapses - namely by the couplings - that are effectively considered as frozen on the short timescale involved by neural dynamics), and let them participate in the thermal equilibrium. This terminology comes from metallurgy and the thermal processing of materials: a "quench" corresponds in this jargon to preparing a sample by quickly bridging it from high to low temperatures, so that atoms do not change their positions, apart from small vibrations. In the "annealing" process, on the contrary, the cooling down is slower and gradual, so that atoms can rearrange and find favorable positions.

Remark 1.6. A quite interesting analogy between spin glasses and neural networks lies in this adiabaticity of the timescales regarding spins (neurons) and links (synapses), hidden behind the concept of quenched variables: indeed, in neural networks, in order to preserve the learning ability of the net, it is pivotal that neurons and synapses evolve on very different timescales and, for the (well known) biological side, these are order 10^2 milliseconds for the neural firing rate and *from days to months* for synaptic plasticity, hence we can safetly consider quenched the synapses while interested in neural dynamics, much as in glassy physics couplings do not evolve while spins (try to) thermalize.

The computation of the annealed free energy is trivial, since the Boltzmann factor in this case can be written as the product of $N(N-1)/2$ statistically independent terms, one for each pair of sites, so that

$$Z_N(\beta|h;\boldsymbol{J}) = \sum_{\boldsymbol{\sigma}} \prod_{1\le i<j\le N} \exp\Big(\frac{\beta}{\sqrt{N}} J_{ij}\sigma_i\sigma_j\Big) \times \exp\Big(\beta h \sum_{1\le k\le N} \sigma_k\Big),$$

and the disorder average factorizes as

$$\begin{aligned}\mathbb{E}Z_N(\beta|h;\boldsymbol{J}) &= \sum_{\boldsymbol{\sigma}} \exp\left(\frac{\beta^2}{2N}\frac{N(N-1)}{2}\right) \exp\Big(\beta h \sum_{1\le k\le N} \sigma_k\Big) \\ &= 2^N \cosh^N(\beta h) \exp\Big(\frac{\beta^2}{4}(N-1)\Big).\end{aligned}$$

Finally, the annealed free energy per site is

$$f_N^{\mathrm{A}}(\beta|h) = -\frac{1}{\beta}\log 2\cosh(\beta h) - \frac{\beta}{4}\frac{N-1}{N}, \tag{1.50}$$

and in the thermodynamic limit we have the next

Proposition 1.2. *The infinite volume limit of the annealed pressure of the SK model reads as*

$$f^A(\beta|h) = \lim_{N\to\infty} f_N^A(\beta|h) = -\frac{1}{\beta}\log 2\cosh(\beta h) - \frac{\beta}{4}. \tag{1.51}$$

Remark 1.7. Since the function $x \to \log x$ is concave, by the Jensen inequality we can immediately say that the quenched free energy is always greater or equal than the annealed one

$$-\frac{1}{\beta N}\mathbb{E}\log Z_N(\beta|h;\boldsymbol{J}) \ge -\frac{1}{\beta N}\log \mathbb{E}Z_N(\beta|h;\boldsymbol{J}).$$

Remark 1.8. It is also immediate to see that the annealed free energy cannot be the correct one, at least at low temperatures, if we look at the corresponding annealed entropy. In the zero-field case, in fact, this is given by

$$s^{\mathrm{A}}(\beta) = \beta^2\partial_\beta f^{\mathrm{A}}(\beta) = \log 2 - \frac{\beta^2}{4}, \tag{1.52}$$

and in particular it becomes negative for $\beta < \beta^* = 2\sqrt{\log 2}$. But entropy is by definition the logarithm of the number of configurations, and it cannot be negative for a discrete system.

1.3.3 Replicas and overlap

Previously, we vaguely introduced the concept of overlap, as defined in Eq. 1.43, by considering two copies (or more precisely *replicas*) of the system. In general, we can consider a generic number n of independent copies of the system, characterized by the spin configurations $\boldsymbol{\sigma}^{(1)}, ..., \boldsymbol{\sigma}^{(n)}$, distributed according to the product state

$$\Omega_{\boldsymbol{J}} = \omega_{\boldsymbol{J}}^{(1)} \times \omega_{\boldsymbol{J}}^{(2)} \times ... \times \omega_{\boldsymbol{J}}^{(n)}, \tag{1.53}$$

where each $\omega_{\boldsymbol{J}}^{(a)}$ acts on the corresponding $\sigma_i^{(a)}$ variables. We stress again that all the replicas are all subject to the same sample $\boldsymbol{J} = \{J_{ij}\}$ of the external disorder: These copies

of the system are usually called *replicas* [19]. When considering such a replicated system, the Boltzmann factor is simply given by the product of the corresponding Boltzmann factor for the single n replicas

$$\exp\Big(-\beta\Big(H_N(\boldsymbol{\sigma}^{(1)}|h;\boldsymbol{J})+H_N(\boldsymbol{\sigma}^{(2)}|h;\boldsymbol{J})+...+H_N(\boldsymbol{\sigma}^{(n)}|h;\boldsymbol{J})\Big)\Big). \tag{1.54}$$

Definition 1.9. Given a generic observable, represented by a smooth function $A=A(\boldsymbol{\sigma})$ of the configuration of the n replicas, we define the $\langle\cdot\rangle$ averages as

$$\langle A(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)},...,\boldsymbol{\sigma}^{(n)})\rangle=\mathbb{E}\Omega_{\boldsymbol{J}}(A(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)},...,\boldsymbol{\sigma}^{(n)})). \tag{1.55}$$

Replica overlaps are the quantities that one usually measures in numerical experiments. It is important to note that if we consider Boltzmann averages $\Omega_{\boldsymbol{J}}$ over different groups of replicas they factorize:

$$\Omega_{\boldsymbol{J}}(q_{12}q_{34})=\Omega_{\boldsymbol{J}}(q_{12})\Omega_{\boldsymbol{J}}(q_{34}).$$

It is instead the average over disorder which introduces correlations between them, since in general

$$\langle q_{12}q_{34}\rangle\neq\langle q_{12}\rangle\langle q_{34}\rangle.$$

On the other hand, these averages are invariant under permutation of replica indices, for instance

$$\langle q_{12}q_{23}\rangle=\langle q_{24}q_{45}\rangle.$$

The whole physical content of the theory is encoded in the distribution of overlap [19], and the averages of many physical quantities can be expressed as $\langle\cdot\rangle$ averages over overlap polynomials. For example, let us consider the disorder average of the internal energy per spin $N^{-1}\omega_{\boldsymbol{J}}(H_N)$ for $h=0$. Using the integration by parts formula

$$\mathbb{E}(JA(J))=\mathbb{E}\Big(\frac{\partial}{\partial J}A(J)\Big), \tag{1.56}$$

which is valid for a centered unit Gaussian variable J and any smooth function $A(J)$, it is straightforward to check that the energy density does not scale with the system size N:

$$E\equiv\frac{\langle H_N\rangle}{N}=\frac{1}{N}\mathbb{E}\,\omega_{\boldsymbol{J}}(H_N)=-\frac{\beta}{2}(1-\langle q_{12}^2\rangle). \tag{1.57}$$

Another example is given by its β derivative, which can be easily evaluated as

$$\begin{aligned}N^{-1}\partial_\beta\langle H_N\rangle &= -N^{-1}\left(\langle H_N^2\rangle-\langle H_N\rangle^2\right)\\ &= -\frac{1}{2}\left(1-\langle q_{12}^2\rangle\right)+\frac{N\beta^2}{2}\left(\langle q_{12}^4\rangle-4\langle q_{12}^2q_{23}^2\rangle+3\langle q_{12}^2q_{34}^2\rangle\right).\end{aligned}$$

1.3.4 The thermodynamic limit

The problem of proving the existence of the thermodynamic limit of the SK free energy remained open for more than twenty years, until the work by Guerra and Toninelli [12]. In order to prove the existence of the thermodynamic limit, as for the Curie-Weiss model we divide the N sites in two blocks N_1,N_2, with $N_1+N_2=N$, and define the auxiliary partition function

$$\begin{aligned}Z_N(\beta,t) &= \sum_{\boldsymbol{\sigma}}\exp\beta\Big(\sqrt{\frac{t}{N}}\sum_{1\le i<j\le N}J_{ij}\sigma_i\sigma_j+\sqrt{\frac{1-t}{N_1}}\sum_{1\le i<j\le N_1}J'_{ij}\sigma_i\sigma_j\\ &\quad+\sqrt{\frac{1-t}{N_2}}\sum_{N_1\le i<j\le N}J''_{ij}\sigma_i\sigma_j\Big),\end{aligned} \tag{1.58}$$

depending on the parameter $t \in [0,1]$. The external disorder is represented by the independent families of unit Gaussian random variables $\boldsymbol{J}$, $\boldsymbol{J}'$ and $\boldsymbol{J}''$. Let us stress that the two subsystem are subject to an external disorder which is independent with respect to the original system, but the probability distributions are the same. As in the previous case, the boundary values of the auxiliary partition function correspond respectively to the original system at $t = 1$, and to the two independent subsystems at $t = 0$:

$$Z_N(\beta, 1) = Z_N(\beta), \tag{1.59}$$
$$Z_N(\beta, 0) = Z_{N_1}(\beta) Z_{N_2}(\beta). \tag{1.60}$$

Consequently, the free energies are realized as

$$\mathbb{E} \log Z_N(\beta, 1) = -N\beta f_N(\beta), \tag{1.61}$$
$$\mathbb{E} \log Z_N(\beta, 0) = -N_1 \beta f_{N_1}(\beta) - N_2 \beta f_{N_2}(\beta). \tag{1.62}$$

Here, the disorder average is performed on all the variables $\boldsymbol{J}$, $\boldsymbol{J}'$ and $\boldsymbol{J}''$. The derivative with respect to t of the auxiliary free energy is given by

$$-\frac{d}{dt}\frac{1}{N\beta}\mathbb{E} \log Z_N(\beta, t) = -\frac{1}{2N}\mathbb{E}\Big(\frac{1}{\sqrt{tN}} \sum_{1 \le i < j \le N} J_{ij}\omega_t(\sigma_i\sigma_j) \tag{1.63}$$
$$-\frac{1}{\sqrt{(1-t)N_1}} \sum_{1 \le i < j \le N_1} J'_{ij}\omega_t(\sigma_i\sigma_j) - \frac{1}{\sqrt{(1-t)N_2}} \sum_{N_1 \le i < j \le N} J''_{ij}\omega_t(\sigma_i\sigma_j)\Big),$$

where $\omega_t(\cdot)$ is the Gibbs average corresponding to the auxiliary partition function (1.58). Using again the integration by parts formula on the previous expression, we have

$$\begin{aligned} -\frac{d}{dt}\frac{1}{N\beta}\mathbb{E} \log Z_N(\beta, t) &= -\frac{\beta}{4N^2} \sum_{1 \le i < j \le N} \mathbb{E}\left(1 - \omega_t^2(\sigma_i\sigma_j)\right) \\ &\quad + \frac{\beta}{4NN_1} \sum_{1 \le i < j \le N_1} \mathbb{E}\left(1 - \omega_t^2(\sigma_i\sigma_j)\right) \\ &\quad + \frac{\beta}{4NN_2} \sum_{N_1 \le i < j \le N} \mathbb{E}\left(1 - \omega_t^2(\sigma_i\sigma_j)\right) \\ &= \frac{\beta}{4}\langle q_{12}^2 - \frac{N_1}{N}(q_{12}^{(1)})^2 - \frac{N_2}{N}(q_{12}^{(2)})^2\rangle_t, \end{aligned} \tag{1.64}$$

where we wrote $\langle\cdot\rangle_t = \mathbb{E}\omega_t(\cdot)$ and defined the partial two-replica overlaps

$$q_{12}^{(1)} = \frac{1}{N_1} \sum_{1 \le i \le N_1} \sigma_i^1 \sigma_i^2, \tag{1.65}$$
$$q_{12}^{(2)} = \frac{1}{N_2} \sum_{N_1 \le i \le N} \sigma_i^1 \sigma_i^2, \tag{1.66}$$

corresponding to the two subsystems. The overlap plays here a role similar to the magnetization in the non-disordered case. Indeed, q_{12} is a convex linear combination of $q_{12}^{(1)}$ and $q_{12}^{(2)}$ of the form

$$q_{12} = \frac{N_1}{N} q_{12}^{(1)} + \frac{N_2}{N} q_{12}^{(2)}, \tag{1.67}$$

and, because of the convexity of the function $x \to x^2$, we have the inequality

$$\langle q_{12}^2 - \frac{N_1}{N}(q_{12}^{(1)})^2 - \frac{N_2}{N}(q_{12}^{(2)})^2\rangle_t \le 0. \tag{1.68}$$

Therefore, we can state as a preliminary result:

Lemma 1.2. *The quenched average of the logarithm of the interpolating partition function, defined by (1.58), increases in t, i.e.*

$$-\frac{d}{dt}\frac{1}{N\beta}\mathbb{E}\log Z_N(\beta,t)\leq 0. \tag{1.69}$$

Moreover, after integrating over t and recalling the boundary conditions (1.61, 1.62), we get the first main result

Theorem 1.3. *The free energy for the SK model is subadditive:*

$$Nf_N(\beta)\leq N_1 f_{N_1}(\beta)+N_2 f_{N_2}(\beta). \tag{1.70}$$

It is interesting to compare this result with the corresponding (1.27) for the Curie-Weiss model, whose free energy is superadditive. Of course, for the SK model it is the pressure $\alpha_N(\beta)=-\beta f_N(\beta)$ which is superadditive because of the minus sign. Together with an N-independent upper bound on the pressure, which is easy to obtain, one deduces again the existence of the thermodynamic limit (for both the pressure and the free energy density), therefore proving the following

Theorem 1.4. *The infinite volume limit for $f_N(\beta)$ exists and equals its infimum:*

$$f(\beta)\equiv\lim_{N\to\infty}f_N(\beta)=\inf_N f_N(\beta). \tag{1.71}$$

Remark 1.9. Note that this result is easily extended to the p-spin models (in which interactions are more than pairwise) since the overlaps to the square in (1.64) and (1.68) are simply replaced by the overlap to the power p, and the (1.69) still holds: this observation will be useful in the last Chapters of this thesis, when we will face how to overcome the actual state of the art in modeling AI via statistical mechanics.

1.3.5 The replica trick and Parisi theory

Parisi Theory (that resulted in the Nobel Prize to Giorgio Parisi this year) has been really a deep revolution in statistical mechanics, *de facto* opening the study of complex systems with a totally new perspective. Since Parisi developed his theory working on the SK model, it is impossible not to pay a minimal tribute and summarize his main results. However, we must also say that, as the theory itself is really tricky and its usage has not yet percolated in AI, we will not deepen it in all details and we remind the study of replica symmetry breaking to excellent textbooks [19, 31, 35]. The natural starting point to examine Parisi theory are the basic concepts of spontaneous symmetry breaking and phase coexistence in statistical mechanics [2, 5, 36]. We consider a system on a d-dimensional hypercubic lattice, defined by a Hamiltonian $H(\boldsymbol{\sigma})$, depending on the configurations of all spins σ_i, with $i\in\mathbb{Z}^d$. The system is initially restricted to a finite subset Λ of the lattice with partition function $Z_\Lambda(\beta)$, in order to deal with mathematically well-defined objects, and its finite volume free energy per site at the temperature $T=1/\beta$ is

$$f_\Lambda(\beta)=-\frac{1}{|\Lambda|\beta}\log Z_\Lambda(\beta), \tag{1.72}$$

where $|\Lambda|$ is the cardinality of the subset Λ. Then, one lets Λ grow to the whole infinite lattice $\mathbb{Z}^d$ in a suitable way imposing boundary conditions, i.e. the positions of the boundary spins or their interaction with the external world (with a certain arbitrariness). It can be proven that these conditions, if interactions have short range, do not affect the free energy

per site in the limit $\Lambda \to \mathbb{Z}^d$, but the equilibrium thermodynamic state of the system is also determined by all the correlation functions

$$\lim_{\Lambda \to \mathbb{Z}^d} \langle \sigma_{i_1} ... \sigma_{i_n} \rangle_\Lambda, \tag{1.73}$$

for all finite sets indices $i_1, ..., i_n$, where $\langle \cdot \rangle$ is the Boltzmann-Gibbs thermal average at the temperature $1/\beta$. The correlation functions in general depend on the choice of the boundary conditions, also in the infinite volume limit. Another usual and strictly related way to select different equilibrium states is to break a symmetry *explicitly* in the Hamiltonian, i.e. by introducing proper auxiliary external fields α_i which are removed only after the thermodynamic limit has been performed. More precisely, the thermodynamic limit for the free energy and for the correlation functions are computed with the explicitly broken symmetry Hamiltonian, and the external fields are then put to zero. In the Curie-Weiss model, for instance it is possible to select one of the two equilibrium states with positive or negative magnetization by introducing a term $-h \sum_i \sigma_i$ in the Hamiltonian which explicitly breaks the spin-flip symmetry, and taking the limit $h \to 0^{\pm}$ after the thermodynamic limit. The set of all equilibrium states forms a simplex, and every state can be written in an unique way as a convex linear combination of certain *extremal* states, called *pure states* or *pure phases*. They are characterized by the cluster property, or spatial decay of correlations, meaning that their connected correlations functions vanish at large distance (or for different points in mean field models):

$$\langle \sigma_{i_1} ... \sigma_{i_n} \sigma_{j_1} ... \sigma_{j_m} \rangle \to \langle \sigma_{i_1} ... \sigma_{i_n} \rangle \langle \sigma_{j_1} ... \sigma_{j_m} \rangle, \tag{1.74}$$

for

$$\min_{a,b} |i_a - j_b| \to \infty.$$

Pure states correspond to our intuitive idea of an equilibrium state. For example, in the Boltzmann-Gibbs state for water at zero Celsius the system has probability 1/2 of being all water and 1/2 of being all ice, while in a pure state the whole sample is water or ice. First order phase transitions are usually associated with the phenomenon of spontaneous symmetry breaking: the Hamiltonian of the model (and the non-clustering Boltzmann-Gibbs state) is invariant under the action of a symmetry group (for instance, the $\mathbb{Z}_2$ spin-flip transformation in the Curie-Weiss model, or rotational symmetry in the Heisenberg model), but equilibrium states belong to smaller symmetry groups. Therefore, it is the symmetry of the model suggesting the choice of the auxiliary external fields (or boundary conditions) which select the pure states, and applying the symmetry group transformation to a particular symmetry-breaking state one obtains another equilibrium state.

Spin-glasses are much more complicated from this point of view, since at low temperature there is an infinite number of pure phases, and it is not clear *a priori* which should be the right external fields (or boundary conditions) to select them, since the broken symmetry in the phase transition is not obvious. Moreover, due to this infinite number of states, the Gibbs phase rule, which states that $k-1$ thermodynamic parameters have to be fixed in order to have k coexisting pure phases (e.g. temperature and pressure in the triple point of a fluid), does not hold in this case. As Parisi showed, the spin glass phase transition is associated to a very peculiar spontaneous symmetry breaking, i.e. the group of permutations of a set of n identical replicas of the system in the limit $n \to 0$.

To explain this, we need to introduce the *replica trick*, which is the celebrated first method developed for the calculation of the free energy in complex scenarios (mainly statistical

mechanics of spin glasses and statistical field theory). The whole method is based on the representation of the (quenched) free energy as

$$f_N(\beta) = -\frac{1}{\beta N}\lim_{n\to 0}\frac{\mathbb{E}Z^n - 1}{n}. \tag{1.75}$$

The integer moments $\mathbb{E}Z_N^n$ of the partition function in the r.h.s. are simpler to compute than the averaged logarithm $\mathbb{E}\log Z_N$, and the trick consists in considering their analytic continuation to real n, and then taking the limit $n \to 0$. For integer n, the moments are nothing but the average of the partition function of a system of n identical (i.e. with the same disorder) replicas of the original system

$$\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = \mathbb{E}\sum_{\boldsymbol{\sigma}^{(1)}}...\sum_{\boldsymbol{\sigma}^{(n)}}\exp\Big(-\beta\sum_{a=1}^{n}H_N(\boldsymbol{\sigma}^{(a)}|h;\boldsymbol{J})\Big). \tag{1.76}$$

The disorder average can be easily carried out since it involves only independent Gaussian integrals, so we find

$$\begin{aligned}\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = &\exp\Big(\frac{\beta^2 n(N-n)}{4}\Big)\\ &\sum_{\boldsymbol{\sigma}^{(1)}...\boldsymbol{\sigma}^{(n)}}\exp\Big(\frac{\beta^2}{2N}\sum_{1\le a<b\le n}\Big(\sum_i \sigma_i^{(a)}\sigma_i^{(b)}\Big)^2 + \beta h\sum_{a=1}^{n}\sum_i \sigma_i^{(a)}\Big),\end{aligned} \tag{1.77}$$

which involves the square overlaps between replicas. The sum over configurations of replicated systems can be computed by linearizing each of these terms by Gaussian integrals. To do this, we introduce a $n\times n$ symmetric matrix Q_{ab} with zeros on the diagonal, and write the sum in (1.77) as

$$\begin{aligned}\sum_{\boldsymbol{\sigma}^{(1)}...\boldsymbol{\sigma}^{(n)}}\int\prod_{a<b}\Big(\sqrt{\frac{\beta^2 N}{2\pi}}dQ_{ab}\Big)\exp\Big(&-\frac{\beta^2 N}{2}\sum_{a<b}Q_{ab}^2\\ &+\beta^2\sum_{a<b}\Big(\sum_i\sigma_i^{(a)}\sigma_i^{(b)}\Big)Q_{ab} + \beta h\sum_a\sum_i\sigma_i^{(a)}\Big).\end{aligned} \tag{1.78}$$

Since clearly there are no couplings between spins belonging to the same replica, it is possible to define new spin variables $s_a = \pm 1$, with $a = 1,...n$, and observe that

$$\begin{aligned}&\sum_{\boldsymbol{\sigma}^{(1)}...\boldsymbol{\sigma}^{(n)}}\exp\Big(\beta^2\sum_{a<b}\Big(\sum_i\sigma_i^{(a)}\sigma_i^{(b)}\Big)Q_{ab} + \beta h\sum_a\sum_i\sigma_i^{(a)}\Big)\\ &\quad=\Big(\sum_{\{\boldsymbol{s}\}}\exp\Big(\beta^2\sum_{a<b}Q_{ab}s_a s_b + \beta h\sum_a s_a\Big)\Big)^N.\end{aligned}$$

Then, equation (1.77) becomes

$$\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = \int\prod_{a<b}\Big(\sqrt{\frac{\beta^2 N}{2\pi}}dQ_{ab}\Big)\exp(-NA[\boldsymbol{Q}]), \tag{1.79}$$

$$\begin{aligned}A[\boldsymbol{Q}] = &\frac{\beta^2}{2}\sum_{a<b}Q_{ab}^2 - \log\sum_{\{\boldsymbol{s}\}}\exp\Big(\beta^2\sum_{a<b}Q_{ab}s_a s_b + \beta h\sum_a s_a\Big)\\ &-\frac{\beta^2 n(N-n)}{4N},\end{aligned} \tag{1.80}$$

with the functional $A[\boldsymbol{Q}]$ depending on $\boldsymbol{Q}$, n, β and h. Since the exponent in the integrand of (1.79) is proportional to N, in the limit of N going to infinity the n-th moment of Z_N can be evaluated through the saddle point method. The infinite volume free energy, once the saddle point has been determined, is then obtained as

$$f(\beta, h) = \lim_{n\to 0} \frac{1}{\beta n} A[\boldsymbol{Q}_{sp}]. \tag{1.81}$$

Since $\boldsymbol{Q}$ is a symmetric matrix with zeros on the diagonal, the model $n(n-1)/2$ independent order parameters, and for a given choice of $\boldsymbol{Q}$ there are such many saddle-point equations $\partial A/\partial Q_{ab} = 0$, which take the form

$$Q_{ab} = \frac{\sum_{\{\boldsymbol{s}\}} s_a s_b \exp\left(\beta^2 \sum_{a<b} Q_{ab} s_a s_b + \beta h \sum_a s_a\right)}{\sum_{\{\boldsymbol{s}\}} \exp\left(\beta^2 \sum_{a<b} Q_{ab} s_a s_b + \beta h \sum_a s_a\right)} \tag{1.82}$$

In the limit $n \to 0$, it can be shown [19] that the r.h.s. of this equation is equivalent to

$$\mathbb{E}\Omega_{\boldsymbol{J}}(\sigma_i^{(a)}\sigma_i^{(b)}) \equiv \langle \sigma_i^{(a)}\sigma_i^{(b)} \rangle,$$

whence, since all sites i are equivalent for large N, the saddle point equation (1.82) can be written as

$$\lim_{n\to 0} Q_{ab} = \langle q_{ab} \rangle. \tag{1.83}$$

This relation is valid for a replica symmetric solution (as we will shortly see). When this symmetry is broken, if a particular choice of $\boldsymbol{Q}$ is a solution of the saddle point equation, then any matrix obtained with a permutation of rows or columns of $\boldsymbol{Q}$ will also be a solution. Therefore, in general one should divide the l.h.s. by $n(n-1)/2$. In the spin glass phase, the average overlap is expected to be different from zero, since it is the average of the positive quantity $\omega^2_{\boldsymbol{J}}(\sigma_i)$ for different realizations of the disorder (while $\omega_{\boldsymbol{J}}(\sigma_i)$ can be positive or negative depending on the particular realization of $\boldsymbol{J}$, and its average vanishes). On the other hand, in the high temperature phase the thermal average of magnetization in each site is zero for every sample, so that $\langle \sigma_i^{(a)}\sigma_i^{(b)} \rangle = 0$.

1.3.6 Replica Symmetric *Ansatz*

Before solving the saddle point equations, one has to choose a form for $\boldsymbol{Q}$ which is symmetric with respect to permutation of row or columns (due to equivalence among replicas). Then, the most natural idea seems to look for a *replica symmetric* (RS) saddle point, corresponding to a matrix $\boldsymbol{Q}$ whose non-diagonal elements are all equal to the same value q, while diagonal elements vanish identically. The integral in Eq. (1.79) then reduces to an ordinary integral over the real variable q, and the quenched free energy is easily computed as

$$-\beta f_{RS}(\beta, h) = \log 2 + \int_{-\infty}^{+\infty} d\mu(z) \log\cosh(\beta\sqrt{q}z + \beta h) + \frac{\beta^2}{4}(1-q)^2, \tag{1.84}$$

where $d\mu(z) = (2\pi)^{-1/2} e^{-z^2/2} dz$ is the Gaussian measure and q satisfies the saddle point equation

$$q = \int_{-\infty}^{+\infty} d\mu(z) \tanh(\beta\sqrt{q}z + \beta h). \tag{1.85}$$

At zero external field, this equation correctly predicts a phase transition at $1/\beta_c = T_c = 1$, since it has solution $q = 0$ for $\beta < \beta_c$ and it admits a solution with $q \neq 0$ for $\beta > \beta_c$.

However, it is possible to see [19] that the replica symmetric free energy is not physically acceptable for a temperature $T < T_c(h)$, since it violates basic thermodynamic stability conditions (such as, for example, the positivity of entropy [18]). The free energy (1.84) can be expanded near the critical point, where the spin glass parameter q is expected to be small. Then, the coefficient for the q^2 term, which according to Landau theory of phase transitions vanishes at the critical point [2], is found to be proportional to $\beta^2 - 1$, so that, consistently, $\beta_c = 1$. It is interesting to note that this coefficient is negative if $\beta < \beta_c$, so that the paramagnetic solution $q = 0$ maximizes (instead of minimizing) the free energy. The same also holds for a spin glass solution with $q > 0$ in the low-temperature phase $\beta > \beta_c$. This is a consequence of the fact that the number $n(n-1)/2$ of replica pairs becomes negative in the limit $n \to 0$ [19, 31]. Since the RS solution is not physically valid everywhere, one has to look for a form of the $\boldsymbol{Q}$ which breaks symmetry between replicas. The correct solution was given by Parisi by means of a powerful *Ansatz*, i.e. the broken replica symmetry ansatz.
We will now present a brief description of the basic philosophy behind it.
In the Ising model at low temperature and zero magnetic field, there is a symmetry breaking with two pure phases, one with magnetization $+m(\beta)$ and the other with $-m(\beta)$. The overlap (1.43) between two typical configurations belonging to the same phase equals

$$q_{++} = q_{--} = m^2(\beta),$$

while, for two different phases,

$$q_{+-} = -m^2(\beta).$$

We stress that symmetry breaking (as well as phase transitions) can be present, strictly speaking, only in the thermodynamic limit. In the limit of infinite volume, the distribution function of the overlap q_{12} between the configurations of two replicas , picked according to their Boltzmann weights, is given by the sum of two delta functions:

$$\mathcal{P}(q) = \frac{\delta(q - m^2(\beta)) + \delta(q + m^2(\beta))}{2}. \tag{1.86}$$

Above the critical temperature, on the other hand, there is just one pure phase with zero magnetization, and in this case we have

$$\mathcal{P}(q) = \delta(q). \tag{1.87}$$

This means that, looking at $\mathcal{P}(q)$, one is able to detect the phenomenon of non-uniqueness of the state without introducing an explicitly symmetry breaking field or proper boundary conditions. Since for spin glasses there is no obvious symmetry to be broken, with associated order parameter and field, the natural way to proceed is to compute

$$\mathcal{P}(q) = \lim_{N\to\infty} \mathbb{E}\mathcal{P}^{(N)}_{\boldsymbol{J}}(q),$$

where $\mathcal{P}^{(N)}_{\boldsymbol{J}}(q)$ is the finite volume probability distribution of the overlap for a given disorder realization $\boldsymbol{J}$. When $\mathcal{P}(q)$ is a single delta distribution the system is said to be replica symmetric. The same holds when $\mathcal{P}(q)$, in absence of magnetic field, is the sum of two deltas, with the two corresponding states related by spin-flip symmetry. On the contrary, if $\mathcal{P}(q)$ has more than two peaks, or it has a continuous part, replica symmetry is said to be broken. Knowing the distribution $\mathcal{P}(q)$ is then equivalent to know the structure of pure states. Given the average overlap

$$\langle q_{12} \rangle = \frac{1}{N} \sum_i \mathbb{E}\Omega_{\boldsymbol{J}}(\sigma_i^{(1)} \sigma_i^{(2)}),$$

we can think to express the Boltzmann weights $\Omega_{\boldsymbol{J}} = \omega^{(1)} \times \omega^{(2)}$ in terms of pure states, and this decomposition is encoded in the $\mathcal{P}(q)$:

$$\langle q_{12} \rangle = \int dq \mathcal{P}(q) q. \tag{1.88}$$

This equation, combined with 1.83, tells us that in the language of replicas $\mathcal{P}(q)$ represents the fraction of elements of the matrix $\boldsymbol{Q}$ assuming the value q [19].

1.3.7 Guerra's interpolating scheme

The idea behind the method precisely follows the same reasoning of the CW case (exploited in Section 1.2.3), despite obvious mathematical differences: to make them clear, we directly introduce the next

Definition 1.10. The interpolating partition function and the interpolating quenched free energy in the Guerra's scheme read as

$$Z_N(\beta, t) = \sum_{\boldsymbol{\sigma}} \exp\left\{ \sqrt{t} \frac{\beta}{\sqrt{N}} \sum_{i<j} J_{ij} \sigma_i \sigma_j + A\sqrt{1-t} \sum_i z_i \sigma_i \right\}, \tag{1.89}$$

$$f_N(\beta, t) = -\frac{1}{\beta N} \mathbb{E} \log Z_N(\beta, t). \tag{1.90}$$

Of course, one can also defined the (disorder-dependent) Boltzmann factor $B_N(t)$ and the Boltzmann-Gibbs state $\omega_t(\cdot)$ in perfect analogy to the CW model:

$$B_N(t) = \exp\left\{ \sqrt{t} \frac{\beta}{\sqrt{N}} \sum_{i<j} J_{ij} \sigma_i \sigma_j + A\sqrt{1-t} \sum_i z_i \sigma_i \right\},$$

$$\omega_t(F) = \frac{\sum_{\boldsymbol{\sigma}} F(\boldsymbol{\sigma}) B_N(t)}{\sum_{\boldsymbol{\sigma}} B_N(t)}.$$

Finally, one can define the (thermodynamic limit of the) statistical pressure in the usual way $\alpha_N(\beta, t) = -\beta f_N(\beta, t)$. Of course, the original system is reproduced at $t = 1$, while for $t = 0$ we replaced the problem with a one-body interacting system. The quenched free energy of the SK model (in the thermodynamic limit) is therefore given by the sum rule

$$f(\beta) \equiv f(\beta, t = 1) = f(\beta, t = 0) + \int_0^1 ds \left[\partial_t f(\beta, t) \right]_{t=s}. \tag{1.91}$$

Some comments are in order here. First of all, the main difference w.r.t. the CW interpolation scheme is that, here, each spin is subjected to a different external field z_i (which is however chosen to share the same Gaussian distribution for all the sites). In the CW model, this feature was not needed since all the couplings were equal (this can be seen as Gaussian distributions collapsing to Dirac deltas). Then, in order to have a z-independent partition function, we should also average over the z realizations. Moreover, we also stress that, w.r.t. the CW model, the interpolating parameter appears through square roots. This is needed because, in the computation, we should use the integration by parts formula over quenched disorder, so this choice is used to precisely cancel unwanted factors.[1] The coefficient A in the definition of the generalized partition function will be

[1] For a $\mathcal{N}(0,1)$ variable X, we recall that the integration by parts formula is $\mathbb{E}_X X f(X) = \mathbb{E}_X \partial_X f(X)$.

determined later. As a final note, we again omitted the dependence of previous quantities on the quenched disored $\boldsymbol{J}$ and $\boldsymbol{z}$ to make the notation more compact.

The derivative of the generalized free energy with respect to the interpolating parameter t is:

$$\frac{df(\beta,t)}{dt} = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\Big(\frac{1}{2\sqrt{t}}\frac{\beta}{\sqrt{N}}\sum_{i<j}J_{ij}\omega_t(\sigma_i\sigma_j) - \frac{A}{2\sqrt{1-t}}\sum_i z_i\omega_t(\sigma_i)\Big). \tag{1.92}$$

Then, integrating by parts w.r.t. to the variables J_{ij} and z_i, we have

$$\frac{df(\beta,t)}{dt} = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\Big(\frac{\beta^2}{4N}\sum_{ij}(1-\omega_t(\sigma_i\sigma_j)^2) - \frac{A^2}{2}\sum_i(1-\omega_t(\sigma_i)^2)\Big). \tag{1.93}$$

The next point in the resolution is to note that the squares of spin correlation functions can be linked to the order parameter of SK model by expressing them in terms of the $\langle\cdot\rangle$ averages previously defined. Indeed, we have

$$\sum_i \mathbb{E}\omega_t(\sigma_i)^2 = \sum_i \mathbb{E}\,\omega_t^{(1)}\times\omega_t^{(2)}(\sigma_i^{(1)}\sigma_i^{(2)}) = N\langle q_{12}\rangle_t, \tag{1.94}$$

$$\sum_i \mathbb{E}\omega_t(\sigma_i\sigma_j)^2 = \sum_i \mathbb{E}\,\omega_t^{(1)}\times\omega_t^{(2)}(\sigma_i^{(1)}\sigma_i^{(2)}\sigma_j^{(1)}\sigma_j^{(2)}) = N^2\langle q_{12}^2\rangle_t. \tag{1.95}$$

Therefore, the derivative of the interpolating free energy is

$$\frac{df(\beta,t)}{dt} = -\frac{\beta}{4}\lim_{N\to\infty}\mathbb{E}\Big(1-\langle q_{12}^2\rangle_t - \frac{2A^2}{\beta^2}(1-\langle q_{12}\rangle_t)\Big). \tag{1.96}$$

Choosing now $A=\beta\sqrt{q}$, where q is the thermodynamic value of the overlap (meaning that we are assuming the replica symmetric *Ansatz* since, in the thermodynamic limit, it does not fluctuate), we have

$$\frac{df(\beta,t)}{dt} = \frac{\beta}{4}\lim_{N\to\infty}\mathbb{E}\Big(\langle(q_{12}-q)^2\rangle_t - (1-q)^2\Big). \tag{1.97}$$

In the thermodynamic limit and in the replica symmetry regime, the overlap assumes its thermodynamic value q with probability 1. Therefore, the first term at the r.h.s. in the last equation goes to zero, leaving only with

$$\frac{df(\beta,t)}{dt} = -\frac{\beta}{4}(q-1)^2. \tag{1.98}$$

The computation of the $t=0$ case is straightforward, since it is a one-body problem with Gaussian disorder. Indeed, we easily get

$$\begin{aligned} f(\beta,0) &= -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big(A\sum_i z_i\sigma_i\Big) = \\ &= -\lim_{N\to\infty}\frac{1}{\beta N}\sum_i\mathbb{E}\log 2\cosh(Az_i). \end{aligned} \tag{1.99}$$

In this last equation, the quenched average involves only the z variables. The result of this integration is actually independent on the index i. Therefore, by recalling the choice for the parameter A, this directly implies that

$$f(\beta,0) = -\frac{1}{\beta}\mathbb{E}\log 2\cosh(\beta\sqrt{q}z). \tag{1.100}$$

By putting everything together according to the sum rule (1.91) and making the Gaussian integration explicit, we get the next

Theorem 1.5. *The explicit expression for the SK quenched free energy in terms of the two replica overlap, in the thermodynamic limit and under the replica symmetric assumption, reads as*

$$f_{RS}(\beta) = -\frac{1}{\beta}\int_{-\infty}^{+\infty} d\mu(z)\log 2\cosh(\beta\sqrt{q}z) - \frac{\beta}{4}(1-q)^2. \tag{1.101}$$

The latter equation precisely reproduce the replica trick prediction (1.84) with vanishing external field $h = 0$.

So far we equipped ourselves with a methodology, statistical mechanics, and two archetypal models (the Curie-Weiss and the Sherrington-Kirkpatrick), with the whole related package of concepts (e.g., replicas, overlaps, etc.): we collected the minimal knowledge to tackle Theoretical Artificial Intelligence with these tools hence, in the next section, we address one of the most famous neural network (that works both as an associative memory and as a pattern recognition network), namely the Hopfield model.

1.4 Generalities on the Hopfield neural network

Neural network models are complex systems designed on the basis on the associative memory notion and on the principle that stable neural activities encode retrieved patterns of information (e.g. images). By *associative memory* we mean the ability of cortical modules in mammals' brain to remember names, objects, faces, schemes, etc. (i.e. *patterns of information* generally speaking) starting from incomplete or corrupted data supply. Let us illustrate hereafter a very minimal description about how the neural system works (following the milestone by Amit [11]) obviously, still from a modelling perspective.

Neurons can be considered as big cells, called *soma*, covered by a membrane to which are attached different fibres emitting electrical spikes generated from the neuron itself. The outgoing signal passes through a bigger fibre conduct called the *axon*. The latter splits into smaller channels that are attached, through the *dendrites*, to the external membrane of other neurons. The point of conjunction of the dendrites with the recipient neuron is called *synapse*. When a neuron is active, it emits an electrical wave propagating across the different dendrites. At the end of this process, a new electrical potential on the synapse of the recipient neurons. The emission of these packs happen when the total synaptic potential, i.e. the sum of the potentials received from other neurons, is higher than a certain *activation threshold* $\bar{h}$ and are active at random times (asynchronous dynamic). In 1949, D. Hebb pointed out the fact that neural pathways are strengthened each time they are used, a concept fundamentally essential to the ways in which humans learn. If two nerves fire at the same time - he argued - then the connection between them is enhanced [37]. The total number of neurons in the human brain is between 10^9 and 10^{10}, and each neuron is generally connected to $10^4/10^5$ other neurons through dendrites. A bridge between neuron dynamics and memory processes has been made thanks to Y. Miyashita's experiments (1988) [38], in which a trained monkey showed neural activity in a well defined region once a picture is presented for the first time. The same group of neurons reactivates when the monkey sees the same typology of images.

The theoretical prototype for a wide class of associative memory models is the Hopfield network [39]. It is a strongly stylized version of a cortical module which is based on the basic assumptions that

- There are essentially two types of variables: neurons (nodes in the neural network) and synapses (links between the nodes). These variables live on very separate time scales, so that we can question about neural dynamics and emerging properties of networks of interacting neurons keeping quenched the synapses;

- There is just one type of neurons and it is represented as a binary variables (e.g. Ising spins or Boolean variables), whose possible values represent respectively its firing (+1) or its quiescent (-1) states;

- The synapses are both excitatory and inhibitory. On average, the 50% of them are positive (excitatory) and the remaining 50% negative, i.e. inhibitory, leaving the bulk of the Hopfield paradigm stable. While the different nature of the synapses is a biological must, the balanced ratio is instead biologically unreasonable, since we know that there is a larger fraction in inhibitory contributions (but this simplification has been already overcome a long time ago [11]);

- The interactions are assumed to be symmetric, i.e. $J_{ij} = J_{ji}$. Again, this is false from the biological point of view (Dale law actually states the opposite [37]). However,

as masterfully discussed by Amit, this wrong assumption is one of the most clever starting point in order to construct a reference framework: this is because - as long as the couplings are symmetric - the *detailed balance* holds and any - reasonably not pathological - stochastic neural dynamics converges to the Gibbs measure for an opportune cost-function, e.g. the Hopfield Hamiltonian [8].

In the first part of the present Section, we first give a mathematical glance at the Hopfield network and the statistical mechanical quantities that we need to tackle its emergent properties. After that, we illustrate the connection between the models that we studied in the previous chapters (i.e. the Curie-Weiss model and the Sherrington-Kirkpatrick mean field spin glass) and the Hopfield network, thus justifying the previous discussion and therefore motivating the key role they (i.e. CW and SK) actually play as "limiting cases" for the behaviour of the Hopfield model (respectively, for too few and too many stored patterns). Finally, we will address the problem of pattern storage via the signal-to-noise technique, closing the descriptive part of the properties of the Hopfield network. In the second part, we will address the problem of obtaining a phase diagram for Hopfield model by heavily relying upon the statistical mechanical techniques we have shown so far (mainly replica trick and interpolation method), focusing on various types of information processing (ranging from storing digital to real patterns).

We consider a fully connected neural network consisting in N neurons. To each of them i is assigned a dichotomic variable σ_i whose possible values represent the active ($\sigma_i = +1$) or quiescent ($\sigma_i = -1$) states. It is worth noticing that the mean field approximation is here not as rude as in Physics of many-body systems (since neurons are effectively highly connected and each neuron in the cortex may share connections with up to $O(10^4/10^5)$ peers). Of course, we shall not consider this as a model of the brain network as a whole, but rather of the small different regions involved with the memorization of patterns.

We start our discussion by giving the following

Definition 1.11. The synaptic potential h_i that the i-th neuron receives from the other $N-1$ is defined as

$$h_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} J_{ij}\sigma_j,$$

where J_{ij}, the synaptic matrix, codes the intensity of the synaptic action of neuron j over neuron i.

Associative memory models are built to recognize a certain group of words or images or concepts, i.e. *patterns*, so the next step is to formalize how the information is encoded in neural networks. A *pattern* is defined as a sequence of random variables $\xi = (\xi_1, \dots, \xi_N)$. In this thesis, we will mainly work with Boolean and Gaussian patterns, namely patterns whose entries are extracted according to a given probability distribution, respectively $\mathcal{P}(\xi_i = +1) = \mathcal{P}(\xi_i = -1) = 1/2 \ \forall i$ for the Boolean case and $\mathcal{P}(\xi_i) = \mathcal{N}(0,1) \ \forall i$ in the Gaussian one. All the patterns we will deal with will share the same length N. Since we want to store several patterns, we have to introduce another index -μ- for labelling different patterns: $\{\xi^1, \dots, \xi^P\}$. In doing this, we shall assume that each ξ_i^μ is independent from the others.

The choice of the synaptic coupling $J_{ij} \ \forall i, j = 1, \dots, N$ ensuring the local attractiveness of each pattern under the neural dynamics (see [34]) is the one incorporating Hebb's

learning rule, i.e.

$$J_{ij} := \frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu. \tag{1.102}$$

Once we specified the nature of dynamical variables and the interaction matrix, we can continue by introducing the Hamiltonian for the Hopfield model.

Definition 1.12. The Hamiltonian (or *cost function* in Machine Learning jargon) of the Hopfield model equipped with N neurons σ_i, $i \in (1, ..., N)$ and P patterns ξ^μ, $\mu \in (1, ..., P)$ is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) := -\frac{1}{N}\sum_{\mu=1}^{P}\sum_{1\leq i<j\leq N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j. \tag{1.103}$$

The next step is to introduce a set of spin-dependent quantities measuring the resemblance of a given network configuration with the stored patterns. These quantities will clearly play the role of order parameters for the Hopfield model and are provided in the next

Definition 1.13. We define P overlaps m_μ, $\mu \in (1, ..., P)$ between the patterns and the neurons, also called Mattis magnetizations, as

$$m_\mu(\boldsymbol{\sigma}) :\doteq m_\mu = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i \in [-1, 1]. \tag{1.104}$$

Note that the Hamiltonian of the Hopfield model can be nicely written in terms of these order parameters as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) \sim -\frac{N}{2}\sum_{\mu=1}^{P}m_\mu^2.$$

It is then crystal clear that, in order for the energy to be minimized, it is more convenient for some m_μ to equal to $+1$ (or -1 because of the spin-flip symmetry $\sigma_i \to -\sigma_i$) meaning that the neurons are all parallel to the pattern, thus eventually indicating a retrieving behaviour.

1.4.1 The CW and the SK limits

In this Section, we illustrate the crucial connection between the Hopfield model and the two already analyzed models, namely the Curie-Weiss and the Sherrington-Kirkpatrick.

The mathematical models of associative memory systems are built in such a way that the distribution of neural activity at an equilibrium state is a codification of a recognized image or notion. In particular, the act of retrieving stored data from partial informations is strictly correlated to finding the minimum values of the system energy. The Sherrington-Kirkpatrick model displays a large number of energy minima (as expected for a cognitive system), yet it is not suitable to act as a associative memory model since its equilibrium states are too "disordered". The Hamiltonian introduced above presents global minima which are not purely random like those in SK (since they must represent ordered stored patterns, a feature which resembles the CW model), but the amount of these minima must be possibly extensive in the number of spins/neurons N. Therefore, a reasonable associative neural network should be designed in order to retain a "ferromagnetic flavor" within a "glassy panorama", i.e. we need something in between. Remarkably, the Hopfield model defined by (1.103) lies exactly in between a Curie-Weiss model and a Sherrington-Kirkpatrick model. Let us clarify this point.

From the CW to Hopfield

By comparing (1.16) and (1.103), and in particular their expression through the order parameters, we can firstly observe that CW model can be interpreted as an (actually very rudimental) model of a neural network where N neurons collaborate to store one pattern of information (together with its spin-flip symmetric partner). Such information patterns, which are built of by all the same numbers (for instance, the sequences $+1, +1, ..., +1$ and $-1, -1, ..., -1$), beyond containing no information by Shannon compression arguments, in turn they represent pathological behaviours (since all the neurons are simultaneously firing or silent). This last criticism can be easily overcome thanks to the Mattis-gauge, namely a re-definition of the neurons as

$$\sigma_i \longmapsto \xi_i \sigma_i,$$

where $\xi_i = \pm 1$ are quenched random entries extracted with equal probability.

Definition 1.14. The Mattis Hamiltonian reads as

$$H_N^{Mattis}(\boldsymbol{\sigma}, \xi) = -\frac{1}{N} \sum_{i=1}^{N} \xi_i \xi_j \sigma_i \sigma_j.$$

The Mattis magnetization is defined as

$$m_M = \frac{1}{N} \sum_{i=1}^{N} \xi_i \sigma_i.$$

In order to inspect the network properties in its lowest energy minima, we perform a comparison with the CW model in the noiseless case $\beta \to \infty$. In terms of the (standard) magnetization,the Curie-Weiss model reads as $H_N(\boldsymbol{\sigma}) \simeq -Nm^2/2$ and, analogously for $H_N^M(\boldsymbol{\sigma}, \xi)$ we have

$$H_N^M(\boldsymbol{\sigma}, \xi) \simeq -\frac{N}{2} m_M^2.$$

It is then clear that, in the low noise limit (where collective properties may emerge), as the minimum of free energy is achieved in the Curie-Weiss model for $m \to \pm 1$, the same holds in the Mattis model for $m_M \to \pm 1$. The only difference lies in the fact that, in the latter case, spins tend to align in parallel (or anti-parallel) to the vector ξ. For instance, if the pattern ξ is, say, $\xi = (+1, -1, -1, -1, +1, +1)$ in a model with $N = 6$, the equilibrium configurations of the network will be $\boldsymbol{\sigma} = (+1, -1, -1, -1, +1, +1)$ and the spin-flip symmetric partner $\boldsymbol{\sigma} = (-1, +1, +1, +1, -1, -1)$. Thus, the network relaxes autonomously to a state where some of its neurons are firing while others are quiescent, as prescribed by the stored pattern ξ. We stress that, as the entries of the vectors ξ are chosen randomly to be ± 1 with equal probability, the retrieval of free energy minimum now corresponds to a spin configuration which is also the most entropic for the Shannon-McMillan argument. Thus, both the most likely and the most difficult to handle (as its information compression is no longer possible).

Two remarks are in order. At this point, one would be tempted to call the spins σ_i neurons, but it is definitely inconvenient to build a network via N spins/neurons, which are further meant to be diverging (i.e. $N \to \infty$), in order to handle one stored pattern of information only. Along the theoretical physics route, overcoming this limitation is quite natural (as provides the Hebbian prescription): if we want a network able to cope with P

patterns, the simplest Hamiltonian should simply be the sum of Mattis Hamiltonians over these stored patterns, namely

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{1\leq i,j\leq N}\Big(\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\Big)\sigma_i\sigma_j,$$

thus recovering the definition (1.103) for the Hopfield network Hamiltonian. Therefore, we can conclude that the Curie-Weiss network can be interpreted as a Hopfield neural network where solely one trivial pattern can be handled.

From the SK to Hopfield

Despite the extension to the case $P > 1$ is formally straightforward, the investigation of the system as P grows becomes by far more tricky. Indeed, neural networks belong to the so-called "complex system" realm. Complex properties can be distinguished by simple behaviours with the fact fact that for the latter the number of free-energy minima of the system does not scale with the volume N, while for complex systems the opposite feature takes place according to a proper function of N. In particular, the Curie-Weiss/Mattis model has two minima only, whatever N (even if $N \to \infty$), thus constituting the paradigmatic example for a simple system. On the other side, we introduced the prototype of complex systems, the Sherrington-Kirkpatrick model, that presents an amount of minima scaling as $\sim e^{cN}$ (with c not depending on N).

We showed above how, when $P = 1$ the Hopfield model (with boolean patterns) recovers the Mattis model (which is nothing but a gauge-transformed Curie-Weiss model). Conversely, when $P \to \infty$,

$$\frac{1}{\sqrt{N}}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu \longrightarrow \mathcal{N}(0,1),$$

by virtue of the standard central limit theorem, so that the Hopfield model recovers the Sherrington-Kirkpatrick one. To understand this point, we start by considering the Hebb construction of the synaptic strength

$$J_{ij} = \frac{1}{N}\sum_{\mu}\xi_i^\mu\xi_j^\mu, \tag{1.105}$$

where each pattern bit is extracted (in our analysis) with probability $\mathcal{P}(\xi_i^\mu = \pm 1) = 1/2$. Since each pattern independently and identically distributed (i.i.d.), this directly implies that $\mathcal{P}(\xi_i^\mu\xi_j^\mu = \pm 1) = 1/2$ itself, meaning that $\mathbb{E}\,\xi_i^\mu\xi_j^\mu = 0$ and $\mathrm{Var}(\xi_i^\mu\xi_j^\mu) = 1$. When summing a large number of such variables, they should be described (in agreement with the central limit theorem, CLT) with a Gaussian distribution. Indeed

Theorem 1.6 (Central Limit Theorem)**.** *Consider a set $X_1, \dots, X_n$ of i.i.d. random variables with mean μ_i and variance $\sigma_i^2 < \infty$, and call*

$$s_n^2 = \sum_{i=1}^{n}\sigma_i^2. \tag{1.106}$$

If, for some $\delta > 0$, the Lyapunov condition is satisfied

$$\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}[|X_i - \mu_i|^{2+\delta}] = 0 \tag{1.107}$$

then the quantity $s_n^{-1}\sum_i(X_i - \mu_i)$ converges (in distributional sense) to $\mathcal{N}(0,1)$.

The Hebb coupling matrix can be rewritten as $J_{ij} = \sqrt{\frac{\alpha_N}{N}}\tilde{J}_{ij}$, where

$$\tilde{J}_{ij} = \frac{1}{\sqrt{P}} \sum_{\mu} \xi_i^\mu \xi_j^\mu, \tag{1.108}$$

and $\alpha_N = P/N$ is the storage capacity (at finite N).[1] Now, since the variables $\xi_i^\mu \xi_j^\mu$ have zero mean and variance 1, we have $s_n = P^{-1/2}$. It is straightforward to verify that such a sample of variables satisfy the Lyapunov condition for all $\delta > 0$. Thus, for large P the coupling matrix J converges in probability to $\mathcal{N}(0,1)$.

Remark 1.10. These result is mathematically rigorous only if P is sent into infinity *independently* on the network size N.

The argument presented above suggests that, when the numbers of stored patterns is too large with respect to the network size, the Hebb coupling matrix behaves (apart for a constant prefactor) as

$$J_{ij} \sim \frac{1}{\sqrt{N}} \tilde{J}_{ij}, \tag{1.109}$$

where $\mathcal{P}(\tilde{J}_{ij}) = \mathcal{N}(0,1)$ for all i, j. This is indeed the form of the coupling matrix for the Sherrington-Kirkpatrick model. Therefore, Hopfield model with a too high stored information is expected to behave as a spin glass network. This naive argument turns out to be true: for α high enough, Hopfield model behaves as a spin glass model, with some differences with respect to the SK case. Such a crossover between CW (or Mattis) and SK models signals that, in order to investigate its statistical properties, we need both the P Mattis magnetizations m_μ (quantifying retrieval of the whole stored patterns, that is the vocabulary), and the two-replica overlaps Q_{ab} (to control the glassiness growth if the vocabulary gets enlarged). Moreover, we also a tunable parameter measuring the ratio between the stored patterns and the amount of available neurons, namely $\alpha = \lim_{N\to\infty} P/N$, i.e. the *storage capacity* at large N. As far as P scales sub-linearly with N (i.e. in the low storage regime with $\alpha = 0$), the phase diagram is ruled by the noise level β only: for $\beta < \beta_c$ the system is a paramagnet (with both $m_\mu = 0$ and $Q_{ab} = 0$), while for $\beta > \beta_c$ the system performs as an attractor, with $m_\mu \neq 0$ for a given $\mu \in (1, \dots, P)$. In this regime, no dangerous glassy phase is lurking, yet the model is able to store only a tiny amount of patterns. Conversely, when P scales linearly with N, i.e. in the high-storage regime defined by $\alpha > 0$, the phase diagram lives in the α, β plane. When α is small enough, the system is expected to behave similarly to $\alpha = 0$ case, hence as an associative network (with a particular non-vanishing Mattis magnetization but also with the two-replica overlap slightly positive, since the glassy nature is intrinsic for $\alpha > 0$). However, for α large enough, the Hopfield model collapses on the Sherrington-Kirkpatrick model as expected, with the Mattis magnetizations brutally reduced to zero and the two-replicac overlap close to one. The transition to the spin-glass phase is often called "blackout scenario" in neural network community [11, 40, 41].

We can summarize the content of the Hopfield model capabilities through its phase diagram as follows.[2] First of all, if the thermal noise $T = \beta^{-1}$ and the storage capacity α are sufficiently low, the system works with almost no errors as an associative neural network

[1] Notice that, throughout the rest of the thesis, we will use simply α also if we are working at finite size N, but the rigorous definition of the storage capacity is $\alpha := \lim_{N\to\infty} P/N$.

[2] What follows is strictly true only in the thermodynamic limit, replica symmetric regime and uncorrelated patterns.

(or pattern recognizer), meaning that the attractors associated to stored patterns are very stable (they are global minima in the quenched free energy landscape). In particular, in the noiseless case $\beta \to \infty$, the critical capacity bounding such a regime is $\alpha_c \simeq 0.051$. Outside this region, the network could still work as an associative memory, but the stored patterns are just local minima (with the spin glass states starting to dominate the landscape): this is the scenario provided that the storage capacity $0.0051 \leq \alpha \leq \alpha_c \simeq 0.138$. For $\alpha > 0.138$, the minima related to the patterns are destroyed and solely the spin-glass panorama remains stable.

Re-introducing the noise in the discussion, the network can escape from the retrieval region in the phase diagram, essentially in one more way. If the noise in the network is above the critical line $T_c = 1 + \sqrt{\alpha}$, the network lies in its ergodic phase: making these predictions quantitative is a non-trivial task in statistical mechanics as we will see in details soon. With respect to the storage capacity α, we distinguish between the following two regimes:

1.4.2 A heuristic digression about the phase space structure

Let us now get more acquainted with the statistical mechanical picture of the Hopfield model. To recall the notation, we have a set of P digital patterns ξ^μ with $\mu = 1, \dots, P$ of length N, and we want to store them in a network composed by N boolean spins $\sigma_i = \pm 1$ for $i = 1, \dots, N$. According to the Hebb rule, the memory is allocated in the synaptic strength by building up the coupling matrix as

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu. \tag{1.110}$$

Then, if we assume that the network evolves sequentially according to the update rule[1]

$$\sigma_i(t+1) = \text{sign}(\tanh(\beta \sum_{j \neq i} J_{ij} \sigma_j(t)) + \eta_i), \tag{1.111}$$

then, thanks to symmetry of its interaction (ultimately to convergence theorem in Markov processes guaranteed by Detailed Balance), its dynamics will end in an equilibrium configuration, which is described by the probability distribution $\mathcal{P}(\boldsymbol{\sigma}) \sim \exp(-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}))$ with

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\sum_{i,j<i}^{N} \Big(\frac{1}{N} \sum_{\mu}^{P} \xi_i^\mu \xi_j^\mu \Big) \sigma_i \sigma_j. \tag{1.112}$$

The whole thermodynamical properties of Hopfield neural networks are therefore completely determined and derived starting from this Hamiltonian (or cost function in neural network jargon).

1.4.3 Stored patterns as attractors

As we said, the basic principle lying behind the functionality of Hopfield networks as associative memory prototype is that stored patterns are associated to system configurations which are attractors for the network dynamics. To make it simple, the situation is the following. Once the P pattern are stored according to the Hebb rule, the system

[1] Here, we set the thresholds for firing $h_i = 0$ since we want to deal only with spontaneous magnetization properties.

should associate the input with the corresponding stored pattern. However, in general the presented input is affected by some external (and not removable) noise, or it is an imperfect realization of the corresponding pattern. Because of the noise, it is easy to understand that an associative memory could not work by comparing each bit in the input with those of all possible stored patterns. There should be a dynamics (internal to the network) finding out the nearest pattern associated to the prescribed input. This motivates the attracting character of stored patterns. If the system receive a (sufficiently low) noisy input, then - by autonomous dynamics - the network is able to reconstruct the pattern we want to be retrieved. This is the *pattern recognition* or *reconstruction* capability of Hopfield model.

In the theory of dynamical systems, the concept of attractor can be introduced in various ways. The definition we will use requires a metric characterization of the phase space. To fulfill this requirement, one should endow the configuration space of the Hopfield network with the Hamming distance:

Definition 1.15. Given two network configurations $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, the Hamming distance is defined as

$$d(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \frac{1}{2N} \sum_{i=1}^{N} |\sigma_{1,i} - \sigma_{2,i}|. \tag{1.113}$$

Remark 1.11. It is easy to show that this definition clearly fulfils all the requirements for a distance. Moreover, when the network size is large, it is possible to define the concept of arbitrarily near configurations. This makes the concept of neighbourhood mathematically well-defined (at least in the thermodynamic limit).

Then, by looking at the previous discussion about pattern recognition, we can introduce the concept of attractor with the following [42]

Definition 1.16. Given a dynamical system whose dynamics is parametrized by a (continuous or discrete) time t and a dynamical function T_t,[1] a set A of the phase space is attracting if it has a neighbourhood $U \neq \emptyset$ (called the *attraction basin*) such that

- For every neighbourhood V of A, then $T_t(U) \subset V$ for sufficiently large t;
- It is dynamically invariant, i.e. $T_t(A) = A$ for all t.

To go deeper in the characterization of stored patterns as attractors for the network dynamics, let us write the Hamiltonian as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) \sim -\frac{1}{2} \sum_{i,j=1}^{N} \Big(\frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \Big) \sigma_i \sigma_j = -\frac{N}{2} \sum_{\mu=1}^{P} m_\mu^2, \tag{1.114}$$

where we used the symbol $\sim$ as "apart for a $\mathcal{O}(1/N)$" error. Now, let us randomly extract configuration $\boldsymbol{\sigma}$ which is uncorrelated to the patterns for all $\mu = 1, \dots, P$. This means that each term in the sum are boolean variables with probability $\mathcal{P}(\xi_i^\mu \sigma_i = \pm 1) = 1/2$. Then, the evaluation of the associated Mattis magnetizations is equivalent to the computation of the displacement in a one-dimensional random walk. Since the net displacement has zero mean (because of the independence of random steps), one should estimate the Mattis magnetization with the square root of its variance, meaning that

$$m_\mu \sim \sqrt{\mathbb{E} m_\mu^2} = \sqrt{\frac{1}{N^2} \sum_{ij} \mathbb{E} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} = \frac{1}{\sqrt{N}}, \tag{1.115}$$

[1] Here, the notation T stands for the "transfer" map, which is endowed with semi-group properties: $T_0 = \mathbb{I}$ and $T_t \cdot T_s = T_{t+s}$, where in the case under consideration $t \in \mathbb{Z}_+$.

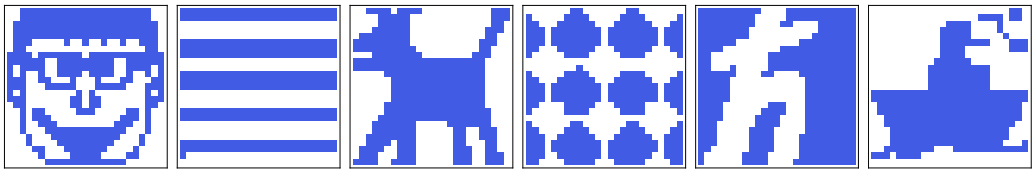

Figure 1.2: **Set of $P = 6$ patterns stored in a Hopfield network of $N = 625$ spins.** Patterns are black and while images: the network is dealing with digital storage of information [8].

since $\mathbb{E}(\xi_i^\mu \xi_j^\mu \sigma_i \sigma_j) = \delta_{ij}$, with $\mathbb{E}$ being the average of the random walk. Then, we can evaluate the Hamiltonian for network configurations which are uncorrelated to all the patterns as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu=1}^{P} m_\mu^2 \sim \mathcal{O}(1), \tag{1.116}$$

provided that the number of patterns P is finite. On the other hand, let us assume now that the network configuration is strongly correlated to a stored pattern (say for example $\boldsymbol{\sigma} = \xi^1$) and uncorrelated to all the others, meaning that $m_1 = 1$ and $m_\mu \sim N^{-1/2}$ for $\mu \geq 2$. Then, the Hamiltonian can be estimated as

$$H_N(\xi^1|\boldsymbol{\xi}) \simeq -\frac{N}{2} + \mathcal{O}(1). \tag{1.117}$$

Then, configurations aligned to the patterns are very convenient from an energetic point of view, with their stability growing with the network size. Moreover, they are the most stable configurations, since $0 \leq |m_\mu| \leq 1$. This implies that (if the number of stored patterns is finite), such configurations are *global* minima for the energy. Now, since the Hamiltonian is a Lyapunov function for the network dynamics (meaning that its temporal derivative is always non-negative, and vanishing at the equilibrium points), as a consequence they are fixed point, and the network evolves towards such configurations: they are attractors for the network dynamics.

An example of attractive power of stored patterns is reported in Fig. 1.3. Here, we consider a Hopfield network consisting in $N = 625$ spins in which we stored the set of $P = 6$ patterns reported in Fig. 1.2, organized in square lattices of 25×25 size. According to the previous discussion, such configurations are associated to attractors for the network dynamics, meaning that, if the network is prepared sufficiently near to a given pattern (i.e. in its attraction basin), then the network dynamics will end in a fixed point coincident with that pattern. To verify this statement, we initially prepared the network aligned to the first pattern (the smiling face), then we flip each spin with probability 0.2 (which means that we have a 20% noise level in the presented input). In the first row of Fig. 1.3, it is resumed the recognition of the first pattern for different evolution time steps starting from a noisy initial condition. In particular, we see that at $t = 1800$ the original pattern is almost reconstructed. In the plot below in the same figure, we see the time evolution of the Mattis magnetizations. The order parameter m_1 starts from an initial value ~ 0.6, and - as time flows - it approach the value 1, while all the other Mattis magnetization are always close to zero. What we discussed so far could lead to an optimistic overestimation of the associative power of Hopfield model. Indeed, by simple performances/processing resources arguments, one could be tempted to store more and more patterns for a given network size. However, as we already said, Hopfield networks behave very well for $P < 0.051N$ (and

moderately well for $P < 0.138N$).[1] The reason behind this limitations are however clear to researchers working in the field, and it is two-fold. First of all, the energetic arguments presented above are strictly true for a *finite* number of patterns for given N. On the other side, when the number of patterns is extensive in N (meaning that $P = \alpha N$), they are no longer valid, so a detailed analysis of equilibrium statistical mechanics of Hopfield model is needed (and this will be the subject of the following Sections). Furthermore, we said that such configurations are *global* minima for the energy function. However, it is not excluded that others fixed point arises when applying the Hebb learning rule. Indeed, this turns out to be the case, also if the information stored is low ($P/N \ll 1$). These additional minima have no counterpart in terms of stored patterns, so they are traditionally called *spurious* fixed points. An example of spurious attractor is given by the configuration

$$\tilde{\xi} = \text{sign}(\xi^1 + \xi^2 + \xi^3). \tag{1.118}$$

The Achille's heel of Hopfield network is that the number of such configurations grows very fast with the number of stored patterns (indeed, the growth is exponential in P, to be compared to the linear abundance of pure fixed points). From the dynamical point of view, this is suddenly a tragedy, since it means that, storing more patterns, the probability for the network dynamics to be trapped in the attraction basins of spurious states gets higher and higher. As a consequence, the attracting power of pure fixed points is dramatically downsized. A pictorial representation of this situation is reported in Fig. 1.5.

[1] Again, we stress that it is valid for a huge number of neurons in the network.

