Università del Salento

DIPARTIMENTO DI MATEMATICA E FISICA
'ENNIO DE GIORGI'

PhD Thesis in Theoretical Physics

Statistical mechanics for Artificial Intelligence: Learning, Retrieving, Unlearning and Sleeping.

Alberto Fachechi

Advisors:	**Referees:**
Prof. Adriano Barra	**Prof. Francesco Guerra**
Prof. Elena Agliari	**Prof. Ido Kanter**

XXXI CICLO

List of publications

Publications included in the thesis

1. A. Barra, M. Beccaria, A. Fachechi, *A new mechanical approach to handle generalized Hopfield neural networks*, published in Neural Networks (2018).

2. A. Fachechi, E. Agliari, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, Neural Networks, in press (2018).

3. E. Agliari, F. Alemanno, A. Barra, A. Fachechi, *Dreaming neural networks: rigorous results*, J. Stat. Phys., in press (2019).

Publications not included in the thesis

1. M. Beccaria, A. Fachechi, G. Macorini, *Virasoro vacuum block at next-to-leading order in the heavy-light limit*, published in JHEP (2016).

2. M. Beccaria, A. Fachechi, G. Macorini, *On the cusp anomalous dimension in the ladder limit of* $\mathcal{N} = 4$ *SYM*, published in JHEP (2016).

3. M. Beccaria, A. Fachechi, G. Macorini, L. Martina, *Exact partition functions for deformed* $\mathcal{N} = 2$ *theories with* $N_f = 4$ *flavours*, published in JHEP (2016).

4. E. Alfinito, M. Beccaria, A. Fachechi, G. Macorini, *Reactive immunization on complex networks*, published in EPL (2017).

5. M. Beccaria, A. Fachechi, G. Macorini, *Chiral trace relations in* Ω*-deformed* $\mathcal{N} = 2$ *theories*, published in JHEP (2017).

6. E. Alfinito, M. Beccaria, A. Fachechi, G. Macorini, *Probing complexity with epidemics: a new reactive immunization strategy*, Proceedings of COMPLEXIS 2017, 2nd International Conference on Complexity, Future Information Systems and Risk (2017).

7. A. Fachechi, M. Beccaria, G. Macorini, *Chiral trace relations in* Ω*-deformed* $\mathcal{N} = 2$ *theories*, Proceedings of International Conference on Integrable Systems and Quantum symmetries (ISQS25, 2017), published in Journal of Physics: Conference Series 965 (2018).

8. E. Alfinito, A. Barra , M. Beccaria, A. Fachechi, G. Macorini, *An evolutionary game model for behavioral gambit of loyalists: Global awareness and risk-aversion*, published in EPL (2018).

9. A. Fachechi, M. Beccaria, G. Macorini, *Chiral trace relations in* $\mathcal{N} = 2^*$ *supersymmetric gauge theories*, Proceedings of the conference Physics and Mathematics of Nonlinear Phenomena (PMNP2017), published in Theoretical and Mathematical Physics 196 (2018).

10. E. Agliari, F. Alemanno, A. Barra, A. Fachechi, *A novel derivation of the Marchenko-Pastur law through analog bipartite spin-glasses*, Proceedings of the conference Disordered serendipity: a glassy path to discovery (2018), to be published in Journal of Physics A (2019).

Contents

Introduction

The spontaneous mechanisms behind the sophisticated information processing tasks taking place in the brain has always been a fascinating subject of study, and many of their characteristics are yet to be understood. Starting from the 1940s, this field has developed with the contributions and efforts of an incredibly large variety of scientists [91], ranging from engineers (mainly involved in electronics and robotics), physicists (mainly involved in statistical mechanics and stochastic processes), and mathematicians (mainly working in logics and graph theory), to neurobiologists and cognitive psychologists. Thanks to their work, we now know that the cerebral cortex can be described as a *neural network*, namely an interconnected web of nerve cells (*neurons*) transmitting electrical signals.[1] Indeed, the interest in studying brain functionalities and neural networks is mainly three-fold, depending on the relevant aspects one would like to consider. From the biological side, one aims to understand information processing in real biological nervous tissue. From the technological point of view (mainly conducted by engineers and computer scientists), one would apply the principles of neural functionalities to design "intelligent" systems [132] exhibiting learning capabilities and taking benefits of massive parallel calculus characterizing the brain. Finally, from the perspective of mathematicians and theoretical physicists (which is the one we shall adopt throughout the entire thesis), the challenge is to understand how neural networks (real systems but also artificial models) develop highly non-trivial emergent behaviours (which is the natural object of study of statistical mechanics).

A remarkable feature of neural networks is certainly its *adaptibility.* From

[1] In a nutshell, neurons, which are excitable and rather noisy elements, can produce electrical pulses (*spikes*) which are used to communicate among themselves [11]. This mutual interconnection of the neurons leads to a brain network structure with a highly non-trivial topology which can vary consistently depending on the regions (e.g. those in the *cortical modules* are rather uniform and we will use them as candidate to be modeled) and their associated cerebral functionalities. For instance, regions devoted to data preprocessing are quite regular, while those associated to cognitive functions are indeed almost amorphous.

one side, we can for example recognize objects even if they are partially visible, deformed or in low visibility conditions. On the other side, our brain system is also able to "re-wire" the nerve fibers in order to bypass damages in specific areas. With these facts in mind and once building blocks (e.g. neurons) of neural systems are understood, the central question in the field is to understand how neural networks self-organize to implement their capabilities. The key point is that these tasks are not operated by the neurons themselves. Indeed they are *very* simple systems as they can either be active (i.e. produce an electrical signal) or quiescent (i.e. not producing electric signals), hence their state can be mimicked by Boolean variables 0/1 or - in physical jargon - *Ising spins* ±1). The complexity of tasks which a neural network can accomplish is therefore a result of brain cells interactions. A fundamental step was done by J. Von Neumann [133] in 1957, who proposed that, in view of the large number of interacting neurons ($\sim 10^{10}$, each of which communicating with roughly 10^4 colleagues) and the intrinsic stochasticity of neural processes, neural network models and operations has to be described in statistical language. In particular, the theoretical approach that has since then been used, and that we will follow in this thesis, is *statistical mechanics*. In fact, the general strategy of statistical mechanics is to abandon any (reductionist) ambition to solve models of such systems at the microscopic level of individual elements, and to use the macroscopic vision to derive laws describing the behaviour of a suitably chosen set of global observables. When applied to neural network models, such an approach, which turned out to be very successful to describe the dynamics of particle and matter systems (see e.g. [28, 43, 78]), reveals the possibility to well-describe collective phenomena (e.g. ordered behaviours and phase transitioning), serving as a guide in choosing the macroscopic observables to consider and in establishing the difference between relevant mathematical subtleties and irrelevant ones. However, as in any statistical theory, clean and transparent mathematical laws are expected to emerge only for large (preferably infinitely large) systems.

Once neurophysiologists were able to give a complete description of the neuron's microscopic behaviour (thanks to work of A. Hodgkin and A.F. Huxley [37]), 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 components alone) was definitely appealing in neuroscience. Remarkably, although rather trivial with respect to the overall cerebral functionalities like learning or computation, the neural dynamics (describing the state of a neuron in terms of the state of the neighbouring ones through the so-called activation function) was found to be particularly apt to a thermodynamic formulation and ultimately to reveal possible emergent capabilities.

It should be stressed that building up such a theory required many concepts and tools originally developed in the field of condensed matter. In fact, theoretical physicists quickly realized that the purely kinetic Hamiltonian, introduced for perfect gases (or Hamiltonian with mild potentials allowing for real gases), is no longer suitable for solids, where atoms do not move freely and the main energy contributions come from potentials. However, as experimentally revealed by crystallography, nuclei are arranged according to regular lattices, hence motivating mathematicians to study periodical structures and help physicists in this modeling, but merging statistical mechanics with lattice theories resulted soon in practically intractable models.[1] It is just due to an effective shortcut to bypass this problem, namely the so called *mean field approximation*, that statistical mechanics approached complex systems and, in particular, *artificial intelligence*, as we will thoroughly see during this thesis.

Let us now address the subject of the thesis more specifically. Artificial Intelligence is trivially intelligence exhibited by machines. It is built along humans' *congnition* mechanisms, which are basically the mental actions or processes of acquiring knowledge and understanding experience and sensations. Despite we dare to enlarge the following *minimal paradigm* with the research summarized in the last Chapters of this thesis, at present, the two pillars of the cognitive process are the abilities to *learn* and *retrieve* information: one is useless without the other, because there is no reason why we should gather information if there is no way to recall it, and we cannot recover notions if we have not previously learnt them. These two aspects of human cognition have been naturally and successfully implemented into machines.

Some of the disciplines that have been attracted towards the study of Artificial Intelligence, actually specialized in either learning or retrieval: engineering worked mainly on the former and mathematics and theoretical physics most studied the latter. In the last few decades, machine learning has strongly developed and engineers have reached remarkable results (from speech and object recognition, to robot locomotion and computer vision) and it still is a prolific field of research and applications. The most recent advancements have been reached through the evolution of machine learning, commonly called *deep learning* [83, 115].

[1] For example, the magnetic Ising model has been resolved in dimensions 1 and 2 but we're still waiting for a solution for the 3-dimensional case.

Meanwhile, mathematicians and theoretical physicists worked to reach rigorous results concerning the retrieval of a machine's stored data, and to create a theory that illustrates neural networks behaviour under different conditions [11, 37]. In these theoretical developments, neural models are built over the human brain modules activity scheme and preserve the *associative memory* property, namely the ability to reconstruct a piece of information, once supplied with solely partial data (much as we do when we recognize a friend by a glance at just a part of its face). In the network model, data is stored in the form of *patterns* of information, i.e. vectors codifying a particular feature of the memory. For example, a black and white image can be stored as a pattern where each component, that is associated to a pixel in the picture, takes on the value 1 if the pixel white or -1 if the pixel is black.

The prototype example for a vast class of associative memory models is the *Hopfield neural network*, introduced by J. Hopfield in 1982 [65], see also [67]. It is a network of binary neural units $\sigma_i \in \{-1, +1\}$ fully connected by couplings which encode the stored patterns.[1] When looking at this neural network model from a statistical mechanics point of view, an interesting feature emerges, namely its relationship with two different mean-field Ising-spin models, namely the Curie-Weiss (CW) and the Sherrington-Kirkpatrick (SK) models. The former is a mean-field ferromagnetic model and it represents the archetype of a simple system (i.e. the number of free-energy minima of the system does not scale with the volume N), while the latter is a mean-field spin-glass [93] and it representes the archetype of a complex system (i.e. the number of free-energy minima does scale with the volume according to a proper function of N). The Hopfield model merges certain characteristics of these limits in such a way that we can read the CW and the SK as its two extremal cases.[2] As we will see, the mean-field ferromagnetic model can be interpreted as a very basic neural network that can only store one pattern, while the mean-field spin glass represents a system where the stored information is by far too much for the network to be able to recall anything. For these reasons, in neural network literature the Hopfield model is often introduced after the study of these models [11, 37, 102].

What do we know about the standard models of memory behaviour? Mathematicians and theoretical physicists have studied under which conditions this property would emerge, finding that there are two determining factors for the presence of a retrieval phase:

[1] For the scope of this introduction, we shall not go into details. For a comprehensive description of the Hopfield network, see chapter 4.

[2] A thorough analysis of this property is given in section 4.2

- The complexity of the patterns: the vector components are taken as either boolean (for example a black and white picture) or real (for example a coloured picture);

- The patterns amount: we quantify this variable in terms of a network capacity λ, defined as the number of stored patterns P over the number of the available neurons for their handling N, i.e. $\lambda \sim P/N$. Network operational modes are historically split into two main categories: the *low storage* case, when the number of stored patterns grows sublinearly (e.g. logarithmically) with system's size, hence $\lim_{N\to\infty} \lambda = 0$, and the *high storage* case, if the law describing the growth of the number of patterns is linear with respect to the system volume such that $\lim_{N\to\infty} \lambda > 0$.[1]

We will show that, in a low load (this, of course, includes also the case where P is finite), we can always have a retrieval phase, no matter whether patterns are dichotomic or analogical. On the other hand, in the high load regime, if the patterns are Boolean there is a memory loss over a critical λ_c [13]. Finally, if the representing patterns have real components, there is no hope for recovering information [37]. So, from a theoretical point of view (machine retrieval), working with dichotomic memories is more productive because they are easier to retrieve, but the application results (machine learning) underline the importance of having these patterns to be real-valued variables (for instance, solely the latter allow useful *trick* as variational principles and calculus in general - helping machine learning; however, it is intuitive that dealing with real number is by far more expensive then with rational - forcing machine retrieval). To overcome this flaw, beyond a standard study of the Hopfield model in one of these two extrema (i.e. equipped with Boolean or Gaussian patterns), we will study also an *hybrid* network, equipped with both the types of storable information (digital and analogical). In this case, we will be able to prove that this (more useful and realistic) network actually shares the same retrieval region of the standard Hopfield model with binary patterns: the critical capacity is smaller, but it does exist. This results confers overall strength to the unifying picture that learning and retrieval are two inseparable aspects of cognition, a - fundamental - picture emerged mathematically clearly just in recent times [21].

In this thesis, we will sensibly enrich the outlined scenario by deepening the potential beneficial role of another *state of the network* beyond *learning*

[1] By Gardner's Theory [53, 52] we know that super-linear growth as, e.g. $P \propto N^x$ with $x > 1$ are impossible with just pairwise interactions -as in the Hopfield scenario- thus there is no need to investigate such a regime.

and *retrieving*, namely *sleeping*, in particular:

- Regarding the *physical side*, the most important result will be by far the discovery that, allowing the network to "sleep", its critical capacity can reach the maximal one, i.e. $\lambda_c = 1$, as prescribed by Gardner's theory (and we stress that - without sleeping - the network is really far away from this bounds as the standard critical capacity is $\lambda_c \sim 0.14$)! In a nutshell, we will give the associative neural network a *daily prescription* (that we have called the *reiforcement&removal* extension). During its *awake state*, the network is fed by inputs (i.e. *patterns* of information) that are stored in an Hebbian fashion,[1] then, during its *sleep*, it gets rid off the (combinatorial[2]) proliferation of spurious mixtures (unavoidably created as metastable states in the free-energy landscape of the network during the learning stage), and it reinforces the pure ones (makings their free energy minima deeper in this *landscape picture*). This procedure, remarkably, keeps the learning phase Hebbian in its nature but allows the network to saturate the storage capacity λ to its upper bound[3] for symmetric networks (i.e., $\lambda_c = 1$) [53]. Further, in the retrieval phase of its phase diagram, pure states are global minima up to $\lambda \sim 0.85$: a much broader range w.r.t. the classical Hopfield counterpart, where they remain global minima solely for $\lambda < 0.05$).

- Regarding the *mathematical side*, we sensibly extended Guerra's interpolation techniques by adapting them to work with the more challenging models we introduced to account for sleeping in neural networks,. These techniques are basically based on complex statistical mechanical and PDE tools. In particular, statistical mechanics of spin glasses [93] has been playing a primary role in the investigation of neural networks, as for the description of both their learning phase [49, 117] and their retrieval properties [11, 37]. Along the past decades, beyond the bulk of results achieved via the so-called replica-trick [93] (the first celebrated method exploited to tackle these systems as pioneered by Giorgio Parisi), a considerable amount of rigorous results exploiting alterna-

[1] We stress that, given the equivalence between restricted Boltzmann machines and Hopfield neural networks [21], also learning via e.g. *contrastive divergence* [113] ultimately falls in the Hebbian category [9, 8].

[2] This means that the number of spurious states roughly grows exponentially in the number of stored patterns, that is, in the high storage regime, in the number of neurons.

[3] Actually the network seems to perform even *better*, returning its maximal capacity to be $\lambda_c \sim 1.07 > 1$: this is obviously not possible and, as explained by Dotsenko and Tirozzi [47, 46], it is a chimera of the replica-symmetric regime at which the theory is developed.

tive routes (possibly mathematically more transparent) were also developed mainly due to the interpolation techniques by Francesco Guerra (see e.g. [3, 4, 10, 29, 30, 32, 18, 25, 24, 47, 46, 126, 125, 101, 102, 59, 55] and references therein). Typically, in our approach, we first obtain new results *heuristically* with the replica trick, and we then confirm - in any detail - these findings by adapting the interpolation scheme to the case, step by step.

As this is a Ph.D. thesis in (theoretical) physics, beyond the two main routes paved respectively by Parisi and Guerra, particular efforts and care will be spent also to adapt an entirely *mechanistic approach to complex systems, and neural networks in particular*, the Hamilton-Jacobi technique (and, more in general, the non-linear PDE theory as a concrete alternative to the standard statistical mechanical approaches). We will show, at first on simple and complex paradigm - as the mean field ferromagnet (i.e. the Curie-Weiss model) and the mean field spin glass (i.e. the Sherrington-Kirkpatrick model), then on harmonic oscillators for learning and retrieval in Artificial Intellicenge (namely the Boltzmann machine for the former and the Hopfield network for the latter), how this technique allows to get an exhaustive picture of the spontaneous information processing skills these network display. Beyond the new results that we will present in this thesis, this observation by itself gives an entire new argument favoring the extensive usage of Theoretical Physics as a methodological guide to understand Artificial Intelligence, a must that our society will have to face in the incoming decade.

The thesis is is split into three main parts:

- Part One (Fundamentals of the required knowledge in Theoretical Phyics), where - once introduced a mandatory know-how (mainly about Statistical Mechanics and Statistical Inference) - we analyze the emerging properties of two archetypes of simple and complex systems (respectively the Curie-Weiss and the Sherrington-Kirkpatrick models);
- Part Two (Fundamental of Artificial Intelligence), where we show, extensively relying upon the concepts exposed in the first part, the state of the art in Neural Networks and Machine Learning (from a statistical mechanical perspective);
- Part Three (Novel results in Artificial Intellicence), where, once highlighted a relation between a larger critical storage capacity for pattern

retrieval with an enhanced skill in avoiding over-fitting during machine learning, we will extensively model sleeping phenomena in these network. In light of the results we will achieve (the saturation of the critical capacity), we will speculate that *Cognition* could be a composed and dynamical phenomenon shown by neural networks among whose salient ingredients, sleeping must have a weight, as learning and retrieval.

As a final remark, as this is a doctoral thesis in Theoretical Physics, we would like to emphasize that, when trying to enlarge the actual state of the art, we have been entirely driven by the underling Physics. In a nutshell, we proved that, in order to approach the statistical mechanical picture of these models, it can be possible to relate it to a mechanical system obeying an Hamilton-Jacobi evolution. Remarkably, it is just by pursuing this route - an effective Lagrangian mechanics description of these networks - that we naturally discover limitations of the previous framework (i.e. cognition split between just learning and retrieval) and we naturally enlarge such a scheme by introducing also *sleeping.* This fact highlights once more the importance for such a young discipline as Artificial Intelligence of being built on solid pillars; in these regards, those Theoretical Physics offer have always played a primary role, so we do hope the latter to become the main route in this (or the few) decisive decade(s) of investigation.

Part One: Tools from Theoretical Physics

Chapter 1

Statistical Mechanics and Statistical Inference

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 functioning 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). Instead of focusing on each particular particle trajectory to characterize the state of the system, 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, it is possible to infer the macroscopic behaviour 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. arrow of time coded in 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 [28, 78].

One step forward beyond the perfect gas, 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.

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 [131]) 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 a merely 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 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 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 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 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. However, 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

[1] Of course we are tacitely assuming the reader to be by far 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.

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

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

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 states 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:

$$\begin{aligned} \frac{1}{N} \ln \mathcal{N} &= \frac{1}{N}\Big(N \ln N - \sum_{i=1}^{K} N_i \ln N_i\Big) = \sum_{i=1}^{K} \frac{N_i}{N}\Big(\ln N - \ln N_i\Big) = \\ &= -\sum_{i=1}^{K} \frac{N_i}{N} \ln \frac{N_i}{N} = -\sum_{i=1}^{K} \mathcal{P}_i \ln \mathcal{P}_i = S_N[\mathcal{P}], \end{aligned}$$

in which the probability $\mathcal{P}_i$ has been identified with the frequency N_i/N thanks to the law of large numbers. 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 will show how Gibbs measure has the ability of maximizing entropy function. 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

$$\begin{cases} \max_{\mathcal{P}_j} S_N[\mathcal{P}], \\ \sum_{i=1}^{N} \mathcal{P}_i E_i = E_N, \\ \sum_{i=1}^{N} \mathcal{P}_i = 1. \end{cases} \tag{1.2}$$

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

$$dF_N = \frac{\partial F_N}{\partial \beta} d\beta + \sum_{i=1}^{N} \frac{\partial F_N}{\partial E_i} dE_i = -E_N d\beta - \beta \sum_{i=1}^{N} \frac{N_i}{N} dE_i, \tag{1.3}$$

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 (1.3) as

$$d\big(F_N + E_N\beta\big) = \beta\Big(dE_N - \sum_{i=1}^{N} \frac{N_i}{N} dE_i\Big), \tag{1.4}$$

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 free energy of the system:

$$F_N(\beta) \equiv -\frac{1}{\beta} \ln Z_N = E_N - TS_N. \tag{1.5}$$

The importance of free energy F_N is that it is a state function that can be expressed through the system order parameters.

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 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 (1.2). Plugging the definitions of entropy (1.1) and average energy into (1.5) and imposing the minimum conditions, we have

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

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

$$f(\beta) \doteq \lim_{N\to\infty} \frac{1}{N} F_N(\beta) = \lim_{N\to\infty} -\frac{1}{\beta N} \ln Z_N. \tag{1.6}$$

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

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

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 pressure function with respect to each order parameter in order to find its critical points where we have a minimum (or a maximum in the case we are dealing with the free energy). We obtain a state equation for the order parameters whose solution can be determined either analytically or numerically. Thanks to this procedure, we will be able to trace a phase diagram and analyze potential phase transitions in the so-called *thermodynamic limit*.

It is time to clarify some points. Firstly, 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 [102], as, along the same reasoning, just in that limit phase transitions do exist [112]), 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 an elementary description of the system under study. Finally, we would like to spend some words about *phase transitions*. We shall refer to this concept whenever we have an order parameter describing the state of the system that, depending on some tunable parameters characterizing the model, changes its value from zero to not-null values (or vice versa). In that value of the tunable parameters, whenever the free energy is continuous and its first derivative with respect to the order parameter investigated is not continuous, we say that we deal with a *first order* phase transition. On the other side, if the free energy and its first derivative are continuous, but the second derivative is not, we speak of *second order* phase transition, or *criticality*, and so on. Once that all the phase transitions have been identified in the space of the tunable parameters, it is possible to trace the phase diagram of the model and start working on another model.

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 ramification. Namely, 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 $\{\mathcal{P}_i, i = 1, \ldots, N\}$ 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 [72] in the 1950s, is based on the observation that the (Shannon [118]) entropy associated to a random variable X, that is

$$S(X) = -k \sum_{x \in A} \mathcal{P}(x) \ln \mathcal{P}(x), \tag{1.8}$$

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 [37, 85, 136]). 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 the available information.

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

$$S[\mathcal{P}^*] = \max_{\mathcal{P}} \{S[\mathcal{P}]\} \quad \text{such that} \quad \langle f_\alpha(x)\rangle = \bar{f}_\alpha. \tag{1.9}$$

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

  $$S_0[\mathcal{P}] = S[\mathcal{P}] + k\lambda_0\Big(\sum_{x\in A}\mathcal{P}(x) - 1\Big).$$

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

  $$\begin{aligned}\frac{\partial S_0[\mathcal{P}]}{\partial \mathcal{P}(x)} &= -k\ln\mathcal{P}(x) - k + k\lambda_0 = 0, && (1.10)\\ \frac{\partial S_0[\mathcal{P}]}{\partial \lambda_0} &= k\Big(\sum_{x\in A}\mathcal{P}(x) - 1\Big) = 0, && (1.11)\end{aligned}$$

  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).

- 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 function to maximize can immediately be written in Lagrangian form as

  $$\begin{aligned}S_2[\mathcal{P}] = S[\mathcal{P}] &+ k\lambda_0\Big(\sum_{x\in A}\mathcal{P}(x) - 1\Big) + k\lambda_1\Big(\sum_{x\in A}x\mathcal{P}(x) - \mu_1\Big)\\ &+ k\lambda_2\Big(\sum_{x\in A}\mathcal{P}(x)(x-\mu_1)^2 - \mu_2\Big).\end{aligned} \tag{1.12}$$

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

  $$\begin{aligned}\frac{\partial S_0[\mathcal{P}]}{\partial \mathcal{P}(x)} &= 0, && (1.13)\\ \frac{\partial S_0[\mathcal{P}]}{\partial \lambda_s} &= 0, && (1.14)\end{aligned}$$

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

$$\mathcal{P}^*(x) = \frac{1}{Z}e^{\lambda_1 x + \lambda_2 x^2} = \frac{1}{Z}e^{-(x-\hat{\lambda}_1)^2/2\hat{\lambda}_2^2},$$

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

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)!

[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.

Chapter 2

Simple systems: the Curie-Weiss paradigm

2.1 Generalities

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 [28, 78]. 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.[1]

[1] We would like to stress also that another reason to introduce the mean-field approximation of the Ising model is that, in one dimension, the latter is unable to explain phase transitions in ferromagnetic materials, since the free energy is an analytic function of the order parameters of the theory for $T \neq 0$. This is due to the fact that spin-spin correlations vanish very fast (i.e. exponentially) for $T > 0$, so they are not sufficient to provide an ordered phase of the system.

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 its 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]

2.2 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 [37]). Therefore, we can give the following

Definition 2.1. The Hamiltonian function $H_N(\boldsymbol{\sigma})$ of the mean field ferromagnetic model (CW) is:

$$H_N(\boldsymbol{\sigma}) = -\frac{1}{N} \sum_{1 \leq i < j \leq N} \sigma_i \sigma_j = -\frac{1}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j + \frac{1}{2}. \tag{2.1}$$

Remark 2.1. In the last definition, the last term $1/2$ can be ignored since it is irrelevant in the thermodynamic limit.

Remark 2.2. 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 2.2. The order parameter for the CW model is the (global) magnetization m defined as

$$m(\boldsymbol{\sigma}) \equiv m = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \in [-1, 1]. \tag{2.2}$$

[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].

Using this definition, we can also rewrite the Hamiltonian (2.1) 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 $H_N(m) \propto N \cdot \text{const}(N)$, where $\text{const}(N)$ means that the quantity is N-independent.

Definition 2.3. For a given inverse temperature $\beta = 1/T$, the partition function $Z_N(\beta)$ is defined as

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} B_N(\beta) = \sum_{\boldsymbol{\sigma}} e^{-\beta H_N(\boldsymbol{\sigma})} = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N}\sum_{ij}\sigma_i\sigma_j}, \tag{2.3}$$

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

Definition 2.4. The Gibbs measure $\omega_N(\cdot)$ for a generic function F depending on $\boldsymbol{\sigma}$ is

$$\omega_N(F) \doteq \frac{\sum_{\boldsymbol{\sigma}} F(\boldsymbol{\sigma}) B_N(\beta)}{\sum_{\boldsymbol{\sigma}} B_N(\beta)}. \tag{2.4}$$

Definition 2.5. The thermodynamic 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 (which we will use later on for several times); 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 give also this method of resolution as a preparatory step to its application to the mean field spin-glass and to 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.

2.3 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 uniqueness is 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 [112]. 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 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 (2.3), 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:

$$\begin{aligned} \frac{1}{N} \ln Z_N(\beta) \geq & \frac{1}{N} \ln \sum_{\boldsymbol{\sigma}} e^{\beta M \sum_i \sigma_i} + \frac{1}{N} \ln e^{-\frac{1}{2}\beta M^2 N} \geq \\ \geq & \frac{1}{N} \ln \Big(\prod_{i=1}^{N} \sum_{\boldsymbol{\sigma}} e^{\beta M \sigma_i} \Big) - \frac{1}{2} \beta M^2 \geq \\ \geq & \sup_{M \in [-1,1]} \Big\{ \ln 2 + \ln \cosh(\beta M) - \frac{1}{2} \beta M^2 \Big\}, \end{aligned} \tag{2.5}$$

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:

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \sum_{M} \delta_{mM} e^{\frac{1}{2}\beta N m^2}, \tag{2.6}$$

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 (2.6) 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 (2.5), we have the resulting upper bound:

$$\frac{1}{N}\ln Z_N(\beta) \leq \ln\frac{N+1}{N} + \sup_{M\in[-1,1]}\Big\{\ln 2 + \ln\cosh(\beta M) - \frac{1}{2}\beta M^2\Big\}. \qquad (2.7)$$

The upper (2.7) and lower (2.5) 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 [60]. 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

$$m_1(\boldsymbol{\sigma}) = \frac{1}{N_1}\sum_{i=1}^{N_1}\sigma_i, \qquad (2.8)$$

$$m_2(\boldsymbol{\sigma}) = \frac{1}{N_2}\sum_{i=N_1+1}^{N}\sigma_i, \qquad (2.9)$$

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

$$m(\boldsymbol{\sigma}) = \frac{N_1}{N}m_1(\boldsymbol{\sigma}) + \frac{N_2}{N}m_2(\boldsymbol{\sigma}). \qquad (2.10)$$

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

$$Z_N(\beta) \leq \sum_{\boldsymbol{\sigma}} \exp\beta\left(N_1 m_1^2(\boldsymbol{\sigma}) + N_2 m_2^2(\boldsymbol{\sigma})\right) = Z_{N_1}(\beta)Z_{N_2}(\beta), \qquad (2.11)$$

hence

$$Nf_N(\beta) = -\frac{1}{\beta}\log Z_N(\beta) \geq N_1 f_{N_1}(\beta) + N_2 f_{N_2}(\beta). \qquad (2.12)$$

This is the well known property of superadditivity of the free energy in the system size. 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. (2.7), 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

$$Z_N(\beta,t) = \sum_{\boldsymbol{\sigma}} \exp\beta \left(NtJm^2(\boldsymbol{\sigma}) + N_1(1-t)Jm_1^2(\boldsymbol{\sigma}) + N_2(1-t)Jm_2^2(\boldsymbol{\sigma})\right). \tag{2.13}$$

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

$$\begin{aligned} -\frac{1}{N\beta}\log Z_N(1) &= f_N(\beta), && (2.14)\\ -\frac{1}{N\beta}\log Z_N(0) &= \frac{N_1}{N} f_{N_1}(\beta) + \frac{N_2}{N} f_{N_2}(\beta). && (2.15)\end{aligned}$$

Taking the derivative respect to t, we obtain

$$-\frac{d}{dt}\frac{1}{N\beta}\log Z_N(\beta,t) = -\omega_t\Big(m^2(\boldsymbol{\sigma}) - \frac{N_1}{N}m_1^2(\boldsymbol{\sigma}) - \frac{N_2}{N}m_2^2(\boldsymbol{\sigma})\Big) \geq 0, \tag{2.16}$$

where $\omega_t(\cdot)$ denotes the Boltzmann-Gibbs thermal average corresponding to the t-dependent partition function (2.13). Then, integrating over t between 0 and 1 and recalling the boundary conditions (2.14, 2.15), one finds again the superadditivity property (2.12).