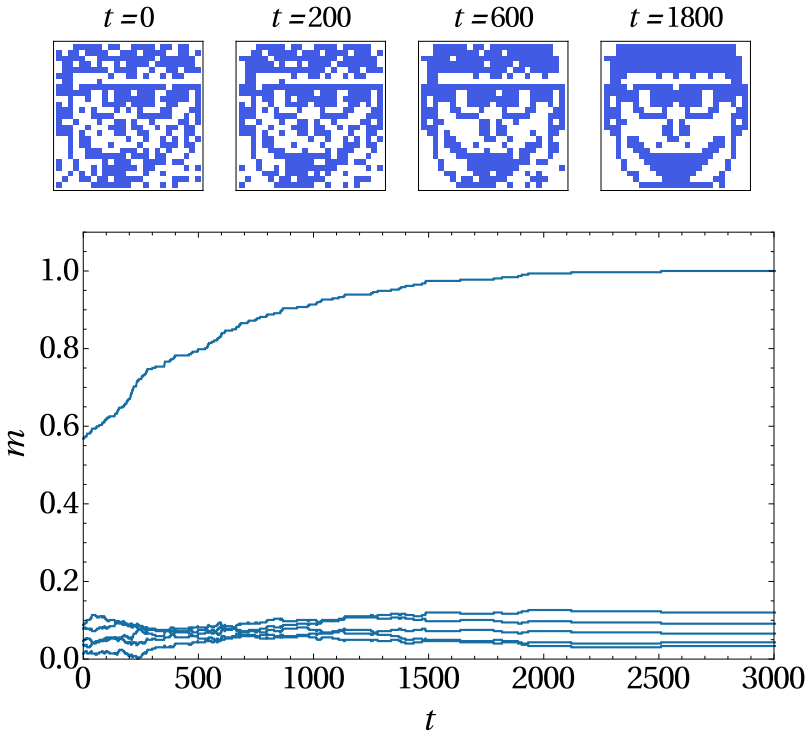

Figure 1.3: **Example of pattern reconstruction in a Hopfield network of $N = 625$ spins that stored $P = 6$ patterns.** Starting with a corrupted information, the Hopfield network is able to retrieve the associated pattern. We observe that, among the six Mattis magnetizations dedicated to quantify the retrieval of the six stored patterns, just one out of them grows up to one and its corresponding pattern is indeed retrieved by the network.

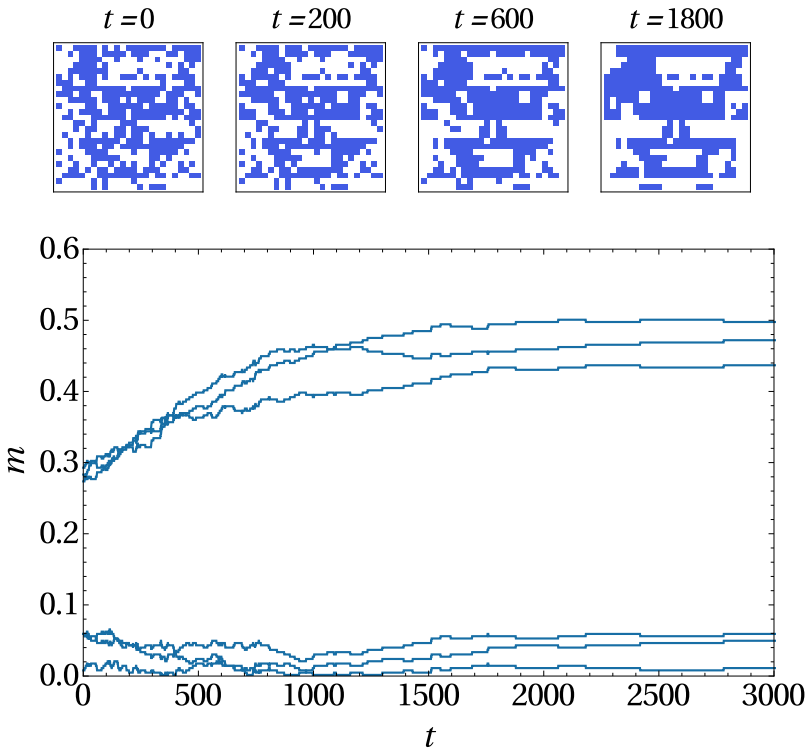

Figure 1.4: **Example of dynamics ending in a spurious state in a Hopfield network of $N = 625$ spins that stored $P = 6$ patterns.** In this example, it is possible to observe that several (three) Mattis magnetization raise sensibly over the noise due to the finite size effects and, correspondingly, the network has not been able to properly retrieve a single pattern, rather obtaining a useless mixture of the stored patterns.

An example of dynamics ending in spurious configurations is reported in Fig. 1.4. In this case, we prepared the network in the spurious configuration (1.118), then we flip again each spin with probability 0.2 and let the network evolve for a sufficient long time. In the first row, we see that the system reaches a configuration which is not in the stored patterns set, and which is indeed a fixed point since all of the order parameters settle on constant values (the Mattis magnetizations with highest equilibrium values are those associated to the first three patterns used to build up the spurious configuration). At this point, it is strongly needed a more careful understanding of pure and spurious fixed points for the network dynamics. This is possible with the so-called *signal/noise* analysis.

1.4.4 Signal-to-noise for Hebbian Storing

To get started with this analysis, we need to go back the Hamiltonian (1.103). By preparing the system near a given pattern, say ξ^1, we can express it as (again including self-interactions)

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{i,j=1}^{N}\xi_i^1\xi_j^1\sigma_i\sigma_j - \frac{1}{2N}\sum_{\mu\geq 2}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j. \tag{1.119}$$

It is clear that the first term tends to align the network configuration with the first pattern, and can therefore be interpreted as a signal contribution. On the other hand, since in general interactions are frustrated, the second term has the effect to destroy the correlation

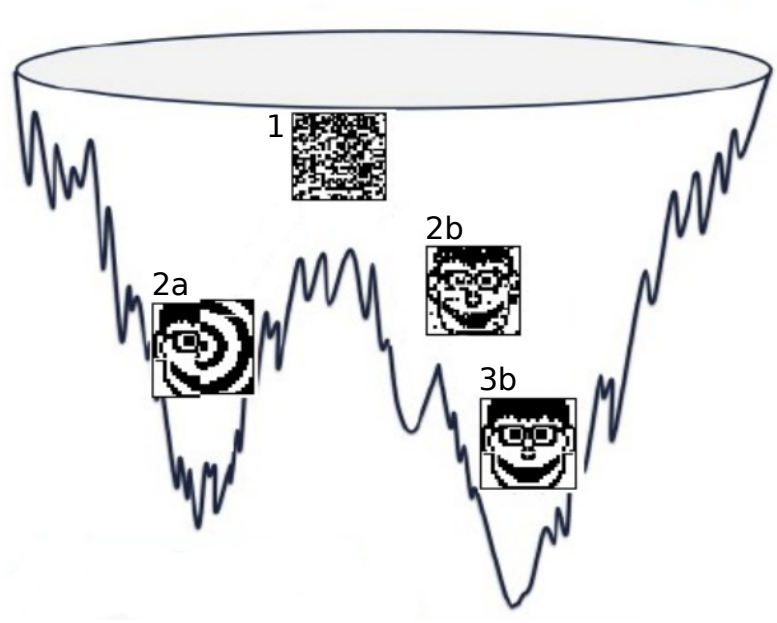

Figure 1.5: **Pictorial representation of minima landscape for the Hopfield model.** Starting with a noisy initial condition (1), the Hopfield network succeeds if the internal dynamics ends in a pure state configuration (with the evolution $1 \to 2\text{b} \to 3\text{b}$). However, the network could end in a metastable state 2a, therefore failing to retrieve the desired pattern.

of the configuration $\boldsymbol{\sigma}$ and the first pattern. Therefore, it can be interpreted as an intrinsic noise contribution. Thus, the goal of signal/noise analysis is to establish under which conditions a given network configuration is stable with respect to the *intrinsic* noise (in doing this, external thermal noise is set to zero: $\beta \to \infty$). The condition for a given configuration to be dynamically stable is

$$h_i \sigma_i \geq 0 \quad \text{for each } i, \tag{1.120}$$

where $h_i = \sum_{j \neq i} J_{ij} \sigma_j = \frac{1}{N} \sum_{j \neq i} \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu} \sigma_j$ is the internal field acting on the i-th neuron.

First of all, we would like to analyze the stability of pure attractors, so we set $\boldsymbol{\sigma} = \xi^1$. In this case, we have

$$h_1 \xi_1^1 = \frac{1}{N} \sum_{j>1} \sum_{\mu} \xi_1^{\mu} \xi_j^{\mu} \xi_j^1 \xi_1^1 = \frac{N-1}{N} + \frac{1}{N} \sum_{j>1} \sum_{\mu>1} \xi_1^{\mu} \xi_j^{\mu} \xi_j^1 \xi_1^1, \tag{1.121}$$

where we separated the signal and the noise contributions and used the dichotomic nature of the patterns. The same analysis can be carried out for all the other spins i. Clearly, the former term is, in the thermodynamic limit, equal to 1. On the other hand, the noise term is a sum of $(N-1)(P-1) \simeq N(P-1)$ variables taking values ± 1 with equal probability.[1] Therefore, the noise term is a random walk of $N(P-1)$ unitary steps. With this observation, we can evaluated the displacement of the random walk with the square root of the variance, which leads to

$$\left| \frac{1}{N} \sum_{j>1} \sum_{\mu>1} \xi_1^{\mu} \xi_j^{\mu} \xi_j^1 \xi_1^1 \right| \sim \sqrt{\frac{P-1}{N}}. \tag{1.122}$$

[1] This fact holds since each bit of different patterns at the same site i and of the same pattern μ at different sites are uncorrelated.

From this simple computations, we arrive to an important conclusion: the pure attractor configurations are stable (i.e. the intrinsic noise of the network is negligible) provided that $P \ll N$ (this also holds in the thermodynamic limit). This is no longer that the high storage regime ($P = \alpha N$), which thus requires a separate analysis. A similar results holds also if we flip a fraction d of the spins in the initial configuration, giving $h_i\sigma_i \sim 1 - 2d + \text{noise}$. In the low storage regime, the noise is still of order $N^{-1/2}$, then the system will quickly align to the pattern in order to increase the signal term (i.e. lower the energy), ending therefore in the pure attractor. This implies that pure attractors have a large attraction basins for $P \ll N$.

A similar analysis can be carried out also for spurious attractors, but a little more cumbersome since they are particular combinations of the stored patterns. To illustrate this point, let us consider the 3-symmetric mixture configuration (1.118). Without loss of generality, we can consider only a single spin $i = 1$ and fix $\xi_1^1 = 1$, so we have four possibilities corresponding $\xi_1^{2,3} = \pm 1$. Among these, only three would give $\sigma_1 = 1$, therefore we have $\mathcal{P}(\sigma_1 = 1) = 3/4$ (recall that patterns are supposed to be uncorrelated). Thus, in general

$$\mathcal{P}(\sigma_1 = \xi_1^\mu) = \frac{3}{4}, \quad \mathcal{P}(\sigma_1 = -\xi_1^\mu) = \frac{1}{4} \quad \text{for} \quad \mu = 1, 2, 3. \tag{1.123}$$

This implies that, in the thermodynamic limit, we have $3N/4$ spins aligned with each of the $\mu = 1, 2, 3$ pattern and $N/4$ with opposite orientation. Then

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i = \frac{1}{N}\Big(\frac{3N}{4} - \frac{N}{4}\Big) = \frac{1}{2}, \quad \mu = 1, 2, 3, \tag{1.124}$$

while $m_\mu \sim \mathcal{O}(N^{-1/2})$ for $\mu > 3$. This result should be compared with the numerical results reported in Fig. 1.4. The stability of the spurious configuration in this case is given by

$$h_1\sigma_1 = \sum_\mu m_\mu \xi_1^\mu \sigma_1 = \sigma_1(m_1\xi_1^1 + m_2\xi_1^2 + m_3\xi_1^3 + \sum_{\mu>3} m_\mu \xi_1^\mu). \tag{1.125}$$

Again, we have a signal contribution (given by the explicit terms in brackets) and a noise term (the sum over $\mu > 3$). For the former, we have

$$\text{Signal} = 0.5\left(\xi_1^1 + \xi_1^2 + \xi_1^3\right) \text{sgn}\left(\xi_1^1 + \xi_1^2 + \xi_1^3\right) = 0.5|\xi_i^1 + \xi_i^2 + \xi_i^3|. \tag{1.126}$$

The lowest value of the signal is 0.5 (corresponding to the case in which two of the bits have the same orientation while the other has opposite sign). Clearly, spurious attractors have a lower signal contribution with respect to the pure ones, making smaller the relative attraction basins (despite they are still large, as can be seen again from Fig. 1.4). However, in order for the initial state to be in the attraction basin of these particular 3-mixture states, the former has to present a large overlap with all the three patterns rather than a single one (which is possible only if the patterns are strongly correlated or when they are high in number). Concerning the intrinsic noise term, it is again a one-dimensional random walk with $N(P-3)$ values. Therefore, with the same arguments as above, it is evaluated to be of the order of $\sqrt{(P-3)/N}$, with the same conclusions as before.

Of course, spurious attractors can have more intricated structure, given by combination of all possible subsets of the patterns. If we consider combinations of the form $\tilde{\xi}_n \sim \sum_{\mu=1}^{n} \xi^\mu$, the taxonomy of the associated energies do respect the following classification [11]

$$E_1 < E_3 < E_5 < \cdots < E_\infty < \ldots E_4 < E_2. \tag{1.127}$$

1.4.5 High storage of Boolean patterns

It is time to turn to the complex case, which is the high storage limit $P = \alpha N$ with $\alpha \in \mathbb{R}^+$. Before focusing on the explicit expression of the quenched free energy for the Hopfield model in the high load regime, let us stress a little detail on the energy function, rewriting it as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{i,j<i}^{P}\sum_{\mu=1}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{1}{2N}\sum_{ij\mu}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \frac{P}{2}. \tag{1.128}$$

In the high storage case, also the last term is of order $\mathcal{O}(N)$ and contributes to the free energy. However, this contribution is constant and equals $\alpha/2$, so we can forget about it during the calculations (thus including also self-interactions during the calculations) and then correcting the obtained expression at the end by reintroducing this term. Of course, the definitions (1.114) hold also in this case, so we avoid to repeat them here. The only difference is that, here, we make explicit the dependent on the storage capacity α (previously, it was not needed because $\alpha = 0$ in the low storage regime).

Rather, we would like to stress an important point on methodology we will use in the following. Since we are interested in the retrieval regime, in which at least one pattern (as usual, we suppose it is ξ^1) is candidate to be retrieved, we will separate a ξ^1-dependent signal term from all the other $P-1$ contributions by the not-retrieved patterns accounting for the genesis of the intrinsic slow noise in the network. As a consequence, we should not average over all possible pattern realizations, but only on those contributing to the internal noise: in other words, we should consider (taking into account the self-interactions correction) the quenched free energy

$$f(\beta,\alpha) = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}'\log Z_N(\beta,\alpha) + \frac{\alpha}{2}, \tag{1.129}$$

where the average over quenched disorder is

$$\mathbb{E}' \equiv \mathbb{E}_{\xi^2}\dots\mathbb{E}_{\xi^P}. \tag{1.130}$$

Thus, in the replica trick approach (where the logarithm of the partition function is represented as a limit of zero replica of the replicated partition function) the relevant quantity is $\mathbb{E}'Z_N^n(\beta,\alpha)$. Introducing the replica index a running over different equivalent realization of the same system, we can write it as

$$\begin{aligned}\mathbb{E}'Z_N^n(\beta,\alpha) &= \mathbb{E}'\sum_{\boldsymbol{\sigma}^{(1)}}\dots\sum_{\boldsymbol{\sigma}^{(n)}}\exp\Big(\frac{\beta}{2N}\sum_{ija\mu}\xi_i^\mu\xi_j^\mu\sigma_i^{(a)}\sigma_j^{(a)}\Big) = \\ &= \mathbb{E}'\sum_{\boldsymbol{\sigma}^{(1)}}\dots\sum_{\boldsymbol{\sigma}^{(n)}}\int\Big(\prod_{a\mu}d\mu(z_\mu^{(a)})\Big)\exp\Big(\sqrt{\frac{\beta}{N}}\sum_{i\mu a}\xi_i^\mu\sigma_i^{(a)}z_\mu^{(a)}\Big),\end{aligned} \tag{1.131}$$

where in the last line we linearized the spin-dependence by using a Gaussian representation of the partition function. Here, we have of course

$$\int d\mu(z) = \int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}\exp(-z^2/2). \tag{1.132}$$

Since the average over the quenched disorder only involves not-retrieved patterns, we can split the replicated Boltzmann factor in two distinct factors, incorporating respectively the signal and the intrinsic noise. Thus, we can write

$$\mathbb{E}'Z_N^n(\beta,\alpha) = \sum_{\boldsymbol{\sigma}^{(1)}}\dots\sum_{\boldsymbol{\sigma}^{(n)}}z_{\text{signal}}[\boldsymbol{\sigma}]z_{\text{noise}}[\boldsymbol{\sigma}], \tag{1.133}$$

where

$$\begin{aligned} z_{\text{signal}}[\boldsymbol{\sigma}] &= \int \Big(\prod_{a=1}^{n} d\mu(z_1^{(a)}) \Big) \exp\Big(\sqrt{\frac{\beta}{N}} \sum_{ia} \xi_i^1 \sigma_i^{(a)} z_1^{(a)} \Big), \\ z_{\text{noise}}[\boldsymbol{\sigma}] &= \int \Big(\prod_{a,\mu\geq 2} d\mu(z_\mu^{(a)}) \Big) \mathbb{E}' \exp\Big(\sqrt{\frac{\beta}{N}} \sum_{ia,\mu\geq 2} \xi_i^\mu \sigma_i^{(a)} z_\mu^{(a)} \Big). \end{aligned} \tag{1.134}$$

The signal contribution is easy to handle with, so we start by considering the noise factor. On the latter, we can easily perform the average over not-retrieved patterns. This produces a $\log\cosh(\sqrt{\beta/N}\sum_a \sigma_i^{(a)} z_\mu^{(a)})$ in the exponential. The argument of this function is a quantity of order $\mathcal{O}(N^{-1/2})$, since the sum involves only the replica index, so we can therefore expand the function at the leading order. After some trivial rearrangements, the whole noise factor can be therefore rewritten as

$$z_{\text{noise}}[\boldsymbol{\sigma}] = \prod_{\mu\geq 2} \int \Big(\prod_a d\mu(z_\mu^{(a)}) \Big) \exp\Big(\frac{\beta}{2N} \sum_{iab} \sigma_i^{(a)} \sigma_i^{(b)} z_\mu^{(a)} z_\mu^{(b)} \Big). \tag{1.135}$$

The crucial point in this expression is that the argument of the exponential accounts for two kind of overlaps: the first one $\sim \sum_i \sigma_i^{(a)} \sigma_i^{(b)}$ is the overlap of different spin replicas; the second one $\sim \sum_\mu z_\mu^{(a)} z_\mu^{(b)}$ is an analogous quantity for replicas of the *hidden variables* z_μ (to use a Machine Learning jargon). We can therefore introduce these overlaps directly into the partition function by insertion of multiple Dirac deltas, therefore obtaining

$$\begin{aligned} z_{\text{noise}}[\boldsymbol{\sigma}] = \prod_{\mu\geq 2} \int \Big(\prod_{a=1}^{n} d\mu(z_\mu^{(a)}) \Big) \Big(\prod_{ab} dQ_{ab} \delta(Q_{ab} - \tfrac{1}{N} \sum_i \sigma_i^{(a)} \sigma_i^{(b)}) \Big) \cdot \\ \cdot \exp\Big(\frac{\beta}{2N} \sum_{ab} Q_{ab} z_\mu^{(a)} z_\mu^{(b)} \Big). \end{aligned} \tag{1.136}$$

The integral over the z variables is Gaussian, so we can easily evaluate it. Using the Fourier representation of the Dirac deltas, we finally found the following form for the noise term:[1]

$$\begin{aligned} z_{\text{noise}}[\boldsymbol{\sigma}] = \int \Big(\prod_{ab} dQ_{ab} \frac{N dP_{ab}}{2\pi} \Big) \exp\Big(iN \sum_{ab} P_{ab} Q_{ab} - i \sum_{iab} P_{ab} \sigma_i^{(a)} \sigma_i^{(b)} \\ - \frac{P}{2} \log\det(\mathbf{1} - \beta \boldsymbol{Q}) \Big). \end{aligned} \tag{1.137}$$

where $\mathbf{1}$ and $\boldsymbol{Q}$ are respectively the $n \times n$ identity and overlap matrices. Again, we note here that - as in the SK case - there are no couplings between spins belonging to the same replicas, so that we can reintroduce new spin variables $s_a = \pm 1$ with $a = 1, \ldots, n$. This allows to further simplify the expression. Including the singal term, with some manipulations we arrive (after some trivial rescalings $z_1^{(a)} \to \sqrt{\beta N} m_1^{(a)}$, $P_{ab} \to i\frac{\alpha\beta^2}{2} P_{ab}$) at the final result

$$\mathbb{E}' Z_N^n(\beta, \alpha) = \int d\mu(\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}) \exp(-NA[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}]), \tag{1.138}$$

[1] Note that, to be precise, since we have $P-1$ integration variables z, the prefactor of the last term should be $P-1$. However, since we want to deal with the high storage limit, the difference between P and $P-1$ is negligible in the thermodynamic limit.

where

$$\begin{aligned} A[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}] = \frac{\beta}{2}\sum_a (m_1^{(a)})^2 + \frac{\alpha\beta^2}{2}\sum_{ab} P_{ab}Q_{ab} + \frac{\alpha}{2}\log\det(\mathbf{1} - \beta\boldsymbol{Q}) \\ - \mathbb{E}\log\sum_{\boldsymbol{s}} \exp\Big(\beta\sum_a \xi^1 m_1^{(a)} s_a + \frac{\alpha\beta^2}{2}\sum_{ab} P_{ab}s_a s_b\Big), \end{aligned} \tag{1.139}$$

and $d\mu(\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P})$ is the measure over the order parameters (apart for constant factors, it is simply given by the Euclidean measure). Of course, the free energy of Hopfield model is recovered by taking the limit

$$f(\beta, \alpha) = \lim_{n\to 0} \frac{1}{\beta n} A[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}]. \tag{1.140}$$

At this point, we can no longer proceed without assuming a precise form for the overlap order parameters.

The replica symmetric solution

In the Hopfield model, the RS *Ansatz* is realized by taking the value of the Mattis magnetization independent on the replica realization. On the other side, the overlap are suppose to have equal non-diagonal elements. Moreover, we set the diagonal entries of the $\boldsymbol{Q}$ matrix equal to 1 (meaning that each replica has maximal overlap with itself), while for the $\boldsymbol{P}$ overlap we can set it to zero.[1] In mathematical terms, this leads to the choice

$$\begin{aligned} m_1^{(a)} &= m_1 \quad \forall a, \\ Q_{ab} &= \delta_{ab} + q(1 - \delta_{ab}), \\ P_{ab} &= p(1 - \delta_{ab}). \end{aligned} \tag{1.141}$$

Therefore, we are left only with three order parameters. With this *Ansatz*, it is possible to compute the replica symmetric free energy $f_{RS}(\beta, \alpha)$. Although the first terms in $A[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}]$ are actually easy to evaluate in the $n \to 0$ (and we refer to [8] to an exhaustive description), we stress that the last one (involving the quenched averaged $\mathbb{E}$) can be estimated as

$$-\frac{\alpha\beta^2}{2}np + n\,\mathbb{E}\int d\mu(z)\log 2\cosh(\beta m_1\xi^1 + \beta z\sqrt{\alpha p}) + \mathcal{O}(n^2). \tag{1.142}$$

Putting everything together and including the correction term $\alpha/(2\beta)$ as prescribed above, we are finally able to state the following [43]

Theorem 1.7. *The replica symmetric free energy for the Hopfield model in the high storage regime is*

$$\begin{aligned} f_{RS}(\beta, \alpha) = \frac{m_1^2}{2} + \frac{\alpha\beta}{2}p(1 - q) + \frac{\alpha}{2\beta}\Big(\beta + \log[1 - \beta(1 - q)] - \frac{q\beta}{1 - \beta(1 - q)}\Big) \\ - \frac{1}{\beta}\int d\mu(z)\log 2\cosh\Big(\beta m_1 + \beta z\frac{\sqrt{\alpha q}}{1 - \beta(1 - q)}\Big), \end{aligned} \tag{1.143}$$

[1] In general, one can choose to set the diagonal entries of the $\boldsymbol{P}$ equal to a fixed value p_D. However, it is possible to show that, under the RS assumption, when extremizing the free energy such an order parameter is not dynamical (meaning that its self-consistency equation is trivial), so one can consistently set it to 0.

where the order parameters satisfy the self-consistency equations

$$\begin{aligned} m_1 &= \int_{-\infty}^{+\infty} d\mu(z) \tanh\Big(\beta m_1 + \beta z \frac{\sqrt{\alpha q}}{1-\beta(1-q)}\Big), \\ q &= \int_{-\infty}^{+\infty} d\mu(z) \tanh^2\Big(\beta m_1 + \beta z \frac{\sqrt{\alpha q}}{1-\beta(1-q)}\Big). \end{aligned} \tag{1.144}$$

at the equilibrium states.

Remark 1.12. We highlight here two points. First of all, the self-consistency equation for the overlap p is algebraic, so it can be easily eliminated on the saddle point when evaluating the free energy. Therefore, we are left only with two order parameters satisfying coupled integral equations. The second point is that it was possible to directly evaluate the quenched average $\mathbb{E}$ since we assumed from the beginning that we are working with only one pattern ξ^1 candidate to be retrieved. In this way, because of the invariance of Gaussian measure under parity transformation and since the function $\log\cosh$ is even, we can trivially compute the quenched average. The extension of this equations to the case of l condensed patterns ξ^μ (with $\mu \in (1, \dots, l)$) is

$$\begin{aligned} m_\mu &= \int_{-\infty}^{+\infty} d\mu(z)\, \mathbb{E}\, \xi^\mu \tanh\Big(\beta \boldsymbol{m}\cdot\boldsymbol{\xi} + \beta z \frac{\sqrt{\alpha q}}{1-\beta(1-q)}\Big), \\ q &= \int_{-\infty}^{+\infty} d\mu(z)\, \mathbb{E} \tanh^2\Big(\beta \boldsymbol{m}\cdot\boldsymbol{\xi} + \beta z \frac{\sqrt{\alpha q}}{1-\beta(1-q)}\Big). \end{aligned} \tag{1.145}$$

Remark 1.13. Solving the Hopfield model in the high storage case beyond the replica symmetric assumption is a very hard task. At present time, the best knowledge we have about it stops at the 2RSB step [44]. However, it has been shown that the modification to the value of the critical capacity due to the replica symmetry breaking is negligible to a first approximation (we refer to [8, 45] for further details).

Of pivotal importante for a mature development of Theoretical Artificial Intelligence, and as the main reward in approching neural networks by the statistical mechanical perspective lies the concept of phase diagram, see Figure 1.6: the knowledge of the phase diagram of the network allows optimal setting of the system *a priori* -before any training or retrieval is tried (for instance it is pointless using the Hopfield network loaded at $\alpha \sim 0.5$ as, whetever the noise level in the network, its collective computational capabilities are lost for such a strong load). This is eventually one of the two the main rewards in the usage of this approach to Theoretical Artificial Intelligence because it allows for an *Optimized AI*, the other main reward lying in XAI *eXplainable AI* as, as we will deepen in the next two Chapters, by tracing clear bridges between artificial and biological information processing, *cracking the black box* (i.e. explaining the behavior of neural networks) ultimately results comprhensibe thanks to these parallelisms.

1.5 Generalities on the restricted Boltzmann machine

The restricted Boltzmann machine (RBM) is a two-layer network, where one layer -referred to as visible- receives input data from the outside world, while the other layer -referred to as hidden- is dedicated to figure out correlations in these input data (see Figure 1.7). Typically, a set of $M > 0$ data vectors $\{\sigma_1, \sigma_2, ..., \sigma_M\}$ (i.e., the so-called *training set* [46]) is presented to the machine and, under the assumption that these data have been generated by the same probability distribution $Q(\sigma)$, the ultimate goal of the machine is

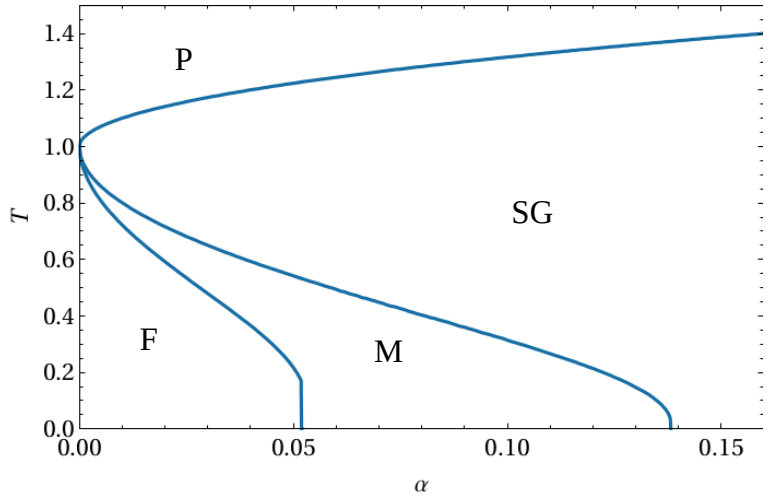

Figure 1.6: **Phase diagram of the Hopfield network.** As the noise ($T := 1/\beta$) in the network is tuned and the storage $\alpha = P/N$ varied, the network's behavior changes, being split into four macroscopic regions: a paramagnetic one (where $m_\mu = 0$ and $q = 0$), a spin glass region (where $m_\mu = 0$ and $q \sim 1$) and two regions where the network properly works as a pattern recognition associative memory, namely the two retrieval regions (one stable and one meta-stable): in both these regions $q \sim 1$ but also $m_\mu \sim 1$ (in the stable retrieval region the patterns are global free energy minma, while in the metastable retrieval region the patterns are just local free energy minima).

to make an inner representation of $Q(\sigma)$, say $P(\sigma|\xi, \theta)$, that is as close as possible to the original one. Clearly, in order to get a good representation, the more complicated $Q(\sigma)$, the larger the training set[1].
Each layer is composed by spins (also called neurons in this context), N for the visible layer and P for the hidden layer, and these spins can be chosen with high generality, ranging from discrete-valued (e.g., Ising spins), to real-valued (e.g., Gaussian spins). The thermodynamic limit of the ratio between the layer sizes, denoted as $\alpha = \lim_{N\to\infty} P/N$, is a control parameter[2] and usually one splits the case $\alpha = 0$ (possibly yielding to under-fitting) and the case $\alpha \in \mathbb{R}^+$ (possibly yielding to over-fitting) [46], the latter being mathematically much more challenging.
Analogously, the entries of the weight matrix can be either real or discrete. Generally speaking, continuous weights allows for learning rules (e.g., the contrastive divergence involving weight derivatives) which are more powerful than their discrete counterparts (the typical learning rule for binary weights is the Hebbian one [11]) and are therefore more convenient during the learning stage; on the other hand, binary weights are more performing in the so-called retrieval phase, that is, once the machine has learnt and is ready to perform the task it has been trained for. This trade-off gave rise to a number of variations on theme within the world of RBMs, where the extremal cases are probably given by a machine with binary (i.e., Boolean) versus real (i.e., Gaussian) weights, equipped with

[1] Actually, in order to optimize the training stage, one should also properly set the internal parameters of the machine such as the ratio between the sizes of the visible and hidden layer, the kind of the neurons, etc. [46, 47].

[2] It is not a case that the ratio between the layers is called α and that the two layer sizes are sharply P and N respectively: we will see along the manuscript that RBM and Hopfield networks are two archetypale faces (the former artificial, the latter biological) of the same coin that is *shallow information processing networks*.

a binary visible layer and a real hidden layer: in the present work we will focus on both these cases and we will try to highlight equivalences (but also crucial differences) among these extrema.

It is useful to summarize the mechanisms underlying the functioning of a standard RBM and, to this aim, we now introduce its definition.

Definition 1. *The Hamiltonian (or "cost function" in a machine learning jargon) of the restricted Boltzmann machine -equipped with a digital (Boolean) visible layer and an analog (Gaussian) hidden layer- reads as*

$$H_N(\sigma, z|\xi, \theta) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} \xi_i^\mu \sigma_i z_\mu - \sum_{i=1}^{N} \theta_i \sigma_i, \tag{1.146}$$

where σ_i ($i \in [1, ..., N]$) denotes the state of the i-th visible unit, z_μ ($\mu \in [1, ..., P]$) denotes the state of the μ-th hidden unit, ξ_i^μ denotes the weight associated to the link connecting the neurons labelled i and μ, and the factor $1/\sqrt{N}$ ensures the linear extensivity of the Hamiltonian with respect to the system volume.

The scalars θ_i ($i \in [1, ..., N]$) can be interpreted as external fields acting on the visible units and provide thresholds for neuron firing: given a certain internal field $\sum_\mu \xi_i^\mu z_\mu/\sqrt{N}$ over σ_i, the larger θ_i and the more likely for the i-th neuron to fire, namely to be in an active state $\sigma_i = +1$.
Now, this system is made to evolve by applying algorithms mimicking cognitive processes [39, 48]. More precisely, one splits *cognition* into two separate acts, namely distinguishing between *learning* (information) and *retrieval* (of the learnt information). The former occurs on a slower time scale and implies a synaptic dynamics which is modeled by properly rearranging the set of weights and thresholds. The latter occurs on a faster time scale and implies a neuronal dynamics which is modeled by properly rearranging the spin configuration, while keeping the weights quenched. Given the gap between the time scales characterizing these dynamical processes[1], one can treat them adiabatically, as done in the following subsections: the next one is dedicated to synaptic dynamics (i.e., rearrangement of the weights), while the successive one to neural dynamics (i.e., rearrangement of the spins).

1.5.1 A brief digression on slow variable's dynamics: learning

In this subsection we focus on the algorithms underlying the learning stage and which imply the dynamic of weights (we refer to [8] for a more extensive treatment). As mentioned in the beginning of Sec. 1.5, the goal is to obtain an inner representation $P(\sigma|\xi, \theta)$ which approximates $Q(\sigma)$; this is usually achieved by the minimization of the Kullback-Leibler cross entropy $D(Q, P)$, defined as

$$D(Q, P) = \sum_\sigma Q(\sigma) \ln \left[\frac{Q(\sigma)}{P(\sigma)}\right], \tag{1.147}$$

where the sum runs over all the possible configurations of the visible layer and we have dropped the dependence on the parameters (ξ, θ) of $P(\sigma|\xi, \theta)$ to lighten the notation. To the same purpose we also introduce $\tilde{\xi}_i^\mu \doteq \xi_i^\mu/\sqrt{N}$, $\forall i, \mu$. Notice that $D(Q, P)$ is minimal

[1] In the biological scenario the time scale for neuronal spikes is order of 50 ms, while the time scale for synaptic rearrangement is order of hours and it takes order of weeks to consolidate.

(and equal to zero) if and only if $P(\sigma)$ and $Q(\sigma)$ are identical. Now, by updating the weights and the thresholds by a gradient descent rule

$$\Delta\tilde{\xi}_i^\mu = -\epsilon\frac{\partial D(Q,P)}{\partial\tilde{\xi}_i^\mu}, \tag{1.148}$$

$$\Delta\theta_i = -\epsilon\frac{\partial D(Q,P)}{\partial\theta_i}, \tag{1.149}$$

where ϵ is a small parameter (also called learning rate), we get

$$\begin{aligned}\Delta D(Q,P) &= \sum_{i,\mu}\frac{\partial D(Q,P)}{\partial\tilde{\xi}_i^\mu}\Delta\tilde{\xi}_i^\mu + \sum_i\frac{\partial D(Q,P)}{\partial\theta_i}\Delta\theta_i \\ &= -\epsilon\left[\sum_{i,\mu}\left(\frac{\partial D(Q,P)}{\partial\tilde{\xi}_i^\mu}\right)^2 + \sum_i\left(\frac{\partial D(Q,P)}{\partial\theta_i}\right)^2\right] \leq 0,\end{aligned} \tag{1.150}$$

that is the cross-entropy $D(Q,P)$ decreases monotonically until a stationary state is reached (which, still, does not necessarily correspond to $D(Q,P)=0$). Now, in order to make this learning rule an explicit, operational, algorithm a bit of work is still necessary. A key point is that weights in the RBM are symmetric (i.e., its graph is undirected) and this, for (non-pathologic) stochastic dynamics, implies *detailed balance* which, in turn, ensures that the invariant measure is the Gibbs one given by

$$P(\sigma,z) = \frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\xi,\theta)}, \tag{1.151}$$

where $Z_{P,N}(\beta|\xi,\theta)$ is a normalization factor (or "partition function" in a Statistical Mechanics jargon [8, 11]) and $\beta\in\mathbb{R}^+$ encodes for the noise (in Physics β plays as an inverse temperature, in proper units). Now, marginalizing $P(\sigma,z)$ over the hidden layer z, we get $P(\sigma)$. Therefore, the internal representation of the probability distribution is formally known and this allows the construction of explicit learning algorithms, among which the *contrastive divergence* that we are going to derive is probably the most applied [46]. In order to proceed with the construction of a learning algorithm we explicitly define

$$Z_{P,N}(\beta|\xi,\theta) = \int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu)\sum_\sigma e^{-\beta H_N(\sigma,z|\xi,\theta)}, \tag{1.152}$$

$$Z_{P,N}(\beta|\sigma,\xi,\theta) = \int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu)e^{-\beta H_N(\sigma,z|\xi,\theta)}, \tag{1.153}$$

$$\begin{aligned}P(\sigma) &= \int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu)P(\sigma,z) = \int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu)\frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\xi,\theta)} \\ &= \frac{Z_{P,N}(\beta|\sigma,\xi,\theta)}{Z_{P,N}(\beta|\xi,\theta)}, \\ P(z|\sigma) &= \frac{P(\sigma,z)}{P(\sigma)} = \frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\sigma,\xi,\theta)},\end{aligned} \tag{1.154}$$

where, summations are meant over all possible spin configurations and $d\mu(z_\mu)$ is the Gaussian measure ($d\mu(z_\mu) = \exp(-z_\mu^2\beta/2)\sqrt{\beta/(2\pi)}$, for $\mu = 1,...,P$). Thus, $Z_{P,N}(\beta|\xi,\theta)$ is the partition function of a system where both variable sets are free to evolve, while

$Z_{P,N}(\beta|\sigma,\xi,\theta)$ is the partition function of a system where the visible layer is "clamped", namely forced to be in the configuration $\{\sigma\}$ encoded by one of the input data. Also, $P(\sigma)$ is the marginalized distribution and $P(z|\sigma)$ is the distribution for the configuration of the hidden layer being the visible layer clamped. At this point we have to evaluate each single term inside (1.150):

$$\begin{aligned}
\frac{\partial D(Q,P)}{\partial \tilde{\xi}_i^\mu} &= -\sum_\sigma Q(\sigma)\frac{\partial \ln P(\sigma)}{\partial \tilde{\xi}_i^\mu} \\
&= -\sum_\sigma Q(\sigma)\frac{\partial}{\partial \tilde{\xi}_i^\mu}\left(\ln Z_{P,N}(\beta|\sigma,\xi,\theta) - \ln Z_{P,N}(\beta|\xi,\theta)\right) \\
&= \beta\sum_\sigma Q(\sigma)\Bigg(\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z_\mu)P(z|\sigma,\xi,\theta)\frac{\partial H_N(\sigma,z|\xi,\theta)}{\partial \tilde{\xi}_i^\mu} \\
&\quad -\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z'_\mu)\sum_{\sigma'}P(z',\sigma'|\xi,\theta)\frac{\partial H_N(\sigma',z'|\xi,\theta)}{\partial \tilde{\xi}_i^\mu}\Bigg) \\
&= -\beta\Bigg(\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z_\mu)\sum_\sigma Q(\sigma)P(z|\sigma,\xi,\theta)\sigma_i z_\mu \\
&\quad -\sum_\sigma Q(\sigma)\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z'_\mu)\sum_{\sigma'}P(z',\sigma'|\xi,\theta)\sigma'_i z'_\mu\Bigg), \\
&= -\beta\left(\langle\sigma_i z_\mu\rangle_{clamped} - \langle\sigma_i z_\mu\rangle_{free}\right), \qquad (1.155)
\end{aligned}$$

where, in the first passage we used the definition (1.147), recalling that $Q(\sigma)$ does not depend on ξ; in the second passage we used (1.154); in the third passage we used (1.152) and (1.153); in the fourth passage we recalled that $\partial H_N(\sigma,z|\xi)/\partial\tilde{\xi}_i^\mu = -\sigma_i z_\mu$ and the subscript *clamped* means that the averages of the two-points correlation functions must be evaluated when the visible layer is forced to assume data values, while *free* means that the averages are the standard, statistical-mechanical ones. For the updating rule of the thresholds $\theta_i (i = 1, ..., N)$, one performs analogous calculations and, recalling $\partial H_N(\sigma,z|\xi)/\partial\theta_i = -\sigma_i$, one gets

$$\frac{\partial D(Q,P)}{\partial \theta_i} = -\beta\left(\langle\sigma_i\rangle_{clamped} - \langle\sigma_i\rangle_{free}\right). \qquad (1.156)$$

Thus, the learning rule (1.150) ultimately tries to make the theoretical one-point and two-point correlation functions as close as possible to the empirical ones[1]. Under this rule the machine will eventually be able to reproduce the statistics of the training data correctly, and this means that the parameters (ξ,θ) have been rearranged such that, if the machine is now asked to generate vectors with $P(\sigma)$, the statistical properties of these vectors will coincide with those of the input data generated by $Q(\sigma)$. In this case we say that the machine *has learnt* a representation of the reality it has been fed with. Note that this approach allows a proper statistical reproduction of mean averages and variances, hence, when $Q(\sigma)$ violates the central limit theorem, a two-layer RBM is no longer suitable for statistical inference and deep or dense networks are preferred.

[1] This argument can be expanded up to arbitrarily N-points correlation functions by paying the price of adding extra hidden layers and this kind of extension is a basic principle underlying Deep Learning [49].

1.5.2 A brief digression on fast variable's dynamics: retrieval

After the learning stage, the machine undergoes a final check over another bulk of data, referred to as *test set*, which stems from the same distribution that has generated the training set [46]. To fix ideas, let us assume that the machine was trained for retrieval tasks: if the trained machine is able to retrieve correctly the items in the test set, then the test is passed and the machine is ready for the usage. In order to move from the learning mode to the retrieval mode, the hidden layer is marginalized over: as we are going to show, following this procedure we end up with a Hopfield model (that is the standard model for pattern retrieval [39]), where each feature learnt by the hidden layer corresponds to one of the learnt patterns and the optimal parameters (ξ, θ) store information about the whole set of learnt patterns.

To see this duality between the RBM and the Hopfield model we look at the temporal evolution of the neurons which can be described by the following stochastic differential equation and map (to fix ideas we take hidden units as continuous and visible units as binary, as before)

$$\frac{dz_\mu(t)}{dt} = -z_\mu(t) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i^\mu\sigma_i + \sqrt{\frac{2}{\beta}}\eta_\mu(t), \tag{1.157}$$

$$\sigma_i(t) = \text{sign}\left[\tanh\left(\frac{\beta}{\sqrt{N}}\sum_{\mu=1}^{P}\xi_i^\mu z_\mu + \beta\theta_i\right) + \tilde{\eta}_i(t)\right]. \tag{1.158}$$

In the previous equation we used t to denote the time and we set the typical timescale of the variables (σ, z) as unitary; also, we denoted with η, $\tilde{\eta}$ standard Gaussian white noises with zero mean and covariance $\langle\eta_\mu(t)\eta_\nu(t')\rangle = \delta_{\mu\nu}\delta(t-t')$. Notice that, in the temporal evolution of the visible (respectively hidden) units, the hidden (respectively visible) units are taken as fixed (see also [8]).
Let us now focus on the hidden layer dynamics: the first term in the right-hand side of eq. (1.157) is the standard leakage term and the second term is the input signal over the hidden layer. This dynamics overall defines an Ornstein-Uhlembeck process, whose equilibrium distribution, at fixed σ's, reads as

$$P(z_\mu|\sigma) = \sqrt{\frac{\beta}{2\pi}}\exp\left[-\frac{\beta}{2}\left(z_\mu - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i^\mu\sigma_i\right)^2\right]. \tag{1.159}$$

Since the hidden units are independent in the RBMs under study, we can write $P(z|\sigma) = \prod_{\mu=1}^{P}P(z_\mu|\sigma)$.
As for the dynamics of the visible layer, each spin perceives an effective field (that is the sum of the overall signal and the threshold for firing) that is compared with the noise in such a way that if the signal prevails over the noise the neuron spikes. Hence, for the σ's, we can write

$$P(\sigma_i|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sigma_i\sum_\mu^P\xi_i^\mu z_\mu + \beta\theta_i\sigma_i}}{2\cosh\left(\beta\sum_\mu^P\xi_i^\mu z_\mu/\sqrt{N} + \beta\theta_i\right)}, \tag{1.160}$$

and, again, $P(\sigma|z) = \prod_{i=1}^{N}P(\sigma_i|z)$. In order to get the joint distribution $P(\sigma, z)$ and the marginal distributions $P(\sigma)$, we use Bayes' Theorem, i.e. $P(\sigma, z) = P(\sigma|z)P(z) =$

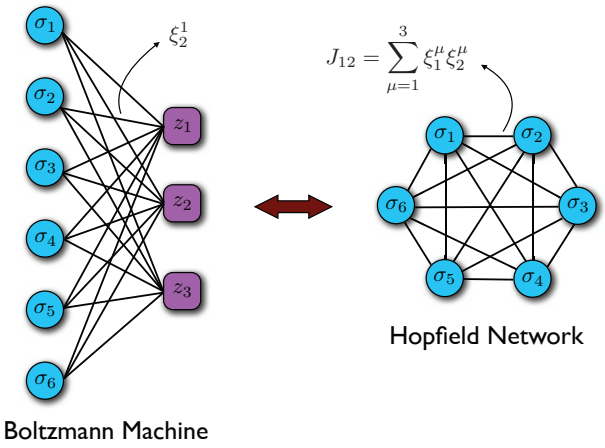

Figure 1.7: **Schematic representation of the (restricted) Boltzmann machine (left panel) and its corresponding dual, the Hopfield network.** Left panel: example of a RBM equipped with six neurons in the visible layer, $\sigma_1, ..., \sigma_6$ and three neurons in the hidden layer $z_1, ..., z_3$. The weights among the two layers are coded by the $N \times P$ matrix ξ_i^μ. Right panel: dual example of the corresponding Hopfield model, obtained by marginalization over the hidden variables. This network uses solely the $\sigma_1, ..., \sigma_6$ neurons, whose links however are now arranged according to the Hebb prescription for learning, that is $J_{ij} = \sum_{\mu=1}^{3} \xi_i^\mu \xi_j^\mu$.

$P(z|\sigma)P(\sigma)$, hence getting

$$P(\sigma, z) \propto \exp\left[-\frac{\beta}{2}\sum_{\mu=1}^{P} z_\mu^2 + \frac{\beta}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu \sigma_i z_\mu\right], \tag{1.161}$$

$$P(\sigma) \propto \exp\left[\frac{\beta}{2N}\sum_{i,j=1}^{N}\left(\sum_{\mu=1}^{P}\xi_i^\mu \xi_j^\mu\right)\sigma_i\sigma_j\right]. \tag{1.162}$$

Remarkably, one can see that the features learnt by the machine (see eq. (1.159)) correspond to the patterns that the machine will successively be able to retrieve (see eq. (1.162)), as this last equation is nothing but the Gibbs probability distribution for the original Hopfield model [11, 39].

Remark 1.14. We presented the RBM architectyre equipping the hidden layer with real-valued neurons with a Gaussian prior because, starting from this architecture, to prove the duality between RBMs and Hopfield networks we just need to integrate out already factorized Gaussian integrals, hence we preferred this choice to guarantee simplicitly in the calculations while briding the two worlds of biological and artificial information processing neworks, but the duality (between RBMs in general and generalized Hebbian networks) continues to hold whatever the details of the networks, it is just more tricky to prove without the Gaussian setting for the hidden neurons. In the next Chapter for instance, that is dedicated to our research findings in Theoretical Artificial Intelligence, we will use a fully binary RBM, where both the layers -visible and hidden- are equipped with digitial neurons (i.e. Ising spins).

We close this section pointing out that for *discrete* weights/patterns the contrastive divergence algorithm shown in the learning section can not be applied as it requires stochastic descent over the weights that must therefore be *real* (and differentiable). For discrete

pattern's entries, given the outlined duality between Restricted Boltzmann machines and Hopfield neural networks, the most natural learning rule would probably be the Hebbian learning from examples: calling J_{ij} the effective coupling between the two neurons σ_i and σ_j, given M examples -labelled with $a \in (1, ..., M)$- for any given μ of the P patterns $\xi^{\mu,a}$, namely P vectors of length N of binary entries, such a prescription results in

$$J_{ij} = \sum_{a=1}^{M} \sum_{\mu=1}^{P} \xi_i^{\mu,a} \xi_j^{\mu,a}, \tag{1.163}$$

that, formally at least, resembles the coupling emerging from the duality between RBMs and Hopfield networks, yet -as we have seen in this first Chapter- the Hopfield network does not learn from examples $\xi^{\mu,a}$, rather it stores already defined patterns ξ^{μ} hence we have to entirely prove our assertion, namely that the above learning rule is truly a learning rule and that the learning objects -inferred by the inspections of the examples $\xi^{\mu,a}$- are effectively the patterns ξ^{μ} (that in turn would coincide with the weights of the dual RBM representation): this will be the starting point of the new research results I present in this thesis as investigated in the next Chapter.

Chapter 2

Part 2: Theoretical Artificial Intelligence

From now on we report results from our research experience during the Ph.D. training time. As we assumed the reader to be familiar with Statistical Mechanics concept during the first Chapter, in this Chapter we shall assume that the reader has familiarity with Machine Learning concepts. The a-priori knowledge of the minimal data-set size to ensure a successful learning, is not yet known in general, despite the pivotal importance of such information en route toward optimized artificial intelligence. Given the duality between Boltzmann machines and Hopfield networks we presented in the closure of the last Chapter, yet, in this thesis we propose a novel approach to quantify the goodness of the learning stage that completely bypasses Bayesian posterior evaluation (i.e. the standard route in classical machine learning): this evaluation is again heavily biologically inspired (and ultimately suggested by the fact that learning and retrieval of information are two inseparable faces of a unique phenomenon that is *cognition*) and we can rely on this observation as follows: once a RBM has accomplished learning, the learnt information will shine when tested the machine in pattern recognitions problems (where its retrieval capabilities are guaranteed by its dual representation in terms of the Hopfield network) and -forcing the network in the retrieval region- if the Mattis magnetizations that we obtain different from zero are the expected ones we can conclude that the training stage has been succesfull, otherwise a plethora of possibilities can be evaluated to overcome the partial achieved learning (e.g. we can consider larger and less noisy data-sets or we can change the architecture of the network, relying on more sophisticated models as those that will be deepened in this research chapter): indeed in this chapter we try to contribute towards this goal and -for the sake of simplicity- we wil restrict to the simplest random data-sets scenario, where shallow networks suffice and a general theory can be worked out.
First, we prove that the supervised Boltzmann learning based on the grandmother cell setting mirrors unsupervised Hopfield learning: the grand-mother cell scenario [50, 51] was also originally introduced in a biological context (and adopted here to the machine learning counterpart) and it assumes that one single neuron (here one single hidden neuron in the hidden layer of the RBM) gets active when a pattern is presented to the network (here when a pattern is presented to the visible layer); while this theory was criticized in the biological context (as it presents flaws for structured datasets because multiple hidden neurons typically rise dealing with structured information), this simple setting naturally works here for structureless data-sets as all the patterns are equivalent under permutations and we can arbitrarily associate any of the patterns to be learnt to any of the hidden neurons, hence we have that the maximal storage capacity $\alpha_c \sim 0.14$ for the

Hopfield model sets also the maximal ratio between visible and hidden layer in machine learning before pushing the Boltzmann machine toward a glassy phase, where inference is still possible but more tricky (for instance decisional majority rules could be implemented). The unsupervised Hopfield learning generalizes the standard Hopfield model in the case where, instead of having a set of definite patterns (archetypes), only a sample of blurred versions are available and these are overall combined in a Hebbian kernel as proposed at the end of the previous Chapter.

2.1 Hebbian Learning: existence of a dataset threshold size

Indeed we closed the last Chapter questioning on the duality between Restricted Botzmann machines (RBM) and Hopfiled neural networks that, given the manifest rewards we achieve if properly established, must be faced now and in detail. Before starting to deepen such an equivalence, let us briefly comment on the rewards: as Machine Learning became pervasive in countless aspects of societal and working habits of our lifes, there is an urgent need in several worldwide research group's agenda toward both eXplainable AI (XAI) and Optimized AI (OAI). The former is due to *crack the black box* as *understading machine learning* is a central question in ethics concerning AI (if a self-driving car is in an *Aut-Aut* and it has to decide if invest an old man or a child we need to know how and why it took an option rather than the other) and the latter is pivotal for a safe massive usage of AI without contributing to global warming as, at present, machine learning algorithms are far from being optimized and, even worse, often we require the machine to accomplish learning that can not simply be achieved given the amount of information provided to the network or due to its inner architectural organization: in these regards, providing phase diagrams where different operational modes of the machines are presented as regions in this plot split by computational phase transitions, would allow us to set the machine in the optimal operational mode a priori, without high energy consumption empirical trials.
Hence, in this Chapter, at first (in this section) we show numerically that a RBM trained over a sample of blurred examples and a Hopfield model learning from the same sample of blurred examples are eventually (as the dataset gets large enough) able to generalize, namely, the former can be used as an archetype classifier/generator and the latter as an archetype retriever. Remarkably, we rigorously obtain a threshold $M_\times$ in the dataset size for the emergence of such a skill and this threshold turns out to be sharply the same for both models, yet -as a matter of presentation- in the following subsections we will address the two problems separately, starting with RBM and then moving to Hopfield networks.

2.1.1 RBM learning from blurred samples

We denote with $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\dots,K}$ the K archetypes, namely the patterns that we would like to see learnt by the RBM and with $\mathcal{S} = \{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$ the related examples that the machine is actually supplied with. These objects are codified in terms of binary vectors of length N; pattern entries are Rademacher random variables drawn with probability

$$\mathcal{P}(\xi_i^\mu = +1) = \mathcal{P}(\xi_i^\mu = -1) = 1/2, \quad \forall i, \mu \tag{2.1}$$

while example entries are defined, $\forall i, \mu, a$, as

$$\begin{aligned} \eta_i^{\mu,a} &= \xi_i^\mu \chi_i^{\mu,a}, \\ \text{with} \quad \mathcal{P}(\chi_i^{\mu,a} = 1) &= 1 - \mathcal{P}(\chi_i^{\mu,a} = -1) = p \in [1/2, 1], \end{aligned} \tag{2.2}$$

in such a way that the closer p gets to $1/2$, the farther from the pattern gets the example[1]. We also introduce

$$r := 2p - 1 \in [0, 1], \tag{2.3}$$

as an index for the dataset quality: r ranges from 0 (any example and the related archetype are uncorrelated) to 1 (any example coincides with the related archetype).

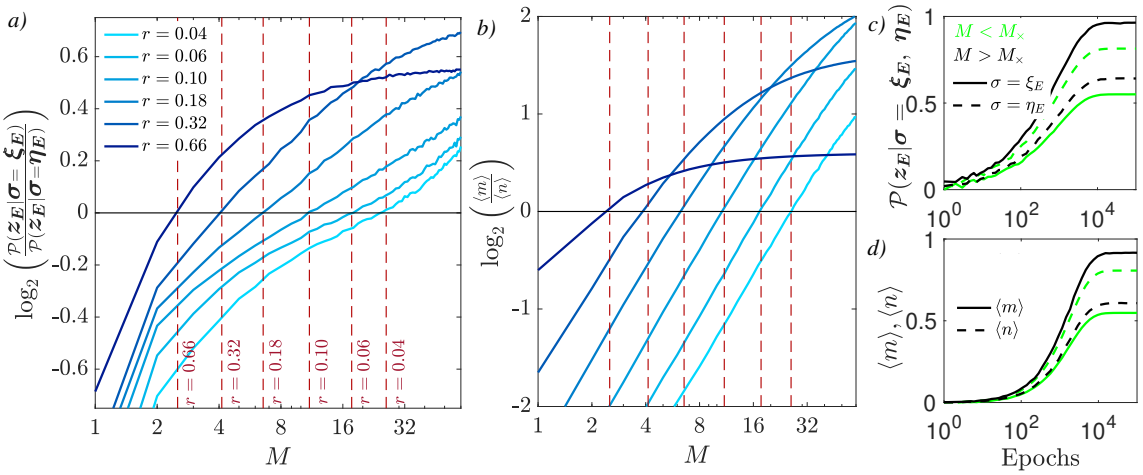

Figure 2.1: The larger panels provide a picture of the performance of a trained RBM used as a classifier (panel a) and as a generative model (panel b); in particular, the logarithm of $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)/\mathcal{P}(\boldsymbol{z}|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ and the logarithm of $\langle m\rangle/\langle n\rangle$, respectively, are shown versus M, for different choices of the parameters r (as explained by the common legend in panel a), which quantifies the dataset quality. The threshold value $M_\times$ corresponds to the interception between the curves and the horizontal axis. In both panels the vertical dashed lines are obtained analytically by studying the dual Hopfield network and asking for the minimum value of M such that archetype retrieval prevails over example retrieval (i.e., $\bar{m} > \bar{n}$, see Sec. 2.1.2 and SM); this estimate is obtained for different choices of r, as reported. Note that the vertical lines intersect the experimental curves always when they also cross the horizontal line, showing that the threshold size is the same for these machines. The smaller panels provide a picture of the training routine for the RBM. As epochs run, we show the evolution of the classification probabilities (panel c) $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ and $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ (respectively, solid and dashed lines) and of the overlaps (panel d) $\langle m\rangle$ and $\langle n\rangle$ (respectively solid and dashed lines), distinguishing between the case $M > M_\times$ (bright color) and $M > M_\times$ (dark color), as explained in the legend.

The machine is made of two layers: the visible one made of N binary neurons $\sigma_i = \pm 1$, $i \in (1, ..., N)$ and a hidden one built of by K binary neurons $z_\mu = \pm 1$, $\mu \in (1, ..., K)$; we denote with $(\boldsymbol{\sigma}, \boldsymbol{z}) \in \{-1, +1\}^{N\times K}$ the overall configuration. Note that there are as many hidden neurons as archetypes and, as we will explain in the following, this architecture allows us to allocate one hidden neuron per archetype, configuring the network in the *grandmother cell* scenario [50, 51].
We also introduce the weight matrix $\boldsymbol{w} \in \mathbb{R}^{N\times K}$, whose entry w_i^μ represents the weight associated to the connection between neurons i and j belonging to different layers. The cost function (or Hamiltonian to keep a physical jargon) $H_{N,K}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w})$ related to this RBM

[1] Although here we are working with random data-sets, for intuition guidance, we could look at a certain pattern $\boldsymbol{\xi}$ as the archetype of, say, a German Shepherd, while the set $\boldsymbol{\eta}^a$ would be a set of pictures of this dog, and similarly for the other patterns.

reads as

$$H_{N,K}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w}) = -\frac{1}{\sqrt{N}} \sum_{\mu=1}^{K} \sum_{i=1}^{N} w_i^\mu \sigma_i z_\mu, \tag{2.4}$$

where the factor $\sqrt{N}$ ensures the linear scaling of the cost function with respect to the size in the thermodynamic limit $N \to \infty$. The equilibrium distribution for such a system is given by the Boltzmann-Gibbs measure

$$\mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w}) = \frac{1}{Z_{N,K}} e^{-\beta H_{N,K}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w})}, \tag{2.5}$$

where $Z_{N,K}$ is the suitable normalization factor obtained by summing the exponential term over all possible configurations $(\boldsymbol{\sigma}, \boldsymbol{z}) \in \{-1, +1\}^{N \times K}$.

We train this machine in a supervised mode, that is, during the training phase the clamped setting involves both the visible and the hidden degrees of freedom, namely $\boldsymbol{\sigma}$ is set to one of the examples in the dataset, say the (ν, a)-th one, i.e., $\sigma_i = \eta_i^{\nu,a}$ for $i = 1, ..., N$, while $\boldsymbol{z}$ is set to a one-hot vector where the entry related to the correct archetype is 1 and the others 0, i.e., $z_\mu = \delta_{\mu,\nu}$ for $\mu = 1, ..., K$; we call $\mathcal{Z}$ the set of all possible K one-hot vectors. This kind of setting can be interpreted as a grandmother-cell setting, namely we establish a one-to-one correspondence between hidden neurons and archetypes and – in the clamped state – we force solely one hidden neuron per archetype to be active, whence the constraint on the number of archetypes equal to the number of hidden neurons.

More specifically, the machine training is accomplished by means of the following Hinton's scheme of contrastive divergence that we derived in Section 1.5.1 (see eq. 1.168):

$$\Delta w_i^\mu \propto \left(\langle \sigma_i z_\mu \rangle_{\text{clamped}} - \langle \sigma_i z_\mu \rangle_{\text{free}} \right), \quad \forall (i, \mu) \in (N \times K),$$

where for each training step the "free" average is sampled via a single step of *alternative Gibbs sampling*, i.e. a random training example $(\boldsymbol{\sigma}_E, \boldsymbol{z}_E) \in \mathcal{S} \times \mathcal{Z}$ is selected, then the free mean is calculated single shot via a pair $(\boldsymbol{\sigma}_{free}, \boldsymbol{z}_{free})$ sampled using the Gibbs-chain $\boldsymbol{z}_E \to \boldsymbol{\sigma}_{free} \to \boldsymbol{z}_{free}$; the "clamped" average is also evaluated single shot using the same pair $(\boldsymbol{\sigma}_E, \boldsymbol{z}_E)$.

The trained machine can be used as a classifier (i.e., as a pattern recognition device, by feeding the machine a noisy $\boldsymbol{\sigma}$ configuration and letting the machine recover the $\boldsymbol{z}$ configuration whose entries indicate how the input signal has been classified), or as a generative model (by feeding the machine a $\boldsymbol{z}$ configuration and letting the machine output the $\boldsymbol{\sigma}$ configuration of the corresponding archetype). We inspect the success of the learning procedure by testing the machine as a classifier and as a generative model, as reported in panels a and b of Fig. 2.1. More precisely, we choose as performance measure for classification the logarithm of $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)/\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$, where $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ (respectively $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$) is the probability of reaching a correct hidden state $\boldsymbol{z}_E$ given a visible state clamped as $\boldsymbol{\sigma} = \boldsymbol{\xi}_E$ (respectively $\boldsymbol{\sigma} = \boldsymbol{\eta}_E$); the ratio between the two terms allows us to assess when one prevails over the other (see Fig. 2.1, panel a). To evaluate computationally $\mathcal{P}(\boldsymbol{z}|\boldsymbol{\sigma} = \boldsymbol{\xi})$ (and analogously $\mathcal{P}(\boldsymbol{z}|\boldsymbol{\sigma} = \boldsymbol{\eta})$) we provide the network with, respectively, the archetype and the example on the visible layer and we study the distribution of activations within the hidden layer (i.e., the entries of the $\boldsymbol{z}$ vector): as training followed the grandmother-cell setting we expect to have just one positive entry – the hidden neuron coupled to the selected archetype – if learning has been properly accomplished.
When looking at the Boltzmann machine as a generative model, we use a different performance measure: having trained the machine as specified above, we clamp the hidden

layer on a certain one-hot vector $\boldsymbol{z}_E \in \mathcal{Z}$ and we let the machine thermalize allowing visible neurons to evolve freely; we expect that the system relaxes to configurations where $\boldsymbol{\sigma}$ corresponds to the related archetype $\boldsymbol{\xi}_E$. To check whether this is the case we measure the overlap between the visible neuron configuration $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}_E$ and compare it with the overlap between $\boldsymbol{\sigma}$ and the examples corresponding to the class of $\boldsymbol{\xi}_E$. To fix ideas, let us set $\boldsymbol{\xi}_E = \boldsymbol{\xi}^\nu$, then we introduce $n^{\nu,a}$ as the overlap between $\boldsymbol{\sigma}$ and the (ν, a)-th example $\boldsymbol{\eta}^{\nu,a}$ for $a = 1, ..., M$, namely

$$n^{\nu,a} := \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu \chi_i^{\nu,a} \sigma_i, \tag{2.6}$$

and m^ν as the overlap between $\boldsymbol{\sigma}$ and the ν-th archetype $\boldsymbol{\xi}^\nu$, namely

$$m^\nu := \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu \sigma_i. \tag{2.7}$$

To evaluate computationally these overlaps we first evaluate $n^{\nu,a}$ and m^ν as normalized dot product between the thermalized configuration $\boldsymbol{\sigma}$ and, respectively, $\boldsymbol{\eta}^{\nu,a}$ and $\boldsymbol{\xi}^\nu$ as per definitions (2.6) and (2.7), then we average over different choices of clamped states (namely by varying $\nu \in [1, K]$) and over different realizations of archetypes (this is the analogous of a quenched average). These mean values are denoted as $\langle n \rangle$ and $\langle m \rangle$. Their comparison allows us to evaluate whether the system is more prone to generate one of the examples it has been exposed to or to generate the unseen archetype (see Fig. 2.1 , panel b).

For both operational modes we see that, if the number of examples provided to the network is relatively small, the system fails, that is the system can classify examples better than archetypes (i.e., $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E) > \mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$) or the system generates examples rather than archetypes (i.e., $\langle n \rangle > \langle m \rangle$). However, if the number of examples is relatively large, the system succeeds, that is the system can classify archetypes better than examples or the system generates archetypes rather than examples. The threshold between a "small" and a "large" dataset is denoted with $M_\times$ and, as expected, $M_\times$ grows as the sample quality r decreases. Empirically, we find that $M_\times \sim r^{-1}$. On the other hand, the two extreme cases $M = 1$ and $M \to \infty$ are trivial as for $M = 1$ there is no difference between examples and archetypes (the example is also the archetype) while for $M \to \infty$ the archetype always prevails over examples by a standard central limit theorem argument.

Analogous remarks can be drawn also from Fig. 2.1 panels c, d where we show the evolution of the classification probabilities $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ and $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ and of the mean overlaps $\langle m \rangle$ and $\langle n \rangle$ as the training is running. Interestingly, as long as $M < M_\times$, the saturation values for $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ and $\langle n \rangle$ are larger than those obtained for $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ and $\langle m \rangle$; the opposite holds as $M > M_\times$.

We conclude this section recalling that we can recast the problem of archetypes generation and classification exploiting the duality between Boltzmann machines and Hopfield networks: as largely discussed in the past decade [52, 53, 54, 55, 56, 57, 58, 59], we have so far understood that by marginalizing the probability distribution $\mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w})$ over the hidden layer, we end up with the probability distribution of a Hopfield network as long as we identify the weights w_i^μ in the former with the entries of the patterns stored by the latter, and we suitably rescale the temperature; in formulae

$$\begin{aligned} \mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{w}) &= \sum_{\sigma}^{2^N} \sum_{z}^{2^K} e^{\frac{\beta}{\sqrt{N}} \sum_{\mu=1}^{K} \sum_{i=1}^{N} w_i^\mu \sigma_i z_\mu} \\ &\to \mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{w}) \propto \sum_{\sigma}^{2^N} e^{\frac{\beta^2}{2N} \sum_{i,j}^{N,N} \sum_{\mu}^{K} (w_i^\mu w_j^\mu) \sigma_i \sigma_j}. \end{aligned} \tag{2.8}$$

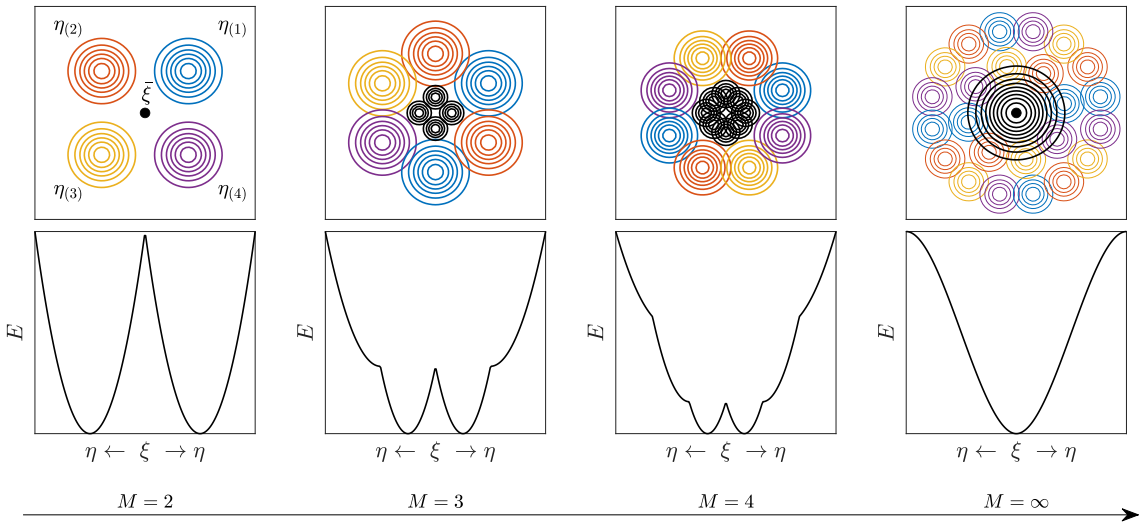

Figure 2.2: Schematic representation of the emergence of the archetype minimum in the energy landscape. The network is supplied with M examples of the pattern $\boldsymbol{\xi}$, which, instead, is never presented to the network. As M grows, from left to right, the network at first stores each single example but it is unable to retrieve $\boldsymbol{\xi}$ (left, $M < M_\times$), then, in the energy landscape, new minima, close to $\boldsymbol{\xi}$, appear and coexist with the minima corresponding to examples (center, $M > M_\times$) and, finally, a unique stable minimum corresponding to the archetype emerges (right, $M \approx M_c$).

From this perspective we may want to check the ability of the system to retrieve an archetype, namely if we initialize the Hopfield network in a configuration $\boldsymbol{\sigma}$ corresponding to an example, say $\boldsymbol{\eta}^{\nu,a}$, and let it thermalize towards equilibrium, does it eventually end up "close" to the archetype $\boldsymbol{\xi}^\nu$? This problem is faced in the next section.

2.1.2 Hopfield network learning from blurred samples

Let us consider a Hopfield neural network made of N binary neurons, whose overall configuration is denoted with $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$, and supplied with the sample $\mathcal{S}$ made of $K \times M$ examples $\eta_i^{\mu,a} = \xi_i^\mu \chi_i^{\mu,a}$ as defined in (2.2). We want to apply Hebb's rule to this sample and check whether the resulting system is able to generalize, namely to retrieve the archetype $\{\boldsymbol{\xi}^\mu\}_{\mu=1,...,K}$ once provided with an example[1]. We write the coupling J_{ij} between the neurons i and j as

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{K} \sum_{a=1}^{M} \eta_i^{\mu,a} \eta_j^{\mu,a}. \tag{2.9}$$

Notice that, in this definition, we are simply summing over all the instances making up the sample $\mathcal{S}$, without caring of the class each term belongs to, in this sense, this kind of Hebbian learning is *unsupervised*. The cost function (or Hamiltonian to keep a physical jargon) $H_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi})$ of the model reads as follows

$$H_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi}) = -\frac{1}{2N} \sum_{\mu,a}^{K,M} \Big(\sum_{i=1}^{N} \xi_i^\mu \chi_i^{\mu,a} \sigma_i \Big)^2 = -\frac{N}{2} \sum_{a=1}^{M} (\boldsymbol{n}^a)^2, \tag{2.10}$$

[1]When this can be accomplished we say that also the Hopfield network can generalize because, starting from the inferred archetype, it generates variations on theme by taking advantage of the fast noise β

where $\boldsymbol{n}^a = (n^{1,a}, ..., n^{K,a})$ with entries defined in (2.6). Analogously, we pose $\boldsymbol{m} = (m^1, ..., m^K)$ with entries defined in (2.7). In this context, we shall also refer to $\boldsymbol{n}^a$ and $\boldsymbol{m}$ as Mattis magnetizations related to, respectively, examples and archetypes.
As we will see, these quantities play as key order parameters to quantify how (and what kind of) pattern recognition is accomplished by the network, implicitly quantifying the goodness of its learning too. In the following we will denote with $\bar{m}$ and $\bar{n}$ their expectations with respect to the Boltzmann-Gibbs distribution related to the cost function (2.10), namely

$$\mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi}) = \frac{1}{Z_{N,K,M}} e^{-\frac{\beta}{2N} H_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi})}, \tag{2.11}$$

where $Z_{N,K,M}$ is the suitable normalization factor obtained by summing the exponential term over all possible configurations $\boldsymbol{\sigma} \in \{-1, +1\}^N$.

As detailed in the next sections, this model can be addressed analytically and we can obtain – in the thermodynamic limit and in the high-storage regime (i.e., $\alpha = \lim_{N\to\infty} K/N$ finite) – self-consistent equations for its order parameters that can be then solved numerically. Following this route, we can compare $\bar{n}$ and $\bar{m}$ and check whether $\bar{m} > \bar{n}$ finding that for this condition to hold, M must be larger than a certain threshold, represented by the vertical dashed lines in Fig. 2.1: remarkably, this threshold corresponds to the threshold value $M_\times$ of the RBM.

Before proceeding we anticipate that the analytical investigation performed on the Hopfield model (2.10) highlights a rich phenomenology that here we try to summarize by means of Fig. 2.2 that sketches the evolution of (a cross section of) the cost-function landscape $E := H_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi})$ as the dataset size is made larger; in this landscape, we especially care of minima since they play as attraction basins for the neural configuration $\boldsymbol{\sigma}$. When M is small the landscape exhibits $K \times M$ minima[1] corresponding to the examples provided; as M is made larger, minima get denser and their attraction basins possibly overlap; when $M > M_\times$ the minima corresponding to examples are only local while new and deeper minima emerge, whose location is closer to the archetype rather than any other example; as M is further increased local minima get less and less stable while global minima get closer and closer to the archetypes; finally when M is large enough, configurations corresponding to archetypes become stable. As we will see in the next section, the last passage can be related to a critical value for M, that scales with r and that we denote with M_c. Interestingly, we also find out that setting $M = M_c$ determines the onset of a critical phase transition.
Therefore, for the Hopfield network defined in (2.10), in addition to the traditional tuneable parameters, namely the fast noise β and the load $\alpha = \lim_{N\to\infty} K/N$, we have the sample size M and the sample quality r; we expect the system to correctly retrieve archetypes as long as $\alpha < \alpha_c(\beta)$ and as long as $M > M_c(r)$. When translating this knowledge into the RBM scenario, we derive restrictions in the data-dimensionality reduction ability of Boltzmann machines (note that α corresponds to the ratio between the sizes of the hidden and the visible layers in RBM) [60] and an interplay between dataset quality and quantity.

2.1.3 Signal-to-Noise for Hebbian Learning

As anticipated, our analytical investigations shall focus on the Hopfield counterpart for which we can rely on solid mathematical methods.

We start with the signal-to-noise analysis to check for local stability of the configurations $\boldsymbol{\sigma} = \boldsymbol{\eta}^{\mu,a}$ and $\boldsymbol{\sigma} = \boldsymbol{\xi}^\mu$, for arbitrary μ and a, in the noiseless limit $\beta \to \infty$. This is

[1] The number of minima is actually $2 \times K \times M$ due to the gauge symmetry.

accomplished by studying if the internal field $h_i = \sum_{j=1}^{N} J_{ij}\sigma_j$, experienced by the neuron i, is aligned with the neural activity σ_i and by monitoring the evolution of the relative energies associated to these configurations (see Figure 2). We find that by increasing M, archetypes (examples) progressively gain (loose) stability at a rate depending on r. In particular, as for archetypes, the stability threshold M_c increases according to the following scaling

$$M_c \sim (2p-1)^{-4}, \tag{2.12}$$

To get sharper estimates and a characterization of a possible phase transition, we need to solve for the quenched free-energy of the model and inspect the related self-consistent equations for order parameters. In the limit of infinite volume N, but finite dataset size M, the quenched pressure (i.e., $-\beta$ times the free energy) of the model (2.10) is defined as

$$A_M(\alpha,\beta) := \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\ln Z_{N,K,M}(\beta|\boldsymbol{\chi},\boldsymbol{\xi}), \tag{2.13}$$

where $\mathbb{E} := \mathbb{E}_\chi\mathbb{E}_\xi$ averages over both the quenched variables χ,ξ, and $Z_{N,K,M}$ is the partition function given by

$$Z_{N,K,M}(\beta|\boldsymbol{\chi},\boldsymbol{\xi}) := \sum_{\sigma}^{2^N} \exp\Big[\frac{\beta}{2N}\sum_{a=1}^{M}\sum_{\mu=1}^{K}\big(\sum_{i=1}^{N}\xi_i^\mu\chi_i^{\mu,a}\sigma_i\big)^2\Big]. \tag{2.14}$$

Note that by a trivial Hubbard-Stratonovich transformation, this partition function coincides with that of a RBM equipped with Gaussian prior. At the replica symmetric level of description, keeping M fixed but sending both K and N to infinity in such a way that α is finite, and focusing on the retrieval of $\boldsymbol{\xi}^1$ with no loss of generality, we reach the following expressions for the quenched statistical pressure

$$A = \log 2 - \frac{\beta\alpha M}{2}\bar{p}(1-\bar{q}) - \frac{\beta M}{2}\bar{n}^2 - \frac{\alpha M}{2}\big(\log[1-\beta(1-\bar{q})]$$
$$-\frac{\beta\bar{q}}{1-\beta(1-\bar{q})}\big) + \mathbb{E}_{\phi\chi}\log\cosh\big(\bar{n}\beta\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big),$$

where ϕ is an auxiliary, mute, Gaussian field, while $\bar{p}$, $\bar{q}$ and $\bar{n}$ are, respectively, the expectation values for the order parameters p_{12}, q_{12}, $n_{1,a}$, being p_{12} and q_{12} overlaps between different replicas of the system. These expectation values can be obtained by looking for the stationary points of the quenched pressure $\nabla_{\bar{n},\bar{q},\bar{p}}A_{N,M}(\alpha,\beta) = 0$ and turn out to fulfil the following self-consistent equations

$$\bar{n} = \mathbb{E}_{\phi\chi}\left(\frac{1}{M}\sum_{a=1}^{M}\chi_a\right)\tanh\big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big), \tag{2.15}$$

$$\bar{q} = \mathbb{E}_{\phi\chi}\tanh^2\big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big), \tag{2.16}$$

$$\bar{p} = \frac{\beta\bar{q}}{[1-\beta(1-\bar{q})]^2}. \tag{2.17}$$

Note that, the example magnetization n is embedded right in the expression of the model cost-function (see (2.10)), much as the Mattis magnetization for the standard AGS theory. On the other hand, m does not play as a natural observable for the model as the system

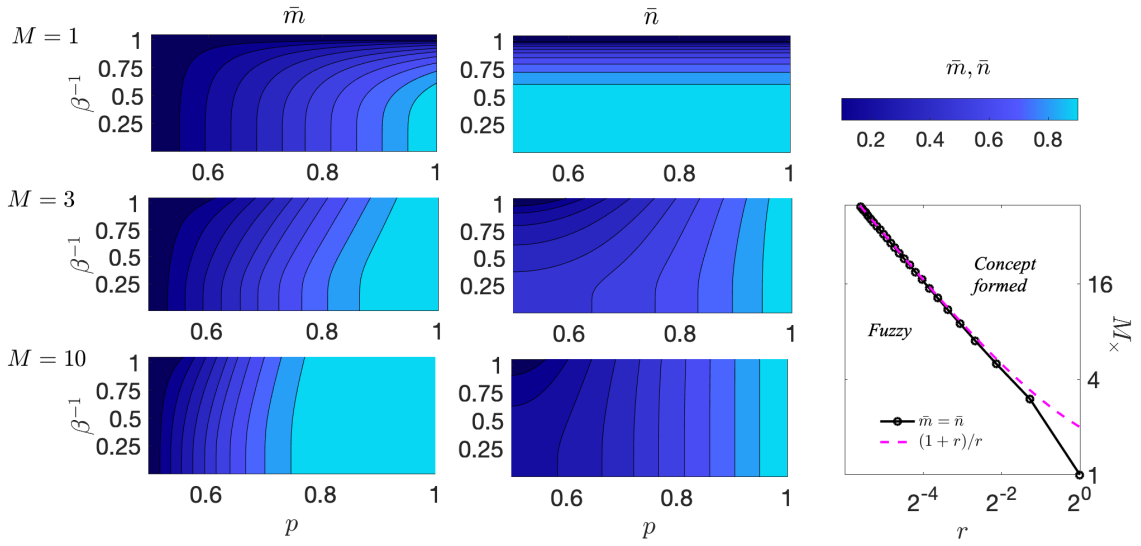

Figure 2.3: Left: Contour plots for the magnetization of the archetype $\bar{m}$ (left panels) and of the examples $\bar{n}$ (right panels), obtained by solving the self-consistencies in eqs. (2.15) and (2.18) for $\alpha = 0$ and for $M = 1, 3, 10$ (from top to bottom), versus p (x-axis) and β^{-1} (y-axis); analogous results are obtained for $\alpha > 0$, see the SM. By comparing the values of $\bar{m}$ and $\bar{n}$ we see that, as the number of examples exceeds a bound $M_\times(r)$, the archetype retrieval dominates over the example retrieval. Right: the condition $\bar{m} = \bar{n}$ is recognized as the boundary between a *fuzzy* regime where the sample size is not enough for the archetype to be inferred and the network is only able to retrieve the examples it has been presented to, and a *concept-formed* regime where the network "forgets" about the examples and can retrieve the archetype. Consistency between the theoretical (dashed line) and the empirical (bullets, solid line is a guide for eyes) estimates is provided. Notice that, at $\alpha = 0$, the function $M_\times(r)$ is temperature independent.

is, in principle, unaware of the archetypes. Having access to the archetypes, a practical way to compute $\bar{m}$ is to insert in the model a small field J coupled to m and then evaluate $\partial_J A_{N,M}(\alpha, \beta, J)$ as $J \to 0$. More interestingly, in the limit of large M, $\bar{m}$ spontaneously emerges and occurs to be directly related to $\bar{n}$; in particular, the two magnetizations, $\bar{m}$ and $\bar{n}$, get related as

$$\bar{n} = \frac{\bar{m} r}{1 - \beta(1 - \bar{q})(1 - r^2)}. \tag{2.18}$$

In the next subsections we analyze the self-consistent equations under different conditions and try to derive analytically the existence of a threshold size $M_\times$ and of a critical size M_c that determine the onset of different regimes as for the system ability to generalize.

Finite dataset size

Let us resume eqs. (2.15)-(2.17) and let us focus on the zero fast noise limit $\beta \to \infty$. Recalling that M is large, we can introduce the random variable $S := \frac{1}{M}\sum_{a=1}^{M} \chi_a = r + \sqrt{\frac{1-r^2}{M}} Z$ with $Z \sim \mathcal{N}(0,1)$, and, posing

$$\delta\bar{Q} = \mathbb{E}_Z \frac{2}{\sqrt{\pi}} \exp\left[-\left(\frac{\bar{n} M S(Z)}{\delta\bar{Q} + \sqrt{2\alpha M}} \right)^2 \right], \tag{2.19}$$

where $\mathbb{E}_Z$ denotes the expectation over Z, the self-consistency equations for the magnetizations $\bar{m}$ and $\bar{n}$ become

$$\bar{n} = \mathbb{E}_Z S(Z) \operatorname{erf}\left(\frac{\bar{n} M S(Z)}{\delta \bar{Q} + \sqrt{2\alpha M}}\right), \tag{2.20}$$

$$\bar{m} = \mathbb{E}_Z \operatorname{erf}\left(\frac{\bar{n} M S(Z)}{\delta \bar{Q} + \sqrt{2\alpha M}}\right). \tag{2.21}$$

Via these equations it is possible to obtain an analytic expression for the threshold $M_\times$: by requiring $\bar{m} > \bar{n}$, we obtain the following inequality

$$\mathbb{E}_Z \left[1 - S(Z)\right] \operatorname{erf}\left(\frac{\bar{n} M S(Z)}{\delta \bar{Q} + \sqrt{2\alpha M}}\right) > 0 \tag{2.22}$$

which, to first order in $\bar{n}$, is satisfied if $\mathbb{E}_Z \left[1 - S(Z)\right] S(Z) > 0$, namely, recalling the definition of S,

$$\mathbb{E}_Z \left[1 - r - \sqrt{\frac{1-r^2}{M}} Z\right] \left[r + \sqrt{\frac{1-r^2}{M}} Z\right] > 0, \tag{2.23}$$

whence $(1-r)r - \frac{1-r^2}{M} \mathbb{E}_Z Z^2 > 0$. The latter inequality yields to

$$M > \frac{1+r}{r} = M_\times. \tag{2.24}$$

Therefore, as expected, in order for the archetype magnetization to prevail over the example magnetization, the dataset size needs to be larger and larger as the sample gets more and more blurred, according to the above scaling. This finding is corroborated by extensive computational checks and its robustness with respect to the fast noise is also tested, as we solved numerically the self-consistency equations for arbitrary, finite β and derived an estimate of $M_\times$ by comparing the solutions of $\bar{m}$ and $\bar{n}$ obtaining analogous results as reported in Fig. 2.3.

Finally, we tested the validity of these results in the RBM framework, also exploring the robustness with respect to different loads. In the left panels of Fig. 2.4 we compare the classification probabilities $\mathcal{P}(\boldsymbol{z}_E | \boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ and $\mathcal{P}(\boldsymbol{z}_E | \boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ versus M and for different choices of r. The two probabilities display a monotonical behaviour as a function of M that is, respectively, increasing and decreasing. This can be intuitively explained invoking the central limit theorem and recalling Fig. 2.2: as M increases, minima in the energy landscape become denser and denser in such a way that the system may eventually fall into a state other than $\boldsymbol{\eta}$, and this gets more and more likely as the dataset quality r is lower. In the right panel of Fig. 2.4 the threshold values obtained for different loads α, analytically (i.e., investigating the Hopfield network, see Eq. (2.24)) and numerically (i.e., investigating the RBM) are shown to be perfectly consistent.

Infinite dataset size

Let us now retain a finite noise β and apply the rescaling of the noise $\beta \to \frac{\beta}{r^2 + \beta(1-q)(1-r^2)}$ to eqs. (2.15)-(2.17), thus, we reach expressions for the magnetization and the overlap whose content finally shines:

$$\bar{m} = \mathbb{E}_Z \tanh\left[\beta \bar{m} M + Z\beta \sqrt{M \frac{1-r^2}{r^2} \bar{m}^2 + \frac{\alpha \bar{p}}{r^4 \beta} M}\right] \tag{2.25}$$

$$\bar{q} = \mathbb{E}_Z \tanh^2\left[\beta \bar{m} M + Z\beta \sqrt{M \frac{1-r^2}{r^2} \bar{m}^2 + \frac{\alpha \bar{p}}{r^4 \beta} M}\right],$$

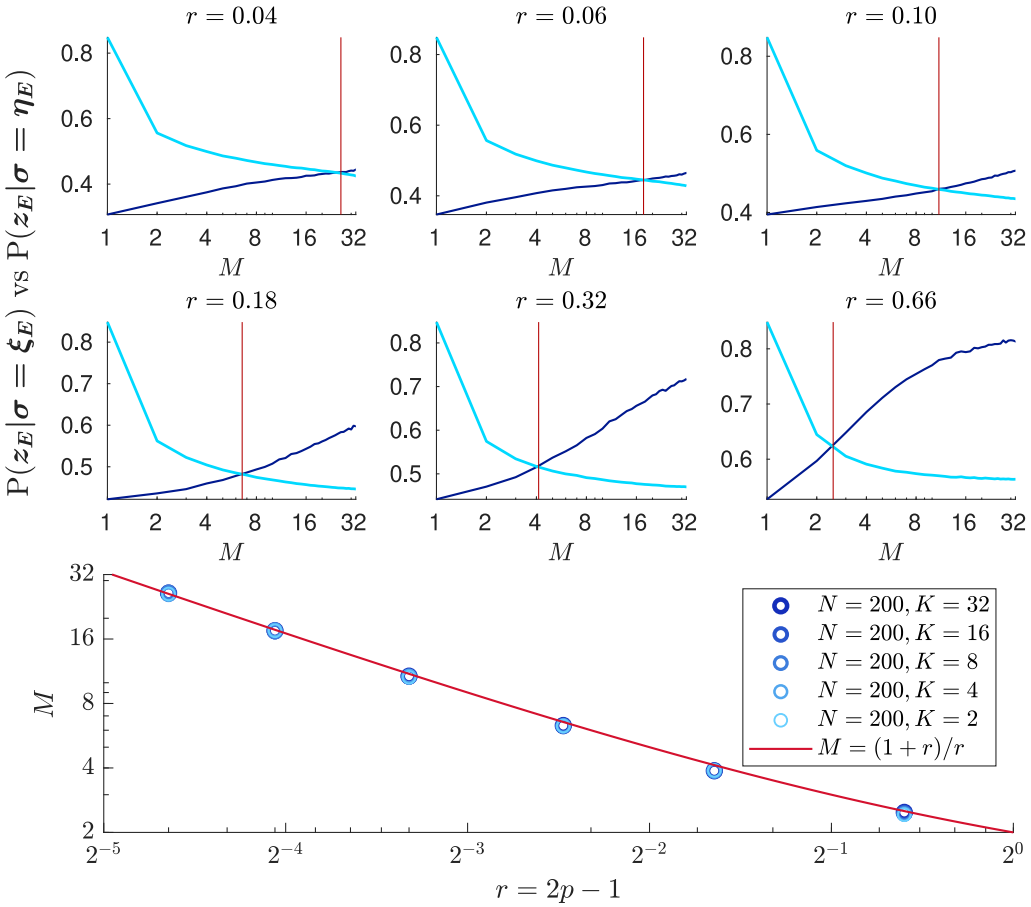

Figure 2.4: Each plot shows the probability of correctly classifying either an example or an archetype: blue lines are computational and drawn from Boltzmann machine learning (dark blue for the archetype, i.e. $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\xi}_E)$, light blue for the example, i.e. $\mathcal{P}(\boldsymbol{z}_E|\boldsymbol{\sigma} = \boldsymbol{\eta}_E)$) while red lines are theoretical and draft from Hopfield network learning. Different plots show different noise $r = 2p - 1$ levels and the vertical red line is evaluated via the relation $M_\times = (1 + r)/r$. Right: This phase diagram shows two regions split by the threshold line $M_\times = (1 + r)/r$ (where $\bar{m} = \bar{n}$) and above that threshold the concept of the archetype emerges (and the network can successfully generalizes) while below a fuzzy misture where all the examples still preserve their characteristics persists. The threshold is shown to be universal: different values of α are computationally simulated for Boltzmann learning and shown as spots (different blue circles), while the theoretical prediction by the Hopfield network is presented as a continuous red line and the two perfectly coincide as expected. On the vertical axes we report the critical size of the training set required for a successful learning while on the horizontal axes we report the degree of noise in the data-set.

where $Z \sim \mathcal{N}(0, 1)$ and $\bar{p}$ was given in (2.17). As arguments of the hyperbolic tangent there are now three contributions and no longer just two as in the standard AGS theory. Indeed, beyond the signal carried by $\bar{m}$ there are two sources of (slow) noise: a classic one, proportional to α, stemming from the other patterns not retrieved (pattern interference), and a new one stemming from the examples making up the sample related to the pattern (example interference). Note that, as consistency check, if the network is not provided with datasets, but just noiseless patterns (i.e. $M = 1$ and $r = 1$), the whole theory collapses over the standard AGS one of the Hopfield model as it should. Further we stress that at $\alpha = 0$ there

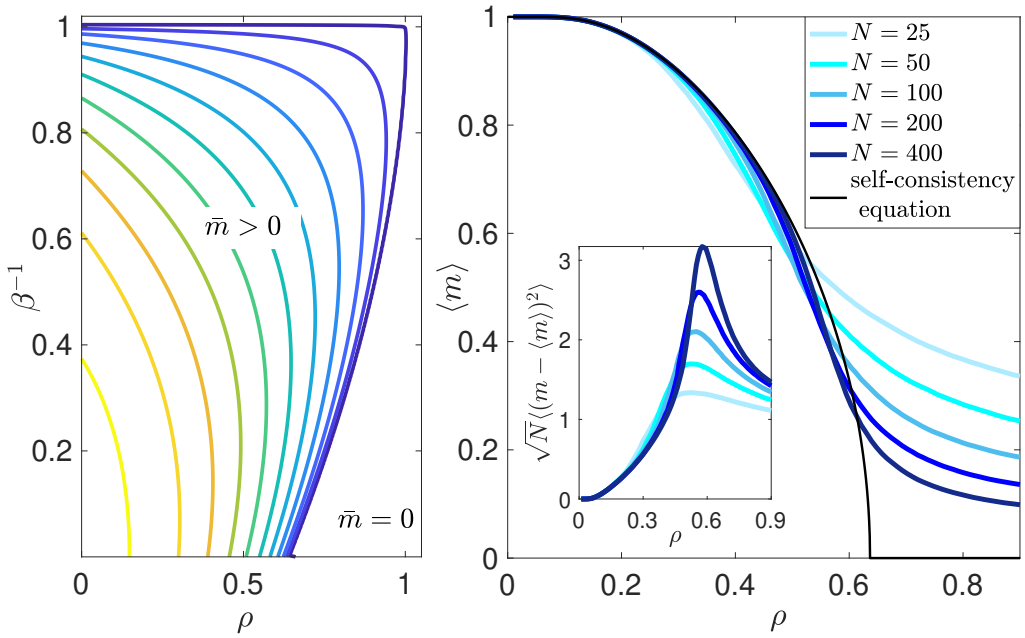

Figure 2.5: Left panel: Phase diagram in the (β, ρ) plane obtained by solving numerically equations (2.27)-(2.28). The outer, darkest line corresponds to the onset of a non-null magnetization $\bar{m} > 0$, the remaining contour lines, in brighter and brighter colors, correspond to larger and larger values of magnetization. Right panel: the main plot shows a comparison between the numerical solution of the self-consistency equation (2.27) in the noiseless limit (thin and darkest solid line), and zero-temperature Monte Carlo runs at different sizes (from brighter to darker nuances, $N = 25, 50, 100, 200, 400$, as shown by legend), while the inset shows the same finite-size-scaling for the susceptibility; in both figures, we set $\alpha = 0.08$, $M = 80$, and $r^2 = \sqrt{\alpha/(M\rho)}$, and, to determine each point, a quenched average over 50 independent coupling matrices was performed.

is not a real phase transition (as a glance at these self-consistencies reveal), rather we need $\alpha > 0$ (namely examples of different archetypes produce reciprocal attenuation of their retrieval, promoting as a result the emergence of the archetypes themselves). We now inspect in more details the self-consistency for $\bar{m}$ and we check when the signal contribution prevails over the noise, namely we require that $\beta M \bar{m} > \beta\sqrt{M}|Z|\sqrt{\bar{m}^2(1-r^2)/r^2 + \alpha\bar{p}/(r^4\beta)}$ holds almost surely. A solution to this inequality is given by

$$M > \frac{\gamma^2}{r^2}\left[1 - r^2 + \frac{\bar{q}}{\bar{m}^2(1-\beta(1-\bar{q}))^2}\frac{\alpha}{r^2}\right], \tag{2.26}$$

where γ assesses the confidence level (in fact, the last condition implies $|Z| < \gamma$ which can be satisfied up to an exceedingly small probability at finite M). Setting $\beta \to \infty$ this result recovers the scaling in (2.12)) obtained via signal-to-noise analysis. Therefore, a large enough database ensures the stability of the archetype.

In order to evidence a possible, genuine phase transition, we have to study the limit $(N, M, K) \to \infty, r \to 0$ and rephrase the whole theory intensive. In this limit we find that $\rho := \alpha/(Mr^4)$ is a suitable control parameter (ruling the overall slow noise) able to trigger

a phase transition and the self-consistency equations can be recast as

$$\bar{m} = \mathbb{E}_Z \tanh(\beta\bar{m} + \beta Z\sqrt{\rho\bar{q}}) \xrightarrow{\beta\to\infty} \mathrm{erf}\left(\frac{\bar{m}}{\sqrt{2\rho}}\right), \tag{2.27}$$

$$\bar{q} = \mathbb{E}_Z \tanh^2(\beta\bar{m} + \beta Z\sqrt{\rho\bar{q}}) \xrightarrow{\beta\to\infty} 1. \tag{2.28}$$

The numerical solution of eq. (2.27) is sketched in Figure 2.5 (left panel), where we highlight a region in the (β, ρ) plane where $\bar{m}$ is non null.

Focusing on the fast noiseless limit $\beta \to \infty$ and expanding at $\bar{m} = 0$, a critical behavior is found at $\rho_c = \frac{2}{\pi}$ with the critical exponent 1/2 (i.e., $\bar{m} \sim \sqrt{\frac{3}{\pi}}\sqrt{2 - \pi\rho}$ near the critical point): the concept is not abruptly formed, rather it stems gradually by a continuous contribution provided by all the examples.
The scenario painted above is corroborated by numerics: in Figure 2.5 (right panel) we plot a finite-size-scaling of the Mattis magnetization of the archetype, along with the related susceptibility, obtained via Monte Carlo simulations. Signatures of criticality just occur at $\rho \approx \rho_c$, according to the theory.
We remark that we dedicate the whole appendix to report in detail the whole statistical mechanical treatment of the network.

At this point we have a minimal satisfactory theory for the simpest Artificial Intelligence, namely shallow networks at work on random settings: indeed working with random datasets (to have a general theory, despite the limitation that this implies) shallow networks suffice (as there are no correlations functions longer than two-points to be inferred) hence Hopfield neural networks and restricted Boltzmann machines alone should be able to handle a minimal process of *cognition* (thought of as *learning from experience some information and then using the latters for some pourposes*) restricting to a world respecting what we have called the *statistical reductionism*. Indeed we have seen that machine learning (e.g. a RBM trained via contrastive divergece) produces information storage very similar to biological learning (e.g. an Hopfield network fed by examples via the generalization of the Hebbian kernel we proposed), further, once something has been learnt by the neural network (whatever it is, artificial or biological), later on the network can use the learnt experience to generalize toward other examples, to accomplish pattern recognition, etc.
A positive aspect of what we have so far achieved is indeed the resemblance between artificial and biological information processing (that is something somehow wanted, en route toward an eXaplainable Artificial Intelligence, XAI), yet the negative aspect is that the maximal trheshold these networks must respect are far from optimal (for instance, the critical storage is $\alpha_c \sim 0.14$, much less than the value 1 -for symmetric networks- or even 2 -for general non symmetric networks, that is an upper bound information-theory derived a long time ago by Elisabeth Gardner [67]). Persisting in a biologically-driven approach to Artificial Intelligence (to guarantee eXplainable AI) and by keeping the usage of statistical mechanics of complex systems (to guaranteed Optimized AI via the production of its related phase diagrams), in the next Section (focusing directly on archetypes rathen than examples for the sake of simplicity) we want to show that -by taking inspiration regarding sleeping mechanisms of mammals- it is possible to force the Hopfield network to sleep and in this way the network learns while on-line and optimizes while off-line thus reaching the upper critical storage of $\alpha_c \sim 1$: we called this properly *ultra-memory* for -at this point- obvious reason.
Finally, in the ultimate section of this Chapter devoted to Theoretical Artificial Intelligence, we will remove another strong limitation of the standard Hopfield reference, namely the requirement that the signal must prevail over the noise (or, at worst, they must have

the same intensity) in order for the pattern to be detected: again inspired by biological information processing, we will show that if we equip the RBM with two identical input layers (thought of as *eyes* of a human), they vehiculate redundant information to the cortical (i.e. hidden) neurons: this redundancy of representation (that mathematically -quite naturally- will pushes away from pairwise Hebbian coupling toward dense networks its modeling) allows the network to enjoy a tunable signal-to-noise threshold for pattern recognition, a phenomenon that we named *ultra-detection*, but let us start deepening the first extension, toward a by-far-enhanced storage capacity (i.e. *ultra-memory*) that we will obtain by forcing the Hopfield model *to take a nap*.

2.2 Neural networks equipped with Ultra-Memory

The idea beyond this *sleeping mechanism* is old and it has been driven in Neural Network's Literature by the works of Personnaz, Guyon, Dreyfus [61] and of Dotsenko et al. [62, 63] in the late 80s and early 90s. Yet, at that time, the importance of remotion of spure states was clear (and was the driving force of that research at that time when modeling was focused on stylized implementation of *random eye moviments* REM-like mechanisms) but the importance of the consolidation of pure memories was not yet understood (while in this manuscript we will consider also this aspect of sleep, by suitably schematizing *slow wave sleep* SWS-like mechanisms). From the *physical side*, the partial knowledge of the first wave of modeling sleeping mechanisms gave rise to tentative cost-functions whose phase diagrams resulted persistently in lackage of a stable retrieval phase, further, from the *mathematical counterpart*, the Guerra's interpolation technique appeared later in the Literature, at the beginning of the present century and with just heuristic methods it was much harder to deal with the models, hence the interest for the field of research dedicated to *unlearning protocols* diminished in the closure of the past century. In a nutshell, the core idea behind *unlearning* is to overcome the following limitation of the standard *always awake* Hopfield model: a linear increase in N of patterns to be stored (e.g. $P = \alpha N$) implies also an exponential proliferation of metastable spurious states (in other words, the price to pay to have these patterns as minima is a huge generation of spurious local minima or saddles that eventually push the network in the spin-glass state). However, if we *thermalize from an uncorrelated input at random in this landscape* -namely if we do a *quench* from high temperature to zero temperature at random- with high probability we end up in one of these spurious states -say $\sigma^{\mathbf{spurious}}$- and, once we know it, we can remove it by an unlearning rule, such as $J_{ij} \sim \sum_\mu^K \xi_i^\mu \xi_j^\mu - \langle \sigma_i^{spurious} \sigma_j^{spurious} \rangle$ (here the random quench that selects the spurious states to discard somehow mimicks the inconscious action of the random eye movement during sleep in mammals). Viceversa, if we luckily end up in a true pattern $\sigma^{\mathbf{pattern}}$ after the quenching, we aim at consolidating that memory generalizing the Hebbian scheme toward $J_{ij} \sim \sum_\mu^K \xi_i^\mu \xi_j^\mu + \langle sigma_i^{pattern} \sigma_j^{pattern} \rangle$: this feature was not implemented in past trial models and we called the algorithm of this sleeping neural network the *consolidation -of pure states- and remotion -of spurious states*. In [64] we introduced the following generalization of the standard Hopfield paradigma [39], referred to as "reinforcement&removal" (RR) algorithm, that was finally working, i.e. the critical threshold for storage shifts from $\alpha_c \sim 0.138$ to saturation $\alpha_c = 1$ while the network preserves robustness w.r.t. fast thermal noise, providing an enlarged retrieval region w.r.t. the always awake Hopfield reference framework.

Consider a network composed by N Ising neurons $\{\sigma_i\}_{i=1,...,N}$ and P patterns $\{\xi^\mu\}_{\mu=1,...,P}$ (i.e., random vectors of the same length N), and denote with $t \in \mathbb{R}^+$ the sleep extent (such that for $t = 0$ the network has never slept, while for $t \to \infty$ an entire sleeping session has occurred), we can then introduce the following

Definition 2. *The Hamiltonian of the reinforcement&removal model reads as:*[1]

$$H_{N,P}^{(RR)}(\sigma|\xi, t) := -\frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N \sum_{\mu=1}^P \sum_{\nu=1}^P \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{\mathbb{I} + tC} \right)_{\mu,\nu} \sigma_i \sigma_j, \tag{2.29}$$

where $\sigma_i = \pm 1 \ \forall i \in (1, ..., N)$, ξ^1 -that is the pattern candidate to be retrieved- has binary entries $\xi_i^1 \in \{-1, +1\}$ drawn from $P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = \frac{1}{2}$, while the remaining

[1] As a matter of notation, we stress that the denominator $1/(\mathbb{I}+tC)$ in the generalized kernel is intended as the inverse matrix $(\mathbb{I} + tC)^{-1}$.

$P-1$ patterns $\{\xi^\mu\}_{\mu=2,...,P}$, have i.i.d. standard Gaussian entries $\xi_i^\mu \sim \mathcal{N}[0,1]$, and the correlation matrix $\boldsymbol{C}$ is defined as

$$C_{\mu,\nu} := \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu.$$

remark 1. *We stress that we used the* universality property *of spin glasses for the sake of mathematical convenience. As deepened in [26], while we defined all the pattern entries to be digital (i.e. ± 1), during the calculations involved in the statistical mechanical treatment of the network, we keep digital solely the pattern candidate for retrieval (i.e. the* signal*), while all the remaining ones (acting as* slow noise *on the retrieval) are chosen as standard Gaussian $\mathcal{Z}[0,1]$: although neural networks, in general, do not exhibit the universality properties of spin glasses [65], this is no longer true if we confine our focus solely to the structure of the slow noise generated by not-retrieved patterns.*

remark 2. *Note that the matrix $\xi^T \left(\frac{1+t}{\mathbb{I}+tC}\right)\xi$, encoding the neuronal coupling, recovers the Hebbian kernel for $t=0$, while it approaches the pseudo-inverse matrix for $t \to \infty$ (see [64] for the proof). Accordingly, the model described by the Hamiltonian (2.29) spans, respectively, from the standard Hopfield model ($t \to 0$) to the Kanter-Sompolinksy model [66] ($t \to \infty$).*
During the sleeping session, both reinforcement and remotion take place: oversimplifying, in the generalized synaptic coupling appearing in (2.29), the denominator (i.e., the term $\propto (1+tC)^{-1}$) yields to the remotion of unwanted mixture states, while the numerator (i.e., the term $\propto 1+t$) reinforces the pure memories.

We are interested in obtaining the phase diagram of the model coded by the cost function (2.29), solely in the thermodynamic limit and under the replica symmetric assumption. To achieve this goal the following definitions are in order.

Definition 3. *Using $\beta \in \mathbb{R}^+$ as a parameter tuning the level of* fast noise *in the network (with the physical meaning of inverse temperature, i.e. calling T the temperature, $\beta \equiv T^{-1}$ in proper units,), the partition function of the model (2.29) is introduced as*

$$Z_{N,P}(\sigma|\xi,t) := \sum_{\{\sigma\}} e^{-\beta H_{N,P}^{(RR)}(\sigma|\xi,t)} = \sum_{\{\sigma\}} \exp\left[\frac{\beta}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu} \sigma_i\sigma_j\right]. \tag{2.30}$$

Definition 4. *Denoting with $\mathbb{E}_\xi$ the average over the quenched patterns, for a generic function $O(\sigma,\xi)$ of the neurons and the couplings, we can define the Boltzmann $\langle O(\sigma,\xi)\rangle$ as*

$$\langle O(\sigma,\xi)\rangle \quad := \quad \frac{\sum_{\{\sigma\}} O(\sigma,\xi) e^{-\beta H_{N,P}^{(RR)}(\sigma|\xi,t)}}{Z_{N,P}(\sigma|\xi,t)}, \tag{2.31}$$

$$\tag{2.32}$$

such that its quenched average reads as $\mathbb{E}_\xi\langle O(\sigma,\xi)\rangle$.

Definition 5. *Once introduced the partition function $Z_{N,P}(\sigma|\xi,t)$, we can define the infinite volume limit of the intensive quenched free-energy $F_N(\alpha,\beta,t)$ and of the intensive quenched statistical pressure $A(\alpha,\beta,t)$ associated to the model (2.29) as*

$$-\beta F(\alpha,\beta,t) \equiv A(\alpha,\beta,t) := \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\ln Z_{N,P}(\sigma|\xi,t). \tag{2.33}$$

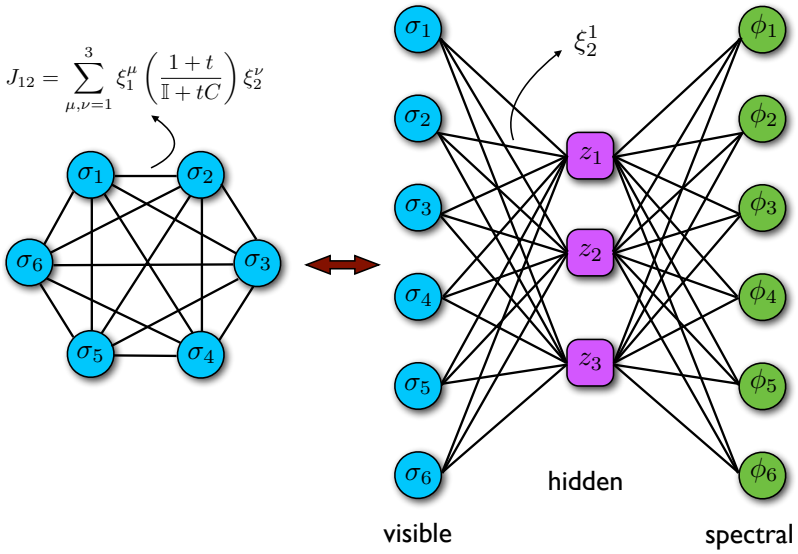

Figure 2.6: Stylized representation of the generalized Hopfield network (left) and its dual generalized (restricted) Boltzmann machine (right), namely the three-partite spin-glass under study: in machine learning jargon these parties are called *layers* and, here, they are respectively the visible, hidden and spectral layers. Note further that, as it should, when $t \to 0$ the duality above reduces to the standard picture of Hopfield networks and restricted Boltzmann machines [26, 52].

remark 3. *The partition function defined in (2.30) can be represented in Gaussian integral form as*

$$\begin{aligned} Z_{N,P}(\sigma|\xi,t) := \sum_{\{\sigma\}} \int \Big(\prod_{\mu=1}^{P} d\mu(z_\mu)\Big)\Big(\prod_{i=1}^{N} d\mu(\phi_i)\Big)\cdot \\ \cdot \exp\left(\sqrt{\frac{\beta}{N}(t+1)}\sum_{\mu,i}^{P,N} z_\mu \xi_i^\mu \sigma_i + i\sqrt{\frac{t}{N}}\sum_{\mu,i}^{P,N} z_\mu \xi_i^\mu \phi_i\right), \end{aligned} \tag{2.34}$$

where $d\mu(z_\mu)$ and $d\mu(\phi_i)$ are the standard Gaussian measures.

This relation shows that the partition function of the reinforcement&removal model is equivalent to the partition function of a tripartite spin-glass -a generalized RBM- where the intermediate party (or *hidden layer* to keep a machine learning jargon) is made of real neurons $\{z_\mu\}_{\mu=1,\dots,P}$ with $z_\mu \sim \mathcal{N}[0,1], \forall \mu$, while the external layers are made, respectively, of a set of Boolean neurons $\{\sigma_i\}_{i=1,\dots,N}$ (the *visible layer*) and of a set of imaginary neurons with magnitude $\{\phi\}_{i=1,\dots,N}$, being $\phi_i \sim \mathcal{N}[0,1], \forall i$ (the *spectral layer*), see Fig. 2.6.

2.2.1 Guerra's interpolating framework for the free energy

Plan of this subsection is to find out an explicit expression of the free energy in terms of the order and control parameters of the theory. En route toward the generation of the self-consistencies for the order parameters we then have to extremize the free energy w.r.t. them and, finally, by ispecting the evolution of the order parameters -by studyig numerically the solutions of the self-consistencies- in the space of the control parameter we

can work out the phase diagram of the *reinforcement & removal* algorithm.
From a methodological point of view we rely upon the Guerra's interpolation technique, as presented for the Sherrington-Kirkpatrick model in Section 1.3.7, suitably adapted to the case. To this task we start our treatmet by introducing the next

Definition 6. *Once expressed the partition function (2.30) in its integral representation (2.34), we can define the related tripartite spin glass Hamiltonian as*

$$H_{N,P} := \frac{a}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} z_\mu \xi_i^\mu k_i, \tag{2.35}$$

where we introduced the "multi-spin" $k_i = \sigma_i + b\phi_i$ and where

$$a = \sqrt{\beta(t+1)}, \quad b = i\sqrt{\frac{t}{\beta(t+1)}}. \tag{2.36}$$

remark 4. *Note that the cost function (2.35) and the one associated to the original model (2.29) share the same partition function and therefore exhibit the same Thermodynamics. By a practical perspective, as we will see soon, the latter is more suitable for understanding the retrieval capabilities of the network, the former for dealing with its learning skills [52, 53].*

In the following we consider the high storage case with $P = \alpha N$ for large N and we aim to obtain an expression for the quenched statistical pressure (2.33) in terms of the order parameters introduced in the next

Definition 7. *The natural order parameters for the neural network model (2.29) -as suggested by its integral representation (2.35)- are the overlaps q_{ab} and p_{ab} between the k's and the z's variables, respectively, as functions of two replicas (a,b) of the system, and the generalized Mattis overlap[1] m_1, namely*

$$q_{ab} := \frac{1}{N} \sum_{i=1}^{N} k_i^{(a)} k_i^{(b)}, \tag{2.37}$$

$$p_{ab} := \frac{1}{P} \sum_{\mu \geq 2} z_\mu^{(a)} z_\mu^{(b)}, \tag{2.38}$$

$$m_1 := \frac{1}{N} \sum_{i=1}^{N} \xi_i^1 k_i. \tag{2.39}$$

remark 5. *The replica symmetric approximation (RS) is imposed by requiring that the order-parameters of the theory do not fluctuate in the thermodynamic limit[2], i.e.*

$$q_{ab} \overset{RS}{\to} W\delta_{ab} + q(1-\delta_{ab}), \tag{2.40}$$

$$p_{ab} \overset{RS}{\to} X\delta_{ab} + p(1-\delta_{ab}), \tag{2.41}$$

$$m_1 \overset{RS}{\to} m, \tag{2.42}$$

where we called, respectively, W, q, X, p, m the replica symmetric values of the diagonal and off-diagonal overlap q, the diagonal and off-diagonal overlap p and the Mattis magnetization m_1.

[1] We arbitrarily (but with no loss of generality) nominated the first pattern as the retrieved one.

[2] This request is obviously perfectly consistent with the replica-symmetric ansatz when approaching the problem via the replica trick [8, 64].

Now the plan is to get an explicit expression for the pressure (2.33) in terms of these order parameters, to extremize the former over the latter and get a phase diagram for the network. To reach this goal we generalize a Guerra's interpolation scheme [20] as exposed in Section 1.3.7 for the SK model: the idea is to compare the original system, as represented in eq. (2.35) (namely a three-layer correlated spin glass), with three random single-layers, where each layer experiences, statistically, the same mean-field that would have been produced by the other layers over it. To this aim we introduce the following

Definition 8. *Being $s \in [0,1]$ an interpolating parameter, $\{\eta_i\}_{i\in(1,\dots,N)}$ a set of N i.i.d. Gaussian variables, $\{\lambda_\mu\}_{\mu\in(2,\dots,P)}$ a set of $P-1$ i.i.d. Gaussian variables, and the scalars C_1, C_2, C_3, C_4, C_5 to be set a posteriori, we use as interpolating pressure the following quantity*

$$\begin{aligned}\mathcal{A}(s) \;:=\; & \frac{1}{N}\mathbb{E}\,\xi,\eta,\lambda \ln \sum_{\sigma}\int d\mu\,(z,\phi)\exp\Big[\sqrt{s}\frac{a}{\sqrt{N}}\sum_{i,\mu\geq 2} z_\mu \xi_i^\mu k_i + \sqrt{s}\frac{a}{\sqrt{N}}\sum_i z_1\xi_i^1 k_i \qquad (2.43)\\ + \;& \sqrt{1-s}\Big(C_1\sum_i^N \eta_i k_i + C_2\sum_{\mu\geq 2}\lambda_\mu z_\mu\Big) + \frac{1-s}{2}\Big(C_3\sum_{\mu\geq 2} z_\mu^2 + C_4\sum_i k_i^2 + C_5 a\sum_i \xi_i^1 k_i\Big)\Big].\end{aligned}$$

remark 6. *When $s=1$ we recover the original model, namely $A(\alpha,\beta,t) = \lim_{N\to\infty}\mathcal{A}(s=1)$, while for $s\to 0$ we are left with a one-body problem, and, consequently, the probabilistic structure of $\mathcal{A}(s=0)$ is more tractable.*

remark 7. *We note the importance of splitting the sum on the ξ's into ξ^1 (i.e. the* signal*) and the $\xi^2\cdots\xi^P$ (i.e. the* quenched noise*) since the quenched average treats them differently, and so we will need to address them separately.*

Proposition 1. *The infinite volume limit of the quenched pressure related to the model (2.29) can be obtained by using the Fundamental Theorem of Calculus as*

$$A(\alpha,\beta,t) \equiv \lim_{N\to\infty}\mathcal{A}(s=1) = \lim_{N\to\infty}\left(\mathcal{A}(s=0) + \int_0^1 \frac{d\mathcal{A}(s)}{ds}\,ds\right). \qquad (2.44)$$

To follow this approach, two calculations are in order: the streaming $d_s\mathcal{A}(s)$ (and its successive back-integration) and the evaluation of the Cauchy condition $\mathcal{A}(s=0)$. Let us start with $d_s\mathcal{A}(s)$:

$$\begin{aligned}\frac{d\mathcal{A}(s)}{ds} \;=\; & \frac{1}{2N}\mathbb{E}\,\xi,\lambda,\eta\Big[\frac{a}{\sqrt{sN}}\sum_{i,\mu\geq 2}\xi_i^\mu\omega_s(z_\mu k_i) + \\ & -\frac{1}{\sqrt{1-s}}\Big(C_1\sum_i \eta_i\omega_s(k_i) + C_2\sum_{\mu\geq 2}\lambda_\mu\omega_s(z_\mu)\Big) + \\ & +\frac{a}{\sqrt{sN}}\sum_i \xi_i^1\omega_s(z_1 k_i) - C_3\sum_{\mu\geq 2}\omega_s(z_\mu^2) + \\ & -C_4\sum_i\omega_s(k_i^2) - C_5 a\sum_i\omega_s(\xi_i^1 k_i)\Big].\end{aligned}$$

We can proceed further by using Wick's Theorem [$\mathbb{E}_x x F(x) = \mathbb{E}_x(x^2)\cdot\mathbb{E}_x\partial_x F(x)$] on the

fields $z^1,\ \xi^{2\cdots P},\ \lambda_\mu,\ \eta_i$, obtaining

$$\begin{aligned}\frac{d\mathcal{A}(s)}{ds} &= \frac{1}{2N}\mathbb{E}\,\xi,\lambda,\eta\Big[\frac{a^2}{N}\sum_{i,\mu\geq 2}\Big(\omega_s(z_\mu^2 k_i^2)-\omega_s(z_\mu k_i)^2\Big)+\frac{a^2}{N}\omega_s((\sum_i \xi_i^1 k_i)^2)+\\ &- C_1^2\sum_i\Big(\omega_s(k_i^2)-\omega_s(k_i)^2\Big)-C_2^2\sum_{\mu\geq 2}\Big(\omega_s(z_\mu^2)-\omega_s(z_\mu)^2\Big)+\\ &- C_3\sum_{\mu\geq 2}\omega_s(z_\mu^2)-C_4\sum_i\omega_s(k_i^2)-C_5 a\sum_i\omega_s(\xi_i^1 k_i)\Big].\end{aligned} \tag{2.45}$$

Using the definition of the order parameters (2.39) we can write $d_s\mathcal{A}(s)$ as

$$\begin{aligned}\frac{d\mathcal{A}(s)}{ds} &= \frac{1}{2}\mathbb{E}\,\xi,\lambda,\eta\Big[a^2\alpha\omega_s(q_{11}p_{11})+a^2\omega_s(m_1^2)-a^2\alpha\omega_s(q_{12}p_{12})-C_1^2\omega_s(q_{11})+\\ &+C_1^2\omega_s(q_{12})-C_2^2\alpha\omega_s(p_{11})+C_2^2\alpha\omega_s(p_{12})-\alpha C_3\omega_s(p_{11})+\\ &-C_4\omega_s(q_{11})-aC_5\omega_s(m_1)\Big].\end{aligned} \tag{2.46}$$

It is now convenient to fix the free scalars $C_{1,..,5}$ as

$$C_1^2=a^2\alpha p,\quad C_2^2=a^2 q,\quad C_3=a^2(W-q),\quad C_4=a^2\alpha(X-p),\quad C_5=2ma, \tag{2.47}$$

such that we can recast the streaming $d_s\mathcal{A}(s)$ as

$$\begin{aligned}\frac{d\mathcal{A}(s)}{ds} &= \frac{1}{2}\mathbb{E}\,\xi,\lambda,\eta\Big[a^2\alpha\omega_s((q_{11}-W)(p_{11}-X))+a^2\omega_s((m_1-m)^2)+\\ &-a^2\alpha\omega_s((q_{12}-q)(p_{12}-p))\Big]+\frac{\alpha a^2}{2}(qp-WX)-\frac{a^2}{2}m^2.\end{aligned} \tag{2.48}$$

remark 8. *When requiring replica symmetry, we have that $\langle q_{11}\rangle\to W$, $\langle p_{11}\rangle\to X$, $\langle m_1\rangle\to m$, $\langle q_{12}\rangle\to q$ and $\langle p_{12}\rangle\to p$, hence the evaluation of the integral in eq. (3.174) becomes trivial as the r.h.s. of eq. (2.48) reduces to*

$$d_s\mathcal{A}(s)=\frac{\alpha a^2}{2}(qp-WX)-\frac{a^2}{2}m^2 \tag{2.49}$$

that does not depend on s any longer.

We must now evaluate the one-body contribution $\mathcal{A}(s=0)$: this can be done by directly setting $s=0$ in (2.43)

$$\begin{aligned}\mathcal{A}(s=0) =& \frac{1}{N}\mathbb{E}\,\xi,\eta,\lambda\ln\sum_\sigma\int d\mu(z,\phi)\exp\Big[C_1\sum_i\eta_i k_i+\frac{C_4}{2}\sum_i k_i^2+\frac{C_5 a}{2}\sum_i\xi_i^1 k_i+\\ &+C_2\sum_{\mu\geq 2}\lambda_\mu z_\mu+\frac{C_3}{2}\sum_{\mu\geq 2}z_\mu^2\Big].\end{aligned} \tag{2.50}$$

Performing standard Gaussian integrations we obtain

$$\begin{aligned}\mathcal{A}(s=0) =& -\frac{\alpha}{2}\ln(1-C_3)-\frac{1}{2}\ln(1-C_4b^2)+\frac{\alpha}{2}\frac{C_2^2}{1-C_3}+\frac{C_4}{2}+\mathbb{E}\,\eta\ln\cosh\Big[\frac{C_1\eta+\frac{C_5 a}{2}}{1-C_4b^2}\Big]+\\ &+b^2\frac{C_1^2+C_4^2+\frac{C_5^2a^2}{4}}{1-C_4b^2}+\ln 2.\end{aligned} \tag{2.51}$$

Keeping in mind the expressions for the parameters $C_1, ..., C_5$ as prescribed in the relations 2.47, by plugging eq. (2.49) and eq. (3.114) into the sum rule (3.174) we finally get an expression for the quenched pressure of the model (2.29) in terms of the replica-symmetric order parameters

$$\begin{aligned} A_{RS}(\alpha,\beta,t) = \; & \tfrac{\alpha a^2}{2}\big(qp - WX\big) - \tfrac{a^2}{2}m^2 - \tfrac{\alpha}{2}\ln\big[1-a^2(W-q)\big] - \tfrac{1}{2}\ln\big[1-a^2b^2\alpha(X-p)\big] + \\ & + \tfrac{\alpha}{2}\tfrac{a^2 q}{1-a^2(W-q)} + \tfrac{\alpha a^2}{2}\big(X-p\big) + \tfrac{a^2b^2}{2}\cdot\tfrac{\alpha p + m^2a^2 + a^2\alpha^2(X-p)^2}{1-a^2b^2\alpha(X-p)} + \\ & + \ln 2 + \mathbb{E}\,\eta \ln\cosh\Big[\tfrac{a\eta\sqrt{\alpha p}+ma^2}{1-\alpha a^2b^2(X-p)}\Big]. \end{aligned} \tag{2.52}$$

To match exactly the notation in [64] there is still a short way to go: it is convenient to re-scale m, p and X as

$$X \to \frac{\beta^2}{a^2}X, \quad p \to \frac{\beta^2}{a^2}p, \quad m \to \frac{\beta}{a^2}m, \tag{2.53}$$

as this allows us to introduce the composite order parameter $\Delta = 1 - \alpha\beta^2b^2(X-p)$ used in [64].
After these transformations, remembering the definition of the free energy (see (2.33)) and the definition of (a,b) (see (2.36)), we obtain exactly the same expression for the quenched free energy as that achieved in [64] via the replica trick, as stated by the next main

theorem 1. *In the infinite volume limit, the replica symmetric statistical pressure related to the neural network defined by eq. (2.29) can be expressed in terms of the natural order parameters of the theory (see def.s (2.39)) as*

$$\begin{aligned} F_{RS}(\alpha,\beta,t) = & -\frac{\beta m^2}{2(1+t)}\Big(1+\frac{t}{\Delta}\Big) - \frac{(1+t)(\Delta-1)}{2t}\beta W - \frac{\alpha\beta^2}{2}p(W-q) \\ & - \frac{\alpha}{2}\Big(\log[1-\beta(1+t)(W-q)] + \frac{q\beta^2(1+t)}{1-\beta(1+t)(W-q)}\Big) \\ & - \frac{\log\Delta}{2} - \frac{\alpha\beta p t}{2(1+t)\Delta} + \mathbb{E}\,\eta\log\cosh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big] + \\ & + \log 2 - \frac{(1+t)(1-\Delta)\beta}{2t\Delta}. \end{aligned} \tag{2.54}$$

Proposition 2. *Using the standard variational principle $\vec{\nabla}F_{RS} = 0$ on the statistical pressure (2.54), namely by extremizing the latter over the order parameters, we obtain the following set of self-consistent equations for these parameters, whose behavior is outlined in the plots of Fig. 2.7.*

$$m = \frac{1+t}{\Delta+t}\,\mathbb{E}\,\eta\tanh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big], \tag{2.55}$$

$$p = \frac{q(1+t)^2}{[1-\beta(1+t)(W-q)]^2}, \tag{2.56}$$

$$\Delta = 1 + \frac{\alpha t}{1-\beta(1+t)(W-q)}, \tag{2.57}$$

$$q = W + \frac{t}{\beta(1+t)\Delta} - \frac{1}{\Delta^2}\,\mathbb{E}\,\eta\cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big], \tag{2.58}$$

$$\begin{aligned} W\Delta^2 = & \; 1 - \frac{t\Delta}{\beta(1+t)} + \frac{\alpha p t^2 - m^2 t(t+2\Delta)}{(1+t)^2} \\ & - \frac{2\alpha\beta p t}{(1+t)\Delta}\,\mathbb{E}\,\eta\cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big]. \end{aligned} \tag{2.59}$$

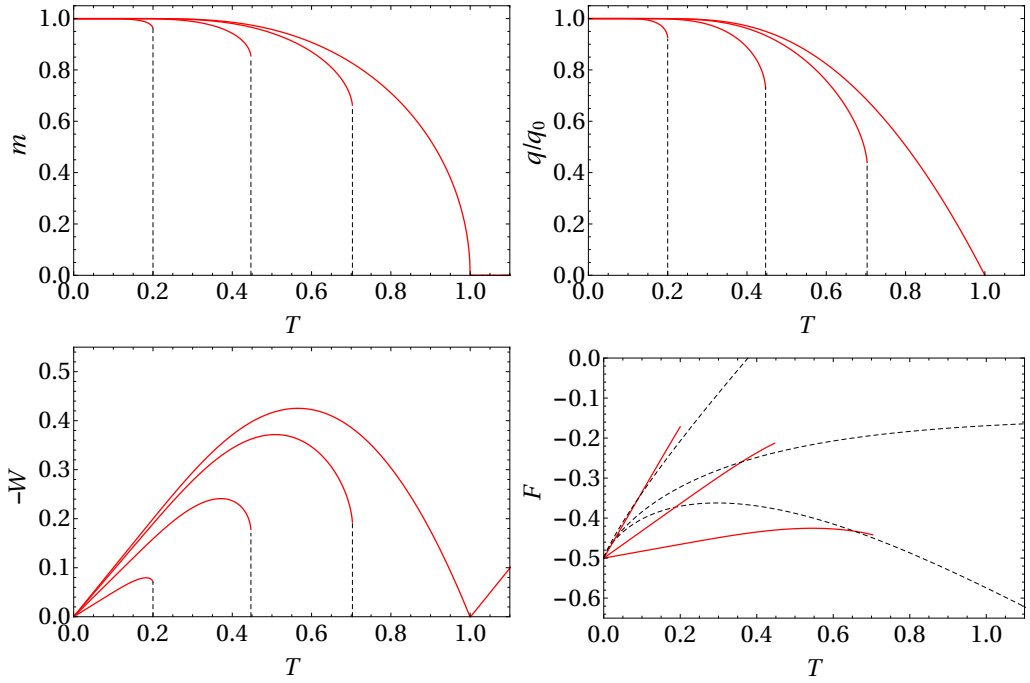

Figure 2.7: **Retrieval state solution for the order parameters and free energy at $t = 1000$.** First row: on the left, the plot shows the Mattis magnetization m as a function of the temperature for various storage capacity values ($\alpha = 0$, 0.05, 0.2 and 0.5, going from the right to the left). The vertical dotted lines indicates the jump discontinuity identifying the critical temperature $T_c(\alpha)$ which separates the retrieval region from the spin-glass phase; on the right, the plot shows the solutions of the non-diagonal overlap q (normalized to the zero-temperature value $q_0 = q(T = 0)$), for the same capacity values. The solution is computed in the retrieval region (*i.e.* $T < T_c(\alpha)$). Second row: on the left, the plot shows the solution for the diagonal overlap $-W$ in the retrieval region for $\alpha = 0$, 0.05, 0.2 and 0.5, finally, on the right the plot shows the free-energy as a function of the temperature for various storage capacity values ($\alpha = 0.05$, 0.2 and 0.5, going from the bottom to the top) for both the retrieval (red solid lines) and spin-glass (black dashed lines) states.

remark 9. *We stress that we obtained exactly the same self-consistencies previously appeared in [64], thus all the consequences stemming by them, as reported in that paper, are here entirely confirmed.*

2.2.2 Replica symmetric phase diagram

In order to contribute to OAI (Optimized Artificial Intelligence), statistical mechanics of complex systems provided one of the main rewards, namely it allows painting phase diagrams for the various neural architectures under study, where the operational regimes of the machine appear in the space of the tunable parameters (e.g. storage load, noise, etc.) split by *computational phase transitions* much as the various regimes of water (e.g. solid, liquid, vapour) shine in its relative phase diagram lying in its space of its tunable parameters (pressure, temperature and volume in that scenario). In these regards, the Hopfield model able to sleep (or alternatively equipped with the reinforcement-&-remotion algorithm) is accompained as well with a phase diagram (see Figure 2.8), obtaied by

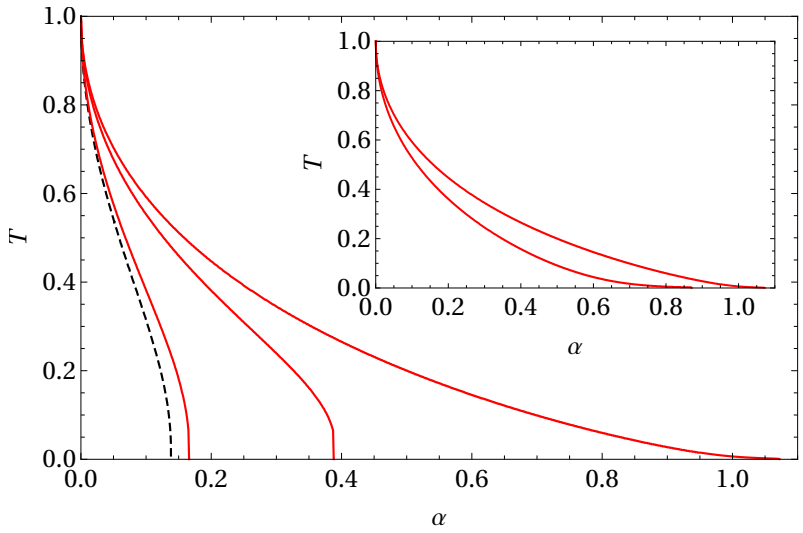

Figure 2.8: Critical line for the transition between retrieval and spin-glass phases for various values of the unlearning time. From the left to the right: $t = 0$ (Hopfield, black dashed line), 0.1, 1 and 1000. The inset shows two curves tracing the boundary of the maximal retrieval regions where patterns are global free energy minima (inner boundary) or local free energy minima (outer boundary) in the long sleep limit.

inspecting the solution of the above self-consistencies (see eq.s (2.55)): the most striking property of this network is that the retrieval region extendes from $\alpha_c \sim 0.14$ up to the maximal one as prescribed by bounds in Information Theory namely $\alpha_c = 1$ (one bit per binary neuron, a quite intuitive saturation) and the convergence to this value is quite fast in the sleeping time -we plotted various retrieval regions as the sleeping time is smoothly increased (suggesting that also a simple "nap" can significantly help the network to optimize its storage). Further, we do not deepen the way in which this order takes place, but it is (tricky yet) possible to show that this form of dreaming concretely implements the Gram-Smidth ortonormalization in the space of the patterns (that are ortogonal solely in the infinite volume limit as they are build at random, hence $\lim_{N\to\infty}\langle \xi^\mu \xi^\nu \rangle = \delta_{\mu,\nu}$ but at finite N fluctuations (going to zero as $1/\sqrt{N}$ do contribute and the sleeping mechanism ensures that -also at finite volume- the storage of the patterns is always kept *ortogonal* (hence optimal in this random setting).

A far from trivial consequence of this optimization in the space of the patterns that the network achieves by sleeping is that the spin-glass states are completely destroyed as we show by a fluctuation analysis and an inspection of criticality as performed in the next two subsections): roughly speaking, the way in which this network stores the patterns guarantees minimal frustration level in the network, that in turn preserve its behavior to get stuck into unwanted spurious spin-glass metastable states. Finally, we remark that -beyond Optimized AI- thanks to a continuous bridge with biological information processing- the explanation of the optimal storage in terms of the Gram-Smidth technique significantly helps also our understanding of artificial information processing, also in this more challenging case.

2.2.3 Study of the overlap fluctuations

As proved in the previous section, the reinforcement&removal algorithm makes the retrieval region in the (α, β) plane wider and wider as t is increased (see Fig. 2.8). As the

retrieval region pervades the spin-glass region, one therefore naturally wonders whether the opposite boundary of the spin-glass region (namely the critical line depicting the transition where ergodicity breakdowns) is as well deformed. To address this point, we now study the behavior of the overlap fluctuations, suitably centered around the thermodynamic values of the overlaps and properly rescaled in order to allow them to diverge when the system approaches the critical line. In fact, they are meromorphic functions and their poles identify the evolution of the critical surface $\beta_c(\alpha, t)$ (if any).
It is worth recalling that the critical line for the standard Hopfield model [39] as predicted by the AGS theory [11] is $\beta_c(\alpha, t = 0) = (1 + \sqrt{\alpha})^{-1}$.

The idea is the same exploited in the previous sections, namely to use the generalized Guerra's interpolation scheme (see eq. (2.43)) to evaluate the evolution of the order parameter's correlation functions from $s = 0$ (where they do not represent the real fluctuations in the system, but their evaluation should be possible) up to $s = 1$ (where they reproduce the true fluctuations). To achieve this goal for the generic correlation function O, we need to evaluate the Cauchy condition $\langle O(s = 0)\rangle$ and the derivative $\partial_s\langle O(s)\rangle$. However, in contrast with the previous section where we imposed replica symmetry, here -as we just want to infer the critical line- we impose ergodic behavior, namely, we assume that the system is approaching this boundary from the high fast-noise limit. This allows us to set all the mean values of the overlaps to zero and to achieve explicit solutions.