2.4 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 (2.1) and the related partition function (2.3) we introduce the following generalized partition function

$$\begin{aligned} Z_N(\beta,t) \doteq& \sum_{\boldsymbol{\sigma}} \exp\left\{\frac{\beta t}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j + (1-t)\psi\sum_{i=1}^{N}\sigma_i\right\} = \\ =& \sum_{\boldsymbol{\sigma}} \exp\left\{\frac{\beta t}{2}Nm^2 + (1-t)\psi N m\right\}, \end{aligned} \tag{2.17}$$

with m defined in (2.2), $t \in [0,1]$ 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 (2.4). The key observation is enclosed in the next

Proposition 2.1. *The statistical pressure for a finite volume N can be written in the following way thanks to the fundamental theorem of calculus:*

$$\alpha_N(\beta) \equiv \alpha_N(\beta, t = 1) = \alpha_N(\beta, t = 0) + \int_0^1 ds \, \Big[\partial_t \alpha_N(\beta, t)\Big]_{t=s}. \tag{2.18}$$

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

$$\begin{aligned}\alpha_N(\beta, t = 0) =& \frac{1}{N} \ln Z_N(\beta, t = 0) = \frac{1}{N} \ln \Big(\sum_{\boldsymbol{\sigma}} e^{\psi \sum_i \sigma_i} \Big) = \\ =& \frac{1}{N} \ln \Big(\prod_{i=1}^{N} \sum_{\boldsymbol{\sigma}} e^{\psi \sigma_i} \Big) = \ln 2 + \ln \cosh(\psi),\end{aligned} \tag{2.19}$$

while the derivative in (2.18) is

$$\begin{aligned}\frac{\partial}{\partial t} \alpha_N(t) =& \frac{1}{N} \frac{\partial_t Z_N(\beta, t)}{Z_N(\beta, t)} = \frac{1}{N Z_N(\beta, t)} \Big[\sum_{\boldsymbol{\sigma}} \Big(\frac{\beta N}{2} m^2 - \psi N m \Big) B_N(t) \Big] = \\ =& \frac{\beta}{2} \omega_t(m^2) - \psi \omega_t(m).\end{aligned} \tag{2.20}$$

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 $M \in [-1, +1]$. Then, trivially we have

$$\omega_t\big((m - M)^2\big) = \omega_t\big(m^2\big) + M^2 - 2M \omega_t(m). \tag{2.21}$$

[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.

Looking back at the final result of equation (2.20), 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 (2.21), we can write a convenient expression for the pressure derivative as

$$\frac{\partial}{\partial t}\alpha_N(\beta, t) = \frac{\beta}{2}\Big(\omega_t\big(m^2\big) - \frac{2\psi}{\beta}\omega_t(m)\Big) = \frac{\beta}{2}\omega_t\Big((m-M)^2\Big) - \frac{1}{2}\beta M^2. \quad (2.22)$$

Finally, plugging the results of equations (2.19) and (2.22) into (2.18), we can state that the pressure function is defined by the next

Theorem 2.1. *The infinite volume limit of the the Curie-Weiss statistical pressure $\alpha(\beta)$ can be written in terms of the magnetization as*

$$\alpha_N(\beta) = \sup_{M\in[-1,1]}\Big\{\ln 2 + \ln\cosh(\beta M) - \frac{1}{2}\beta M^2 + \frac{\beta}{2}\omega\Big((m-M)^2\Big)\Big\}, \quad (2.23)$$

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*

$$M = \tanh(\beta M), \quad (2.24)$$

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

2.5 The Hamilton-Jacobi formalism

This next method is not as immediate as the previous one, but it has a much wider range of usage and it is by far conceptually deeper. The approach we want to use is to “extend” the parameter space and investigate which PDEs are obeyed by the model in such a space, in order to inherit the technology for their resolution from classical mechanics. In particular, we want to relate the two-body and one-body interactions with respectively a fictitious time and space coordinates and check if the free energy derivatives w.r.t. spacetime combine as it happens in classical mechanics for the action).

In order to exploit this idea, we give the following

Definition 2.6. The generalized partition function in the Hamilton-Jacobi framework is defined as

$$Z_N(t,x) = \sum_{\boldsymbol{\sigma}} B_N(t,x) = \sum_{\boldsymbol{\sigma}} \exp\Big\{ \frac{t}{2N} \sum_{i,j=1}^{N} \sigma_i\sigma_j + x\sum_{i=1}^{N}\sigma_i \Big\},$$

in which the exponential is the generalized Boltzmann factor $B_N(t,x)$.

With this generalization, the definitions of $\alpha_N(t,x)$ and of the Gibbs average $\omega_{t,x}(\cdot)$ naturally follow. Classical statistical mechanics is of course recovered in the free field case by setting $t=\beta$ (or $t=\beta J$ if we work with not-unitary couplings) and $x=\beta h$ (or $x=0$ in case of zero external field). In the same way, the averages $\omega_{t,x}(\cdot)$ will be denoted by $\omega(\cdot)$ whenever evaluated in the sense of statistical mechanics. Let us introduce the following

Definition 2.7. The action $S_N(t,x)$ of this *mechanical analogy* mathematically shares the same structure of the pressure, as it reads

$$S_N(t,x) = -\frac{1}{N}\ln \sum_{\boldsymbol{\sigma}} \exp\Big\{ \frac{t}{2N} \sum_{i,j=1}^{N} \sigma_i\sigma_j + x\sum_{i=1}^{N}\sigma_i \Big\}. \qquad (2.25)$$

As anticipated, the variables t,x can be thought of as fictitious time and space coordinates. Moreover, calling $S(t,x)=\lim_{N\to\infty} S_N(t,x)$, we have of course $S(t=\beta,x=0)=\beta f(\beta)=-\alpha(\beta)$.

In order to highlight this approach, we now need to work out the spatial and temporal derivatives of $S_N(t,x)$, which read

$$\begin{aligned}
\frac{\partial S_N(t,x)}{\partial t} &= -\frac{1}{2}\omega_N(m^2)_{t,x},\\
\frac{\partial S_N(t,x)}{\partial x} &= -\omega_N(m)_{t,x},\\
\frac{\partial^2 S_N(t,x)}{\partial x^2} &= N\left(\omega_N(m^2)_{t,x} - \omega_N(m)^2_{t,x}\right).
\end{aligned}$$

Following Guerra's prescription [17] and noticing the form of the previous derivatives, it is possible to build a Hamilton-Jacobi equation for $S_N(t,x)$ as stated in the next

Proposition 2.2. *The pressure of the CW model in statistical mechanics plays the role of the Guerra's action in Lagrangian mechanics. Indeed, it obeys the following Hamilton-Jacobi PDE*

$$\partial_t S_N(t,x) + \frac{1}{2}\left(\partial_x S_N(t,x)\right)^2 + V_N(t,x) = 0, \qquad (2.26)$$

with potential $V_N(t,x) = -(1/2N)\partial^2_{xx} S_N(t,x) = (1/2)\big(\omega_N(m^2) - \omega_N(m)^2\big)$.

It is important to stress that, by virtue of the self-average of the order parameter m (i.e. $|\omega_N(m^2) - \omega_N(m)^2| \to 0$ for $N \to \infty$), the potential vanishes in the infinite size limit. Therefore, in this Lagrangian mechanistic equivalence, the thermodynamics of the CW model is painted as a free Galilean trajectory.

Moreover, deriving equation (2.26) with respect to the x variable and calling $u_N(t,x) = \partial_x S_N(t,x) = -\omega_N(m)$, we get the Burgers equation for the velocity (i.e. the magnetization in the statistical mechanics framework apart the minus sign):

$$\partial_t u_N(t,x) + u_N(t,x)\partial_x u_N(t,x) - \frac{1}{2N}\partial^2_{xx} u_N(t,x) = 0, \tag{2.27}$$

where the viscous term is controlled by the system size. In particular, this equation becomes naturally inviscid in the thermodynamic limit, since $S_N(t,x)$ admits the thermodynamic limit thanks to boundaries we have given in Section (2.3). Performing the Cole-Hopf transformation

$$\Phi_N(t,x) = \exp\Big\{ -N \int dx\, u_N(t,x)\Big\} = e^{-NS_N(t,x)}, \tag{2.28}$$

we derive this last expression with respect to the t variable, we use expression (2.26) and we obtain

$$\begin{aligned}\partial_t \Phi_N(t,x) =& -N\partial_t S_N(t,x)\cdot \Phi_N(t,x) = \\ =&\frac{N}{2}\big(\partial_x S_N(t,x)\big)\Phi_N - \frac{1}{2}\partial^2_{xx} S_N(t,x)\Phi_N(t,x) = \\ =&\frac{1}{2N}\partial^2_{xx}\Phi_N(t,x).\end{aligned}$$

Therefore, Φ_N satisfies the heat equation:

$$\frac{\partial \Phi_N(t,x)}{\partial t} - \frac{1}{2N}\frac{\partial \Phi_N(t,x)}{\partial^2 x} = 0, \tag{2.29}$$

with the initial condition

$$\begin{aligned}\Phi_N(0,x) =& e^{-NS_N(0,x)} = \exp\Big\{\ln \sum_\sigma e^{x\sum_i \sigma_i}\Big\} = \\ =& \exp\Big\{N \ln(2\cosh(x))\Big\},\end{aligned} \tag{2.30}$$

The heat equation can be easily solved in the Fourier space, through the Green propagator and the convolution theorem. The Fourier transform $\hat{\Phi}_N(t,k)$ of $\Phi_N(t,x)$

$$\hat{\Phi}_N(t,k) = \frac{1}{\sqrt{2\pi}}\int dx\; e^{-ikx}\Phi_N(t,x), \tag{2.31}$$

satisfies the algebraic equation

$$\partial_t \hat{\Phi}_N(t,k) + \frac{k^2}{2N}\hat{\Phi}_N(t,k) = 0,$$

whose solution is

$$\hat{\Phi}_N(t,k) = \hat{\Phi}_N(0,k) e^{-\frac{k^2}{2N}t}. \tag{2.32}$$

For the sake of completeness, we shall remind the statement of the convolution theorem:

Theorem 2.2. *For two functions $f, g \in \mathcal{L}^1(\mathbb{R}^d)$ the following expression subsists:*

$$\mathcal{F}\big(f * g\big) = (2\pi)^{d/2}\hat{f}\cdot\hat{g},$$

where $\mathcal{F}$ denotes the Fourier transform.

Thus, in order to write down the solution of the heat equation, it is sufficient to find the function whose Fourier transform is the exponential term in (2.32). It is easy to check that such a function is equal to $\sqrt{(N/t)}e^{-Nx^2/(2t)}$ and, in the convolution, it gives us the Green propagator. Therefore, in the original space, the solution of equation (2.29) is simply

$$\Phi_N(t,x) = \int dy\ G_t(x-y)\Phi_N(0,y),$$

where the Green propagator is given by

$$G_t(x) = \sqrt{\frac{N}{2\pi t}} e^{-N\frac{x^2}{2t}}. \tag{2.33}$$

Recalling the relation between Φ_N and S_N given in (2.28), and the initial condition (2.30), overall we get

$$S_N(t,x) = -\frac{1}{N}\ln\left(\sqrt{\frac{N}{2\pi t}}\int dy\ e^{-N\left(\frac{(x-y)^2}{2t} - \ln 2 - \ln\cosh(y)\right)}\right). \tag{2.34}$$

Since the exponent in equation (2.34) is proportional to the volume, for large N, we can apply now the saddle point method to get

$$\begin{aligned}\alpha(t,x) = -\lim_{N\to\infty} S_N(t,x) = \sup_y\left\{-\frac{(x-y)^2}{2t} + \ln 2 + \ln\cosh(y)\right\} = \\ = -\frac{(x-\hat{y})^2}{2t} + \ln 2 + \ln\cosh(\hat{y}),\end{aligned}$$

where $\hat{y}$ is the maximizer satisfying the condition

$$x = \hat{y} - t\tanh(\hat{y}) = \hat{y} + u(t,x)t = \hat{y} - \omega(m)t.$$

The second equality holds because the Burgers equation becomes inviscid in the thermodynamic limit and, for $x = 0$ (where statistical mechanics is recovered) the above conidition implies $\hat{y}_0 = \omega(m)t$, leading to the well-known Curie-Weiss self-consistency (properly evaluated at $t = \beta$)

$$\omega(m) = \tanh\left(\beta\omega(m)\right), \tag{2.35}$$

when imposing the extremality condition for the statistical pressure. Therefore, we can state the following

Theorem 2.3. *The infinite volume limit of the the Curie-Weiss statistical pressure $\alpha(\beta)$ can be written in terms of the magnetization as the maximal value of*

$$\alpha(\beta) = \sup_{\omega(m)} \Big\{ \ln 2 + \ln\cosh\left(\beta\omega(m)\right) - \frac{\beta}{2}\omega(m)^2 \Big\}.$$

The Curie-Weiss model undergoes a phase transition of the second order from an ergodic (paramagnetic) phase to a ferromagnetic one at $\beta = 1$ ($\beta = 1/J$ in case of non-unitary coupling) and an external field $h = 0$, with the critical exponent equal to 1/2. Let us show this from equation (2.35): if we expand the hyperbolic tangent close to $\omega(m) = 0$, assuming continuity for m (this is justified since the transition is of the second order), we have

$$\omega(m) = \tanh\big(\beta\omega(m)\big) \sim \beta\omega(m) - \frac{\big(\beta\omega(m)\big)^3}{3},$$

from which one gets

$$\omega(m)(1-\beta) + \frac{1}{3}\big(\beta\omega(m)\big)^3 \sim 0.$$

The first solution is trivially $\omega(m) = 0$ (which is also the only solution in the ergodic phase) while the other two solutions can be obtained by solving

$$\omega(m)^2 \sim \frac{(\beta-1)^3}{\beta^3} \sim 3\Big(1 - \frac{1}{\beta}\Big),$$

close to the critical point $\beta = 1$, obtaining

$$\omega(m) \sim (\beta - 1)^{1/2},$$

from which we get the critical exponent $\gamma = 1/2$.

Chapter 3

Complex systems: the Sherrington-Kirkpatrick paradigm

Spin glasses, besides constituting “a challenge for mathematician” [127], 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. [6, 116, 130] and, last but not least, Artificial Intelligence [49, 11].
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 [119] 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 [93]. 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 [25, 27] and Talagrand [126, 125]. Apart from a few exceptions [29, 30, 102], 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 [60]. 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 [3, 5, 8, 24, 19], as we will extensively deepen in this thesis.

3.1 Generalities

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. [93, 62, 45]), 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 $\sigma_i = (\vec{\sigma}_i)_z = \pm 1$:

$$H_N(\boldsymbol{\sigma}|\boldsymbol{J}) = -\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j, \tag{3.1}$$

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,

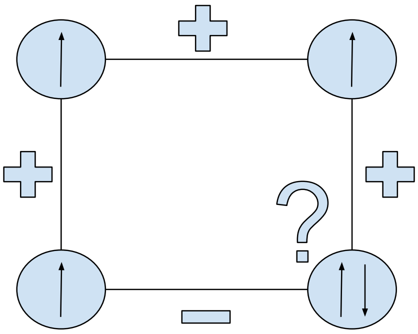

Figure 3.1: **A very simple example of a frustrated system**. The spins tend to be parallel when they interact with a positive coupling and anti-parallel when the interaction is negative. Obviously, not all the conditions can be met simultaneously, meaning that interaction is frustrated.

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. 3.1). We have seen before (see Sec. 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]

3.2 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 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. [93, 62]), whose Hamiltonian is given by the next

Definition 3.1. The mean field spin glass is introduced by the following Sherrington-Kirkpatrick Hamiltonian

$$H_N(\boldsymbol{\sigma}|h;\boldsymbol{J}) = -\frac{1}{\sqrt{N}}\sum_{1\leq i<j\leq N} J_{ij}\sigma_i\sigma_j - h\sum_{1\leq i\leq N}\sigma_i. \tag{3.2}$$

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 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

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

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 energy, entropy and 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

$$\begin{aligned}
\mathbb{E}(H_N(\boldsymbol{\sigma}^{(1)}|\boldsymbol{J})H_N(\boldsymbol{\sigma}^{(2)}|\boldsymbol{J})) &= \frac{1}{N}\sum_{i<j}^{1,N}\sum_{k<l}^{1,N}\mathbb{E}(J_{ij}J_{kl})\sigma_i^{(1)}\sigma_j^{(1)}\sigma_k^{(2)}\sigma_l^{(2)} \\
&= \frac{1}{N}\sum_{1\leq i<j\leq N}\sigma_i^{(1)}\sigma_j^{(1)}\sigma_i^{(2)}\sigma_j^{(2)} \\
&= \frac{N}{2}\Big(\frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)}\Big)^2 - \frac{1}{2}. \qquad (3.3)
\end{aligned}$$

The quantity

$$q_{12} = q(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)}, \qquad (3.4)$$

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

$$\mathbb{E}\left(H_N(\boldsymbol{\sigma}|\boldsymbol{J})\right)^2 = \frac{N}{2} - \frac{1}{2}, \qquad (3.5)$$

showing that the normalization factor is correct.

3.2.1 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 3.2. 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

$$\begin{aligned} Z_N(\beta|h;\boldsymbol{J}) &= \sum_{\boldsymbol{\sigma}} \exp(-\beta H_N(\boldsymbol{\sigma}|h;\boldsymbol{J})), && (3.6)\\ f_N(\beta|h) &= -\frac{1}{\beta N}\mathbb{E}\log Z_N(\beta|h;\boldsymbol{J}), && (3.7)\\ \omega_{\boldsymbol{J}}(A) &= Z_N(\beta,h;\boldsymbol{J})^{-1}\sum_{\boldsymbol{\sigma}} A(\boldsymbol{\sigma})\exp(-\beta H_N(\boldsymbol{\sigma}|h;\boldsymbol{J})), && (3.8) \end{aligned}$$

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

$$\alpha_N(\beta|h) = \frac{1}{N}\mathbb{E}\log Z_N(\beta|h;\boldsymbol{J}) = -\beta f_N(\beta|h), \tag{3.9}$$

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 3.1. 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 come back to these features 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* [101], meaning that its deviations from the quenched value vanish in the thermodynamic limit with probability one.

Definition 3.3. One can also consider the so-called *annealed* free energy

$$f_N^{\mathrm{A}}(\beta|h) = -\frac{1}{\beta N}\log \mathbb{E}Z_N(\beta|h;\boldsymbol{J}), \tag{3.10}$$

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.

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

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

Finally, the annealed free energy per site is

$$f_N^{\mathrm{A}}(\beta|h) = -\frac{1}{\beta}\log 2\cosh(\beta h) - \frac{\beta}{4}\frac{N-1}{N}, \tag{3.11}$$

and in the thermodynamic limit we have the next

Proposition 3.1. *The infinite volume limit of the annealed pressure of the SK model reads as*

$$f^A(\beta|h) = \lim_{N\to\infty} f_N^A(\beta|h) = -\frac{1}{\beta}\log 2\cosh(\beta h) - \frac{\beta}{4}. \tag{3.12}$$

Remark 3.2. 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

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

Remark 3.3. 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

$$s^{\mathrm{A}}(\beta) = \beta^2\partial_\beta f^{\mathrm{A}}(\beta) = \log 2 - \frac{\beta^2}{4}, \tag{3.13}$$

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.

3.2.2 Replicas and overlap

Previously, we vaguely introduced the concept of overlap, as defined in Eq. 3.4, 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

$$\Omega_{\boldsymbol{J}} = \omega_{\boldsymbol{J}}^{(1)} \times \omega_{\boldsymbol{J}}^{(2)} \times ... \times \omega_{\boldsymbol{J}}^{(n)}, \tag{3.14}$$

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* [93]. 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

$$\exp\left(-\beta\left(H_N(\boldsymbol{\sigma}^{(1)}|h;\boldsymbol{J}) + H_N(\boldsymbol{\sigma}^{(2)}|h;\boldsymbol{J}) + ... + H_N(\boldsymbol{\sigma}^{(n)}|h;\boldsymbol{J})\right)\right). \tag{3.15}$$

Definition 3.4. 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

$$\langle A(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}, ..., \boldsymbol{\sigma}^{(n)})\rangle = \mathbb{E}\Omega_{\boldsymbol{J}}(A(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}, ..., \boldsymbol{\sigma}^{(n)})). \tag{3.16}$$

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 [93], 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

$$\mathbb{E}(JA(J)) = \mathbb{E}\Big(\frac{\partial}{\partial J}A(J)\Big), \tag{3.17}$$

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:

$$E \equiv \frac{\langle H_N\rangle}{N} = \frac{1}{N}\mathbb{E}\,\omega_{\boldsymbol{J}}(H_N) = -\frac{\beta}{2}(1 - \langle q_{12}^2\rangle). \tag{3.18}$$

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

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

3.3 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 [60]. However, it was earlier noticed that the disorder fluctuations of the free energy vanish when taking the infinite volume limit: the free energy of the Sherrington-Kirkpatrick model was first proved to be self-averaging by Pastur, Shcherbina and Tirozzi [101], by using martingale techniques. They found that

$$\mathbb{E}\Big(\frac{1}{N}\log Z_N(\beta;J)\Big)^2-\Big(\mathbb{E}\frac{1}{N}\log Z_N(\beta;J)\Big)^2\leq\frac{C}{N}+O\left(\frac{1}{N^2}\right), \tag{3.19}$$

for some constant C. This result was later improved by Guerra [59], who gave a more precise estimate for the upper bound, showing that

$$C\leq\beta^2\frac{q_{12}^2}{2}. \tag{3.20}$$

This does not necessarily implies convergence, since the mean value could oscillate as the system size grows. The property of absence of fluctuations for a physical quantity in the thermodynamic limit is called *self-averaging.* This property is usually expected, in ordinary statistical mechanics, for intensive quantities (such as magnetization or free energy per site) with respect to thermal fluctuations and away from phase transition points. In spin-glass systems, there is a somewhat different scenario [93], and one expects some quantities (such as free and internal energy) to be self-averaging, and others, in particular the overlap between the configurations of two replicas, to fluctuate even in the thermodynamic limit at low temperature. As we have seen in the previous chapter, this is an indication of the occurrence of Replica Symmetry Breaking.

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

$$\begin{aligned}Z_N(\beta,t) \;&=\; \sum_{\boldsymbol{\sigma}}\exp\beta\Big(\sqrt{\frac{t}{N}}\sum_{1\leq i<j\leq N}J_{ij}\sigma_i\sigma_j+\sqrt{\frac{1-t}{N_1}}\sum_{1\leq i<j\leq N_1}J'_{ij}\sigma_i\sigma_j\\ &\quad+\sqrt{\frac{1-t}{N_2}}\sum_{N_1\leq i<j\leq N}J''_{ij}\sigma_i\sigma_j\Big),\end{aligned} \tag{3.21}$$

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 $t = 0$:

$$\begin{aligned} Z_N(\beta, 1) &= Z_N(\beta), &(3.22)\\ Z_N(\beta, 0) &= Z_{N_1}(\beta) Z_{N_2}(\beta). &(3.23)\end{aligned}$$

Consequently, the free energies are realized as

$$\begin{aligned} \mathbb{E} \log Z_N(\beta, 1) &= -N\beta f_N(\beta), &(3.24)\\ \mathbb{E} \log Z_N(\beta, 0) &= -N_1\beta f_{N_1}(\beta) - N_2\beta f_{N_2}(\beta). &(3.25)\end{aligned}$$

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

$$-\frac{d}{dt}\frac{1}{N\beta}\mathbb{E}\log Z_N(\beta, t) = -\frac{1}{2N}\mathbb{E}\Big(\frac{1}{\sqrt{tN}}\sum_{1\le i<j\le N} J_{ij}\omega_t(\sigma_i\sigma_j) \tag{3.26}$$
$$-\frac{1}{\sqrt{(1-t)N_1}}\sum_{1\le i<j\le N_1} J'_{ij}\omega_t(\sigma_i\sigma_j) - \frac{1}{\sqrt{(1-t)N_2}}\sum_{N_1\le i<j\le N} J''_{ij}\omega_t(\sigma_i\sigma_j)\Big),$$

where $\omega_t(\cdot)$ is the Gibbs average corresponding to the auxiliary partition function (3.21). Using again the integration by parts formula on the previous expression, we have

$$\begin{aligned} -\frac{d}{dt}\frac{1}{N\beta}\mathbb{E}\log Z_N(\beta, t) &= -\frac{\beta}{4N^2}\sum_{1\le i<j\le N}\mathbb{E}\left(1-\omega_t^2(\sigma_i\sigma_j)\right) &(3.27)\\ &\quad +\frac{\beta}{4NN_1}\sum_{1\le i<j\le N_1}\mathbb{E}\left(1-\omega_t^2(\sigma_i\sigma_j)\right)\\ &\quad +\frac{\beta}{4NN_2}\sum_{N_1\le i<j\le N}\mathbb{E}\left(1-\omega_t^2(\sigma_i\sigma_j)\right)\\ &= \frac{\beta}{4}\langle q_{12}^2 - \frac{N_1}{N}(q_{12}^{(1)})^2 - \frac{N_2}{N}(q_{12}^{(2)})^2\rangle_t, \end{aligned}$$

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

$$\begin{aligned} q_{12}^{(1)} &= \frac{1}{N_1}\sum_{1\le i\le N_1}\sigma_i^1\sigma_i^2, &(3.28)\\ q_{12}^{(2)} &= \frac{1}{N_2}\sum_{N_1\le i\le N}\sigma_i^1\sigma_i^2, &(3.29)\end{aligned}$$

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

$$q_{12} = \frac{N_1}{N} q_{12}^{(1)} + \frac{N_2}{N} q_{12}^{(2)}, \tag{3.30}$$

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

$$\langle q_{12}^2 - \frac{N_1}{N} (q_{12}^{(1)})^2 - \frac{N_2}{N} (q_{12}^{(2)})^2 \rangle_t \leq 0. \tag{3.31}$$

Therefore, we can state as a preliminary result:

Lemma 3.1. *The quenched average of the logarithm of the interpolating partition function, defined by (3.21), increases in t, i.e.*

$$-\frac{d}{dt} \frac{1}{N\beta} \mathbb{E} \log Z_N(\beta, t) \leq 0. \tag{3.32}$$

Moreover, after integrating over t and recalling the boundary conditions (3.24, 3.25), we get the first main result

Theorem 3.2. *The free energy for the SK model is subadditive:*

$$N f_N(\beta) \leq N_1 f_{N_1}(\beta) + N_2 f_{N_2}(\beta). \tag{3.33}$$

It is interesting to compare this result with the corresponding (2.12) 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 3.3. *The infinite volume limit for $f_N(\beta)$ exists and equals its infimum:*

$$f(\beta) \equiv \lim_{N \to \infty} f_N(\beta) = \inf_N f_N(\beta). \tag{3.34}$$

Remark 3.4. 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 (3.27) and (3.31) are simply replaced by the overlap to the power p, and the (3.32) 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.

3.4 The replica trick and Parisi theory

Parisi Theory 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 but simply remind to excellent textbooks [81, 62, 93].

3.4.1 The Replica Trick

The natural starting point to examine Parisi theory are the basic concepts of spontaneous symmetry breaking and phase coexistence in statistical mechanics [78, 112, 48]. 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

$$f_\Lambda(\beta) = -\frac{1}{|\Lambda|\beta} \log Z_\Lambda(\beta), \tag{3.35}$$

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

$$\lim_{\Lambda \to \mathbb{Z}^d} \langle \sigma_{i_1} ... \sigma_{i_n} \rangle_\Lambda, \tag{3.36}$$

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 λ_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):

$$\langle \sigma_{i_1}...\sigma_{i_n}\sigma_{j_1}...\sigma_{j_m}\rangle \to \langle \sigma_{i_1}...\sigma_{i_n}\rangle\langle \sigma_{j_1}...\sigma_{j_m}\rangle, \tag{3.37}$$

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

$$f_N(\beta) = -\frac{1}{\beta N} \lim_{n\to 0} \frac{\mathbb{E}Z^n - 1}{n}. \tag{3.38}$$

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

$$\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = \mathbb{E}\sum_{\boldsymbol{\sigma}^{(1)}} ... \sum_{\boldsymbol{\sigma}^{(n)}} \exp\Big(-\beta\sum_{a=1}^n H_N(\boldsymbol{\sigma}^{(a)}|h;\boldsymbol{J})\Big). \tag{3.39}$$

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

$$\begin{aligned}\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = \exp\Big(\frac{\beta^2 n(N-n)}{4}\Big)\\ \sum_{\boldsymbol{\sigma}^{(1)}...\boldsymbol{\sigma}^{(n)}} \exp\Big(\frac{\beta^2}{2N}\sum_{1\le a<b\le n}\Big(\sum_i \sigma_i^{(a)}\sigma_i^{(b)}\Big)^2 + \beta h\sum_{a=1}^n\sum_i \sigma_i^{(a)}\Big),\end{aligned} \tag{3.40}$$

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 (3.40) as

$$\begin{aligned}\sum_{\boldsymbol{\sigma}^{(1)}...\boldsymbol{\sigma}^{(n)}} \int\prod_{a<b}\Big(\sqrt{\frac{\beta^2 N}{2\pi}}dQ_{ab}\Big)\exp\Big(-\frac{\beta^2 N}{2}\sum_{a<b}Q_{ab}^2\\ +\beta^2\sum_{a<b}\Big(\sum_i \sigma_i^{(a)}\sigma_i^{(b)}\Big)Q_{ab} + \beta h\sum_a\sum_i \sigma_i^{(a)}\Big).\end{aligned} \tag{3.41}$$

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

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

Then, equation (3.40) becomes

$$\mathbb{E}Z_N^n(\beta|h;\boldsymbol{J}) = \int\prod_{a<b}\Big(\sqrt{\frac{\beta^2 N}{2\pi}}dQ_{ab}\Big)\exp(-NA[\boldsymbol{Q}]), \tag{3.42}$$

$$\begin{aligned}A[\boldsymbol{Q}] &= \frac{\beta^2}{2}\sum_{a<b}Q_{ab}^2-\log\sum_{\{\boldsymbol{s}\}}\exp\Big(\beta^2\sum_{a<b}Q_{ab}s_a s_b+\beta h\sum_a s_a\Big)\\ &\quad-\frac{\beta^2 n(N-n)}{4N},\end{aligned} \tag{3.43}$$

with the functional $A[\boldsymbol{Q}]$ depending on $\boldsymbol{Q}$, n, β and h. Since the exponent in the integrand of (3.42) 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

$$f(\beta,h)=\lim_{n\to 0}\frac{1}{\beta n}A[\boldsymbol{Q}_{sp}]. \tag{3.44}$$

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

$$Q_{ab}=\frac{\sum_{\{\boldsymbol{s}\}}s_a s_b\exp\left(\beta^2\sum_{a<b}Q_{ab}s_a s_b+\beta h\sum_a s_a\right)}{\sum_{\{\boldsymbol{s}\}}\exp\left(\beta^2\sum_{a<b}Q_{ab}s_a s_b+\beta h\sum_a s_a\right)} \tag{3.45}$$

In the limit $n\to 0$, it can be shown [93] 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 (3.45) can be written as

$$\lim_{n\to 0}Q_{ab}=\langle q_{ab}\rangle. \tag{3.46}$$

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 $\omega_{\boldsymbol{J}}^2(\sigma_i)$ 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 $\langle \sigma_i^{(a)} \sigma_i^{(b)} \rangle = 0$.

3.4.2 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. (3.42) then reduces to an ordinary integral over the real variable q, and the quenched free energy is easily computed as

$$-\beta f_{RS}(\beta, h) = \log 2 + \int_{-\infty}^{+\infty} d\mu(z) \log\cosh(\beta\sqrt{q}z + \beta h) + \frac{\beta^2}{4}(1-q)^2, \quad (3.47)$$

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

$$q = \int_{-\infty}^{+\infty} d\mu(z) \tanh(\beta\sqrt{q}z + \beta h). \quad (3.48)$$

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 [93] 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 [119]). The free energy (3.47) 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 [78], 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 $n \to 0$ [93, 62]. 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 (3.4) 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:

$$\mathcal{P}(q) = \frac{\delta(q - m^2(\beta)) + \delta(q + m^2(\beta))}{2}. \tag{3.49}$$

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

$$\mathcal{P}(q) = \delta(q). \tag{3.50}$$

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 $\mathcal{P}(q)$:

$$\langle q_{12}\rangle = \int dq\mathcal{P}(q)q. \tag{3.51}$$

This equation, combined with 3.46, 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 [93].

3.4.3 Broken Replica Symmetry *Ansatz*

Before to proceed, we would like to make the following

Remark 3.5. We repeat that, while hereafter we quickly revise the Parisi description of spontaneous replica symmetry breaking, times are not ripe for such a level of resolution in AI and, as a matter of fact, the bulk of results in research on neural networks and machine learning is pursued at the replica symmetric level of description, as also all our ones will be. Hence, the reader can skip this Section without eliminating any essential knowledge required for the second part of this thesis, devoted to AI applications of statistical mechanics of disordered systems.

We have just seen that the replica symmetric solution is not adequate because it violates thermodynamic stability conditions. The correct way to construct a matrix $\boldsymbol{Q}$ breaking replica symmetry has been discovered by Parisi. Operatively, the procedure consists in dividing the n replicas in n/m groups of m, where m is obviously a submultiple of n. Then, one takes $Q_{ab} = q_2$ if a and b belong to the same group (with $a \neq b$), and $Q_{ab} = q_1$ if they belong to different replicas. For example, if $n = 4$ we can have a matrix with the form

$$\left(\begin{array}{cc|cc} 0 & q_2 & q_1 & q_1 \\ q_2 & 0 & q_1 & q_1 \\ \hline q_1 & q_1 & 0 & q_2 \\ q_1 & q_1 & q_2 & 0 \end{array}\right).$$

With such an *Ansatz*, the overlap distribution is given by [93]:

$$\mathcal{P}(q) = (1-m)\delta(q-q_2) + m\delta(q-q_1), \tag{3.52}$$

which is not negative only if $0 \le m \le 1$. The free energy corresponding to this first step of broken replica symmetry (1-RSB) is given by

$$\begin{aligned} -\beta f_{1RSB}(\beta|h) &= \log 2 + \frac{\beta^2}{4}\left[(1-m)q_2^2 + mq_1^2 + 1 - 2q_2\right] \\ &\quad + \frac{1}{m}\int d\mu(u)\log\int d\mu(v)\cosh^m\Theta, \quad (3.53)\\ \Theta &= \beta\left(\sqrt{q_1}u + \sqrt{q_2-q_1}v + h\right), \quad (3.54)\end{aligned}$$

where the parameters q_1 and q_2 are the solutions of the self-consistence (saddle point) equations

$$\begin{aligned} q_1 &= \int d\mu(u)\left(\frac{\int d\mu(v)\cosh^m\Theta\tanh\Theta}{\int d\mu(v)\cosh^m\Theta}\right)^2, \quad (3.55)\\ q_2 &= \int d\mu(u)\frac{\int d\mu(v)\cosh^m\Theta\tanh^2\Theta}{\int d\mu(v)\cosh^m\Theta}. \quad (3.56)\end{aligned}$$

We refer to [93] for a detailed treatment of the interpretation and for the physical consequences of the RSB *Ansatz*. This solution turns out to be better than the RS one below the critical temperature, but it is not yet the right one. However, one can apply this procedure iteratively. Indeed, in a second step the off-diagonal blocks are left untouched, while the diagonal blocks are further divided into m_1/m_2 blocks, with the matrix elements assuming the value $Q_{ab} = q_3$ for $a \neq b$ inside the same block and $Q_{ab} = q_2$ otherwise. In order to find the proper free energy, one has to apply this procedure an infinite number of times (full-RSB or ∞-RSB [93], [125, 126]). Since we have seen how to compute the free energy and the distribution $P(q)$ in the setting of the replica method, in the following we try to continue presenting the main results of the theory but using a slightly different language (referring for instance to [58] for a presentation along these lines). Let us introduce the convex space $\mathcal{X}$ of the functional order parameters x, as non-decreasing functions of the auxiliary variable q, with both x and q taking values on the real interval $[0,1]$, i.e.

$$\mathcal{X} \ni x : [0,1] \ni q \to x(q) \in [0,1]. \tag{3.57}$$

Notice that we call x the non-decreasing function, and $x(q)$ its values. A metric on $\mathcal{X}$ is introduced through the $L^1([0,1], dq)$ norm, where dq is the

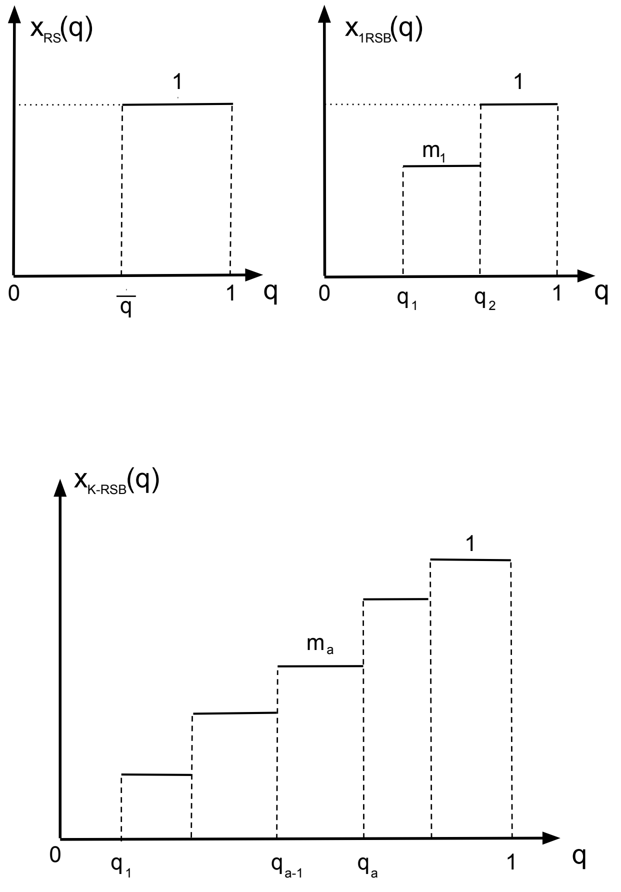

Figure 3.2: **Schematical representation of replica symmetry breaking.** The upper plot corresponds to the RS *Ansatz* (left) and 1RSB (right). The plot on the bottom corresponds to a generic K-step RSB.

Lebesgue measure. We will consider piecewise constant functional order parameter, since every regular function in this interval can be approximated with arbitrary precision in this way. Then, given an integer number K of intervals, we have two sequences $q_0, q_1, ..., q_K$ and $m_1, m_2, ..., m_K$ satisfying

$$\begin{aligned} 0 &= q_0 \leq q_1 \leq ... \leq q_{K-1} \leq q_K = 1, && (3.58) \\ 0 &\leq m_1 \leq m_2 \leq ...m_K \leq 1, && (3.59) \end{aligned}$$

and such that

$$x(q) = \begin{cases} m_1 & \text{for} \quad 0 = q_0 \leq q < q_1, \\ m_2 & \text{for} \quad 0 = q_1 \leq q < q_2, \\ ... & \\ m_K & \text{for} \quad 0 = q_{K-1} \leq q < q_K. \end{cases} \tag{3.60}$$

as it is shown in Figure 3.2. The choice of a piecewise constant functional order parameter corresponds to consider replica symmetry breaking to a finite number K of steps in the frame of Parisi theory. For instance, the replica symmetric case is reconstructed by taking

$$K = 2, \ q_1 = q, \ m_1 = 0, \ m_2 = 1, \tag{3.61}$$

while for the 1-RSB distribution one has to take $K = 3$, and so on (see Figure 3.2). Let us now introduce the function $f = f(q, y; x, \beta)$, depending on the variables $q \in [0, 1]$, $y \in \mathbb{R}$, on the functional order parameter x and on the inverse temperature β. This function should not be confused with the free energy per site in the thermodynamic limit $f(\beta|h)$. The formed is indeed defined as the solution of the nonlinear antiparabolic equation

$$\frac{\partial}{\partial q} f(q, y) + \frac{1}{2} \frac{\partial^2}{\partial y^2} f(q, y) + \frac{1}{2} x(q) \left(\frac{\partial}{\partial y} f(q, y) \right)^2 = 0, \tag{3.62}$$

with final condition

$$f(1, y) = \log \cosh(\beta y). \tag{3.63}$$

For the sake of clearness, here we only stressed the dependence of f on q and y. This equation, if we consider q corresponding to the time and y to the position in space, is formally equivalent to a diffusive heat equation when $x(q) \equiv 0$, while it is equivalent to a Hamilton-Jacobi equation with varying mass $(x(q))^{-1}$ if the second-order derivative in y vanishes identically. Let us consider the solution is some simple cases.

- $x \equiv 0$

The solution can be easily obtained starting from Eq. (3.63), adding to y a gaussian variable z weighted with the root $\sqrt{1-q}$, which vanishes at the end of the interval, and integrating over z:

$$\begin{aligned} f(q, y) &= \int d\mu(z) \log \cosh \beta \left(y + z\sqrt{1-q} \right), \\ d\mu(z) &\equiv e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}}. \end{aligned} \tag{3.64}$$

- $x \equiv 1$

In this case, taking $f(q,y) = \log\cosh\beta y + a(q)$, with $a(1) = 0$, and solving Eq. (3.62) respect to a, we find the solution

$$f(q,y) = \log\cosh\beta y + \frac{1}{2}\beta^2(1-q). \tag{3.65}$$

- $x \equiv x_{\bar{q}} = \begin{cases} 0 & \text{for} \quad 0 \leq q < \bar{q}, \\ 1 & \text{for} \quad \bar{q} \leq q \leq 1 \end{cases}$

Starting from Eq. (3.65), which is valid in the interval $\bar{q} \leq q \leq 1$, we get the final condition for $f(q,y)$ in this interval:

$$f(\bar{q},y) = \log\cosh\beta y + \frac{1}{2}\beta^2(1-\bar{q}). \tag{3.66}$$

The solution for $q \in [0,\bar{q}]$ can be found, similarly to the case of $x \equiv 0$, starting from the final condition in $q = \bar{q}$ and then adding to y a properly weighted variable z, over which a Gaussian integration is performed. This leads to

$$f(q,y) = \int d\mu(z) \log\cosh\beta\left(y + z\sqrt{\bar{q}-q}\right) + \frac{1}{2}\beta^2(1-\bar{q}). \tag{3.67}$$

In the general case, for a piecewise constant x with $x(q) = m_a$ for $q_{a-1} \leq q < q_a$ ($m_a > 0$), it is convenient to introduce the auxiliary function $g_a(q,y) = \exp m_a f(q,y)$ satisfying the equation

$$\frac{\partial}{\partial q} g_a(q,y) + \frac{1}{2}\frac{\partial^2}{\partial y^2} g_a(q,y) = 0. \tag{3.68}$$

The final condition (3.63) for g_K in the last interval is $g_K(q_K,y) = \cosh^{m_K}\beta y$, and - as we have jsut seen for $f(q,y)$ in the case $x \equiv 0$ - the solution in the interval $[q_{K-2}, q_{K-1}]$ is obtained from the final condition, adding to y a properly weighted gaussian variable, and then integrating it

$$g_{K-1}(q,y) = \int d\mu(z_K) g_K\left(q_K, y + z_K\sqrt{q_K - q}\right), \tag{3.69}$$

whence for $q \in [q_{K-2}, q_{K-1}]$ we have the solution

$$\exp f(q,y) = \left(\int d\mu(z_K) \exp\left(m_K f\left(q_K, y + z_K\sqrt{q_K - q}\right)\right)\right)^{\frac{1}{m_K}}. \tag{3.70}$$

The general solution for all previous intervals is then found by iterating such an algorithm. Notice that, if $m_1 = 0$, the solution in the corresponding interval can be computed, as we saw for $x = x_{\bar{q}}$, starting from the one valid for $q \in [q_1, q_2]$ and integrating it

$$f(q, y) = \int d\mu(z_1) f(q_1, y + z_K\sqrt{q_K - q_{K-1}} + ... + z_1\sqrt{q_1 - q}), \qquad (3.71)$$

which is equivalently obtained from the general formula for a finite m_a, in the limit $m_1 \to 0$. Since any functional order parameter can be approximated (in the L^1 norm) with piecewise constant x, and since it can be shown that f is pointwise continuous with respect to x, we can handle mostly with piecewise constant order parameters. This important result is stated in the following

Theorem 3.4. *The function f is monotone in x in the sense that*

$$x(q) \leq \bar{x}(q) \;\; \forall q \in [0, 1] \quad \Rightarrow \quad f(q, y; x, \beta) \leq f(q, y; \bar{x}, \beta)$$

for any $q \in [0, 1]$, $y \in R$. Moreover, f is pointwise continuous in the $L^1([0, 1], dq)$ norm. In fact, for generic x and $\bar{x}$, we have

$$|f(q, y; x, \beta) - f(q, y; \bar{x}, \beta)| \leq \frac{\beta^2}{2} \int_q^1 |x(q') - \bar{x}(q')| dq'.$$

Once the function f is introduced, we are now ready for the following important definitions.

Definition 3.5. The trial auxiliary function $\bar{\alpha}$, depending on the functional order parameter x, is defined as

$$\bar{\alpha}(\beta, h; x) \equiv \log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2} \int_0^1 q x(q) dq. \qquad (3.72)$$

Let us observe that, in this definition, the function f appears evaluated at $q = 0$, and $y = h$, where h is the value of the external magnetic field.

Definition 3.6. The Parisi spontaneously broken replica symmetry solution is defined by

$$\bar{\alpha}(\beta, h) \equiv \inf_x \bar{\alpha}(\beta, h; x), \qquad (3.73)$$

where the infimum is taken with respect to all functional order parameters x.

The main prediction of Parisi theory is that, for the Sherrington-Kirkpatrick model, this infimum is related to the free energy in the thermodynamic limit as

$$-\beta f(\beta|h) = \lim_{N\to\infty} N^{-1}\mathbb{E}\log Z_N(\beta,h;J) = \bar{\alpha}(\beta,h). \tag{3.74}$$

Moreover, the functional parameter x realizing the infimum in (3.73) was interpreted by Parisi as the cumulative distribution function associated to the overlap probability distribution $\mathcal{P}(q)$, i.e.

$$x(q) = \int_0^q \mathcal{P}(q')dq'. \tag{3.75}$$

In other words, if replica symmetry holds, then $\mathcal{P}(q) = \delta(q-\bar{q})$ and the optimal order parameter is just a step function, as in Figure 3.2. The 2-step choice of $x(q)$ corresponds to the first level of broken replica symmetry, and so on. When discussing the replica symmetric solution, we already noticed that the RS free energy is maximized by the proper choice of the parameter q. Here, we stress again that, according to (3.73), the trial functional $-\beta\bar{\alpha}(\beta,h;x)$ has to be maximized over the space of functional order parameter in order to obtain the infinite volume free energy $f(\beta|h)$. The usual variational principle of statistical mechanics, which follows from the second principle of thermodynamics, states that the free energy can be obtained through minimization of a suitable free energy functional, on all possible trial states. This means that, for any order parameter x different from the optimal one, $-\beta\bar{\alpha}(\beta,h;x)$ *cannot* be interpreted as the free energy associated to some trial state. However, it has been shown that the the value given in the Parisi *Ansatz* is a lower bound for the quenched average of the free energy, uniformly in the size of the system [58]. Furthermore, in the same paper a sum rule for the difference between the Parisi formula (3.73) and the real free energy was given. Afterwards, this difference has been showed to be vanishing in the thermodynamic limit [126].

3.5 Guerra's interpolating scheme

The idea behind the method precisely follows the same reasoning of the CW case (exploited in Section 2.4), despite obvious mathematical differences: to make them clear, we directly introduce the next

Definition 3.7. The interpolating partition function and the (thermodynamic limit of) the quenched free energy in the Guerra's scheme read as

$$Z_N(\beta,t) = \sum_{\boldsymbol{\sigma}} \exp\left\{\sqrt{t}\frac{\beta}{\sqrt{N}}\sum_{i<j} J_{ij}\sigma_i\sigma_j + A\sqrt{1-t}\sum_i z_i\sigma_i\right\}, \tag{3.76}$$

$$f_N(\beta,t) = -\frac{1}{\beta N}\mathbb{E}\log Z_N(\beta,t). \tag{3.77}$$

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

$$f(\beta) \equiv f(\beta,t=1) = f(\beta,t=0) + \int_0^1 ds\ \Big[\partial_t f(\beta,t)\Big]_{t=s}. \tag{3.78}$$

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 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.

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

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

$$\frac{df(\beta,t)}{dt} = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\Big(\frac{1}{2\sqrt{t}}\frac{\beta}{\sqrt{N}}\sum_{i<j}J_{ij}\omega_t(\sigma_i\sigma_j) - \frac{A}{2\sqrt{1-t}}\sum_i z_i\omega_t(\sigma_i)\Big). \tag{3.79}$$

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

$$\frac{df(\beta,t)}{dt} = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\Big(\frac{\beta^2}{4N}\sum_{ij}(1-\omega_t(\sigma_i\sigma_j)^2) - \frac{A^2}{2}\sum_i(1-\omega_t(\sigma_i)^2)\Big). \tag{3.80}$$

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

$$\sum_i \mathbb{E}\omega_t(\sigma_i)^2 = \sum_i \mathbb{E}\,\omega_t^{(1)}\times\omega_t^{(2)}(\sigma_i^{(1)}\sigma_i^{(2)}) = N\langle q_{12}\rangle_t, \tag{3.81}$$

$$\sum_i \mathbb{E}\omega_t(\sigma_i\sigma_j)^2 = \sum_i \mathbb{E}\,\omega_t^{(1)}\times\omega_t^{(2)}(\sigma_i^{(1)}\sigma_i^{(2)}\sigma_j^{(1)}\sigma_j^{(2)}) = N^2\langle q_{12}^2\rangle_t. \tag{3.82}$$

Therefore, the derivative of the interpolating free energy is

$$\frac{df(\beta,t)}{dt} = -\frac{\beta}{4}\lim_{N\to\infty}\mathbb{E}\Big(1-\langle q_{12}^2\rangle_t - \frac{2A^2}{\beta^2}(1-\langle q_{12}\rangle_t)\Big). \tag{3.83}$$

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

$$\frac{df(\beta,t)}{dt} = \frac{\beta}{4}\lim_{N\to\infty}\mathbb{E}\Big(\langle(q_{12}-q)^2\rangle_t - (1-q)^2\Big). \tag{3.84}$$

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

$$\frac{df(\beta,t)}{dt} = -\frac{\beta}{4}(q-1)^2. \tag{3.85}$$

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

$$\begin{aligned} f(\beta,0) &= -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big(A\sum_i z_i\sigma_i\Big) = \\ &= -\lim_{N\to\infty}\frac{1}{\beta N}\sum_i \mathbb{E}\log 2\cosh(Az_i). \end{aligned} \tag{3.86}$$

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

$$f(\beta, 0) = -\frac{1}{\beta}\mathbb{E}\log 2\cosh(\beta\sqrt{q}z). \tag{3.87}$$

By putting everything together according to the sum rule (3.78) and making the Gaussian integration explicit, we get the next

Theorem 3.5. *The explicit expression for the SK pressure in terms of the two replica overlap, in the thermodynamic limit and under the replica symmetric assumption, reads as*

$$f_{RS}(\beta) = -\frac{1}{\beta}\int_{-\infty}^{+\infty} d\mu(z)\log 2\cosh(\beta\sqrt{q}z) - \frac{\beta}{4}(1-q)^2. \tag{3.88}$$

The latter equation precisely reproduce the replica trick prediction (3.47) with vanishing external field $h = 0$.

3.6 The Hamilton-Jabobi formalism

In this Section, we will use the Hamilton-Jacobi framework used in section 2.5 in order to the Sherrington-Kirkpatrick mean field spin glass model. Again, we are interested in an explicit expression for the quenched free energy $f(\beta)$ (or better the pressure $\alpha(\beta)$) in the thermodynamic limit.

Mirroring Section 2.5, we introduce two fictitious spacetime coordinates t and x, so that we can make the following

Definition 3.8. The Guerra interpolating function in the Hamilton-Jacobi approach to the mean field spin-glass model is

$$\alpha_N(t, x) = \frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\left\{\sqrt{\frac{t}{N}}\sum_{i<j}J_{ij}\sigma_i\sigma_j + \sqrt{x}\sum_{i=1}^{N}J_i^1\sigma_i\right\}, \tag{3.89}$$

where J_i^1, $\forall i = 1, \dots, N$, are independently and identically distributed unitary Gaussian random variables.

The pressure (in the vanishing external field case) is recovered whenever evaluating $\alpha_N(t, x)$ at $t = \beta^2$, $x = 0$. However, w.r.t. the CW model, for the SK the pressure $\alpha_N(\beta, h)$ is directly connected to the action $S_N(t, x)$ of a Hamilton-Jacobi equation, but it is not the action itself. Indeed, the action is obtained by performing a linear transformation in the $\{t, x\}$ plane on the pressure. This fact is formalized in the following

Definition 3.9. The Guerra action $S_N(t,x)$ for the SK model is

$$S_N(t,x) = 2\alpha_N(t,x) - x - \frac{t}{2}. \tag{3.90}$$

By direct computations, we can see that the following relations hold:

$$\begin{aligned} \partial_t S_N(t,x) &= -\frac{1}{2}\langle q_{12}^2\rangle_{t,x}, \\ \partial_x S_N(t,x) &= -\langle q_{12}\rangle_{t,x}. \end{aligned} \tag{3.91}$$

To proceed, we need to introduce the potentials

$$\begin{aligned} V_0(t,x) &= \frac{1}{2}\left(\langle q_{12}^2\rangle - \langle q_{12}\rangle^2\right), \\ V_1(t,x) &= -\frac{1}{2}\left(\langle q_{12}^2\rangle - 4\langle q_{12}q_{23}\rangle + 3\langle q_{12}q_{34}\rangle\right). \end{aligned}$$

By direct construction, it is straightforward to check that the following proposition holds:

Proposition 3.2. *The Guerra action for the SK model obeys the following Hamilton-Jacobi PDE:*

$$\frac{\partial S_N(t,x)}{\partial t} + \frac{1}{2}\left(\frac{\partial S_N(t,x)}{\partial x}\right)^2 - \frac{1}{2}\left(\langle q_{12}^2\rangle - \langle q_{12}\rangle^2\right) \equiv -V_0(t,x). \tag{3.92}$$

If we add a vanishing (in the thermodynamic limit) potential, containing the second derivative of $S_N(t,x)$, such as

$$\lim_{N\to\infty} \frac{1}{2N}\frac{\partial^2 S_N(t,x)}{\partial x^2} \equiv V_1(t,x) = 0,$$

and assuming the replica symmetric scheme, where $\lim_{N\to\infty}\left(\langle q_{12}^2\rangle - \langle q_{12}\rangle^2\right) = 0$, we can easily check that the action $S_N(t,x)$ satisfies the differential equation

$$\lim_{N\to\infty}\left(\partial_t S_N(t,x) + \frac{1}{2}\left(\partial_x S_N(t,x)\right)^2 - \frac{1}{2N}\partial^2_{xx} S_N(t,x)\right) = 0.$$

A direct comparison shows that this differential equation is the same as the CW case, so that we can solve it easily with the usual Cole-Hopf transform. A reminder to Parisi theory is in order here: as discussed in detail in [19], since the overlap is not self-averaging in the true solution of the SK model [93], we force $V_0(t,x)$ to be zero in order to get straightforwardly the replica-symmetric solution. On the other hand, $V_1(t,x)$ is always zero in the thermodynamic limit (and of course reduces to an elementary identity once read in the RS framework [56]). We stress that it is not strictly necessary to solve this problem where $V_0(t,x)$ and $V_1(t,x)$ are pasted in the same equation, since we could split the standard Hamilton-Jacobi equation for the Guerra action from the constraint $\frac{1}{2N}\frac{\partial^2 S_N(t,x)}{\partial^2 x}=0$. However, such a "compact procedure" allows to obtain the RS free energy solving a Fourier problem (with all its related know-now) for its Cole-Hopf transform.

To compute explicitly $V_1(t,x)$, it is convenient to introduce the x-streaming relative to a generic observable F which depends on s replicas as [59]

$$\partial_x\langle F_s\rangle = N\langle F\Big(\sum_{ab}^{s} q_{ab} - s\sum_{a}^{s} q_{as+1} + \frac{s(s+1)}{2}q_{s+1,s+2}\Big)\rangle.$$

Hence, remembering from (3.91) that $\partial_x S_N(x,t)=-\langle q_{12}\rangle$, we get

$$-\lim_{N\to\infty}\frac{1}{N}\partial^2_{xx}S_N(t,x) = \lim_{N\to\infty}\Big(\langle q_{12}^2\rangle - 4\langle q_{12}q_{23}\rangle + 3\langle q_{12}q_{34}\rangle\Big) = 0. \tag{3.93}$$

As we did for the CW model, we can solve the Burgers-like equation for the action

$$\partial_t S_N(t,x) + \frac{1}{2}\big(\partial_x S_N(t,x)\big)^2 - \frac{1}{2N}\partial^2_{xx}S_N(t,x) = 0,$$

mapping the latter into a Fourier equation via the Cole-Hopf transform, namely

$$\Phi_N(t,x) = e^{-NS_N(t,x)}.$$

Of course, the Cole-Hopf transform of Guerra action satisfies the heat equation

$$\frac{\partial\Phi_N(t,x)}{\partial t} - \frac{1}{2N}\frac{\partial^2\Phi_N(t,x)}{\partial x^2} = 0. \tag{3.94}$$

Denoting with $\hat{\Phi}_N(t,k)$ the Fourier transform of $\Phi_N(t,x)$, defined by equation (2.31), as in the CW case we have the algebraic equation

$$\partial_t\hat{\Phi}_N(t,k) + \frac{k^2}{2N}\hat{\Phi}_N(t,k) = 0.$$

Using $\Phi_0(k)$ to denote the Cauchy initial condition, we arrive at the solution in the Fourier space

$$\hat{\Phi}_N(t,k) = \hat{\Phi}_0(k)e^{-\frac{k^2}{2N}t}. \tag{3.95}$$

The solution in the original space is obtained by simple application of the Convolution Theorem:

$$\Phi_N(t,x) = \int dy\ G_t(x-y)\Phi_0(y) = \sqrt{\frac{N}{2\pi t}}\int dy\ e^{-N\frac{(x-y)^2}{2t}}\Phi_0(y), \tag{3.96}$$

where $G_t(x-y)$ is the Green propagator (defined in (2.33)) and

$$\Phi_0(y) = e^{-NS_0(y)}.$$

From the definition of the interpolating function (3.89) and from the definition (3.90), we can directly get the expression for $S_0(y)$, which reads

$$S_0(y) = 2\ln 2 + 2\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \ln\cosh(\sqrt{y}z) - y.$$

By direct substitution, we easily obtain

$$\begin{aligned} S_N(t,x) = -\frac{1}{N}\ln\Phi_N(t,x) = -\frac{1}{N}\ln\sqrt{\frac{N}{2\pi t}}\int dy\ \exp\Big\{-N\Big(\frac{(x-y)^2}{2t}+ \\ + 2\ln 2 + 2\int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\ln\cosh(z\sqrt{y}) - y\Big)\Big\}. \end{aligned}$$

Once computed in the thermodynamic limit $N \to \infty$ by means of the saddle point method, this expression reads

$$\begin{aligned} S(t,x) &= \inf_y \Big\{\frac{(x-y)^2}{2t} + S_0(y)\Big\} = \\ &= \inf_y \Big\{\frac{(x-y)^2}{2t} + 2\ln 2 + 2\int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\ln\cosh(z\sqrt{y}) - y\Big\}. \end{aligned} \tag{3.97}$$

We now denote by $\hat{y}(t,x)$ the location where the infimum in (3.97) is achieved. As we are going to show, its inverse $x(t,y)$ is the location at time t of the fictitious particle initially at y. To do this, we observe that the previous maximizing condition can also be expressed as:

$$S(t,x) - \frac{x^2}{2t} = -\sup_y\Big\{\phi_0(y) + \frac{xy}{t}\Big\}, \tag{3.98}$$

where $\phi_0(y) = -y^2/2t - S_0(y)$. Hence, the solution of the Burgers-like equation can be expressed again in terms of a Legendre transform of $\phi_0(y)$. From the extremal condition, we get

$$x = \hat{y} - t\int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}\ e^{-z^2/2}\tanh^2(\sqrt{\hat{y}}z) = \hat{y} + tu(t,x), \tag{3.99}$$

with $\hat{y}$ maximizer. The last equality holds because - in the thermodynamic limit - the Burgers equation becomes inviscid, hence trajectories represent

Galilean motion with a velocity $u(t,x)$ given explicitly by the previous expression. Under the replica symmetric assumption, from equation (3.91) and the definition of $u(t,x) = \partial_x S(t,x)$, in the thermodynamic limit we can finally recover the self-consistent equation for the overlap

$$u(t,x) = \partial_x S(t,x) = -\langle q_{12}\rangle_{t,x} = -\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh^2(z\sqrt{\hat{y}(t,x)}),$$

whose are the order parameter values minimizing the free energy, thus giving the exact value of this quantity in the thermodynamic limit. To recover statistical mechanics, we need to evaluate everything at $x = 0$ (e.g. in equation (3.99)), so that the value of $\hat{y}_0 = tq$ that maximizes the expression (3.98) is

$$\hat{y}_0 = tq = t\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh^2(z\sqrt{tq}).$$

This translates in the self-consistency equation for the order parameter q given by

$$q = \int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh^2(z\sqrt{tq}), \tag{3.100}$$

Going back to equation (3.97), with $\hat{y}$ maximizer, we finally have

$$S(x,t) = \frac{(x-\hat{y})^2}{2t} + 2\ln 2 + 2\int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \ln\cosh\big(z\sqrt{\hat{y}}\big) - \hat{y}.$$

Evaluating everything at $x = 0$ and $t = \beta^2$ (thus we use the relation $\hat{y}_0 = q\beta^2$) we can finally state:

Theorem 3.6. *The replica symmetric thermodynamic limit of the SK free energy density of the mean field spin-glass model is determined by the minimum value of*

$$f(\beta) = -\frac{1}{\beta}\alpha(\beta),$$

where

$$\alpha(\beta) = \ln 2 + \int_{-\infty}^{+\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \ln\cosh\big(\beta\sqrt{q}z\big) + \frac{\beta^2}{4}(1-q)^2. \tag{3.101}$$

Part Two: Statistical Mechanics for Information Storage and Retrieve

Chapter 4

Retrival phase of AI: the Hopfield 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, scheme, 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 [128]. 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 other neurons through dendrites. A bridge between neuron dynamics and memory processes has been made thanks to Y. Miyashita's experiments (1988) [95], 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 [65]. 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 [128]). 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 [37].

In the first part of the present Chapter, 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). Then, we will approach the Gibbs measure of the Hopfield model from a purely inferential perspective (this is to justify why we should keep an equilibrium effective description of a phenomenon that appears far from equilibrium). 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).

4.1 Generalities

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 $O(10^6)$ 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 4.1. 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 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 to labelling different words: $\{\xi^1, \dots, \xi^P\}$. 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 [101]) is the one incorporating Hebb's learning rule, i.e.

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu. \tag{4.1}$$

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

Definition 4.2. 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

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N} \sum_{\mu=1}^{P} \sum_{1 \leq i < j \leq N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j. \tag{4.2}$$

The next step is to introduce a set 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.

Definition 4.3. We define P overlaps m_μ, $\mu \in (1, ..., P)$ between the patterns and the neurons, also called Mattis magnetizations, as

$$m_\mu(\boldsymbol{\sigma}) \doteq m_\mu = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i \in [-1, 1]. \tag{4.3}$$

The Hamiltonian 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 indicating a retrieving behaviour.