Definition 9. *The centered and rescaled overlap fluctuations θ_{lm} and ρ_{lm} are introduced as*

$$\theta_{lm} = \sqrt{N}\big[q_{lm} - \delta_{lm}W - (1 - \delta_{lm})q\big] \tag{2.60}$$
$$\rho_{lm} = \sqrt{P}\big[p_{lm} - \delta_{lm}X - (1 - \delta_{lm})p\big]. \tag{2.61}$$

remark 10. *As we will address the problem of the overlap fluctuations in the ergodic region, the signal is absent, thus there is no need to introduce a rescaled Mattis order parameter: only the boundary between the ergodic region and the spin-glass region is under study here.*

Proposition 3. *It is convenient to introduce the $r-$replicated interpolating pressure $\mathcal{A}^r_J(s)$, where we further added a source field J, coupled to an observable O (that is a smooth function of the neurons of the r-replicas) as*

$$\begin{aligned}\mathcal{A}^r_J(s) &= \mathbb{E}_{\xi,\eta,\lambda}\ln\sum_{\sigma_R}\int d\mu\,(z_R,\phi_R)\exp\Big[\frac{a\sqrt{s}}{\sqrt{N}}\sum_{l=1}^{r}\sum_{i,\mu} z_\mu^{(l)}\xi_i^\mu k_i^{(l)}\\ &+ \sqrt{1-s}\Big(C_1\sum_{l=1}^{r}\sum_i \eta_i k_i^{(l)} + C_2\sum_{l=1}^{r}\sum_\mu \lambda_\mu z_\mu^{(l)}\Big)\\ &+ \frac{1-s}{2}\Big(C_3\sum_{l=1}^{r}\sum_\mu (z_\mu^{(l)})^2 + C_4\sum_{l=1}^{r}\sum_i (k_i^{(l)})^2\Big) + J\hat{O}\Big].\end{aligned} \tag{2.62}$$

where k_i is the same as in Definition 5 and the interpolation constants $C_{1,2,3,4}$ are the same given in the previous section (see eq. ((2.47))).

By definition

$$\omega_s(O(s)) = \frac{\partial\mathcal{A}^r_J(s)}{\partial J}\Big|_{J=0}, \qquad \partial_s\omega_s(O(s)) = \frac{\partial(\partial_s\mathcal{A}^r_J)}{\partial J}\Big|_{J=0}. \tag{2.63}$$

Therefore, in order to evaluate the fluctuations of O we need to evaluate first $\partial_s \mathcal{A}_J^r$ and, by a routine calculation, we get

$$\partial_s \mathcal{A}_J^r = \frac{1}{2}\sqrt{\alpha}\beta(1+t)\sum_{l,m=1}^{r}\Big[\omega_s(g_{l,m}) - \omega_s(g_{l,m+r})\Big], \qquad g_{l,m} = \theta_{l,m}\rho_{l,m}. \tag{2.64}$$

To evaluate the fluctuations of a general operator O, function of $r-$replicas, we must use the results (2.63) and perform the same rescaling that we did in the previous section, namely

$$(X,p) \to \frac{\beta^2}{a^2}(X,p). \tag{2.65}$$

Overall this brings to the next

Proposition 4. *Given O as a smooth function of r replica overlaps $(q_1,\dots,q_r)$ and $(p_1,\dots,p_r)$, the following streaming equation holds:*

$$\begin{aligned} d_\tau \omega_s(O) =& \frac{1}{2}\sum_{a,b}^{r}\omega_s(O\cdot g_{a,b}) - r\sum_{a=1}^{r}\omega_s(O\cdot g_{a,r+1}) + \\ &+ \frac{r(r+1)}{2}\omega_s(O\cdot g_{r+1,r+2}) - \frac{r}{2}\omega_s(O\cdot g_{r+1,r+1}), \end{aligned} \tag{2.66}$$

where we used the operator d_τ defined as

$$d_\tau = \frac{1}{\beta(1+t)\sqrt{\alpha}}\frac{d}{ds}, \tag{2.67}$$

in order to simplify calculations and presentation.

2.2.4 Criticality and ergodicity breaking

To study the overlap fluctuations we must consider the following correlation functions (it is useful to introduce and link them to capital letters in order to simplify their visualization):

$$\omega_s(\theta_{12}^2)_s = A(s), \omega_s(\theta_{12}\theta_{13})_s = B(s), \omega_s(\theta_{12}\theta_{34})_s = C(s), \tag{2.68}$$
$$\omega_s(\theta_{12}\rho_{12})_s = D(s), \omega_s(\theta_{12}\rho_{13})_s = E(s), \omega_s(\theta_{12}\rho_{34})_s = F(s), \tag{2.69}$$
$$\omega_s(\rho_{12}^2)_s = G(s), \omega_s(\rho_{12}\rho_{13})_s = H(s), \omega_s(\rho_{12}\rho_{34})_s = I(s), \tag{2.70}$$
$$\omega_s(\theta_{11}^2)_s = J(s), \omega_s(\theta_{11}\rho_{11})_s = K(s), \omega_s(\rho_{11}^2)_s = L(s), \tag{2.71}$$
$$\omega_s(\theta_{11}\theta_{12})_s = M(s), \omega_s(\theta_{11}\rho_{12})_s = N(s), \omega_s(\rho_{11}\theta_{12})_s = O(s), \tag{2.72}$$
$$\omega_s(\rho_{11}\rho_{12})_s = P(s), \omega_s(\theta_{11}\rho_{22})_s = Q(s), \omega_s(\theta_{11}\theta_{22})_s = R(s). \tag{2.73}$$
$$\omega_s(\rho_{11}\rho_{22})_s = S(s), \tag{2.74}$$

Since we are interested in finding the critical line for ergodicity breaking *from above* we can treat $\theta_{a,b}, \rho_{a,b}$ as Gaussian variables with zero mean (this allows us to apply Wick-Isserlis theorem inside averages) as we can also treat both the k_i and z_μ as zero mean random variables in the ergodic region (thus all averages involving uncoupled fields are vanishing): this considerably simplifies the evaluation of the critical line (as expected since we are approaching criticality from the *trivial* ergodic region [68]).
We can thus reduce the analysis to

$$\omega_s(\theta_{12}^2)_s = A(s), \omega_s(\theta_{12}\rho_{12})_s = D(s), \omega_s(\rho_{12}^2)_s = G(s), \tag{2.75}$$
$$\omega_s(\theta_{11}^2)_s = J(s), \omega_s(\theta_{11}\rho_{11})_s = K(s), \omega_s(\rho_{11}^2)_s = L(s), \tag{2.76}$$
$$\omega_s(\theta_{11}\rho_{22})_s = Q(s), \omega_s(\theta_{11}\theta_{22})_s = R(s), \omega_s(\rho_{11}\rho_{22})_s = S(s). \tag{2.77}$$

According to (2.66) and to the previous reasoning we obtain:

$$d_\tau A = 2AD, \tag{2.78}$$
$$d_\tau D = D^2 + AG, \tag{2.79}$$
$$d_\tau G = 2GD. \tag{2.80}$$

Suitably combining A and G in (2.80) we can write

$$d_\tau \ln \frac{A}{G} = 0 \implies A(\tau) = r^2 G(\tau), \quad r^2 = \frac{A(0)}{G(0)}. \tag{2.81}$$

Now we are left with

$$d_\tau D = D^2 + r^2 G^2, \tag{2.82}$$
$$d_\tau G = 2GD. \tag{2.83}$$

The trick here is to complete the square by summing $d_\tau D + r d_\tau G$ thus obtaining

$$d_\tau Y = Y^2, \tag{2.84}$$
$$Y = D + rG, \tag{2.85}$$
$$d_\tau G = 2G(Y - rG). \tag{2.86}$$

The solution is trivial and it is given by

$$Y(\tau) = \frac{Y_0}{1 - \tau Y_0}, \quad Y_0 = D(0) + \sqrt{A(0)G(0)}. \tag{2.87}$$

So we are left with the evaluation of the correlations at $s = 0$: namely the Cauchy

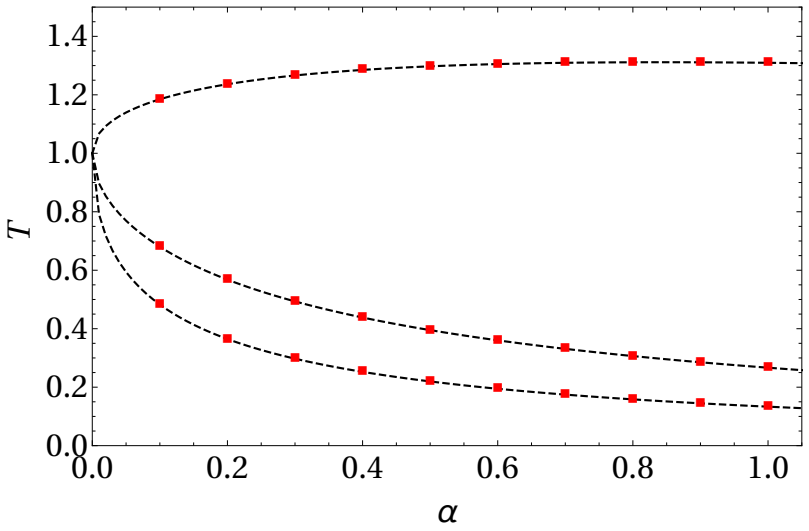

Figure 2.9: **Ergodicity breaking critical line.** The plot shows a comparison between the theoretical predictions (black dashed lines) for the ergodicity breaking critical line according to Eq. (2.97) and numerical solutions for spin glass states (red markers). The latter are evaluated by solving the self-consistency equations with $m = 0$ with α fixed and searching for the temperature T above which the solution has $q = 0$. Going from top to bottom of the plot, the sleep extent is $t = 0.1$, 1 and 2.

conditions related to the solution coded in eq. (2.87). To this task we introduce a one-body generating function for the momenta of z, k: this can be done by setting inside (2.62) $s = 0, r = 1$ and adding source fields (j_i, J_μ) coupled respectively to (k_i, z_μ), with $i \in (1, ..., N)$, $\mu \in (1, ..., P)$. Since we are approaching the critical line from the high fast noise limit we can set $m, p, q = 0$ (when we explicitly make use of the coefficients (2.47)), overall writing

$$F(j, J) = \ln \sum_\sigma \int d\mu\,(z, \phi) \exp\Big[\sum_i j_i k_i + \sum_\mu J_\mu z_\mu + \tfrac{a^2 W}{2} \sum_\mu z_\mu^2 + \tfrac{1-\Delta}{2b^2} \sum_i k_i^2\Big] \quad (2.88)$$

Clearly, we took great advantage in approaching the ergodic region from above, since even the one-body problem (for the Cauchy condition) has been drastically simplified: showing only the relevant terms in j, J we have

$$F(j, J) = \frac{b^2\Delta + 1}{2\Delta^2} \sum_i j_i^2 + \frac{1}{2(1 - a^2 W)} \sum_\mu J_\mu^2 + O(j^3). \quad (2.89)$$

As anticipated, all the observable averages needed at $s = 0$ can now be calculated simply as derivatives of $F(j, J)$, thus the $s = 0$ correlation functions are finally given by

$$D(0) = \sqrt{NP}\,\big(\partial_j F\big)^2 \big(\partial_J F\big)^2\Big|_{j,J=0} = 0, \quad (2.90)$$

$$A(0) = \big(\partial_j^2 F\big)^2\Big|_{j,J=0} = \Big[\frac{\beta(1+t) - t\Delta}{\beta(1+t)\Delta^2}\Big]^2 = W^2, \quad (2.91)$$

$$G(0) = \big(\partial_J^2 F\big)^2\Big|_{j,J=0} = (1 - \beta(1+t)W)^{-2}. \quad (2.92)$$

Inserting this result in (2.87), we get

$$Y(\tau) = \frac{W}{1 - \beta(1+t)W - \tau W}. \quad (2.93)$$

Upon evaluating $Y(\tau)$ for $\tau = \beta(1+t)\sqrt{\alpha}s$, $s = 1$ and reporting the relevant ergodic self-consistent equations we obtain the following system:

$$Y(s=1) = \frac{W}{1 - \beta(1+t)W(1+\sqrt{\alpha})}, \quad (2.94)$$

$$W\Delta^2 = 1 - \frac{t\Delta}{\beta(1+t)}, \quad (2.95)$$

$$\Delta = 1 + \frac{\alpha t}{1 - \beta(1+t)W}. \quad (2.96)$$

Since we are interested in obtaining the critical temperature for ergodicity breaking, where fluctuations (in this case Y) grow arbitrarily large we can check where the denominator at the r.h.s. of the first eq. (2.94) becomes zero and recast this observation as follows

theorem 2. *The ergodic region of the model defined by the cost function (2.29) is delimited by the following critical surface in the (α, β, t) space of the tunable parameters*

$$\beta_c = \frac{1}{1+t}\Big[\frac{\Delta^2}{1+\sqrt{\alpha}} + t\Delta\Big] \quad \textit{with} \quad \Delta = 1 + \sqrt{\alpha}(1+\sqrt{\alpha})t. \quad (2.97)$$

remark 11. *At $t = 0$, where the model reduces to Hopfield's scenario, the critical surface correctly collapses over the Amit-Gutfreund-Sompolinsky critical line $\beta_c = (1+\sqrt{\alpha})^{-1}$, but*

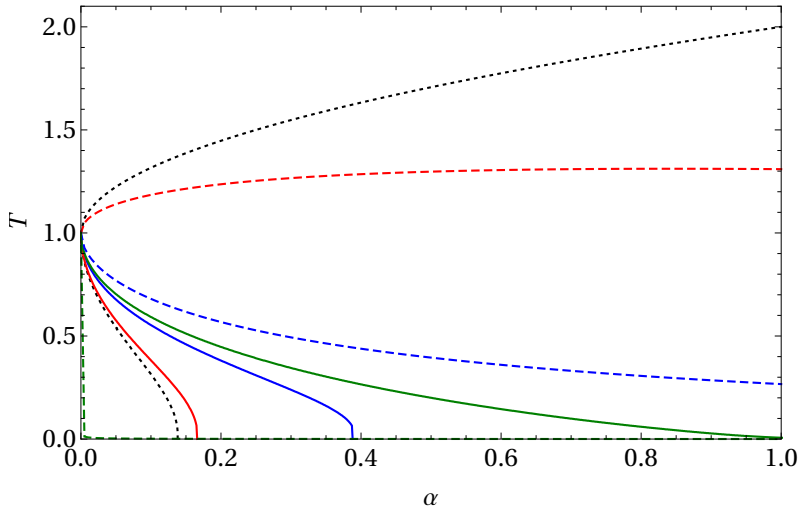

Figure 2.10: Critical lines for ergodicity breaking (dotted curves) and retrieval region boundary (solid curves) for various values of the unlearning time. From the top to the bottom: $t = 0$ (black lines, i.e. the Hopfield phase diagram), $t = 0.1$ (red lines), 1 (blue lines) and 1000 (green lines).

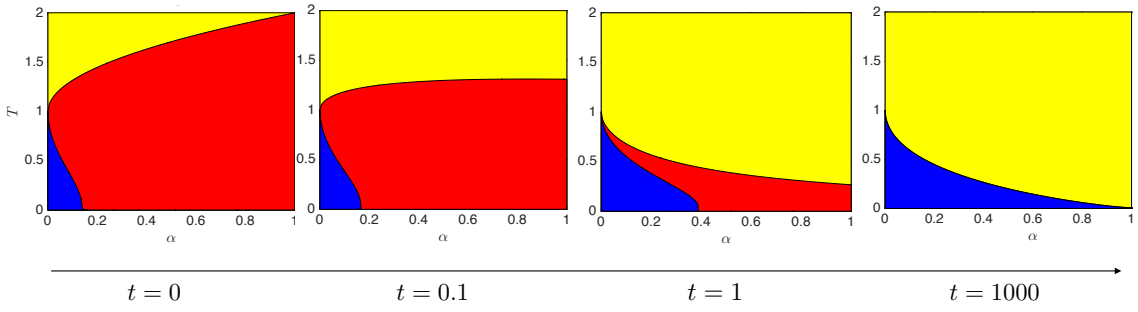

Figure 2.11: The phase diagram is depicted for different choices of t, namely, from left to right, $t = 0, 0.1, 1, 1000$. Notice that, as t grows, the retrieval region (blue) and the ergodic region (yellow) get wider at the cost of the spin-glass region (red) which progressively shrinks up to collapse as $t \to \infty$. Also notice the change in the concavity of the critical line which separates ergodic and spin-glass region.

in the large t limit the ergodic region collapses to the axis $T = 0$: this may have a profound implication, namely that the ergodic region -during the sleep state- phagocytes *the spin-glass region.*
Since we have already seen that also the retrieval region phagocytes *the spin-glass region* [1] *this means that spurious states are entirely suppressed with a proper rest, allowing the network to achieve perfect retrieval, as suggested in the pioneering study by Kanter and Sompolinsky [66].*

[1] Note that the ergodic line does not affect the retrieval region, they simply *fade* one into the other. This is because the critical surface is calculated assuming an ergodic regime (hence, it does not takes into account the signal) and, more importantly, the retrieval region is delimited by a first order phase transition, that is not detected by a second order inspection as that needed for criticality.

2.2.5 Discussion on *ultra-memory* as an emergent skill

Summarizing the current Section where we reported on our research findings on Theoretical Artificial Intelligence to overcome the huge storage limitation of the bare Hopfield reference, we extended the previous schemes and architectural designs of sleeping mechanisms for neural networks as modeled by Crick & Mitchinson [69], Hopfield himself [70] and by many others in the Neuroscience Literature, see e.g [71, 72, 73, 74]) by accounting -beyond forgetting spurious memories via REM-like mechanisms- also the consolidation of pure memories by accounting also for SWS-like mechanisms. Mathematically we described the phenomenon of sleeping via a novel algorithm, the *reinforcement & removal* generalization of the Hopfield network: interestingly, such mechanisms have been evidenced to lead to a severe improvement of the retrieval capacity of the system, reaching the theoretical bound of symmetric networks prescribed by $\alpha_c = 1$. In particular, we showed that the Hopfield network able to take some rest reaches the expected upper critical capacity $\alpha_c = 1$, still preserving robustness with respect to fast noise. To paint this scenario we extended a Guerra's interpolation scheme [20], originally developed to deal with the standard Hopfield model (i.e. equipped with the canonical Hebbian synaptic coupling), to deal with this generalization: at first we showed the equivalence of this model with a three-layer spin-glass where some links among different layers are cloned (hence introducing correlation in the network and in the random fields required for the interpolation) and the third, and novel (w.r.t. the standard equivalence between Hopfield models and two-layers Boltzmann machines we built in the first Chapter), layer is equipped with imaginary real-valued neurons (best suitable to perform spectral analysis[1]). As a consequence, the resulting interpolating architecture is rather tricky, by far richer than its classical limit yet it turns out to be managable and actually a sum rule for the quenched free energy related to the model can be written and even integrated, under the assumption of replica symmetry, proving the saturation of the theoretical bound to $\alpha_c = 1$. We further systematically developed a fluctuation analysis of the overlap correlation functions, searching for critical behaviour, in order to inspect where ergodicity breaks down and in this investigation we found a very interesting result: as long as the Hopfield model is awake, the critical line is the one predicted by Amit-Gutfreund-Sompolinksy (as it should and as it is known by decades). However, as the network sleeps, the ergodic region starts to invade the spin glass region, ultimately destroying the spin glass states entirely, thus allowing the network (at the end of an entire sleep session) to live solely within a -quite large- retrieval region, surrounded by ergodicity: noticing that at this final stage of sleeping the network approached the Kanter-Sompolinsky model [66], it shines why these Authors called their model *associative recall of memory without errors.*
Let us also remark the importance nowadays of an OAI (Optimized Artificial Intelligence) and how, in these regards, statistical mechanics allows painting the phase diagram of the machine, information that -in turn- allows setting the machine in its optimal operational regime for a given task: this kind of information has been provided by the present research for this sleepy Hopfield network (see Figure 2.11) and deepened in the relative section (see Section 2.2.3).
While somehow this extension closes the discussion on the storage capacity for symmetric networks, yet -continuously driven by the inspection of biological information processing (en route also for XAI, eXplainable Artificial Intelligence)- it is still questionable that the signal-to-noise threshold in the standard Hopfield model (or equivalently in its dual representation in terms of the RBM) has to be order one, namely the signal has to shine a bit

[1] We plan to report soon on the learning algorithms for this generalized restricted Boltzmann machine, where the properties of the spectral layers will spontaneously shine.

in the sea of the fluctuating noise, a rather unsatisfactory limitaton that in humans is not present: for instance, if we have to recognize a signal in the fog or in the dark, we can adapt our signal-to-noise threshold by suitably changing the focus of our eyes... yet we have two eyes while the input layer for the RBM we studied is just one. In the next Section the last generalization of the Hopfiled reference we achieved during my PhD research time will be presented: networks that can tune their signal-to-noise threshold for pattern detection, achieving *ultra-detection*.
As we will see, starting by the Boltzmann machine dual representation, we will equip the latter with two input layers (namely one input and one mirror layer) rather than just one input layer to prove that the resulting redundancy of information stemming from this doubled source allows for tunable signal-to-noise thresholds. To understand this result the duality between standard RBM and Hopfield networks will be extended beyond the *statistical reductionism*, toward a three-layer RBM (often called Sejinowski machine), whose dual network is a dense Hebbian kernel -namely a P-spin Hopfield network with P=4. Crucially, while the latter is known to store a maximal amount of patterns K scaling as $K \sim \gamma(P)N^{P-1}$ (and the standanrd Hopfield limit is recovered for $P = 2$ where $\gamma(2) \to \alpha_c$), it could however sacrifice memory storage and handle *just* N^1 patterns (rathen than the maximal amount $\propto N^3$) and -in this relatively *low-storage setting*- it can lower its signal-to-noise threshold, as -intuitively- by dealing with *solely* N^1 patterns they can be much more noisy than the standard ones, in particular we will prove that their noise can even diverge in the thermodynamic limt, yet the network will still be able to perform the recognition of the pattern. Let us deepen this concepts in the next Section.

2.3 Neural Networks equipped with Ultra-Detection

2.3.1 The idea beyond *redundant representations*

Here we consider a minimal extension of the basic architecture for machine learning discussed so far(i.e. the RBM), namely the *restricted Sejnowski machine* (RSM) [75], that is a third-order Boltzmann machine [76], where triples of units interact symmetrically; in the jargon of Statistical Mechanics, this is just a three-layer spin-glass with (P=3)-wise interactions. In particular, we equip this network with a standard hidden layer and with two visible layers (a primary and a mirror channel, see Fig.2.12 left), which possibly mimic the typical presence of two input sources in biological networks (i.e., the *eyes*). As we show, the RSM displays, as a dual representation, a bipartite DAM, i.e., a bipartite Hopfield model with ($P = 4$)-wise interactions, see Fig. 2.12 right. In this dual representation, the K features embedded in the RSM correspond to the K patterns stored in the DAM.
It is worth recalling that a neural network with ($P > 2$)-wise interactions among its units does not need to fulfil Gardner's bound: the latter holds solely for quadratic cost functions and implies that, being $K_{\max}(N)$ the largest number of random i.i.d. patterns that a network built of N binary neurons can store, then $\lim_{N\to\infty} K_{\max}(N)/N = \alpha_c < 2$ [67]. In fact, Baldi & Venkatesh [77] proved that, for a P-spin associative memory built of N binary neurons, $K_{\max}(N) \propto N^{P-1}$ (a result made rigorous by Bovier & Niederhauser [78]); clearly for $P = 2$ we recover the standard Hopfield scenario.
In the last decades, the quest for enhanced storage capacities has strongly biased the statistical mechanical investigations, possibly limiting alternative inspections of the computational capabilities of these networks, which is the main focus of this work, as summarized hereafter.

In the standard Hopfield model it is possible to retrieve a number K of patterns that

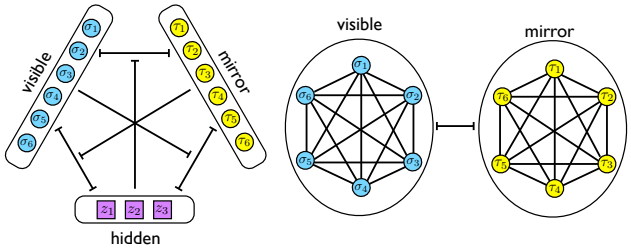

Figure 2.12: Schematic representations of the Restricted Seinowskj Machine (left) and its dual representation in terms of a bipartite Dense Associative Network (right). In the former, neurons i, μ, ρ interact 3-wisely through the coupling $\xi^{\rho}_{i\mu}$ (see also eq. 2.98), while, in the latter, neurons i, μ, j, ν interact 4-wisely through the coupling $J^{j\nu}_{i\mu}$ (see also eq. 2.102).

is extensive in N (i.e., $K = \alpha N$ with $\alpha \leq 0.14$) by pushing the signal-to-noise ratio to its limit, namely by letting the magnitude $\mathcal{S}$ of the signal – stemming from the pattern to be retrieved – and the magnitude $\mathcal{N}$ of the (quenched) noise – stemming from the remaining patterns providing an intrinsic glassiness – share the same order. Should the information encoded by patterns be affected by some source of noise, the condition $\mathcal{S}/\mathcal{N} \sim \mathcal{O}(1)$ would be deranged in favour of the noise and retrieval capabilities would be lost. On the other hand, as we show, if we let dense ($P = 4$) networks operate with a load $K = \alpha N$ (with $\alpha > 0$), these turn out to be able to retrieve the information ($\sim \mathcal{O}(1)$) encoded by patterns is perturbed by extensive noise ($\sim \mathcal{O}(\sqrt{N})$). This is ultimately due to the possibility of redundant representation of patterns [79, 80], which implies a storage cost of $\mathcal{O}(N^2)$ bits per pattern.
In the following we give more technical details to prove the previous statements.

The RSM [75] considered here is built on three layers, two of which – referred to as *visible* and *mirror*, respectively (see Fig.2.12, left panel) – are digital and made up of N Ising neurons per layer, $\boldsymbol{\sigma} \in \{-1, +1\}^N$ and $\boldsymbol{\tau} \in \{-1, +1\}^N$, while the third layer – referred to as *hidden* – is analog and made of K neurons $\boldsymbol{z}$, whose states are i.i.d. Gaussians $\mathcal{N}(0, \beta^{-1})$ ($\beta > 0$ tuning the level of the fast noise in the net [11]). The model presents third-order interactions among neurons of different layers but no intra-layer interactions (whence the *restriction*). Its cost function H_{RSM} is given by

$$H_{\mathrm{RSM}}(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{z}|\boldsymbol{\xi}) = -\frac{1}{N^{3/2}} \sum_{i,\mu,\rho=1}^{N,N,K} \xi^{\rho}_{i\mu} \sigma_i \tau_\mu z_\rho, \tag{2.98}$$

with $i, \mu = 1, .., N$ and $\rho = 1, .., K$. In the thermodynamic limit each layer size diverges such that $\lim_{N\to\infty} K/N = \alpha > 0$ and the factor $N^{-3/2}$ keeps the mean value of the cost function (under the quenched Gibbs measure [81]) linearly extensive in N. The interaction between each triplet of neurons is encoded in the $K \times N \times N$ tensor $\boldsymbol{\xi}$ whose ρ-th element will be written as

$$\xi^{\rho}_{i\mu} = \xi^{\rho}_i \xi^{\rho}_\mu, \quad i, \mu = 1, ..., N, \tag{2.99}$$

where $\xi^{\rho}_i \in \{-1, +1\}$ is meant as the i-th entry of the ρ-th pattern to be retrieved in the dual bipartite DAM. Notice that the factorization (2.99) ensures the symmetry of $\xi^{\rho}_{i\mu}$ for any ρ and this lies at the core of the pattern redundancy scheme pursued here. In fact, the information contained into a set of K binary patterns of length N is inflated into a symmetric tensor of size KN^2.

Given a small learning rate $\epsilon > 0$, we obtain for this network the following contrastive-divergence [48] learning rule (see Section 2.3.2 for details on its derivation and performances)

$$\Delta\xi_{i\mu}^{\rho} = \epsilon\beta\left(\langle\sigma_i\tau_\mu z_\rho\rangle_+ - \langle\sigma_i\tau_\mu z_\rho\rangle_-\right), \tag{2.100}$$

where the subscript "+" means that both visible and mirror layers are set at the data input (i.e., they are *clamped*), while the subscript "−" means that all neurons in the network are left free to evolve; importantly, while clamped, visible and mirror layers are always exposed to the same information (i.e., $\boldsymbol{\sigma} = \boldsymbol{\tau} = \boldsymbol{\xi}^{\rho}$).
Using the symbol Dz_ρ to denote the Gaussian measure with variance β^{-1} (i.e., $Dz_\rho \equiv dz_\rho \exp(-\beta z_\rho^2/2)\sqrt{\beta/2\pi}$), the partition function Z related to the cost function (2.98) reads

$$Z = \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \int \prod_{\rho=1}^{K} Dz_\rho \exp\left(\frac{\beta}{N^{3/2}} \sum_{i,\mu,\rho=1}^{N,N,K} \xi_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\right). \tag{2.101}$$

By construction, the couplings are symmetric ($\xi_{i\mu}^{\rho} = \xi_{\mu i}^{\rho}$) and detailed balance ensures that the long term relaxation of any (not-pathological) neural dynamics is described by the related Gibbs measure [11, 52]. Marginalizing over the hidden layer,

$$P(\boldsymbol{\sigma},\boldsymbol{\tau}|\boldsymbol{\xi}) = \frac{\int D\boldsymbol{z} e^{-\beta H_{\mathrm{RSM}}(\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{z}|\boldsymbol{\xi})}}{Z} \equiv \frac{e^{-\beta H_{\mathrm{DAM}}(\boldsymbol{\sigma},\boldsymbol{\tau}|\boldsymbol{\xi})}}{Z},$$

where the last equation tacitely defines the cost function of the DAM, namely

$$\begin{aligned} H_{\mathrm{DAM}}(\boldsymbol{\sigma},\boldsymbol{\tau}|\boldsymbol{\xi}) &:= -\frac{1}{2N^3}\sum_{\rho=1}^{K}\left(\sum_{i,\mu=1}^{N,N}\xi_{i\mu}^{\rho}\sigma_i\tau_\mu\right)^2 \\ &= -\frac{1}{2N^3}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{N,N} J_{i\mu}^{j\nu}\sigma_i\sigma_j\tau_\mu\tau_\nu, \end{aligned} \tag{2.102}$$

where $J_{i\mu}^{j\nu} = \left(\sum_\rho \xi_{i\mu}^{\rho}\xi_{j\nu}^{\rho}\right)$. This decomposition shows that the $\boldsymbol{\xi}$'s play as eigenvectors for the tensor $\boldsymbol{J}$, whose symmetry with respect to an exhange of indices (i,μ) and (j,ν) mirrors the symmetry between the $\boldsymbol{\sigma}$ and the $\boldsymbol{\tau}$ variables underlying the learning rule (2.100). Notice that H_{DAM} corresponds to a (P=4)-wise bipartite Hopfield model (see Fig. 2.12, right panel), namely a minimal generalization of the Hebbian kernel in the classic Hopfield reference (quite similar to auto-encoders in Engineering jargon [60]). Also, this equivalence generalizes the standard duality between restricted Boltzmann machines and (pairwise) Hopfield neural networks [52, 82].
To start dealing with network's capabilities, it is convenient to introduce generalized Mattis order parameters M_ρ defined as

$$M_\rho \equiv \frac{1}{N^2}\sum_{i,\mu=1}^{N,N}\xi_{i\mu}^{\rho}\sigma_i\tau_\mu. \tag{2.103}$$

The signal-to-noise analysis for this system can be obtained by requiring the dynamic stability of the neural state recalling, without loss of generality, the pattern $\rho = 1$, that is,

$\sigma_i\tau_\mu = \xi^1_{i\mu}$. Therefore, denoting with $h_{i\mu}$ the internal field acting on σ_i and τ_μ we get

$$\begin{aligned}\sigma_i\tau_\mu h_{i\mu} &= \mathcal{S} + \mathcal{N} = \frac{1}{2N}\sum_{\rho=1}^{K} M_\rho \xi^\rho_{i\mu}\xi^1_{i\mu} \\ &= \frac{1}{2N}\left[M_1 + \sum_{\rho>1}^{K} M_\rho \xi^\rho_{i\mu}\xi^1_{i\mu}\right]. \end{aligned} \tag{2.104}$$

As the signal term inside the brackets in (2.104) is $M_1 \sim \mathcal{O}(1)$, while the noise term corresponds to a sum of $(K-1)$ stochastic and uncorrelated contributions, each of order $\mathcal{O}(N^{-1})$, exploiting the central limit theorem it is immediate to check that the quenched noise due to non-retrieved patterns can be amplified by a factor $\sqrt{N}$ still preserving the stability condition $\mathcal{S}/\mathcal{N} \sim \mathcal{O}(1)$. We can therefore introduce noisy patterns yielding to the noisy tensor $\boldsymbol{\eta}$ with entries

$$\eta^\rho_{i\mu} \equiv \xi^\rho_{i\mu} + \sqrt{K}\tilde{\xi}^\rho_{i\mu}, \tag{2.105}$$

where the information is carried by the Boolean entries of $\xi^\rho_{i\mu}$, while the noise is coded in the real $\tilde{\xi}^\rho_{i\mu}$ that are i.i.d. standard Gaussian variables for $i, \mu = 1, ..., N$ and $\rho = 1, ..., K$. Notice that the information encoded by the patterns is perturbed by adding a stochastic term $\tilde{\boldsymbol{\xi}}$ on $\boldsymbol{\xi}$ (eq.2.105) rather than directly on $\boldsymbol{J}$; the latter choice would have a lower impact on network capacity and is therefore less challenging. In analogy with (2.103) we also define

$$\tilde{M}_\rho \equiv \frac{1}{N^2}\sum_{i,\mu=1}^{N,N} \tilde{\xi}^\rho_{i\mu}\sigma_i\tau_\mu. \tag{2.106}$$

Replacing the Boolean tensor (2.99) in eq. (2.102) with the noisy tensor (2.105) and exploiting the definitions (2.103) and (2.128), we get $H_{\mathrm{DAM}} = -\frac{N}{2}\sum_\rho (M_\rho + \sqrt{K}\tilde{M}_\rho)^2$. Then, in the limit of large N, splitting the signal and the noise contributions, the Boltzmann factor in eq. (2.101) reads as (see Section 2.3.4 for all the details in the statistical mechanical treatment of this network)

$$\exp\left(-\beta H_{\mathrm{RSM}}\right) \underset{N\to\infty}{\sim} \exp\left(\beta\frac{N}{2}M_1^2 + \beta\frac{\alpha N^2}{2}\sum_{\rho\geq 2}^{K}\tilde{M}_\rho^2\right).$$

Let us now handle the two terms appearing as argument of the exponential in the r.h.s.: exploiting the redundancy $\xi^1_{i\mu} = \xi^1_i\xi^1_\mu$ and calling m_σ and m_τ the Mattis magnetization related to the visible layer $\boldsymbol{\sigma}$ and to the mirror layer $\boldsymbol{\tau}$ respectively, we get $M_1 = \left(\frac{1}{N}\sum_i \xi^1_i\sigma_i\right)\left(\frac{1}{N}\sum_\mu \xi^1_\mu\tau_\mu\right) \equiv m_\sigma m_\tau$, in such a way that $\beta N M_1^2/2 = \beta N m_\sigma^2 m_\tau^2/2$; by performing a Hubbard-Stratonovich transformation, the quenched noise given by the non-retrieved $K-1$ patterns is linearized as $\sqrt{\alpha\beta}\sum_{i,\mu,\rho\geq 2}\tilde{\xi}^\rho_{i\mu}\sigma_i\tau_\mu z_\rho/N$. After these passages one can address the evaluation of the intensive quenched pressure of the model, defined as,

$$A(\alpha,\beta) := \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\boldsymbol{\eta}}\ln\sum_{\boldsymbol{\sigma},\boldsymbol{\tau}}\int\prod_{\rho=1}^{K} Dz_\rho \exp\left(-\beta H_{\mathrm{RSM}}\right),$$

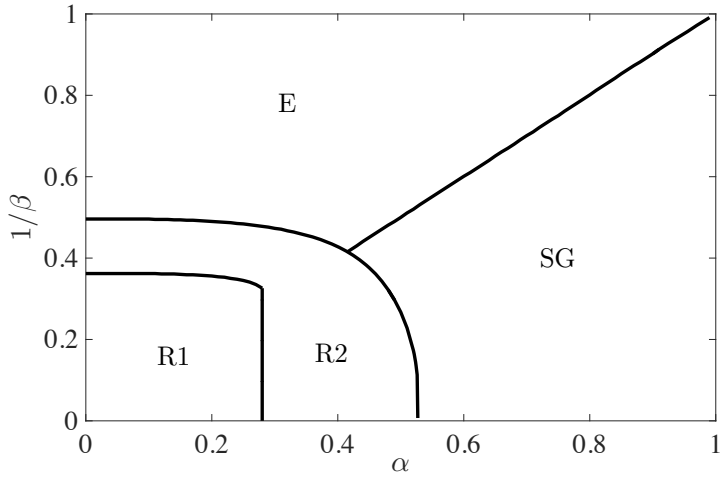

Figure 2.13: Phase diagram for the DAM with $(P = 4)$-wise interactions among the N neurons and a load $K = \alpha N$, as a function of the capacity α and of the the noise level $1/\beta$. This diagram was obtained by solving the self-consistent equations (2.108)-(2.112) and by identifying the retrieval region as the region where each neural configurations corresponding to the stored patterns (and their symmetric version) is a maximum of the pressure – either global, (R1) or local (R2) – the spin-glass (SG) region as the region where retrieval capabilities are lost due to prevailing "slow noise" α, and the ergodic (E) region as the region where retrieval capabilities are lost due to prevailing "fast noise" $1/\beta$ (see Section 2.3.4 for further details).

exploiting Guerra's interpolation techniques [26, 52]. Under the Replica Symmetric (RS) ansatz, the quenched pressure reads as (see Section 2.3.4 for technical details in the statistical mechanical treatment)

$$\begin{aligned} A^{\mathrm{RS}} =& 2\ln 2 + \frac{\alpha^2\beta^2}{2} p(2qr - r - q) - \frac{3}{2}\beta \overline{m}_\sigma^2 \overline{m}_\tau^2 \\ &+ \int Dx \ln\cosh\left(\alpha\beta x\sqrt{rp} + \beta \overline{m}_\sigma \overline{m}_\tau^2\right) \\ &+ \int Dx \ln\cosh\left(\alpha\beta x\sqrt{qp} + \beta \overline{m}_\sigma^2 \overline{m}_\tau\right) \\ &- \frac{\alpha}{2}\ln[1 - \alpha\beta(1 - qr)] + \frac{\alpha^2\beta}{2}\frac{qr}{1 - \alpha\beta(1 - qr)}, \end{aligned} \tag{2.107}$$

where $\overline{m}_\sigma$ and $\overline{m}_\tau$ are the RS values of the Mattis magnetizations, while q, p and r are the RS values for the two-replica overlaps for each layer (visible, hidden and mirror respectively). Its extremization returns the following self-consistency equations for the

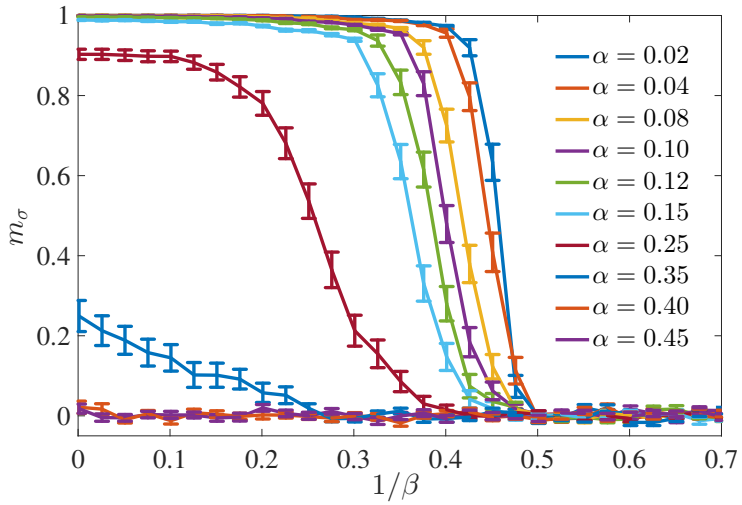

Figure 2.14: Expected Mattis magnetization obtained from Monte Carlo simulations run for $N = 150$ and for different values of α, as a function of $1/\beta$
. Notice that, as α is tuned from 0.25 to 0.35, the magnetization abruptly drops even at small values of $1/\beta$, consistently with the transition from the region R1 to the region R2 found theoretically (see Fig. 2.13 and Section 2.3.6 for further details and discussions).

order parameters

$$q = \int Dx \tanh^2\left(\alpha\beta\sqrt{rp}x + \beta\overline{m}_\sigma\overline{m}_\tau^2\right), \tag{2.108}$$

$$r = \int Dx \tanh^2\left(\alpha\beta\sqrt{qp}x + \beta\overline{m}_\sigma^2\overline{m}_\tau\right), \tag{2.109}$$

$$p = \frac{\alpha qr}{\left[1-\alpha\beta(1-qr)\right]^2}, \tag{2.110}$$

$$\overline{m}_\sigma = \int Dx \tanh\left(\alpha\beta\sqrt{rp}x + \beta\overline{m}_\sigma\overline{m}_\tau^2\right), \tag{2.111}$$

$$\overline{m}_\tau = \int Dx \tanh\left(\alpha\beta\sqrt{qp}x + \beta\overline{m}_\sigma^2\overline{m}_\tau\right). \tag{2.112}$$

whose solution paints the phase diagram in Fig. 2.13 (see the Appendix for more details).

The theory is also corroborated via Monte Carlo simulations; a sample of this analysis is shown in Fig. 2.14, while more extensive discussions can be found in Section 2.3.6.

To summarise, we considered a Sejnowski machine equipped with two visible layers and we showed that it can perform pattern-redundant representation via a suitable generalization of the standard contrastive divergence. Further, we proved that this machine has a dual representation in terms of a bipartite DAM in such a way that the features learnt by the former correspond to the patterns stored in the latter and, whatever the learning mode (adaptive versus Hebbian), in the operational mode these networks achieve pattern recognition always in a Hebbian fashion. We studied these nets via statistical mechanical tools obtaining (under the RS ansatz) a phase diagram, where their remarkable capabilities shine. In particular, there exists a region in the parameter space where they can retrieve patterns although these are (apparently) overpowered by the noise. This may contribute to explain the high-rate ability of deep/dense networks in pattern recognition, as empirically

evidenced in a variety of tasks. Indeed, at finite volumes (as standard dealing with real data-sets), it is not obvious which regime of operation the network is actually set at: to see this one can notice that at finite N and K one has only access to the ratio $\alpha(K,N)=K/N$ which can possibly be compatible with different scalings (e.g., $K=\alpha_1 N^{P-1}$ or $K=\alpha_2 N$). Hence, we speculate that such impressive detection skills emerge when these nets are away from the memory storage saturation. Further, we have shown by a pure statistical mechanical perspective, how pattern recognition power and memory storage are strongly related. For the sake of completeness, we report that also in the purely engineering counterpart, pattern redundancy is exploited to cope with high noise rate (e.g., in white Gaussian additive channels [83, 84]). In particular, our approach is close to the so called *channel access method* in telecommunications, namely a set-up where more than two terminals connected to the same transmission medium are allowed to share its capacity.

2.3.2 A new Contrastive-Divergence learning rule

While the reward in the paradigm shift from the pairwise Hamitonian toward deep or dense networks is expected to be huge (and far from being entirely explored at present, *ultra-detection* being just one out of several aspects of the increased performances of these modern architectures), as a matter of fact all the existing celebrated learning rules - originally derived within the *statistical reductionism* framework- must now sutably be enlarged to work also for these networks. Aim of this section is to generalize the classical contrastive divergence scheme -introduced by Ackley, Hinton and Sejnowski to deal with the standard, pairwise, RBM in order to work also for this generalization of the RBM machine introduced by Sejnowski.
Let us recall the partition function (4) of the Sejnowski machine we are inspecting

$$Z=\sum_{\boldsymbol{\sigma},\boldsymbol{\tau}}\int\Big(\prod_{\rho=1}^{K}\frac{dz_\rho}{\sqrt{2\pi\beta^{-1}}}\Big)\times \\ \times\exp\Big(-\frac{\beta}{2}\sum_{\rho=1}^{K}z_\rho^2+\beta\sum_{i,\mu,\rho=1}^{N,N,K}\hat{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\Big), \tag{2.113}$$

where $\hat{\xi}_{i\mu}^{\rho}=N^{-3/2}\xi_{i\mu}^{\rho}$. Such expression suggests that learning should act on the couplings $\xi_{i\mu}^{\rho}$ rather than on the information patterns ξ_i^ρ (as it happens in the simpler pairwise scenario [48, 52]). In order for the learning procedure to create free-energy (we recall that the pressure A is simply related to the free energy F by $A=-\beta F$, in such a way that the two functions exhibit the same extreme points, yet maxima in the pressure just corresponds to minima in the free-energy) minima placed at $\boldsymbol{\sigma}=\boldsymbol{\tau}=\boldsymbol{\xi}^\rho$, both the visible and mirror layers should be set according to the data, namely the two *eyes* of the machine do look at the same outside world.
In this Section, we prove that the learning rule reads as

$$\Delta\hat{\xi}_{i\mu}^{\rho}=\epsilon\beta\left(\langle\sigma_i\tau_\mu z_\rho\rangle_+-\langle\sigma_i\tau_\mu z_\rho\rangle_-\right), \tag{2.114}$$

where the subscript "+" means that both visible and mirror layers are set at the data input (i.e., they are *clamped*), while the subscript "−" means that all neurons in the network are left free to evolve. Let us write explicitly the probability distribution for a given

configuration state:

$$P(\boldsymbol{\sigma},\boldsymbol{z},\boldsymbol{\tau}) = Z^{-1}\left(\frac{1}{\sqrt{2\pi\beta^{-1}}}\right)^{K}\times \\ \times\exp\Big(-\frac{\beta}{2}\sum_{\rho=1}^{K} z_\rho^2+\beta\sum_{i,\mu,\rho=1}^{N,N,K}\hat{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\Big), \tag{2.115}$$

and suppose the set of data is made of i.i.d. entries generated by a probability distribution $Q(\boldsymbol{\sigma})$, whose features we aim to extract.
Since the mirror layer, by definition, should mimic the activity of the visible layer (as we want to put the information content in pure minima of the free energy given by configurations of the form $\boldsymbol{\tau}=\boldsymbol{\sigma}$), we have to build a representation of the couplings $\hat{\xi}_{i\mu}^{\rho}$ such that the marginal distribution $P(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})$ is the best approximation for $Q(\boldsymbol{\sigma})$, where

$$P(\boldsymbol{\sigma},\boldsymbol{\tau}) = Z^{-1}\int\Big(\prod_{\rho=1}^{K}\frac{dz_\rho}{\sqrt{2\pi\beta^{-1}}}\Big)\times \\ \times\exp\Big(-\frac{\beta}{2}\sum_{\rho=1}^{K} z_\rho^2+\beta\sum_{i,\mu,\rho=1}^{N,N,K}\hat{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\Big) \\ := \frac{Z(\boldsymbol{\sigma},\boldsymbol{\tau})}{Z}. \tag{2.116}$$

Therefore, we introduce the Kullback-Leibler cross-entropy as

$$D(P,Q)=\sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma})\log\frac{Q(\boldsymbol{\sigma})}{P(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})}.$$

Under a gradient-descent approach, we have to compute the derivative of the cross-entropy w.r.t. the couplings, that reads as

$$\frac{\partial D}{\partial\hat{\xi}_{i\mu}^{\rho}} = -\sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma})\times \\ \times\left[Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})^{-1}\frac{\partial Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})}{\partial\hat{\xi}_{i\mu}^{\rho}} - Z^{-1}\frac{\partial Z}{\partial\hat{\xi}_{i\mu}^{\rho}}\right]. \tag{2.117}$$

The first term in the square brackets of eq. (2.117) can be written as:

$$Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})^{-1}\frac{\partial Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})}{\partial\hat{\xi}_{i\mu}^{\rho}} = \\ = Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})^{-1}\int\Big(\prod_{\rho=1}^{K}\frac{dz_\rho}{\sqrt{2\pi\beta^{-1}}}\Big)\beta\sigma_i\tau_\mu z_\rho\times \\ \times\exp\Big(-\frac{\beta}{2}\sum_{\rho=1}^{K} z_\rho^2+\beta\sum_{i,\mu,\rho=1}^{N,N,K}\hat{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\Big) = \\ = Z(\boldsymbol{\sigma},\boldsymbol{\tau}=\boldsymbol{\sigma})^{-1}\int\Big(\prod_{\rho=1}^{K}dz_\rho\Big)\beta\sigma_i\tau_\mu z_\rho\ P(\boldsymbol{\sigma},\boldsymbol{z},\boldsymbol{\tau}=\boldsymbol{\sigma}), \tag{2.118}$$

and, using Bayes' theorem under the constraint $\boldsymbol{\tau} = \boldsymbol{\sigma}$,

$$P(\boldsymbol{\sigma}, \boldsymbol{z}, \boldsymbol{\tau} = \boldsymbol{\sigma}) = P(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma}) P(\boldsymbol{z}|\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma}) = \\ = \frac{Z(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma})}{Z} P(\boldsymbol{z}|\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma}). \tag{2.119}$$

Therefore, combining (2.118) and (2.119)

$$Z(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma})^{-1} \frac{\partial Z(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma})}{\partial \hat{\xi}^{\rho}_{i\mu}} = \\ = \int \Big(\prod_{\rho=1}^{K} dz_\rho \Big) \beta \sigma_i \tau_\mu z_\rho P(\boldsymbol{z}|\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma}). \tag{2.120}$$

When taking the Q-weighted sum, we have

$$\sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma}) Z(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma})^{-1} \frac{\partial Z(\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma})}{\partial \hat{\xi}^{\rho}_{i\mu}} = \\ = \sum_{\boldsymbol{\sigma}} \int \Big(\prod_{\rho=1}^{K} dz_\rho \Big) \beta \sigma_i \tau_\mu z_\rho Q(\boldsymbol{\sigma}) P(\boldsymbol{z}|\boldsymbol{\sigma}, \boldsymbol{\tau} = \boldsymbol{\sigma}) = \\ = \beta \langle \sigma_i \tau_\mu z_\rho \rangle_+, \tag{2.121}$$

since data are extracted with probability $Q(\boldsymbol{\sigma})$.
The second term in the square brackets of eq. (2.117) can be written as:

$$Z^{-1} \frac{\partial Z}{\partial \hat{\xi}^{\rho}_{i\mu}} = \sum_{\boldsymbol{\sigma}', \boldsymbol{\tau}} \int \Big(\prod_{\rho=1}^{K} dz_\rho \Big) \beta \sigma'_i \tau_\mu z_\rho \; P(\boldsymbol{\sigma}', \boldsymbol{z}, \boldsymbol{\tau}), \tag{2.122}$$

where now there are no constraints on the mirror and visible layers. Since there is no dependence on $\boldsymbol{\sigma}$, the Q-weighted sum can be trivially performed, as
$\sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma}) = 1$, leading to

$$\sum_{\boldsymbol{\sigma}} Q(\boldsymbol{\sigma}) Z^{-1} \frac{\partial Z}{\partial \hat{\xi}^{\rho}_{i\mu}} = \\ = \sum_{\boldsymbol{\sigma}', \boldsymbol{\tau}} \int \Big(\prod_{\rho=1}^{K} dz_\rho \Big) \beta \sigma'_i \tau_\mu z_\rho \; P(\boldsymbol{\sigma}', \boldsymbol{z}, \boldsymbol{\tau}) = \\ = \beta \langle \sigma_i \tau_\mu z_\rho \rangle_-. \tag{2.123}$$

All together, we have

$$\frac{\partial D}{\partial \hat{\xi}^{\rho}_{i\mu}} = -\beta (\langle \sigma_i \tau_\mu z_\rho \rangle_+ - \langle \sigma_i \tau_\mu z_\rho \rangle_-). \tag{2.124}$$

The gradient descent rule (2.114) can therefore be expressed in a contrastive divergence (CD) form as $\Delta \hat{\xi}^{\rho}_{i\mu} = -\epsilon \frac{\partial D}{\partial \hat{\xi}^{\rho}_{i\mu}}$.
In order to check the performance of this network we proceed as follows: we consider the Restricted Sejnowski Machine (RSM) and, for comparison, a standard Restricted Boltzmann Machine (RBM) and, for both the networks, we arbitrarily choose two random configurations $(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$ to be the patterns to be learnt. Via Gibbs-sampling we generate

a training set (the same for both the networks) by producing corrupted versions of these patterns (with a level of corruption up to 30%). The latter are thus learnt simultaneously via CD and, once the training stage is over, pattern retrieval is further examined. The overlaps $m_{1,2}$ are measured and compared in the training and in the validation stages. In all the tests we performed – a sample of which is shown in Fig. 2.15 – the RSM outperforms the standard RBM (all the tests produced results similar to those reported in Fig. 2.15). In particular, beyond being more accurate, the CD-algorithm for the RSM is significantly faster with respect to its RBM counterpart, that is, it reaches large values of $m_{1,2}$ already for a relatively small number of CD steps.
As a last remark we notice that, in the very initial stage (when the number of CD steps is small), the RBM displays a large overlap with respect to the RSM. This effect is of purely stochastic nature as the RBM is fed with a vector of N entries while the RSM is fed with a matrix of N^2 entries, in such a way that a random initial configuration will exhibit a larger alignment in the former case. This remark further highlights the higher speed of the RSM.

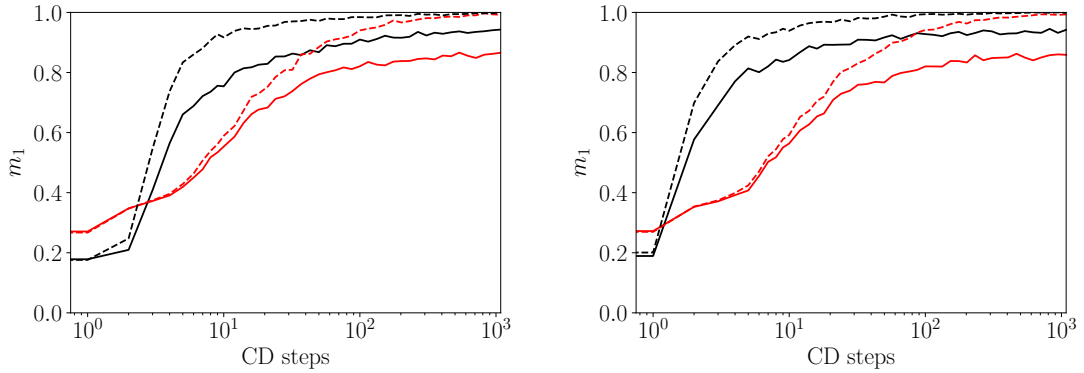

Figure 2.15: The two plots show a comparison between learning performances of a RSM (black lines) and RBM (red lines). Dashed lines are for comparison of the performances of the machines during the training stage while solid lines are for comparison during the validation stage. On the horizontal axes, we report the number of CD-steps while on the vertical axes we show the overlap between the visible layer $\boldsymbol{\sigma}$ and the retrieved test-pattern $\boldsymbol{\xi}^1$ (i.e., the magnetization m_1). In the left plot the two networks work with the optimal learning rate for the RBM (evaluated as $\epsilon = 0.266$), nonetheless the RSM outperforms the RBM both in the training and in the validation stages. In the right plot, the two networks operate with their respective optimal learning rates (that is $\epsilon_{\text{RBM}} = 0.266$ and $\epsilon_{\text{RSM}} = 0.52$) and the difference in the performances is further enhanced. In both cases, the network size is fixed to $N = 20$ (however, we obtained analogous results up to $N = 200$). Similar results also hold for the overlap with pattern $\boldsymbol{\xi}^2$.

2.3.3 Signal-to-noise stability analysis

In this Section we perform a signal to noise analysis [11] for the Dense Associative Memory (DAM) with $(P = 4)$-wise interactions among spins and in the linear storage regime $K = \alpha N, \alpha > 0$. Its Hamiltonian, or *cost function* to keep a Machine Learning

jargon, appearing in eq. (5) in the main text, can be rewritten as

$$H_{\rm DAM} = -\sum_{i,\mu=1}^{N,N} h_{i\mu}\sigma_i\tau_\mu, \tag{2.125}$$

where

$$h_{i\mu} = \frac{1}{2N^3}\sum_{j,\nu=1}^{N,N}\sum_{\rho=1}^{K}\eta_{i\mu}^{\rho}\eta_{j\nu}^{\rho}\sigma_j\tau_\nu, \tag{2.126}$$

are the internal fields acting on the dimer $\sigma_i\tau_\mu$.
The tensors $\boldsymbol{\eta}$ (see also eq.2.105) in the main text and [85]) are expressed as the sum of a Boolean contribution $\boldsymbol{\xi}$ providing the signal and a real contribution $\tilde{\boldsymbol{\xi}}$ accounting for a noise source:

$$\eta_{i\mu}^{\rho} = \xi_{i\mu}^{\rho} + \sqrt{K}\tilde{\xi}_{i\mu}^{\rho} = \xi_{i\mu}^{\rho} + \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}, \tag{2.127}$$

where $\mathbb{P}(\xi_{i\mu}^{\rho} = \pm 1) = 1/2$ and $\mathbb{P}(\tilde{\xi}_{i\mu}^{\rho}) = \mathcal{N}(0,1)$ for each $i,\mu = 1,\dots,N$ and $\rho = 1,\dots,K$.
We recall the definition of the $2K$ generalized Mattis magnetizations as

$$M_\rho \;\; = \tfrac{1}{N^2}\textstyle\sum_{i,\mu=1}^{N,N}\xi_{i,\mu}^{\rho}\sigma_i\tau_\mu, \tag{2.128}$$

$$\tilde{M}_\rho \;\; = \tfrac{1}{N^2}\textstyle\sum_{i,\mu=1}^{N,N}\tilde{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu, \tag{2.129}$$

with $\rho = 1,...,K$. In terms of these overlaps, the internal fields (2.126) can be written as

$$\begin{aligned}
h_{i\mu} =& \frac{1}{2N^3}\sum_{j,\nu=1}^{N,N}\sum_{\rho=1}^{K}\left(\xi_{i\mu}^{\rho} + \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}\right)\left(\xi_{j\nu}^{\rho} + \sqrt{\alpha N}\tilde{\xi}_{j\nu}^{\rho}\right)\sigma_j\tau_\nu \\
=& \frac{1}{2N^3}\sum_{j,\nu=1}^{N,N}\sum_{\rho=1}^{K}\Big(\xi_{i\mu}^{\rho}\xi_{j\nu}^{\rho}\sigma_j\tau_\nu + \sqrt{\alpha N}\xi_{i\mu}^{\rho}\tilde{\xi}_{j\nu}^{\rho}\sigma_j\tau_\nu + \\
&+ \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}\xi_{j\nu}^{\rho}\sigma_j\tau_\nu + \alpha N\tilde{\xi}_{i\mu}^{\rho}\tilde{\xi}_{j\nu}^{\rho}\sigma_j\tau_\nu\Big) \\
=& \frac{1}{2N}\sum_{\rho=1}^{K}\Big(\xi_{i\mu}^{\rho}M_\rho + \sqrt{\alpha N}\xi_{i\mu}^{\rho}\tilde{M}_\rho + \\
&+ \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}M_\rho + \alpha N\tilde{\xi}_{i\mu}^{\rho}\tilde{M}_\rho\Big).
\end{aligned} \tag{2.130}$$

We aim to check the stability of configurations where the dimers $\sigma_i\tau_\mu$ are aligned to a given element of the tensor, say $\xi_{i\mu}^1$; the resulting contribution to the energy function (2.125) is

$$\begin{aligned}
h_{i\mu}\xi_{i\mu}^1 =& \frac{1}{2N}\sum_{\rho=1}^{K}(\xi_{i\mu}^{\rho}\xi_{i\mu}^1 M_\rho + \sqrt{\alpha N}\xi_{i\mu}^{\rho}\xi_{i\mu}^1\tilde{M}_\rho + \\
&+ \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}\xi_{i\mu}^1 M_\rho + \alpha N\tilde{\xi}_{i\mu}^{\rho}\xi_{i\mu}^1\tilde{M}_\rho) \\
=& \frac{1}{2N}\Big[M_1 + \sqrt{\alpha N}\tilde{\xi}_{i\mu}^1\xi_{i\mu}^1 M_1 + \\
&+ \sum_{\rho=2}^{K}(\xi_{i\mu}^{\rho}\xi_{i\mu}^1 M_\rho + \sqrt{\alpha N}\tilde{\xi}_{i\mu}^{\rho}\xi_{i\mu}^1 M_\rho) + \\
&+ \sum_{\rho=1}^{K}(\sqrt{\alpha N}\xi_{i\mu}^{\rho}\xi_{i\mu}^1\tilde{M}_\rho + \alpha N\tilde{\xi}_{i\mu}^{\rho}\xi_{i\mu}^1\tilde{M}_\rho)\Big]
\end{aligned} \tag{2.131}$$

$$= (\mathcal{S} + \mathcal{N})/2N, \tag{2.132}$$

where in the first line of eq. (2.131) we used the trivial identity $(\xi^1_{i\mu})^2 = 1$ and in eq. (2.132) we split the energy contribution $h_{i\mu}\xi^1_{i\mu}$ into a signal $\mathcal{S}$ and a noise $\mathcal{N}$ term:

$$\mathcal{S} = M_1, \tag{2.133}$$

$$\begin{aligned}\mathcal{N} = \sqrt{\alpha N}\Big[\tilde{\xi}^1_{i\mu}\xi^1_{i\mu}M_1 + \sum_{\rho=2}^{K}\Big(\frac{\xi^\rho_{i\mu}\xi^1_{i\mu}M_\rho}{\sqrt{\alpha N}} + \tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}M_\rho\Big) + \\ + \sum_{\rho=1}^{K}(\xi^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho + \sqrt{\alpha N}\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho)\Big].\end{aligned} \tag{2.134}$$

We now compare the scaling behaviours of these two terms, by computing their ratio. We anticipate that this ratio depends on the realization of the noisy patterns and we should average in some way over the variables $\tilde{\xi}^\rho_{i\mu}$. If not interested in the magnitude of fluctuations (mirroring the statistical mechanical side, where the model is kept mean-field and analyzed at the replica symmetric level), one can simply consider the ratio $\mathbb{E}_{\tilde{\xi}}(\mathcal{S})/\mathbb{E}_{\tilde{\xi}}(\mathcal{N})$, where $\mathbb{E}_{\tilde{\xi}}(\cdot)$ is the average over the internal noise realizations $\tilde{\xi}^\rho_{i\mu}$. In this way, fluctuations are averaged out and we are only left with the magnitudes of the first moment. Let us now turn to the evaluation of the scaling behaviours of $\mathcal{S}$ and $\mathcal{N}$ in (2.133) and (2.134), respectively.
First, under the perfect retrieval hypothesis, we have $M_1 = 1$, whence

$$\mathcal{S} = M_1 = 1. \tag{2.135}$$

As for $\mathcal{N}$, we can preliminary notice that, among its five contributions appearing in (2.134), the second term $\sum_{\rho>1}\xi^\rho_{i\mu}\xi^1_{i\mu}M_\rho$ can be neglected as it is vanishing as $\mathcal{O}(N^{-1/2})$ in the thermodynamic limit. This can be seen by expanding the magnetizations, that is

$$\begin{aligned}\sum_{\rho=2}^{K}\xi^\rho_{i\mu}\xi^1_{i\mu}M_\rho &= \frac{1}{N^2}\sum_{\rho=2}^{K}\sum_{j,\nu=1}^{N,N}\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{j\nu}\xi^1_{j\nu} = \\ &= \frac{1}{N^2}\sum_{\rho=2}^{K}\Big(\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{i\mu}\xi^1_{i\mu} + \sum_{\substack{\nu=1\\ \nu\neq\mu}}^{N}\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{i\nu}\xi^1_{i\nu} + \\ &\qquad + \sum_{\substack{\nu,j=1\\ j\neq i}}^{N,N}\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{j\nu}\xi^1_{j\nu}\Big) \\ &= \frac{1}{N^2}\Big[\sum_{\rho=2}^{K}(\xi^\rho_{i\mu})^2(\xi^1_{i\mu})^2 + \sum_{\rho=2}^{K}\sum_{\substack{\nu=1\\ \nu\neq\mu}}^{N}\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{i\nu}\xi^1_{i\nu} + \\ &\qquad + \sum_{\rho=2}^{K}\sum_{\substack{\nu,j=1\\ j\neq i}}^{N,N}\xi^\rho_{i\mu}\xi^1_{i\mu}\xi^\rho_{j\nu}\xi^1_{j\nu}\Big],\end{aligned} \tag{2.136}$$

and checking that the first term in square brackets is of order $\mathcal{O}(N)$, due to the trivial equality $(\xi^\rho_{i\mu})^2 = 1$ and to the fact that the sum includes $K-1$ terms with $K \sim \alpha N$; the remaining two terms can be looked at as the displacement covered by simple random walkers performing, respectively, $\sim N^2$ and $\sim N^3$ steps on a linear chain, in such a way that for large enough N they are Gaussian distributed with standard deviation of order

$\mathcal{O}(N)$ and $\mathcal{O}(N^{3/2})$, respectively.
Therefore, in the large N limit, the leading contribution in the noise term (2.134) is given by

$$\mathcal{N} \underset{N\to\infty}{\sim} \sqrt{\alpha N}\tilde{\xi}^1_{i\mu}\xi^1_{i\mu}M_1 + \sum_{\rho=2}^{K}\sqrt{\alpha N}\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}M_\rho + \\ + \sum_{\rho=1}^{K}\sqrt{\alpha N}\xi^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho + \sum_{\rho=1}^{K}\alpha N\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho, \tag{2.137}$$

and, when taking the average with respect to the pattern internal noise, only the last term survives, since it is the only one with even powers of $\tilde{\xi}^\rho_{i\mu}$.
Then, focusing on the last term, we get

$$\sum_{\rho=1}^{K}\alpha N\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho = \frac{\alpha}{N}\sum_{\rho=1}^{K}\sum_{j,\nu=1}^{N,N}\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}\tilde{\xi}^\rho_{j\nu}\xi^1_{j\nu}, \tag{2.138}$$

and, introducing the variables $\tilde{\tilde{\xi}}^\rho_{i\mu} = \xi^1_{i\mu}\tilde{\xi}^\rho_{i\mu}$, which are obviously Gaussian-distributed,

$$\begin{aligned}\sum_{\rho=1}^{K}\alpha N\tilde{\xi}^\rho_{i\mu}\xi^1_{i\mu}\tilde{M}_\rho &= \frac{\alpha}{N}\sum_{\rho=1}^{K}\sum_{j,\nu=1}^{N,N}\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{j\nu} = \\ &= \frac{\alpha}{N}\sum_{\rho=1}^{K}\left(\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{i\mu} + \sum_{\substack{\nu=1\\ \nu\neq\mu}}^{N}\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{i\nu} + \sum_{\substack{\nu,j=1\\ j\neq i}}^{N,N}\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{j\nu}\right) = \\ &= \frac{\alpha}{N}\sum_{\rho=1}^{K}\left(\tilde{\tilde{\xi}}^\rho_{i\mu}\right)^2 + \frac{\alpha}{N}\sum_{\rho=1}^{K}\Big(\sum_{\substack{\nu=1\\ \nu\neq\mu}}^{N}\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{i\nu} + \sum_{\substack{\nu,j=1\\ j\neq i}}^{N,N}\tilde{\tilde{\xi}}^\rho_{i\mu}\tilde{\tilde{\xi}}^\rho_{j\nu}\Big).\end{aligned} \tag{2.139}$$

Furthermore, in the expression (2.139), only the first term in the last line gives non-vanishing contribution (since the other two terms are product of uncorrelated random variables). Therefore

$$\mathbb{E}_{\tilde{\xi}}(\mathcal{N}) = \frac{\alpha}{N}\sum_{\rho=1}^{K}\mathbb{E}_{\tilde{\xi}}(\tilde{\tilde{\xi}}^\rho_{i\mu})^2 \sim \alpha\frac{K}{N} = \alpha^2, \tag{2.140}$$

which is $\mathcal{O}(1)$, that is the same scaling behaviour of the signal M_1.