4.2 The CW and the SK limits

In this Section, we illustrate the connection between the Hopfield model and the two models we studied in Chapters 2 and 3, 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 (4.2) 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 (2.1) and (4.2), 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 4.4. 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 (4.2) 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, in the previous Chapter 3 we introduced the prototype of complex systems, the Sherrington-Kirkpatrick model. It presents an amount of minima scaling as $\sim e^{cN}$ (with c not depending on N).

We remind that the SK Hamiltonian (3.2) is built with a interaction matrix $\boldsymbol{J}$ whose entries $J_{ij} \sim \mathcal{N}(0, 1)$. This implies that couplings can be either positive (hence favouring parallel spin configurations) as well as negative (encouraging anti-parallel spin configuration). Thus, in the thermodynamic limit, spins will receive conflicting signals with large probability, so we speak about "frustrated networks". Indeed frustration, the hallmark of complexity, is fundamental in order to split the phase space in several disconnected zones, i.e. in order to have several minima (or several stored patterns in neural network language). The mean field statistical mechanics for the low-noise behavior of spin-glasses has been first described by Parisi and it predicts a hierarchical organization of states and a relaxation dynamic spread over many timescales. Here, we just need to know that their natural order parameter is no longer the magnetization (since these systems do not magnetize), but the replica overlap Q_{ab} introduced in the previous Chapter. Spin-glasses are balanced ensembles of ferromagnets and antiferromagnets and, as a consequence, the magnetization m is always equal to zero. On the other hand, a comparison between two realizations of the system (pertaining to the same coupling set) is meaningful, since at large temperatures it is expected to be zero, as everything is uncorrelated, but at low temperature their overlap is strictly non-vanishing, as spins freeze in disordered but correlated states. More precisely, given two "replicas" of the system, labeled as a and b and with overlap Q_{ab}, the mean-field spin glass has a completely random paramagnetic phase with $\langle q \rangle \equiv 0$, and a "glassy phase" with $\langle q \rangle > 0$. These phase are split by a phase transition at $\beta_c = T_c = 1$.

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

$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu, \tag{4.4}$$

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 4.1 (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*

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2. \tag{4.5}$$

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

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|X_i - \mu_i|^{2+\delta}] = 0 \tag{4.6}$$

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 $J_{ij} = \sqrt{\frac{\lambda_N}{N}} \tilde{J}_{ij}$, where

$$\tilde{J}_{ij} = \frac{1}{\sqrt{P}} \sum_{\mu} \xi_i^\mu \xi_j^\mu, \tag{4.7}$$

and $\lambda_N = P/N$ 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 $\tilde{\boldsymbol{J}}$ converges in probability to $\mathcal{N}(0,1)$.

Remark 4.1. 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

$$J_{ij} \sim \frac{1}{\sqrt{N}} \tilde{J}_{ij}, \tag{4.8}$$

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 $\lambda = \lim_{N\to\infty} P/N$, 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 $\lambda = 0$), 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 $\lambda > 0$, the phase diagram lives in the λ, β plane. When λ is small enough, the system is expected to behave similarly to $\lambda = 0$ 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 $\lambda > 0$). However, for λ large enough, the Hopfield model collapses on the

[1] Notice that, throughout the rest of the thesis, we will use simply λ also if we are working at finite size N.

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, 89, 51].

We can summarize the content of the Hopfield model capabilities through its phase diagram as follows.[1] First of all, if the thermal noise $T = \beta^{-1}$ and the storage capacity λ are sufficiently low, the system works with almost no errors as an associative neural network (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 $\lambda_c \simeq 0.051$. 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 $0.0051 \leq \lambda \leq \lambda_c \simeq 0.138$. For $\lambda > 0.138$, 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 $T_c = 1 + \sqrt{\lambda}$, 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 λ, we distinguish between the following two regimes:

- Low storage (or low load) regime. This is the regime we investigate under the assumption that the patterns stored in the network grow slowly with the system size, i.e. $P \sim \log N$, such that $\lambda = 0$. Tools closer to the statistical mechanics of ferromagnets suffice to investigate this regime.

- High storage (or high load) regime. This is the regime we investigate under the hypotesis that the number of patterns stored in the network grows quickly with the system size, i.e. $P = \lambda N$ for large N, with $\lambda \in \mathbb{R}^+$. Tools closer to the statistical mechanics of spin glasses are needed to address its investigation.

Remark 4.2. We stress that, despite we know how to prove the existence of thermodynamic limit of the quenched free energy both for mean field

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

ferromagnets as well as mean field spin glasses, at presen, no proof of such an existence is available for the Hopfield model in the high storage regime. This is essentially because its quenched free energy shares properties with those of both the models, but the SK free energy is super-extensive while the CW free energy is sub-extensive in the system size. Their *mixture* actually escapes a rigorous treatment with the mathematical tools available at present. We will not give a proof of the existence of such a limit in the low storage, as we will give such a proof for an improved version of the Hopfield model we will discuss in Section 6.3.

4.3 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

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu. \tag{4.9}$$

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

$$\sigma_i(t+1) = \mathrm{sign}(\tanh(\beta \sum_{j \neq i} J_{ij}\sigma_j(t)) + \eta_i), \tag{4.10}$$

then 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

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\sum_{i,j<i} \Big(\frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu \Big) \sigma_i \sigma_j. \tag{4.11}$$

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

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 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 "concept". 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 4.5. Given two network configurations $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, the Hamming distance is defined as

$$d(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \frac{1}{2N} \sum_{i=1}^{N} |\sigma_{1,i} - \sigma_{2,i}|. \tag{4.12}$$

Remark 4.3. 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 [111]

Definition 4.6. 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

[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 $t \in \mathbb{Z}_+$.

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

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) \sim -\frac{1}{2}\sum_{i,j=1}^{N}\Big(\frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\Big)\sigma_i\sigma_j = -\frac{N}{2}\sum_{\mu=1}^{P} m_\mu^2, \tag{4.13}$$

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

$$m_\mu \sim \sqrt{\mathbb{E}m_\mu^2} = \sqrt{\frac{1}{N^2}\sum_{ij}\mathbb{E}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j} = \frac{1}{\sqrt{N}}, \tag{4.14}$$

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

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu=1}^{P} m_\mu^2 \sim \mathcal{O}(1), \tag{4.15}$$

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

$$H_N(\xi^1|\boldsymbol{\xi}) \simeq -\frac{N}{2} + \mathcal{O}(1). \tag{4.16}$$

Then, configurations aligned to the patterns are very convenient from an energetic point of view, with their stability growing with the network size.

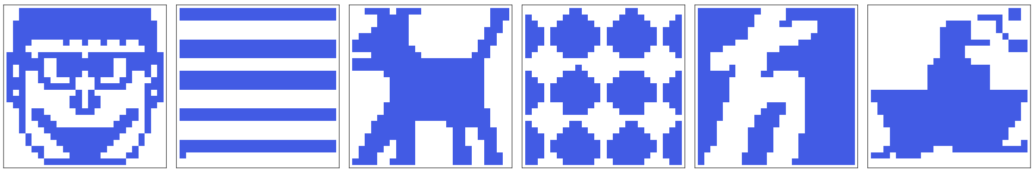

Figure 4.1: **Set of $P = 6$ patterns stored in a Hopfield network of $N = 625$ spins.** Patterns are black and while images: the network is dealing with digital storage of information [37].

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. 4.2. Here, we consider a Hopfield network consisting in $N = 625$ spins in which we stored the set of $P = 6$ patterns reported in Fig. 4.1, 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. 4.2, 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

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

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 $P = \lambda N$), 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

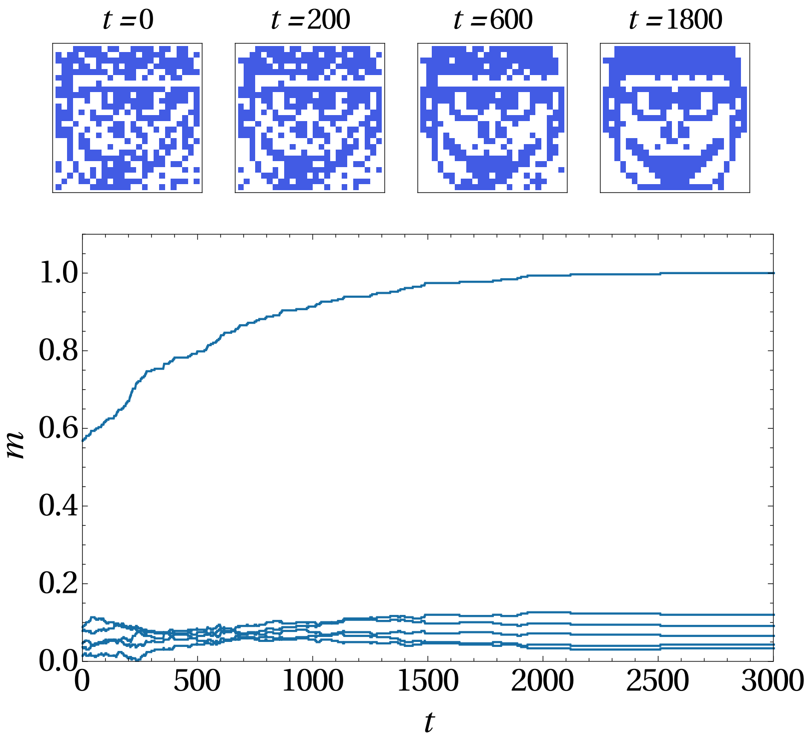

Figure 4.2: **Example of pattern reconstruction in a Hopfield network of $N = 625$ spins that stored $P = 6$ patterns.** Starting with a corrupted information, the Hopfield network is able to retrieve the associated pattern. We observe that, among the six Mattis magnetizations dedicated to quantify the 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.

the configuration

$$\tilde{\xi} = \text{sign}(\xi^1 + \xi^2 + \xi^3). \tag{4.17}$$

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. 4.4.

An example of dynamics ending in spurious configurations is reported in Fig. 4.3. In this case, we prepared the network in the spurious configuration (4.17), 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.

Signal-to-noise analysis

To get started with this analysis, we need to go back the Hamiltonian (4.2). By preparing the system near a given pattern, say ξ^1, we can express it as (again including self-interactions)

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{i,j=1}^{N} \xi_i^1 \xi_j^1 \sigma_i \sigma_j - \frac{1}{2N}\sum_{\mu\geq 2}\sum_{i,j=1}^{N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j. \tag{4.18}$$

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 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

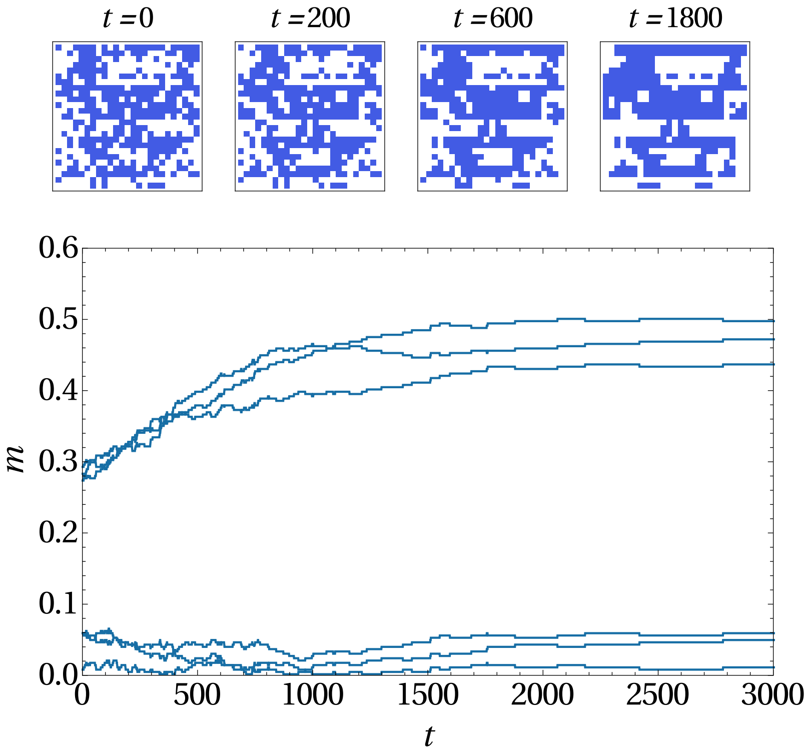

Figure 4.3: **Example of dynamics ending in a spurious state in a Hopfield network of $N = 625$ spins that stored $P = 6$ patterns.** In this example, it is possible to observe that several (three) Mattis magnetization raise sensibly over the noise due to the finite size effects and, correspondingly, the network has not been able to properly retrieve a single pattern, rather obtaining a useless mixture of the stored patterns.

doing this, external thermal noise is set to zero: $\beta \to \infty$). The condition for a given configuration to be dynamically stable is

$$h_i \sigma_i \geq 0 \quad \text{for each } i, \tag{4.19}$$

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

$$h_1 \xi_1^1 = \frac{1}{N} \sum_{j>1} \sum_{\mu} \xi_1^\mu \xi_j^\mu \xi_j^1 \xi_1^1 = \frac{N-1}{N} + \frac{1}{N} \sum_{j>1} \sum_{\mu>1} \xi_1^\mu \xi_j^\mu \xi_j^1 \xi_1^1, \tag{4.20}$$

where we separated the signal and the noise contributions and used the dichotomic nature of the patterns. The same analysis can be carried out

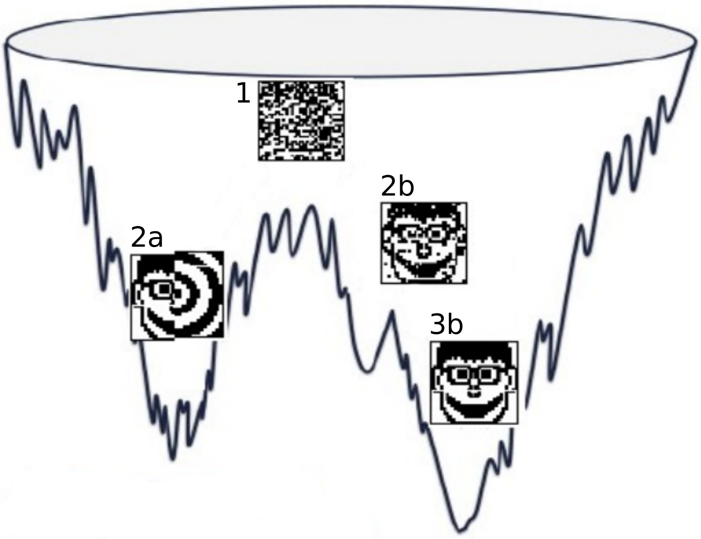

Figure 4.4: **Pictorial representation of minima landscape for the Hopfield model.** Starting with a noisy initial condition (1), the Hopfield network succeeds if the internal dynamics ends in a pure state configuration (with the evolution 1 → 2b → 3b). However, the network could end in a metastable state 2a, therefore failing to retrieve the desired pattern.

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

$$\Big|\frac{1}{N}\sum_{j>1}\sum_{\mu>1}\xi_1^\mu\xi_j^\mu\xi_j^1\xi_1^1\Big| \sim \sqrt{\frac{P-1}{N}}. \tag{4.21}$$

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 ($P = \lambda N$), 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

[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.

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 (4.17). 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

$$\mathcal{P}(\sigma_1 = \xi_1^\mu) = \frac{3}{4}, \quad \mathcal{P}(\sigma_1 = -\xi_1^\mu) = \frac{1}{4} \quad \text{for} \quad \mu = 1,2,3. \tag{4.22}$$

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

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i = \frac{1}{N}\Big(\frac{3N}{4} - \frac{N}{4}\Big) = \frac{1}{2}, \quad \mu = 1,2,3, \tag{4.23}$$

while $m_\mu \sim \mathcal{O}(N^{-1/2})$ for $\mu > 3$. This result should be compared with the numerical results reported in Fig. 4.3. The stability of the spurious configuration in this case is given by

$$h_1\sigma_1 = \sum_\mu m_\mu \xi_1^\mu \sigma_1 = \sigma_1(m_1\xi_1^1 + m_2\xi_1^2 + m_3\xi_1^3 + \sum_{\mu>3} m_\mu \xi_1^\mu). \tag{4.24}$$

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

$$\text{Signal} = 0.5\left(\xi_1^1 + \xi_1^2 + \xi_1^3\right)\text{sgn}\left(\xi_1^1 + \xi_1^2 + \xi_1^3\right) = 0.5|\xi_i^1 + \xi_i^2 + \xi_i^3|. \tag{4.25}$$

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. 4.3). 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]

$$E_1 < E_3 < E_5 < \cdots < E_\infty < \ldots E_4 < E_2. \tag{4.26}$$

4.4 The Hopfield model from statistical inference

As one may question about the validity of the equilibrium statistical mechanical approach since the problem we are facing ultimately addresses steady states of out-of-equilibrium current flows, it is very instructive to check that the probability distribution resulted from an inferential procedure exactly matches the Gibbs measure of the Hopfield model [36, 69]. 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

$$\langle m_\mu \rangle_{\text{exp}} = \frac{1}{N}\sum_i \xi_i^\mu \langle \sigma_i \rangle_{\text{exp}}, \qquad \langle m_\mu^2 \rangle_{\text{exp}} = \frac{1}{N^2}\sum_{ij} \xi_i^\mu \xi_j^\mu \langle \sigma_i \sigma_j \rangle_{\text{exp}}. \tag{4.27}$$

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 Chapter.

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]

$$\begin{aligned} S_{A,\beta,h}[\mathcal{P}] = &- \sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \log \mathcal{P}(\boldsymbol{\sigma}) + AN\Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) - 1 \Big) + \\ &+ hN \sum_\mu \Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \frac{1}{N} \sum_i \xi_i^\mu \sigma_i - \langle m_\mu \rangle \Big) \\ &+ \frac{\beta N}{2} \sum_\mu \Big(\sum_{\boldsymbol{\sigma}} \mathcal{P}(\boldsymbol{\sigma}) \frac{1}{N^2} \sum_{ij} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j - \langle m_\mu^2 \rangle \Big), \end{aligned} \tag{4.28}$$

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,

$$\frac{\delta S[\mathcal{P}]}{\delta \mathcal{P}(\boldsymbol{\sigma})} = -\log \mathcal{P}(\boldsymbol{\sigma}) - 1 + AN + h \sum_{i\mu} \xi_i^\mu \sigma_i + \frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j = 0, \tag{4.29}$$

which means that

$$\mathcal{P}(\boldsymbol{\sigma}) = \mathrm{cost} \exp\Big(\frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + h \sum_{i\mu} \xi_i^\mu \sigma_i \Big) \tag{4.30}$$

By putting the constant equal to $\mathrm{cost} = Z_N(\beta)^{-1}$, we prove the following

Theorem 4.2. *The partition function associated to the probability distribution $\mathcal{P}(\boldsymbol{\sigma})$ maximizing the Shannon entropy* (4.28) *with the costraints* (4.27) *for the first and the second moment of neural activity is*

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \exp\Big(\frac{\beta}{2N} \sum_{ij\mu} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + h \sum_{i\mu} \xi_i^\mu \sigma_i \Big). \tag{4.31}$$

Remark 4.4. Clearly, the second term is a bias term driving the systems towards a precise configuration (upon the breaking of parity symmetry). Of course, in neural networks applications one is often interested in the development of a spontaneous magnetization in absence of biases. In such a case, one has to chose $\langle m_\mu \rangle = 0$ (this of course does not mean that all final configurations have $m_\mu = 0$, but the sum over all the possible configurations would give $\langle m_\mu \rangle = 0$). In other words, we have to set $h = 0$, then we perfectly have the Hopfield partition function, as we will see in a moment.

[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.

4.5 Low storage of Boolean and Gaussian patterns

In this Section, we will concern with the statistical mechanics treatment of the low storage regime (i.e. $\lambda = 0$ case) of the Hopfield model. Before to proceed, we would like to stress that we can rewrite the Hamiltonian as (4.11)

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{i,j<i}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{1}{2N}\sum_{ij\mu}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \frac{1}{2N}\sum_{i\mu}(\xi_i^\mu)^2\sigma_i^2. \tag{4.32}$$

It is clear that the second term equals $P/2$ (while the energy is extensive in the network size), and it can be neglected in the thermodynamic limit in the low storage regime. Thus, we will omit it in the following computations, since it trivially contributes to the partition function (and therefore to the free energy). In order to make this Chapter self-contained, also in this case we give the fundamental definition for the statistical mechanics picture of Hopfield model.

Definition 4.7. The (pattern realization dependent) partition function of the Hopfield model equipped with N neurons σ_i, $i \in (1, ..., N)$ and P patterns ξ^μ, $\mu \in (1, ..., P)$ and described by the Hamiltonian (4.2) is

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \exp(-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi})). \tag{4.33}$$

The associated Boltzmann factor and Boltzmann-Gibbs average are respectively $B_N(\beta,\boldsymbol{\sigma}) = \exp(-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}))$ and

$$\omega_{\boldsymbol{\xi}}(F) = \frac{\sum_{\boldsymbol{\sigma}} F(\boldsymbol{\sigma})B_N(\beta,\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} B_N(\beta,\boldsymbol{\sigma})}, \tag{4.34}$$

for any arbitrary function $F(\boldsymbol{\sigma})$ of the spins. The statistical pressure is $\alpha_N(\beta) = -\beta f_N(\beta)$, where

$$f_N(\beta) = -\frac{1}{\beta N}\mathbb{E}\log Z_N(\beta), \tag{4.35}$$

is the (quenched) free energy.

The main interest of this Section (and of the following one, for which the same definitions hold) is to find the expression of the free energy in the thermodynamic limit $f(\beta) = \lim_{N\to\infty} f_N(\beta)$ in terms of the order parameters, both in the low and high storage regimes.

Saddle point method

The simplest way to determine the free energy for Hopfield model (and consequently the self-consistency equations for the order parameters) is the saddle point method. To apply this technique, we firstly express the (pattern realization dependent) partition function $Z_N(\beta) = \sum_{\boldsymbol{\sigma}} B_N(\beta, \boldsymbol{\sigma})$ by introducing the density of the states in the following way:

$$\begin{aligned} Z_N(\beta) &= \sum_{\boldsymbol{\sigma}} \exp\Big\{\frac{\beta}{2N}\sum_{ij\mu}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\Big\} = \\ &= \sum_{\boldsymbol{\sigma}} \int \Big(\prod_{\mu=1}^P dm_\mu \delta(m_\mu - \tfrac{1}{N}\sum_i \xi_i^\mu \sigma_i)\Big) \exp\Big(\frac{\beta N}{2}\sum_\mu m_\mu^2\Big) = \\ &= \sum_{\boldsymbol{\sigma}} \int \Big(\prod_{\mu=1}^P dm_\mu \frac{N d\bar{m}_\mu}{2\pi}\Big) \exp\Big(iN\sum_\mu \bar{m}_\mu m_\mu - i\sum_{i\mu}\bar{m}_\mu \xi_i^\mu \sigma_i \\ &+ \frac{\beta N}{2}\sum_\mu m_\mu^2\Big). \end{aligned}$$

This trick allows to linearize the spin-dependent part of the partition function, so that we can directly sum over the network configurations to get

$$\begin{aligned} Z_N(\beta) = \int \Big(\prod_{\mu=1}^P dm_\mu \frac{N d\bar{m}_\mu}{2\pi}\Big) \exp\Big(iN\sum_\mu \bar{m}_\mu m_\mu \\ + \sum_i \log 2\cos(\sum_\mu \bar{m}_\mu \xi_i^\mu) + \frac{\beta N}{2}\sum_\mu m_\mu^2\Big). \end{aligned}$$

On the $N \to \infty$ saddle point, we can replace $\bar{m}_\mu$ with the extremality condition of the quantity in the exponent with respect to m_μ (which is $\bar{m}_\mu = i\beta m_\mu$) and thus evaluated the integrals over the $\bar{m}$ variables. Then, we can write

$$Z_N(\beta) \underset{N\to\infty}{\sim} \int \Big(\prod_{\mu=1}^P dm_\mu\Big) \exp\Big(-\frac{\beta N}{2}\sum_\mu m_\mu^2 + \sum_i \log 2\cosh(\beta\sum_\mu m_\mu \xi_i^\mu)\Big).$$

On the r.h.s. of the last equation, we can again apply the saddle point formula, so finally we get

$$Z_N(\beta) \underset{N\to\infty}{\sim} \Big(\frac{2\pi}{N}\Big)^{\frac{P}{2}} \exp\Big(-\frac{\beta N}{2}\sum_\mu m_\mu^2 + \sum_i \log 2\cosh(\beta\sum_\mu m_\mu \xi_i^\mu)\Big).$$

Then, the quenched free energy in the thermodynamic limit is therefore

$$f(\beta) = \frac{1}{2}\sum_{\mu} m_{\mu}^2 - \frac{1}{\beta N}\sum_{i} \mathbb{E}\log 2\cosh(\beta \sum_{\mu} m_{\mu}\xi_i^{\mu}), \tag{4.36}$$

where order parameters satisfy the extremality condition. However, when commuting the sum over the spin index and the average over external noise, it's easy to understand that the only non-trivial operations concern with the sum over the configurations the vector $\boldsymbol{\xi}_i = (\boldsymbol{\xi}_i^1, \ldots, \boldsymbol{\xi}_i^P)$ for fixed i. Then, the net average $\mathbb{E}$ in (4.36) is

$$\mathbb{E} \equiv \frac{1}{2^P}\sum_{\xi_i^1=\pm 1}\cdots\sum_{\xi_i^P=\pm 1} . \tag{4.37}$$

Since we are summing over μ and since each pattern is extracted with equal probability with no regard on the index i, each term in the sum is equal,[1] so we prove the following [12]

Theorem 4.3. *The (quenched) free energy of the Hopfield model in the thermodynamic limit of the low storage regime is*

$$f(\beta) = \frac{1}{2}\sum_{\mu} m_{\mu}^2 - \frac{1}{\beta}\mathbb{E}\log 2\cosh(\beta \sum_{\mu} m_{\mu}\xi^{\mu}), \tag{4.38}$$

where the Mattis magnetizations satisfy the self-consistency equations

$$m_{\mu} = \mathbb{E}\,\xi^{\mu}\tanh(\beta\sum_{\nu} m_{\nu}\xi^{\nu}), \tag{4.39}$$

at the equilibrium states.

Remark 4.5. In particular, if we assume the only a single pattern is candidate to be retrieved, namely $m_1 \neq 0$ while $m_{\mu} = 0$ for all $\mu > 1$, we have the simpler self-consistency equation

$$m_1 = \tanh(\beta m_1), \tag{4.40}$$

which is precisely the Curie-Weiss law. To obtain this, we used the fact that the hyperbolic tangent is a odd function, and since $\xi_i^1 = \pm 1$ we have $\tanh(\beta m_1 \xi_i^1) = \xi_i^1 \tanh(\beta m_1)$, therefore compensating the ξ_i^1 prefactor. Since in this way there is no explicit dependence on the patterns, the $\mathbb{E}$ average is trivially computed.

[1] This is strictly true in the thermodynamic limit, in which the frequencies of pattern extractions equal the probabilities.

The Hamilton-Jacobi framework

We recall that, for CW model, the Guerra mechanical analogy consists in interpreting the statistical pressure as the Hamilton-Jacobi action for a classical particle propagating freely in a $1+1$-dimensional space, while the magnetization is interpreted as the classical spatial momentum. In the Hopfield case, we have P Mattis magnetizations, so it is natural to expect that, in this case, the dual mechanical system consists in a classical particle traveling in a $P+1$-dimensional space. Then, our basic quantity is the generalized partition function introduced in the next

Definition 4.8. The generalized partition function $Z_N(\beta; t, \boldsymbol{x}) \equiv Z_N(t, \boldsymbol{x})$ of the Hopfield model in the low storage regime, suitable for the Hamilton-Jacobi analysis, reads as

$$Z_N(t, \boldsymbol{x}) = \sum_{\boldsymbol{\sigma}} \exp\Big\{ -\frac{tN}{2} \sum_{\mu=1}^{P} m_\mu^2 + N \sum_{\mu=1}^{P} x_\mu m_\mu \Big\}, \tag{4.41}$$

where $\boldsymbol{x} = (x_1, \dots, x_P)$. The Boltzmann-Gibbs average with respect to this partition function will be denoted by $\omega_{t,\boldsymbol{x}}(\cdot)$. The generalized statistical pressure is

$$\alpha_N(t, \boldsymbol{x}) = \frac{1}{N} \mathbb{E} \log Z_N(t, \boldsymbol{x}). \tag{4.42}$$

Then, the derivatives of the statistical pressure with respect to the space-time coordinates are

$$\frac{\partial \alpha_N(t, \boldsymbol{x})}{\partial t} = -\frac{1}{2} \sum_{\mu=1}^{P} \mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu^2), \quad \nabla_{\boldsymbol{x}} \alpha_N(t, \boldsymbol{x}) = \sum_{\mu=1}^{P} \mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu). \tag{4.43}$$

With a simple check and by imposing the self-averaging of the order parameters in the thermodynamic limit ($\sum_{\mu=1}^{P} [\mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu^2) - \mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu)^2] \to 0$ as $N \to \infty$), we find that the next proposition holds:

Proposition 4.1. *The Guerra's action $S_N(t, \boldsymbol{x}) = \alpha_N(t, \boldsymbol{x})$ for the Hopfield model in the low storage regime obeys the following Hamilton-Jacobi PDE*

$$\frac{\partial S_N(t, \boldsymbol{x})}{\partial t} + \frac{1}{2} (\nabla_{\boldsymbol{x}} S_N(t, \boldsymbol{x}))^2 + V_N(t, \boldsymbol{x}) = 0, \tag{4.44}$$

with $V_N(t, \boldsymbol{x}) = \sum_{\mu=1}^{P} [\mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu^2) - \mathbb{E}\omega_{t,\boldsymbol{x}}(m_\mu)^2]$.

Note that this is the straightforward generalization of the CW model. As in the latter case, the potential vanishes in the thermodynamic limit because of the self-averaging property of the order parameters, leaving us with the Hamilton-Jacobi equations for a free particle. Of course, the classical solution of such an equation (4.44) are straight trajectories (in the $P+1$-dimensional space this time), i.e. $x_\mu = x_{0,\mu} + (t-t_0)m_\mu$, where we called m_μ the thermodynamic value of the Mattis magnetization.[1]

Proposition 4.2. *The (thermodynamic limit of the) action can be computed via the fundamental theorem of calculus as*

$$S(t,\boldsymbol{x}) = S(t_0,\boldsymbol{x}_0) + \int_{t_0}^{t} dt'\,\mathcal{L}. \tag{4.46}$$

Also in this case, the Lagrangian $\mathcal{L} = \frac{1}{2}\sum_{\mu=1}^{P} m^2$ is constant over the classical trajectories, so the integral is trivial. Again, we fix the initial condition $t_0 = 0$ and express the starting point $\boldsymbol{x}_0$ is terms of $\boldsymbol{x}$ and t. Then, for the one-body term, we have

$$\begin{aligned} S(0,\boldsymbol{x}_0) &= \frac{1}{N}\sum_i \mathbb{E}\log 2\cosh(\boldsymbol{x}_0\cdot\xi_i) = \\ &= \mathbb{E}\log 2\cosh(\boldsymbol{x}_0\cdot\xi), \end{aligned} \tag{4.47}$$

where in the last line we used the argument in the previous section. With these straightforward computation, we find the statistical pressure

$$S(t,\boldsymbol{x}) = \frac{t}{2}\sum_{\mu=1}^{P} m^2 + \mathbb{E}\log 2\cosh(\boldsymbol{x}_0\cdot\xi). \tag{4.48}$$

Finally, imposing $t=-\beta$, $x_\mu = x_{0,\mu} + (t-t_0)m_\mu$ and set $\boldsymbol{x}=0$, we obtain the following

Theorem 4.4. *The thermodynamic limit of the Guerra's action (i.e. the statistical pressure) for the Hopfield model in the low storage regime reads as*

$$\alpha(\beta) = S(-\beta,0) = -\frac{\beta}{2}\sum_{\mu=1}^{P} m^2 + \mathbb{E}\log 2\cosh(\beta\sum_{\mu=1}^{P} m_\mu\xi^\mu). \tag{4.49}$$

Recalling that $\alpha(\beta) = -\beta f(\beta)$, this is exactly the expression of the statistical pressure previously obtained with a pure statistical mechanical treatment (4.38).