The arguments just exposed allow us to introduce the "pattern recognition power" as the maximal extent of noise that can affect the information encoded by patterns (supposed $\mathcal{O}(1)$) still allowing pattern retrieval. This is strongly related to the memory storage: if we load the network with $K \sim N^3$ patterns then the pattern recognition power is $\mathcal{O}(N^0)$, if $K \sim N^2$, then the pattern recognition power is $\mathcal{O}(N^{1/4})$, if $K \sim N^1$, then the pattern recognition power is $\mathcal{O}(N^{1/2})$, and so on. Therefore, if $K \sim N^3$ the pattern recognition power of this net is the same as the one of the standard Hopfield model in high load, but if we sacrifice pattern storage letting $K \sim N^1$, then the pattern recognition power of this net is much higher than the one of the standard Hopfield model.

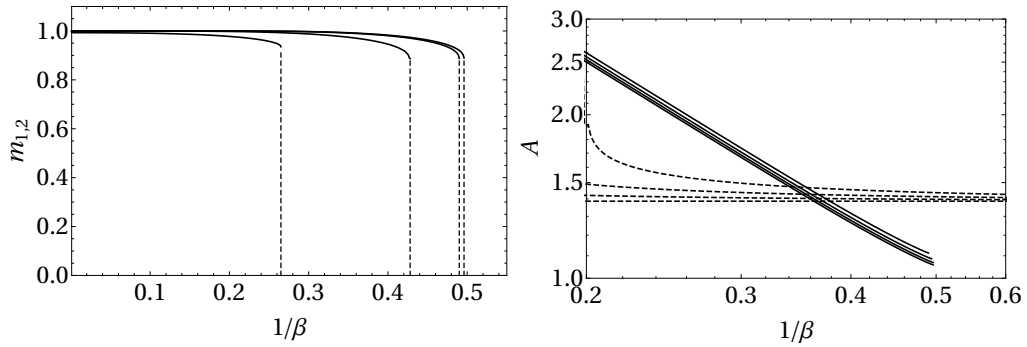

Figure 2.16: Mattis magnetization(s) and free-energy. Left: the plot shows the Mattis magnetization m (we stress that, on the self-consistency solutions, $m_1 = m_2$) as a function of the fast noise $1/\beta$ for various storage capacity values ($\alpha = 0, 0.20, 0.40, 0.50$, going from the right to the left). The vertical dotted lines indicates the jump discontinuity identifying the critical noise level $1/\beta(\alpha)$ that traces the boundary between the retrieval region and the pure spin-glass phase. Right: the plot shows the corresponding pressure as a function of the fast noise level $1/\beta$ at the storage capacity values $\alpha = 0, 0.10, 0.15, 0.20$ (going from the bottom to the top) in the retrieval (continuous lines) and spin-glass (dotted lines) states. Note that the sampled α-values are different among the two plots for a matter of best visualization (for too low values of α all the magnetizations heavily overlap and it is hard to distinguish them by eye inspection). Note: the solutions always share the symmetry $m_1 = m_2$.

2.3.4 Replica symmetric phase diagram

The phase diagram shown in Figure 2.13 exhibits four qualitatively different phases as explained hereafter:

- Ergodic phase (E)
  The "fast" noise $1/\beta$ in the system is too strong for the neurons to reciprocally feel each other, in such a way that they tend to behave randomly and no emergent collective property can be appreciated. In this region, the solution of the self-consistency equations [i.e., eqs. (11)-(15) in the main text] which maximizes the pressure [i.e., eq. (10) in the main text] is given by $m = 0,\ q = 0$.

- Spin-glass phase (SG)
  The "slow" noise α is too large for the neurons to correctly handle the whole set of patterns, and again the system fails to retrieve information, although the thermalized configurations are not purely random. In this region, the solution of the self-consistency equations which maximizes the pressure is given by $m = 0,\ q \neq 0$.

- Retrieval phase (R1)
  Both "fast" and "slow" noise are small enough for neural collective capabilities to spontaneously appear. In particular, the most likely configurations, namely those corresponding to the global maxima of the pressure, are those corresponding to stored patterns. In this region the solution of the self-consistent equations which maximizes the pressure is therefore given by $m \neq 0,\ q \neq 0$.

- Retrieval phase (R2)
  Both "fast" and "slow" noise are still relatively small hence neural collective capabil-

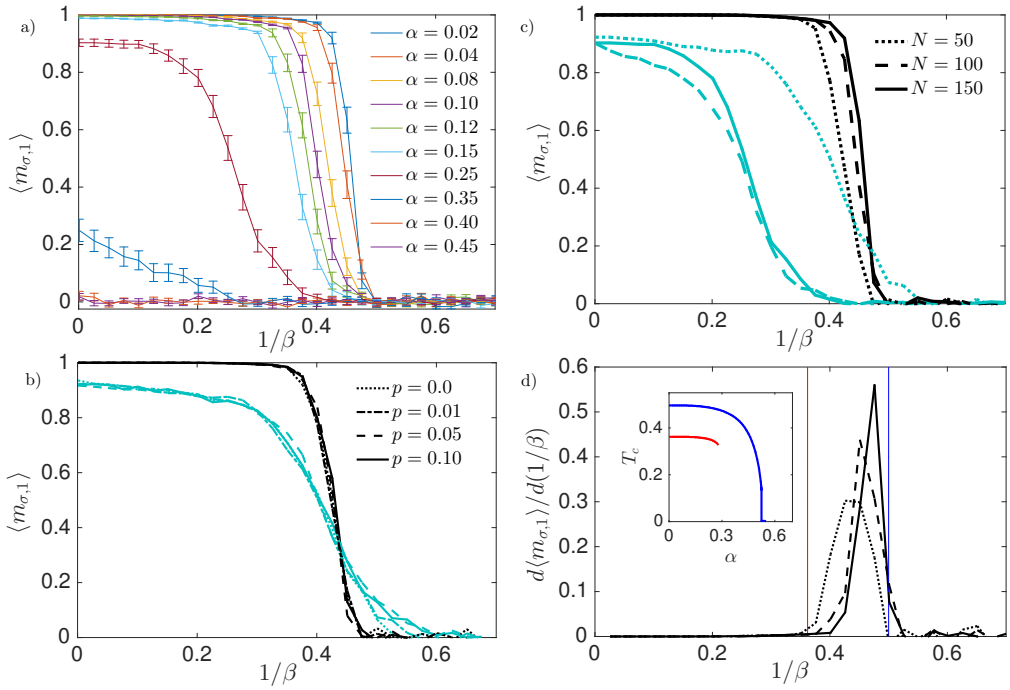

Figure 2.17: This figure shows results obtained through Monte Carlo simulations. Seeking for clarity, only $\langle m_{\sigma,\rho} \rangle$ is shown, but quantitatively analogous values are obtained also for $\langle m_{\tau,\rho} \rangle$. Errorbars (reported only in panel a, seeking for clarity) stem from the average of thermal noise and quenched noise. All cases depicted here correspond to $Q = 100$ realizations. Panel a: Expected Mattis magnetization $\langle m_{\sigma,\rho} \rangle$ for $N = 150$ and $p = 0.01$ as a function of $1/\beta$ and for different values of α as explained by the legend. Panel b: Comparison of the expected Mattis magnetization $\langle m_{\sigma,\rho} \rangle$ for different initial configurations $p = 0.01, 0.05, 0.1$ (depicted in different line style as explained by the legend), for fixed $N = 150$ and for $\alpha = 0.02$ (dark curves) and $\alpha = 0.25$ (bright curves). Panel c: Comparison of the expected Mattis magnetization $\langle m_{\sigma,\rho} \rangle$ for different sizes $N = 50, 100, 150$ (depicted in different line style as explained by the legend), for fixed $p = 0.01$ and for $\alpha = 0.02$ (dark curves) and $\alpha = 0.25$ (bright curves). Panel d: The critical noise T_c is estimated by taking the discrete derivative of the expected Mattis magnetization $\langle m_{\sigma,\rho} \rangle$ with respect to the noise and by selecting the value of noise $1/\beta$ (if any) where the derivative peaks. Such estimates are obtained for $N = 50, 100, 150$ (same legend as panel c) and for $p = 0.01$, $\alpha = 0.02$; the corresponding theoretical values are recalled in the inset.

ities can still spontaneously appear. However, here configurations corresponding to stored patterns are only local maxima of the pressure in such a way that patterns can be retrieved as far as the initialization of the system is not too far (in the sense of the Hamming distance) with respect to the target pattern. In this region the self-consistent equations admit as solution $m \neq 0,\ q \neq 0$ as well as $m = 0, q \neq 0$, both corresponding to maxima of the pressure, the former being local maxima, the latter being global ones.

A sketch of the analysis underlying the definition of the various regions is provided in Fig. 2.16, while a numerical confirmation via Monte Carlo runs is shown in Figure 2.17 and discussed hereafter.
We performed also Monte Carlo simulations to mimic the evolution of a finite-size DAM network made of N neurons interacting ($P = 4$)-wisely and $K = \alpha N$ patterns, described by the cost function (3.163): see Figure 2.17. The reason behind the need of numerical simulations does not lie in the necessity to check heuristic derivations within our treatment as we worked out the whole theory under the Guerra's schemes that are mathematically sound, rather to confirm that the effect of replica symmetry breaking is rather mild on these neural networks (as we worked out the whole theory under the replica symmetric ansatz).
We first fixed the parameters (N, α, β) where K has to be meant as the integer part of αN. Then, we drew the i.i.d. Boolean variables ξ_i^ρ, with $i = 1, ..., N$ and $\rho = 1, ..., K$ as well as the related Gaussian noise $\tilde{\xi}_{i\mu}^\rho$ with $i, \mu = 1, ..., N$ and $\rho = 1, ..., K$. Then, the tensor $\boldsymbol{\eta}$ is built following the prescription (3.164). Next, we initialize the system configuration in such a way that $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are aligned with $\boldsymbol{\xi}^1$, except for a fraction p of misaligned entries, and we let the system evolve by a single spin-flip Glauber dynamics. Once the equilibrium state is reached, we collect data for the instantaneous Mattis magnetizations $m_{\sigma,\rho} = \sum_i \xi_i^\rho \sigma_i / N$ and $m_{\tau,\rho} = \sum_\mu \xi_\mu^\rho \tau_i / N$ to obtain the thermal average referred to as $\langle m_{\sigma,\rho} \rangle$ and $\langle m_{\tau,\rho} \rangle$, with $\rho = 1, ..., K$ (notice that, initially, one has $m_{\sigma,1} = m_{\tau,1} = 1 - 2p$, while $m_{\sigma,\rho\neq 1}, m_{\tau,\rho\neq 1} \approx 0$). This is repeated for $Q = 100$ different realizations of the patterns $\boldsymbol{\xi}$ and the noise $\tilde{\boldsymbol{\xi}}$, over which thermal averages are accordingly averaged. The resulting values provide our numerical estimate for the expectation of the Mattis magnetizations $\langle m_{\sigma,\rho} \rangle$ and $\langle m_{\tau,\rho} \rangle$ to be compared with the solution of the self-consistent equations. Different parameters (N, α, β, p) are considered and, for each choice, the same procedure applies. A sample of our results for $N = 150, p = 0.01$ and different values of α, β is shown in Fig.2.17a, where one can check that the Mattis magnetization $m_{\sigma,1}$ corresponding to the retrieved pattern $\boldsymbol{\xi}^1$ vanishes at large values of the noise$T \equiv 1/\beta$ and/or at large values of α; as expected from the theoretical analysis, the larger α and the smaller the critical temperature T_c above which no retrieval takes place. We also notice that for $T = 0$, when α increases beyond $\alpha \approx 0.3$ the magnetization abruptly vanishes. Then, in Fig.2.17b we compare results stemming from different choices of p and for $\alpha = 0.02$ (dark curves) and $\alpha = 0.25$ (bright curves): the initial configuration (as long as close enough to $\boldsymbol{\xi}^1$) does not influence quantitatively the final outcome. Next, in Fig. 2.17c we perform a finite-size-scaling considering, again, $\alpha = 0.02$ (dark curves) and $\alpha = 0.25$ (bright curves): the curves for $N = 50$, $N = 100$ and $N = 150$ are slightly shifted and the shift gets more significant as α is increased. Finally, in Fig. 2.17d, main plot, we show the numerical derivative of $m_{\sigma,1}$ for $\alpha = 0.02$ and for the three sizes analyzed before: the peak in the derivative can be used to estimate T_c and this can in turn be compared with the theoretical results highlighted by the vertical lines and recalled in the inset.

2.3.5 Discussion on *ultra-detection* as an emergent skill

Summarizing alsomresults regarding this extension from the bare Hopfield reference, we have shown that -if we abandon the pairwise interactions (quite familiar to Physicists, as tacitely underlie linear forces -where superposition principle applies- and thus the whole, partially obsolete, reductionistic scaffold of Hard Science)- rewards are huge and far from trivial: for instance, by generalizing the restricted Boltzmann machine toward a network with two input layers (hence accounting for redundant information), the emergent properties of such an architecture are those of a dense Hopfield network whose neurons interact in cliques made of four neurons per time: this architecture significantly outperform w.r.t. the Hopfield scenario: for instance, without changing the signal-to-noise ratio for signal detection (keeping the standard one that holds for the Hopfield reference), such a network can store an amount of patterns that scales as $\propto N^3$ (while for the Hopfield network this scaling is just linear, e.g. $K = \alpha N$), further, as shown by the new contrastive divergence learning scheme (see equation 2.203) -and as deepened in the relative section (see Section 2.3.2)- the network can infer more than solely single- and pair-wise correlation functions hidden in the datasets: this is mandatory at work with structured datasets in real-world problems in machine learning[1].
Last but not least, a remarkably property that these dense networks share is that they can sacrifice storage space in order to lower their trheshold for signal detection: for instance, in the analyzed case, the network -that in principle can store $O(N^3)$ patterns (whose signal-to-noise ratio for detection must however be $O(1)$ int hat case)- can retain *solely* $O(N^1)$ patterns (that in the thermodynamic limiti it is still a diverging amount of information), but -in this *low-storage* regime- the network is able to detect a pattern shining with an intensity of $O(1)$ even if floating in a sea of noise whose intensity is $O(\sqrt{N})$: the mathematical mechanism that we proved in this thesis lies at the core of the remarkable modern patterrn recognition scores that these networks are collecting (e.g. they outpeform w.r.t. experienced doctors in melanoma recnognition on skin and a plethora of similar key problems and are indeed expected to play a pivotal role in the Personalized Medicine that is approaching and expected to revoluzionize Healthcare withihn our modern societies at the international level.
Indeed, it is exactly to this point that the remaining chapter of the manuscript is due to: far from being solely a thesis in Theoretical Artificial Intelligence, in the next pages a summarize about how our group tackled two biological problems is presented. These biological problems are extremely actual in modern Healthcare, the former dealing with understanding cancerogenesis, the latter dealing with pathologies related to hearth: let us deepen these findings!

[1] Indeed, this is one of the main reasons why -in concrete problems- if we keep the minimal architecture of the RBM we must stuck one on the top of the other, toward *deep learning architectures*, in order to infer broadly generated correlations contained in the inspected datasets.

Chapter 3

Part 3: Applications in Biological Complexity

From now on, I will report on results I obtained at work in the laboratories dealing with health-related problems in Biological Complexity.
The underlying motivation that prompted me to add this third part to the thesis (beyond my genuine interest in Biological Sciences) is essentially the will to show also experimental examples where it is possible to appreciate the benefit in mastering automatized high-dimensional statistical inference modi operandi for complex systems, namely modern algorithmic prescriptions inspired by neural networks and ultimately stemming from Parisy Complexity Theory, declined within the Jaynes inferential perspective on entropy maximization, namely the two pillars above which the whole thesis stands.
Regarding the choice of the biological complexity to unveil, I selected two problems quite far away within the Biological world and spatiotemporally complementary (but quite similar from the above modeling perspective on complex and automated inference):

- in the former -a *spatial problem*- we aim to detect the interactions among cancerous cells and their sourrounding both in presence and absence of a chemoterapeutic drug (we will consider two different pancreatic cancerous lineages to show that they give rise to quite different outcomes). By comparison of the results in the two cases, the present research becomes a cheap and powerful method to inspect the role and validity of a particular chemoterapy against a particular cancer type. Concretely we will deal with pictures of ensembles of broadly interacting cells in vitro, left free to evolve under time-lapse confocal microscopy and we infer the existence of cellular dialogues that affect their kinetics (namely signals that result in locomotion) by suitably mixing maximum entropy statistical inference and stochastic processes theory. This is a new protocol effective in order to study the crosstalk between cancerous cells and their surrounding cells.

- in the latter, instead, we deal with a *temporal problem*, namely we want to characterize the hear rate variability of healthy and pathogenic patients (we consider also here two pathologies, atrial fibrillation and cardiac decompensation) ultimately with the will of correlating differencies in the inferred observables with the pathologies they are coupled to, in order to provide to Clinicians a cheap computational protocol that can highlight new (and complementary w.r.t. achieved by standard routes) information helpful in early characterization of heart diseases. Concretely, once provided with Holter's registrations of cardiac performances of large collections of labelled patients, by inspecting heart rate variability with a suitable adaptation of the max-

imum entropy technique at work on these historical series, we will show that -while there universal scalings both in the temporal and frequency domains (already well known in the Literature and confirmed also by our findings)- there are also second order variations from these scalings and these variations are pathology-dependent; further, ultimately the picture that emerges from our approach is that of a *glassy hearth*, namely a picture where the healthy heart behavior is chaotic and complex, systemically accompained by a power-law statistics typical of frustrated systems. In particular we congecture that the typical $1/f$ scalings (vide infra) are (typical glassy) responses suggesting that the intrinsic variability in heat rate ultimately stems from the interplay of the sympathetic and the parasympathetic nervous system: as Holter's like recordings will be largely available in the next generation of wearable diagnostic tools, this kind of investigation aims to contribute to the (already started) growth of a *personalized medicine*.

3.1 Preamble: The Hopfield model from statistical inference

Before addressing the two biological problems this Chapter is due to, as a mandatory excercise we now obtain the Hopfield cost function from the Jaynes inferential perspective about the maximum entropy principle [86, 87]: this is important both for the mathematical as well as from the physical perspectives.
From the mathematical side it is instructive to see that, while we introduced the Hofield model simply by implementing in mathematical language the Hebb prescription for (biological) learning (i.e. *cells that fire together wire together*, also driven solely by Information Theory argument we would reach the same cost function. Indeed if we want that minima of the free energy to be highly correlated with the patterns ξ, we should introduce a cost function that has lowest values when the configuration of the network σ is close to ξ, but the simplest way to express this mathematically is exactly by writing $H(\sigma|\xi) \propto -(\sigma \cdot \xi)^2$ that is nothing but the Hopfield model. Further, via this inferential route, it becomes crystal clear that the pairwise Hopfield model searches only for one-point and two-points correlation functions in the datasets (thus its strength is confined within the *statistical reductionism*).
From the physical counterpart, instead, it is mandatory to read the second principe within the Jaynes perspective, simply because once the network ends up in a stable retrieval minimum, it gets stuck not in a thermodynamic minimum but in a cycle (a steady state where a circuitry of current becomes stationary) that is not at all the equilibrium condition analyzed by (equilibrium) statistical mechanics. In other words, the maximum entropy principle from a pure physical perspective is not enough to guarantee a sound theory simply because this is a network of (live) neurons and not (dead) atoms.

In an experimental scenario, in order to check retrieval performances of an associative neural network, one should measure at least two (series of) numbers: the mean values of the overlaps between the final output and the stored patterns and their relative variances. In other words, the experimental setup requires the observation of the quantities

$$\langle m_\mu \rangle_{\text{exp}} = \frac{1}{N} \sum_i \xi_i^\mu \langle \sigma_i \rangle_{\text{exp}}, \qquad \langle m_\mu^2 \rangle_{\text{exp}} = \frac{1}{N^2} \sum_{ij} \xi_i^\mu \xi_j^\mu \langle \sigma_i \sigma_j \rangle_{\text{exp}}. \tag{3.1}$$

The subscript exp means that we are considering experimentally evaluated quantities on some given sample. In order to make the notation more clear, we shall omit it, but the

averages $\langle\cdot\rangle$ should not be confused with the theoretical expectation values $\langle\cdot\rangle \equiv \mathbb{E}\Omega_{\boldsymbol{J}}$ introduced in the previous theoretical Chapters.

The goal is then to determine the probability distribution $\mathcal{P}(\boldsymbol{\sigma})$ accounting for these data. To do this, the standard tool coming from statistical inference is the *maximum entropy principle* discussed in the first Chapter. The basic idea is that $\mathcal{P}(\boldsymbol{\sigma})$ is obtained by maximizing the relative Shannon entropy $\mathcal{S}[\mathcal{P}] = -\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \log \mathcal{P}(\boldsymbol{\sigma})$. However, we have to impose some other constraints via a Lagrange multiplier problem. First of all, $\mathcal{P}(\boldsymbol{\sigma})$ should be a probability distribution, so the sum on the whole space should be equal to 1. Furthermore, we have to require that the mean values of the overlap m_μ and its square m_μ^2 equal the experimental data. In other words, we should maximize the quantity[1]

$$\begin{aligned} S_{A,\beta,h}[\mathcal{P}] = &-\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \log \mathcal{P}(\boldsymbol{\sigma}) + AN\Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) - 1\Big) + \\ &+ hN \sum_{\mu} \Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \frac{1}{N} \sum_i \xi_i^\mu \sigma_i - \langle m_\mu \rangle \Big) \\ &+ \frac{\beta N}{2} \sum_{\mu} \Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \frac{1}{N^2} \sum_{ij} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j - \langle m_\mu^2 \rangle \Big), \end{aligned} \tag{3.2}$$

with respect to $\mathcal{P}(\boldsymbol{\sigma})$ and the parameters A, h, β. The constraint $\partial_A S = 0$ is equivalent to require $\mathcal{P}(\boldsymbol{\sigma})$ is indeed a probability distribution, while the requirements $\partial_h S = \partial_\beta S = 0$ effectively fix the theoretical observables with the experimental data. Finally,

$$\frac{\delta S[\mathcal{P}]}{\delta \mathcal{P}(\boldsymbol{\sigma})} = -\log \mathcal{P}(\boldsymbol{\sigma}) - 1 + AN + h \sum_{i\mu} \xi_i^\mu \sigma_i + \frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j = 0, \tag{3.3}$$

which means that

$$\mathcal{P}(\boldsymbol{\sigma}) = \text{cost} \exp\Big(\frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + h \sum_{i\mu} \xi_i^\mu \sigma_i \Big) \tag{3.4}$$

By putting the constant equal to $\text{cost} = Z_N(\beta)^{-1}$, we prove the following last theorem of the thesis (than we move to real experiments).

Theorem 3.1. *The partition function associated to the probability distribution $\mathcal{P}(\boldsymbol{\sigma})$ maximizing the Shannon entropy* (3.2) *with the constraints* (3.1) *for the first and the second moment of neural activity is*

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \exp\Big(\frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + h \sum_{i\mu} \xi_i^\mu \sigma_i \Big), \tag{3.5}$$

namely the partition function related to the Hopfield cost function.

[1] Note that we added some extra N factor in order to ensure that all terms have the same order. Indeed, in the case of a constant probability distribution, i.e. $\mathcal{P}(\boldsymbol{\sigma}) = \prod_i \mathcal{P}(\sigma_i) = 2^{-N}$, therefore the logarithm in the Shannon entropy would give a factor N in the first term.

3.2 Problem One: Maximum entropy for stroma-cancer cross-talk

Cells in our body are not single entities but part of a community, they interact between themselves and with the surrounding environment thus experiencing social life [88]. This is pivotal in maintaining tissue organization and homeostasis, so as to coordinate an appropriate response to dangerous perturbations. When these dialogues go wrong diseases may rise, thus highlighting the importance of predicting and modeling cellular interactions, which provide insight into the mechanism of diseases development and progression [89, 90, 91]. The best examples is cancer, which can be defined of social dysfunction within cellular community and used as a model system to study intercellular communications. The search for signalling pathways, from direct cellular contact to soluble mediators, is highly non trivial involving advanced techniques ranging from gene expression measurements, yeast two-hybrid screening, co-immunoprecipitation, proximit labelling proteinomics, fluorescence resonance energy transfer imaging, X-ray crystallografy and more [92].
In this dedicated subsection, restricting solely to signalling affecting cell's kinetics, we present a novel and cheap computational approach -whose strength is its simplicity (it uses solely fluorescence microscopy and flow cytometry as experimental needs)- that allows quantifying the existence and intensity of interactions ruling cell's dynamics and spatio-temporal coordination: in a nutshell, via standard fluorescence microscopy imaging and cell's tracking, we collect the *phase space* of the experiment, namely the ordered time-series of all the cell's positions and velocities and we fed the computational protocol (based on maximum entropy extremization [15, 93, 94, 95]) by this knowledge. The output the algorithms returns is the effective intensity of interactions among the various cells and a quantitative description of the global motion: note that the method can predict the existence and magnitude of the interaction, but not its biological nature (we can state that there is a signaling protein at work but we can not identify it). Yet, despite this limitation, it can play as a powerful tool to compare how different drugs affect the kinetics of the same ensemble of cells, hence it plays possibly as a new approach to quantify kinetic cancer's drug response.
To prove this statement, we select pancreatic ductal adenocarcinoma (PDAC) cell lines as benchmark cases (specifically L3.6pl and AsPC-1 lines) and pancreatic stellate cells (PSCs) as cellular culture system, since their mutual interaction is known to be critical for PDAC progression: in this setting many efforts have been devoted to assess whether soluble mediators produced by carcinoma cells stimulated motility, proliferation and matrix synthesis of PSCs, and how these interactions enhanced tumour growth and progression [96, 97, 98]. Bachem and colleagues recently performed the wound assay in absence and in presence of tumour cell supernatants or in co-culture experiments with PSCs and tumour cells, observing random motility of PSCs in the wound assay and directed PSCs migration towards tumour cells in the co-culture experiments [99]. These data are supported by *in vivo* studies in which PDAC cells were orthotopically injected and their activity was seen to be promoted by PSCs [99, 100, 101, 102].
A major part of literature affirms that chemoresistance in PDAC is partially due to a unique presence of fibrous, stiff extracellular matrix (desmoplasia) surrounding the tumour, that could affect the intratumoural drug penetration [103]. However, the role of desmoplasia in cancer progression is complex and remains somehow controversial; in 2014, Gore and Korc went through the available literature trying to clarify whether the stroma is friend or foe in PDAC [104]: indeed, in that period several studies had demonstrated how targeting the stroma resulted in undifferentiated and more aggressive pancreatic cancer [105, 106].

Desmoplasia mainly derives from pancreatic stellate cells (PSCs) that are activated to proliferate and produce collagens, laminin, and fibronectin [107]; consequently, besides the physical role played by desmoplasia, another key aspect to consider is the molecular cross-talk between stroma and cancer cells, that regulates each cell type's survival, migration and other pro-tumourigenic properties. The lack of proper experimental models and approaches contributed to enhance the poor knowledge related to PDAC underlying mechanisms. Indeed, despite the need to study the complex interactions between PDAC cells and PSCs, very limited *in vitro* options currently exist [108, 109].

3.2.1 Automatic inference of cell's cross-talk

In our approach, at first, to to be sure about the choice of cell lines to investigate, we inspect the effect of PDAC cell lines on PSC kinetics by performing a standard wound healing experiment (see Figure 3.1, panel A): we let PSCs grow in a medium conditioned by L3.6pl, by AsPC-1 or nothing (as a control reference), we make a scratch and then we inspect the time cells need to migrate and fill the empty region. As shown in Figure 3.1 panel B, PSCs cell migration is highly conditioned by the medium resulting from AsPC-1 or L3.6PL cells, confirming the dependency of the dynamic behavior of PSCs from factors secreted by PDAC cells, thus highlighting the existence of specific stroma-cancer interactions. En route toward their quantification, we then perform the following series of experiments: for each cellular line (i.e. AsPC-1 or L3.6PL) we mix the PDAC cells with the PSC cells homogeneously and, via time lapse fluorescent microscopy, we collect cell's position and velocities. This knowledge suffices to infer the possible existence of kinetics interactions ruling the overall cellular dynamics via maximum-entropy statistical analysis and to characterize the kind of diffusion these cells give rise to by stochastic process theory. Note that we started from a mixed scenario to inspect how cells orchestrate their coordination to form larger aggregates (e.g. for the ADPC lines) and/or to inspect if and how PSC infiltrate these aggregates. We then repeat the experiments providing in the medium 5mg of gemcitabine and by comparison among the two series of experiments we can deduce the role of the chemoterapeutic treatment in cancer-stroma kinetics. In these experiments (Figure 3.1, panel C) the two cellular populations are left free to interact (with or without drug) keeping the ratio 25% of pancreatic cancer cells (*tumour* from now on) and 75% of PSCs pancreatic stellate cells (*stroma* from now on) whatever the malignant line (e.g. both for the L3.6pl as well as for the AsPC-1 cell lines).
In both sets of experiments described, the tumour and stroma cells were labeled with different tracking dyes used for fluorescent cell staining and time-lapse confocal imaging was applied to produce two distinct datasets containing all the cell's positions at given time points (Figure 3.1, panel D) and thus, by differentiating two consecutive time frames, also cell's velocities, namely the *phase-space* of the whole experiment (Figure 3.1, panel E): this is a typical input for several statistical methods inspired by statistical physics [110] (Figure 3.1, panels F, G), first of all the maximum entropy principle.
To obtain a clear scenario of the cell's kinetics, the key observable we investigate is cell's velocity: we split the study of this vector by analyzing its direction by means of maximum-entropy inference and by inspecting its modulus with stochastic processes theory: for the former, we adapt the Jaynes maximum entropy inference, in its Bayesian formulation [111], to infer existence and magnitude of interactions among cells, while, for the latter, we frame cell's dynamics as a Wiener process [6] (hence we evaluate its diffusion, drift, fluctuations, persistency, turning angles, etc. [112, 113, 114]) and, taking advantage by the homogeneous initial state, we inspect if and how the two cellular population tend to form aggregates (e.g. tumor cells can give rise to spheroids [89]), to mix (e.g. stroma can infiltrate within

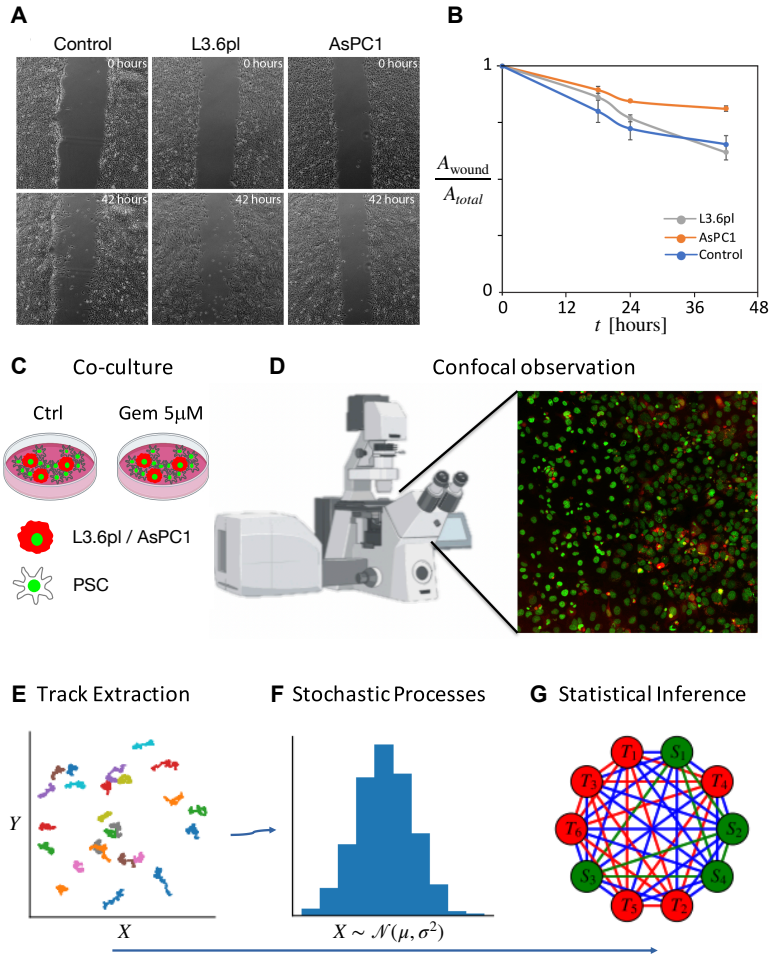

Figure 3.1: *Cartoon depicting the whole experimental and computational setups.* In the first row we inspect how PSCs migrate in a wound healing assay with 2D indirect co-culture (panel A). PSCs were grown in a cell monolayer and exposed, after a scratch, to conditioned medium from AsPC-1 or L3.6PL cells. As shown in panel B, where the vertical axes quantifies the ratio between the area of the scratch and the total area, contaminated medium sensibly affects PSCs cell migration confirming the presence of information exchanges among the various cellular lines . Prompted by this preliminary check, we define the following protocol to quantify such interactions: PDAC tumour cells (L3.6pl or AsPC-1, red symbols) and stromal cells (PSC, green symbols) are co-cultured in a cell culture dish with or without gemcitabine ($5\mu M$) up to 58 hours (panel C); Time-lapse confocal fluorescence microscopy is applied to track the positions of the cells versus time (panel D); Trajectories of each cell are reconstructed and, by temporal differentiation, the whole phase space of the experiment is acquired (panel E), namely the time ordered series of all the cell's positions and velocities: this information is the input to our algorithmic approach, split in stochastic process analysis (panel F) and maximum-entropy statistical inference (panel G).

the tumoral clumps [115]), etc.: merging their results and comparing experiments with and without gemcitabine we finally conclude on the role of the drug in governing the overall kinetics under investigation.

3.2.2 On cell's sensing and interactions

Interactions can be inferred by studying the directional aspects of cell's velocities, namely focusing on their reciprocal influence in turning: as standard in this case [15, 94, 95], we study the normalized orientational order parameter

$$\hat{n}_i(t) := \frac{\vec{v}_i(t)}{||\vec{v}_i(t)||} = \frac{\vec{r}_i(t+\Delta t) - \vec{r}_i(t)}{||\vec{r}_i(t+\Delta t) - \vec{r}_i(t)||}$$

defined as the velocity of a given cell i at time t divided by its modulus, namely *the angle* tracing the orientation, or simply *the direction* of that cell. Do these cells tend to cooperate, to align or, rather, to move independently? And how their coordination -if any- is affected by the drug? To answer these questions we need to know the collective properties of cell's directions, probabilistically coded by some unknown probability distribution $P(\hat{n})$ that we aim to find out by maximum entropy inference [15].
In a stylized way, given a dataset $\hat{n} = \hat{n}_1(1), ..., \hat{n}_i(t), ..., \hat{n}_N(T)$ (composed of multiple observations of the quantity $\hat{n}_i(t)$ from $t = 1$ to $t = T$ and for all the cells we tracked, i.e. $i \in (1, ..., N)$), this approach allows to reconstruct the probability distribution $P(\hat{n})$ from a limited number of empirical observations (too small to reconstruct the probability distribution directly from the data) by requiring such a probability distribution to reproduce all the experimental measurements (e.g. the one-point correlation function, $C_1 = \langle \hat{n}_i \rangle$ and two-point connected correlation function, $C_2 = \langle \hat{n}_i(t) \cdot \hat{n}_j(t) \rangle - \langle \hat{n}_i(t) \rangle \langle \hat{n}_j(t) \rangle$) yet being minimally structured, in a standard Occam razor way (namely, at the maximum entropy). We refer to the next sub-subsections for a detailed explanation of this declination of the maximum entropy principle (in particular section *Maximum entropy extremization for one and two point correlations* for its construction suitable for the present analysis and section *Bayesian marginalizatio: en route to the posterior* for the related resolution). Specifically, given a set of observables (e.g. C_1, C_2) related to the variable $\hat{n}_i(t)$, their experimental and model estimates are, respectively

$$\langle C_1(\hat{n}) \rangle_{experimental} = \frac{1}{T} \sum_{i=1}^{T} \hat{n}_i(t), \tag{3.6}$$

$$\langle C_1(\hat{n}) \rangle_{computational} = \int d\hat{n} P(\hat{n}) \hat{n}_i(t), \tag{3.7}$$

and likewise for C_2: the maximum entropy method constructs $P(\hat{n})$ as the least-structured probability distribution that matches the experimental averages above with its theoretical outcomes, i.e. $\langle C_k(\hat{n}) \rangle_{experimental} = \langle C_k(\hat{n}) \rangle_{computational}$ for $k \in (1, 2)$, the amount of *structure* in $P(\hat{n})$ being quantified by the Shannon entropy

$$S[P] = -\int d\hat{n} P(\hat{n}) \ln P(\hat{n})$$

such that, the higher the value of $S[P]$, the less structured $P(\hat{n})$ results.
As we are dealing with two cellular populations (namely those belonging to tumor and those to stroma), we need to enlarge the above standard (single population) maximum entropy toward a multi-population generalization: using the labels S and T for *stroma*

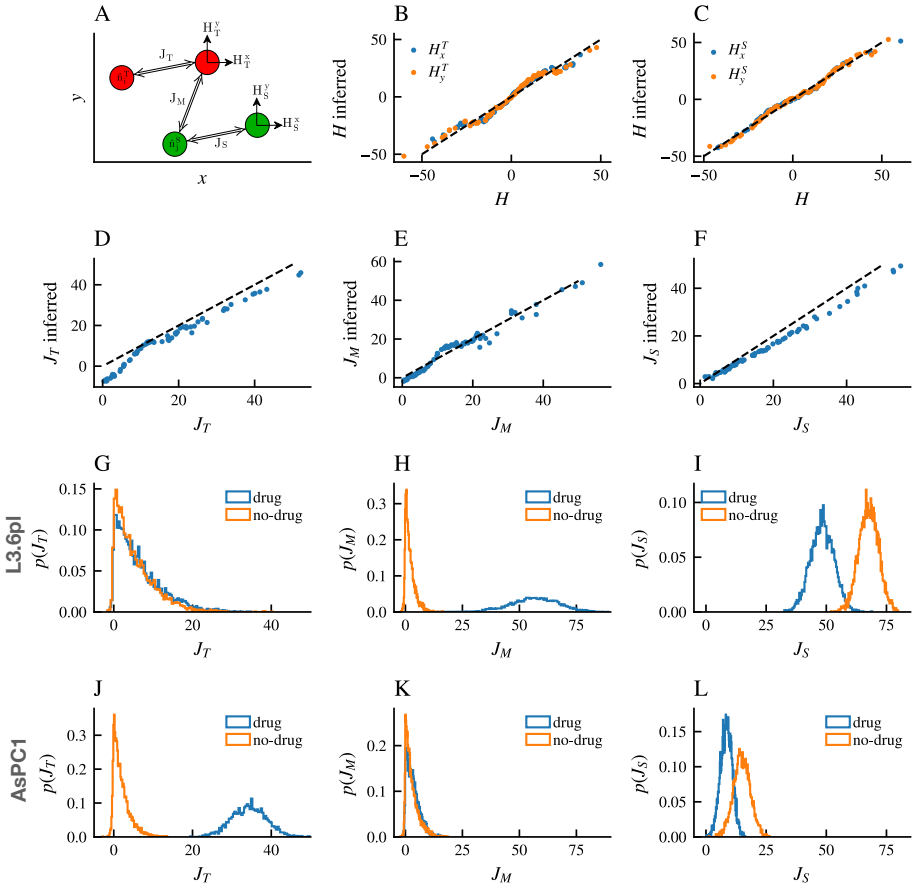

Figure 3.2: *Inferring the interactions.* Panel A: Cartoon stylizing two cells per cellular lineage (different lineages in different colors) interacting via the thee possible couplings J_T (tumor-tumor interaction), J_S (stroma-stroma interactinos) and J_M (mixed interactions tumor-stroma) and, eventually, perceiving a planar field (e.g. a chemotactic gradient) H_x, H_y.

Panels B-F: results of the maximum entropy inference on synthetic datasets simulated by the Kuramoto-Heisemberg model. We simulated 20000 synthetic trajectories -whose parameters were known- and analyzed their phase space. We plot on the horizontal line the true value of the parameters and on the vertical line the inferred ones. In particular external fields are reported in panels B and C, while the interactions among tumor-tumor, tumor-stroma and stroma-stroma are reported in panels D, E, F respectively.

Panel G-I: results from the L3.6pl experiments.: distributions of the inferred couplings in the two datasets (without drug in orange and with drug in blue). While J_T is roughly left invariant by the drug (panel G) and stable on low intensities (suggesting poor tumor-tumor interactions), J_M is by far increased (panel H) and J_S is sensibly decreased (panel I) by the presence of gemcitabine, the whole suggesting that an effect of the drug is to diminish stroma-stroma interactions to enrich mixed ones.

Panel J-L: results from the AsPC-1 experiments.: distributions of the inferred couplings in the two datasets (without drug in orange and with drug in blue). While J_T is sensibly increased by the presence of the drug (panel J), mixed interactions are almost absent with or without the drug (panel K) and stroma-stroma interactions mildly leveraged by the drug.

and *tumour* respectively, we need to match the empirical and computational expectations of the correlation functions $\langle C_1(\hat{n^S})\rangle$, $\langle C_1(\hat{n^T})\rangle$, $\langle C_2(\hat{n^S},\hat{n^S})\rangle$, $\langle C_2(\hat{n^T},\hat{n^T})\rangle$, $\langle C_2(\hat{n^S},\hat{n^T})\rangle$ hence we have a *coupling* J_S (resulting from the constraint on $\langle C_2(\hat{n^S},\hat{n^S})\rangle$) accounting for interactions among two stroma cells, a *coupling* J_T (resulting from the constraint on $\langle C_2(\hat{n^T},\hat{n^T})\rangle$) accounting for interactions among two cancerous cells and a *mixed coupling* J_M (resulting from the constraint on $\langle C_2(\hat{n^S},\hat{n^T})\rangle$) accounting for interactions among S and T cells to be inferred. Further $\mathbf{H}_S = (H_{S,x}, H_{S,y})$ and $\mathbf{H}_T = (H_{T,x}, H_{T,y})$ are two bi-dimensional extra-parameters (i.e. simple homogeneous external fields) that we should infer as well to deal with a possible persistency coded in the one-point correlation functions $\langle C_1(\hat{n^S})\rangle, \langle C_1(\hat{n^T})\rangle$: see the cartoon in panel A of Fig. 3.2 to capture the meaning of the various parameters.
In the mean-field limit, the extremization of the Shannon entropy returns the probability distribution $P(\hat{n})$ in terms of a Gibbs measure with a given cost function $\mathcal{H}(\hat{n}_S, \hat{n}_T | J_S, J_T, J_M, H_S, H_T)$, that is an explicit function of these couplings $\{J_S, J_T, J_M\}$ and fields $\{H_S, H_T\}$ and that reads as

$$P(\hat{n}) = \frac{e^{-\mathcal{H}(\hat{n}_S,\hat{n}_T|J_S,J_T,J_M,H_S,H_T)}}{Z(J_S,J_T,J_M,H_S,H_T)}, \tag{3.8}$$

$$\begin{aligned}\mathcal{H} \sim \frac{-1}{N(N-1)}[&\sum_{i\neq j}^{N_S,N_S} J_S \hat{n}_i \hat{n}_j + \sum_{i\neq j}^{N_T,N_T} J_T \hat{n}_i \hat{n}_j \\ &+ \sum_{i\neq j}^{N_S,N_T} J_M \hat{n}_i \hat{n}_j] - \frac{1}{N}\left(\mathbf{H}_S \cdot \sum_i^{N_S} \hat{n}_i + \mathbf{H}_T \cdot \sum_j^{N_T} \hat{n}_j\right)\end{aligned} \tag{3.9}$$

where $N = N_S + N_T$ and $Z(J_S, J_T, J_M, H_S, H_T)$ -the partition function in statistical physics- plays here as a simple normalization factor: we obtained the cost-function (or *Hamiltonian* to keep the statistical physics jargon) of a bipartite Heisenberg-Kuramoto model [94, 116].
Hereafter we comment the result of such inferential procedure.
In the first row of Fig. 3.2, beyond the picture in panel A, we report results on synthetic datasets –generated accordingly to the Heisemberg-Kuramoto and Vicsek models (see Supplementary Material for details)– to calibrate the computational approach: the maximum entropy inference reconstructs with high accuracy the (known) values of the drifts H (shown in panels B and C of Fig. 3.2) as well as the interactions (shown in panels D,E,F of Fig. 3.2 respectively for J_T, J_M, J_S). In the second and third rows of Fig. 3.2, panels D,E,F, and panels G,H,I, respectively, the real distributions of the key parameters J_S, J_T, J_M for the L3.6pl and the AsPC-1 cases are shown: by inspecting these plots we conclude that

- Interactions among L3.6pl cancerous cells are not influenced by the drug (panel G), while interactions among AsPC-1 cancerous cells are heavily enhanced by the drug (panel J) highlighting a significant heterogeneity these cells manifest in the kinetic response to the drug.
- Interaction among stroma cells and L3.6pl cancerous cells are deeply influenced by the drug (panel H): in particular, without gemcitabine, there is roughly no interaction among stroma and cancer, while -in the presence of the drug- pronounced interactions do appear. At contrary, for the AsPC-1 case, there are no net interactions nor without neither with the drug, whose effect seems rather marginal on the overall dynamics (panel K).

- Interaction within the stroma are deeply influenced by the drug for the L6.3p, scenario (panel I): in particular, PSC cells diminish to interact reciprocally in presence of gemcitabine (possibly to enhance interactions with the cancerous counterpart). This effect is barely observable in the AsPC-1 scenario, as reported in panel L highlighting a strong heterogeneity also in the response of these cells to the presence of the drug.

3.2.3 Algorithmic implementation

The input of all our algorithms is made of four datasets, as there are two cancerous lines, i.e. L3.6pl and AsPC-1, that interact with the stroma and experiments are made in presence of gemcitabine, D (i.e. *with drug*), and in absence of gemcitabine, ND (i.e. *without drug*), for comparison and for all of them notation is as follows:

- time t
- cell number i
- coordinates $r_i(t) = (x_i(t),\, y_i(t))$
- normalized red intensity $I_i(t) = \frac{1}{2} + \frac{1}{2}\tanh\left[\frac{I_i^{Red}(t) - median(I_i^{Red}(t))}{IQR(I_i^{Red}(t))}\right]$

Since cells of different types can be distinguished by the presence or absence of red light over the cell nucleus, at first we classify cell's lineages according to a predefined threshold λ on the recorded normalized red intensities I^{Red} such that if the latter satisfies $I^{Red} > \lambda$ the cell will be assigned to population T, otherwise to population S. However, we stress that by dealing with approximate Bayesian inference as in the present approach, it is possible to remove this external tuning by producing an optimal estimate also over λ, thus eliminating the need of establishing an arbitrary threshold a-priori: we checked a posteriori that results are in full agreement whatever the approach.

Maximum entropy extremization for one and two point correlations

The one-point, C_1, and two-points, C_2, correlation functions for the angles $\hat{n}$, where

$$\hat{n}_i(t) = \frac{\vec{r}_i(t + \Delta t) - \vec{r}_i(t)}{||\vec{r}_i(t + \Delta t) - \vec{r}_i(t)||},$$

are defined as

$$C_1(A) = \langle \hat{n}_A(t) \rangle_t \tag{3.10}$$

$$C_2(A, B) = \langle \hat{n}_A(t) \cdot \hat{n}_B(t) \rangle_t - \langle \hat{n}_A(t) \rangle_t \cdot \langle \hat{n}_B(t) \rangle_t \tag{3.11}$$

where $A \in (S, T)$ as well as $B \in (S, T)$: we now constraint, via the maximum entropy, the experimental and computational evaluations of these averages to match.
Despite we started with the Shannon expression for the constrained entropy $S[P]$, once understood that the inferential procedure returns the Gibbs measure of a suitable cost-function, we can deal directly with the Lagrangian functional for the constrained Gibbs

entropy that reads as

$$\mathcal{L}(\{\hat{n}_k\}, J, H) = -\int d\{\hat{n}_k\}\, P(\{\hat{n}_k\}) \log P(\{\hat{n}_k\})$$
$$+\frac{1}{N(N-1)} \sum_{i<j=1}^{N,N} J_{i,j}\Big(\int d\{\hat{n}_k\}\, P(\{\hat{n}_k\})\hat{n}_i \cdot\, \hat{n}_j - \langle \hat{n}_i \cdot\, \hat{n}_j\rangle_t\Big)$$
$$+\frac{1}{N}\sum_{i=1}^{N} \vec{H}_i \cdot \Big(\int d\{\hat{n}_k\}\, P(\{\hat{n}_k\})\hat{n}_i - \langle \hat{n}_i\rangle_t\Big)$$
$$+\lambda\Big(\int d\{\hat{n}_k\}\, P(\{\hat{n}_k\}) - 1\Big)$$

Via standard functional extremization we can at first check that, at the stationary point, P correctly reproduces the correlation functions; moreover if we extremize $\mathcal{L}$ w.r.t. P we deduce (up to normalization) the expression for P, that is

$$P(\{\hat{n}_k\}|J,H) \propto \exp\Big(\sum_{i<j=1}^{N,N} J_{ij}\frac{\hat{n}_i \cdot\, \hat{n}_j}{N(N-1)} + \sum_{i=1}^{N} \vec{H}_i \cdot \frac{\hat{n}_i}{N}\Big). \tag{3.12}$$

This model depends on the matrix J and on the vectors $\vec{H}_i$: at present there are overall $O(N)$ variables and $O(N^2)$ parameters. Clearly to make use of this model we need to coarse grain: as there are only two lineages of cells and it is reasonable to believe that similar cells share similar statistics, we partition the set $\{\hat{n}_k\}$ into two sets $\{\hat{n}_k^T\}$ and $\{\hat{n}_k^S\}$ (the former regarding the tumour, the latter the stroma) and we take the entries of J to be the constant J_T if both the row index and column index belong to the set of T cells, the constant J_S for S cells and J_M for mixed interactions T-S; of course we apply the same argument for the vectors $\vec{H}$.
These assumptions turn the model into a bi-partite mean-field Heisenberg-Kuramoto pairwise model whose solution can be achieved analytically: the cost function of the model reads as

$$\mathcal{H}(n^T, n^S|J,H) = \tag{3.13}$$
$$= \frac{-J^T}{N_T(N_T-1)}\sum_{i\neq j=1}^{N_T} \hat{n}_i^T \cdot \hat{n}_j^T - \frac{J^S}{N_S(N_S-1)}\sum_{i\neq j=1}^{N_S} \hat{n}_i^S \cdot \hat{n}_j^S$$
$$\frac{-J^M}{N_T N_S}\sum_{i,j=1}^{N_T,N_S} \hat{n}_i^T \cdot \hat{n}_j^S + \frac{1}{N_T}\vec{H}^T \cdot \sum_{i=1}^{N_T} \hat{n}_i^T + \frac{1}{N_S}\vec{H}^S \cdot \sum_{i=1}^{N_S} \hat{n}_i^S$$

where $\hat{n}$ are unit vectors in $\mathbb{R}^2$.
Now we are left with only seven free parameters opening up the possibility of inferring them: the probability of observing a configuration $\hat{n} := (n^T, n^S)$ is

$$P(n^T, n^S|J,H) = \frac{\exp\big(-\mathcal{H}(n^T, n^S|J,H)\big)}{Z(J,H)} \tag{3.14}$$

where $Z(J,H)$ is the partition function:

$$Z(J,H) = \int_{\mathcal{C}^{N_T+N_S}} d^{N_T}\hat{n}^T\; d^{N_S}\hat{n}^S\; \exp\big(-H(\hat{n}|J,H)\big) \tag{3.15}$$

and $\mathcal{C}$ is the set of unit vectors in $\mathbb{R}^2$.

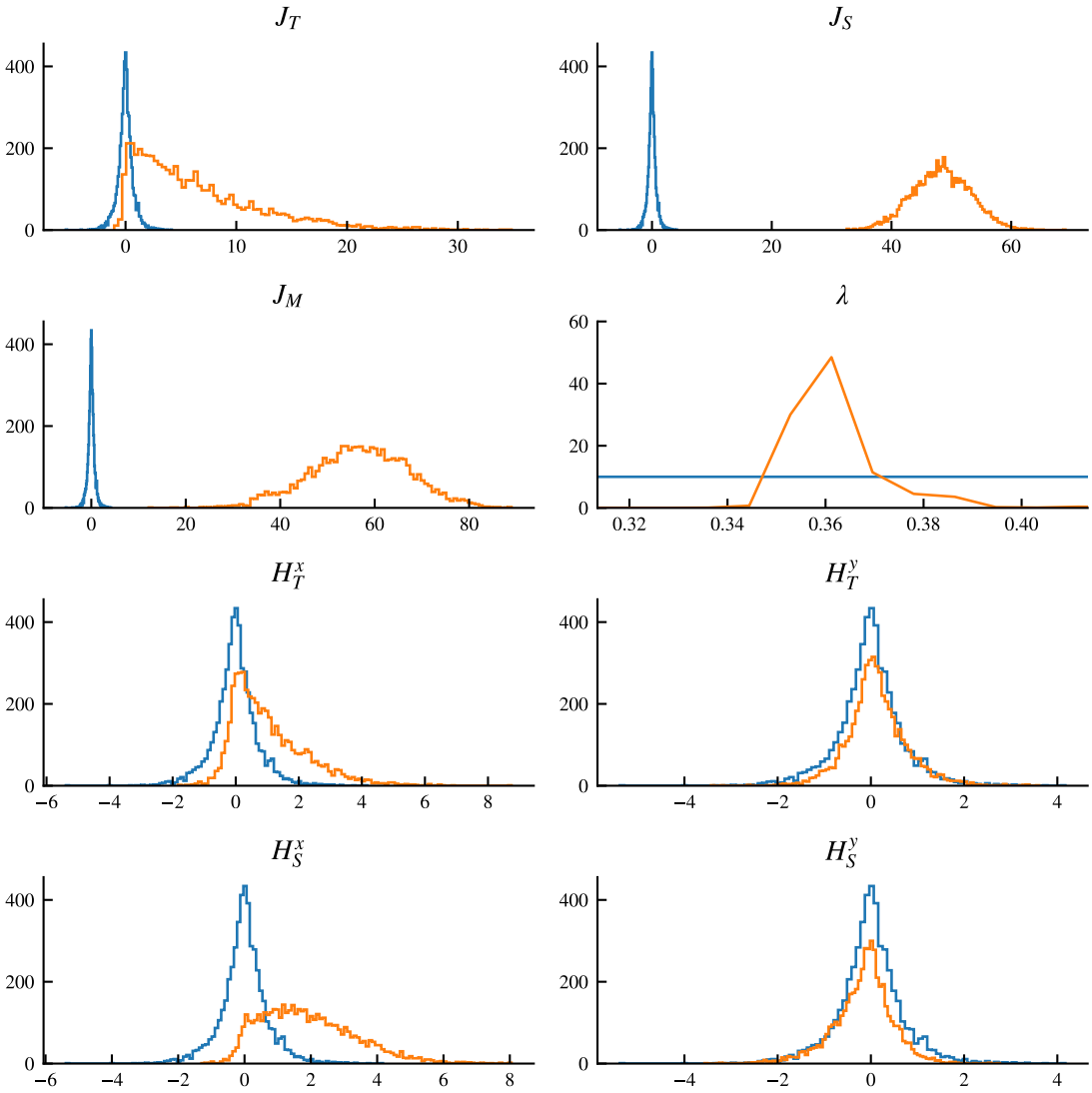

Figure 3.3: Case L3.6pl. Dataset D: Parameters inferred for the Maximum Entropy model (3.27), the blue curves are the prior distributions, while the orange curves are the posterior distributions.

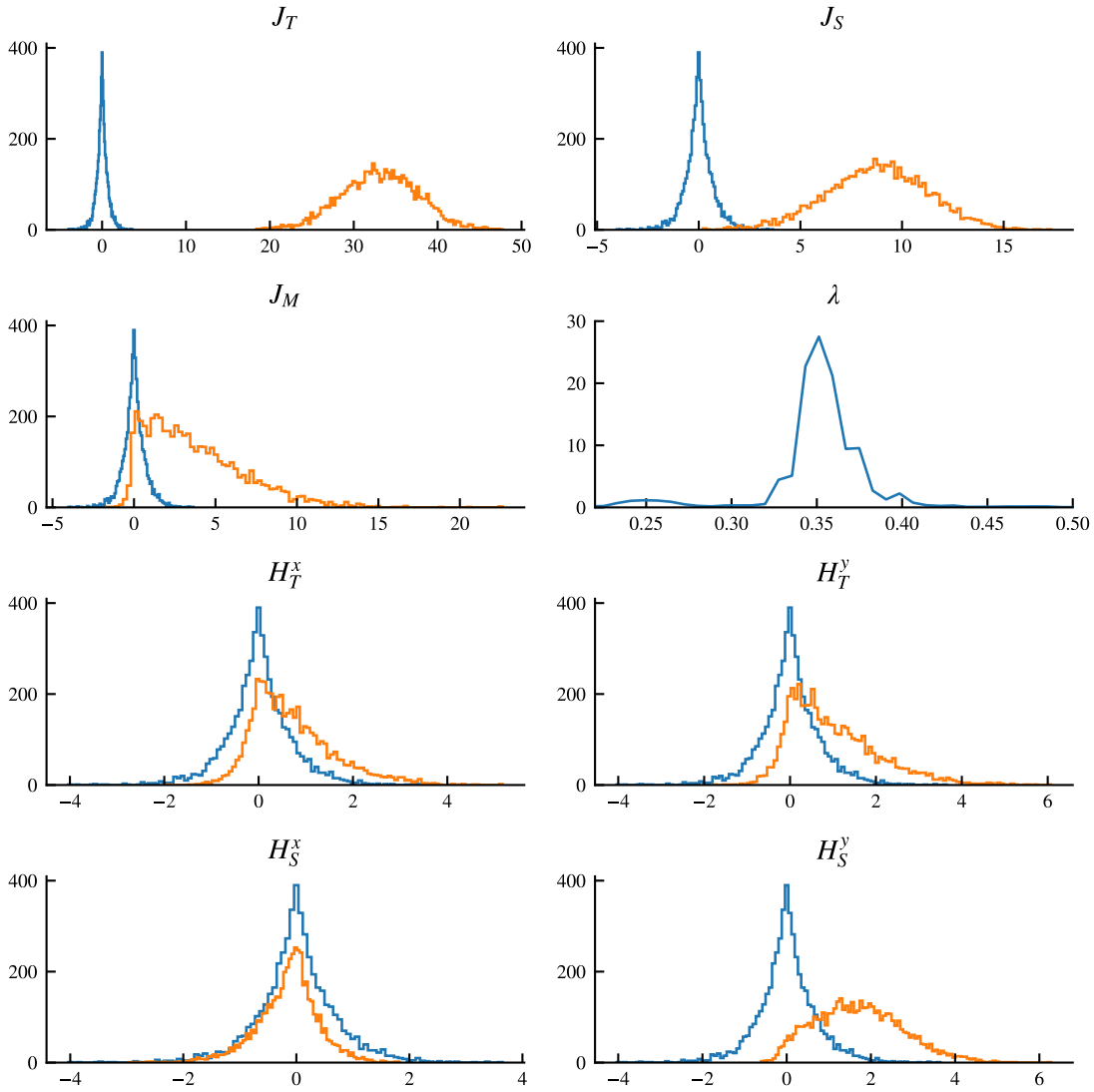

Figure 3.4: Case AsPC-1. Dataset D: Parameters inferred for the Maximum Entropy model (3.27), the blue curves are the prior distributions, while the orange curves are the posterior distributions.

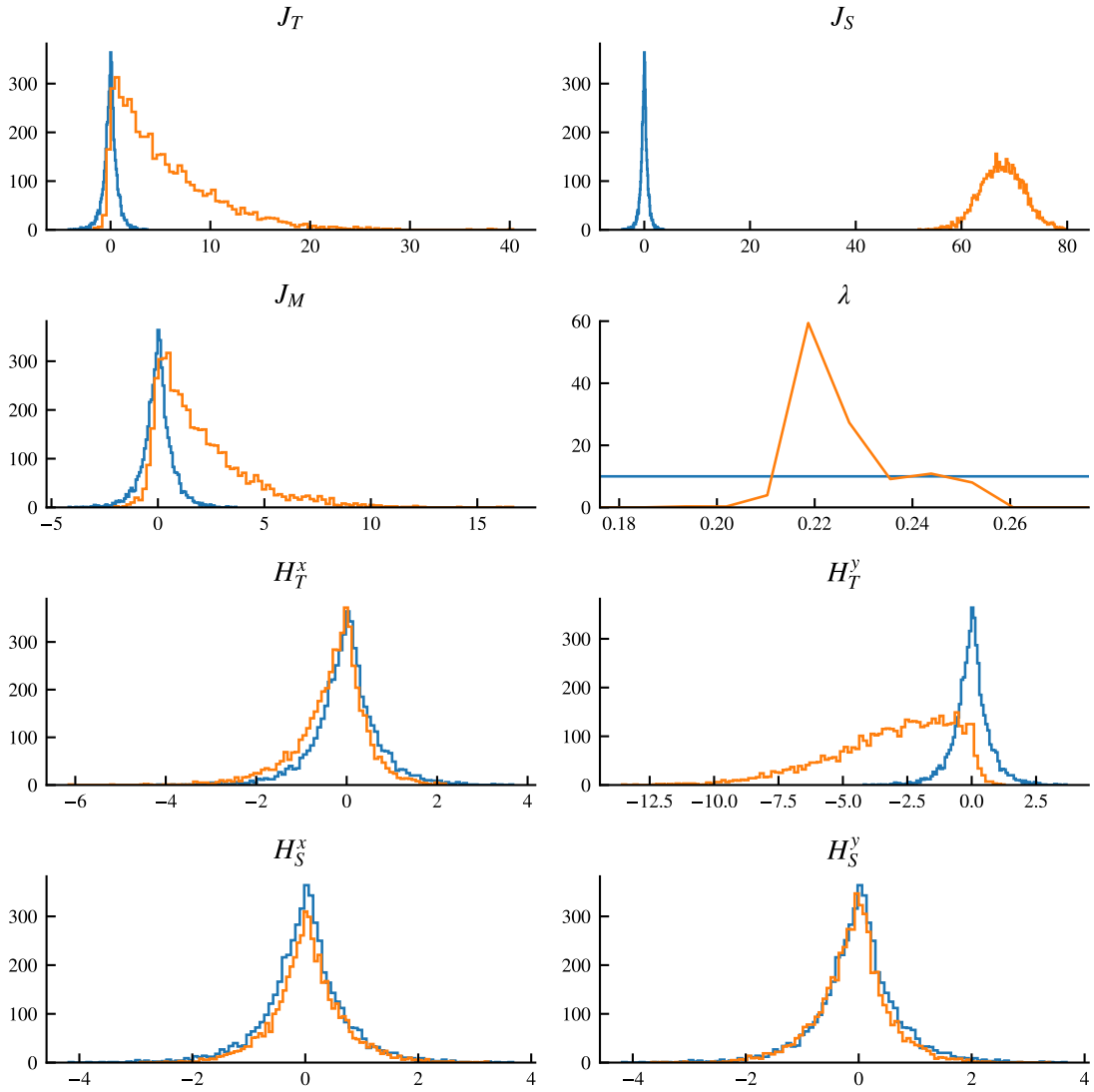

Figure 3.5: Case L3.6pl. Dataset ND: Parameters inferred for the Maximum Entropy model (3.27), the blue curves are the prior distributions, while the orange curves are the posterior distributions.

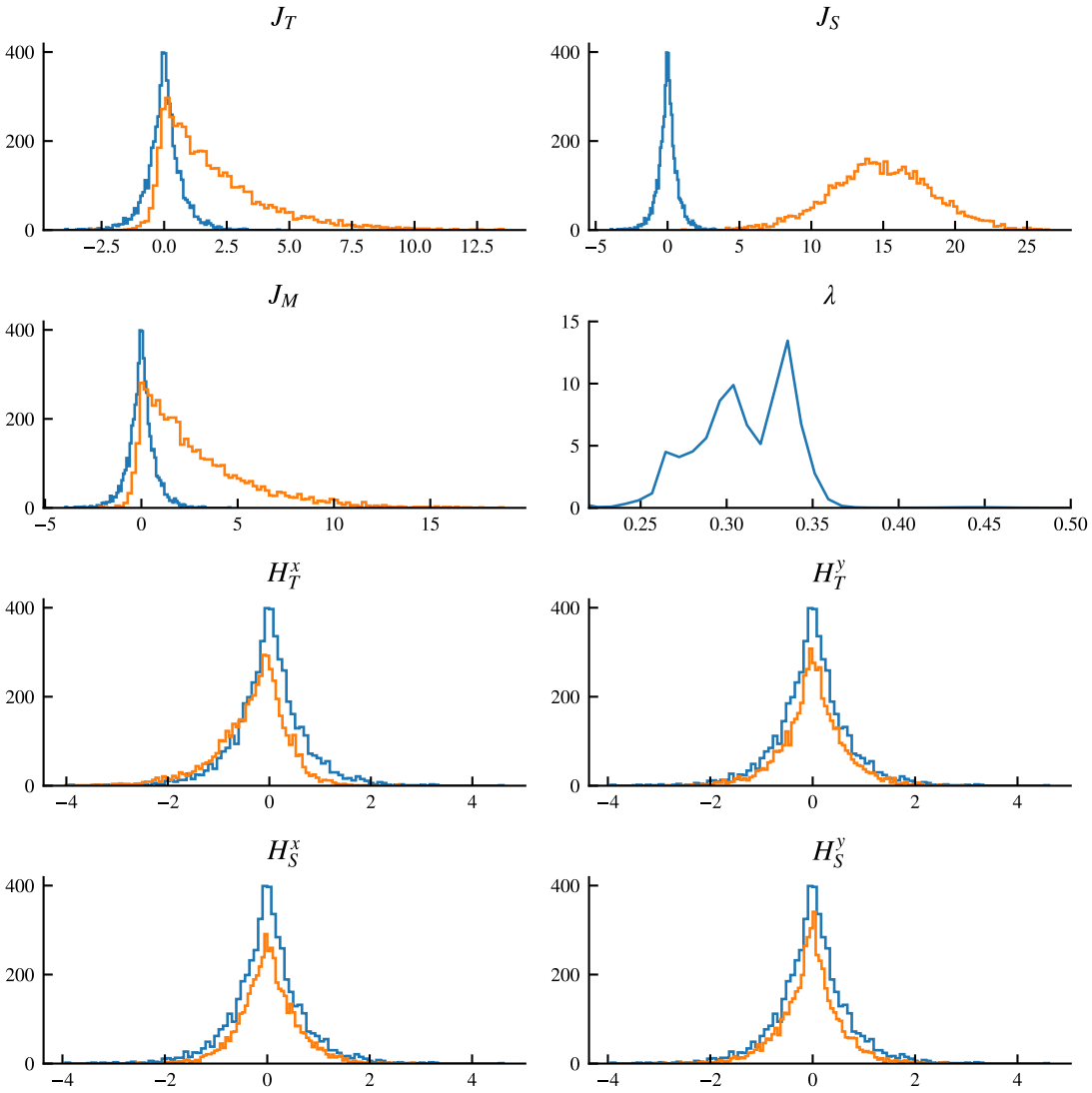

Figure 3.6: Case AsPC-1. Dataset ND: Parameters inferred for the Maximum Entropy model (3.27), the blue curves are the prior distributions, while the orange curves are the posterior distributions.

Bayesian marginalization: en route for the posterior

Given the acquired observations $\{\hat{n}_k\}$ of the $\hat{n}$'s, to estimate the parameters $J_S, J_M, J_T, \vec{H}_S, \vec{H}_T$ we rely on Bayes theorem, which dictates

$$P(\{\hat{n}_k\}|J,H)P(J,H) = P(J,H|\{\hat{n}_k\})P(\{\hat{n}_k\}) \tag{3.16}$$

thus

$$P(J,H|\{\hat{n}_k\}) = \frac{P(\{\hat{n}_k\}|J,H)P(J,H)}{P(\{\hat{n}_k\})} \tag{3.17}$$

so the log-posterior density is

$$l_P(J,H|\{\hat{n}_k\}) = \log P(\{\hat{n}_k\}|J,H) + \log P(J,H) - \log P(\{\hat{n}_k\}) \tag{3.18}$$

where $\log P(\hat{n}_1, \cdots, \hat{n}_T|J,H)$ is the likelihood of the set of observation of $\hat{n}$: more precisely, assuming $k \in (1,...,T)$, we can write $\log P(\{\hat{n}_k\}|J,H)$ as

$$\sum_{i=1}^{T} \log P(\hat{n}_i|J,H) = \sum_{i=1}^{T} \mathcal{H}(\hat{n}_i|J,H) - T \log Z(J,H). \tag{3.19}$$

As well known, there is a glaring problem with this approach, the partition function Z is intractable. This problem can be tackled via the pseudo-likelihood approximation with great accuracy: the pseudo-likelihood approximation consists in writing the log-density

$$\log P(\hat{n}|J,H) \approx \sum_{i \in 1,\cdots,N_A+N_B} \log P(\hat{n}_i|J,H,\hat{n}_{\setminus i}) \tag{3.20}$$

as a sum of conditional log-densities and the great advantage of this approximation is that the term $\log P(\hat{n}_i|J,H,\hat{n}_{\setminus i})$ admits a closed form solution; indeed via the identity

$$P(\hat{n}|J,H) = P(\hat{n}_i|J,H,\hat{n}_{\setminus i})P(\hat{n}_{\setminus i}|J,H) = \frac{\exp\big(-\mathcal{H}(\hat{n}|J,H)\big)}{Z(J,H)} \tag{3.21}$$

we get

$$P(\hat{n}_i|J,H,\hat{n}_{\setminus i}) = \frac{\exp\big(-\mathcal{H}(\hat{n}|J,H)\big)}{\int_{\mathcal{C}} d\hat{n}_i \ \exp\big(-\mathcal{H}(\hat{n}|J,H)\big)}. \tag{3.22}$$

Since $\hat{n}_i$ can either be in the set of $\hat{n}^T$ or in the set of $\hat{n}^S$ we must distinguish these cases: we do so by introducing the vector quantity

$$F_Y(X,\hat{n}) = \frac{2J^X}{N_X(N_X-1)} \sum_{i=1}^{N_X} \hat{n}_i^X + \frac{J^M}{N_X N_Y} \sum_{i=1}^{N_Y} \hat{n}_i^Y + \frac{1}{N_X}\vec{H}_X. \tag{3.23}$$

This quantity allows to express both conditional densities conveniently as

$$P(\hat{n}_i^T|\hat{n}_{\setminus i}^T,\hat{n}^S,J,H) = \frac{\exp(F_S(T,\hat{n})\cdot\hat{n}_i^T)}{2\exp(\frac{2J^T}{N_T(N_T-1)})I_0(|F_S(T,\hat{n})|)} \tag{3.24}$$

$$P(\hat{n}_i^S|\hat{n}_{\setminus i}^S,\hat{n}^T,J,H) = \frac{\exp(F_T(S,\hat{n})\cdot\hat{n}_i^S)}{2\exp(\frac{2J^S}{N_S(N_S-1)})I_0(|F_T(S,\hat{n})|)} \tag{3.25}$$

where I_0 is the modified Bessel function of type "I" and order 0.
With this approximation the (pseudo) log posterior density becomes tractable

$$
\begin{aligned}
& l_P(J,H|\hat{n}_1,\cdots,\hat{n}_T) \approx \\
& \sum_{n\in\mathcal{D}}\Big[\sum_{i=1}^{N_T}\log P(\hat{n}_i^T|\hat{n}_{\setminus i}^T,\hat{n}^S,J,H) + \\
& \sum_{i=1}^{N_S}\log P(\hat{n}_i^S|\hat{n}_{\setminus i}^S,\hat{n}^T,J,H)\Big] + \\
& \log P(J,H) - \log P(\hat{n}_1,\cdots,\hat{n}_T) = \\
& \sum_{d\in\mathcal{D}}\Big[F_S(T,\hat{n})\cdot\sum_{i=1}^{N_T}\hat{n}_i^T + F_T(S,\hat{n})\cdot\sum_{i=1}^{N_S}\hat{n}_i^S - \\
& N_T(\log 2 + \frac{2J^T}{N_T(N_T-1)} - \log I_0(|F_S(T,\hat{n})|)) - \\
& N_S(\log 2 + \frac{2J^S}{N_S(N_S-1)} - \log I_0(|F_T(S,\hat{n})|))\Big] + \\
& \log P(J,H) - \log P(\hat{n}_1,\cdots,\hat{n}_T)
\end{aligned}
\tag{3.26}
$$

This pseudo-log-posterior density is finally suitable for sampling the variable J,H: we have done so via Hamiltonian Monte Carlo method and the results are summarised in Fig. 3.3 for the drugged dataset of L3.6pl case, in Fig. 3.4 for the drugged dataset of AsPC-1 case, in Fig. 3.5 for the not-drugged dataset of 6.3pl case and in Fig. 3.6 for the not-drugged dataset of AsPC-1 case.
The results show the relevant heterogeneity in response to the drug by the two inspected cellular lines: in the L3.6pl scenario, the presence of the drug increases the cross-talk between stroma and tumour (J_M is drastically drifted away from zero, indicating that the probability distribution no longer factorizes over the cell lineages) while sacrificing stroma-stroma dialogues (J_S gets weaker by the presence of gemcitabine). In the AsPC-1 counterpart, instead, cross-talk is absent without and stays absent in the presence of the drug, rather tumor-tumor interactions -that are barely pronounced without the drug, become significantly predominant, in response to the drug. Via stochastic processes and, finally, cell's counting we will correlated these interactions with overall global dynamics of the cells and their survival.

3.2.4 On cell's diffusion and crowding

As the two types of cells are homogeneously mixed together, by a trivial symmetry argument, there is no global chemotactic gradient (nor in the experiment with no drug (ND) neither in the one with the drug (D)), hence, in the long run limit, cells should overall perform Brownian motion (their dynamics is expected asymptotically diffusive): this is confirmed in the first row of panels in Fig. 3.8 where we show the temporal evolution of the ratio between the empirical root mean square displacement of the two lineages and that of a pure Brownian diffusion (the *control* in the panels) for both the dataset without the drug (panel A) and the dataset with the drug (panel B): while on the short timescale cells deviate from pure diffusion (and we will see soon that their motion can actually be locally ballistic), for long enough times the two perfectly collapse on the control.
However, looking at shorter times, it is also evident that interactions among cells take place and that these are enhanced by the presence of the drug: to inspect their effects, e.g.

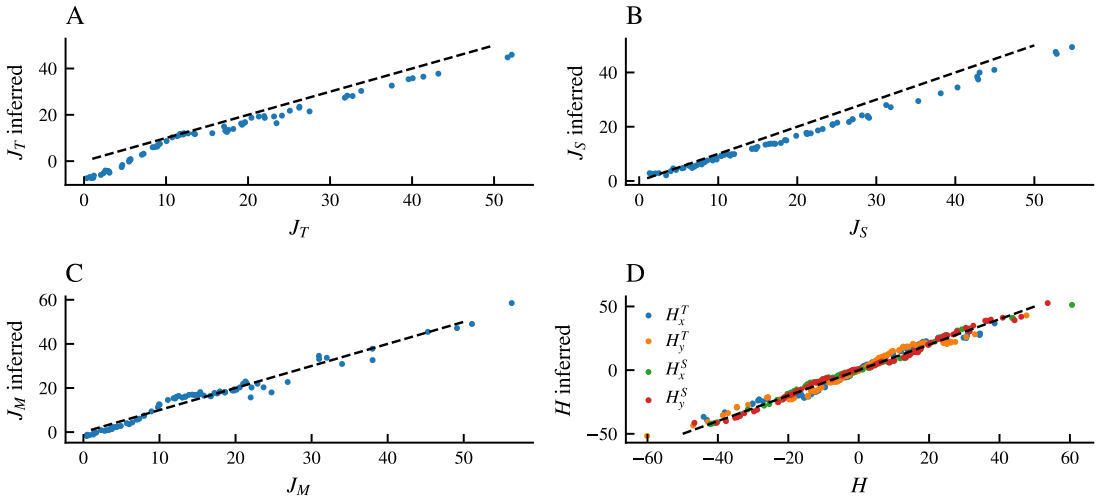

Figure 3.7: *Validation of the inferential procedure with the two-population Heisenberg-Kuramoto model*
We simulated overall 20000 synthetic trajectories by varying systematically the coupling J_S, J_T, J_M and the fields H_x^S, H_y^S, H_x^T, H_y^T, and we report the scatter plot among the original parameters and the inferential outcomes for the various couplings and fields as the various labels explain along the panels: as it shines even by a visual glance at the plots, the algorithm almost perfectly reconstructs the correct interactions and fields.

if and how cells thicken, we study the average intercellular distance $D(t)$, as a function of time t, defined as

$$D_{A,B}(t) = \langle ||\vec{r}_a(t) - \vec{r}_b(t)|| \rangle_{a \in A, b \in B}, \tag{3.27}$$

where the averages are restricted to the cellular type such that $A := (S, T)$ and $B := (S, T)$ giving rise to three quantifiers: $D_{S,S}(t)$, $D_{S,T}(t)$, $D_{T,T}(t)$.
If there is no crowding, these quantifiers are expected to fluctuate around a constant value over time, conversely if -say- S and T types are merging, $D_{S,T}(t)$ should be a monotonously decreasing function (likewise, if those cells are spreading away, $D_{S,T}(t)$ is expected to increase in time): these markers are depicted in the second line of panels in Fig. 3.8 for the L3.6pl case for both the datasets, without drug (panel C) and with drug (right, panel D) and in the third line of panels in Fig. 3.8 for the AsPC-1 case for both the datasets, without drug (panel E) and with drug (panel F).
Remarkably, for the L3.6pl scenario, while $D_{S,S}(t)$ is kept (approximately) constant in both the experiments, $D_{S,T}(t)$ and $D_{T,T}(t)$ are (approximately) constant solely in the dataset without the drug, while in presence of gemcitabine these are monotonically decreasing functions of time. In particular, more than $D_{T,T}(t)$, $D_{S,T}(t)$ heavily experiences this phenomenon, suggesting that while tumour cells tend to form agglomerations also stromal ones strongly tend to join in due to the presence of the drug. This is no longer true in the AsPC-1 counterpart where $D_{S,T}(t)$ stays constant even in presence of the drug.
A quite remarkable behavior we highlight is that, while the motion of these cells is globally diffusive at the macroscale (as shown in panels A and B of Figure 3.8), local interactions give rise to ballistic motion, typical of sensing cells [94, 112] as the best fit for their (average) reciprocal distances versus time returns a roughly linear dependence of time for $D_{a,b}(t)$ vs t: a local ballistic shortage, suggests that the T and S lineages are actually interacting, as it happens in the L3.6pl case (Figure 3.8, panel D) in complete agreement with the

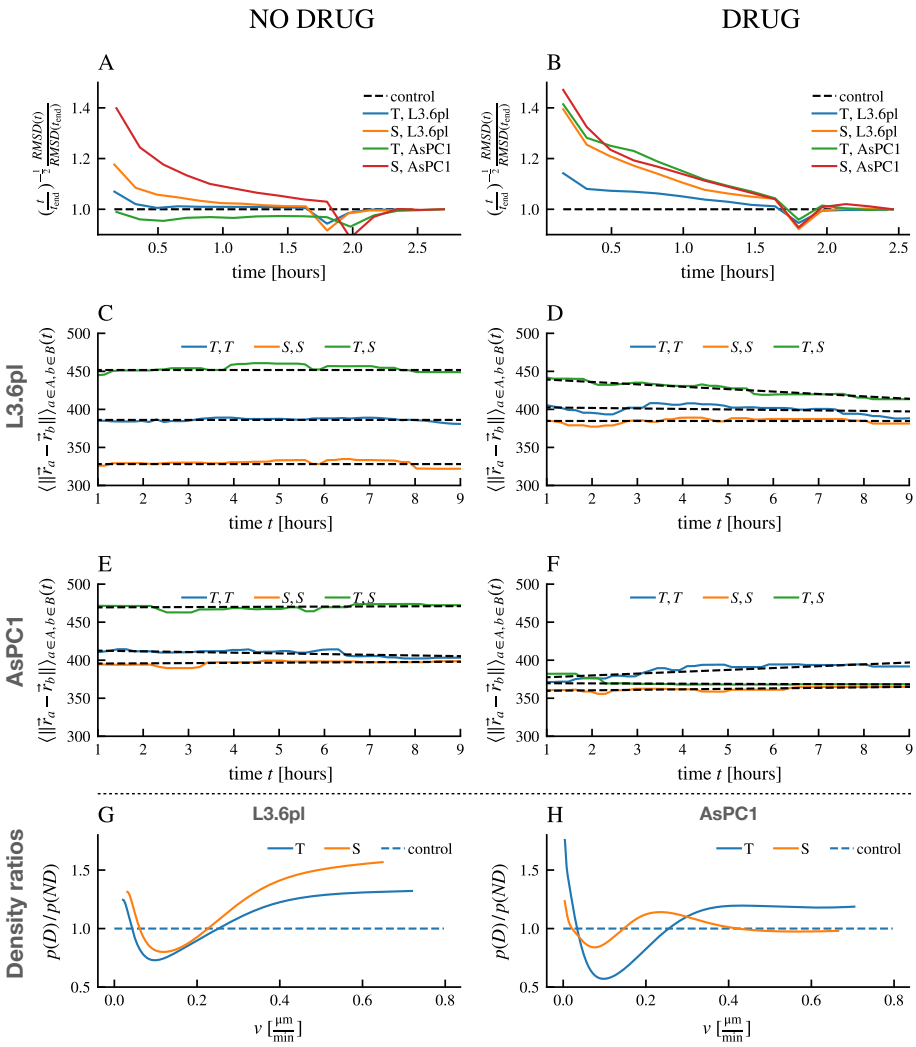

Figure 3.8: *Evolution of Inter-Cellular Distances (ICD) vs time.* Upper line panels: ratio of the RMSD over a standard diffusion $\langle x^2 \rangle \sim t$ for both tumour (blue) and stromal (orange) cells (dataset with drug: panel A; dataset without drug: panel B). The control (dashed black line) is the Brownian pure diffusion reference.
Second line panels: distributions, for the L3.6pl case , of mean intercellular distances $D_{T,T}(t), D_{T,S}(t), D_{S,S}(t)$ (see Eq. (5). Dataset with drug: panel C; dataset without drug: panel D. The trajectories, that are almost ballistic, show drastic differences in the evolution of these quantifiers over time: overall, in the dataset with gemcitabine, cells show more activity and more capacity of reducing relative ICD w.r.t. the cells belonging to the drug-less dataset, suggesting that -due to gemcitabine- the two different populations of cells tend to stick together (i.e. $D_{T,S}(t)$ is a monotonic decreasing function in time).
Third line panels: distributions, for the AsPC-1 case , of mean intercellular distances $D_{T,T}(t), D_{T,S}(t), D_{S,S}(t)$ (see Eq. (5). Dataset with drug: panel E; dataset without drug: panel F. The main difference w.r.t. the L3.6pl counterpart is that it is no longer true that $D_{T,S}(t)$ decreases in time, it remains roughly constant (suggesting that dialogues among different cell lines is suppressed in this case).
Fourth line panels: ratio among the velocity distribution in presence of drug over distribution of velocities in absence of drug, for the L3.6pl case (panel H) and the AsPC-1 case (panel H): we highlight that, while in panel G both the stroma and the tumor acquire motility (as both the orange and blue curve are above one for higher values of velocity ν), this does not happen in the AsPC-1 case, where solely the tumoral line acquires motility.

inferential outcomes by maximum entropy extremization of the previous section.
Indeed, if we plot the ratio of the distributions for the two cellular lineages, namely if we plot the density ratio *(drug distribution)/ (no drug distribution)* for both stromal and tumour cells, as presented in panels G and H for the L3.6pl and the AsPC-1 cases respectively, we see that -for the L3.6pl kinetics- the effect of gemcitabine is to speed up above a critical threshold (that is slightly different between S and T cells resulting in $\sim 0.3\mu m/min$ and $0.2\mu m/min$ respectively) the bulk of all the cells, that sensibly acquire motility (this phenomenon is by far more pronounced in the stromal lineage, as the latter is possibly approaching cancerous clumps and it is coherent with the raise of the mixed interactions we inferred in the previous section, see Figure 3.2 panel H). In the AsPC-1 counterpart, instead, stromal cell's dynamics result almost unaffected by the presence of gemcitabine also from this perspective (coherently with panel K of Figure 3.2 where mixed interactions have not been detected).
Finally, we can correlate the outcomes of the effects of the drug by counting live/dead cells by flow cytometry and relating these results with previous findings: $5\mu M$ gemcitabine decreased L3.6pl cell proliferation as compared to the control group but it did not affect AsPC-1 proliferation. In particular, manual and automatic counting of dead and live cells showed high cell death in the tumour core (roughly 50%) for the L3.6pl case, while the stromal core resulted highly resistant to the treatment (Figure 3.9 panels A, B, C).

Crowding as a stochastic process

Called $\vec{r}_i(t) := (x_i(t), y_i(t))$ the position of the i^{th} cell at time t, the first quantifier we consider stems from stochastic process theory and is the root mean square displacement of any cell, defined by

$$RMSD[\vec{r}_i(t)](\Delta t) = \sqrt{\langle ||\vec{r}_i(t+\Delta t) - \vec{r}_i(t)||^2 \rangle_t} \tag{3.28}$$

where the $\langle\rangle_t$ is an average over all possible choices of time t for which we have $\vec{r}_i(t+\Delta t)$ and $\vec{r}(t)_i$ and $i \in (1, ..., N)$ labels the N cells: the underlying idea in its usage is that, without the presence of a macroscopic gradient to sense [112], in the long term limit cells are expected to behave randomly in accordance with Brownian motion, whose formalization is usually mathematically achieved via a Wiener process, hence we expect that roughly $RMSD[\vec{r}_i(t)](\Delta t) \sim \sqrt{(\Delta t)}$, possibly perturbed by some persistence cells may display [88], see the first rows of Figures $8 - 11$.
The Wiener process can be described via the following equation

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \vec{\mu}\ \Delta t + \sigma\, \vec{\epsilon}_t\ \sqrt{\Delta t} \tag{3.29}$$

where $\vec{\mu}$ is a vector accounting for the ballistic component of motion (i.e., a possible drift) while ϵ_t is a random vector whose entries are distributed according to the standard probability density $\mathcal{N}(0, 1)$ that acts as the source of fluctuations. This very simple model has the advantage that the $RMSD(\Delta t)$ can be obtained analytically

$$RMSD(\Delta t) = \sqrt{\langle \|\vec{r}(t+\Delta t) - \vec{r}(t)\|^2 \rangle_\epsilon} = \sqrt{\mu^2 (\Delta t)^2 + d\, \sigma^2 \Delta t} \tag{3.30}$$

where $d = 2$ is the number of dimensions in which the vectors lie. The $RMSD$ obtained from (3.30) has to be compared to the empirical $RMSD$ of each observed path generated by each individual cell. As each cell can obey a different Weiner process (thus there can be in principle N different values of μ, σ coupled to the N cells), we can actually think to have distributions $P(\mu),\ P(\sigma)$ (whose extractions return the observed μ, σ values) that

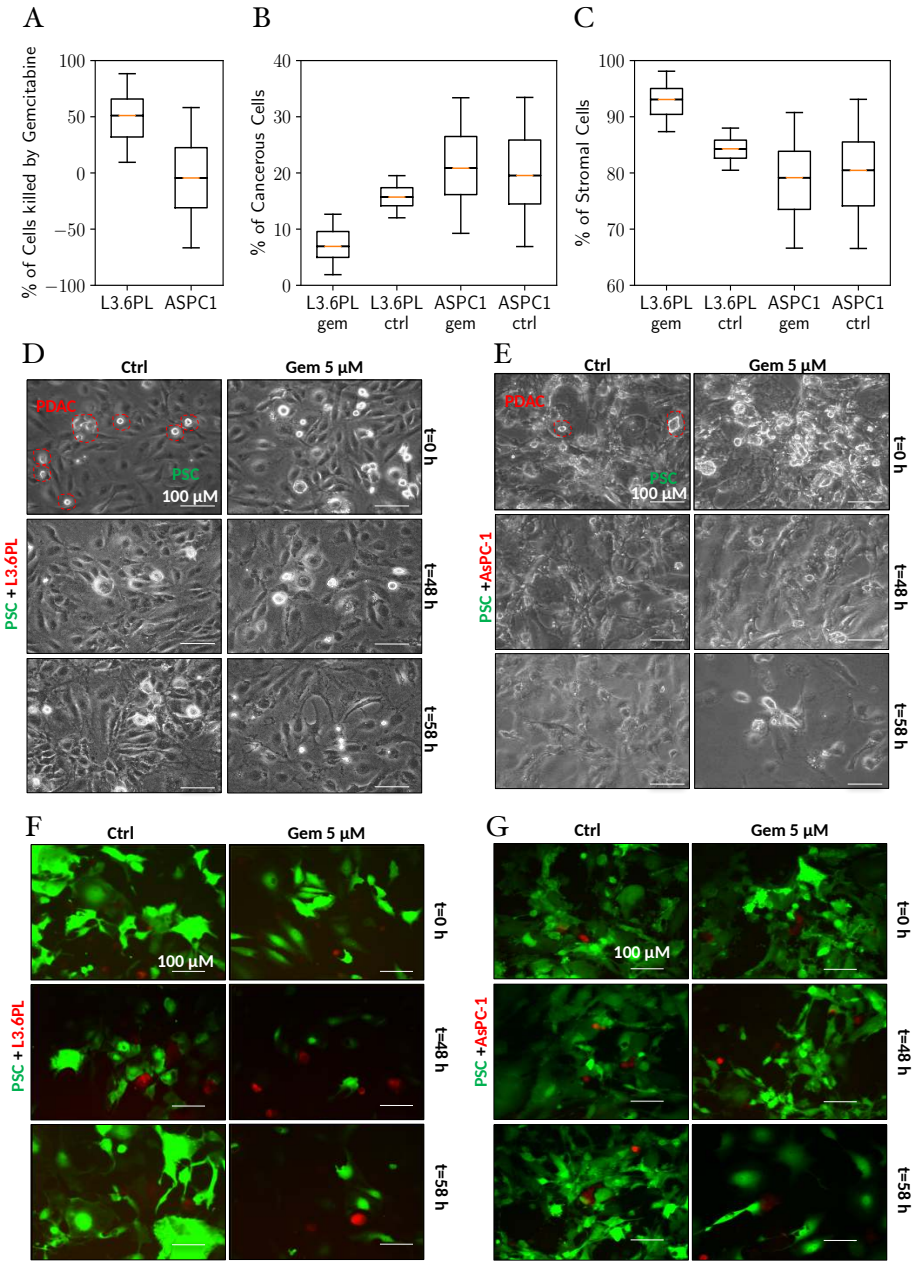

Figure 3.9: *Tumor cells and stromal cells co-cultured for 58 hours.* In the first raw, panels A, B and C report on the count of cells killed by gemcitabine (in panel B the percentage of cells treated or untreated with $5\mu M$ gemcitabine is counted with haemocytometer at the indicated times, $p < 0:05$, $n \geq 3$ and in panel C the percentage of cells treated or untreated with $5\mu M$ gemcitabine counted with flow cytometer at 58 hours. $p < 0.05$, $n \geq 3$: it shines that the L3.6pl line is highly affected by the drug, while the AsPC-1 counterpart is not. In panels D and F we show the representative brightfield for L3.6pl and AsPC-1 cases respectively, while panels E and G show fluorescent images of L3.6pl -panel F- and AsPC-1 -panel G- (green) and PSCs (red) co-cultured cells growth in the presence or absence of $5\mu M$ gemcitabine for 0, 48 and 58 hours. Scale bars: $100\mu m$.

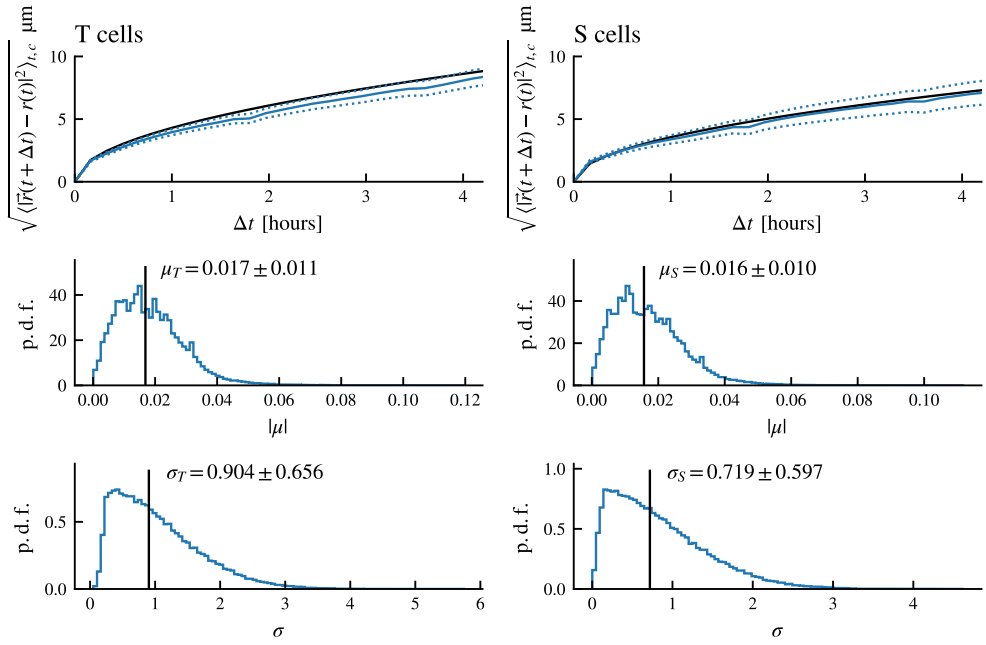

Figure 3.10: case L3.6pl. Fit of the model (3.29) to the D dataset $RMSD$ for both cell types, tumour cells on the left, stromal cells on the right. For each parameter μ and σ the values of median and interquartile range are available as a measure of location and dispersion, it is evident that the drift component in the Wiener process is very weak as $\frac{\mu}{\sigma} \ll 1$..

we assume as Gaussian $\mathcal{N}(m_\mu, v_\mu)$ with unknown mean m_μ and variance v_μ. These means and variances are assumed uniformly distributed over a physically plausible range of values (hence they are centered in zero and share an unreasonably large variance of value 4 that ensures that it acts as an upper bound on the real one) that we reduce self-consistently (vide infra). Calling $\mathcal{U}$ the uniform distribution, we can write

$$\begin{aligned} m_\mu &\sim \mathcal{U}(0,4),\ v_\mu \sim \mathcal{U}(0,4) && (3.31)\\ m_\sigma &\sim \mathcal{U}(0,4),\ v_\sigma \sim \mathcal{U}(0,4) && (3.32)\\ \mu &\sim \mathcal{N}(m_\mu, v_\mu),\ \sigma \sim \mathcal{N}(m_\sigma, v_\sigma) && (3.33)\\ RMSD(\vec{r}, \Delta t) &\sim \sqrt{\mu^2(\Delta t)^2 + d\,\sigma^2 \Delta t} \quad \forall \Delta t, \forall \vec{r} \in Cells. \end{aligned}$$

To calculate the distributions of these parameters μ, σ we exploited the framework of Approximate Bayesian Computation (in particular we used the algorithm developed in [117]): we assumed as a metric distance between the simulated datasets and the original datasets the maximum of Kolmogorov-Smirnoff distances between the empirical RMSD and the simulated RMSD at each time point. Optimal values are reported in the second and third lines of panels in Fig. 3.10 for the L3.6pl drugged dataset, in Fig. 3.11 for the AsPC-1 drugged dataset, in Fig. 3.12 for the L3.6pl non-drugged dataset and in in Fig. 3.13 for the AsPC-1 non-drugged dataset.
This preliminary analysis shows that cells do not exhibit purely ballistic motion (as expected as there is no global chemotactic gradient in the experiments), although the distribution for μ is systematically biased toward positive (but small) values in every plot in accordance with mild persistency.

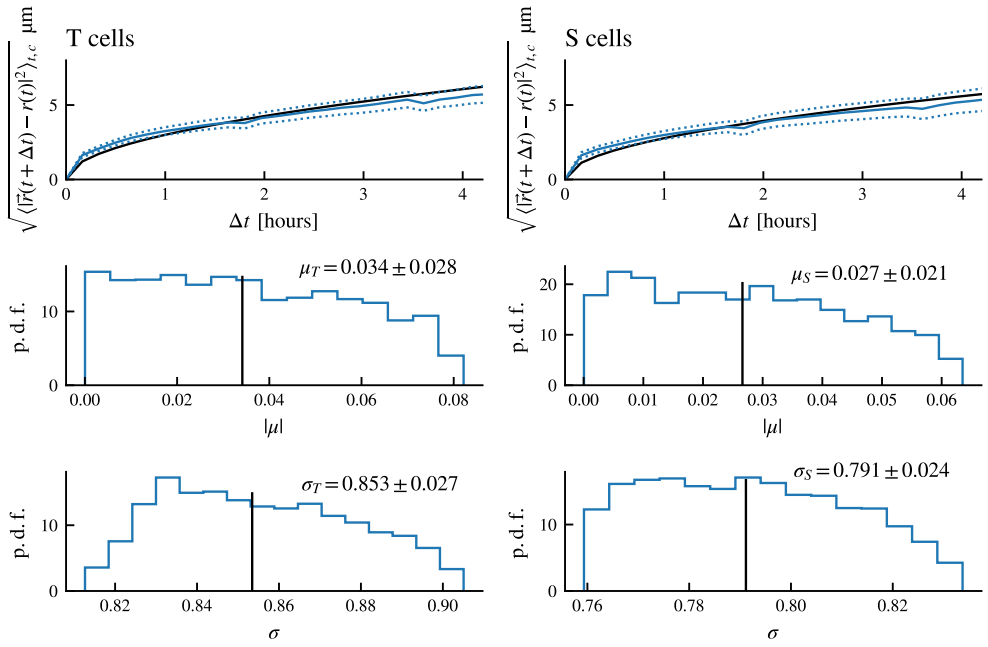

Figure 3.11: case AsPC-1. Fit of the model (3.29) to the D dataset $RMSD$ for both cell types, tumour cells on the left, stromal cells on the right. For each parameter μ and σ the values of median and interquartile range are available as a measure of location and dispersion, it is evident that the drift component in the Wiener process is very weak as $\frac{\mu}{\sigma} \ll 1$..

3.2.5 Discussion on the first experiment

By means of time-lapse confocal imaging we were able to track two different cell populations (tumour and stroma cells) co-cultured in presence (or in absence, for comparison) of a chemoterapeutic agent (i.e. gemcitabine): the resulting data-sets storing cell's positions and velocities resulted in sufficient information to infer the effect of the drug on stroma-cancer kinetics, as well as their dynamical cross-talk, due to a novel computational algorithm we developed. Focusing on cell's velocities, we analyze the directions and moduli of these vectors separately: the former are investigated via maximum-entropy inference, the latter are studied via stochastic processes, resulting overall into a unified synergic approach where global coherence can be appreciated.
By performing the same analysis with and without the presence of gemcitabine on two different malignant lineages, namely the L3.6pl and the AsPC-1 test cases, by comparison of their kinetic responses we can quantify the effect of the drug on these dynamics: we prove that, for the L3.6pl case, the drug added to the cell medium highly increases interactions among cancer and stroma, much more than interactions within the same lineage (that is almost left invariant for the tumour and it is actually diminished for the stroma). As a result of such enhanced interactions, cells tend to form cluster and, locally, the dynamics of the involved cells is no longer diffusive but ballistic, resulting in a marked acquired motility. In the AsPC-1 counterpart, instead, the effect of the drug is sensibly milder: nor mixed interactions raise due to gemcitabine, neither the stochastic dynamics of the cells acquires enhanced motility. Correlating these findings with counts on dead/live cells, we find that while in the AsPC-1 case, cancer progression kept almost unperturbed, in the L3.6pl scenario, the drug killed roughly $\sim 50\%$ of the cancerous cells without affecting the

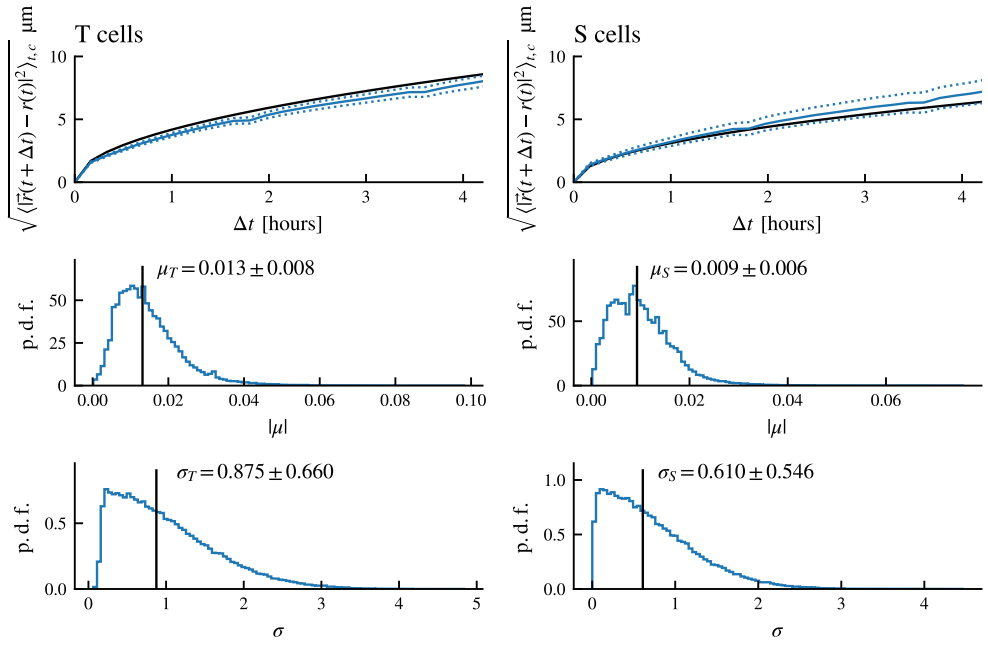

Figure 3.12: case L3.6pl. Fit of the model (3.29) to the ND dataset $RMSD$ for both cell types, tumour cells on the left, stromal cells on the right. For each parameter μ and σ the values of median and interquartile range are available as a measure of location and dispersion, it is evident that the drift component in the Wiener process is very weak as $\frac{\mu}{\sigma} \ll 1$.

vitality of the stroma.
Focusing on research aspects, these results contribute to enlarge the amount of techniques and tools (integrating those mainly -omics derived) we have to quantify drug response to cancer: while it is true that we can not identify sharply eventual signaling molecules, yet, as it is well known that stellate stromal cells of the pancreas, when activated, migrate and secrete components of the extracellular matrix such as type I collagen, chemokines and cytokines to which the movement of the cell is also linked, our approach can be used in broad generality to help profiling which of these molecules (or related receptors) contribute to the interactions at the core of cell's kinetic coordination: for example, if we get a high J_M -as for the L6.36pl cells case in presence of gemcitabine (as shown by the maximum entropy inference approach)- and then we could measure the concentration of various molecules and/or the expression levels of various proteins with specific assays, we can get information on which molecules (or related receptors) is involved in the cross-talk and how the latter results pivotal under the administration of a particular drug.
Further, focusing on clinical aspects, as stroma can play a very broad critical role –ranging from cancer fighter to cancer facilitator– our protocol could help (at a very cheap cost) to quickly understand whether the stroma-tumor interaction harms the therapy or not: for example, in the case of L6.36pl cells we see that the interactions between tumor and stroma raise sensibly when treated with gemcitabine and, coherently, also crowding effects do appear (as shown by the stochastic process approach). As these data correlate with cell-counting data indicating that a high percentage of L6.36pl cells are killed by gemcitabine, it can be hypothesized that in this case the stromal cells do not inhibit the action of the chemotherapeutic agent and that the related crowding among the two cellular populations

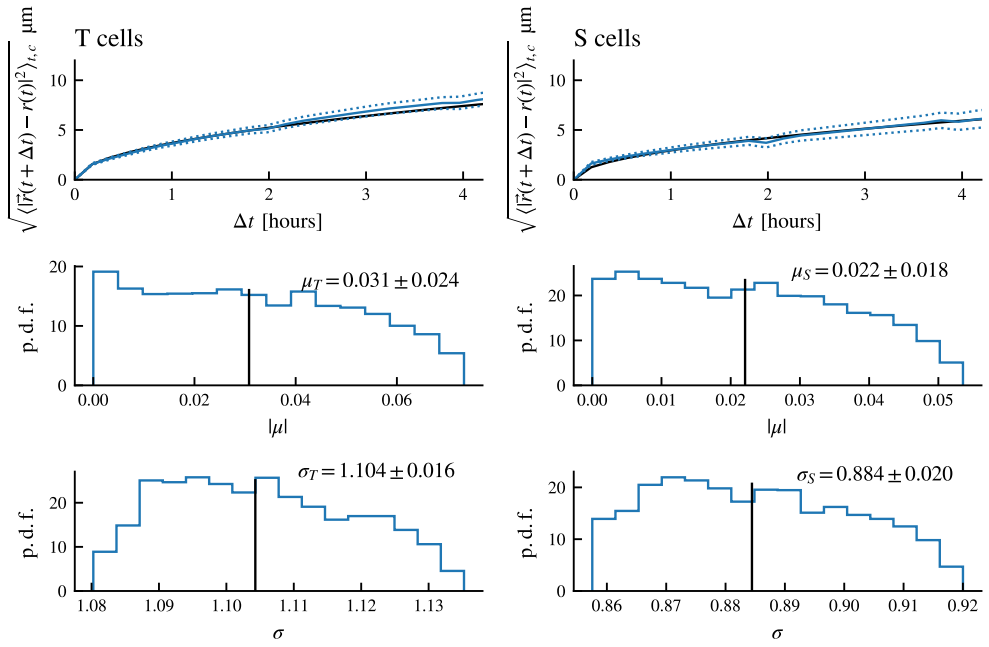

Figure 3.13: case AsPC-1. Fit of the model (3.29) to the ND dataset $RMSD$ for both cell types, tumour cells on the left, stromal cells on the right. For each parameter μ and σ the values of median and interquartile range are available as a measure of location and dispersion, it is evident that the drift component in the Wiener process is very weak as $\frac{\mu}{\sigma} \ll 1$.

is likewise important for a successful therapy.

3.3 Problem Two: Maximum Entropy for Heart Rate Variability

Heart-rate variability (HRV) analysis constitutes a major tool for investigating the mechanisms underlying the complex and chaotic cardiac dynamics as well as for identifying general features discriminating the clinical status of patients [118, 119, 120, 121, 122, 123, 124]. To this aim a fundamental observable is the RR series $\mathbf{r} = \{r_1, r_2, ...\}$, where r_n is the temporal distance between the n-th and the $(n+1)$-th R peaks in a ECG recording (see Fig. 3.14 left panel). Several approaches have been carried out in the past in order to address the HRV analysis from this observale (see e.g., [125, 126, 127]) about the RR series: by paving this route the problem of classification of heart failures via time-series analysis is translated into a search for clusterization in a high-dimensional space[1].

Interestingly, the intrinsic variability in heat rate ultimately stems from the interplay of the sympathetic and the parasympathetic nervous system. In this work, exploiting glassy statistical inference approaches, we aim to unveil any signature of this underlying autonomic neural regulation. To this scope we will study HRV in the temporal and in the frequency domain, and at different levels of aggregation (in the higher one the sample is made of all available data, in the lower one we build different sub-samples pertaining to patients displaying a different clinical status: healthy, suffering from cardiac decompensation, suffering from atrial fibrillation).

The statistical inference approach we adapt to the present case of study is the leight-motif of the whole thesis, namely the maximum entropy framework. By this technique, we search for the minimal structured probabilistic model compatible to our data; more precisely, we consider the family of probability distributions $P(\mathbf{r})$ over the sequences of inter-beat intervals $\mathbf{r}$ whose lowest momenta match the empirical ones and, among all the elements of this family we select the one corresponding to the maximum entropy. As a direct consequence of the definition of entropy in terms of the logarithm of the probability distribution $P(\mathbf{r})$ over the inter-beat sequences $\mathbf{r}$, this approach returns an exponential family $P(\mathbf{r}) \sim e^{-H(\mathbf{r})}$, where $H(\mathbf{r})$ can be interpreted as a cost function (or Hamiltonian in a physical jargon). By requiring a match on the first two moments only (i.e., by requiring that the theoretical average and two-point correlation provided by the model are quantitatively consistent with the empirical ones), $H(\mathbf{r})$ results in a pairwise (r_n, r_m) cost-function, as standard in Physics (see the first Section of this Chapter and the Section 1.1.2 at the beginning of the thesis.

Of course, recovering the complex structure hidden in RR series by a relatively simple pairwise model has several advantages: on the one hand, the low number of parameters prevents from over-fitting, on the other hand, the inferred cost-function can be framed in a statistical mechanics context (see e.g., [15, 93, 95, 129, 130]) and we can therefore rely on several powerful techniques and on a robust Literature (as for instance those provided in the first Chapter of the present thesis). In particular, we will show that, despite its simplicity, such a pair-wise model is able to capture the complex nature of the temporal correlation between beats which emerges experimentally; in fact, the coupling between two beat-intervals r_n and $r_{n+\tau}$ turns out to be long-range (i.e., displays a power-law decay with the distance τ) and frustrated (i.e., the couplings between two beats can be positive and negative). In a statistical-mechanical jargon, this system is referred to as a two-body spin-glass with power-law quenched interactions.

Remarkably, frustration in couplings, which is a key feature of spin-glasses, means the existence of competitive driving forces and it is natural to look at this emerging feature in our

[1] Via machine learning approaches, our group addressed that perspective in [124, 128].

model as the hallmark of the interplay between the parasympathetic and orthosympathetic systems (indeed, while the first one tends to increase the distance between RR peaks, i.e., to lower the heart rate [122], the latter tends to decrease it [131]). We speculate that these competing interactions may be responsible for the well-known $1/f$ noise shown by HRV [132, 133, 134]: spin-glasses typically display a chaotic dynamics [135, 136, 137] spread over several timescales [138] and their power spectrum density is power-law [139, 140, 141]. In fact, here we show that the autocorrelation in the $\{r_n\}$ series decays in the beat number as n^{-1} and its related power spectrum decays in frequecy as f^{-1}. Incidentally, we notice that variables whose fluctuations display $1/f$ noise are widespread, ranging from inorganic (e.g., condense [142], granular [143], etc.) to organic matter (e.g. in DNA sequences [144], membrane channels [145]) and, even broadly, in Nature (e.g. ranging from earthquakes [146] to off-equilibrium flows of current trough resistors [147], to the whole self-organized criticality [148, 149]).

3.3.1 Summary of experimental data

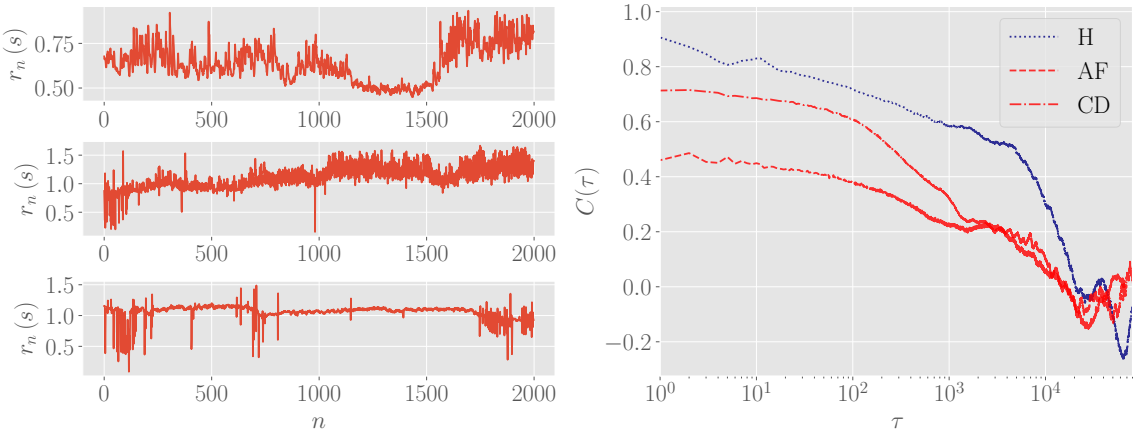

Figure 3.14: Left: examples of the bare RR time series for a single patient for each class; the window depicted is restricted to the first 2000 beats. Right: examples of autocorrelation functions for a single patient for each class. The dotted blue line refers to a healthy patients, while red are patients with AF (dashed curve) and CD (dash-dotted line).

In this Section, we give some details about the data and the quantities considered in our analysis.
The database is made of ECG recordings on $M = 348$ patients, wearing an Holter device for nominal 24 hours. From these recordings we extract the RR series

$$\{\mathbf{r}(i)\}_{i=1,\dots,M} = \{r_n(i)\}_{\substack{i=1,\dots,M \\ n=1,\dots,N_i}}, \tag{3.34}$$

where i labels the patient and n labels the number of beats in each sequence (which is order of 10^5 and depends on the patient). Patients belong to three classes, according to their clinical status: healthy individuals (H), individuals with atrial fibrillation (AF) and individuals with congestive heart failure (hereafter simplified as *cardiac decompensation*) (CD). Their number is $M_H = 149$, $M_{AF} = 139$, and $M_{CD} = 60$, respectively; of course, $M = M_H + M_{AF} + M_{CD}$. In Fig. 3.14 (left) we show examples of the series $\mathbf{r}(i)$ for three patients belonging to the different classes.
In order to make a meaningful comparison of the variability among the RR series $\mathbf{r}(i)$ of

different patients, we standardize them with respect to their temporal mean and standard deviation, so that the study of HRV is recast in the study of fluctuations of the standardized RR series around the null-value. More precisely, we introduce

$$z_n(i) = \frac{r_n(i) - \langle \mathbf{r}(i) \rangle}{\mathrm{std}[\mathbf{r}(i)]}, \quad \text{for} \quad n = 1, ..., N \tag{3.35}$$

or, in vectorial notation,

$$\mathbf{z}(i) = \frac{\mathbf{r}(i) - \langle \mathbf{r}(i) \rangle}{\mathrm{std}[\mathbf{r}(i)]}, \tag{3.36}$$

where we defined

$$\langle \mathbf{r}(i) \rangle = \frac{1}{N_i} \sum_{n=1}^{N_i} r_n(i), \quad \langle \mathbf{r}^2(i) \rangle = \frac{1}{N_i} \sum_{n=1}^{N_i} r_n^2(i), \quad \mathrm{std}[\mathbf{r}(i)] = \sqrt{\langle \mathbf{r}^2(i) \rangle - \langle \mathbf{r}(i) \rangle^2}. \tag{3.37}$$

The raw histograms for the standardized inter-beat intervals in the three classes of patients are shown in Fig. 3.15: notice that the frequency distributions exhibit heavy-tails.

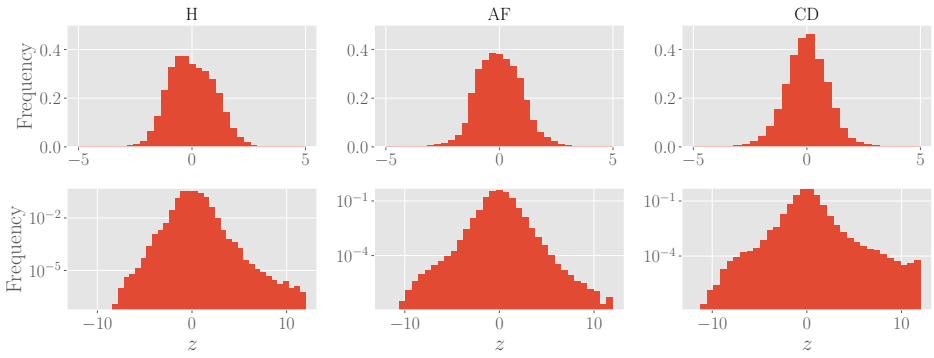

Figure 3.15: Histograms of the standardized values $\{\mathbf{z}(i)\}$ divided by classes: left panels are build by collecting data from healthy patients, middle panels are build by collecting data from patients suffering from atrial fibrillation and right panels are build by collecting data from patients suffering from cardiac decompensation. In the first row, we reported relative frequencies in the natural scale, while the second row we reported relative frequencies in the logarithmic scale.

We consider the points in the standardized RR series as random variables sampled by a hidden stochastic process, in such a way that the value of $z_n(i)$ at a given step n depends in principle on all the values $\{z_m(i)\}_{m<n}$ taken in the previous steps $m < n$ since the beginning of sampling. From this perspective, a meaningful observable to look at is the auto-correlation function at a distance τ, defined as

$$C(i, \tau) = \frac{1}{N} \sum_{n=1}^{N_i - \tau} \big(z_n(i) - \langle z(i) \rangle_+\big)\big(z_{n+\tau}(i) - \langle z(i) \rangle_-\big), \tag{3.38}$$

where $\langle z(i) \rangle_+ = \frac{1}{N_i - \tau} \sum_{n=1}^{N_i - \tau} z_n(i)$ and $\langle z(i) \rangle_- = \frac{1}{N_i - \tau} \sum_{n=\tau}^{N_i} z_n(i)$. Given the standardization over the whole segment $[1, N_i]$, as long as $\tau \ll N_i$, we expect that $\langle z(i) \rangle_+$ and $\langle z(i) \rangle_-$ are both close to zero and shall be neglected in the following (indeed, we checked that this is the case, since $\langle z(i) \rangle_+, \langle z(i) \rangle_- \sim 10^{-15} \div 10^{-17}$). Then, the auto-correlation

function we measure simply reduces to

$$C(i,\tau) = \frac{1}{N}\sum_{n=1}^{N_i-\tau} z_n(i) z_{n+\tau}(i). \tag{3.39}$$

Some examples of the autocorrelation function for patients of the three classes are reported in the right plot of Fig. 3.14, where we stress that the autocorrelation is non-null over a large τ window and its shape is patient-dependent.

Finally, we introduce a further average operation, this time on the sample of patients, namely, we define

$$\mathbb{E}_{class}(\mathbf{z}) = \frac{1}{M_{class}}\sum_{i\in class} \mathbf{z}(i), \tag{3.40}$$

$$\mathbb{E}_{class}(\mathbf{z^2}) = \frac{1}{M_{class}}\sum_{i\in class} \mathbf{z}^2(i), \tag{3.41}$$

$$\mathbb{E}_{class}(C(\tau)) = \frac{1}{M_{class}}\sum_{i\in class} C(i,\tau). \tag{3.42}$$

where $class \in \{H, AF, CD\}$ and, with "$i \in class$" we mean all the indices corresponding to patients belonging to a certain class. In the following we will consider the vectors $\mathbf{z}$ as random variables sampled from an unknown probability distribution $P^{true}_{class}(\mathbf{z})$, which we will estimate by the probability distribution $P_{class}(\mathbf{z})$ characterized by a minimal structure and such that its first and second moments are quantitatively comparable with $\mathbb{E}_{class}(\mathbf{r})$, $\mathbb{E}_{class}(\mathbf{z^2})$ and $\mathbb{E}_{class}(C(\tau))$, respectively.

3.3.2 On the model and on the inferential procedure

Our atomic variable is the sequence $\{z_1, z_2, ..., z_N\}$ and, as anticipated above, we denote with $P(\mathbf{z})$ the related probability distribution emerging from the inferential operations on the sample of experimental data. The Shannon entropy $\widetilde{H}[P(\mathbf{z})]$ associated to $P(\mathbf{z})$ is

$$\widetilde{H}[P(\mathbf{z})] = -\int \mathrm{d}\mathbf{z} P(\mathbf{z}) \ln P(\mathbf{z}). \tag{3.43}$$

According to the maximum entropy principle, we look for the distribution $P(\mathbf{z})$ that maximizes $\widetilde{H}[P(\mathbf{z})]$ and such that its moments match those evaluated experimentally, in particular, here the we choose to apply the constraints on the one-point and two-points correlation function that is, $\mathbb{E}_{class}(\mathbf{z})$, $\mathbb{E}_{class}(\mathbf{z^2})$ and $\mathbb{E}_{class}(C(\tau))$, respectively. To lighten the notation hereafter these moments shall be referred to simply as, respectively, $\mu^{(1)}$, $\mu^{(2)}$ and $C(\tau)$, without specifying the class. In fact, the inferential procedure works analogously regardless of the class, the latter affecting only the quantitative value of the parameters occurring in $P(\mathbf{z})$. Constraints are set via Lagrange multipliers $(\lambda_0, \lambda_1, \lambda_2, \lambda_\tau)$ in such a

way that the problem is recast in the maximization of the functional

$$\begin{aligned}\widetilde{H}_{\lambda_0,\lambda_1,\lambda_2,\lambda_\tau}[P(\mathbf{z})] = \widetilde{H}[P(\mathbf{z})] &+ \lambda_0\left(\int \mathrm{d}\mathbf{z}\ P(\mathbf{z}) - 1\right) + \\ &+ \lambda_1\left(\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n - N\mu^{(1)}\right) + \\ &+ \lambda_2\left(\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n^2 - N\mu^{(2)}\right) + \\ &+ \sum_{\tau=1}^{N}\lambda_\tau\left(\sum_{n=1}^{N-\tau}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n z_{n+\tau} - (N-\tau)C(\tau)\right),\end{aligned} \tag{3.44}$$

where integration is made over $\mathbb{R}^N$. Note that, while the derivation with respect to λ_1, λ_2 and λ_τ ensure, respectively, the agreement between the theory and the experiments at the two lowest orders, i.e. the temporal average $\mu^{(1)}$, the second moment $\mu^{(2)}$ and the auto-correlation function $C(\tau)$, λ_0 guarantees that $P(\mathbf{z})$ is normalized, so that $P(\mathbf{z})$ is a probability distribution function. In the asymptotic limit of long sampling ($N \to \infty$) and under a stationarity hypothesis (see [150] for a similar treatment), the solution of the extremization procedure, returning the probability of observing a certain sequence $\mathbf{z}$, is given by

$$P(\{z_n\}_{n=1}^{\infty}) = \frac{1}{Z}\left(\prod_{n=1}^{\infty}P_0(z_n)\right)\exp\left(\sum_{n=1}^{\infty}\sum_{\tau=1}^{\infty}J(\tau)z_n z_{n+\tau} + h\sum_{n=1}^{\infty}z_n\right), \tag{3.45}$$

where h and $J(\tau)$ can be estimated from available data (*vide infra*). Here, P_0 is the $\mathcal{N}(0,1)$ distribution and plays the role of prior for the variable z_n, the parameter $J(\tau)$ represents the pairwise interaction between elements at non-zero distance τ in the series (notice that each element occurs to be coupled to any other), and the parameter h represents the bias possibly affecting the single value in the sequence (and it is expected to be zero as we standardized the RR series). The factor Z plays here as a normalization constant, like the partition function in the statistical mechanics setting [19]. Notice that the interaction between two elements r_n and r_m depends on the distance $\tau = n - m$, but not on the particular couple considered. This stems from a"stationary hypothesis", meaning that one-point and two-point correlation functions calculated on a segment spanning $O(\tau \ll N)$ elements along the series are approximately the same and since the starting time of sampling is arbitrary, we get that $J(n,m) = J(m-n)$.

The standard inference setup for the model parameters is based on a *Maximum Likelihood Estimation* (MLE), i.e. the maximization of the function

$$\mathcal{L}(\mathbf{J}, h|\mathcal{D}) = -\frac{1}{M}\sum_{\mathbf{z}\in\mathcal{D}}\log P(\mathbf{z}|\mathbf{J}, h), \tag{3.46}$$

where $\mathcal{D}$ is the time-series database (of a given class) and where we made clear the dependence of P on the model parameters. However, such an approach requires the computation of the whole partition function Z, which is numerically hard in this case. Then, we chose to adopt as objective function for the inference procedure the *pseudo-(log-)likelihood* function[150]:

$$\mathcal{L}(\mathbf{J}, h|\mathcal{D}) = -\frac{1}{M}\sum_{\mathbf{z}\in\mathcal{D}}\log P(z_{L+1}|\{z_n\}_{n=1}^{L}), \tag{3.47}$$

that is, given L observation in a fixed time-series $\mathbf{z}$, we maximize the conditional probability to observe the value z_{T+1} at the successive time step. Further, we make two main modifications with respect to the standard pseudo-likelihood approach: i) in order to use the entire available time-series in our database, we also adopt a window average procedure; ii) we add regularization terms in order to prevent divergence for the model parameters. A detailed discussion is reported in Section 3.3.5. Our objective function is therefore given by

$$\begin{aligned}\mathcal{L}^{(\text{reg})}(\mathbf{J}, h|\mathcal{D}) =& \frac{1}{M}\sum_{\mathbf{z}\in\mathcal{D}}\Bigg[-\frac{1}{2(N-T)}\sum_{n=T}^{N-1}\Big(z_{n+1}-h-\sum_{\tau=1}^{T}J(\tau)z_{n+1-\tau}\Big)^2 \\ &-\frac{\lambda}{2}h^2-\frac{\lambda}{2}\sum_{\tau=1}^{T}f(\tau)J(\tau)^2\Bigg],\end{aligned} \tag{3.48}$$

where T is the largest τ we want to consider (namely T must be larger that the maximal decorrelation time), λ is the regularization weight and $f(\tau)$ is a temporal regularizer that prevents the elements of $J(\tau)$ to get too large for large τ (see Section 3.3.5 for a detailed description).

This inference method allows us to determine the values of the parameters $J(\tau)$ and h as well as their uncertainties $\sigma_{J(\tau)}$ and σ_h. As for the parameter h, due to series standardization, its value, evaluated over the different classes, is expected to be vanishing (this is indeed the case as it turns out to be $h \sim 10^{-3}$ with a related uncertainty of the same order). As for the pairwise couplings, we find that for all the classes considered, $J(\tau)$ is significantly non-zero only for relatively small values of τ, with a cutoff at $T \sim 10^2$, and, for a given $\tau < T$, the coupling does not display a definite sign, that is, for pairs $(z_n, z_{n+\tau})$ and $(z_m, z_{m+\tau})$ at the same distance τ the related couplings can be of opposite signs. These results are shown in Fig. 3.16: in the left column we reported the inferred $J(\tau)$ with the associated uncertainties for all τ, and in each panel in the right we reported the frequency distributions for the first values of τ as examples.

In order to study how decorrelation of RR intervals takes place, it is interesting to study how interaction vector $J(\tau)$ (regardless of its sign) vanishes as the delay time τ increases. We found that the long delay time behavior of the magnitude (i.e. disregarding the signs and oscillatory characteristics) of the interactions is well-described by a power law of the form

$$|J(\tau)|_{\text{leading}} \sim A \cdot \tau^{-\beta}. \tag{3.49}$$

We thus fitted the tails of the inferred $J(\tau)$ with this trial function (tails are chosen in order to maximize the adjusted R^2 score). In Table 3.1, we report best-fit parameters, the adjusted R^2 score and the reduced χ^2. It is interesting to note that the scaling parameter β is around 1, meaning that, for each of the three classes, the leading behavior of the interactions at large τ is $\sim 1/\tau$ (as we will see later, the same scaling also characterize the power spectral density in the Fourier domain, see Fig. 3.20). In the upper row of Fig. 3.17, we depicted with red circles the results of the inference procedure (once taken their absolute values), while the best fit of the general trend is represented with dashed black lines. In the lower row of the same figure, we also reported the residuals of the experimental values with respect to the best fit (normalized to the corresponding uncertainty). A part for the first few points in the AF and CD cases, we see that the residuals are distributed in a range of at most $2\sigma_{J(\tau)}$ (where $\sigma_{J(\tau)}$ is the standard deviation at each τ point), and, in particular, for sufficiently long τ (where oscillations are softened), experimental values are always contained in the range $[-\sigma_{J(\tau)}, \sigma_{J(\tau)}]$, implying that the leading behavior of the

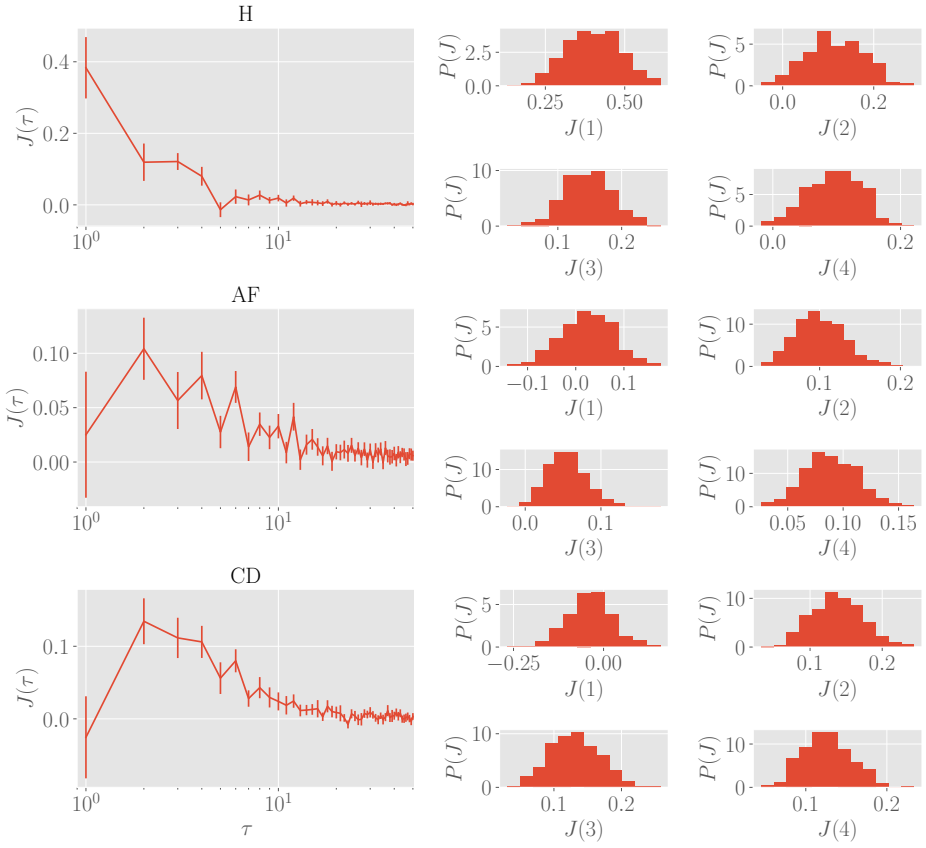

Figure 3.16: Inference results for delayed interactions. Left column: the plots show the results of the inference procedure (distinguishing between the clinical status) for the first 50 τs. Right: frequency distribution of the Js for some selected values of τ (i.e. $\tau = 1, 2, 3, 4$). In both cases,the statistics consists in $M = 500$ different realizations of the $J(\tau)$ which are realized by randomly extracting different mini-batches, each with size $n = 20$. We stress that some frequency distributions present tails on negative values of J for some τ. This means that frustrated interactions are also allowed, implying that the system is fundamentally complex, i.e. a *glassy hearth*.

delayed interaction is well-captured by $\sim 1/\tau$ noise both in our datasets as well as in the model's prediction.

To summarize, a few comments are in order here.

- couplings can be both positive and negative (see Fig. 3.16), defining the heart as a complex glassy system.
- the coupling magnitude decays in τ as a power-law whose leading order is $\sim 1/\tau$ (see Fig. 3.17).
- the coupling magnitude displays a sharp scaling $1/\tau$ solely in healthy patients, while for the remaining patients it display a *bump* in the short time-scales (see Fig. 3.16 and the residual plots in Fig. 3.17).
- by fitting data via Eq. (3.49), we obtain refined estimates for the exponent β (as reported in Tab. 3.1): interestingly, different classes (i.e. different pathologies) are

Class	A	β	$\bar{R}^2$	χ^2/DOF
H	0.38 ± 0.04	1.41 ± 0.04	0.75	0.33
AF	0.20 ± 0.03	0.96 ± 0.05	0.79	0.42
CD	0.37 ± 0.05	1.2 ± 0.05	0.78	0.46

Table 3.1: **Best fit values regarding the scaling reported in Eq.** (3.49). For each class, we report the best-fit parameters A and β, as well as the adjusted R^2 and the reduced χ^2 scores quantifying the fit goodness.

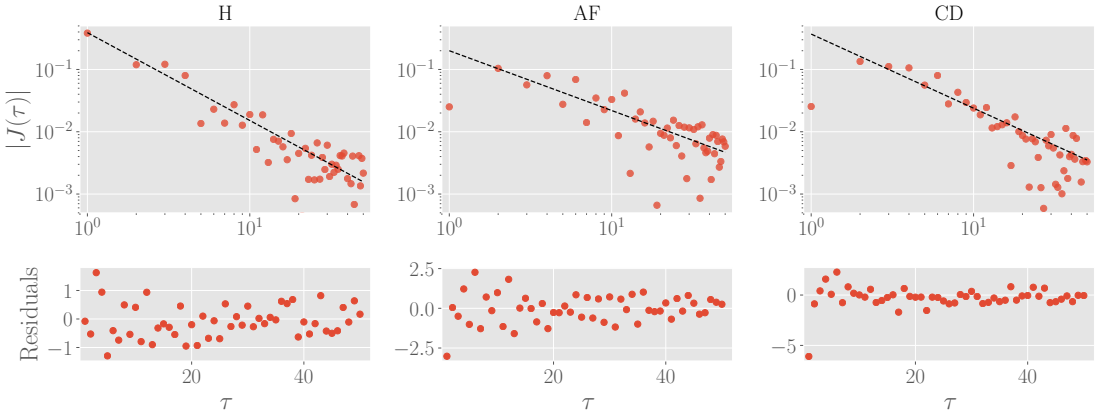

Figure 3.17: Leading behavior of magnitude of the delayed interactions. In the upper panel, we reported the absolute value of the delayed interactions $J(\tau)$ and the relative best fit. In the lower panel, we reported the residuals (normalized by the uncertainty at each point τ) of the experimental data with respect to the best fit function. We stress that, even if the interactions $J(\tau)$ are far from the fitting curve (in the log-log scale, see first row), they are compatible within the associated uncertainties, as remarked by the residual plots.

robustly associated to different best-fit values of β, in such a way that classification of cardiac failures via HRV via this route seems possible.

3.3.3 On the model and on the generalization procedure

Once the model and the related parameters are inferred for each of the three classes, we can use the original sequences $\{\mathbf{z}\}$ to generate synthetic sequences $\{\tilde{\mathbf{z}}\}$ of length $N-T$. The intuition behind the procedure followed to get the synthetic sequence is briefly described hereafter.