[1] This means that

$$P(\boldsymbol{m}(\boldsymbol{\sigma})) \underset{N\to\infty}{\longrightarrow} \delta(\boldsymbol{m}(\boldsymbol{\sigma}) - \boldsymbol{m}), \tag{4.45}$$

with $\boldsymbol{m} = (m_1,\dots,m_P)$ and δ being the P-dimensional Dirac delta distribution and $\boldsymbol{m} = \lim_{N\to\infty}\mathbb{E}\omega_{t,\boldsymbol{x}}(\boldsymbol{m}(\boldsymbol{\sigma}))$.

4.6 High storage of Boolean patterns: replica trick.

It is time to turn to the complex case, which is the high storage limit $P = \lambda N$ with $\lambda \in \mathbb{R}^+$. 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

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{i,j<i}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{1}{2N}\sum_{ij\mu}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \frac{P}{2}. \tag{4.50}$$

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 $\lambda/2$, 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 4.7 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 λ (previously, it was not needed because $\lambda = 0$ 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, while all the other $P-1$ contributions by the not-retrieved patterns accounts 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

$$f(\beta,\lambda) = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}'\log Z_N(\beta,\lambda) + \frac{\lambda}{2}, \tag{4.51}$$

where the average over quenched disorder is

$$\mathbb{E}' \equiv \mathbb{E}_{\xi^2}\dots\mathbb{E}_{\xi^P}. \tag{4.52}$$

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 $\mathbb{E}'Z_N^n(\beta,\lambda)$. Introducing the replica index a running over different equivalent realization of the same system, we can write it as

$$\begin{aligned}\mathbb{E}'Z_N^n(\beta,\lambda) &= \mathbb{E}'\sum_{\boldsymbol{\sigma}^{(1)}}\cdots\sum_{\boldsymbol{\sigma}^{(n)}}\exp\Big(\frac{\beta}{2N}\sum_{ija\mu}\xi_i^\mu\xi_j^\mu\sigma_i^{(a)}\sigma_j^{(a)}\Big) = \\ &= \mathbb{E}'\sum_{\boldsymbol{\sigma}^{(1)}}\cdots\sum_{\boldsymbol{\sigma}^{(n)}}\int\Big(\prod_{a\mu}d\mu(z_\mu^{(a)})\Big)\exp\Big(\sqrt{\frac{\beta}{N}}\sum_{i\mu a}\xi_i^\mu\sigma_i^{(a)}z_\mu^{(a)}\Big),\end{aligned} \tag{4.53}$$

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

$$\int d\mu(z) = \int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}\exp(-z^2/2). \tag{4.54}$$

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

$$\mathbb{E}'Z_N^n(\beta,\lambda) = \sum_{\boldsymbol{\sigma}^{(1)}}\cdots\sum_{\boldsymbol{\sigma}^{(n)}} z_{\text{signal}}[\boldsymbol{\sigma}]z_{\text{noise}}[\boldsymbol{\sigma}], \tag{4.55}$$

where

$$\begin{aligned}z_{\text{signal}}[\boldsymbol{\sigma}] &= \int\Big(\prod_{a=1}^{n}d\mu(z_1^{(a)})\Big)\exp\Big(\sqrt{\frac{\beta}{N}}\sum_{ia}\xi_i^1\sigma_i^{(a)}z_1^{(a)}\Big), \\ z_{\text{noise}}[\boldsymbol{\sigma}] &= \int\Big(\prod_{a,\mu\geq 2}d\mu(z_\mu^{(a)})\Big)\mathbb{E}'\exp\Big(\sqrt{\frac{\beta}{N}}\sum_{ia,\mu\geq 2}\xi_i^\mu\sigma_i^{(a)}z_\mu^{(a)}\Big).\end{aligned} \tag{4.56}$$

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

$$z_{\text{noise}}[\boldsymbol{\sigma}] = \prod_{\mu\geq 2}\int\Big(\prod_a d\mu(z_\mu^{(a)})\Big)\exp\Big(\frac{\beta}{2N}\sum_{iab}\sigma_i^{(a)}\sigma_i^{(b)}z_\mu^{(a)}z_\mu^{(b)}\Big). \tag{4.57}$$

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

$$\begin{aligned} z_{\text{noise}}[\boldsymbol{\sigma}] = \prod_{\mu\geq 2}\int \Big(\prod_{a=1}^{n} d\mu(z_\mu^{(a)})\Big)\Big(\prod_{ab} dQ_{ab}\delta(Q_{ab} - \tfrac{1}{N}\sum_i \sigma_i^{(a)}\sigma_i^{(b)})\Big)\cdot \\ \cdot \exp\Big(\frac{\beta}{2N}\sum_{ab} Q_{ab} z_\mu^{(a)} z_\mu^{(b)}\Big). \end{aligned} \tag{4.58}$$

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]

$$\begin{aligned} z_{\text{noise}}[\boldsymbol{\sigma}] = \int \Big(\prod_{ab} dQ_{ab}\frac{N dP_{ab}}{2\pi}\Big) \exp\Big(iN\sum_{ab} P_{ab}Q_{ab} - i\sum_{iab} P_{ab}\sigma_i^{(a)}\sigma_i^{(b)} \\ - \frac{P}{2}\log\det(\mathbf{1} - \beta\boldsymbol{Q})\Big). \end{aligned} \tag{4.59}$$

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 $a = 1,\dots,n$. This allows to further simplify the expression. Including the singal term, with some manipulations we arrive (after some trivial rescalings $z_1^{(a)} \to \sqrt{\beta N} m_1^{(a)}$, $P_{ab} \to i\frac{\lambda\beta^2}{2}P_{ab}$) at the final result

$$\mathbb{E}' Z_N^n(\beta,\lambda) = \int d\mu(\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P})\exp(-NA[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}]), \tag{4.60}$$

where

$$\begin{aligned} A[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}] = \frac{\beta}{2}\sum_a (m_1^{(a)})^2 + \frac{\lambda\beta^2}{2}\sum_{ab} P_{ab}Q_{ab} + \frac{\lambda}{2}\log\det(\mathbf{1} - \beta\boldsymbol{Q}) \\ - \mathbb{E}\log\sum_{\boldsymbol{s}}\exp\Big(\beta\sum_a \xi^1 m_1^{(a)} s_a + \frac{\alpha\beta^2}{2}\sum_{ab} P_{ab}s_a s_b\Big), \end{aligned} \tag{4.61}$$

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

[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.

free energy of Hopfield model is recovered by taking the limit

$$f(\beta, \lambda) = \lim_{n\to 0} \frac{1}{\beta n} A[\boldsymbol{m}_1, \boldsymbol{Q}, \boldsymbol{P}]. \tag{4.62}$$

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

$$\begin{aligned} m_1^{(a)} &= m_1 \quad \forall a, \\ Q_{ab} &= \delta_{ab} + q(1-\delta_{ab}), \\ P_{ab} &= p(1-\delta_{ab}). \end{aligned} \tag{4.63}$$

Therefore, we are left only with three order parameters. With this *Ansatz*, it is possible to compute the replica symmetric free energy $f_{RS}(\beta, \lambda)$. 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 [37] to an exhaustive description), we stress that the last one (involving the quenched averaged $\mathbb{E}$) can be estimated as

$$-\frac{\lambda\beta^2}{2}np + n\,\mathbb{E}\int d\mu(z) \log 2\cosh(\beta m_1 \xi^1 + \beta z\sqrt{\lambda p}) + \mathcal{O}(n^2). \tag{4.64}$$

Putting everything together and including the correction term $\lambda/(2\beta)$ as prescribed above, we are finally able to state the following [13]

Theorem 4.5. *The replica symmetric free energy for the Hopfield model in the high storage regime is*

$$\begin{aligned} f_{RS}(\beta,\lambda) = \frac{m_1^2}{2} + \frac{\lambda\beta}{2}p(1-q) + \frac{\lambda}{2\beta}\Big(\beta + \log[1-\beta(1-q)] - \frac{q\beta}{1-\beta(1-q)}\Big) \\ - \frac{1}{\beta}\int d\mu(z) \log 2\cosh\Big(\beta m_1 + \beta z \frac{\sqrt{\lambda q}}{1-\beta(1-q)}\Big), \end{aligned} \tag{4.65}$$

[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

$$
\begin{aligned}
m_1 &= \int_{-\infty}^{+\infty} d\mu(z) \tanh\Big(\beta m_1 + \beta z \frac{\sqrt{\lambda q}}{1-\beta(1-q)}\Big),\\
q &= \int_{-\infty}^{+\infty} d\mu(z) \tanh^2\Big(\beta m_1 + \beta z \frac{\sqrt{\lambda q}}{1-\beta(1-q)}\Big).
\end{aligned}
\tag{4.66}
$$

at the equilibrium states.

Remark 4.6. 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

$$
\begin{aligned}
m_\mu &= \int_{-\infty}^{+\infty} d\mu(z)\, \mathbb{E}\, \xi^\mu \tanh\Big(\beta \boldsymbol{m}\cdot\boldsymbol{\xi} + \beta z \frac{\sqrt{\lambda q}}{1-\beta(1-q)}\Big),\\
q &= \int_{-\infty}^{+\infty} d\mu(z)\, \mathbb{E} \tanh^2\Big(\beta \boldsymbol{m}\cdot\boldsymbol{\xi} + \beta z \frac{\sqrt{\lambda q}}{1-\beta(1-q)}\Big).
\end{aligned}
\tag{4.67}
$$

Remark 4.7. 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 [121]. However, it has been shown that the modification due to the replica symmetry breaking is negligible to a first approximation (we refer to [37, 122] for further details).

4.7 High storage of Gaussian patterns: interpolation method

The Hopfield neural network presents different thermodynamic behaviours depending on the amount of noise the network is embedded in, the amount of load the network has to face but also the nature of the patterns coding the information the network is dealing with, as we will prove in this Section. In particular we are going to show that if the patterns have real entries (rather than digital as previously assumed) a retrieval region in the phase diagram is no longer possible: this is not really surprising since, if we think that with digital patterns the model can handle at most a fraction $O(N)$ of them, the amount of information supplied to the network with real patterns if real patterns are supplied to the network is not even comparable. Consequently, the system (when not in the ergodic phase) is always a spin-glass, and it can never work as a pattern recognizer.

In this Section, using an analogy among neural networks and bipartite spin glasses, we move the interpolating techniques (essentially based on two different stochastic perturbations) which we use to give a complete description of the analogical Hopfield model phase diagram in the replica symmetric approximation in the high storage regime, and we'll show that the network doesn't present a retrieval phase. The solely difference w.r.t. the model analyzed in the previous Section is the pattern probability distribution (which previously was $\mathcal{P}(\xi_i^\mu) = (1/2)\delta_{\xi_i^\mu,+1} + (1/2)\delta_{\xi_i^\mu,-1}$), is here replaced by a Gaussian probability distribution, i.e.

$$\mathcal{P}(\xi_i^\mu) = \frac{1}{\sqrt{2\pi}} e^{-(\xi_i^\mu)^2/2}.$$

We can apply the Gaussian integration to linearize with respect to the bilinear quenched memories carried by $\xi_i^\mu \xi_j^\mu$ the Hopfield partition function, thus obtaining

$$Z_N(\beta,\lambda) = \sum_{\boldsymbol{\sigma}} \int \prod_{\mu=1}^{P} d\mu(z_\mu) \ \exp\Big\{\sqrt{\frac{\beta}{N}} \sum_{\mu=1}^{P} \sum_{i=1}^{N} \xi_i^\mu \sigma_i z_\mu\Big\}, \tag{4.68}$$

where $d\mu(z_\mu)$ is again the standard Gaussian measure for all the z_μ. Taken F as a generic function of the neurons, we define the Gibbs measure $\omega(F)$ like in equation (2.4) at a given level of noise β. The s-replicated Gibbs measure is defined as in (3.14) in which we replace the coupling matrix $\boldsymbol{J}$ with the pattern vectors $\boldsymbol{\xi}$. All the single Gibbs measures are independent at the same noise level β^{-1}, and share an identical distribution of quenched memories ξ.

Here, the quenched average $\mathbb{E}$ is obviously defined as

$$\mathbb{E}[F(\boldsymbol{\xi})] = \int \prod_{\mu=1}^{P} \prod_{i=1}^{N} \frac{d\xi_i^\mu e^{-\frac{(\xi_i^\mu)^2}{2}}}{\sqrt{2\pi}} F(\boldsymbol{\xi}) = \int F(\boldsymbol{\xi}) d\mu(\boldsymbol{\xi}),$$

for a generic function of these memories $F(\boldsymbol{\xi})$. Of course, $\mathbb{E}[\xi_i^\mu] = 0$ and $\mathbb{E}[(\xi_i^\mu)^2] = 1$.Reflecting the bipartite nature of the Hopfield model expressed by Eq. (4.68), we again introduce two other order parameters:

Definition 4.9. The overlap between the replicated neurons (first party overlap), is defined as

$$Q_{ab} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(a)}\sigma_i^{(b)} \in [-1,+1].$$

The overlap between the replicated Gaussian variables z (second party overlap), is defined as

$$P_{ab} = \frac{1}{P}\sum_{\mu=1}^{P} z_\mu^{(a)} z_\mu^{(b)} \in (-\infty,+\infty).$$

Both the two order parameters above play a considerable role in the theory, since they can express thermodynamical quantities.

We now pay attention to the structure of the free energy: as standard in the interpolation scheme, we want to obtain the it via a sum rule in which we may isolate explicitly the order parameter fluctuations in order to be able to neglect them achieving a replica-symmetric behaviour. Due to the equivalence among neural networks and bipartite spin-glasses [27], we need to generalize the way cavity field and the stochastic stability techniques [10] work on spin glasses to these structures by introducing the following interpolation scheme. For the sake of clearness, in order to exploit the interpolation method adapted to the physics of the model, we introduce 3 free parameters in the interpolating structure (i.e. A, B, C) that we fix *a fortiori*, once the sum rule is almost achieved. In a pure stochastic stability fashion, we need to introduce also two classes of i.i.d. $\mathcal{N}(0,1)$ variables, namely N variables η_i and P variables θ_μ, whose average is still encoded into the $\mathbb{E}$ operator. Then, we make the following

Definition 4.10. The (pattern realization dependent) interpolating statistical $\alpha_N(\beta;t) \equiv \alpha_N(t)$ pressure is

$$\begin{aligned}\alpha_N(t) =& \frac{1}{N}\mathbb{E}\log\sum_{\sigma}\int\prod_{\mu=1}^{P}d\mu(z_\mu)\ \exp\Big\{\sqrt{t}\frac{\beta}{N}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\sigma_i z_\mu\Big\}\cdot\\ &\cdot\exp\Big\{A\sqrt{1-t}\sum_{i=1}^{N}\eta_i\sigma_i\Big\}\cdot\exp\Big\{B\sqrt{1-t}\sum_{\mu=1}^{P}\theta_\mu z_\mu\Big\}\cdot\\ &\cdot\exp\Big\{C\frac{1-t}{2}\sum_{\mu=1}^{P}z_\mu^2\Big\},\end{aligned} \tag{4.69}$$

where $\eta_i, \theta_\mu \sim \mathcal{N}(0,1)$.

Remark 4.8. Of course, in this way also the interpolating partition function $Z_N(t)$ and the Boltzmann factor $B_N(t)$ are straightforwardly defined.

As usual, we stress that $t \in [0,1]$ interpolates between $t = 0$ (where the interpolating quenched pressure becomes made of non-interacting systems, i.e. a series of one-body problems whose integration is straightforward) and the opposite limit, $t = 1$ recovering the original quenched free energy. The plan is again to evaluate the t-streaming of such a quantity and then obtain the Hopfield model free energy by using the fundamental theorem of calculus, just like we did for the Curie-Weiss model in Section 2.4 and for Sherrington-Kirkpatrick spin glass in 3.5. To formalize this procedure, we state the following

Proposition 4.3. *The quenched free energy of the Hopfield model, equipped with real-valued patterns, in the high storage regime, is realized as*

$$\alpha_N(\beta,\lambda) = \alpha_N(t=1) = \alpha_N(t=0) + \int_0^1 ds \Big[\partial_t \alpha_N(t)\Big]_{t=s}. \tag{4.70}$$

When evaluating the streaming $\partial_t \alpha$, we get the sum of four terms, which we call I, II, III and IV. Each one of them comes as a consequence of the derivation of a corresponding exponential term appearing in the interpolating pressure (4.69). Once introduced the averages $\omega_t(\cdot)$ and $\langle\cdot\rangle_t = \mathbb{E}\Omega_t$ that naturally extend the Gibbs measures encoded in the interpolating scheme (and reduce to the proper one whenever setting $t = 1$), we can write them down as

$$
\begin{aligned}
\mathrm{I} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(z)\ \sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\sigma_i z_\mu\cdot\frac{1}{2\sqrt{t}}B_N(t)\Big] = \\
&= \frac{\sqrt{\beta}}{2N\sqrt{Nt}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\mathbb{E}\Big[\xi_i^\mu\omega_t(\sigma_i z_\mu)\Big] = \frac{\sqrt{\beta}}{2N\sqrt{Nt}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\mathbb{E}\Big[\partial_{\xi_i^\mu}\omega_t(\sigma_i z_\mu)\Big] = \\
&= \frac{\beta}{2N}\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2) - \frac{\lambda\beta}{2}\langle q_{12}p_{12}\rangle_t;
\end{aligned}
\tag{4.71}
$$

$$
\begin{aligned}
\mathrm{II} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(z)\frac{-A}{2\sqrt{1-t}}\sum_{i=1}^{N}\eta_i\sigma_i B_N(t)\Big] = \\
&= \frac{-A}{2N\sqrt{1-t}}\sum_{i=1}^{N}\mathbb{E}\Big[\eta_i\omega_t(\sigma_i)\Big] = \frac{-A}{2N\sqrt{1-t}}\sum_{i=1}^{N}\mathbb{E}\Big[\partial_{\eta_i}\omega_t(\sigma_i)\Big] = \\
&= -\frac{A^2}{2}\big(1-\langle q_{12}\rangle_t\big);
\end{aligned}
\tag{4.72}
$$

$$
\begin{aligned}
\mathrm{III} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(z)\ \frac{-B}{2\sqrt{1-t}}\sum_{\mu=1}^{P-1}\theta_\mu z_\mu B_N(t)\Big] = \\
&= \frac{-B}{2N\sqrt{1-t}}\sum_{\mu=1}^{P-1}\mathbb{E}\Big[\theta_\mu\omega_t(z_\mu)\Big] = \frac{-B}{2N\sqrt{1-t}}\mathbb{E}\Big[\partial_{\theta_\mu}\omega_t(z_\mu)\Big] = \\
&= -\frac{B^2}{2N}\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2) + \frac{\lambda B^2}{2}\langle p_{12}\rangle_t.
\end{aligned}
\tag{4.73}
$$

In the computation of these three terms, we used Wick theorem with respect to the auxiliary fields η_i and η_μ. Finally, the term IV is easily calculated as

$$
\mathrm{IV} = \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(z)\frac{-C}{2}\sum_{\mu=1}^{P-1}z_\mu^2 B_N(t)\Big] = -\frac{C}{2N}\sum_{\mu=1}^{P}\mathbb{E}\omega_t(z_\mu^2),
\tag{4.74}
$$

In the replica symmetric ansatz, the order parameters m, q_{12}, p_{12} do not fluctuate with respect to the quenched average, so let us define their thermodynamic values as $\langle m\rangle_t = m$, $\langle q_{12}\rangle_t = q$, $\langle p_{12}\rangle_t = p$. Summing all the contributions given by I, II, III and IV, and adding conveniently and subtracting the term $\lambda\beta qp/2$, we get

$$\frac{\partial \alpha_N(t)}{\partial t} = (\beta - B^2 - C)\frac{1}{2N}\mathbb{E}\sum_{\mu=1}^{P}\omega_t(z_\mu^2) - \frac{\lambda\beta}{2}\langle q_{12}p_{12}\rangle_t$$
$$- \frac{A^2}{2}(1 - \langle q_{12}\rangle_t) + \frac{\lambda B^2}{2}\langle p_{12}\rangle_t + \frac{\lambda\beta}{2}qp - \frac{\lambda\beta}{2}qp.$$

Since A, B and C are tunable parameters, we can make the following choice:

$$A = \sqrt{\lambda\beta p}, \quad B = \sqrt{\beta q}, \quad C = \beta(1-q),$$

so that we get

$$\frac{\partial \alpha_N(t)}{\partial t} = -\frac{\lambda\beta}{2}\langle (q_{12} - q)(p_{12} - p)\rangle_t - \frac{\lambda\beta}{2}p(1-q). \tag{4.75}$$

Remark 4.9. In the definition of the overlaps Q_{ab} and P_{ab}, we showed that these quantities can also take negative values. This might seem in contradiction with the definition of the parameters A and B because we have a square root of a potentially negative term. Going on with the discussion, precisely once we get to the self-consistency equations, we will verify that these quantities can only take non-negative values, thus justifying our procedure.

In order to get the replica symmetric solution $\alpha_N(\beta)$ we impose the self-averaging of the overlaps in the thermodynamic limit, so that we need to evaluate only

$$\alpha_N(\beta) = \alpha_N(t=0) - \frac{\lambda\beta}{2}p(1-q) - \frac{\lambda\beta}{2},$$

where we have reinserted the factor that comes from the diagonal term of the first party as explained previously. The evaluation of $\alpha_N(t=0)$ is easily performed because it is a one-body calculation. With simple manipulations, we have

$$\alpha_N(t=0) = \frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\left\{\sqrt{\lambda\beta p}\sum_{i=1}^{N}\eta_i\sigma_i\right\}$$
$$+ \frac{1}{N}\mathbb{E}\log\int\prod_{\mu=1}^{P}dz_\mu\ \exp\left\{-\frac{1}{2}\sum_{\mu=1}^{P}z_\mu^2(1-\beta(1-q)) + \sqrt{\beta q}\sum_{\mu=1}^{P}\theta_\mu z_\mu\right\} =$$
$$= \int d\mu(z)\ \log 2\cosh\left(\sqrt{\lambda\beta p}z\right) + \frac{\lambda}{2}\log\left(1 - \beta(1-\bar{q})\right)$$
$$+ \lambda\mathbb{E}\log\int dr e^{-r^2/2}e^{\sqrt{\frac{\beta\bar{q}}{1-\beta(1-\bar{q})}}\eta r},$$

where we introduced $r = \sigma z$, with σ defining the standard Gaussian variance such that $\sigma^2 = (1-\beta(1-q))^{-1}$. As a consequence, we get

$$\begin{aligned}\alpha_N(t=0) = &\log 2 + \int d\mu(z) \log\cosh(\sqrt{\lambda\beta p}z)\\ &+ \frac{\lambda}{2}\log\Big(\frac{1}{1-\beta(1-q)}\Big) + \frac{\lambda\beta}{2}\frac{q}{1-\beta(1-q)}.\end{aligned}$$

Overall, we can state the next theorem.

Theorem 4.6. *The thermodynamic limit of the replica symmetric pressure function of the analogical Hopfield neural network is given by the following expression*

$$\begin{aligned}\alpha(\beta,\lambda) = \log 2 + \int d\mu(z)\log\cosh(\sqrt{\lambda\beta p}z) + \frac{\lambda}{2}\log\Big(\frac{1}{1-\beta(1-q)}\Big)\\ + \frac{\lambda\beta}{2}\frac{q}{1-\beta(1-q)} - \frac{\lambda\beta}{2}p(1-q) - \frac{\lambda\beta}{2},\end{aligned} \tag{4.76}$$

where q and p satisfy equations (4.77) *and* (4.78) *respectively at the equilibrium states.*

Indeed, self-consistency relations can be found by imposing equal to zero the partial derivatives of the free energy with respect to its order parameters. Therefore, we obtain

$$\frac{\partial\alpha}{\partial q} = \frac{\lambda\beta}{2}\Big(p - \frac{\beta q}{(1-\beta(1-q))^2}\Big) = 0, \tag{4.77}$$

$$\frac{\partial\alpha}{\partial p} = \frac{\lambda\beta}{2}\Big(\int d\mu(z)\tanh^2(\sqrt{\lambda\beta p}z) - q\Big) = 0. \tag{4.78}$$

Upon eliminating p on the saddle point, we have

$$q = \int d\mu(z)\tanh^2\Big(\frac{\sqrt{\lambda q}\beta z}{1-\beta(1-q)}\Big). \tag{4.79}$$

Remark 4.10. We would like to point out two key observations:

- the quenched noise is universal. In fact, if we look back at Eq. (3.100) (with $t = 1$) and compare it to equation (4.79), we see that the part identifying the value of q for the phase transition has the same structure. We will verify this universal property also for the hybrid neural network, thus deducing that the SG-paramagnetic line is the same in every model showing these properties.

- the signal is not universal. In particular, while a retrieval phase (coded by a positive Mattis magnetization in the thermodynamic limit) is available when the stored patterns are digital, in the high storage this region is destroyed if the stored patterns have real-valued entries.

We are now ready to analyze *hybrid* Hopfield neural networks, whose patterns are mixtures of Boolean and Gaussian samples. Before entering the (quite lenghty) calculations, we would to motivate why we are going to spend a lot of efforts for tackling this problem. In a nutshell, as we will see in the next Chapter (which is devoted to learning capabilities of neural networks), we will introduce the Restricted Boltzmann Machine (RBM) as the archetype of machine learning: remarkably we will show that RBMs and Hopfield networks share the same marginal distributions, suggesting the intuitive concept that learning and retrieval are two aspects of a single phenomenon that is cognition. However, while we equipped the Hopfield model with both real valued and discrete patterns, patterns in the RBM - or *weights* in Machine Learning jargon - must necessarily be real since learning requires making derivatives w.r.t. these weights. However, we already saw that Hopfield networks with solely real-valued patterns do not have a retrieval region. Thus, in order for the learning algorithms of the RBMs to be able to harmonically coexist with the retrieval phase in the dual Hopfield models, we will see that an hybrid Hopfield network still preserves the phase diagramù of the Boolean Hopfield model, despite having the bulk of patterns with real-valued entries. Hence, the overall theoretical scaffold of AI - when analyzed through statistical mechanics - is preserved.

4.7.1 The hybrid case: a Boolean pattern in a real sea

We will begin our discussion by setting up the characteristics of the model and the statistical tools. We recall here

Definition 4.11. The Hamiltonian function for the analog Hopfield neural network with N Ising spins $\sigma_i = \pm 1$, $i = 1, \dots, N$ and P patterns is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N} \sum_{1 \le i < j \le N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j, \tag{4.80}$$

where $\xi_i^\mu \sim \mathcal{N}(0,1)$ $\forall i = 1, \dots, N$, $\mu = 1, \dots, P$.

In our hybrid model, we will instead introduce one binary pattern $\tilde{\xi}$ and $P - 1 \sim \lambda N$, with $\lambda > 0$, real patterns ξ_i^μ with the following probability distributions:

$$\begin{cases} \tilde{\xi}_i \colon \mathcal{P}\{\tilde{\xi}_i = 1\} = \mathcal{P}\{\tilde{\xi}_i = -1\} = \frac{1}{2} & \forall i = 1 \dots N, \\ \xi_i^\mu \sim \mathcal{N}(0,1) & \forall i = 1 \dots N, \mu = 1, \dots, P-1. \end{cases}$$

The Hamiltonian is therefore naturally splitted in two terms separating the part that concerning the binary pattern from the real ones. Therefore, we have the following

Definition 4.12. The Hamiltonian for an hybrid Hopfield neural network with N Ising spins $\sigma_i = \pm 1$, $i = 1, \dots, N$, one Boolean $\tilde{\xi}$ and $P-1$ Gaussian patterns ξ_i^μ is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}, \tilde{\xi}) = -\frac{1}{N}\sum_{i<j}^{N} \tilde{\xi}_i \tilde{\xi}_j \sigma_i \sigma_j - \frac{1}{N}\sum_{i<j}^{N}\sum_{\mu=1}^{P-1} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j, \tag{4.81}$$

where $\mathcal{P}(\tilde{\xi}_i = \pm 1) = 1/2$ and $\mathcal{P}(\xi_i^\mu) = \mathcal{N}(0,1)$.

It is clear that the such an Hamiltonian can be expressed as[1]

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}, \tilde{\xi}) = -\frac{1}{2N}\sum_{i,j=1}^{N} \tilde{\xi}_i \tilde{\xi}_j \sigma_i \sigma_j - \frac{1}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P-1} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + \frac{1}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{P-1} (\xi_i^\mu)^2. \tag{4.82}$$

It is important to notice that, if we perform a Mattis gauge on the Boolean term, the Hamiltonian is written as the sum of a Curie-Weiss and an analog Hopfield term. Such structure will be convenient in preparation for the next Section, where we will combine the Guerra interpolation techniques done for both of the models in sections 2.4 and 4.7. Then:

Definition 4.13. The partition function associated to the Hamiltonian (4.81) is

$$Z_N(\beta, \lambda) = e^{-\frac{\beta}{2N}\sum_{i\mu}(\xi_i^\mu)^2} \sum_{\boldsymbol{\sigma}} \exp\left\{ \frac{\beta}{2N}\sum_{i,j=1}^{N} \tilde{\xi}_i \tilde{\xi}_j \sigma_i \sigma_j + \frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P-1} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \right\}. \tag{4.83}$$

[1] In this equality, we used the identity $\sum_{i,j} x_i x_j = \sum_i x_i^2 + 2\sum_{i<j} x_i x_j$ and included the self-interaction on the Boolean part with an error vanishing in the thermodynamic limit.

Even though the partition function Z_N, and the functions that we will obtain from it, is also a function of the patterns, to simplify the notation we only leave visible (as standard in this thesis) the dependency from the parameters β and λ.

Our ultimate goal is to write an explicit expression for the thermodynamic limit of the quenched free energy density $f_N(\beta,\lambda)$ or of the statistical pressure $\alpha_N(\beta,\lambda)$ in the order parameters. After that, we want to find the minimizing measures for the free energy by deriving this expression with respect to the order parameters. As a result, we will get the self-consistency equations for the latters, whose solutions will be the ones we are looking for to have the actual value of the free energy of the system at the equilbrium. For practical reasons, we choose to work on the quenched intensive pressure $\alpha_N(\beta,\lambda)$, of which we remind the definition:

Definition 4.14. The statistical pressure and the free energy of the hybrid Hopfield model (4.81) are

$$\alpha_N(\beta,\lambda) = -\beta f_N(\beta) = \frac{1}{N}\mathbb{E}\log Z_N(\beta), \tag{4.84}$$

where $\mathbb{E}$ stands for the average over the quenched memories for any generic function $F(\xi,\tilde{\xi})$ depending on ξ and $\tilde{\xi}$, that is

$$\mathbb{E}[F(\xi,\tilde{\xi})] = \int \prod_{\mu=1}^{P-1}\prod_{i=1}^{N}\frac{d\xi_i^\mu}{\sqrt{2\pi}}e^{-\frac{(\xi_i^\mu)^2}{2}}\cdot \prod_{j=1}^{N}\sum_{\{\tilde{\xi}_j\}}\frac{1}{2}F(\xi,\tilde{\xi}). \tag{4.85}$$

As we did for standard Hopfield model, it is now useful to apply the Gaussian integration to linearize the second factor of equation (4.83) with respect to the bilinear quenched memories carried by $\xi_i^\mu\xi_j^\mu$:

$$\begin{aligned} Z_N(\beta,\lambda) =& \exp\Big(-\frac{\beta}{2N}\sum_i\sum_\mu(\xi_i^\mu)^2\Big)\cdot \\ &\cdot\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\tilde{\xi}_i\tilde{\xi}_j\sigma_i\sigma_j + \frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P-1}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\Big\} = \\ =& \exp\Big(-\frac{\beta}{2N}\sum_{i\mu}(\xi_i^\mu)^2\Big)\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\tilde{\xi}_i\tilde{\xi}_j\sigma_i\sigma_j\Big\}\cdot \\ &\cdot\int_{\mathbb{R}^{P-1}}d\mu(\boldsymbol{z})\exp\Big\{\sqrt{\frac{\beta}{N}}\sum_{\mu=1}^{P-1}\sum_{i=1}^{N}\xi_i^\mu\sigma_i z_\mu\Big\}, \end{aligned} \tag{4.86}$$

where $d\mu(\boldsymbol{z}) = \prod_{\mu=1}^{P-1} \frac{dz_\mu}{\sqrt{2\pi}} e^{z_\mu^2/2}$ is the $(P-1)$-dimensional Gaussian measure. Therefore, using the definition of α_N given in (4.84) and equations (4.83) and (4.86), we have

$$\begin{aligned}\alpha_N(\beta,\lambda) = &-\frac{\lambda\beta}{2} + \frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{ij}\tilde{\xi}_i\tilde{\xi}_j\sigma_i\sigma_j\Big\}\cdot \\ &\cdot\int_{\mathbb{R}^{P-1}} d\mu(\boldsymbol{z})\ \exp\Big\{\sqrt{\frac{\beta}{N}}\sum_{\mu=1}^{P-1}\sum_{i=1}^{N}\xi_i^\mu\sigma_i z_\mu\Big\}.\end{aligned} \tag{4.87}$$

Given expression (4.87), we can notice that the second term is factorized into an exponential that refers to the binary pattern and the other one referring to the gaussian patterns. As mentioned previously after equation (4.82), the presence of these two distinct factors recalling the Curie-Weiss and the Hopfield models, suggests the choice of an interpolation function that combines the ones used in these cases. Therefore, we shall proceed in perfect analogy with sections 2.4 and 4.7. To do so, we have to introduce four tunable parameters A, B, C, ψ (to be fixed later) and two classes $\{\eta_i\}_{i=1}^N$ and $\{\theta_\mu\}_{\mu=1}^{p-1}$ of i.i.d. $\mathcal{N}(0,1)$ variables. Therefore, with the use of a parameter $t \in [0,1]$, we can define the following interpolating functions.

Definition 4.15. The interpolating partition function for the simple hybrid Hopfield network is the following:

$$\begin{aligned}Z_N(\beta,\lambda;t) \doteq Z_N(t) = e^{-\frac{\beta}{2N}\sum_i\sum_\mu(\xi_i^\mu)^2}&\sum_{\boldsymbol{\sigma}}\int d\mu(\boldsymbol{z})\ \exp\Big\{t\frac{\beta}{2}Nm^2\Big\}\cdot \\ &\cdot\exp\Big\{(1-t)\psi Nm + \sqrt{t}\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P-1}\xi_i^\mu\sigma_i z_\mu\Big\}\cdot \\ &\cdot\exp\Big\{A\sqrt{1-t}\sum_{i=1}^{N}\eta_i\sigma_i + B\sqrt{1-t}\sum_{\mu=1}^{P-1}\theta_\mu z_\mu\Big\}\cdot \\ &\cdot\exp\Big\{(1-t)\frac{C}{2}\sum_{\mu=1}^{P-1}z_\mu^2\Big\},\end{aligned} \tag{4.88}$$

and the interpolating quenched intensive pressure is

$$\alpha_N(\beta,\lambda;t) \doteq \alpha_N(t) = \frac{1}{N}\mathbb{E}\log Z_N(t). \tag{4.89}$$

Again, it is simple to check that we recover the same interpolating properties that we had for the separate models. In fact, $\alpha_N(t=1) = \alpha_N(\beta,\lambda)$

and $\alpha_N(t=0)$ is made of a series of one-body systems. Furthermore, thanks to equation (4.88), we can extend the Gibbs measures ω and Ω to their interpolating counterparts ω_t and Ω_t. We can therefore introduce the average $\langle\cdot\rangle_t = \mathbb{E}\Omega_t$ recovering the proper measures for $t=1$. Hence, assuming that $\alpha_N(t)$ is sufficiently regular, we give the expression for the quenched intensive pressure using as usual the fundamental theorem of calculus:

Proposition 4.4. *The thermodynamic limit of the quenched free energy of the hybrid Hopfield model is realized as*

$$\alpha_N(\beta,\lambda) = \alpha_N(t=1) = \alpha_N(t=0) + \int_0^1 ds\Big[\partial_t\alpha_N(t)\Big]_{t=s}. \tag{4.90}$$

Since $\alpha_N(t=0) = \frac{1}{N}\mathbb{E}\log Z_N(t=0)$ consists in one-body systems, we can directly evaluated it. Thus, we obtain

$$\begin{aligned}
Z_N(t=0) &= e^{-\frac{\beta}{2N}\sum_{i\mu}(\xi_i^\mu)^2}\sum_{\boldsymbol{\sigma}}\exp\Big\{\psi\sum_{i=1}^N\tilde{\xi}_i\sigma_i + A\sum_{i=1}^N\eta_i\sigma_i\Big\}\cdot\\
&\quad\cdot\int d\mu(\boldsymbol{z})\exp\Big\{\frac{C}{2}\sum_{\mu=1}^{P-1}z_\mu^2 + B\sum_{\mu=1}^{P-1}\theta_\mu z_\mu\Big\} =\\
&= e^{-\frac{\beta}{2N}\sum_{i\mu}(\xi_i^\mu)^2}\Big(\prod_{i=1}^N\sum_{\boldsymbol{\sigma}}e^{(\psi+A\eta_i)\sigma_i}\Big)\int\prod_{\mu=1}^{P-1}\frac{dz_\mu}{\sqrt{2\pi}}e^{-\left(\frac{1-C}{2}\right)z_\mu^2+B\theta_\mu z_\mu} =\\
&= e^{-\frac{\beta}{2N}\sum_{i\mu}(\xi_i^\mu)^2}\Big(2^N\prod_{i=1}^N\cosh(\psi\tilde{\xi}_i + A\eta_i)\Big)\frac{e^{\frac{P-1}{2(1-C)}B^2\theta^2}}{(1-C)^{(P-1)/2}},
\end{aligned} \tag{4.91}$$

where $\theta\sim\mathcal{N}(0,1)$. Consequently, the associated intensive pressure is

$$\alpha_N(t=0) = -\frac{\lambda\beta}{2} + \mathbb{E}\log 2\cosh(\psi\tilde{\xi}+A\theta) - \frac{\lambda}{2}\log(1-C) + \frac{\lambda B^2}{2(1-C)}. \tag{4.92}$$

Finally, we have to calculate $\partial_t\alpha_N(t) = \partial_t\frac{1}{N}\mathbb{E}\log Z_N(t)$. In this computation, it is convenient to observe that it is the sum of six terms (I, II, III, IV, V, VI), each coming from the derivation of the corresponding exponential term in equation (4.88).

Terms I and II come from the boolean section and their calculation is straightforward:

$$\begin{aligned}\mathrm{I} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\int d\mu(\boldsymbol{z})\sum_{\boldsymbol{\sigma}}\frac{\beta}{2}Nm^2B_N(t)\Big] = \\ &= \frac{\beta}{2}\mathbb{E}\omega_t(m^2)\end{aligned} \tag{4.93}$$

$$\begin{aligned}\mathrm{II} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\int d\mu(\boldsymbol{z})\sum_{\boldsymbol{\sigma}}(-\psi Nm)B_N(t)\Big] = \\ &= -\psi\mathbb{E}\omega_t(m),\end{aligned} \tag{4.94}$$

where $B_N(t)$ is the generalized Boltzmann state. For terms III, IV and V, we proceed in perfect analogy with section 4.7 by using Wick theorem on the patterns and the auxiliary fields:

$$\begin{aligned}\mathrm{III} &= \frac{1}{N}\mathbb{E}\Big[\sum_{\boldsymbol{\sigma}}\int d\mu(\boldsymbol{z})\,\sqrt{\frac{\beta}{N}}\sum_{i,\mu}^{N,P-1}\xi_i^\mu\sigma_i z_\mu\cdot\frac{1}{2\sqrt{t}}B_N(t)\Big] = \\ &= \frac{\sqrt{\beta}}{2N\sqrt{Nt}}\sum_{i,\mu}^{N,P-1}\mathbb{E}\Big[\xi_i^\mu\omega_t(\sigma_i z_\mu)\Big] = \\ &= \frac{\sqrt{\beta}}{2N\sqrt{Nt}}\sum_{i,\mu}^{N,P-1}\mathbb{E}\Big[\partial_{\xi_i^\mu}\omega_t(\sigma_i z_\mu)\Big] = \\ &= \frac{\beta}{2N}\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2) - \frac{\lambda\beta}{2}\langle q_{12}p_{12}\rangle_t;\end{aligned} \tag{4.95}$$

$$\begin{aligned}\mathrm{IV} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(\boldsymbol{z})\,\frac{-A}{2\sqrt{1-t}}\sum_{i=1}^{N}\eta_i\sigma_i B_N(t)\Big] = \\ &= \frac{-A}{2N\sqrt{1-t}}\sum_{i=1}^{N}\mathbb{E}\Big[\eta_i\omega_t(\sigma_i)\Big] = \\ &= \frac{-A}{2N\sqrt{1-t}}\sum_{i=1}^{N}\mathbb{E}\Big[\partial_{\eta_i}\omega_t(\sigma_i)\Big] = \\ &= -\frac{A^2}{2}\big(1-\langle q_{12}\rangle_t\big);\end{aligned} \tag{4.96}$$

$$\begin{aligned}\mathrm{V} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(\boldsymbol{z})\ \frac{-B}{2\sqrt{1-t}}\sum_{\mu=1}^{P-1}\theta_\mu z_\mu B_N(t)\Big] = \\ &= \frac{-B}{2N\sqrt{1-t}}\sum_{\mu=1}^{P-1}\mathbb{E}\Big[\theta_\mu\omega_t(z_\mu)\Big] = \frac{-B}{2N\sqrt{1-t}}\mathbb{E}\Big[\partial_{\theta_\mu}\omega_t(z_\mu)\Big] = \\ &= -\frac{B^2}{2N}\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2) + \frac{\lambda B^2}{2}\langle p_{12}\rangle_t.\end{aligned} \tag{4.97}$$

Finally, the term VI is easily computed with standard Gaussian integration, so

$$\begin{aligned}\mathrm{VI} &= \frac{1}{N}\mathbb{E}\Big[\frac{1}{Z_N(t)}\sum_{\boldsymbol{\sigma}}\int d\mu(\boldsymbol{z})\ \frac{-C}{2}\sum_{\mu=1}^{P-1}z_\mu^2 B_N(t)\Big] = \\ &= -\frac{C}{2N}\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2).\end{aligned} \tag{4.98}$$

Summing the final expressions of equations (4.93), (4.94), (4.95), (4.96), (4.97) and (4.98), we get

$$\begin{aligned}\frac{\partial\alpha_N}{\partial t}(t) &= \mathbb{E}\omega_t\Big(\frac{\beta}{2}\big(m^2 - \frac{2\psi}{\beta}m\big)\Big) + \frac{1}{2N}\big(\beta - B^2 - C\big)\sum_{\mu=1}^{P-1}\mathbb{E}\omega_t(z_\mu^2) + \\ &- \frac{\lambda\beta}{\langle q_{12}p_{12}\rangle_t} - \frac{A^2}{2}\big(1 - \langle q_{12}\rangle_t\big) + \frac{\lambda B^2}{2}\langle p_{12}\rangle_t.\end{aligned} \tag{4.99}$$

Again, to assume the the replica symmetric ansatz, we require that the order parameters m, q_{12}, p_{12} do not fluctuate with respect to the quenched average in the thermodynamic limit, so we introduce the only values that they can acquire:

$$\langle m\rangle_t = \bar{m}, \quad \langle q_{12}\rangle_t = q, \quad \langle p_{12}\rangle_t = p.$$

We now fix the four parameters A, B, C, ψ in order to simplify the expression of $\alpha_N(t)$, then making the RS *Ansatz* in the thermodynamic limit. We choose

$$A = \sqrt{\lambda\beta p}, \quad B = \sqrt{\beta q}, \quad C = \beta(1-q), \quad \psi = \bar{m}\beta, \tag{4.100}$$

where the choice of ψ can be justified for the same reasons that we illustrated in section 2.4. Replacing these values and adding and subtracting $\frac{\lambda\beta}{2}qp$ in Eq. (4.99), we have

$$\frac{\partial\alpha_N}{\partial t}(t) = \frac{\beta}{2}\mathbb{E}\omega_t((m-\bar{m}^2)) - \frac{1}{2}\beta\bar{m}^2 - \frac{\lambda\beta}{2}\langle(q_{12}-q)(p_{12}-p)\rangle_t - \frac{\lambda\beta}{2}p(1-q). \tag{4.101}$$

According to Eq. (4.90), we can finally write the explicit expression for the quenched pressure at a finite volume N using expressions (4.92) and (4.101)

$$\begin{aligned}\alpha_N(\beta,\lambda) =& -\frac{\lambda\beta}{2} + \log 2 + \mathbb{E}\log\cosh(\beta\bar{m} + \sqrt{\lambda\beta p}\theta) - \frac{\lambda}{2}\log\big(1-\beta(1-q)\big) + \\ & -\frac{\lambda\beta}{2}\cdot\frac{q}{1-\beta(1-q)} + \frac{\beta}{2}\mathbb{E}\omega_t((m-\bar{m})^2) - \frac{\beta}{2}\bar{m}^2 + \\ & -\frac{\lambda\beta}{2}\langle(q_{12}-q)(p_{12}-p)\rangle_t - \frac{\lambda\beta}{2}p(1-q).\end{aligned} \tag{4.102}$$

Finally, performing the thermodynamic limit of the previous equation and making the RS ansatz, we have that

$$\mathbb{E}\omega_t((m-\bar{m})^2) \underset{N\to\infty}{\longrightarrow} 0,$$
$$\langle(q_{12}-q)(p_{12}-p)\rangle_t \underset{N\to\infty}{\longrightarrow} 0,$$

so we can state the following theorem

Theorem 4.7. *The replica symmetric thermodynamic limit of the free energy density of the hybrid Hopfield neural network with N Ising spins* $\sigma_i \in \{-1,+1\}$ $\forall i = 1,\dots,N$, *one binary pattern and a high load of* $P-1$ *real patterns, described by the Hamiltonian* (4.82)*, is determined by the minimum value of the following function:*

$$f(\beta,\lambda) = -\frac{1}{\beta}\alpha(\beta,\lambda),$$

where

$$\begin{aligned}\alpha(\beta,\lambda) =& -\frac{\lambda\beta}{2} + \log 2 + \mathbb{E}\log\cosh(\beta m + \sqrt{\lambda\beta p}\theta) - \frac{1}{2}\beta m^2 + \\ & -\frac{\lambda}{2}\log\big(1-\beta(1-q)\big) + \frac{\lambda\beta q}{2\big(1-\beta(1-q)\big)} - \frac{\lambda\beta}{2}p(1-q),\end{aligned} \tag{4.103}$$

where the order parameter m, q and p are respectively the magnetization, the binary and real overlaps.

Note that, for the sake of notation homogeneity, in previous the we dropped the bar over the thermodynamic value for the magnetization.

Remark 4.11. For $\lambda = 0$, we obtain the classic intensive pressure of the Curie-Weiss model, see Chapter 2. Furthermore, eliminating the Boolean pattern, i.e. putting $m = 0$, we recover the expression for the Hopfield network with a high load of analogical patterns (4.76).

To find the value of the free energy of the system, according to the minimum energy principle and the maximum entropy principle, we have to values for the order parameters in which f is minimized (or α is maximized). To achieve this, we derive equation (4.103) with respect to its order parameters and set the result equal to zero, thus obtaining the self-consistency relations.

$$\begin{aligned}\frac{\partial\alpha}{\partial p} &= \frac{\partial}{\partial p}\int_{-\infty}^{+\infty}\frac{d\theta}{\sqrt{2\pi}}e^{-\theta^2/2}\log\cosh(\beta m+\sqrt{\lambda\beta p}\theta)-\frac{\lambda\beta}{2}(1-q)=\\ &=\frac{\sqrt{\lambda\beta}}{2\sqrt{p}}\int_{-\infty}^{+\infty}\frac{d\theta}{\sqrt{2\pi}}\theta e^{-\theta^2/2}\tanh(\beta m+\sqrt{\lambda\beta p}\theta)=\\ &=-\frac{\lambda\beta}{2}\int_{-\infty}^{+\infty}d\mu(\theta)\tanh^2(\beta m+\sqrt{\lambda\beta p}\theta)+\frac{\lambda\beta}{2}-\frac{\lambda\beta}{2}(1-q)=0,\end{aligned} \tag{4.104}$$

where we have performed an integration by parts in to get to the third equation, and

$$\frac{\partial\alpha}{\partial q}=\frac{\lambda\beta}{2}\left[-\frac{1}{1-\beta(1-q)}+\frac{1-\beta}{(1-\beta(1-q))^2}+p\right]=0 \tag{4.105}$$

$$\begin{aligned}\frac{\partial\alpha}{\partial m} &= \frac{\partial}{\partial m}\int d\mu(\theta)\log\cosh(\beta m+\sqrt{\lambda\beta p}\theta)-\beta m=\\ &=\int_{-\infty}^{+\infty}d\mu(\theta)\tanh(\beta m+\sqrt{\lambda\beta p}\theta)\beta-\beta m=0.\end{aligned} \tag{4.106}$$

Hence, we get the three self-consistency equations

$$m=\int_{\mathbb{R}}d(\theta)\tanh(\beta m+\sqrt{\lambda\beta p}\theta), \tag{4.107}$$

$$q=\int_{\mathbb{R}}d\mu(\theta)\tanh^2(\beta m+\sqrt{\lambda\beta p}\theta), \tag{4.108}$$

$$p=\frac{\beta q}{\big(1-\beta(1-q)\big)^2}. \tag{4.109}$$

From these equations we find that q has a second order transition phase, thus giving the SG-paramagnetic transition, while we the magnetization $\bar{m}$ doesn't only have the null value as a solution to (4.107), thus showing a first order transition phase from a ferromagnetic state to a paramagnetic one.

We can finally assert that the hybrid Hopfield neural network with one boolean pattern and a high load of real patters presents a retrieval phase of the boolean memory. The details of these last considerations will be illustrated in the next section with an appropriate phase diagram, because this simple model is just a particular case of the general one with a low storage of boolean patterns.

4.7.2 The hybrid case: many Boolean patterns in a real sea

After studying the hybrid model with one boolean pattern and a high load of real patterns, it comes naturally to wonder whether the non trivial case of a hybrid network with the same amount of real patters but with a low load of binary ones will behave in the same way. To answer to this question, we have combined two different solving strategies: the stochastic stability and the Hamilton-Jacobi technique.

We will assign the variables $\tilde{\xi}^\nu$, $\nu = 1, \dots, K \sim \gamma \log N$ to the dychotomic patterns and ξ^μ, $\mu = 1, \dots, P \sim \lambda N$ to the real ones (obviously $\gamma, \lambda \geq 0$). The probability distribution generating the patterns are respectively

$$\begin{cases} \mathcal{P}\{\tilde{\xi}_i^\nu = 1\} = \mathcal{P}\{\tilde{\xi}_i^\nu = -1\} = \frac{1}{2} & \forall i = 1, \dots, N \text{ and } \nu = 1, \dots, k \\ \xi_i^\mu \sim \mathcal{N}(0,1) & \forall i = 1, \dots, N \text{ and } \mu = 1, \dots, p. \end{cases}$$

Following the description of the generic Hopfield neural network given in (4.2), we then define

Definition 4.16. The Hamiltonian function for the hybrid Hopfield neural network with a high load P of real patterns and a low load K of boolean memories is

$$\begin{aligned} H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}) &= -\frac{1}{N} \sum_{1 \leq i < j \leq N} \sum_{\nu=1}^{K} \tilde{\xi}_i^\nu \tilde{\xi}_j^\nu \sigma_i \sigma_j - \frac{1}{N} \sum_{1 \leq i < j \leq N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j = \\ &= -\frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\nu=1}^{K} \tilde{\xi}_i^\nu \tilde{\xi}_j^\nu \sigma_i \sigma_j - \frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + \frac{k}{2} \\ &\quad + \frac{1}{2N} \sum_{i=1}^{N} \sum_{\mu=1}^{P} (\xi_i^\mu)^2 . \end{aligned} \tag{4.110}$$

Of course, the average over the quenched memories $\{\tilde{\xi}_i^\nu\}_{i,\nu}$ and $\{\xi_i^\mu\}_{i,\mu}$ for a generic function $F(\boldsymbol{\xi}, \tilde{\boldsymbol{\xi}})$ is now

$$\mathbb{E}\left[F(\boldsymbol{\xi}, \tilde{\boldsymbol{\xi}})\right] = \int \prod_{\mu=1}^{P} \prod_{i=1}^{N} \frac{d\xi_i^\mu}{\sqrt{2\pi}} e^{-\frac{(\xi_i^\mu)^2}{2}} \cdot \prod_{\nu=1}^{k} \prod_{j=1}^{N} \sum_{\{\tilde{\xi}_j^\nu\}} \frac{1}{2} F(\xi, \tilde{\xi}).$$

In the partition function, we first notice that we can isolate the diagonal factor as follows:

$$
\begin{aligned}
Z_N(\beta,\lambda) =& \exp\Big\{-\frac{\beta K}{2}+\frac{\beta}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{P}(\xi_i^\mu)^2\Big\}\cdot \\
&\cdot\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\nu=1}^{K}\tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j+\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\Big\}.
\end{aligned}
\tag{4.111}
$$

Therefore, the pressure density function is

$$
\begin{aligned}
\alpha_N(\beta,\lambda) =& \frac{1}{N}\mathbb{E}\log Z_N(\beta,\lambda) = \\
=&\frac{1}{N}\mathbb{E}\Big[-\frac{\beta K}{2}-\frac{\beta}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{P}(\xi_i^\mu)^2\Big]+ \\
&+\frac{1}{N}\mathbb{E}\log\Big(\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\nu=1}^{K}\tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j+\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\Big\}\Big) = \\
=&-\mathcal{O}(\log N/N)-\frac{\lambda\beta}{2}+ \\
&+\frac{1}{N}\mathbb{E}\log\Big(\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\nu=1}^{K}\tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j+\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\Big\}\Big),
\end{aligned}
\tag{4.112}
$$

in which $\lambda = P/N$ is kept finite for $N \to \infty$. The simple Gibbs measure and the s-replicated one are defined in the standard way (using the appropriate Hamiltonian obviously for the weights). Again, sometimes it will be convenient to use the notation $\langle\cdot\rangle$ instead of $\mathbb{E}\omega(\cdot)$ or $\mathbb{E}\Omega(\cdot)$, depending on the context. Finally, we define the order parameters of our model. For tackling the quenched noise, we keep the two-replica overlaps q and p we used so far, while to handle with the signal carried by the digital stored patterns, we introduce (one for each boolean pattern) the standard Mattis magnetizations as

$$
m_\nu = \frac{1}{N}\sum_{i=1}^{N}\tilde{\xi}_i^\nu\sigma_i \in [-1,1]. \tag{4.113}
$$

As always, we are interested in finding an explicit expression for the pressure density in terms of the order parameters and the set of self-consistency equations. Looking at expression (4.112), we have to work on the third term of the equation. One possible way to proceed is to combine the Hamilton-Jacobi procedure, that will be the main tool for what concerns the Boolean section of the model, and the stochastic stability, that we will use for the analogical part.

Summarizing the process, we will first study a generalized model depending on the interpolating parameters $t \in \mathbb{R}$, $x \in \mathbb{R}^K$, $\psi \in [0,1]$ such that, once set to $t = \beta$, $x = 0$ and $\psi = 1$, we recover our original hybrid network. Subsequently, we will apply the stochastic stability method to simplify the expression of the pressure density function. The Hamilton-Jacobi formalism comes in hand when dealing with the explicit calculation of $\alpha_N(t, x, \psi = 0)$ in which we will interpret one term as the density pressure of a Hopfield network with binary patterns and an external field. We also highlight the fact that the order in which we apply these two methods is interchangeable, and in the next Section we show how, reasonably proceeding the other way around, we obtain the same results.

It is useful to repeat what we have done in equation (4.86) to linearize the gaussian section of the pressure density function $\alpha_N(\beta)$. Following the same steps, we define

Definition 4.17. The generalized partition function for the hybrid Hopfield neural network in this framework is

$$\begin{aligned}
Z_N(t, \boldsymbol{x}, \psi) = \exp\Big\{ -\frac{\beta K}{2} - \frac{\beta}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{P}(\xi_i^\mu)^2\Big\}\cdot \\
\sum_{\boldsymbol{\sigma}}\int_{\mathbb{R}^P} d\mu(\boldsymbol{z})\ \exp\Big\{\frac{t}{2N}\sum_{i,j=1}^{N}\sum_{\nu=1}^{K}\tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j + \sum_{\nu=1}^{K}x_\nu\sum_{i=1}^{N}\tilde{\xi}_i^\nu\sigma_i\Big\}\cdot \\
\cdot\exp\Big\{\sqrt{\psi}\sqrt{\frac{\beta}{N}}\sum_{\mu=1}^{P}\sum_{i=1}^{N}\xi_i^\mu\sigma_i z_\mu + A\sqrt{1-\psi}\sum_{i=1}^{N}\eta_i\sigma_i\Big\}\cdot \\
\cdot\exp\Big\{B\sqrt{1-\psi}\sum_{\mu=1}^{P}\theta_\mu z_\mu + C\frac{1-\psi}{2}\sum_{\mu=1}^{P}(z_\mu)^2\Big\},
\end{aligned} \tag{4.114}$$

with $\theta_\mu, \eta_i \sim \mathcal{N}(0,1)$ $\forall\mu = 1,\dots,P$, $i = 1,\dots,N$.

In the same fashion as previous cases, we can extend the Gibbs measures to $\omega_{t,\boldsymbol{x},\psi}$ and $\Omega_{t,\boldsymbol{x},\psi}$ and the overall average to $\langle\cdot\rangle_{t,\boldsymbol{x},\psi}$. We stress the fact these quantities maintain the same interpolation properties of the partition function, thus recovering standard statistical mechanics measures ones we set $t = \beta$, $\boldsymbol{x} = 0$ and $\psi = 1$.

Proposition 4.5. *As standard at this point, in order to obtain an explicit expression for the quenched free energy of the hybrid Hopfield model, we apply the Fundamental Theorem of Calculus to $\alpha_N(t,\boldsymbol{x},\psi)$ in the ψ variable:*

$$\begin{aligned}\alpha_N(t,\boldsymbol{x}) &\doteq \alpha_N(t,\boldsymbol{x},\psi=1) = \\ &= \alpha_N(t,\boldsymbol{x},\psi=0) + \int_0^1 d\varphi\Big[\partial_\psi \alpha_N(t,\boldsymbol{x},\psi)\Big]_{\psi=\varphi}.\end{aligned} \tag{4.115}$$

To compute the first term we only have to go through a standard Gaussian integration, hence

$$\begin{aligned}\alpha_N(t,\boldsymbol{x},\psi=0) =& -\mathcal{O}(\log N/N) - \frac{\lambda\beta}{2} + \\ &+ \frac{1}{N}\mathbb{E}\Big[\log\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{t}{2N}\sum_{i,j=1}^N\sum_{\nu=1}^K \tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j + \sum_{\nu=1}^K x_\nu\sum_{i=1}^N\tilde{\xi}_i^\nu\sigma_i + A\sum_{i=1}^N\eta_i\sigma_i\Big\}\cdot \\ &\cdot\int_{\mathbb{R}^P}\frac{dz_1\cdots dz_P}{(2\pi)^{P/2}}\exp\Big\{\sum_{\mu=1}^P\Big(B\theta_\mu z_\mu + \frac{C-1}{2}z_\mu^2\Big)\Big\}\Big] = \\ =& -\mathcal{O}(\log N/N) - \frac{\lambda\beta}{2} + \frac{1}{N}\mathbb{E}\log\Big(\frac{1}{(1-C)^{P/2}}e^{\frac{B^2\theta^2}{2(1-C)}P}\Big) + \\ &+ \frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{t}{2N}\sum_{i,j=1}^N\sum_{\nu=1}^K\tilde{\xi}_i^\nu\tilde{\xi}_j^\nu\sigma_i\sigma_j + \sum_{\nu=1}^K x_\nu\sum_{i=1}^N\tilde{\xi}_i^\nu\sigma_i + A\sum_{i=1}^N\eta_i\sigma_i\Big\}.\end{aligned} \tag{4.116}$$

Is now important to notice that the fourth term of Eq. (4.116), i.e. the second expected value, can be interpreted as the generalized pressure density function $\tilde{\alpha}_N(t,\boldsymbol{x})$ of a Hopfield network with K binary patterns $\{\tilde{\xi}^\nu\}$ and an external field $h_i = A\sum_i \eta_i$, with a generalized partition function

$$\tilde{Z}_N(t,\boldsymbol{x}) = \sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{tN}{2}\sum_{\nu=1}^K m_\nu^2 + N\sum_{\nu=1}^K x_\nu m_\nu + A\sum_{i=1}^N\eta_i\sigma_i\Big\},$$

in which we explicit the dependency on the Mattis magnetizations, the order parameter defined in Eq. (4.113). These observations lead to the choice of applying of the Hamilton-Jacobi formalism to solve the matter in hand, and we are now going to follow the technique's guidelines we have illustrated in the previous chapters to tackle this issue. For the sake of clearness, we shall summarize the main steps: we first check the properties of $\tilde{\alpha}_N(t,\boldsymbol{x})$'s derivatives and notice that we can define a Hamilton-Jacobi action that is equal to $\tilde{\alpha}_N(t,\boldsymbol{x})$ (but with the opposite sign). As a second step, we introduce a vector $\Gamma_N(t,x)$ depending on the generalized pressure and satisfying a Burgers-like equation. By performing a Cole-Hopf transform, we arrive to a heat equation which we solve using standard methods. The solution to this equation can be obtained via saddle point method and, being $\tilde{\alpha}_N(t,\boldsymbol{x})$ in a logarithmic relation with this function, so we can compute the thermodynamic limit for the pressure function and set the interpolating parameters to the ones recovering our model. Now let ys go into details.