For any class, we consider our estimate for $J(\tau)$, along with the estimate $\sigma_{J(\tau)}$ of its uncertainty, and we build the noisy estimate for $J(\tau)$, that is $\bar{J}(\tau) = J(\tau) + \delta J(\tau)$ where $\delta J(\tau) = \eta \sigma_{J(\tau)}$ and η is a $\mathcal{N}(0,1)$ random variable. Next, taken a certain $\{\mathbf{z}\}$, we convolve it with $\bar{J}(\tau)$ and this returns $\{\tilde{\mathbf{z}}\}$. Of course, due to the initial standardization of the RR series, the inference procedure returned a vanishing bias parameter h, hence the synthetic series will also be centered at zero. However, a synthetic sequence is no longer standardized and this is done by hand.

Then, it is natural to compare the synthetic sequences and the experimental ones. We generate a sample of data with the same size of the experimental data available, and we compute the empirical cumulative distribution function for both the experimental and syn-

thetic data in order to compare them: results are reported in Fig. 3.18. In the first row, we directly compare the experimental (red solid line) and the synthetic (black dashed line) cumulative distributions highlighting an excellent agreement for all the classes. This is then corroborated by checking the probability plots in the same figure (second row): here, the red solid line shows the synthetic cumulative distribution versus experimental cumulative distribution, while the black dashed curve is the identity line. The green regions in the plot are confidence intervals with $p = 0.95$.
Next, we test whether the model is able to effectively capture correlation in the RR series, in particular by comparing experimental auto-correlation functions and their predicted counterparts. However, since autocorrelation functions are individual-dependent, starting from a single (randomly chosen) RR series we generate 100 synthetic series with different realizations of the $\bar{J}(\tau)$ according to the above mentioned procedure. In this way, we can use our estimation of the uncertainties on $J(\tau)$ in order to give a confidence interval for our predictions. In Fig. 3.19 we compare the auto-correlation functions for the experimental series and for the synthetic series; for the latter we also highlight the confidence interval with $p = 0.68$. More precisely, we depict the experimental (red solid line) and theoretical (black dashed line) auto-correlation functions and see that the former always fall inside the confidence interval of the re-sampled series (the green region). Thus, we can conclude that our inferred minimal pairwise model is able to effectively capture the temporal autocorrelation in the RR series.

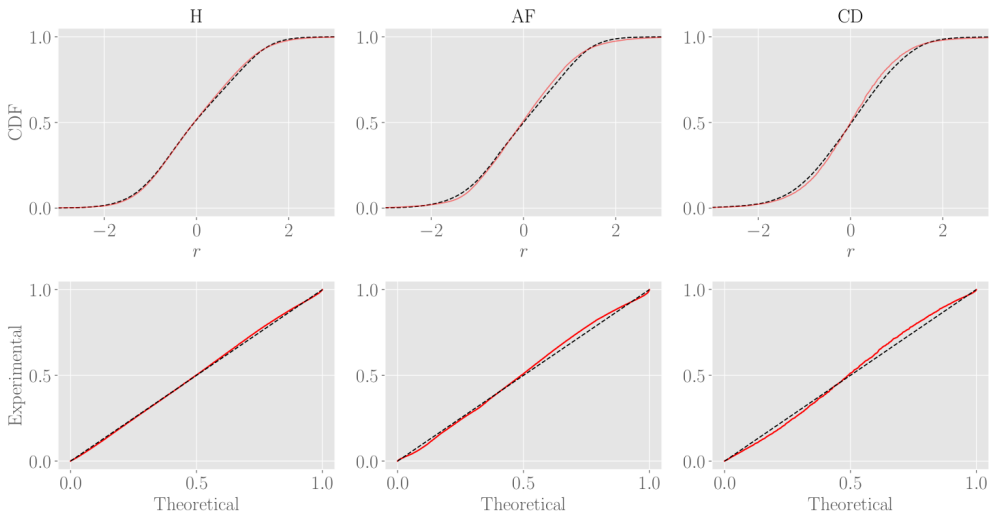

Figure 3.18: Comparison between posterior distributions for experimental and synthetic data. First row: comparisons between the empirical cumulative distributions for both experimental (solid red lines) and resampled (black dashed lines) populations for all of the three classes. Second row: probability plots for the two populations of data (i.e. empirical *versus* theoretical ones, red solid lines) for all of the three classes. The black solid curves are the identity lines for reference. The green region is the confidence interval with $p = 0.95$.

As a final comment, we also looked at the power spectrum density (PSD) of the provided datasets $\{\mathbf{z}\}$ that, as expected (see e.g., [132, 133, 134]), displays the long tail $1/f$ (see Fig. (3.20), upper panel) and we made the following comparison: for all the patients, we evaluated its PSD and in the region $[10^{-4}, 10^{-2}]$ Heartbeat^{-1} we fit with a power-law

$$\mathrm{PSD}(f) = \alpha \cdot f^{-\gamma}$$

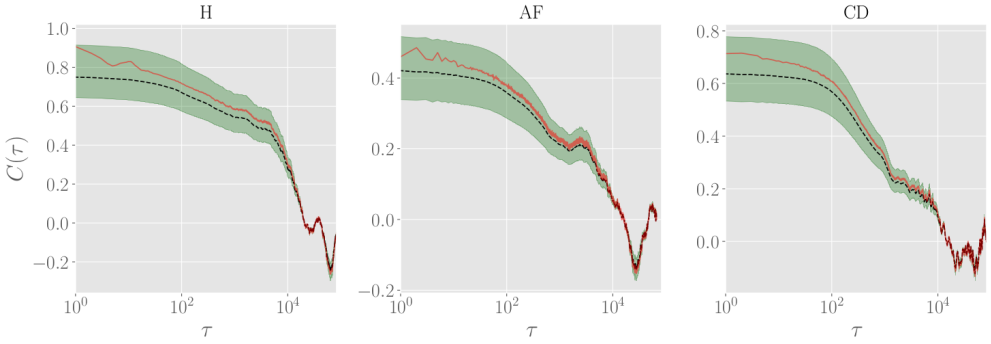

Figure 3.19: Comparison between autocorrelation functions for experimental and synthetic data. The autocorrelation function for one patient randomly extracted from the experimental data-set (red solid lines) is compared with the median autocorrelation function obtained from the synthetic dataset (black dashed lines). Notice that the former falls in the confidence interval with $p = 0.68$ (green region) of the latter.

where $\gamma \sim 1$ and its value is taken as the x-coordinate of that patient in the lower panels of Fig. (3.20). The corresponding y-value is obtained by calculating the PSD over 100 synthetic RR-series generated by convolution with the empirical series playing as seed and using as value of $\boldsymbol{J}$ the one pertaining to the class the patient belongs to (H, AF, CD); results are in good agreement on the diagonal.

3.3.4 Pairwise correlations from maximum entropy principle

The probabilistic model we use to frame the analysis of heart-rate variability contained in the standardized RR series $\{z_n\}_{n=1}^N$ emerges as the solution of extremization procedure of the constrained Shannon entropy functional

$$\begin{aligned}\widetilde{H}_{\lambda_0,\lambda_1,\lambda_2,J}[P(\mathbf{z})] = \widetilde{H}[P(\mathbf{z})] &+ \lambda_0\left(\int \mathrm{d}\mathbf{z}\ P(\mathbf{z}) - 1\right) \\ &+ \lambda_1\left(\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n - N\mu^{(1)}\right) + \lambda_2\left(\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n^2 - N\mu^{(2)}\right) \\ &+ \sum_{\tau=1}^{N}\lambda_\tau\left(\sum_{n=1}^{N-\tau}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n z_{n+\tau} - (N-\tau)C(\tau)\right).\end{aligned} \tag{3.50}$$

Here, we recall that the first term is the standard Shannon entropy for probability distribution for continuous variables, i.e.

$$\widetilde{H}[P(\mathbf{z})] = -\int \mathrm{d}\mathbf{z} P(\mathbf{z})\log P(\mathbf{z}), \tag{3.51}$$

while the other terms are constraints with Lagrangian multipliers λ_0, λ_1, λ_2 and λ_τ. The extremization with respect to these parameters leads to the following conditions:

$$\int \mathrm{d}\mathbf{z}\ P(\mathbf{z}) = 1 \quad , \quad \frac{1}{N}\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n = \mu^{(1)},$$

$$\frac{1}{N}\sum_{n=1}^{N}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n^2 = \mu^{(2)} \quad , \quad \frac{1}{N-\tau}\sum_{n=1}^{N-\tau}\int \mathrm{d}\mathbf{z}\ P(\mathbf{z})z_n z_{n+\tau} = C(\tau),$$

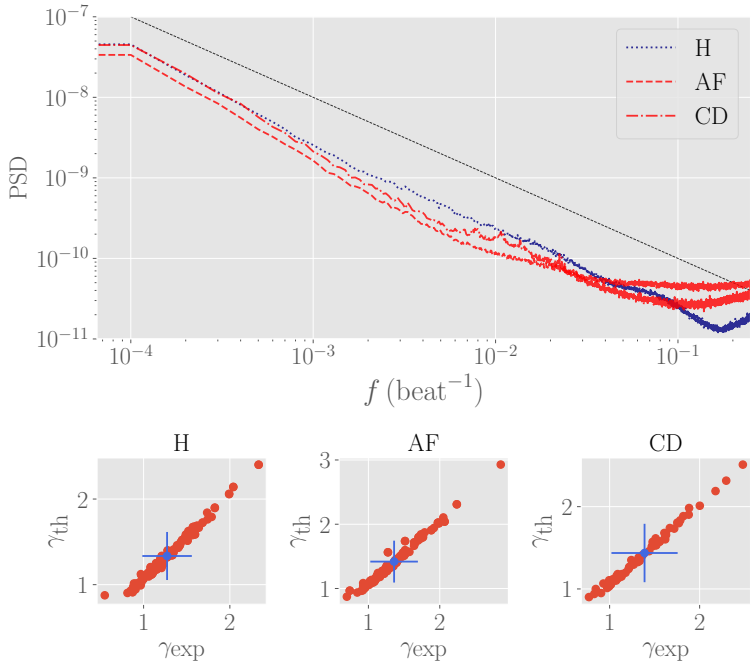

Figure 3.20: Top: empirical power spectral density (PSD). The dotted blue line refers to a healthy patients, while red are patients with AF (dashed curve) and CD (dash-dotted line). The PSD is computed according to the Welch procedure with 50% windows overlap. The black continuous curve is the expected $1/f$-noise distribution for visual comparison. Bottom: scatter plot for the scaling exponent of the PSD (in the region 10^{-4} e 10^{-2} Hz); in particular, we take the simple average over the synthetic realizations, the red spots are the exponent for the single patient (notice that the uncertainties over the synthetic realization are much smaller and are not visible in the plot), and the blue spot marks the average over all patients (both experimental and synthetic) with the relative uncertainties.

i.e. that the function $P(\mathbf{z})$ is a probability distribution and that the moments up to the second order are captures the experimental temporal average $\mu^{(1)}$, the temporal standard deviation $\mu^{(2)}$ and the auto-correlation function $C(\tau)$. The extremization with respect to the function P leads to the explicit form of the solution, i.e.

$$\log P(\mathbf{z}) = \lambda_0 - 1 + \lambda_1 \sum_{n=1}^{N} z_n + \lambda_2 \sum_{n=1}^{N} z_n^2 + \sum_{\tau=1}^{N} \lambda_\tau \sum_{n=1}^{N-\tau} z_n z_{n+\tau}, \tag{3.52}$$

which can be rewritten as

$$P(\mathbf{z}) = \text{cost} \cdot \exp\left(\lambda_1 \sum_{n=1}^{N} z_n + \lambda_2 \sum_{n=1}^{N} z_n^2 + \sum_{\tau=1}^{N} \sum_{n=1}^{N-\tau} \lambda_\tau z_n z_{n+\tau} \right). \tag{3.53}$$

The constant in the latter equation is computed by using the normalization property of the probability distribution $P(\mathbf{z})$, and it is given by

$$\text{cost}^{-1} = Z = \int \mathrm{d}\mathbf{z} \exp\left(\lambda_1 \sum_{n=1}^{N} z_n + \lambda_2 \sum_{n=1}^{N} z_n^2 + \sum_{\tau=1}^{N} \sum_{n=1}^{N-\tau} \lambda_\tau z_n z_{n+\tau} \right), \tag{3.54}$$

where we used the letter Z to make contact with the notion of partition function from the statistical mechanical dictionary. Since the model is essentially Gaussian, we can directly compute the partition function (at least, in formal way) as

$$\begin{aligned} Z &= \int \mathrm{d}\mathbf{z} \exp\left(\lambda_1 \mathbf{E}^T \cdot \mathbf{z} + \mathbf{z}^T(\lambda_2 \mathbb{I} + \boldsymbol{\lambda})\mathbf{z}\right) = \\ &= (-\pi)^{N/2} \det{}^{-1/2}(\lambda_2 \mathbb{I} + \boldsymbol{\lambda}) \exp(-\lambda_1^2 \mathbf{E}^T(\lambda_2 \mathbb{I} + \boldsymbol{\lambda})^{-1}\mathbf{E}). \end{aligned} \tag{3.55}$$

where $\mathbf{E} = (1, 1, \ldots, 1)$ is a N-dimensional vector of ones and we defined the interaction matrix $(\boldsymbol{\lambda})_{n,m} = \sum_{\tau=1}^{N} \delta_{n,m-\tau}\lambda_\tau$ (which turns out to be an upper triangular Toeplitz matrix with zeros on the main diagonal). Because of this, the determinant of the kernel $\lambda_2\mathbb{I} + \boldsymbol{\lambda}$ is trivially $\det(\lambda_2\mathbb{I} + \boldsymbol{\lambda}) = \lambda_2^N$. We can now determine the relation between the temporal average and standard deviation in terms of the model parameters. This relations read as

$$\mu^{(1)} = \langle \frac{1}{N}\sum_{n=1}^{N} z_n \rangle = \frac{1}{N}\frac{\partial \log Z}{\partial \lambda_1} = -\frac{2}{N}\lambda_1 \mathbf{E}^T(\lambda_2\mathbb{I} + \boldsymbol{\lambda})^{-1}\mathbf{E}, \tag{3.56}$$

$$\mu^{(2)} = \langle \frac{1}{N}\sum_{n=1}^{N} z_n^2 \rangle = \frac{1}{N}\frac{\partial \log Z}{\partial \lambda_2} = -\frac{1}{2\lambda_2} - \frac{\lambda_1^2}{N}\frac{\partial}{\partial \lambda_2}\mathbf{E}^T(\lambda_2\mathbb{I} + \boldsymbol{\lambda})^{-1}\mathbf{E}. \tag{3.57}$$

Since $\mathbf{z}$ is temporally standardized, we directly get $\lambda_1 = 0$ and $\lambda_2 = -1/2$. However, we left the former as a free parameter to be inferred and check *a posteriori* that it is consistent with zero. In order to get contact with Physics' dictionary, we rename the Lagrangian multipliers $\lambda_1 = h$ and $\lambda_\tau = J(\tau)$, playing the role of external magnetic field and two-body interactions respectively. Then, the solution of the maximum entropy problem (after some rearrangements of the sum indices) is given by

$$P(\mathbf{z}) = \frac{1}{Z}\left(\prod_{n=1}^{N} P_0(z_n)\right) \exp\left(\sum_{n=1}^{N-1}\sum_{\tau=1}^{N-n} J(\tau) z_n z_{n+\tau} + h\sum_{n=1}^{N} z_n\right), \tag{3.58}$$

where $P_0(z)$ is the Gaussian distribution $\mathcal{N}(0,1)$ and Z is the partition function.
The prior distribution for the values of the r_n elements in the time-series is chosen to be a Gaussian distribution:

$$P_0(r_n) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{r_n^2}{2}\Big), \tag{3.59}$$

and by introducing by hand a temporal correlation between the elements of the series. In other words, we consider the model described by the partition function (in the physics jargon)[1]

$$Z = \int \Big(\prod_{n=1}^{\infty} dr_n P_0(r_n)\Big) \exp\Big(\sum_{n,n'>n} J(n,n') r_n r_{n'} + \sum_{n=1}^{\infty} h(n) r_n\Big). \tag{3.60}$$

Since we assume that the relevant features differentiating the clinical status of the patients are entirely encoded in the HRV series, this directly implies that the parameters J and h should be characteristic of each class. As a general working hypothesis, we assume that the clinical status of each patient does not change during the sampling time and, since the starting time is arbitrary, this implies that the model should be characterized by a

[1] We stress that, in the first sum, the constrained $n' > n$ stands for the fact that correlation do not travel backward in time, *i.e.*, the value of the variable r_n does not affect those at previous time-steps.

translational invariance in time, *i.e.* $J(n,n') \equiv J(n'-n) = J(\tau)$ with $\tau = n'-n$ and $h(n) = h$ for all n. Then, we can rewrite the entire partition function as[1]

$$Z = \int \big(\prod_{n=1}^{\infty} dr_n P_0(r_n)\big) \exp\Big(\sum_{n=1}^{\infty}\sum_{\tau=1}^{\infty} J(\tau) r_n r_{n+\tau} + h\sum_{n=1}^{\infty} r_n\Big). \tag{3.61}$$

Since we are interested in finite-length HRV sequences, we need the truncated version of the partition function, which reads as

$$Z^{(L)} = \int \big(\prod_{n=1}^{L} dr_n P_0(r_n)\big) \exp\Big(\sum_{n=1}^{L-1}\sum_{\tau=1}^{L-n} J(\tau) r_n r_{n+\tau} + h\sum_{n=1}^{L} r_n\Big), \tag{3.62}$$

and yields to the following probability to observe a given finite-lenght sequence $\{r_n\}_{n=1}^{L}$

$$P\big(\{r_n\}_{n=1}^{L}\big) = \frac{1}{Z^{(L)}}\big(\prod_{n=1}^{L} P_0(r_n)\big) \exp\Big(\sum_{n=1}^{L-1}\sum_{\tau=1}^{L-n} J(\tau) r_n r_{n+\tau} + h\sum_{n=1}^{L} r_n\Big). \tag{3.63}$$

3.3.5 The pseudo-likelihood setup

The determination of the model parameters $J(\tau)$ and h is based on a maximum likelihood approach. As stated in the model description, the usage of the full probability is computationally untractable (because of the high dimensionality of the integral in the partition function),thus we use a maximum pseudo-likelihood approach [150] (namely a tractable asymptotic correct estimator of the likelihood), in which the fundamental object to be maximized is the conditional probability that, given the observations $\{z_n\}_{n=1}^{L}$, the successive observation is equal to experimental data z_{L+1}:

$$P\big(z = z_{L+1}|\{z_n\}_{n=1}^{L}\big) = \frac{P\big(\{z_n\}_{n=1}^{L+1}\big)}{P\big(\{z_n\}_{n=1}^{L}\big)} = \tag{3.64}$$
$$\frac{Z^{(L)}}{Z^{(L+1)}} \frac{\big(\prod_{n=1}^{L+1} P_0(z_n)\big) \exp\Big(\sum_{n=1}^{L}\sum_{\tau=1}^{L+1-n} J(\tau) z_n z_{n+\tau} + h\sum_{n=1}^{L+1} z_n\Big)}{\big(\prod_{n=1}^{L} P_0(z_n)\big) \exp\Big(\sum_{n=1}^{L-1}\sum_{\tau=1}^{L-n} J(\tau) z_n z_{n+\tau} + h\sum_{n=1}^{L} z_n\Big)}.$$

The second factor is easy to handle with, while the partition function can be evaluated by using the fact that $\int dz P\big(z|\{z_n\}_{n=1}^{L}\big) = 1$, so that we finally have

$$P\big(z = z_{L+1}|\{z_n\}_{n=1}^{L}\big) = \frac{\exp\Big(\log P_0(z_{L+1}) + \sum_{\tau=1}^{L} J(\tau) z_{L+1-\tau} z_{L+1} + h\, z_{L+1}\Big)}{\int dz \exp\Big(\log P_0(z) + \sum_{\tau=1}^{L} J(\tau) z_{L+1-\tau} z + h\, z\Big)}. \tag{3.65}$$

Since the prior is Gaussian, we can directly integrate the denominator for carrying out a closed form for the conditional probability. Thus, we get

$$P\big(z = z_{L+1}|\{z_n\}_{n=1}^{L}\big) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(z_{L+1} - \sum_{\tau=1}^{L} J(\tau) z_{L+1-\tau} - h\big)^2\Big). \tag{3.66}$$

[1] We stress that, since here the parameter h plays the role of bias (or, in physical jargon, external uniform magnetic field), then its effect should be the appearance of a non-zero average value for the temporal series. However, by definition the HRV series has zero mean, so it can be consistently put to zero (however, the results we obtained were also checked in presence of a non-zero field).

The MLE is based on the maximization of this conditional probability, or equivalently of the pseudo log-likelihood, which is composed by quantities of the form

$$\log P(z = z_{L+1}|\{z_n\}_{n=1}^{L}) = -\frac{1}{2}\big(z_{L+1} - \sum_{\tau=1}^{L} J(\tau) z_{L+1-\tau} - h\big)^2, \tag{3.67}$$

where we discarded unessential constant terms. Since we would like to infer the first values of the delayed interaction vector $J(\tau)$ and since the RR time-series have a size which is of the order of 10^5, it is better to use a sliding-window average approach, whose functioning is ensured by the stationary hypothesis. In this way, we can also perform a temporal average over all a single RR time-series. Supposing we want to infer the first T elements of the delayed interaction $J(\tau)$ (i.e. we truncate long-term correlations) and given the time-series $\{z_n\}_{n=1}^{N}$ of length N, we define the individual log-likelihood as

$$\begin{aligned}\mathcal{L}(\mathbf{J}, h|\{z_n\}_{n=1}^{N}) =& -\frac{1}{N-T}\sum_{L=T}^{N-1} \log P(z_{L+1}|\{z_n\}_{n=L-T+1}^{L}) = \\ =& -\frac{1}{2(N-T)}\sum_{L=T}^{N-1}\big(z_{L+1} - h - \sum_{\tau=1}^{T} J(\tau) z_{L+1-\tau}\big)^2.\end{aligned} \tag{3.68}$$

In order to prevent the parameters to acquire large values, we also introduce some regularization. For the bias, we simply add a quadratic penalization term: $\mathcal{R}(h) = -\lambda h^2/2$. Concerning the interaction vector, in order to discourage the algorithm to generate spurious correlation for high τ, we introduce a penalization which depends on the delay time τ, i.e.

$$\mathcal{R}(\mathbf{J}) = -\frac{\lambda}{2}\sum_{\tau=1}^{T} f(\tau) J(\tau)^2. \tag{3.69}$$

However, in order to ensure not to destroy correlation for interesting values of τ, we adopt a mild regularizer. In all of our tests, we found that a good choice is $f(\tau) = \log^2(1+\tau)$. Putting all pieces together, we have the regularized individual pseudo log-likelihood

$$\begin{aligned}\mathcal{L}^{(\mathrm{reg})}(\mathbf{J}, h|\{z_n\}_{n=1}^{N}) =& -\frac{1}{2(N-T)}\sum_{L=T}^{N-1}\big(z_{L+1} - h - \sum_{\tau=1}^{T} J(\tau) z_{L+1-\tau}\big)^2 \\ & -\frac{\lambda}{2}h^2 - \frac{\lambda}{2}\sum_{\tau=1}^{T} f(\tau) J(\tau)^2.\end{aligned} \tag{3.70}$$

The whole pseudo-likelihood is the average over the set $\mathcal{D}_c$ time-series in each given class (where $c \in \{H, AF, CD\}$):

$$\mathcal{L}^{(\mathrm{reg})}(\mathbf{J}, h|\mathcal{D}_c) = \frac{1}{M_c}\sum_{\mathbf{z}\in\mathcal{D}_c} \mathcal{L}^{(\mathrm{reg})}(\{z_n\}_{n=1}^{N}|\mathbf{J}, h), \tag{3.71}$$

where M_c is the number of examples in the class. By adopting a standard gradient descent (GD) approach, we can derive the following optimization rules:

$$\delta J(\tau) = \sum_{L=T}^{N-1} \Delta_L z_{L+1-\tau} - \lambda f(\tau) J(\tau), \tag{3.72}$$

$$\delta h = \sum_{L=T}^{N-1} \Delta_L - \lambda h, \tag{3.73}$$

where

$$\Delta_L = \frac{1}{N-T}(z_{L+1} - h - \sum_{\tau=1}^{T} J(\tau) z_{L+1-\tau}). \tag{3.74}$$

In order to speed up the inference procedure, we use a AdaGrad[151] adaptation method for the gradient descent rules (3.72). Since we want a uncertainties estimation for the coupling matrix $J(\tau)$, we proceed in the following way: better than to realize a single delayed interaction for the whole database (for each class), we minimize the pseudo log-likelihood to M random subsets of cardinality n of the database of each class (i.e. the gradients are averaged with respect to this minibatch) and then let the inferential algorithm converge towards a fixed point. Then, we compute the mean values and the standard deviation with respect to this M realizations of the interaction vector $J(\tau)$.

3.3.6 Discussion on the second experiment

Several past studies have highlighted that heart-rate fluctuations, in healthy individuals, exhibit the characteristic $1/f$ *noise* (see [133] and references therein). Deviations from this behavior can in fact be associated to cardiac pathologies such as atrial fibrillation or congestive heart failure [132]. In this work we tried to deepen the mechanisms possibly underlying this peculiar behavior, both in healthy as well as in compromised subjects.
To this aim we exploited inferential tools derived from statistical mechanics (i.e. the maximum entropy principle by the Jaynes perspective, one of the two pillars over which this thesis stands) to construct a probability distribution $P(\mathbf{r})$ characterizing the occurrence of a RR series $\mathbf{r}$. By requiring that $P(\mathbf{r})$ is minimally structured (i.e., prescribing the maximum entropy) and that $P(\mathbf{r})$ correctly matches the empirical first and second moments, we end up with a probabilistic model analogous to a spin-glass where quenched couplings $J(\tau)$ among spins exhibit frustration and a power-law decay with the distance τ between spin pairs. This kind of system is known to display chaotic dynamics spread over several timescales and the $1/f$ noise. We thus speculate that the presence of competitive driving force are key features for the emergence of the rich phenomenology displayed by heart-rate and we are naturally tempted to identify the two opposite driving forces with the sympathetic and parasympathetic systems: the ultimate representation of this mechanism is thus a *glassy hearth* whose core mechanism in order for it to work properly is frustration, thus highlighting also the systemic importance of the second pillar this thesis stands on, Parisi theory of complex systems, and the synergism by which these two pillars interplay to give rise to a unique coherent and powerful quantitative theory to describe -both theoretically and experimentally complex systems.
Finally, a last clinician-oriented observation: our data-driven glassy model is robustly checked against extensive available datasets and the preliminary results we reported in this thesis in order to classify heart rate variabiilty in healthy vs pathogenic patients seem to candidate the exponent β controlling the coupling decay $J(\tau) \sim \tau^{-\beta}$ as an indicator for classify the patient clinical status. Should further research confirm these findings, also this second computatonal approach would result in a cheap classifier for (some) heart failures, a new instrument -complementary to those already avaialble by clinicians- to fight diseases developing *personalized weapons*.

Conclusions

In this thesis I tried to summarize my research experience during my PhD time, by logically concatenating selected results I obtained along the way.
Before entering the details of any given problem I faced, the criterion that I followed in planning the exposition has been the systematic usage of *lightmotifs* along the whole manuscript. These are two major concepts (or better *methodological perspectives*) and a mathematical approach, all borred from Glassy Statistical Mechanics, namely Parisi representation of complex systems and Jaynes interpretation of entropy extremization as long as the concepts are regarded and Guerra's interpolation technique (coupled with Signal-to-Noise analysis and Monte Carlo simulations), as long as the methodologies are regarded.
The perspective that stems by merging Parisi Complexity Theory and Jaynes Entropy Extremization resulted in a powerful tool that we used to undertand both theoretical information processing networks (i.e. advances in Artificial Intelligence) as well as experimental information processing networks in biological complexity, ranging from cancer-related analysis to heart failure investigations.
Indeed, regarding the former (namely advances in Theoretical Artificial Intelligence), the picture that is emerging in the past three decades is that *statistical mechanics of disordered and complex systems* -namely Parisi theory framed within Jaynes entropic extremization (with all its package of definitions, concepts and tools - e.g. overlaps, replicas, replica trick, Guerra's interpolation, etc.) is becoming the main methodological exhaustive theoretical approach by which finally a systemic analysis of neural networks and learning machines can be achieved: trough this approach the spontaneous information processing capabilities of the networks emerge as a natural result of the countless interactions among neurons (much as the phases of matter emerge in classical statistical mechanics as a result of the tumultuos and noisy interactions of the molecules at the microscopic scale). Despite initial distrust and skepticism by Computer Scientists, this approach is nowadays particularly welcome in the Communities involved in Machine Learning and Neural Networks because, at present, there is an urgent need of an *Explainable Artificial Intelligence* (XAI) and of an *Optimized Artificial Intelligence* (OAI) and (as largely discussed in this thesis), in these regards, statistical mechanics of complex systems can play a pivotal role: focusing on XAI, it allows clear bridges between biological information processing and artificial information processing (thus helping in *cracking the black box*), further -focusing on OAI- the ultimate pourpose of the statistical mechanical approach is to produce *phase diagrams* and it is a credo of mine (and of several research groups worldwide) that their extensive production and rationalization could truly contribute toward an optimized usage of machine learning algorithms[1] because it allows setting the network in a desired optimal operational mode *a-priori* (selecting the most suitable region in the phase diagram, e.g. the retrieval region

[1]Note that, far from being a marginal aspect of Theroetical Artificial Intelligence, this is a climate-related mandatory imperative in order for these algorithms to be widespread in modern Societies without a dramatic impact on energy consumption and global warming.

if pattern recognition is concerned), thus potentially saving from expensive unsuccessfull trainings or from facing bad-posed tasks in general.
In this thesis, I contrbuted to the field mainly in three aspects as discussed in the first part of this manuscript (the second chapter, dedicated to Theoretical Artificial Intelligence): at first I have shown that there exist a sharp duality between key architectures in biologically inspired neural networks (e.g. the Hopfield model) and celebrated architectures in machine learning (e.g. the Boltzmann machine). By generalizing the Hebbian kernel of the Hopfield network in order to deal with examples of patterns -rather than patterns themselves- I have shown that the learning thresholds for this network do coincide with those of the Boltzmann machine and thus that there are deep similarities between artificial and biological information processing mechanisms. Riding the wave of this result, I tried to generalize it in two directions, namely by increasing the maximal storage capacity of these networks and by lowering their signal-to-noise trheshold for signal detection. I want to remark that in both the extensions, I have been inspired by biological information processing, implementing (suitably stylized versions of) sleeping-like mechanisms in the former and equipping the network with two input layers in the latter (imitating the redundat information sources provided by the two eyes or two ears in mammals).
These are just very small steps forward a Theory for Artificial Intelligence of course, yet, the most important point to me is that finally we are starting to see a route to be paved in order to have such a theory and, in this journey, Parisi theory and Jaynes perspective are expected to play a major role.

Clearly, while there is a giant attention on Artificial Intelligence in these years, hence the first application of the techniques that I have learnt has been focused in that field, the modus operandi that I have studied has a much broader basin of application and, in the second chapter of research in this thesis (chapter three) I applied the same know-how for analyzing experiments in biological complexity: I presented two streamlined examples of what can be learny by analyzing biological datasets via this *glassy maximum entropy approach*.
in the former I used this approach to infer the evolution of the interactions between (pancreatic) cancerous cells and (stroma) surrounding cellls in vitro: by repeating the experients with and without the presence of a chemotherapeutic drug, the ultimate aim was to study how their kinetics behavior and their interactios are affected by the presence of chemotherapy. Note that I framed these cells as networks of interacting element, with possibly complex interaction (hence these were Parisi's complex networks) and I inferred the related interaction by maximum entropy extremization in a genuine Jaynes perspective thus extensing quite naturally the modus operandi of the first part of the thesis to the second part. Via this procedure I obtained a quite net result: there are malignant lineages whose relative dialogue with the stroma is absent before the administration of the drug and persist to be absent also after (and this lack of interactions highly correlates with cancer progression) and there are malignant lineages whose relative dialogues with the stroma are absent before adding the drug in the medium but these are strongly enanched by the presence of the drug (and the raise of these interactions is highly correlated with cancer regression): while those are obviously just preliminary results, the interest for this kind of research is manifest and the methodology is completely general, extremely cheap and can be applied to countless other similar examples.
In the latter I used this methodology to infer long-range correlations among heart beats in historical series (i.e. Holter records) of hart rate variability: this is an *hard problem* from a statistical perspective because the underlying distributions of key-quantities are typi-

cally scale-free. Here we had labelled patients (healthy, suffering from atrial fibrillation and suffering from cardiac decompensation) and we studied the statistics hidden in their historical series again via Jaynes maximum entropy principle to find out a *glassy hearth*. Namely, we effectively described the correlations among heart beats as both long-range and positive as well as negative, much as the behavior of a one-dimension spin glass with long range frustrated couplings. This is at first corroborated by the typical $1/f$ noise (that we evidenced both in the temporal and in the frequency domain), further, we generated syntethic datasets from this data-driven glassy-hearth model and the statistics of synthetic data perfectly reproduces the original one conferring further robustness to our findings. Finally, we searched for refined information hidden within these historical series, with the aim of classification of phatologies as we were provied with labelled patients. Indeed statistically robust different outcomes have been obtained as refinements to the above $1/f$ leading scalings, potentially allowing for new classification criteria for automatized early detection of cardiac pathologies, en route toward a Personalized Medicine.
Cleary there is a long way to go before AI and these related tools will extensively take care of us in the Hospitals, yet the fact that a Ph.D. student in three years of research had the opportunity to inspect these fields and produce some new results is pheraphs an indicator that times are finally ripe for such revolution, where biological and artificial information processing networks will be mixed at various leves and the best has yet to come (but it is approaching fast): while I have certainly learned a lot in these years, from Science toutcourt to producing Science and living in a scientific environment, i.e. a research group - I also dare to believe that, in these three years, I tried my best to help our Society, for the little that was in my possibilities to be done, in this direction.

Appendices

3.4 Statistical mechanics approach to ultra-memory

To better inspect the crossover between archetype and example stabilities evidenced by signal-to-noise analysis and, possibly, to frame such a phenomenon into a classical phase transition setting where fast noise is also accounted, we must rely on statistical mechanics of spin glasses. In particular, in this appendix, we use a reformulation [26, 152] of the celebrated Guerra's interpolation technique [81].

3.4.1 General setting and main definitions

Let us consider a network made of N Ising neurons $\sigma_i = \pm 1$, with $i \in (1, ..., N)$, $K = \alpha N$ archetype patterns $\boldsymbol{\xi}_i^\mu \in \{-1, +1\}$ with $\mu \in (1, ..., K)$, and M noisy examples per archetype $\boldsymbol{\eta}^{\mu,a}$ with $\mu \in (1, ...K)$ and $a \in (1, ..., M)$. The latter constitute a stochastic, perturbed version of the archetypes, that are still binary and the arbitrary i-th component can be written as $\eta_i^{\mu,a} = \xi_i^\mu \chi_i^{\mu,a}$ for $i = 1, ..., N$, where $\chi_i^{\mu,a}$ is a Bernoullian random variable taking value -1 or $+1$. We will assume that, for each component, $\mathcal{P}(\xi = \pm 1) = 1/2$ and $\mathcal{P}(\chi = 1) = 1 - \mathcal{P}(\chi = -1) = p$ namely, the closer p is to $1/2$ and the higher the noise in the example (viceversa for $p \to 0$ and $p \to 1$, as the network stores equally a pattern and its flipped version, due to the spin-flip symmetry $\sigma_i \to -\sigma_i$). The network is fed by the $M \times K$ noisy patterns and has no direct access to the K archetype patterns.

Definition 10. *The Hamiltonian of the model is defined as*

$$\mathcal{H}_{N,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi}) = -\frac{1}{2N} \sum_{a=1}^{M} \sum_{\mu=1}^{K} \big(\sum_{i=1}^{N} \xi_i^\mu \chi_i^{\mu,a} \sigma_i \big)^2. \tag{3.75}$$

The partition function coupled to the Hamiltonian (3.75) *is defined as*

$$Z_{N,M}(\alpha, \beta|\boldsymbol{\chi}, \boldsymbol{\xi}) = \sum_{\sigma} \exp(-\beta \mathcal{H}_{N,M}(\boldsymbol{\sigma}|\boldsymbol{\chi}, \boldsymbol{\xi})) = \sum_{\sigma} \exp\Big[\frac{\beta}{2N} \sum_{a=1}^{M} \sum_{\mu=1}^{K} \big(\sum_{i=1}^{N} \xi_i^\mu \chi_i^{\mu,a} \sigma_i \big)^2 \Big]. \tag{3.76}$$

At finite network volume N and sample size M, the quenched pressure (i.e., the free energy times $-\beta$ [81]) of this model reads as

$$A_{N,M}(\alpha, \beta) = \frac{1}{N} \mathbb{E} \ln Z_{N,M}(\alpha, \beta|\boldsymbol{\chi}, \boldsymbol{\xi}), \tag{3.77}$$

where $\mathbb{E} := \mathbb{E}_\chi \mathbb{E}_\xi$, *being*

$$\begin{aligned}
\mathbb{E}_{\chi_i^{\mu,a}} f(\chi) &= \begin{cases} \int_{-\infty}^{+\infty} d\chi_i^{\mu,a} \left(p\delta(\chi_i^{\mu,a}-1) + (1-p)\delta(\chi_i^{\mu,a}+1)\right) f(\chi) & \mu = 1 \\ \int_{-\infty}^{+\infty} \frac{d\chi_i^{\mu,a}}{\sqrt{2\pi}} \exp(-\frac{(\chi_i^{\mu,a})^2}{2})\, f(\chi) & \mu = 2, \cdots, K \end{cases} \\
\mathbb{E}_\xi\, G(\xi) &= \int_{\mathbb{R}} \Big(\prod_{i=1}^{N} \prod_{\mu=1}^{K} \frac{d\xi_i^\mu}{2} \left[\delta(\xi_i^\mu + 1) + \delta(\xi_i^\mu - 1)\right] \Big) G(\xi) \qquad (3.78) \\
\mathbb{E}_\chi G(\chi) &= \left(\prod_{\mu=1}^{K} \prod_{a=1}^{M} \prod_{i=1}^{N} \mathbb{E}_{\chi_i^{\mu,a}} \right) G(\chi)
\end{aligned}$$

Finally, for a generic observable $O(\boldsymbol{\sigma}|\boldsymbol{\xi}, \boldsymbol{\chi})$, *we define the brackets as* $\langle O \rangle := \mathbb{E}\Omega\left(O(\boldsymbol{\sigma}|\boldsymbol{\xi}, \boldsymbol{\chi})\right)$, *being* Ω *the (replicated) Boltzmann average.*

remark 12. *In equation* (3.78) *we approximated the noise terms* $\chi_i^{\mu,a}$ *for* $\mu = 2, \cdots, K$ *as standard Gaussian variables; in the thermodynamic limit this assumption fits the worst case* ($p = 1/2$) *and, in general, it plays as a bound: if the network is able to infer an archetype out of this noisiest example sample, it will certainly works also in less challenging* ($p > 1/2$) *cases.*

Definition 11. *In order to quantify both the retrieval of the archetype and the retrieval of the examples, we define the related Mattis magnetizations as, respectively,*

$$m_\mu = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i, \qquad (3.79)$$

$$n_{\mu,a} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \chi_i^{\mu,a} \sigma_i. \qquad (3.80)$$

Proposition 5. *The partition function defined in* (3.76) *can be recast as*

$$\begin{aligned}
Z_{N,M}(\beta, \alpha|\boldsymbol{\chi}, \boldsymbol{\xi}) &= \lim_{J\to 0} Z_{N,M}(\beta, \alpha, J|\boldsymbol{\chi}, \boldsymbol{\xi}) = \\
&= \lim_{J\to 0} \sum_{\sigma} \int \prod_{\mu=2,a=1}^{K,M} \frac{dz_{\mu,a}}{\sqrt{2\pi}} \exp\Big[-\frac{1}{2} \sum_{\mu=2,a=1}^{K,M} z_{\mu,a}^2 + \\
&\quad + J \sum_{i=1}^{N} \xi_i^1 \sigma_i + \sqrt{\frac{\beta}{N}} \sum_{a=1}^{M} \sum_{\mu=2}^{K} \sum_{i=1}^{N} \xi_i^\mu \chi_i^{\mu,a} \sigma_i z_{\mu,a} + \\
&\quad + \frac{\beta}{2N} \sum_{a=1}^{M} \big(\sum_{i=1}^{N} \xi_i^1 \chi_i^{1,a} \sigma_i \big)^2 \Big],
\end{aligned}$$

which corresponds to the partition function of a restricted Boltzmann machine with N *visible binary neurons* $\sigma_i \in \{-1, +1\}$, $M \times K$ *hidden Gaussian neurons* $z_{\mu,a} \sim \mathcal{N}(0,1)$, *and weights* $\chi_i^{\mu,a}\xi_i^\mu$, *for any* $i = 1, ..., N$, $\mu = 1, ..., K$, *and* $a = 1, ..., M$.

remark 13. *In the expression above we added the last term $J\sum_{i=1}^{N}\xi_i^1\sigma_i$ to generate the expectations of the Mattis magnetization m_1, by evaluating the derivative of the quenched pressure w.r.t. J at $J=0$. In fact, we need to quantify both the retrieval of the archetype and the retrieval of the examples, but, while the noisy examples* exist *and are supplied to the network (in fact, the Hamiltonian itself can be written in terms of the examples $\{\boldsymbol{\eta}^{\mu,a}\}$), the archetype is a network's abstraction, nor it exists by itself neither it is coded in the Hamiltonian, hence we need to use the functional generator trick. However, as we will see in Sec. 3.4.4, as far as $M \gg 1$, we can bypass this artifice and obtain the expectation value of m by exploiting is direct proportionality with the expectation value of n, which, instead, is a natural order parameter for the model.*

Proof. We chose as "marked" (or "condensate") patterns [8, 11] those related to the archetype labelled as $\mu=1$ and, accordingly, we re-write eq. (3.76) as

$$Z_{N,M}(\beta,\alpha|\boldsymbol{\chi},\boldsymbol{\xi}) = \sum_{\sigma}\exp\Big[\frac{\beta}{2N}\sum_{a=1}^{M}\big(\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i\big)^2+\frac{\beta}{2N}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\big(\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i\big)^2\Big]. \quad (3.81)$$

Since we are interested in extracting the magnetization for both the noisy examples and the archetypes, we introduce a source field J such that the partition function is generalized as

$$\begin{aligned} Z_{N,M}(\beta,\alpha,J|\boldsymbol{\chi},\boldsymbol{\xi}) = &\sum_{\sigma}\exp\Big[\frac{\beta}{2N}\sum_{a=1}^{M}\big(\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i\big)^2+ \\ &+\frac{\beta}{2N}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\big(\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i\big)^2+J\sum_{i=1}^{N}\xi_i^1\sigma_i\Big]. \end{aligned}$$

Then, we apply the relation

$$\exp\left(\frac{X^2}{2}\right) = \int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}+Xz) \quad (3.82)$$

to each squared term appearing in the argument of the exponential and this directly yields to Eq. (3.81). □

3.4.2 Guerra's interpolation for the quenched pressure

The strategy that we follow to solve the model is based on Guerra's interpolation technique [26, 152] and ultimately consists in exploiting the mean-field nature of the model to properly compare the original model with an effective one-body model that shares the same statistical features of the original one in the thermodynamics limit.

Definition 12. *The Guerra interpolating functional for the quenched pressure related to the cost-function 3.75 is defined as*

$$\begin{aligned} A_{N,M}(\alpha,\beta,J;t) = \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\ln\Big[\sum_{\sigma}\int\prod_{\mu=2,a=1}^{K,M}\frac{dz_{\mu,a}}{\sqrt{2\pi}}\exp\Big(-\frac{\psi(t)}{2}\sum_{\mu=2,a=1}^{K,M}z_{\mu,a}^2+ \\ +J\sum_{i=1}^{N}\xi_i^1\sigma_i+\Gamma(t)\sqrt{\frac{\beta}{N}}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i z_{\mu,a}+ \\ +\rho(t)\frac{\beta}{2N}\sum_{a=1}^{M}\big(\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i\big)^2+NW_{N,M}(t)\Big)\Big], \end{aligned} \quad (3.83)$$

where $\psi(t), \Gamma(t), \rho(t)$ are auxiliary fields to be set a posteriori, and $W_{N,M}(t) := W(\boldsymbol{\sigma}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\chi}; t)$ is a source term whose specific expression will be set a posteriori too.

In the following, to lighten the notation, we will set $A_J := A_{N,M}(\alpha, \beta, J; t)$.
Note that the original model can be recovered by setting

$$\psi(t=1) = \Gamma(t=1) = \rho(t=1) = 1, \tag{3.84}$$

$$W_{N,M}(t) = 0, \tag{3.85}$$

and, as standard, we approach $\psi(t=1) = \Gamma(t=1) = \rho(t=1) = 1$ by evaluating the factorized case $\psi(t=0), \Gamma(t=0), \rho(t=0) = 0$ and then integrating back in t from 0 to 1 by using the fundamental theorem of calculus.

To accomplish this plan, denoting by $\langle . \rangle_t$ the averages evaluated in this extended framework (and clearly $\langle \rangle_t \to \langle \rangle$ as $t \to 1$), let us start working out the streaming of A_J:

$$\begin{aligned}
\frac{\partial}{\partial \Gamma} A_J &= \frac{1}{N}\sqrt{\frac{\beta}{N}} \sum_{a=1}^{M}\sum_{\mu=2}^{K}\sum_{i=1}^{N} \mathbb{E}_{\phi,\chi,\xi}\, \omega_s(\xi_i^\mu \chi_i^{\mu,a} z_{\mu,a}\sigma_i)_t = \\
&= \frac{\beta}{N^2}\Gamma_t \sum_{a=1}^{M}\sum_{\mu=2}^{K}\sum_{i=1}^{N} \mathbb{E}\,\phi\, \mathbb{E}\,\chi\, \mathbb{E}\,\xi \big(\omega_s(z_{\mu,a}^2)_t - \omega_s(z_{\mu,a}\sigma_i)_t\big) \\
\frac{\partial}{\partial \rho} A_J &= \frac{\beta}{2}\sum_{a=1}^{M} \mathbb{E}\,\phi\, \mathbb{E}\,\chi\, \mathbb{E}\,\xi \omega_s(\big(\frac{1}{N}\sum_{i=1}^{N}\xi_i^1 \chi_i^{1,a}\sigma_i\big)^2)_t \\
\frac{\partial}{\partial \psi} A_J &= -\frac{1}{2N}\sum_{a=1}^{M}\sum_{\mu=2}^{K} \mathbb{E}\,\phi\, \mathbb{E}\,\chi\, \mathbb{E}\,\xi \omega_s(z_{\mu,a}^2)_t
\end{aligned} \tag{3.86}$$

such that

$$\frac{dA_J}{dt} = \dot{\Gamma}\frac{\partial}{\partial \Gamma}A_J + \dot{\rho}\frac{\partial}{\partial \rho}A_J + \dot{\psi}\frac{\partial}{\partial \psi}A_J + \omega_s(\dot{W})_t. \tag{3.87}$$

We still have the freedom of choice for the source term $W_{N,M}(t)$: the idea is the classical one in Guerra's interpolation, as we are explaining hereafter. By taking advantage of the mean-field nature of the model, it should be possible to linearize the "nasty" quadratic interactions appearing in (3.81) by properly balancing them with extra one-body terms (i.e., those introduced in 3.83) such that each contribution within the source term has to match the second order moments of the order parameters. In this way we can calculate and tune the effective one-body contributions – that are easy to evaluate – and, in the thermodynamic limit, under the replica symmetric assumption disregard fluctuations around those means. To this task we choose $W_{N,M}(t)$ as:

$$W_{N,M}(t) = \frac{\lambda(t)}{N}\sum_{i=1}^{N}\phi_i\sigma_i + \frac{\mu(t)}{N}\sum_{\mu=2}^{K}\sum_{a=1}^{M}\phi_{\mu,a}z_{\mu,a} + \frac{\tau(t)}{N}\sum_{a=1}^{M}\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i \tag{3.88}$$

With this choice for the source term a few more derivatives must be calculated,

$$\frac{\partial}{\partial\lambda}A_J = \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\sum_{i=1}^{N}\phi_i\omega_s(\sigma_i)_t = \frac{\lambda(t)}{N}\mathbb{E}_{\phi,\chi,\xi}\sum_{i=1}^{N}\big(1-\omega_s(\sigma_i)_t^2\big) \tag{3.89}$$

$$\begin{aligned}\frac{\partial}{\partial\mu}A_J &= \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\sum_{\mu=2}^{K}\sum_{a=1}^{M}\phi_{\mu,a}\omega_s(z_{\mu,a})_t = \\ &= \frac{\mu(t)}{N}\sum_{\mu=2}^{K}\sum_{a=1}^{M}\mathbb{E}_{\phi,\chi,\xi}\big(\omega_s(z^2_{\mu,a})_t-\omega_s(z_{\mu,a})_t^2\big)\end{aligned} \tag{3.90}$$

$$\frac{\partial}{\partial\tau}A_J = \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\sum_{i=1}^{N}\sum_{a=1}^{M}\omega_s(\xi_i^1\chi_i^{1,a}\sigma_i)_t \tag{3.91}$$

Proposition 6. *By inspecting the moments generated by differentiating A_J we naturally introduce a complete set of order parameters to characterize the system, namely, the two replica overlaps p_{lm} for the z variables, the two replica overlaps q_{lm} for the σ variables (accounting for the slow noise in the system) and two sets of quantifiers of the retrieval, namely the standard Mattis magnetization of the archetype m_μ and a generalized Mattis magnetization for the noise example $n_{\mu,a}$:*

$$p_{lm} = \frac{1}{KM}\sum_{\mu=1}^{K}\sum_{a=1}^{M}z_{\mu,a}^{(l)}z_{\mu,a}^{(m)} \tag{3.92}$$

$$q_{lm} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(l)}\sigma_i^{(m)} \tag{3.93}$$

$$n_{\mu,a} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\chi_i^{\mu,a}\sigma_i \tag{3.94}$$

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i. \tag{3.95}$$

By these definitions each differential can be rewritten as

$$\begin{aligned}\frac{\partial}{\partial\psi}A_J &= -\frac{KM}{2N}\mathbb{E}_{\phi,\chi,\xi}\omega_s(p_{11})_t;\\ \frac{\partial}{\partial\Gamma}A_J &= \beta\Gamma(t)\frac{KM}{N}\mathbb{E}_{\phi,\chi,\xi}\big(\omega_s(p_{11})_t-\omega_s(p_{12}q_{12})_t\big)\\ \frac{\partial}{\partial\rho}A_J &= \frac{\beta}{2}\sum_{a=1}^{M}\mathbb{E}_{\phi,\chi,\xi}\omega_s(n_{1,a}^2)_t;\\ \frac{\partial}{\partial\lambda}A_J &= \lambda(t)\mathbb{E}_{\phi,\chi,\xi}\big(\omega_s(q_{11})_t-\omega_s(q_{12})_t\big)\\ \frac{\partial}{\partial\mu}A_J &= \mu(t)\frac{KM}{N}\big(\omega_s(p_{11})_t-\omega_s(p_{12})_t\big);\\ \frac{\partial}{\partial\tau}A_J &= \sum_{a=1}^{M}\mathbb{E}_{\phi,\chi,\xi}\omega_s(n_{1,a})_t.\end{aligned}$$

We are now ready to explicitly write $\frac{dA_J}{dt}$:

$$\frac{dA_J}{dt} = -\frac{\alpha}{2}M\omega_s(p_{11})\dot{\psi} + \beta\alpha M\Gamma\dot{\Gamma}(\omega_s(p_{11}) - \omega_s(p_{12}q_{12})) + \dot{\rho}\frac{\beta}{2}\sum_{a=1}^{M}\omega_s(n_{1,a}^2) +$$
$$+\lambda\dot{\lambda}(\omega_s(q_{11}) - \omega_s(q_{12}))\mu\dot{\mu}\alpha M(\omega_s(p_{11}) - \omega_s(p_{12})) + \dot{\tau}\sum_{a=1}^{M}\omega_s(n_{1,a}).$$

As in the replica symmetric regime we can discard fluctuations of the order parameters, assuming the latter to self-average around their mean values, that we indicate by a bar in the following, i.e. $\lim_{N\to\infty}\mathcal{P}(q_{12}) = \delta(q_{12} - \bar{q}), \lim_{N\to\infty}\mathcal{P}(p_{12}) = \delta(p_{12} - \bar{p})$, the strategy now is to write correlations as a source term, made of by mean values (that we will keep in the asymptotic limit), and fluctuations around these means (that will be discarded in the asymptotic limit), thus we write

$$\begin{aligned}\omega_s(p_{12}q_{12}) &= \omega_s((p_{12} - \bar{p})(q_{12} - \bar{q})) - \bar{p}\bar{q} + \bar{p}\omega_s(q_{12}) + \bar{q}\omega_s(p_{12}) \\ \omega_s(n_{1,a}^2) &= \omega_s((n_{1,a} - \bar{n})^2) - \bar{n}^2 + 2\bar{n}\omega_s(n_{1,a}).\end{aligned} \tag{3.96}$$

We plug the previous expressions in the streaming equation for A_J

$$\begin{aligned}\frac{dA_J}{dt} &= -\frac{\alpha}{2}M\omega_s(p_{11})\dot{\psi} + \beta\alpha M\Gamma\dot{\Gamma}(\omega_s(p_{11}) - \omega_s((p_{12} - \bar{p})(q_{12} - \bar{q})) + \bar{p}\bar{q} - \bar{p}\omega_s(q_{12}) - \bar{q}\omega_s(p_{12})) + \\ &+ \dot{\rho}\frac{\beta}{2}\sum_{a=1}^{M}(\omega_s((n_{1,a} - \bar{n})^2) - \bar{n}^2 + 2\bar{n}\omega_s(n_{1,a})) + \lambda\dot{\lambda}(1 - \omega_s(q_{12})) + \\ &+ \mu\dot{\mu}\alpha M(\omega_s(p_{11}) - \omega_s(p_{12})) + \dot{\tau}\sum_{a=1}^{M}\omega_s(n_{1,a})\end{aligned} \tag{3.97}$$

and we set to zero each coefficient coupled to a first order moment of any of the order parameters, namely

$$\omega_s(p_{11}) : -\frac{1}{2}\dot{\psi} + \beta\Gamma\dot{\Gamma} + \mu\dot{\mu} = 0, \tag{3.98}$$
$$\omega_s(p_{12}) : \beta\Gamma\dot{\Gamma}\bar{q} + \mu\dot{\mu} = 0, \tag{3.99}$$
$$\omega_s(q_{12}) : \bar{p}\beta\alpha M\Gamma\dot{\Gamma} + \lambda\dot{\lambda} = 0, \tag{3.100}$$
$$\omega_s(n_{1,a}) : \dot{\tau} + \bar{n}\dot{\rho}\beta = 0. \tag{3.101}$$

This PDE system is under-determined: it is sufficient to find a solution which solves it and that also satisfies the Cauchy condition for A_J (3.84) and

$$\Gamma_{t=0} = 0, \tag{3.102}$$
$$\rho_{t=0} = 0. \tag{3.103}$$

The last two constraints allow us to further simplify the solution of the model, and make it exactly solvable at the replica symmetric level. It is easy to solve this PDE system: one can verify that the solution we are looking for is given by

$$\Gamma(t) = \sqrt{t}, \tag{3.104}$$
$$\rho(t) = t, \tag{3.105}$$
$$\psi(t) = 1 - (1-t)\beta(1-\bar{q}), \tag{3.106}$$
$$\mu(t) = \sqrt{\beta\bar{q}(1-t)}, \tag{3.107}$$
$$\lambda(t) = \sqrt{\alpha\beta\bar{p}M(1-t)}, \tag{3.108}$$
$$\tau(t) = \bar{n}(1-t). \tag{3.109}$$

remark 14. *We point out a difference between our approach and the original Guerra's route: in the latter, the interpolation parameter associated to glassy terms appears under the square root, while when associated to the signal terms it appears linearly; in our approach, the interpolants are general functions of t and we obtain Guerra's prescriptions as the result of the resolution of the differential equation system coded in the eq.s 3.98.*

These terms have to be plugged in the streaming equation for A_J, whose final expression is given by

$$\frac{dA_J}{dt} = -\frac{1}{2}\beta\alpha M\bar{p}(1-\bar{q}) - \frac{\beta M}{2}\bar{n}^2 - \frac{1}{2}\beta\alpha M\omega_s((p_{12}-\bar{p})(q_{12}-\bar{q})) + \frac{\beta}{2}\sum_{a=1}^{M}\omega_s((n_{1,a}-\bar{n})^2). \tag{3.110}$$

As, under the replica symmetric ansatz, we can disregard the fluctuations asymptotically, we can state the next

theorem 3. *In the high storage ($K = \alpha N$) and in the infinite volume of the network limit ($N \to \infty$), but finite dataset size M, the quenched replica symmetric pressure of the model (3.75) is given by the following expression in terms of the natural order parameters of the theory:*

$$\begin{aligned} A_{N,M}(\alpha,\beta,J;t) &= \log 2 - \frac{\beta\alpha M}{2}\bar{p}(1-\bar{q}) - \frac{\beta M}{2}\bar{n}^2 - \frac{\alpha M}{2}\big(\log[1-\beta(1-\bar{q})] - \frac{\beta\bar{q}}{1-\beta(1-\bar{q})}\big) + \\ &+ \mathbb{E}_{\phi\chi}\log\cosh\big(J + \bar{n}\beta\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big) \end{aligned} \tag{3.111}$$

Proof. Note that, with the expression (3.110) for the streaming of A_J we express the flux of A_J in t by two kinds of object: average values of the order parameters, i.e. $\bar{q}$, $\bar{p}$, $\bar{n}$, that contribute to the source term, and all the remaining terms that are fluctuations around these means, i.e. $\langle(p_{12}-\bar{p})(q_{12}-\bar{q})\rangle$ and $\omega_s((n_{1,a}-\bar{n})^2)$: the latter can be discarded in the thermodynamic limit, under replica-symmetric assumption. Note further that, so far, the Mattis magnetization for the archetype has played no role.
For the sake of completeness we write also the interpolating structure in its final form that reads

$$\begin{aligned} A_J &= \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\log\Big[\sum_{\sigma}\int\prod_{\mu=2}^{K}\prod_{a=1}^{M}\frac{dz_{\mu,a}}{\sqrt{2\pi}}\exp\Big(-\frac{1-\beta(1-\bar{q})(1-t)}{2}\sum_{\mu=2}^{K}\sum_{a=1}^{M}z_{\mu,a}^2 + \\ &+ \sqrt{t}\sqrt{\frac{\beta}{N}}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\xi_i^{\mu}\chi_i^{\mu,a}z_{\mu,a}\sigma_i + t\frac{\beta N}{2}\sum_{a=1}^{M}(\frac{1}{N}\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i)^2 + J\sum_{i=1}^{N}\xi_i^1\sigma_i + \\ &+ \sqrt{\alpha\beta\bar{p}M(1-t)}\sum_{i=1}^{N}\phi_i\sigma_i + \sqrt{\beta\bar{q}(1-t)}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\phi_{\mu,a}z_{\mu,a} + \bar{n}\beta(1-t)\sum_{a=1}^{M}\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i\Big)\Big]. \end{aligned} \tag{3.112}$$

The true power of the interpolation scheme now shines: the solution of the model can be recast as a simple integration problem. Recalling that we are interested in the original model (which can be recovered by setting $t = 1, J = 0$ inside the interpolating structure 3.112), we can exploit the fundamental theorem of calculus now, as

$$A_J(t=1) = A_J(t=0) + \int_0^1 ds\ \frac{dA_J}{dt}\Big|_{t=s}, \tag{3.113}$$

thus all that is left to do is evaluating the trivial 1-body problem $A_J(t=0)$: this is a routinely integration procedure and it is performed as follows

$$\begin{aligned} A_J(t=0) &= \frac{1}{N}\mathbb{E}_{\phi,\chi,\xi}\log\Big[\sum_{\sigma}\int\prod_{\mu=2}^{K}\prod_{a=1}^{M}\frac{dz_{\mu,a}}{\sqrt{2\pi}}\exp\Big(-\frac{1-\beta(1-\bar{q})}{2}\sum_{\mu=2}^{K}\sum_{a=1}^{M}z_{\mu,a}^2 + \\ &+ J\sum_{i=1}^{N}\xi_i^1\sigma_i + \sqrt{\alpha\beta\bar{p}M}\sum_{i=1}^{N}\phi_i\sigma_i + \sqrt{\beta\bar{q}}\sum_{a=1}^{M}\sum_{\mu=2}^{K}\phi_{\mu,a}z_{\mu,a} + \bar{n}\beta\sum_{a=1}^{M}\sum_{i=1}^{N}\xi_i^1\chi_i^{1,a}\sigma_i\Big)\Big] = \\ &= -\frac{\alpha M}{2}\left(\log[1-\beta(1-\bar{q})] - \frac{\beta\bar{q}}{1-\beta(1-\bar{q})}\right) + \mathbb{E}_{\phi\chi}\log\cosh\big(J+\bar{n}\beta\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big), \end{aligned} \tag{3.114}$$

thus ending the proof. □

corollary 1. *The self-consistency equations related to the model introduced in Definition (10) are obtained by looking for the stationary points of the quenched pressure* $\nabla_{\bar{n},\bar{q},\bar{p}}A_J|_{J=0} = 0$. *These equations are given by*

$$\bar{p} = \frac{\beta\bar{q}}{[1-\beta(1-\bar{q})]^2}, \tag{3.115}$$

$$\bar{q} = \mathbb{E}_{\phi\chi}\tanh^2\big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big), \tag{3.116}$$

$$\bar{n} = \mathbb{E}_{\phi\chi}\big(\frac{1}{M}\sum_{a=1}^{M}\chi_a\big)\tanh\big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big). \tag{3.117}$$

Further, exploiting the auxiliary field J, inserted by hand in such a way that $\bar{m} = \nabla_J A_J$, we obtain

$$\bar{m} = \mathbb{E}_{\phi\chi}\tanh\big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \sqrt{\alpha\beta\bar{p}M}\phi\big). \tag{3.118}$$

Proof. The proof works by straightforward derivation of A_J in (3.111). □

3.4.3 Network behavior in the noiseless limit $\beta \to \infty$

As standard also for the classic Hopfield scenario, namely within the AGS theory [8, 11], en route to the ground-state solution (namely the self-consistencies for $\beta \to \infty$), we now assume that $\lim_{\beta\to\infty}\beta(1-\bar{q})$ is finite. This gives rise to the following

theorem 4. *The zero-temperature self-consistency equations for the order parameters read as*

$$\bar{K} := \frac{\sqrt{2\alpha M}\beta(1-\bar{q})}{\beta(1-\bar{q})-1} = \mathbb{E}_{\chi}\,\mathrm{erf}'\big(\frac{\bar{n}\sum_{a=1}^{M}\chi_a}{\bar{K}+\sqrt{2\alpha M}}\big) \tag{3.119}$$

$$\bar{n} = \mathbb{E}_{\chi}\frac{\sum_{a=1}^{M}\chi_a}{M}\,\mathrm{erf}\big(\frac{\bar{n}\sum_{a=1}^{M}\chi_a}{\bar{K}+\sqrt{2\alpha M}}\big) \tag{3.120}$$

$$\bar{m} = \mathbb{E}_{\chi}\,\mathrm{erf}\big(\frac{\bar{n}\sum_{a=1}^{M}\chi_a}{\bar{K}+\sqrt{2\alpha M}}\big) \tag{3.121}$$

where erf is the error function and erf′ is it's first derivative $erf'(x) := \frac{2}{\sqrt{\pi}}\exp(-x^2)$.

Proof. As a first step we introduce an additional term βx in the argument of the hyperbolic tangent appearing in the self-consistency equations (3.115):

$$\bar{q} = \mathbb{E}_{\chi,\phi} \tanh^2 \Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M\bar{q}}{[1-\beta(1-\bar{q})]^2}} + \beta x\Big) \tag{3.122}$$

$$\bar{n} = \mathbb{E}_{\chi,\phi}(\frac{1}{M}\sum_{a=1}^{M}\chi_a)\tanh\Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M\bar{q}}{[1-\beta(1-\bar{q})]^2}} + \beta x\Big) \tag{3.123}$$

$$\bar{m} = \mathbb{E}_{\chi,\phi}\tanh\Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M\bar{q}}{[1-\beta(1-\bar{q})]^2}} + \beta x\Big). \tag{3.124}$$

We also recognize that at $\beta\to\infty$ we also have $q\to 1$ thus in order to correctly perform the limit a reparametrization is in order,

$$\bar{q} = 1 - \frac{\delta q}{\beta} \quad \text{as} \quad \beta\to\infty \tag{3.125}$$

Via this reparametrization we obtain

$$1-\frac{\delta q}{\beta} = \mathbb{E}_{\chi,\phi}\tanh^2\Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M(1-\frac{\delta q}{\beta})}{(1-\delta q)^2}} + \beta x\Big) \tag{3.126}$$

$$\bar{n} = \mathbb{E}_{\chi,\phi}(\frac{1}{M}\sum_{a=1}^{M}\chi_a)\tanh\Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M(1-\frac{\delta q}{\beta})}{(1-\delta q)^2}} + \beta x\Big) \tag{3.127}$$

$$\bar{m} = \mathbb{E}_{\chi,\phi}\tanh\Big(\beta\bar{n}\sum_{a=1}^{M}\chi_a + \beta\phi\sqrt{\frac{\alpha M(1-\frac{\delta q}{\beta})}{(1-\delta q)^2}} + \beta x\Big). \tag{3.128}$$

Taking advantage of the new parameter x we can recast the last equation in δq as a derivative of the magnetization $\bar{m}$:

$$\frac{\partial\bar{m}}{\partial x} = \beta[1-(1-\frac{\delta q}{\beta})] = \delta q \tag{3.129}$$

where we used both the self-consistencies for $\bar{m}$ and δq. Thanks to this correspondence between $\bar{m}$ and δq, we can proceed with our limit without worrying about $\bar{q}$: the limiting equations for $\bar{m},\bar{n}$ are now for $\beta\to\infty$:

$$\bar{n} = \mathbb{E}_{\chi,\phi}(\frac{1}{M}\sum_{a=1}^{M}\chi_a)\,\mathrm{sign}\Big(\bar{n}\sum_{a=1}^{M}\chi_a + \phi\sqrt{\frac{\alpha M}{[1-\delta q]^2}} + x\Big), \tag{3.130}$$

$$\bar{m} = \mathbb{E}_{\chi,\phi}\,\mathrm{sign}\Big(\bar{n}\sum_{a=1}^{M}\chi_a + \phi\sqrt{\frac{\alpha M}{[1-\delta q]^2}} + x\Big). \tag{3.131}$$

These equations can be further simplified by evaluating the Gaussian integral in ϕ, via the relation:

$$\mathbb{E}_\phi\,\mathrm{sign}(A\phi+B) = \mathrm{erf}\left(\frac{B}{\sqrt{2}A}\right)$$

to get

$$\bar{m} = \mathbb{E}_\chi\,\mathrm{erf}\Big[(\bar{n}\sum_{a=1}^{M}\chi_a + x)\frac{1-\delta q}{\sqrt{2\alpha M}}\Big] \tag{3.132}$$

$$\bar{n} = \mathbb{E}_\chi(\frac{1}{M}\sum_{a=1}^{M}\chi_a)\,\mathrm{erf}\Big[(\bar{n}\sum_{a=1}^{M}\chi_a + x)\frac{1-\delta q}{\sqrt{2\alpha M}}\Big] \tag{3.133}$$

while δq, thanks to (3.129), becomes

$$\delta q = \frac{\partial \bar{m}}{\partial x} = \mathbb{E}_\chi \frac{2}{\sqrt{\pi}} \frac{1-\delta q}{\sqrt{2\alpha M}} \exp\left\{ - \left[(\bar{n} \sum_{a=1}^{M} \chi_a + x) \frac{1-\delta q}{\sqrt{2\alpha M}} \right]^2 \right\}. \tag{3.134}$$

In order to simplify the equation in δq we make one last change of variables,

$$\delta q = \frac{\delta Q}{\delta Q + \sqrt{2\alpha M}}$$

yielding to

$$\bar{m} = \mathbb{E}_\chi \operatorname{erf}\Big(\frac{\bar{n} \sum_{a=1}^{M} \chi_a}{\sqrt{2\alpha M} + \delta Q} \Big) \tag{3.135}$$

$$\bar{n} = \mathbb{E}_\chi \big(\frac{1}{M} \sum_{a=1}^{M} \chi_a \big) \operatorname{erf}\Big(\frac{\bar{n} \sum_{a=1}^{M} \chi_a}{\sqrt{2\alpha M} + \delta Q} \Big) \tag{3.136}$$

$$\delta Q = \mathbb{E}_\chi \frac{2}{\sqrt{\pi}} \exp\Big[- \Big(\frac{\bar{n} \sum_{a=1}^{M} \chi_a}{\sqrt{2\alpha M} + \delta Q} \Big)^2 \Big] \tag{3.137}$$

where x has been set to 0, allowing to close the proof. □

The solutions of these equations, as p and M are varied, is captured in the plots of Fig. 3.21. Remarkably, there exists a crossover at $\tilde{M}(p)$, such that as $M < \tilde{M}(p)$ ($M > \tilde{M}(p)$) the example magnetization $\bar{n}$ is larger (smaller) than the archetype magnetization $\bar{m}$. We would be tempted to label the crossover points $\tilde{M}(p)$ as candidate markers of a phase transition, yet we still need to further inspect the system and to develop the theory by suitably sending both M and N (and K as well in the high storage) to infinity before we can robustly refer to a phase transition; this work will be achieved in the next subsection.