It is simple to check that $\tilde{\alpha}_N(t,\boldsymbol{x})$ has the following properties:

$$\partial_t \tilde{\alpha}_N(t,\boldsymbol{x}) = \frac{1}{2}\sum_{\nu=1}^{K} \langle m_\nu^2 \rangle_{t,\boldsymbol{x}}, \quad \partial_{x_\nu} \tilde{\alpha}_N(t,\boldsymbol{x}) = \langle m_\nu \rangle_{t,\boldsymbol{x}}, \tag{4.117}$$

so we can now proceed according to the Hamilton-Jacobi formalism. In fact, if we consider $-\tilde{\alpha}_N(t,\boldsymbol{x})$, it is immediate to check that, thanks to the properties (4.117), we built a Hamilton-Jacobi action, hence

$$\partial_t\big(-\tilde{\alpha}_N(t,\boldsymbol{x})\big) + \frac{1}{2}\big(\partial_x \tilde{\alpha}_N(t,\boldsymbol{x})\big)^2 + V_N(t,x) = 0, \tag{4.118}$$

with a potential $V_N(t,\boldsymbol{x}) = \frac{1}{2}\sum_\nu \big(\langle m_\nu^2\rangle_{t,\boldsymbol{x}} - \langle m_\nu\rangle_{t,\boldsymbol{x}}^2\big) = \frac{1}{2N}\nabla_{\boldsymbol{x}}^2 \tilde{\alpha}_N(t,\boldsymbol{x})$. So, introducing the vector $\Gamma_N^\nu(t,\boldsymbol{x}) = -\partial_{x_\nu}\tilde{\alpha}_N(t,\boldsymbol{x})$, we obtain the Burgers equations by deriving equation (4.118) with respect to x_ν:

$$\partial_t \Gamma_N^\nu(t,\boldsymbol{x}) + \sum_{\tau=1}^{k} \Gamma_N^\tau(t,\boldsymbol{x}) \cdot \partial_{x_\tau}\Gamma_N^\nu(t,\boldsymbol{x}) = \frac{1}{2N}\sum_{\tau=1}^{k} \partial^2_{x_\tau x_\tau}\Gamma_N^\nu(t,\boldsymbol{x}), \quad \forall \nu. \tag{4.119}$$

Performing the Cole-Hopf transformation $\Phi_N(t,\boldsymbol{x}) := e^{N\tilde{\alpha}_N(t,\boldsymbol{x})}$, we can assert that solving (4.119) is equivalent to solve the Cauchy problem

$$\begin{cases} \partial\Phi_N(t,\boldsymbol{x}) - \frac{1}{2N}\Delta\Phi_N(t,\boldsymbol{x}) = 0, & t \in \mathbb{R}, x \in \mathbb{R}^K \\ \Phi_N(0,\boldsymbol{x}) = e^{N\tilde{\alpha}_N(0,\boldsymbol{x})}, & x \in \mathbb{R}^K. \end{cases} \tag{4.120}$$

We solve the problem above through the standard techniques we used in sections 2.5 and 3.6 for the CW and the SK models, therefore introducing the general solution as convolution of the initial data and the Green propagator G as

$$\Phi_N(t,\boldsymbol{x}) = \int_{\mathbb{R}^K} dx_1' \cdots dx_K' G(t,\boldsymbol{x}-\boldsymbol{x}')\Phi_N(0,\boldsymbol{x}'), \tag{4.121}$$

where $G(t,\boldsymbol{x}) = (\frac{N}{2\pi t})^{K/2} e^{-\frac{\sum_\nu x_\nu^2 N}{2t}}$. The computations for the initial condition

$\Phi_N(0,\boldsymbol{x})$ are not elaborate:

$$\begin{aligned}\Phi_N(0,\boldsymbol{x}) &= \exp\Big\{\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big\{\sum_i\sum_\nu x_\nu\tilde{\xi}_i^\nu\sigma_i + A\sum_i\eta_i\sigma_i\Big\}\Big\} = \\ &= \exp\Big\{\mathbb{E}\log\prod_{i=1}^N\sum_{\boldsymbol{\sigma}} e^{\left(\sum_\nu x_\nu\tilde{\xi}_i^\nu + A\eta_i\right)\sigma_i}\Big\} = \\ &= \exp\Big\{N\log 2 + \sum_{i=1}^N\mathbb{E}\log\cosh\Big(\sum_{\nu=1}^K\tilde{\xi}_i^\nu x_\nu + A\eta_i\Big)\Big\}.\end{aligned}$$

The solution to the problem (4.120) is therefore given by the following saddle point equation:

$$\Phi_N(t,\boldsymbol{x}) = \Big(\frac{N}{2\pi t}\Big)^{K/2}\int_{\mathbb{R}^K} dx_1'\cdots dx_K' e^{-Ng(t,\boldsymbol{x},\boldsymbol{x}')},$$

where

$$g(t,\boldsymbol{x},\boldsymbol{x}') = \frac{1}{2t}\sum_{\nu=1}^K(x_\nu - x_\nu')^2 - \log 2 - \frac{1}{N}\sum_{i=1}^N\mathbb{E}\log\cosh\Big(\sum_{\nu=1}^K\tilde{\xi}_i^\nu x_\nu' + A\eta_i\Big). \tag{4.122}$$

Recalling that $\tilde{\alpha}_N(t,\boldsymbol{x}) = \frac{1}{N}\log\Phi_N(t,\boldsymbol{x})$, when performing the thermodynamic limit (which is what we are ultimately interested in), we have that $\tilde{\alpha}(t,\boldsymbol{x})$ is determined by

$$\tilde{\alpha}(t,\boldsymbol{x}) = \lim_{N\to\infty}\tilde{\alpha}_N(t,\boldsymbol{x}) = -\min_{\boldsymbol{x}'\in\mathbb{R}^K} g(t,\boldsymbol{x},\boldsymbol{x}'). \tag{4.123}$$

Equation (4.115) is almost all explicit. What is now left to be calculated is the integral term, for which it is sufficient to calculate the ψ-derivative of the pressure function. The process is analogous with the one adopted in Section 4.7.1, and we shall not report all the steps of the calculations:

$$\begin{aligned}\partial_\psi\tilde{\alpha}_N(t,\boldsymbol{x}) =& \frac{1}{N}\mathbb{E}\frac{\partial_\psi Z_N(t,\boldsymbol{x},\psi)}{Z_N(t,\boldsymbol{x},\psi)} = \frac{1}{2N}(\beta - B^2 - C)\sum_{\mu=1}^P\mathbb{E}\omega(z_\mu^2)_{t,\boldsymbol{x}} + \\ & -\frac{\lambda\beta}{2}\langle q_{12}p_{12}\rangle_{t,\boldsymbol{x}} - \frac{A^2}{2}(1-\langle q_{12}\rangle_{t,\boldsymbol{x}}) + \frac{\lambda\beta^2}{2}\langle p_{12}\rangle_{t,\boldsymbol{x}}.\end{aligned} \tag{4.124}$$

Again we have that, in the replica symmetric *Ansatz*, the order parameters m, q_{12}, p_{12} do not fluctuate with respect to the quenched average, so let us define the only values that they can acquire: $\langle m\rangle_{t,\boldsymbol{x}} = \bar{m}$, $\langle q_{12}\rangle_{t,\boldsymbol{x}} = q$, $\langle p_{12}\rangle_{t,\boldsymbol{x}} = p$. Fixing the tunable parameters as

$$A = \sqrt{\lambda\beta p}, \qquad B = \sqrt{\beta q}, \qquad C = \beta(1-q),$$

and adding and subtracting the term $\frac{\lambda\beta}{2}qp$ in Eq. (4.124), we have

$$\partial_\psi \tilde{\alpha}_N(t, \boldsymbol{x}) = -\frac{\lambda\beta}{2}\langle (q_{12} - q)(p_{12} - p)\rangle_{t,\boldsymbol{x}} - \frac{\lambda\beta}{2}p(1-q). \tag{4.125}$$

Also in this case we point out that, even though in the definition of the overlaps q_{ab} and p_{ab} we showed that these quantities can take negative values, we will verify in the self-consistency equations that these quantities can only take non-negative values, without conflicts with the choice of the tunable parameters.

To get the final expression of the thermodynamic limit of the pressure density, we shall make the RS ansatz in (4.125), perform the $N \to \infty$ limit in both (4.125) and (4.116) and set $t = \beta$, $\boldsymbol{x} = 0$. Therefore, performing the minimization of the function (4.122), with the values of t and $\boldsymbol{x}$ that we fixed just now, we have to set $x'_\nu = \beta\bar{m}_\nu$ $\forall \nu = 1, \dots, K$. We can finally conclude the section with the following theorem.

Theorem 4.8. *The replica symmetric thermodynamic limit of the free energy density of the hybrid Hopfield neural network with N Ising spins $\sigma_i \in \{-1, +1\}$ $\forall i = 1, \dots, N$, a low load (i.e. $K \sim \gamma \log N$) of binary patterns and a high load (i.e. $P \sim \lambda N$) real patters, described by the Hamiltonian* (4.110) *is determined by the minimum value of*

$$f(\beta, \lambda) = -\frac{1}{\beta}\alpha(\beta, \lambda),$$

where

$$\begin{aligned}\alpha(\beta,\lambda) = &-\frac{\lambda\beta}{2} - \frac{\lambda}{2}\log\big(1 - \beta(1-q)\big) + \frac{\lambda\beta q}{2\big(1-\beta(1-q)\big)} - \frac{\beta}{2}\sum_\nu m_\nu^2 + \\ &+ \int_{-\infty}^{+\infty} d\mu(\eta)\mathbb{E}\log 2\cosh(\beta\sum_\nu \tilde{\xi}^\nu m_\nu + \sqrt{\lambda\beta p}\eta) - \frac{\lambda\beta}{2}p(1-q),\end{aligned} \tag{4.126}$$

with $\mathbb{E}$ being the average over the binary patterns and $\eta \sim \mathcal{N}(0,1)$.

Also in this case, we dropped the bar over m to make notation coherent with other Chapters, but we remind that it should be always intended as the thermodynamic limit of Mattis magnetization in the replica symmetry assumption.

Remark 4.12. Like for the previous models, we highlight the fact that, for $\lambda = 0$ and $\nu = 1$, we gain the CW pressure density function. If $\lambda > 0$ and $\nu = 1$, we recover the equation for the basic case of one boolean pattern and a high load of analogical memories (4.103). Finally, if $\lambda = 0$ we precisely recover the CW expression for the quenched pressure.

As always, we conclude this Section by deriving the self-consistency for the order parameters minimizing the free energy, or equivalently maximizing the statistical pressure. The derivative expressions are the following:

$$\frac{\partial \alpha}{\partial q} = \frac{\lambda\beta}{2}\Big(- \frac{\beta q}{\big(1-\beta(1-q)\big)^2}\Big) = 0,$$
$$\frac{\partial \alpha}{\partial p} = -\frac{\lambda\beta}{2}(1-q) + \int_{-\infty}^{+\infty} d\mu(\eta)\mathbb{E}\Big(\tanh\Big(\beta\sum_\nu \tilde{\xi}^\nu m_\nu + \sqrt{\lambda\beta p}\eta\Big)\frac{\eta}{2}\sqrt{\frac{\lambda\beta}{p}}\Big) = 0,$$
$$\frac{\partial \alpha}{\partial m_\nu} = \beta m_\nu - \int_{-\infty}^{+\infty} d\mu(\eta)\mathbb{E}\Big(\tanh\Big(\beta\sum_\nu \tilde{\xi}^\nu m_\nu + \sqrt{\lambda\beta p}\eta\Big)\beta\tilde{\xi}^\nu\Big) = 0.$$

Applying an integration by part to the second equation, we have

$$m_\nu = \int_{-\infty}^{+\infty} d\mu(\eta)\mathbb{E}\,\tilde{\xi}^\nu \tanh\Big(\beta\sum_{\nu=1}^{K} \tilde{\xi}^\nu m_\nu + \sqrt{\lambda\beta p}\eta\Big), \tag{4.127}$$
$$q = \int_{-\infty}^{+\infty} d\mu(\eta)\mathbb{E}\tanh^2\Big(\beta\sum_{\nu=1}^{K} \tilde{\xi}^\nu m_\nu + \sqrt{\lambda\beta p}\eta\Big), \tag{4.128}$$
$$p = \frac{\beta q}{\big(1-\beta(1-q)\big)^2}. \tag{4.129}$$

Again, the SG-paramagnetic transition is of the second order with a critical exponent equal to 1, while the paramagnetic-ferromagnetic transition is of the first degree with a critical exponent equal to 1/2. While the SG-paramagnetic line can be found analytically, to be able to detect the transition from the retrieval phase to the SG one we need computer calculations. Considering expression (4.128), one can approximate the square of the hyperbolic tangent with its argument and then use the Taylor expansion of the latter for $q \sim 0$, thus having:

$$q \simeq \frac{\beta^2\lambda}{(1-\beta)^2}q + \mathcal{O}(q^2).$$

Then, the transition line is given by equation

$$\frac{\beta^2\lambda}{(1-\beta)^2} = 1 \quad \Leftrightarrow \quad T = 1 + \sqrt{\lambda}.$$

For the numerical calculations we performed the *pure state* ansatz, meaning that we assumed that we can eventually retrieve one pattern at the time, thus having only one possible $m_\nu \neq 0$. The program calculates the values of $\alpha(\beta, \lambda)$ in many different points of the $\{\lambda, \beta\}$ plane, and plots the points where the minimizing value of the order paramenter m_ν goes from being non vanishing (the retrieval phase) to being equal to zero (the SG phase). What we obtained is the scheme reported in 4.5.

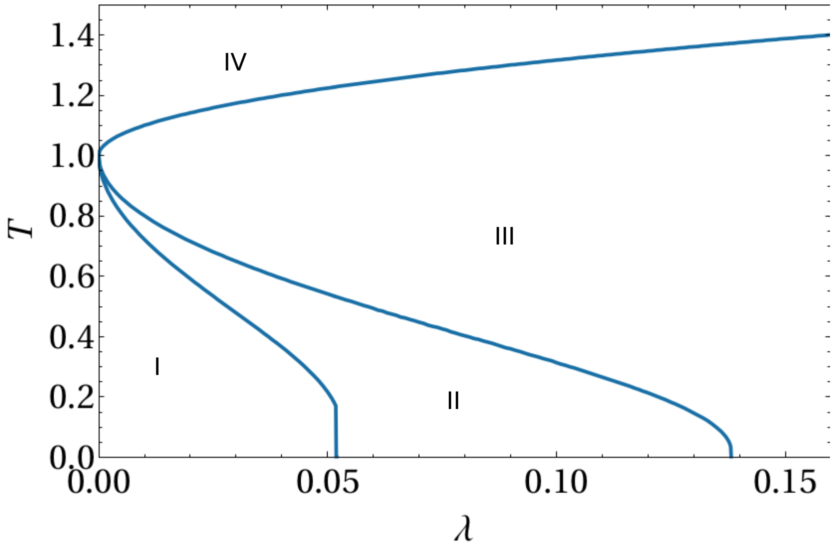

Figure 4.5: **Phase diagram for the hybrid Hopfield neural network**. From the bottom to the top: pure retrieval region (I), in which pure states are global minima for the free energy; mixed retrieval region (II), in which pure states are only local minima for the free energy; spin-glass (III) and ergodic (IV) regions.

The $\{\lambda, T\}$ plane is divided into three regions. The one above the green (SG-paramagnetic) line is where the system behaves as a paramagnet because the noise created by the temperature is too high and the neurons can only behave randomly, and the average magnetization and the average overlap are both null. This line is determined analytically with the calculations we performed above. The points on the SG-paramagnetic line represent the values (λ, T) such that the order parameter q goes from being positive to zero, hence presenting a (second order) phase transition.

Between the green and the blue line we recover a spin-glass behaviour, with the average magnetization equal to zero and a positive average overlap.

Finally, between the plane axes and the blue line, we have the retrieval region. Here, we have a positive value for the average magnetization and for the average overlap. The blue line has been traced thanks to a numerical simulation that locates the (λ, T) points where we have a (first order) phase transition for the magnetization, which goes from being null (spin-glass and paramagnetic behaviour) to noon vanishing (as the temperature decreases).

From this plot we can conclude that the hybrid Hopfield neural network with a high storage of real patterns presents a retrieval phase of the boolean patterns memorized in a low load. Furthermore, we notice how the SG-paramagnetic line is determined by the same equation as the previous basic model and as the one belonging to the classic Hopfield neural network [37].

Chapter 5

Learning phase of AI: the Boltzmann machine

Until now, we focused on the retrieval capabilities of neural networks, but before we can be able to retrieve a pattern of information, we have to learn it. This is the field of the branch of AI termed *machine learning.* The aim of this Chapter is to give the basic definition of a learning algorithm (giving two key examples, namely linear regression and the perceptron training), in order to define the archetype of a learning machine, namely the Restricted Boltzmann Machine, coupled with its learning rule (namely the Hinton's *contrastive divergence*), and link the learning capabilities of this machine with the retrieval skills of the Hopfield network by proving that the two models, ultimately, share the same quenched free energy (and thus the same phase diagram).

Generally speaking, for a (Boltzmann) machine to be able to learn, it must have (at least) a visible layer (where a set of binary data vectors can be presented) and an hidden layer to process the information provided to the visible layer (also another output layer is often considered, however the *minimal model* can still be thought of as an elementary two-party network). In order to learn, the machine must rearrange its internal connections (e.g. weights), ultimately mimicking synaptic dynamics rather than neural one (that is due to information retrieval as we largely saw at this point).

5.1 Generalities

A machine learning algorithm is an algorithm that is able to learn from data. But what do we mean by learning? T. M. Mitchell provides the following definition: "*A computer program is said to learn from experience*

$\mathcal{E}$with respect to some class of tasks $\mathcal{T}$and performance measure $\mathcal{P}$, if its performance at tasks in $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$."

One can imagine a very wide variety of experiences $\mathcal{E}$, tasks $\mathcal{T}$, and performance measures $\mathcal{P}$, and we do not make any attempt in these thesis to provide a formal definition of what may be used for each of these entities. Instead, this Section provides intuitive descriptions and examples of the different kinds of tasks, performance measures and experiences that can be used to construct machine learning algorithms.

The task, $\mathcal{T}$

The process of learning itself is *not* the task, as it is solely our means of attaining the ability to perform the task. For example, if we want a robot to be able to walk, then walking is the task. We could program the robot to learn to walk, or we could attempt to directly write a program that specifies how to walk manually.

Machine learning tasks are usually described in terms of how the machine learning system should process an *example*. An *example* is a collection of features that have been quantitatively measured from some object or event that we want the machine learning system to process. We typically represent an example as a vector $\boldsymbol{x} \in \mathbb{R}^n$ where each entry x_i of the vector is a *feature*. For example, the features of an image are usually the values of the pixels in the image.

Many kinds of tasks can be solved with machine learning. Some of the most common machine learning tasks include the following:

- *Classification*: In this type of task, the computer program is asked to specify which of k categories some input belongs to. To solve this task, the learning algorithm is usually asked to produce a function $f : \mathbb{R}^n \to \{1, \dots, k\}$. When $y = f(\boldsymbol{x})$, the model assigns an input described by vector $\boldsymbol{x}$ to a category identified by numeric code y. There are other variants of the classification task, for example, where f outputs a probability distribution over classes. An example of a classification task is object recognition, where the input is an image (usually described as a set of pixel brightness values), and the output is a numeric code identifying the object in the image. Modern object recognition is best accomplished with deep learning, the ultimate frontier of machine learning (we won't address it directly in this thesis, so we refer to [57, 114, 120]). Object recognition is the same basic technology that allows computers to recognize faces.

- *Classification with missing inputs*: Classification becomes more challenging if the computer program is not guaranteed that every measurement in its input vector will always be provided. In order to solve the classification task, the learning algorithm only has to define a single function mapping from a vector input to a categorical output. When some of the inputs may be missing, rather than providing a single classification function, the learning algorithm must learn a set of functions. Each function corresponds to classifying $\boldsymbol{x}$ with a different subset of its inputs missing. This kind of situation arises frequently in medical diagnosis, because many kinds of medical tests are expensive or invasive. One way to efficiently define such a large set of functions is to learn a probability distribution over all of the relevant variables, then solve the classification task by marginalizing out the missing variables. With n input variables, we can now obtain all 2^n different classification functions needed for each possible set of missing inputs, but we only need to learn a single function describing the joint probability distribution. Many of the other tasks described in this Section can also be generalized to work with missing inputs.

- *Regression*: In this type of task, the computer program is asked to predict a numerical value given some input. To solve this task, the learning algorithm is asked to output a function $f : \mathbb{R}^n \to \mathbb{R}$. This type of task is similar to classification, except that the format of output is different. An example of a regression task is the prediction of the expected claim amount that an insured person will make (used to set insurance premiums), or the prediction of future prices of securities. These kinds of predictions are also used for algorithmic trading.

- *Transcription*: In this type of task, the machine learning system is asked to observe a relatively unstructured representation of some kind of data and transcribe it into discrete, textual form. For example, in optical character recognition, the computer program is shown a photograph containing an image of text and is asked to return this text in the form of a sequence of characters (e.g. ASCII format). Google Street View uses deep learning to process address numbers in this way. Another example is speech recognition, where the computer program is provided an audio wave form and emits a sequence of characters or word ID codes describing the words that were spoken in the audio recording. Deep learning is a crucial component of modern speech recognition systems used at major companies including Microsoft, IBM and Google.

- *Density estimation or probability mass function estimation*: In the density estimation problem, the machine learning algorithm is asked to learn a function $\mathcal{P}_{model} : \mathbb{R}^n \to \mathbb{R}$, where $\mathcal{P}_{model}(\boldsymbol{x})$ can be interpreted as a probability density function (if $\boldsymbol{x}$ is continuous) or a probability mass function (if $\boldsymbol{x}$ is discrete) on the space that the examples were drawn from. To do such a task well (we will specify exactly what that means when we discuss performance measures $\mathcal{P}$), the algorithm needs to learn the structure of the data it has seen. It must know where examples cluster tightly and where they are unlikely to occur. Most of the tasks described above require the learning algorithm to at least implicitly capture the structure of the probability distribution. Density estimation allows us to explicitly capture it. In principle, we can then perform computations on that distribution in order to solve the other tasks as well. In practice, density estimation does not always allow us to solve all of these related tasks, because in many cases the required operations on $\mathcal{P}(x)$ are computationally intractable.

Of course, many other tasks and types of tasks are possible. The types of tasks we list here are intended only to provide examples of what machine learning can do, not to define a rigid taxonomy.

5.1.1 The performance measure, $\mathcal{P}$

In order to evaluate the abilities of a machine learning algorithm, we must design a quantitative measure of its performances. Usually, this performance measure $\mathcal{P}$ is specific to the task $\mathcal{T}$ being carried out by the system. For tasks such as classification, classification with missing inputs, we often measure the accuracy of the model, which is just the proportion of examples for which the model produces the correct output. We can also obtain equivalent information by measuring the error rate, the proportion of examples for which the model produces an incorrect output. We often refer to the error rate as the expected $0-1$ loss. The $0-1$ loss on a particular example is 0 if it is correctly classified and 1 if it is not. On the other hand, for tasks such as density estimation, it does not make sense to measure accuracy, error rate, or any other kind of $0-1$ loss. Instead, we must use a different performance metric that gives the model a continuous-valued score for each example. The most common approach is to report the average log-probability the model assigns to some examples.

Usually, we are interested in how well the machine learning algorithm performs on data that it has not seen before, since this determines how well it will work when deployed in the real world. We therefore evaluate these performance measures using a test set of data that is separate from the data used for training the machine learning system. The choice of performance measure may seem straightforward and objective, but it is often difficult to choose one well-describing the desired behavior of the system. In some cases, this is because it is difficult to decide what should be measured. For example, when performing a transcription task, should we measure the accuracy of the system at transcribing entire sequences, or should we use a more fine-grained performance measure that gives partial credit for getting some elements of the sequence correct? When performing a regression task, should we penalize the system more if it frequently makes medium-sized mistakes or if it rarely makes very large ones? Clearly, these kinds of design choices depend on the application. In other cases, we know what quantity we would ideally like to measure, but measuring it is impractical. For example, such kind of difficulties arises frequently in the context of density estimation. Many of the best probabilistic models represent probability distributions only implicitly. Computing the actual probability value assigned to a specific point in space is intractable in many such models. In these cases, one must design an alternative criterion that still corresponds to the define objectives, or design a good approximation to the desired criterion.

5.1.2 The experience, $\mathcal{E}$

Machine learning algorithms can be broadly categorized as *unsupervised* or *supervised* by what kind of experience they are allowed to have during the learning process. Most of the learning algorithms can be understood as being allowed to experience an entire *dataset*. A *dataset* is a collection of many examples, as defined earlier. *Unsupervised* learning algorithms experience a dataset containing many features, then learn useful properties of the structure of this dataset. In the context of deep learning, we usually want to learn the entire probability distribution that generated a dataset, whether explicitly as in density estimation or implicitly for tasks like synthesis or de-noising. Some other unsupervised learning algorithms perform other roles, like clustering, which consists of dividing the dataset into clusters of similar examples. *Supervised* learning algorithms experience a dataset containing features, but each example is also associated with a label or target. Roughly speaking, unsupervised learning involves observing several examples of a random vector $\boldsymbol{x}$, and attempting to (implicitly or explicitly) learn the probability distribution $\mathcal{P}(\boldsymbol{x})$ or some of its interesting properties, while supervised learning involves observing several examples of a random vector $\boldsymbol{x}$ and an associated value or vector $\boldsymbol{y}$, and learning to predict $\boldsymbol{y}$ from $\boldsymbol{x}$ by estimating $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$. The term supervised learning originates from the view of the target $\boldsymbol{y}$ being provided by an instructor or teacher who shows the machine learning system what to do. In unsupervised learning, there is no instructor or teacher, and the algorithm must learn to make sense of the data without this guide.

Unsupervised learning and supervised learning are not formally defined terms. The lines between them are often blurred. Many machine learning technologies can be used to perform both tasks. For example, the chain rule of probability states that for a vector $\boldsymbol{x} \in \mathbb{R}^n$, the joint distribution can be decomposed as

$$\mathcal{P}(\boldsymbol{x}) = \prod_{i=1}^{n} \mathcal{P}\big(x_i \mid x_1, \dots, x_{i-1}\big).$$

This decomposition means that we can solve the ostensibly unsupervised problem of modeling $\mathcal{P}(\boldsymbol{x})$ by splitting it into n supervised learning problems. Alternatively, we can solve the supervised learning problem of learning $\mathcal{P}\big(\boldsymbol{y}|\boldsymbol{x}\big)$ by using traditional unsupervised learning technologies to learn the joint distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$, and then inferring

$$\mathcal{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}'} \mathcal{P}(\boldsymbol{x}, \boldsymbol{y}')}.$$

Though unsupervised learning and supervised learning are not completely formal or distinct concepts, they do help to roughly categorize some of the things we do with machine learning algorithms. Traditionally, people refer to regression and classification problems as supervised learning. Density estimation in support of other tasks is usually considered unsupervised learning. Other variants of the learning paradigm are possible. For example, in semi-supervised learning, some examples include a supervision target but others do not.

Most machine learning algorithms simply experience a dataset. A dataset can be described in many ways. In all cases, a dataset is a collection of examples, which are in turn collections of features. One common way of describing a dataset is with a *design matrix* containing a different example in each row, while each column of the matrix corresponds to a different feature. Of course, to describe a dataset as a design matrix, it must be possible to describe each example as a vector, and each of these vectors must be the same size. This is not always possible. For example, if you have a collection of photographs with different widths and heights, then different photographs will contain different numbers of pixels, so not all of the photographs may be described with the same length of vector. In such cases, rather than describing the dataset as a matrix with n rows, we will describe it as a set containing n elements: $\{x(1), x(2), ..., x(n)\}$. This notation does not imply that any two example vectors $x(i)$ and $x(j)$ have the same size. In the case of supervised learning, the example contains a label or target as well as a collection of features. For example, if we want to use a learning algorithm to perform object recognition from photographs, we need to specify which object appears in each of the photos. We might do this with a numeric code, with 0 signifying a person, 1 signifying a car, 2 signifying a cat, etc.

Often, when working with a dataset containing a design matrix of feature observations X, we also provide a vector of labels $\boldsymbol{y}$, with y_i providing the label for example i. Of course, sometimes the label may be more than just a single number. For example, if we want to train a speech recognition system to transcribe entire sentences, then the label for each example sentence is a sequence of words. Just as there is no formal definition of supervised and unsupervised learning,there is no rigid taxonomy of datasets or experiences.

5.1.3 An example of learning algorithm: linear regression

Our definition of a machine learning algorithm as an algorithm that is capable of improving the performance of a computer program at some task via experience is somewhat abstract. To make this more concrete, we present an example of a simple machine learning algorithm that we all studied - actually quite similarly to machines - when we were first-year students in Academy: *linear regression.* The goal of the algorithm is to build a system that can take a vector $\boldsymbol{x} \in \mathbb{R}^n$ as input and predict the value of a scalar $y \in \mathbb{R}$ as its output. In the case of linear regression, the output is a linear function of the input. Let $\hat{y}$ be the value that our model predicts y should take on. We define the output to be

$$\hat{y} = \boldsymbol{w}^T \cdot \boldsymbol{x},$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is a vector of *parameters.*

Remark 5.1. Preserving a machine learning notation, we have used the letter $\boldsymbol{w}$ to refer to parameter vector, i.e. the *weight* vector. The choice of swapping $\boldsymbol{w}$ with the letter $\boldsymbol{\xi}$, as we will do in the next Section, is made when we want to underline the correspondence between the Boltzmann machines' weights $\boldsymbol{w}$ and the Hopfield networks patterns $\boldsymbol{\xi}$.

Parameters are values that control the behavior of the system. In this case, w_i is the coefficient that we multiply by feature x_i before summing up the contributions from all the features. We can think of $\boldsymbol{w}$ as a set of weights that determine how each feature affects the prediction. If a feature x_i receives a positive weight w_i, then increasing the value of that feature increases the value of our prediction $\hat{y}$. On the contrary, if a feature receives a negative weight, then increasing the value of that feature decreases the value of our prediction. If a feature weight is large in magnitude, then it has a large effect on the prediction. Otherwise, if a feature weight is zero, it has no effect on the prediction.

We thus have a definition of our task $\mathcal{T}$: to predict y from $\boldsymbol{x}$ by outputting $\hat{y} = \boldsymbol{w}^T\boldsymbol{x}$. Next, we need a definition of our performance measure $\mathcal{P}$. Suppose that we have a design matrix of n example inputs that we will not use for training, only for evaluating how well the model performs. We also have a vector of regression targets providing the correct value of y for each of these examples. Because this dataset will only be used for evaluation, we call it the *test set*. We refer to the design matrix of inputs as $X^{(test)}$ and the vector of regression targets as $y^{(test)}$. One way of measuring the performance of the model is to compute the mean squared error of the model on the test set.