3.4.4 Network behavior in the large dataset limit $M \to \infty$

In the theory developed so far, we assumed that, as K and N are made larger and larger, their ratio α remains finite in such a way that it can be used as an intensive parameter tuning pattern load, however, the parameter M expressing the sample size is still extensive and its tuning is not related to a tuning in the network volume N or in the number K of pattern. In this section we turn the whole theory intensive such that the meaning of the self-consistencies, as well as the nature of the phase transition, can appear manifestly.
This goal is approached by steps: first, setting M as large (but still retaining the parameter M explicit), via the central limit theorem, we approximate the quantity $\frac{1}{M}\sum_{a=1}^{M} \chi_a$, where, we recall $\mathcal{P}(\chi_a) = p\,\delta(\chi_a - 1) + (1-p)\,\delta(\chi_a + 1)$, with a Gaussian random variable, namely

$$\frac{1}{M} \sum_{a=1}^{M} \chi_a \sim 2p - 1 + 2\sqrt{\frac{p(1-p)}{M}} Z, \quad Z \sim \mathcal{N}(0,1). \tag{3.138}$$

This expression can be used to considerably simplify the self-consistency equations. Let us focus on the retrieval of the noisy patterns quantified by $\bar{n}$:

$$\begin{aligned} \bar{n} &= \mathbb{E}_{\phi,Z} \left(2p - 1 + 2\sqrt{\frac{p(1-p)}{M}} Z \right) \tanh\left[\beta M \bar{n} \left(2p - 1 + 2\sqrt{\frac{p(1-p)}{M}} Z \right) + \sqrt{\alpha \beta M \bar{p}} \phi \right] = \\ &= (2p-1)\bar{m} + \beta M \bar{n} \frac{4p(1-p)}{M} (1 - \bar{q}), \end{aligned} \tag{3.139}$$

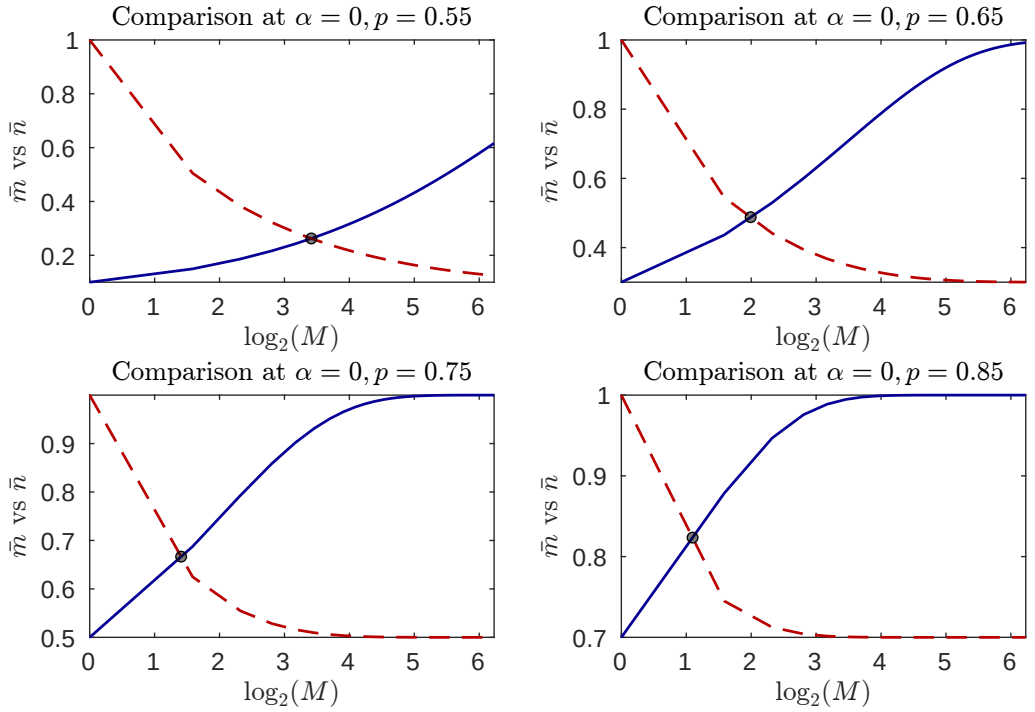

Figure 3.21: We compare the expected magnetizations $\bar{m}$ (solid line) and $\bar{n}$ (dashed line), obtained by numerically solving (3.135) and (3.136), holding in the limit of vanishing temperature $\beta \to \infty$ and infinite size $N \to \infty$, in the low load regime $\alpha = 0$. We notice that, as the size M of the dataset increases, the magnetization of the noisy example diminishes while that of the archetype starts to grow; we denote with $\tilde{M}$ the value of M corresponding to the intersection between the two curves. Different values of p are considered, as reported in the title of the panels.

where the last step has been performed via Wick theorem: $\mathbb{E}_Z Z f(Z) = \mathbb{E}_Z \partial_Z f(Z)$. This equation implies that, for large M, beyond n, the order parameter m – assessing the retrieval of archetypes – also starts to play a fundamental role; in fact, the configurations $\boldsymbol{\sigma} = \boldsymbol{\xi}^\mu$ emerge as ground states. Indeed, we have

$$\bar{n} = \frac{\bar{m} r}{1 - \beta(1 - \bar{q})(1 - r^2)}, \tag{3.140}$$

where, for simplicity, we posed $r = 2p - 1$.
This equation allows us to get rid of n and rather focus on m: by replacing (3.140) in the remaining self-consistencies we find

$$\bar{p} = \frac{\beta \bar{q}}{[1 - \beta(1 - \bar{q})]^2}, \quad G := \frac{\beta r^2}{1 - \beta(1 - r^2)(1 - \bar{q})}, \tag{3.141}$$

$$\bar{m} = \mathbb{E}_{\phi,Z} \tanh\left[G \bar{m} M \left(1 + Z \sqrt{\frac{1 - r^2}{r^2 M}}\right) + \phi \sqrt{\alpha \beta \bar{p} M} \right], \tag{3.142}$$

$$\bar{q} = \mathbb{E}_{\phi,Z} \tanh^2\left[G \bar{m} M \left(1 + Z \sqrt{\frac{1 - r^2}{r^2 M}}\right) + \phi \sqrt{\alpha \beta \bar{p} M} \right], \tag{3.143}$$

where the parameter G has been introduced to lighten the notation.
For a straight comparison to AGS theory, we introduce a more convenient scale for the

temperature, such that

$$\beta \to \frac{\beta}{\beta(q-1)\,(r^2-1)+r^2}. \tag{3.144}$$

Via this rescaling the self-consistent equations become

$$\bar{m} = \mathbb{E}_{\phi,Z} \tanh\left[\beta\bar{m}M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2} + \phi\beta\sqrt{\alpha\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}M}\right], \tag{3.145}$$

$$\bar{q} = \mathbb{E}_{\phi,Z} \tanh^2\left[\beta\bar{m}M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2} + \phi\beta\sqrt{\alpha\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}M}\right] \tag{3.146}$$

These equations can be further simplified as shown in the next

Proposition 7. *For the model introduced in Definition (10), in the thermodynamic limit and for large samples of examples ($M \gg 1$), the order parameters fulfill the following self-consistent equations:*

$$\bar{m} = \mathbb{E}_Z \tanh\left[\beta\bar{m}M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2 + \alpha\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}M}\right], \tag{3.147}$$

$$\bar{q} = \mathbb{E}_Z \tanh^2\left[\beta\bar{m}M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2 + \alpha\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}M}\right]. \tag{3.148}$$

Proof. Given a function F, we introduce the relation

$$\mathbb{E}_{X,Y} F(aX+bY+c) = \mathbb{E}_Z F(\sqrt{a^2+b^2}Z + c), \tag{3.149}$$

where X, Y, Z are assumed to be Gaussian random variables. This relation allows us to reduce any number of averages with the same structure to a single Gaussian average, and, in particular, by appling (3.149) to eqs. (3.145)-(3.146) we get eqs. (3.147)-(3.148). □

remark 15. *The argument of the hyperbolic tangents in (3.147)-(3.148) includes three contributions (and no longer just two as in the standard Hopfield scenario). Indeed, beyond the signal carried by $\bar{m}$ there are two sources of (slow) noise: a classic one given by the other patterns not retrieved (pattern interference) and a new one given by the examples within the dataset related to the pattern the network is retrieving (example interference).*

remark 16. *As a consistency check, we point out that if the network is not provided with datasets, but just patterns (i.e. $M = 1$) and those are assumed noiseless (i.e. $r = 1$), the whole theory collapses over the standard AGS theory of the Hopfield model as expected.*

Proposition 8. *To be sure that the archetype is retrieved over the noisy patterns we can use a simple argument, namely we can require that*

$$\beta M\bar{m} > \beta\sqrt{M}|Z|\sqrt{\frac{1-r^2}{r^2}\bar{m}^2 + \frac{\alpha}{r^4(1-\beta(1-\bar{q}))^2}\bar{q}}, \quad Z \sim \mathcal{N}(0,1) \tag{3.150}$$

holds almost surely: a solution in M to the above equation is given by

$$M > \frac{\gamma^2}{r^2}\left[1 - r^2 + \frac{q}{\bar{m}^2(1-\beta(1-\bar{q}))^2}\frac{\alpha}{r^2}\right] \tag{3.151}$$

where γ establishes the confidence level (indeed the last condition implies $|Z| < \gamma, \quad Z \sim \mathcal{N}(0,1)$ which can be satisfied up to an exceedingly small probability at finite M): these results recover the scaling behaviour achieved via signal to noise analysis in the previous section. In particular, in the low storage $\alpha = 0$ the correct scaling is $M \propto 1/(2p-1)^2$, while in the high storage $\alpha > 0$ the correct scaling is $M \propto 1/(2p-1)^4$.

Proof. The proof works by requiring that the signal term in the argument of tanh (3.147) is on average greater than the noise term, which amounts to the condition:

$$\beta\bar{m}M > |Z|\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2 + \alpha\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}M} \tag{3.152}$$

this condition can be recast as

$$|Z| < \frac{\sqrt{M}}{\sqrt{\frac{1-r^2}{r^2} + \frac{\alpha}{\bar{m}^2}\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}}} =: W(M) \tag{3.153}$$

if we further require

$$|Z| < \gamma < W(M) \tag{3.154}$$

by solving $W(m) > \gamma$ w.r.t M we obtain

$$M > \gamma^2\left[\frac{1-r^2}{r^2} + \frac{\alpha}{\bar{m}^2}\frac{\bar{q}}{r^4(1-\beta(1-\bar{q}))^2}\right] \tag{3.155}$$

concluding the proof. □

Now, to further inspect the competition between m and n, we resume Theorem 4, see in particular equations (3.120)-(3.121), which are used to build Fig. 3.22: the "Fuzzy" phase corresponds to a region in the parameter space where the retrieval of the examples is more effective than the retrieval of the archetype ($\bar{n} > \bar{m}$), no matter how good the retrieval can be. Focusing on the low-load regime, this region is demarcated by the line $\tilde{M}(p) := \tilde{M}(\alpha = 0, p)$; beyond that line the retrieval of the archetype is more effective than the retrieval of the example ($\bar{m} > \bar{n}$) and, by requiring also a high-quality retrieval (i.e., $|\bar{m}| > z$), we get the line $\tilde{M}_z(\alpha, p)$, which detects a region whose volume decreases with z. Focusing on the high-load regime, the "Fuzzy" region is demarcated by the line $\tilde{M}(\alpha, p)$, which is more restrictive that $\tilde{M}(\alpha, p)$.

Finally, we want to deepen the possible existence of a genuine phase transition distinguishing between a region where the system can infer the archetype ($\bar{m} > 0$) and a regione where noise – either fast (i.e., ruled by T) or slow (i.e., ruled by a suitable combination

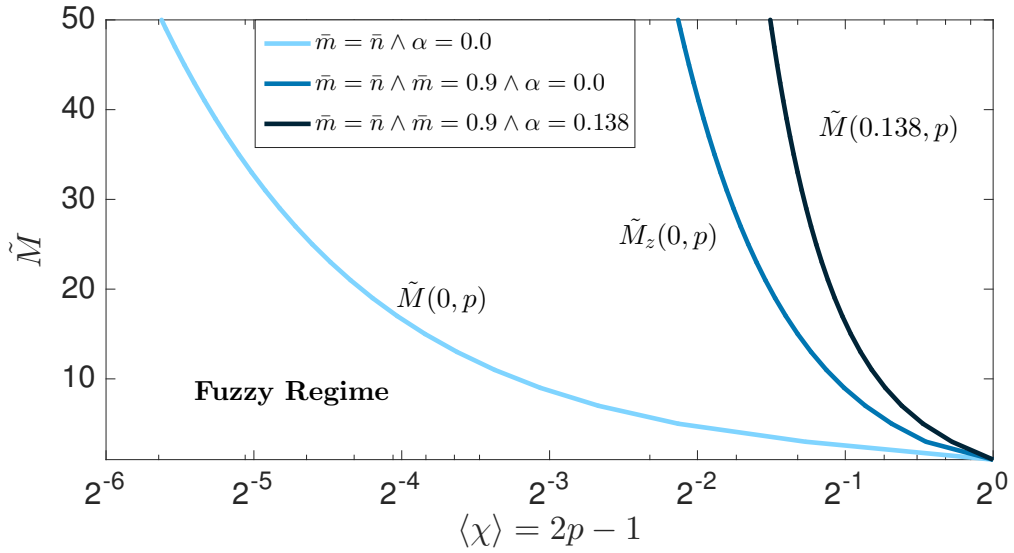

Figure 3.22: In this plot we show the crossovers values for M as a function of $2p-1$ and under different conditions. In particular, given $P = \alpha N$ archetypes and feeding the network with $M \times P$ examples characterized by a noise p, as $M > \tilde{M}(\alpha, p)$, then $\bar{m} > \bar{n}$. As expected, moving from a low load ($\alpha = 0$) to a high load ($\alpha > 0$), the region in this parameter space where $\bar{m} > \bar{n}$ shrinks. Notice that $\tilde{M}(\alpha, p)$ simply signs a crossover between $\bar{n}$ and $\bar{m}$, while no conditions are posed on the magnitude of magnetizations. This kind of information is provided by $\tilde{M}_z(\alpha, p)$ which also requires that $|\bar{m}| > z$. In this way, we can highlight a region where the pattern is better retrieved than examples *and* with high quality.

of α, r and M) – prevails ($\bar{m} = 0$). A close look to the self-consistent equations (3.147)-(3.148) suggests that a suitable, intensive and tuneable parameter able to trigger the phase transition is given by

$$\rho := \frac{\alpha}{Mr^4}. \tag{3.156}$$

In the following analysis we will let $M \to \infty$ and, accordingly, we rescale the temperature as $\beta \to \frac{\beta}{M}$ to ensure the well-definiteness of the model (3.75); this limit also implies that that we are focusing on the limit of high disorder in the dataset ($p \to 1/2$) so to retain a finite ρ.

Proposition 9. *In the limit of large samples ($M \to \infty$) and high disorder in the dataset ($r \to 0$) a critical behaviour is found as ρ approaches $\rho_c = \frac{2}{\pi}$, where $\bar{m} \sim \sqrt{\frac{3}{\pi}}\sqrt{2 - \pi\rho}$.*

Proof. Taking the large-M self-consistency equations (3.147)-(3.148), all that we have to

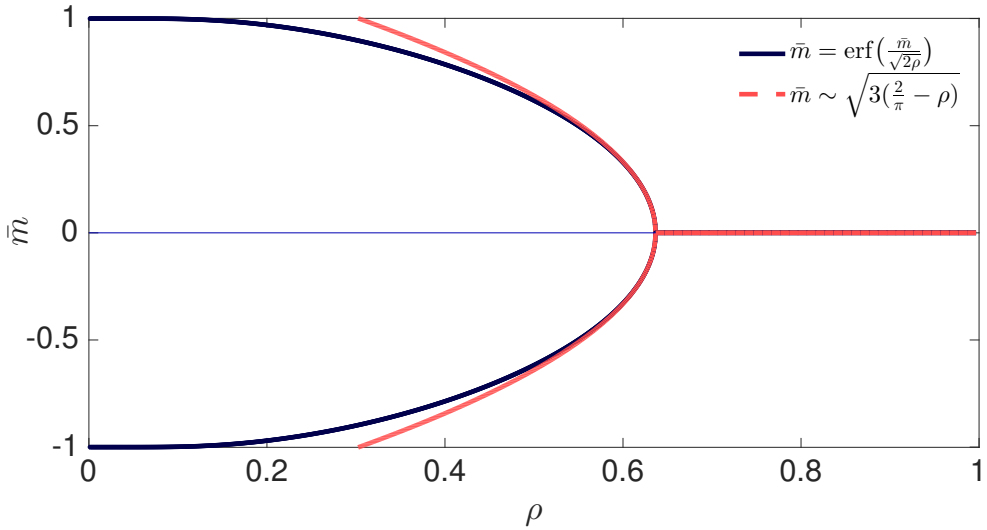

Figure 3.23: Zero-temperaure self-consistency for the Mattis magnetization in the limit of $M, N, K \to \infty$ such that $(0,1) \ni \rho := K/(MNr^4)$ is the tunable control parameter for the dataset density (see eq. 3.156): for values of ρ smaller than $\rho_c = 2/\pi$ the solely solution is $\bar{m} = 0$ while for values of $\rho > \rho_c$ two (gauge-invariant) not-null values of the Mattis magnetization appear. Beyond the exact result given by eq. (3.161), the figure also shows a comparison with the square-root estimate valid nearby the critical point.

do is replace r^2 with $\sqrt{\frac{\alpha}{\rho M}}$ and β with $\frac{\beta}{M}$ obtaining:

$$\bar{m} = \mathbb{E}_Z \tanh\left(\beta\bar{m} + Z\beta\sqrt{\frac{1-\sqrt{\frac{\alpha}{\rho M}}}{M\sqrt{\frac{\alpha}{\rho M}}} + \frac{\rho\bar{q}}{[1-\frac{\beta}{M}(1-\bar{q})]^2}}\right), \tag{3.157}$$

$$\bar{q} = \mathbb{E}_Z \tanh^2\left(\beta\bar{m} + Z\beta\sqrt{\frac{1-\sqrt{\frac{\alpha}{\rho M}}}{M\sqrt{\frac{\alpha}{\rho M}}} + \frac{\rho\bar{q}}{(1-\frac{\beta}{M}[1-\bar{q})]^2}}\right). \tag{3.158}$$

The whole theory now has been rephrased intensive in M, allowing us to take the limit $M \to \infty$:

$$\bar{m} = \mathbb{E}_Z \tanh(\beta\bar{m} + \beta Z\sqrt{\rho\bar{q}}) \quad \text{as} \quad M \to \infty, \tag{3.159}$$

$$\bar{q} = \mathbb{E}_Z \tanh^2(\beta\bar{m} + \beta Z\sqrt{\rho\bar{q}}) \quad \text{as} \quad M \to \infty. \tag{3.160}$$

In particular, the zero-temperature limit of the previous equations, where we send $\beta \to \infty$, reads as

$$\bar{m} = \mathbb{E}_Z \mathrm{sign}(\bar{m} + Z\sqrt{\rho\bar{q}}) = \mathrm{erf}\left(\frac{\bar{m}}{\sqrt{2\rho}}\right) \quad \text{as} \quad \beta, M \to \infty, \tag{3.161}$$

$$\bar{q} = \mathbb{E}_Z \mathrm{sign}(\bar{m} + Z\sqrt{\rho\bar{q}})^2 = 1 \quad \text{as} \quad \beta, M \to \infty. \tag{3.162}$$

By Taylor expanding equation (3.161) around $\bar{m} = 0$, a critical behaviour is found at $\rho_c = \frac{2}{\pi}$ with scaling $\bar{m} \sim \sqrt{\frac{3}{\pi}}\sqrt{2 - \pi\rho}$ near the critical point. □

The behavior of the magnetization $\bar{m}$ versus ρ, in the limit $M, N, K \to \infty$ is shown in Fig. 3.23, where the critical behavior is also corroborated.

remark 17. *As a direct consequence of the previous proposition we can state that concepts, namely* archetypes *of the experienced examples, are formed by the network via a critical behavior and not abruptly (as, for instance, happens to the Hopfield network when forgetting, i.e. the* blackout scenario*).*

3.5 Statistical mechanics approach to ultra-detection

3.5.1 General settings and main definitinos

In this Section, we report the technical details underlying the solution of the model provided by the equations $2.211 - 2.215$. Here, we will adopt a formal style to stress that the analysis is led by rigorous tools. In fact, beyond the signal-to-noise analysis performed in the previous Section, the numerical approach followed in the next Section and analytical non-rigorous methods based on the replica-trick (that we also carried out to check overall consistency, without presenting in details) that still retain a heuristic flavour, the problem can be actually addressed rigorously.

For completeness, let us recall the basic ingredients of the model.

Definition 13. *The Hamiltonian function for the DAM neural network is*

$$H_{DAM} = -\frac{1}{2N^3} \sum_{\rho=1}^{K} \left(\sum_{i,\mu=1}^{N,N} \eta_{i\mu}^{\rho} \sigma_i \tau_\mu \right)^2, \tag{3.163}$$

where $\sigma_i, \tau_\mu \in \{-1, +1\}$ *for* $i, \mu = 1, ..., N$ *and* $K = \alpha N$.

Following the preliminary analysis by the Signal-to-Noise analysis of the previous section, the noisy tensors $\boldsymbol{\eta}$ is given by

Definition 14. *The interaction strength for the dimer* (σ_i, τ_μ) *is defined as*

$$\eta_{i\mu}^{\rho} = \xi_{ij}^{\rho} + \sqrt{K}\tilde{\xi}_{i\mu}^{\rho} = \xi_i^{\rho}\xi_\mu^{\rho} + \sqrt{K}\tilde{\xi}_{i\mu}^{\rho}, \tag{3.164}$$

where the ξ_i^{ρ}*'s are i.i.d. drawn from* $\mathbb{P}(\xi_i^{\rho} = \pm 1) = 1/2$*, while the* $\tilde{\xi}_{i\mu}^{\rho}$*'s are i.i.d. drawn from* $\mathbb{P}(\tilde{\xi}_{i\mu}^{\rho}) = \mathcal{N}(0, 1)$.

3.5.2 Guerra's interpolation for the quenched pressure

Here, we extend Guerra's interpolating scheme [20] to deal with these dense networks. This technique works directly on the main quantity of interest in the Statistical Mechanical analysis, namely the quenched pressure associated to the cost function (3.163), as introduced in the next

Definition 15. *The quenched pressure density associated to the Hamiltonian (3.163) is defined as*

$$A(\alpha, \beta) \equiv \lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\boldsymbol{\eta}} \log Z, \tag{3.165}$$

where Z is the partition function associated to the Hamiltonian (3.163) given by

$$Z \equiv \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \exp\left[\frac{\beta}{2N^3}\sum_{\rho=1}^{K}\left(\sum_{i,\mu=1}^{N,N}\eta_{i\mu}^{\rho}\sigma_i\tau_\mu\right)^2\right], \tag{3.166}$$

and $\mathbb{E}_{\boldsymbol{\eta}}$ denotes the quenched average over the realizations of the tensor $\boldsymbol{\eta}$: for a generic function f of the tensor elements $\{\eta_{i\mu}^{\rho}\}$ this average is defined as standard.

We stress that the partition function can be written in terms of the magnetizations (2.128) as

$$Z = \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \exp\Big[\frac{\beta N}{2}\sum_{\rho=1}^{K}\Big(M_\rho + \sqrt{K}\tilde{M}_\rho\Big)^2\Big]. \tag{3.167}$$

Since we are looking for the retrieval regime, we shall assume that only a single information pattern (say ξ^1) is candidate for the condensation. Then,

$$\begin{aligned} Z = \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \exp\Big[&\frac{\beta N}{2}(M_1 + \sqrt{K}\tilde{M}_1)^2 + \\ &+\frac{\beta N}{2}\sum_{\rho=2}^{K}(M_\rho + \sqrt{K}\tilde{M}_\rho)^2\Big]. \end{aligned} \tag{3.168}$$

The magnetization M_1 associated to the retrieved pattern is of order $\mathcal{O}(1)$, while all the others M_ρ are $\mathcal{O}(N^{-1})$. This implies that, in the thermodynamic limit, we can neglect the subleading contributions, so that (note that this decomposition leads to the same results also in case of $M_1 \sim \mathcal{O}(N^{-1})$, i.e. when the network fails to retrieve the presented pattern. In that case, the overlap of the network with the external noise source $\tilde{\xi}^1$ is not negligible w.r.t. to the signal part. However, discarding it (i.e. only taking the last sum in (3.168)) would lead to a negligible error in the thermodynamic limit. Indeed, it is straightforward to check that the former is of order $\mathcal{O}(1)$, while the remaining contributions scale as $\mathcal{O}(N)$)

$$\begin{aligned} Z &\underset{N\to\infty}{\sim} \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \exp\left(\frac{\beta N}{2}M_1^2 + \frac{\alpha\beta N^2}{2}\sum_{\rho=2}^{K}\tilde{M}_\rho^2\right) = \\ &= \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}}\int D\mathbf{z}\exp\Big(\frac{\beta N}{2}m_\sigma^2 m_\tau^2 + \\ &\qquad + \sqrt{\alpha}\frac{\beta}{N}\sum_{\rho=2}^{K}\sum_{i,\mu=1}^{N,N}\tilde{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\Big), \end{aligned} \tag{3.169}$$

where, in the second line, we used the Hubbard-Stratonovich linearization (by this, $D\mathbf{z}$ is the $(K-1)$-dimension $\mathcal{N}(0,\beta^{-1})$ Gaussian measure) and we posed $M_1 = m_\sigma m_\tau$ with

$$m_\sigma \equiv \frac{1}{N}\sum_{i=1}^{N}\xi_i^1\sigma_i, \qquad m_\tau \equiv \frac{1}{N}\sum_{\mu=1}^{N}\xi_\mu^1\tau_\mu, \tag{3.170}$$

reflecting the factorization of the signal part of the interaction strength, i.e. $\xi_{i\mu}^1 = \xi_i^1\xi_\mu^1$. The expression in the last line of (3.169) is the starting point for our interpolation procedure.

Definition 16. *Guerra's interpolating function related to the quenched pressure of the DAM cost function (3.163) is*

$$\mathcal{A}(t) = \frac{1}{N}\mathbb{E}_{\boldsymbol{\eta}} \log Z_t,$$

where

$$\begin{aligned} Z_t \equiv \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \int D\boldsymbol{z} \exp\Big(& t\frac{\beta}{2} N m_\sigma^2 m_\tau^2 + \\ & + \sqrt{t}\sqrt{\alpha}\frac{\beta}{N} \sum_{\rho=2}^{K} \sum_{i,\mu=1}^{N,N} \tilde{\xi}_{i\mu}^{\rho} \sigma_i \tau_\mu z_\rho + \\ & + \sqrt{1-t}\ \mathcal{W} + (1-t)\ \mathcal{D}\Big) \end{aligned} \tag{3.171}$$

is the generalized partition function and $\mathcal{W}$ and $\mathcal{D}$ are defined as

$$\begin{aligned} \mathcal{D} \equiv & NC_1 m_\sigma + NC_2 m_\tau + C_6 \sum_{\rho=2}^{K} \frac{z_\rho^2}{2}, \\ \mathcal{W} \equiv & C_3 \sum_{i=1}^{N} \tilde{\xi}_i^{(1)} \sigma_i + C_4 \sum_{\mu=1}^{N} \tilde{\xi}_\mu^{(2)} \tau_\mu + C_5 \sum_{\rho=2}^{K} \tilde{\xi}_\rho z_\rho, \end{aligned} \tag{3.172}$$

where the external fields $\tilde{\xi}_i^{(1)}$, $\tilde{\xi}_\mu^{(2)}$ and $\tilde{\xi}_\rho$ are i.i.d. variables and $C_1, ..., C_6$ are suitable constants to be set a posteriori.

Definition 17. *Given a generic function $F(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{z})$ of the neurons, the (generalized) Boltzmann average $\omega_t(F)$ is defined as*

$$\begin{aligned} \omega_t(F) \equiv & Z_t^{-1} \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \int D\boldsymbol{z} F(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{z}) \exp\Big(t\frac{\beta}{2} N m_\sigma^2 m_\tau^2\Big) \times \\ & \times \exp\Big(\sqrt{t}\sqrt{\alpha}\frac{\beta}{N} \sum_{\rho=2}^{K} \sum_{i,\mu=1}^{N,N} \tilde{\xi}_{i\mu}^{\rho} \sigma_i \tau_\mu z_\rho\Big) \times \\ & \times \exp\Big(\sqrt{1-t}\ \mathcal{W} + (1-t)\ \mathcal{D}\Big). \end{aligned} \tag{3.173}$$

Note that, as standard in Guerra's interpolation techniques (see [47] for ferromagnets, [81] for spin glasses and [20] for neural networks), in the function $\mathcal{A}(t)$, the interpolating parameter appears with different exponents (1 and 1/2) mirroring the nature of the interaction (ferromagnetic and glassy, respectively) and, ultimately, the need to apply Wick's theorem.

remark 18. *Guerra's interpolating function evaluated at $t = 1$ corresponds to the original quenched pressure, in such a way that its explicit expression can be recovered via a simple sum rule by using the Fundamental Theorem of Calculus, i.e.*

$$\begin{aligned} A(\alpha, \beta) &= \lim_{N\to\infty} \mathcal{A}(t=1) = \\ &= \lim_{N\to\infty} \left(\mathcal{A}(t=0) + \int_0^1 dt\, \partial_t \mathcal{A}(t) \right). \end{aligned} \tag{3.174}$$

Therefore we now have to evaluate $\partial_t \mathcal{A}$ and $\mathcal{A}(0)$. This calculation is rather lengthy and goes along the same line as [20] without requiring any particular operation; for this reason, we shall report the explicit passages related only to the second term (that is, the most complex) in the argument of the exponential in $\mathcal{A}(t)$ and just for illustrative purposes. Then, let us pose

$$\begin{aligned} \mathcal{A}^{(2)}(t) &\equiv \frac{1}{N}\mathbb{E}_{\boldsymbol{\eta}} \log \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \int D\mathbf{z} \times \\ &\times \exp\left(\sqrt{t}\sqrt{\alpha}\frac{\beta}{N}\sum_{\rho=2}^{K}\sum_{i,\mu=1}^{N,N}\tilde{\xi}_{i\mu}^{\rho}\sigma_i\tau_\mu z_\rho\right), \end{aligned} \tag{3.175}$$

and define the generalized Boltzmann average $\omega_t^{(2)}$ as in Def. 17 (of course, when dealing with the generalized pressure $\mathcal{A}(t)$ rather than $\mathcal{A}^{(2)}(t)$, we have to replace $\omega_t^{(2)}$ with the corresponding Boltzmann average ω_t). Thus, deriving with respect to t, we get

$$\begin{aligned} \partial_t \mathcal{A}^{(2)}(t) &= \frac{\sqrt{\alpha}}{2\sqrt{t}}\frac{\beta}{N^2}\sum_{\rho=2}^{K}\sum_{i,\mu=1}^{N,N}\mathbb{E}_{\boldsymbol{\eta}}\ \tilde{\xi}_{i\mu}^{\rho}\ \omega_t^{(2)}(\sigma_i\tau_\mu z_\rho) \\ &= \frac{\alpha\beta^2}{2N^3}\sum_{\rho=2}^{K}\sum_{i,\mu=1}^{N,N}\mathbb{E}_{\boldsymbol{\eta}}\left[\omega_t^{(2)}(\sigma_i^2\tau_\mu^2 z_\rho^2) - \omega_t^{(2)}(\sigma_i\tau_\mu z_\rho)^2\right] \\ &= \frac{\alpha\beta^2}{2N^3}\sum_{\rho=2}^{K}\sum_{i,\mu=1}^{N,N}\mathbb{E}_{\boldsymbol{\eta}}\omega_t^{(2)}(z_\rho^2) - \frac{\alpha^2\beta^2}{2}\langle q_{12}r_{12}p_{12}\rangle_t^{(2)} \\ &= \frac{\alpha^2\beta^2}{2}\langle p_{11}\rangle_t^{(2)} - \frac{\alpha^2\beta^2}{2}\langle q_{12}r_{12}p_{12}\rangle_t^{(2)}. \end{aligned} \tag{3.176}$$

Here, in the second passage we applied Wick's theorem, in the third passage we exploited the Boolean nature of the σ and τ variables, we defined $\langle\cdot\rangle_t^{(2)} \equiv \mathbb{E}_{\boldsymbol{\eta}}\omega_t^{(2)}(\cdot)$ and introduced two-replica overlaps (one for each layer) to account for the quenched noise. More precisely:

$$q_{12} \equiv \frac{1}{N}\sum_{i=1}^{N}\sigma_i^1\sigma_i^2, \tag{3.177}$$

$$r_{12} \equiv \frac{1}{N}\sum_{\mu=1}^{N}\tau_\mu^1\tau_\mu^2, \tag{3.178}$$

$$p_{12} \equiv \frac{1}{K-1}\sum_{\rho=2}^{K}z_\rho^1 z_\rho^2. \tag{3.179}$$

Further, the first term in (3.176) can be cancelled by suitably choosing the constant C_6 (dedicated to tune the variance z^2). Repeating analogous calculations for the remaining terms making up $\mathcal{A}(t)$, overall we get

$$\partial_t \mathcal{A} = \mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{\beta}{2}\omega_s(m_\sigma^2 m_\tau^2)_t + \frac{\alpha^2\beta^2}{2}\omega_s(p_{11})_t + \tag{3.180}$$

$$-\frac{\alpha^2\beta^2}{2}\omega_s(p_{12}q_{12}r_{12})_t - C_1\omega_s(m_\sigma)_t - C_2\omega_s(m_\tau)_t + \tag{3.181}$$

$$-\frac{\alpha C_6}{2}\omega_s(p_{11})_t - \frac{C_3^2}{2} - \frac{C_4^2}{2} + \frac{C_3^2}{2}\omega_s(q_{12})_t + \tag{3.182}$$

$$+\frac{C_4^2}{2}\omega_s(r_{12})_t - \frac{\alpha C_5^2}{2}\omega_s(p_{11})_t + \frac{\alpha C_5^2}{2}\omega_s(p_{12})_t\Big], \tag{3.183}$$

where now $\langle\cdot\rangle_t \equiv \mathbb{E}_{\boldsymbol{\eta}}\omega_t(\cdot)$ and the quenched average $\mathbb{E}_{\boldsymbol{\xi}}$ applies only on the Boolean variables $\boldsymbol{\xi}$ as the Gaussian variables $\tilde{\boldsymbol{\xi}}$ have been already averaged out (via Wick's theorem). As anticipated, a trivial simplification can be implemented by setting $C_6 = \alpha\beta^2 - C_5^2$ in such a way that we get rid of $\omega_s(p_{11})$. A further simplification can be obtained asking for vanishing fluctuations for the order parameters, as prescribed by the RS approximation in the thermodynamic limit. Then, calling $(\overline{m}_\sigma,\ \overline{m}_\tau)$ and $(q,\ p,\ r)$ the RS values for, respectively, the Mattis magnetizations and the overlaps, the corresponding probability distributions in the thermodynamic limit satisfy

$$\lim_{N\to\infty} \mathbb{P}(m_\sigma) = \delta(m_\sigma - \overline{m}_\sigma), \tag{3.184}$$
$$\lim_{N\to\infty} \mathbb{P}(m_\sigma) = \delta(m_\tau - \overline{m}_\tau), \tag{3.185}$$
$$\lim_{N\to\infty} \mathbb{P}(q_{12}) = \delta(q_{12} - q), \tag{3.186}$$
$$\lim_{N\to\infty} \mathbb{P}(p_{12}) = \delta(p_{12} - p), \tag{3.187}$$
$$\lim_{N\to\infty} \mathbb{P}(r_{12}) = \delta(r_{12} - r). \tag{3.188}$$

Denoting with Δ the fluctuation of the generic observable w.r.t. its thermodynamic value (e.g., $\Delta = m_\sigma - \overline{m}_\sigma$), we can recast the interaction terms appearing in (3.180) as

$$\omega_s(p_{12}q_{12}r_{12})_t = -2pqr + pq\omega_s(r_{12})_t + pr\omega_s(q_{12})_t + \\ + rq\omega_s(p_{12})_t + \mathcal{O}(\Delta^2), \tag{3.189}$$
$$\omega_s(m_\sigma^2 m_\tau^2)_t = -3\overline{m}_\sigma^2\overline{m}_\tau^2 + 2\overline{m}_\sigma^2\overline{m}_\tau\omega_s(m_\tau)_t + \\ + 2\overline{m}_\tau^2\overline{m}_\sigma\omega_s(m_\sigma)_t + \mathcal{O}(\Delta^2). \tag{3.190}$$

Moreover, since in the RS regime fluctuations vanish, we can disregard terms $\mathcal{O}(\Delta^2)$, obtaining

$$\omega_s(p_{12}q_{12}r_{12})_t = -2pqr + pq\omega_s(r_{12})_t + pr\omega_s(q_{12})_t + rq\omega_s(p_{12})_t, \tag{3.191}$$
$$\omega_s(m_\sigma^2 m_\tau^2)_t = -3\overline{m}_\sigma^2\overline{m}_\tau^2 + 2\overline{m}_\sigma^2\overline{m}_\tau\omega_s(m_\tau)_t + 2\overline{m}_\tau^2\overline{m}_\sigma\omega_s(m_\sigma)_t. \tag{3.192}$$

Replacing these expressions inside the streaming term, and choosing our free parameters as

$$C_1 = \beta\overline{m}_\sigma\overline{m}_\tau^2, \tag{3.193}$$
$$C_2 = \beta\overline{m}_\tau\overline{m}_\sigma^2, \tag{3.194}$$
$$C_3^2 = \alpha^2\beta^2 pr, \tag{3.195}$$
$$C_4^2 = \alpha^2\beta^2 pq, \tag{3.196}$$
$$C_5^2 = \alpha\beta^2 qr, \tag{3.197}$$
$$C_6 = \alpha\beta^2(1-qr), \tag{3.198}$$
$$\tag{3.199}$$

we reach the simple result

$$\partial_t\mathcal{A} = \frac{\alpha^2\beta^2}{2}p(2qr - q - r) - \frac{3}{2}\beta\overline{m}_\sigma^2\overline{m}_\tau^2, \tag{3.200}$$

which is independent on t, so that the integration is trivial. Now, we must evaluate the

one-body term:

$$\mathcal{A}(0) = \frac{1}{N}\mathbb{E}_{\boldsymbol{\eta}} \log \sum_{\boldsymbol{\sigma},\boldsymbol{\tau}} \int D\mathbf{z} \times \tag{3.201}$$

$$\times \exp\Big(NC_1 m_\sigma + NC_2 m_\tau + C_6 \sum_{\rho=1}^{K} \frac{z_\rho^2}{2}\Big) \times \tag{3.202}$$

$$\times \exp\Big(C_3 \sum_{i=1}^{N} \tilde{\xi}_i^{(1)} \sigma_i + C_4 \sum_{\mu=1}^{N} \tilde{\xi}_\mu^{(2)} \tau_\mu + C_5 \sum_{\rho=1}^{K} \tilde{\xi}_\rho z_\rho\Big). \tag{3.203}$$

With straightforward computations and recalling the choices (3.193) for the C_i coefficients, we have

$$\mathcal{A}(0) = 2\log 2 + \mathbb{E}_x \log\cosh\left[\alpha\beta x\sqrt{pr} + \beta\overline{m}_\sigma \overline{m}_\tau^2\right] + \tag{3.204}$$

$$+\mathbb{E}_x \log\cosh\left[\alpha\beta x\sqrt{pq} + \beta\overline{m}_\sigma^2 \overline{m}_\tau\right] + \tag{3.205}$$

$$+\frac{\alpha^2\beta}{2}\frac{qr}{1-\alpha\beta(1-qr)} - \frac{\alpha}{2}\log[1-\alpha\beta(1-qr)], \tag{3.206}$$

where x is a standard Gaussian variable. Exploiting the sum rule (3.174) we have the final result

theorem 5. *In the thermodynamic limit, under the replica symmetric approximation, the quenched pressure density related to the cost function (3.163) can be expressed in terms of the natural order parameters of the model (i.e. the two Mattis magnetizations for the visible and mirror layers and the three two-replica overlaps of the visible, hidden and mirror layers) as follows*

$$A_{RS} = 2\log 2 + \mathbb{E}_x \log\cosh\left[\alpha\beta x\sqrt{pr} + \beta\overline{m}_\sigma \overline{m}_\tau^2\right] + \tag{3.207}$$

$$+\mathbb{E}_x \log\cosh\left[\alpha\beta x\sqrt{pq} + \beta\overline{m}_\sigma^2 \overline{m}_\tau\right] + \tag{3.208}$$

$$+\frac{\alpha^2\beta}{2}\frac{qr}{1-\alpha\beta(1-qr)} - \frac{\alpha}{2}\log[1-\alpha\beta(1-qr)] + \tag{3.209}$$

$$+\frac{\alpha^2\beta^2}{2}p(2qr-q-r) - \frac{3}{2}\beta\overline{m}_\sigma^2\overline{m}_\tau^2. \tag{3.210}$$

Its extremization selects the maximum entropy solutions that minimize the cost function 3.163 and yields to the self-consistent equations (11)-(15).

As a final remark, we note that, from a machine-learning perspective, beyond signal detection (involving the Mattis magnetizations), also quenched noise is to be estimated and, since the latter is carried by the overlaps, a first estimate can be obtained by a Plefka-like expansion of the free energy in the high (fast)-noise limit (see e.g., [153, 154, 155] and reference therein).

Bibliography

[1] R.J. Baxter, *Exactly solved models in statistical mechanics*, Courier Dover Publ., (2007).

[2] C. Kittel, *Elementary statistical physics*, Courier Dover Publications, (2004).

[3] H.C. Tuckwell, *Introduction to theoretical neurobiology*, Cambridge University Press (2005).

[4] L. Pastur, M. Shcherbina, B. Tirozzi, *The replica-symmetric solution without replica trick for the Hopfield model*, J. Stat. Phys. **74**.5:1161-1183, (1994).

[5] D. Ruelle, *Statistical mechanics: Rigorous results*, World Scientific (1999).

[6] Jaynes, E.T. *Probability theory: the logic of science*, St. Louis, Washington University, (1996).

[7] C.E. Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal, **27**: 379-423, 623-656, (1948).

[8] A.C.C. Coolen, R. Kuhn, P. Sollich, *Theory of neural information processing systems*, Oxford Press (2005).

[9] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).

[10] L. Zdeborova, F. Krzakala, *Statistical physics of inference: thresholds and algorithms*, Advances in Physics, **65**, 453-552, (2016).

[11] D.J. Amit, *Modeling brain functions*, Cambridge Univ. Press (1989).

[12] F. Guerra, F.L. Toninelli, *The thermodynamic limit in mean field spin glass models*, Comm. Math. Phys. **230**(1):71-79, (2002).

[13] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models*, Springer Science & Business Media, (2003).

[14] E. Agliari, et al, *Immune networks: multitasking capabilities near saturation*, J.Phys.A: Math. & Theor. **46**(41):415003, (2003).

[15] Schneidman, Elad, et al. *Weak pairwise correlations imply strongly correlated network states in a neural population*, Nature **440**(7087):1007-1012, (2006).

[16] J. Tubiana, R. Monasson, *Emergence of compositional representations in restricted Boltzmann machines*, Phys. Rev. Lett. **118**(13), 138301 (2017).

[17] A. Engel, C. Van den Broeck, *Statistical mechanics of learning*, Cambridge University Press (2001).

[18] D. Sherrington, S. Kirkpatrick, *Solvable model of a spin-glass*, Phys. Rev. Lett. **35**(26):1792, (1975).

[19] Mezard, M., Parisi, G. & Virasoro, M.A., Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications, *World Scientific Publishing Company* (1987).

[20] A. Barra, et al., *The replica symmetric approximation of the analogical neural network*, J. Stat. Phys. **140**(4):784, (2010).

[21] A. Barra, F. Guerra, G. Genovese, D. Tantari, *How glassy are neural networks?*, JSTAT P07009, (2012).

[22] M. Talagrand, *Rigorous results for the Hopfield model with many patterns*, Prob. Theor. & Rel. Fiel. **110**(2):177, (1998).

[23] M. Talagrand, *Exponential inequalities and convergence of moments in the replica-symmetric regime of the Hopfield model*, Ann. Prob. 1393-1469, (2000).

[24] A. Bovier, V. Gayrard, *Hopfield models as generalized random mean field models*, Mathematical aspects of spin glasses and neural networks, 3-89, Birkhauser, Boston (1998).

[25] A. Bovier, V. Gayrard, P. Picco, *Gibbs states of the Hopfield model in the regime of perfect memory*, Prob. Theor. & Rel. Fields **100**(3):329, (1994).

[26] E. Agliari, A. Barra, C. Longo, D. Tantari, *Neural Networks retrieving binary patterns in a sea of real ones*, J. Stat. Phys. **168**:1085, (2017).

[27] E. Agliari, et al., *Hierarchical neural networks perform both serial and parallel processing*, Neural Networks **66**, 22-35, (2015).

[28] E. Agliari, et al., *Multitasking attractor networks with neuronal threshold noises*, Neural Networks **49**, 19, (2013).

[29] A. Barra, G. Genovese, F. Guerra, *Equilibrium statistical mechanics of bipartite spin systems*, J. Phys. A (Math. & Theor.) **44**.24:245002, (2011).

[30] A. Barra, A. Di Biasio, F. Guerra, *Replica symmetry breaking in mean field spin glasses trough Hamilton-Jacobi technique*, JSTAT P09006, (2010).

[31] J. Hertz, R. Palmer, *Introduction to the theory of neural networks*, Lecture Notes, (1991).

[32] V. Dotsenko, *An introduction to the theory of spin glasses and neural networks*, World Scientific, (1995).

[33] P. Carmona, Y. Hu, *Universality in Sherrington-Kirkpatrick's spin glass model*, Ann. Henri Poincarè **42**, 2, (2006).

[34] L. Pastur, M. Shcherbina, B. Tirozzi, *On the replica symmetric equations for the Hopfield model*, J. Math. Phys. **40**(8): 3930, (1999).

[35] D. Krotov, J.J. Hopfield, *Dense associative memory is robust to adversarial inputs*, arXiv:1701.00939, (2017).

[36] R. Ellis, *Entropy, large deviations, and statistical mechanics*, Taylor & Francis press (2005).

[37] R.F. Thompson, *The neurobiology of learning and memory*, Science **233**.4767:941-947, (1986).

[38] Y. Miyashita, H.S. Chang, *Neuronal correlate of pictorial short-term memory in the primate temporal cortex*, Nature **331**.6151:68, (1988).

[39] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. **79**(8): 2554-2558 (1982).

[40] M. McCloskey, N.J. Cohen, *Catastrophic interference in connectionist networks: The sequential learning problem*, Psychol. Learn. & Motiv. **24**:109-165, (1989).

[41] J.A. Fodor, Z.W. Pylyshyn, *Connectionism and cognitive architecture: A critical analysis*, Cognition **28**(1):3-71, (1988).

[42] D. Ruelle, *Small Random Perturbations of Dynamical Systems and the Definition of Attractors*, Commun. Math. Phys. **82**, 137-151 (1981).

[43] D. Amit, H. Gutfreund, H. Sompolinsky, *Storing infinite numbers of patterns in a spin-glass model of neural networks*, Phys. Rev. Lett. **55**.14:1530, (1985).

[44] H. Steffan, R. Kühn, *Replica symmetry breaking in attractor neural network models*, Zeitschrift für Physik B Condensed Matter, **95**(2), 249–260 (1994).

[45] T. Stiefvater, K.R. Müller, R. Kühn, *Averaging and finite-size analysis for disorder: The Hopfield model*, Physica A: Statistical Mechanics and its Applications, **232**, 61-73 (1996).

[46] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, M.I.T. press (2017).

[47] A. Barra, *The mean-field Ising model trought interpolating techniques*, J. Stat. Phys. **132**, 406, (2006).

[48] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cognit. Sci. **9**.1:147, (1985).

[49] LeCun, Y., Bengio, Y., Hinton, G. *Deep learning*. Nature **521**.7553 (2015): 436-444.

[50] C.G. Gross, *Genealogy of the "grandmother cell"*, The Neuroscientist **8**.5:512-518, (2012).

[51] J.S. Bowers, *What is a grandmother cell? And how would you know if you found one?*, Connect. Sci. **23**.2:91-95, (2011).

[52] A. Barra, et al., *On the equivalence among Hopfield neural networks and restricted Boltzmann machines*, Neural Networks **34**, 1-9, (2012).

[53] A. Barra, et al., *Phase transitions of Restricted Boltzmann Machines with generic priors*, Phys. Rev. E **96**, 042156, (2017).

[54] F.E. Leonelli, L. Albanese, E. Agliari, A. Barra, *On the effective initialisation for restricted Boltzmann machines via duality with Hopfield model*, Neu. Nets. **143**, 314, (2021).

[55] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, L. Troyansky, *Determining computational complexity from characteristic phase transitions*, Nature **400**(6740), 133, (1999).

[56] A. Decelle, G. Fissore, C. Furtlehner, *Thermodynamics of restricted Boltzmann machines and related learning dynamics*, J. Stat. Phys. **172**.6:1576-1608, (2018).

[57] M. Mézard, *Mean-field message-passing equations in the Hopfield model and its generalizations*, Phys. Rev. E **95**(2), 022117 (2017).

[58] H. Huang, *Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses*, JSTAT 053302, (2017).

[59] C. Marullo, E. Agliari, *Boltzmann Machines as Generalized Hopfield Networks: a Review of Recent Results and Outlooks*, Entropy **23**.1:34, (2021).

[60] G.E. Hinton, R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science **313**, 504-507 (2006).

[61] L. Personnaz, I. Guyon, G. Dreyfus, *Information storage and retrieval in spin-glass like neural networks*, J. Phys. Lett. **46**, L-359:365, (1985).

[62] V. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, *Statistical mechanics of Hopfield-like neural networks with modified interactions*, J. Phys. A **24**, 2419, (1991).

[63] V. Dotsenko, B. Tirozzi, *Replica symmetry breaking in neural networks with modified pseudo-inverse interactions*, J. Phys. A **24**:5163-5180, (1991).

[64] A. Fachechi, E. Agliari, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, submitted to Neural Nets available at arXiv:1810.12217 (2018).

[65] G. Genovese, *Universality in bipartite mean field spin glasses*, J. Math. Phys. **53**(12):123304, (2012).

[66] I. Kanter, H. Sompolinsky, *Associative recall of memory without errors*, Phys. Rev. A **35**.1:380, (1987).

[67] E. Gardner, *The space of interactions in neural network models*, J. Phys. A **21**(1):257 (1988).

[68] A. Barra, F. Guerra, *About the ergodic regime of the analogical Hopfield neural network*, J. Math. Phys. **49**, 125217, (2008)

[69] F. Crick, G. Mitchinson, *The function of dream sleep*, Nature **304**, 111, (1983).

[70] J.J. Hopfield, D.I. Feinstein, R.G. Palmer, *Unlearning has a stabilizing effect in collective memories*, Nature Lett. **304**, 280158, (1983).

[71] S. Diekelmann, J. Born, *The memory function of sleep*, Nature Rev. Neuroscience **11**(2):114, (2010).

[72] J.A. Hobson, E.F. Pace-Scott, R. Stickgold, *Dreaming and the brain: Toward a cognitive neuroscience of conscious states*, Behavioral and Brain Sciences **23**, (2000).

[73] P. Maquet, *The role of sleep in learning and memory*, Science **294**.5544:1048, (2001).

[74] J.L. McGaugh, *Memory - a century of consolidation*, Science **287**.5451:248-251, (2000).

[75] T.J. Sejnowski, *Higher-order Boltzmann machines*, AIP Conference Proceedings 151 on Neural Networks for Computing, 398-403 (1986).

[76] R. Salakhutdinov, G. Hinton, *Deep Boltzmann machines*, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, 448-455 (2009).

[77] P. Baldi, S.S. Venkatesh, *Number of stable points for spin-glasses and neural networks of higher orders*, Phys. Rev. Lett. **58**.9:913 (1987).

[78] A. Bovier, B. Niederhauser, *The spin-glass phase-transition in the Hopfield model with p-spin interactions*, Adv. Theor. Math. Phys. **5**:1001-1046 (2001).

[79] M. Elad, *Sparse and redundant representation modeling: what next?*, IEEE Sign. Proc. Lett.s **19**:12 (2012).

[80] H. Barlow, *Redundancy reduction revisited*, Network: Comp. Neur. Sys. **12**, 241 (2001).

[81] F. Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Comm. Math. Phys. **233**, 1, (2003).

[82] E. Agliari, A. Barra, A. De Antoni, A. Galluzzi, *Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines*, Neur. Net. **38**:52 (2013).

[83] B. Li, R. Di Fazio, A. Zeira, *A low bias algorithm to estimate negative SNRs in an AWGN channel*, IEEE Comm. Lett.s **6**(11):469 (2002).

[84] J. Lehnert, M.B. Pursley, *Error probabilities for binary direct-sequence spread-spectrum communications with random signature sequences*, IEEE Trans. Comm. **35**(1):87 (1987).

[85] J. Barbier, *Overlap matrix concentration in optimal Bayesian inference*, arXiv:1904:02808 (2019).

[86] S. Cocco, R. Monasson, *Adaptive cluster expansion for inferring Boltzmann machines with noisy data*, Phys. Rev. Lett. **106**.9: 090601, (2011).

[87] H. Huang, *Reconstructing the Hopfield network as an inverse Ising problem*, Phys. Rev. E **81**.3:036104, (2010).

[88] S. Vedel, et al., *Migration of cells in a social context*, Proc. Natl. Acad. Sci. **110**.1:129-134, (2013).

[89] J.H. Lee, et al., *Microfluidic co-culture of pancreatic tumor spheroids with stellate cells as a novel 3D model for investigation of stroma-mediated cell motility and drug resistance*, J. Exp. & Clin. Can. Res. **37**.1:1-12, (2018).

[90] S. Sebens, H. Schafer, *The tumor stroma as mediator of drug resistance-a potential target to improve cancer therapy?*, Curr. Pharm. Biotech. **13**.11:2259-2272, (2012).

[91] E.S. Nakasone, et al., *Imaging tumor-stroma interactions during chemotherapy reveals contributions of the microenvironment to resistance*, Cancer cell **21**.4:488-503.

[92] E. Armingol, et al., *Deciphering cell-cell interactions and communication from gene expression*, Nat. Rev. Gen. **22**.2:71-88, (2021).

[93] Mora, Thierry, et al. *Maximum entropy models for antibody diversity*, Proc. Natl. Acad. Sci. **107**.12:5405-5410, (2010)

[94] E. Agliari, A. Barra, M. Castellana, M. Piel, P. Saez, P. Vergas, *A statitical-inference approach to reconstruct intercellular interactions in cell migration experiments*, Science Advances **6**, 11, (2020).

[95] Bialek, William, et al. *Statistical mechanics for natural flocks of birds*, Proc. Natl. Acad. Sci. **109**.13 (2012): 4786-4791.

[96] E. Lonardo, J. Frias-Aldeguer, P.C. Hermann, C. Heeschen, *Pancreatic stellate cells form a niche for cancer stem cells and promote their self-renewal and invasiveness*, Cell Cycle **11**:7, (2012).

[97] R.F. Hwang, et al., *Cancer-associated stromal fibroblasts promote pancreatic tumour progression*, Cancer Res. **68**:(3), 918 (2008).

[98] M. Amrutkar, et al., *Secretion of fibronectin by human pancreatic stellate cells promotes chemoresistance to gemcitabine in pancreatic cancer cells*, BMC Cancer **19**, 596, (2019).

[99] D. Delle Cave, et al., $TGF-\beta 1$ *secreted by pancreatic stellate cells promotes stemness and tumourigenicity in pancreatic cancer cells through* $L1CAM$ *downregulation*, Oncogene 1-15, (2020).

[100] Z. Xu, A. Vonlaufen, P.A. Phillips, E. Fiala-Beer, X. Zhang, *Role of Pancreatic Stellate Cells in Pancreatic Cancer Metastasis*, Am. J. Pathol. **177**(5), 2585, (2010).

[101] S. Suklabaidya, et al., *Experimental models of pancreatic cancer desmoplasia*, Lab. Invest. **98**, 27, (2018).

[102] C. Sousa, et al., *Pancreatic stellate cells support tumour metabolism through autophagic alanine secretion*, Nature **536**, 479, (2016).

[103] P.P. Provenzano, C. Cuevas, A.E. Chang, V.K. Goel, D.D. Von Hoff, S.R. Hingorani, *Enzymatic Targeting of the Stroma Ablates Physical Barriers to Treatment of Pancreatic Ductal Adenocarcinoma*, Cancer Cell **21**(3), 418, (2012).

[104] J. Gore, M. Korc, *Pancreatic Cancer Stroma: Friend or Foe?*, Cancer Cell **25**(6), 711, (2014).

[105] B.C. Ozdemir, et al., *Depletion of Carcinoma-Associated Fibroblasts and Fibrosis Induces Immunosuppression and Accelerates Pancreas Cancer with Reduced Survival*, Cancer Cell **25**(6), 719, (2014).

[106] A.D. Rhim, et al., *Stromal Elements Act to Restrain, Rather Than Support, Pancreatic Ductal Adenocarcinoma*, Cancer Cell **25**(6), 735, (2014).

[107] M. Apte, J. Wilson, A. Lugea, S. Pandol, *A Starring Role for Stellate Cells in the Pancreatic Cancer Microenvironment*, Gastroenterology **144**(6), 1210, (2013).

[108] M.J. Ware, V. Keshishian, J.J. Law, J.C. Ho, C.A. Favela, P. Rees, *Generation of an in vitro 3D PDAC stroma rich spheroid model*, Biomaterials **108**, 129, (2016).

[109] E. Tomas-Bort, M. Kieler, S. Sharma, J.B. Candido, D. Loessner, *3D approaches to model the tumour microenvironment of pancreatic cancer*, Theranostics **10**(11):5074, (2020).

[110] W. Bialek, *Biophysics: searching for principles*, Princeton University Press, (2012).

[111] Skilling, John, ed. *Maximum Entropy and Bayesian Methods*, Cambridge, England, Springer Science, 2013.

[112] E. Agliari, et al., *Cancer driven dynamics of immune cells in a microfluidic enviroment*, Sci. Rep. **4**, 6639, (2014).

[113] E. Biselli, et al., *Organs on chip approach: a tool to evaluate cancer-immune cells interactions.* Sci. Rep. **7**, 1, (2017).

[114] A. Mencattini, et al., *Discovering the hidden messages within cell trajectories using a deep learning approach for in vitro evaluation of cancer drug treatments*, Sci. Rep. **10**, 1, (2020).

[115] L. De Monte, et al., *Intratumor T helper type 2 cell infiltrate correlates with cancer-associated fibroblast thymic stromal lymphopoietin production and reduced survival in pancreatic cancer*, J. Exp. Med. **208**.3:469-478, (2011).

[116] Acebron, J.A., et al., *The Kuramoto model: A simple paradigm for synchronization phenomena*, Rev. Mod. Phys. **77**.1 (2005): 137.

[117] Beaumont, M. A., Cornuet, J. M., Marin, J. M., & Robert, C. P., *Adaptive approximate Bayesian computation.* Biometrika, 96(4), 983-990, (2009).

[118] Shaffer, F. & Ginsberg, J.P., An overview of Heart Rate Variability Metrics and Norms, *Front. Public Health* **5**, 258-287 (2017).

[119] Melillo, P. & al., Automatic prediction of Cardiovascular and Cerebrovascular Events using Heart Rate Variability Analysis, *Plos-One* **10**(3), e0118504-e0118512 (2015).

[120] Laborde, S., Mosley, E. & Thayer, J.F., Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research, *Front. Phychol.* **3**, 213-229 (2017).

[121] Voss, A., Schulz, S., Schroeder, R., Baumert, M. & Caminal, P., Methods derived from non-linear dynamics for analysis heart rate variability, *Phil. Trans. R. Soc A* **367**, 277-296 (2009).

[122] McCraty, R. & Shaffer, F., Heart Rate Variability: new perspectives on physiological mechanisms, assesment of self-regulatory capacity and health risk, *Global Adv. Health Med.* **4**(1), 46-61 (2015).

[123] Beauchine, T.P. & Thayer, J.F., Heart Rate Variability as a transdiagnostic biomarker of psychopathology, *International Journal of Psychophysiology* **98**, 338-350 (2015).

[124] Agliari, E. & al., Detecting cardiac pathologies via machine learning on the clinical markers based on heart rate variability, *submitted to Sci. Rep.* (2020).

[125] Sassi, R. & al., Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Hearth Rhythm Association co-endorsed by the Asia-Pacific Heart Rhytm Society, *Europace* **17**, 1341-1348 (2015).

[126] Melillo, & P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N. & Pecchia, L., Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis, *PLoS ONE* **10**(3), e0118504-e0118510 (2015).

[127] Peltola, M.A., Role of editing of R?R intervals in the analysis of heart rate variability, *Front. Physiol.* **3**, 148-166 (2012).

[128] Barra, O.A. & Moretti, L., The "Life Potential" a new complex algorithm to assess "heart rate variability" from Holter records for cognitive and dignostic aims, *avaiable at arXiv:1310.7230*, (2013).

[129] Tkacik, G. & al., The simplest maximum entropy model for collective behavior in a neural network, *JSTAT* **2013**(03), P03011-P03043 (2013).

[130] Tao, W., Jin, H. & Liu, L., Object segmentation using ant colony optimization algorithm and fuzzy entropy, *Patt. Recog. Lett.* **28**(7), 788-796 (2007).

[131] Li, K., Rudiger, H. & Ziemssen, T., Spectra Analysis of Heart Rate Variability: Time Window Matters, *Front. Neurol.* **10**, 545-561 (2019).

[132] Ivanov, P.C., Rosenblum, M.G., Peng, J.E., Mietus, Havlin, S., Stanley, E.H. & Goldberger, A.L., Scaling and universality in heart rate variability distributions, *Physica A* **249**, 587-593 (1998).

[133] Sarlis, N.V., Skordas, E.S. & Varotos, P.A., Heart rate variability in natural time and $1/f$ noise, *Europhys. Letts.* **87**, 18003-19009 (2009).

[134] Pilgram, B. & Kaplan, D.T., Nonstationarity and $1/f$ noise characteristics in heart rate, *Amer. Phys. Soc.* **0363**, 6119-6125 (1999).

[135] Jonason, K., Vincent, E., Hammann, J., Bouchaud, J.P. & Nordblad, P., Memory and chaos effects in spin glasses, *Phys. Rev. Lett.* **81**(15), 3243-3247 (1998).

[136] Banavar, J.R. & Bray, A.J., Chaos in spin glasses: A renormalization-group study, *Phys. Rev. B* **35**(16), 8888-8900 (1987).

[137] Kondor, I., On chaos in spin glasses, *J. Phys. A* **22**(5), L163-L169 (1989).

[138] Van Mourik, J. & Coolen, A.C.C., Cluster derivation of Parisi's RSB solution for disordered systems, *J. Phys. A* **34**(10), L111-L1116 (2001).

[139] Ocio, M., Bouchiat, H. & Monod, P., Observation of $1/f$ magnetic fluctuations in a spin glass, *Jour. de Phys. Letts.* **46**(14), 647-652 (1985).

[140] Refregier, P. & al., Equilibrium magnetic fluctuations in spin glasses: Temperature dependence and deviations from $1/f$ behaviour, *Europhys. Letts.* **3**(4), 503-509 (1987).

[141] Pytte, E. & Imry, Y., Ubiquity of logarithmic scaling, $1/f$ power spectrum, and the $\pi/2$ rule, *Phys. Rev. B* **35**(3), 1465-1478 (1987).

[142] M.B. Weissman, $1/f$ *noise and other slow nonexponential kinetics in condensed matter*, Rev. Mod. Phys. **60**(2):537, (1988).

[143] P. Gongwen & H.J. Herrmann, *Density waves and* $1/f$ *density fluctuations in granular flow*, Phys. Rev. E **51**(3):1745, (1995).

[144] Peng, C.K.,Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H.E., Long-range correlations in nucleotide sequences, *Nature* **356**(6365), 168-170 (1992).

[145] Mercik, S., Weron, K. & Siwy, Z., Statistical analysis of ionic current fluctuations in membrane channels, *Phys. Rev. E* **60**(6), 7343-7363 (1999).

[146] Baiesi, M. & Paczuski, M., Scale-free networks of earthquakes and aftershocks, *Phys. Rev. E* **69**(6), 066106-066122 (2004).

[147] Rammal, R., Tannous, C. & Tremblay, A.M.S., $1/f$ noise in random resistor networks: fractals and percolating systems, *Phys. Rev. A* **31**(4), 2662-2681 (1985).

[148] Bak, P., Chao & T., Wiesenfeld, K., Self-organized criticality: An explanation of the $1/f$ noise, *Phys. Rev. Lett.* **59**(4), 381-385 (1987).

[149] Christensen, K., Zeev, O. & Bak, P., Deterministic $1/f$ noise in nonconserative models of self-organized criticality, *Phys. Rev. Lett.* **68**(16), 2417-2421 (1992).

[150] Sakellariou, J., Tria, F., Loreto V. & Pachet, F., Maximum entropy models capture melodic styles, *Sci. Rep.* **7**, 9172-9185 (2017).

[151] Duchi, J., Hazan, E. & Singer, Y., Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *J. Mach. Learn. Res.* **12**, 2121-2159 (2011).

[152] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, *Generalized Guerra's interpolating techniques for dense associative memories*, Neur. Netw. **128**, 254, (2020).

[153] A. Barra, *Irreducible free energy expansion and overlaps locking in mean field spin glasses*, J. Stat. Phys. 123(3):601, (2006).

[154] H. Jacquin, A. Rançon, *Resummed mean-field inference for strongly coupled data*, Phys. Rev. E **94**.4:042118, (2016).

[155] V. Sessak, R. Monasson, *Small-correlation expansions for the inverse Ising problem*, J. Phys. A. **42**(5):055001, (2009).