If $\hat{y}^{(test)}$ gives the predictions of the model on the test set, then the mean squared error is given by

$$\mathrm{MSE}_{test} = \frac{1}{n}\sum_i \left(\hat{y}^{(test)} - y^{(test)}\right)_i^2.$$

Intuitively, one can see that this error measure decreases to 0 when $\hat{y}^{(test)} = y^{(test)}$. Alternatively, we can see that

$$\mathrm{MSE}_{test} = \frac{1}{n}\|\hat{y}^{(test)} - y^{(test)}\|_2^2,$$

so the error increases whenever the Euclidean distance between the predictions and the targets increases. To make a machine learning algorithm, we need to design an algorithm that will improve the weights $\boldsymbol{w}$ such that it reduces MSE_{test} when the algorithm is allowed to gain experience by observing a training set $(X^{(train)}, y^{(train)})$. One intuitive way of doing this is just to minimize the mean squared error on the training set, MSE_{train}.
To minimize MSE_{train}, we can simply solve for where its gradient is 0:

$$\begin{aligned} \nabla_{\boldsymbol{w}}\mathrm{MSE}_{train} = 0 \quad &\Rightarrow \quad \frac{1}{n}\nabla_{\boldsymbol{w}}\|\hat{y}^{(train)} - y^{(train)}\|_2^2 = 0, \\ \Rightarrow \quad &\nabla_{\boldsymbol{w}}\left(X^{(train)}\boldsymbol{w} - y^{(train)}\right)^T\left(X^{(train)}\boldsymbol{w} - y^{(train)}\right) = 0, \\ \Rightarrow \quad &2X^{(train)T}X^{(train)}\boldsymbol{w} - 2X^{(train)T}y^{(train)} = 0, \end{aligned}$$

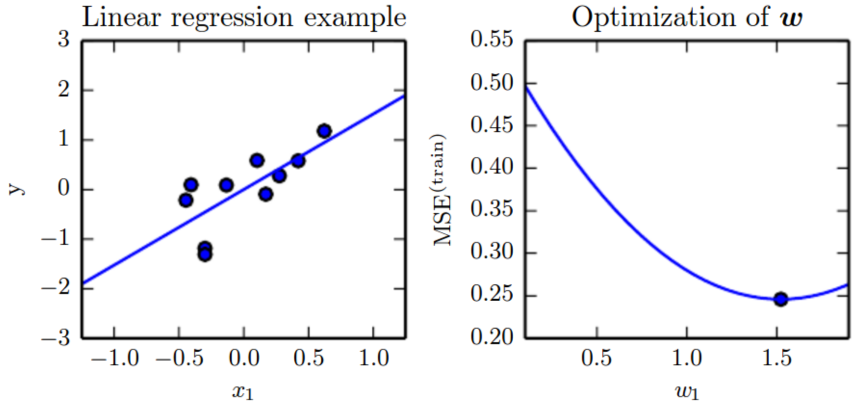

Figure 5.1: **A linear regression problem** for a training set consisting of ten data points, each containing one feature. Because there is only one feature, the weight vector w contains only a single parameter to learn, w_1. Left: bserve in the $\{x, y\}$ plane that linear regression learns to set w_1 such that the line $y = w_1 x$ comes as close as possible to passing through all the training points. Right: the plotted point in the $\{w_1, \text{MSE}_{train}\}$ plane indicates the value of w_1 found by the normal equations, which we can see minimizes the mean squared error on the training set.

which leads to

$$\boldsymbol{w} = \left(X^{(train)T} X^{(train)}\right)^{-1} X^{(train)T} y^{(train)}. \tag{5.1}$$

The system of equations, whose solution is given by equation (5.1), is known as the *normal equations*.

Evaluating equation (5.1) constitutes a simple learning algorithm. For an example of the linear regression learning algorithm in action, see Fig. 5.1. It is worth noting that the term linear regression is often used to refer to a slightly more sophisticated model with one additional parameter - an intercept term b. In this model, $\hat{y} = \boldsymbol{w}^T \cdot \boldsymbol{x} + b$, so the mapping from parameters to predictions is still a linear function but the mapping from features to predictions is now an affine function. This extension means that the plot of the model's predictions still looks like a line, but it need not pass through the origin. Instead of adding the bias parameter b, one can continue to use the model with only weights but augment $\boldsymbol{x}$ with an extra entry that is always set to 1. The weight corresponding to the extra 1 entry plays the role of the bias parameter. The intercept term b is often called the *bias parameter* of the affine transformation. This terminology derives from the point of view that the output of the transformation is biased toward being b in the absence of any input. This term is different from the idea of a statistical bias, in which a statistical estimation algorithm's expected estimate of a quantity is not equal to the true quantity.

Linear regression is of course an extremely simple and limited learning algorithm, but it provides an example of how a learning algorithm can work.

5.1.4 Capicity, overfitting and underfitting

The central challenge in machine learning is that we must perform well on new, previously unseen inputs — not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called *generalization*. Typically, when training a machine learning model, we have access to a training set, we can compute some error measure on the training set called the *training error*, and we reduce this training error. So far, what we have described is simply an optimization problem. What separates machine learning from simple optimization is that we want the generalization error, also called the *test error*, to be low as well. The generalization error is defined as the expected value of the error on a new input, where the expectation is taken across different possible inputs, drawn from the distribution of inputs we expect the system to encounter in practice. Typically, one estimates the generalization error of a machine learning model by measuring its performance on a test set of examples that were collected separately from the training set.

In our linear regression example, we trained the model by minimizing the training error

$$\frac{1}{n^{(train)}}||X^{(train)}\boldsymbol{w} - y^{(train)}||_2^2,$$

but we actually should also care about the test error

$$\frac{1}{n^{(test)}}||X^{(test)}\boldsymbol{w} - y^{(test)}||_2^2.$$

How can we affect performance on the test set when we get to observe only the training set? The field of statistical learning theory provides some answers. If the training and the test set are collected arbitrarily, there is indeed little we can do. If instead we are allowed to make some assumptions about how the training and test set are collected, then we can make some progress. The train and test data are generated by a probability distribution over datasets called the *data generating process*. We typically make a set of assumptions known collectively as the *i.i.d. assumptions*. These assumptions are that the examples in each dataset are independent from each other, and that the train set and test set are identically distributed, drawn from the same probability distribution as each other. This assumption allows us to describe the data generating process with a probability distribution over a single example. The same distribution is then used to generate every train example and every test example. We call that shared underlying distribution the *data generating distribution*, denoted $\mathcal{P}_{data}$. This probabilistic framework and the i.i.d. assumptions allow us to mathematically study the relationship between training error and test error. As a primary consequence, we can observe between the training and test error is that the expected training error of a randomly selected model is equal to the expected test error of that model.

Suppose we have a probability distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{y})$ and we sample from it repeatedly to generate the train set and the test set. For some fixed value $\boldsymbol{w}$, the expected training set error is exactly the same as the expected test set error, because both expectations are formed using the same dataset sampling process. The only difference between the two conditions is the name we assign to the dataset we sample. Of course, when we use a machine learning algorithm, we do not fix the parameters ahead of time, then sample both datasets. We sample the training set, then use it to choose the parameters to reduce training set error, then sample the test set. Under this process, the expected test error is greater than or equal to the expected value of training error. The factors determining how well a machine learning algorithm will perform are its ability to:

1. Make the training error small;

2. Make the gap between training and test error small.

These two factors correspond to the two central challenges in machine learning: *underfitting* and *overfitting. Underfitting* occurs when the model is not able to obtain a sufficiently low error value on the training set. *Overfitting* occurs when the gap between the training error and test error is too large. We can control whether a model is more likely to overfit or underfit by altering its *capacity.* Informally, the *capacity* of a model is its ability to fit a wide variety of functions. Models with low capacity may struggle to fit the training set. Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set.

One way to control the capacity of a learning algorithm is by choosing its *hypothesis space*, the set of functions that the learning algorithm is allowed to select as being the solution. For example, for the linear regression algorithm it is the set of all linear functions. We can relax this requirements by allowing our model to include polynomials, rather than just linear functions, in the hypothesis space. Doing so, the model capacity increases. A polynomial of degree one gives us the linear regression model with which we are already familiar, with prediction $\hat{y} = b + wx$. By introducing x^2 as another feature, we can learn a model that is quadratic as a function of x:

$$\hat{y} = b + w_1 x + w_2 x^2.$$

Though this model implements a quadratic function of its input, the output is still a linear function of the parameters, so we can still use the normal equations to train the model in closed form. We can continue to add more powers of x as additional features, for example to obtain a polynomial of degree p:

$$\hat{y} = b + \sum_{i=1}^{p} w_i x^i.$$

Machine learning algorithms will generally perform best when their capacity is appropriate for the true complexity of the task they need to perform and the amount of training data they are provided with. Models with insufficient capacity are unable to solve complex tasks. On the other hand, models with high capacity can solve complex tasks, but when their capacity is higher than needed to solve the present task they may overfit. Fig. 5.2 shows this principle in action. We compare a linear, quadratic and degree-9 predictor attempting to fit a problem where the true underlying function is quadratic. The linear function is unable to capture the curvature in the true underlying problem, so it underfits. The degree-9 predictor is capable of representing the correct function, but it is also capable of representing infinitely many other functions that pass exactly through the training points, because we have more parameters than training examples. We have little chance of choosing a solution that generalizes well when so many wildly different solutions exist. In this example, the quadratic model is perfectly matched to the true structure of the task so it generalizes well to new data.

So far, we described only one way of changing the model capacity, i.e. the number of input features it has, and simultaneously adding new parameters associated with those features. There are in fact many ways of tuning it. Capacity is not only determined by the choice of model. Indeed, the specifies which family of functions the learning algorithm can choose from when varying the parameters in order to reduce a training objective. This is called the *representational capacity* of the model. In many cases, finding the best function within this family is a very difficult optimization problem. In practice,

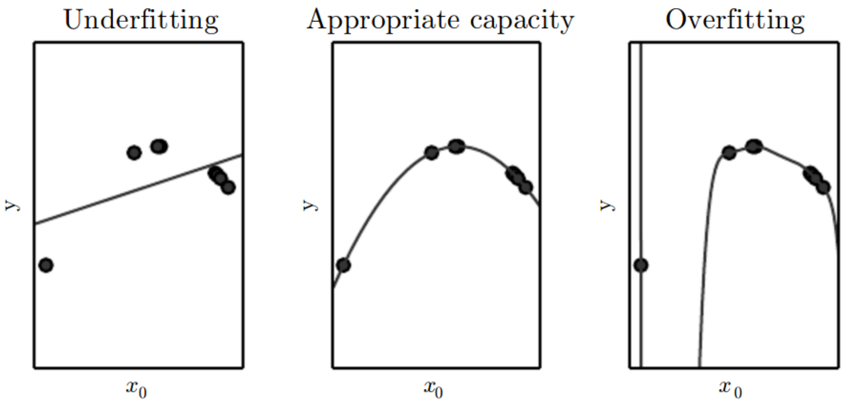

Figure 5.2: **Example of underfitting and overfitting issues**. We fit three models to this example training set. The training data was generated synthetically, by randomly sampling x values and choosing y deterministically by evaluating a quadratic function. Left: a linear function fit to the data suffers from underfitting-it cannot capture the curvature that is present in the data. Center: a quadratic function fit to the data generalizes well to unseen points. It does not suffer from a significant amount of overfitting or underfitting. Right: a polynomial of degree 9 fit to the data suffers from overfitting. The solution passes through all of the training points exactly, but it doesn't extract the correct structure. It now has a deep valley in between two training points that does not appear in the true underlying function. It also increases sharply on the left side of the data, while the true function decreases in this area.

the learning algorithm does not actually find the best function, but merely one that significantly reduces the training error. These additional limitations, such as the imperfection of the optimization algorithm, mean that the learning algorithm effective capacity may be less than the representational capacity of the model family.

Our modern ideas about improving the generalization of machine learning models are refinements of thought dating back to philosophers at least as early as Ptolemy. Many early scholars invoke a principle of parsimony that is now most widely known as *Occam's razor* (c. 1287-1347). This principle states that among competing hypotheses explaining known observations equally well, one should choose the "simplest" one. This idea was formalized (and made more precise) in the 20th century by the founders of statistical learning theory, but we will not deepen this argument for it goes beyond the purpose of this thesis.

Basically, while simpler functions are more likely to generalize (to have a small gap between training and test error), we must still choose a sufficiently complex hypothesis to achieve low training error. In practical applications, training error decreases until it asymptotes to the minimum possible error value as model capacity increases (assuming the error measure has a minimum value). Typically, generalization error has a U-shaped curve as a function of model capacity.

5.2 The Perceptron learning rule

A perceptron [110, 94] is a neural system composed by N input Boolean spins performing a weighted sum of the signals. The answer function $S : \{-1,1\}^N \to \{-1,1\}$ is therefore $S(\boldsymbol{\sigma}) = \mathrm{sign}(\sum_i J_i \sigma_i - h)$, where J_i are the weights and h is the threshold.[1] If we want to train the network in order to encode a *teacher function* $T(\boldsymbol{\sigma})$, we should invoke the Perceptron Convergence Theorem [37]:

Theorem 5.1. *If the teacher function $T(\boldsymbol{\sigma})$ is linearly separable,[2] then the update rule*

$$\begin{aligned} \Delta \boldsymbol{J} &= \frac{1}{2}[T(\boldsymbol{\sigma}) - sign(\boldsymbol{J} \cdot \boldsymbol{\sigma})]\boldsymbol{\sigma}, \\ \Delta h &= T(\boldsymbol{\sigma}), \end{aligned} \tag{5.2}$$

with randomly chosen $\boldsymbol{\sigma}$, will converge to $S(\boldsymbol{\sigma}) = T(\boldsymbol{\sigma})$ in a finite number of training steps.

An instructive point in this theorem is its continuous version, which is obtained by putting a learning strength $\epsilon \ll 1$ (defining a temporal scale in which learning is effective) in front of the r.h.s. of equation (5.2) and taking the $\epsilon \to 0$ limit. In this case, the perceptron learning rule acquires the suggestive form

$$\frac{d\boldsymbol{J}}{dt} = \langle [T(\boldsymbol{\sigma}) - \mathrm{sign}(\boldsymbol{J} \cdot \boldsymbol{\sigma})]\boldsymbol{\sigma} \rangle = -\frac{\partial \rho}{\partial \boldsymbol{J}}, \tag{5.3}$$

where $\rho(\boldsymbol{J}) = \langle (\boldsymbol{J} \cdot \boldsymbol{\sigma})[\mathrm{sign}(\boldsymbol{J} \cdot \boldsymbol{\sigma}) - T(\boldsymbol{\sigma})] \rangle$. Equation (5.3) is a *gradient descent* algorithm, and leads to a fixed point, i.e. to a (local) minima for

[1] For a statistical mechanics approach to binary perceptrons, see e.g. [70].

[2] With linear separability, we mean that inputs corresponding to different teacher function values can be separated by an (hyper-)plane in the configuration space.

the error function ρ. With such a procedure, the error function decreases monotonically with the learning time, as it can be easily checked that

$$\frac{d\rho}{dt} = -\sum_i \left(\frac{\partial \rho}{\partial J_i}\right)^2 \leq 0, \tag{5.4}$$

until a fixed point is reached.

Gradient descent methods (and their stochastic generalizations) are commonly used in optimization problems, since they aim to find minima for a general error measure. As we previously discussed, in realistic cases one should reconstruct the teacher function from a limited number of input/output relations, i.e.

$$T_s = T(\boldsymbol{\sigma}_s) \qquad s = 1, \dots, n, \tag{5.5}$$

where T_s are the teacher answers relative to the inputs $\boldsymbol{\sigma}_s$. Moreover, also the teacher's data could present some degree of noise or errors. The number of available teacher input/output relations is of course less than the full space dimensionality (which in the Boolean case is of course 2^N). On the spirit of the above discussion, one can introduce two possible error measure

$$\begin{aligned} \rho_T &= \frac{1}{n}\sum_s \Delta(T_s, S(\boldsymbol{\sigma_s})), \\ \rho_G &= \sum_{\boldsymbol{\sigma}} \Delta(T(\boldsymbol{\sigma}), S(\boldsymbol{\sigma})), \end{aligned} \tag{5.6}$$

where Δ quantifies the difference between the teacher and student answers. Recall that the function S depends on the network parameters $\boldsymbol{J}$ and $\boldsymbol{b}$, which we call in compact form $\boldsymbol{w}$. The first measure is the training error quantifying the learning error for the data in the sample, while the second one (the generalization error) does it on the whole state space. The second one is the measure one should minimize in order to correctly encode the teacher function. However, since it is not accessible pratically, one has to work on the training error. The natural generalization of the learning perceptron rule can be casted as the gradient descent system

$$\frac{d\boldsymbol{w}}{dt} = -\nabla_{\boldsymbol{w}} \rho_T(\boldsymbol{w}). \tag{5.7}$$

The choice of the model capacity (which in the present case is the number of tunable parameters) is indeed related to under/overfitting issues. In practical applications, they can be faced both with *cross-validation* (i.e. by splitting the whole data set in two or more sections and using part of it to approximate the generalization error and therefore tuning the network parameter or complexity) or *regularization* (which is accomplished by adding some penalty term to the gradient descent rule). We will show at the end of this Chapter that statistical mechanics of spin-glasses offers a theoretical alternative to these empirical trial, by taking advantage of a deep equivalence among learning and retrieval. Besides under/overfitting issues, there is another point to highlight here, which is the fact that generally the training dynamics does not surely end in a global minimum for the error measure (but only in a local one). To face with this problem, it is possible to use stochastic version of gradient descent algorithms.

In other realistic cases (such as density estimation tasks), the teacher function can be only known in probabilistic form, i.e. the objective function is a probability distribution. Suppose we have a set of data $\mathcal{S} = \{\boldsymbol{\sigma}_s, s = 1, \dots, n\}$ generated by an unknown probability distribution $\mathcal{Q}(\boldsymbol{\sigma})$. The aim is to find a distribution $\mathcal{P}(\boldsymbol{\sigma})$ with tunable parameters $\boldsymbol{w}$ within a class of models best approximating the goal solution. Assuming that the data $\mathcal{S}$ are generated by the $\mathcal{P}(\boldsymbol{\sigma})$, one can write their conditional probability as

$$\mathcal{P}(\mathcal{S}) = \prod_s \mathcal{P}(\boldsymbol{\sigma}_s) = \exp(N\mathcal{L}(\boldsymbol{\Lambda}|\mathcal{S})), \tag{5.8}$$

where we defined the *log-likelihood* $\mathcal{L}(\boldsymbol{w}|\mathcal{S}) = N^{-1}\sum_s \log \mathcal{P}(\boldsymbol{\sigma}_s)$. The problem of finding the parameters $\boldsymbol{w}$ maximizing the probability of obtaining the data $\mathcal{S}$ is equivalent to find the maximum of the log-likelihood, and it is insensitive to the addition of a constant. The trick is to take this constant equal to the empirical (Shannon) entropy $-N^{-1}\sum_s \log \mathcal{Q}(\boldsymbol{\sigma}_s)$, so that the function to be maximized is

$$\Delta_N(\mathcal{Q}, \mathcal{P}) = -\frac{1}{N}\sum_s \log \frac{\mathcal{Q}(\boldsymbol{\sigma}_s)}{\mathcal{P}(\boldsymbol{\sigma}_s)}. \tag{5.9}$$

In the large data set limit $n \to \infty$, this quantity can be proven to converge (in probability) towards the so-called Kullback-Leibler (KL) cross entropy, defined as

Definition 5.1. Given two probability measures $\mathcal{Q}$ and $\mathcal{P}$, their related Kullback-Leibler cross entropy reads as

$$\Delta(\mathcal{Q}, \mathcal{P}) = -\sum_{\boldsymbol{\sigma}} \mathcal{Q}(\boldsymbol{\sigma}) \log \frac{\mathcal{Q}(\boldsymbol{\sigma})}{\mathcal{P}(\boldsymbol{\sigma})}, \tag{5.10}$$

and quantifies the *distance* between the two probability distributions.

When $\mathcal{P} = \mathcal{Q}$, the distance is equal to zero. Then, we can design a gradient descent algorithm in order to minimize the KL distance, so that

$$\frac{d\boldsymbol{w}}{dt} = -\nabla_{\boldsymbol{w}} \Delta_N(\mathcal{Q}, \mathcal{P}), \tag{5.11}$$

with all the issue associated to the method.

Another possible approach to the learning problem is the *Bayesian learning*, giving a solid probabilistic basis to training procedures. The basic idea here is that, given the set of input/output relations $\mathcal{S} = \{(\boldsymbol{\sigma}_s, T_s), n = 1, \dots, s\}$ and a prior distribution $\mathcal{P}(\boldsymbol{w})$ for the network parameters, we can compute by standard Bayes' theorem the posterior distribution $\mathcal{P}(\boldsymbol{w}|\mathcal{S})$, measuring the conditional probability for $\boldsymbol{w}$ given the experimental data $\mathcal{S}$. In mathematical terms,

$$\mathcal{P}(\boldsymbol{w}|\mathcal{S}) = \frac{\mathcal{P}(\boldsymbol{w})\mathcal{P}(\mathcal{S}|\boldsymbol{w})}{\mathcal{P}(\mathcal{S})}, \tag{5.12}$$

where $\mathcal{P}(\mathcal{S}|\boldsymbol{w})$ is the data likelihood (i.e. the probability of finding the output T_s by presenting the questions $\boldsymbol{\sigma}_s$ for each s) upon fixing the parameter $\boldsymbol{w}$. Moreover, $\mathcal{P}(\mathcal{S}) = \int d\boldsymbol{w}' \mathcal{P}(\boldsymbol{w}')\mathcal{P}(\mathcal{S}|\boldsymbol{w}')$. After the posterior distribution is determined, we can evaluated the probability of find another output $\bar{T}$ by presenting a new input $\bar{\boldsymbol{\sigma}}$ as

$$\mathcal{P}(\bar{T}|\bar{\boldsymbol{\sigma}}, \mathcal{S}) = \int d\boldsymbol{w}\, \mathcal{P}(\bar{T}|\bar{\boldsymbol{\sigma}}, \mathcal{S})\mathcal{P}(\boldsymbol{w}|\mathcal{S}), \tag{5.13}$$

where $\mathcal{P}(\bar{T}|\bar{\boldsymbol{\sigma}}, \mathcal{S})$ is of course the probability to find the output $\bar{T}$ given the set of parameters $\boldsymbol{w}$ and presented the new input $\bar{\boldsymbol{\sigma}}$.

The Bayesian approach to learning has the great advantage of making the training problem with solid statistical foundations, with the possibility to make precise confidence estimations for single data predictions. Moreover, in Bayesian learning we fix the prior distribution for the parameters (i.e. a class of models), therefore the model complexity is fixed, so that there is no need to cross-validate the model (and consequently all the data can be used for training).

5.3 Restricted Boltzmann Machines and contrastive divergence

In the previous Chapter, we described the Hopfield model as the simplest paradigm for machine retrieval. Since in this thesis we would like to give a general picture of the cognition skills of AI, hereafter we will briefly describe the paradigmatic learning machine, i.e. the Restricted Boltzmann Machine (RBM) [71, 63, 64, 84, 129].

Boltzmann machines are neural networks (multi-partite spin glasses in statistical mechanical jargon) in which mutually connected neurons are organized in different layers, two in its minimal representation (visible and hidden layer), three in the typical realization (visible layer, hidden layer and output layer) and several in deep learning architectures (the so-called Deep Boltzmann Machines, where the majority of layers are hidden). In the following, we will derive the celebrated Hinton's *contrastive divergence* [1, 64, 74] learning algorithm for a standard three layer RBM. In such a network, the first layer (denoted by $\boldsymbol{\sigma}^{\mathrm{I}}$) is the input layer, the last one (whose spins are indicated with $\boldsymbol{\sigma}^{\mathrm{III}}$) is the output layer, while the middle one is the hidden layer (denoted by $\boldsymbol{\sigma}^{\mathrm{II}}$). When there is no danger of confusion, we will use a single variable s_i ($i = 1, \dots, N_I + N_H + N_O$) to denote all of the neurons in the network. Crucially for the arsenal of our weapons to hold, the weights in a RBM are required to be symmetric ($J_{ij} = J_{ji}$) and self-interactions are excluded ($J_{ii} = 0$): under these assumptions Detailed Balance holds and one can prove that the update rule

$$\mathcal{P}(s_i \to -s_i) = \frac{1}{2}[1 - \tanh(\beta s_i h_i(\boldsymbol{s}))], \tag{5.14}$$

converges to a unique stationary distribution described by the partition function

$$\begin{aligned} \mathcal{P}(\boldsymbol{s}) &= Z^{-1} \exp(-\beta H_N(\boldsymbol{s}|\boldsymbol{w})), & (5.15) \\ H_N(\boldsymbol{s}|\boldsymbol{w}) &= -\frac{1}{2}\sum_{ij} J_{ij} s_i s_j - \sum_i b_i s_i, & (5.16) \end{aligned}$$

hence the name *Boltzmann* for these machines. In the following, we shall omit the dependence on the network parameters to make notation clear.

Once we have the probability distribution on the space state of the entire network, we have to setup the training procedure. Here, we have to note that the desired goal is a distribution relating the input and output in a probabilistic form (with no reference on the hidden layer, as it should). We call $\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{III}})$ such a distribution. Then, parameters $\boldsymbol{w}$ in the network must be tuned for $\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{III}})$ to minimize the distance from $\mathcal{Q}$. As discussed above, one can consider the maximum likelihood principle as basic principle. In the discrete picture, the training procedure is therefore described by the prescriptions

$$\begin{aligned}\Delta J_{ij} &= -\epsilon\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}},\\ \Delta b_i &= -\epsilon\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial b_i},\end{aligned}\tag{5.17}$$

where $\epsilon \ll 1$ is the learning strength. The variation of the KL distance is (up to the quadratic order)

$$\begin{aligned}\delta\Delta(\mathcal{Q},\mathcal{P}) &= \sum_{ij}\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}}\Delta J_{ij} + \sum_i \frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial b_i}\Delta b_i + \mathcal{O}(\epsilon^2) =\\ &= -\Big[\sum_{ij}\Big(\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}}\Big)^2 + \sum_i\Big(\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial b_i}\Big)^2\Big] + \mathcal{O}(\epsilon^2),\end{aligned}\tag{5.18}$$

so it will decrease since a fixed point is reached. Before obtaining a concrete prescription for the parameter adjustments, we have to think again to the supervised learning philosophy: for each given input, the student and teacher compare their outputs, and the former update its parameters to minimize the errors. Then, during training, the input (i.e. neurons in the first layer) are fixed in both situations. Then, the desired probability distribution is

$$\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}}) = \sum_{\boldsymbol{\sigma}^{\mathrm{II}}}\mathcal{P}(\boldsymbol{s}).\tag{5.19}$$

Now, since the input layer is fixed, we have to express everything in terms of the conditional probability $\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}})$. Indeed

$$\mathcal{P}(\boldsymbol{s}) = \mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}})\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}}) = \mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}})\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}}),\tag{5.20}$$

since of course the probability $\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}}) = \mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}}) = \sum_{\boldsymbol{\sigma}^{\mathrm{III}}}\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})$ is known. Then, we have

$$\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}}) = \mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}})\sum_{\boldsymbol{\sigma}^{\mathrm{II}}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}}).\tag{5.21}$$

But

$$\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}}) = Z(\boldsymbol{\sigma}^{\mathrm{I}})^{-1}\exp(-\beta H_N(\boldsymbol{s})),\tag{5.22}$$

where

$$Z(\boldsymbol{\sigma}^{\mathrm{I}}) = \sum_{\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\exp(-\beta H_N(\boldsymbol{s})).\tag{5.23}$$

Then

$$\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})=\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}})\frac{Z(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})}{Z(\boldsymbol{\sigma}^{\mathrm{I}})},\tag{5.24}$$

where of course $Z(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})=\sum_{\boldsymbol{\sigma}^{\mathrm{II}}}\exp(-\beta H_N(\boldsymbol{s}))$. Then, the KL distance we have to minimize is[1]

$$\Delta(\mathcal{Q},\mathcal{P})=-\frac{1}{\beta}\sum_{\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}}}\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})[\log Z(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})-\log Z(\boldsymbol{\sigma}^{\mathrm{I}})].\tag{5.25}$$

Now, computing the derivative with respect to J_{ij}, we have

$$\begin{aligned}\frac{\partial}{\partial J_{ij}}\log Z(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})&=-\beta\sum_{\boldsymbol{\sigma}^{\mathrm{II}}}\frac{\partial H_N}{\partial J_{ij}}\frac{e^{-\beta H_N(\boldsymbol{s})}}{Z(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})}=-\beta\sum_{\boldsymbol{\sigma}^{\mathrm{II}}}\frac{\partial H_N}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}}|\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}}),\\ \frac{\partial}{\partial J_{ij}}\log Z(\boldsymbol{\sigma}^{\mathrm{I}})&=-\beta\sum_{\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\frac{\partial H_N}{\partial J_{ij}}\frac{e^{-\beta H_N(\boldsymbol{s})}}{Z(\boldsymbol{\sigma}^{\mathrm{I}})}=-\beta\sum_{\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\frac{\partial H_N}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}}).\end{aligned}\tag{5.26}$$

Putting this results into the KL distance, we have

$$\begin{aligned}\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}}=&\sum_{\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\frac{\partial H_N(\boldsymbol{s})}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}}|\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})\\ &-\sum_{\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}},\bar{\boldsymbol{\sigma}}^{\mathrm{III}}}\frac{\partial H_N(\bar{\boldsymbol{s}})}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\bar{\boldsymbol{\sigma}}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}})\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}}),\end{aligned}\tag{5.27}$$

where $\bar{\boldsymbol{s}}=(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{II}},\bar{\boldsymbol{\sigma}}^{\mathrm{III}})$. The expression can be further simplified as

$$\begin{aligned}\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}}=&\sum_{\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\frac{\partial H_N(\boldsymbol{s})}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}}|\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{III}})\\ &-\sum_{\boldsymbol{\sigma}^{\mathrm{I}},\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}}\frac{\partial H_N(\boldsymbol{s})}{\partial J_{ij}}\mathcal{P}(\boldsymbol{\sigma}^{\mathrm{II}},\boldsymbol{\sigma}^{\mathrm{III}}|\boldsymbol{\sigma}^{\mathrm{I}})\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}}).\end{aligned}\tag{5.28}$$

Recalling the expression of the Hamiltonian for the Boltzmann machine (see eq. 5.16), we finally arrive to the next crucial

Proposition 5.1. *The Contrastive Divergence learning algorithm for RBMs reads as*

$$\frac{\partial\Delta(\mathcal{Q},\mathcal{P})}{\partial J_{ij}}=-(\langle s_is_j\rangle_{\mathit{clumped}}-\langle s_is_j\rangle_{\mathit{free}}),\tag{5.29}$$

[1] We added an extra factor β^{-1} for convenience. Of course, this does not affect the minimization problem.

where clamped *means that the averages are evaluated when the visible layer is forced on a pattern of information (the one we are storing) while* free *is the standard quenched average in statistical mechanics. Of course, a completely analogous derivation holds also for the external fields, so that finally we arrive to the parameters update rule*

$$\begin{aligned} \Delta J_{ij} &= \epsilon(\langle s_i s_j \rangle_{\textit{clamped}} - \langle s_i s_j \rangle_{\textit{free}}), \\ \Delta b_i &= \epsilon(\langle s_i \rangle_{\textit{clamped}} - \langle s_i \rangle_{\textit{free}}). \end{aligned} \tag{5.30}$$

Remark 5.2. These learning rules are crystal clear: what these machines do is that they try to learn the statistical structure of the data they have been exposed to, by reproducing the lowest order correlations functions. Clearly, as discussed in the first Chapter, as long as we deal with Gaussian theories, one-point and two-point correlations functions (accounting for mean and variances in the available data) suffice.

Remark 5.3. From the operational point of view, it is possible to numerically estimate the one- and two-point correlation functions in both clamped and free states. Therefore, the numerical algorithm works as follows. First of all, we fix input and output neurons and operate the network dynamics until the hidden layer relaxes, then we compute the one- and two-point clumped correlation functions. This procedure is repeated for many input/output relations $(\boldsymbol{\sigma}^{\mathrm{I}}, \boldsymbol{\sigma}^{\mathrm{III}})$. After that, we fix only the input layer and leave the network reach the equilibrium until the equilibrium is reached, then we compute the one- and two-point free correlation functions. Again, the procedure is repeated for many inputs $\boldsymbol{\sigma}^{\mathrm{I}}$ generated according to $\mathcal{Q}(\boldsymbol{\sigma}^{\mathrm{I}})$. Finally, we update the network parameters according to the prescribed rule.

5.4 Retrieving what has been learnt: associative neural nets

We mentioned how the Hopfield network is a representative model for retrieval and Boltzmann machines are fundamental systems for learning. From both intuitively and formal point of view, learning and retrieval are not two independent operations, rather two complementary aspects of cognition. Hence, it must be possible to recover the Hebb rule for learning also starting from the (restricted) Boltzmann machines.

To see this, for the sake of simplicity hereafter we will use them in their simplest representation, namely as basic two-layer networks. We use the symbol σ_i, $i \in \{1, ..., N\}$ for neurons in the visible layer, z_μ, $\mu \in \{1, ..., P\}$

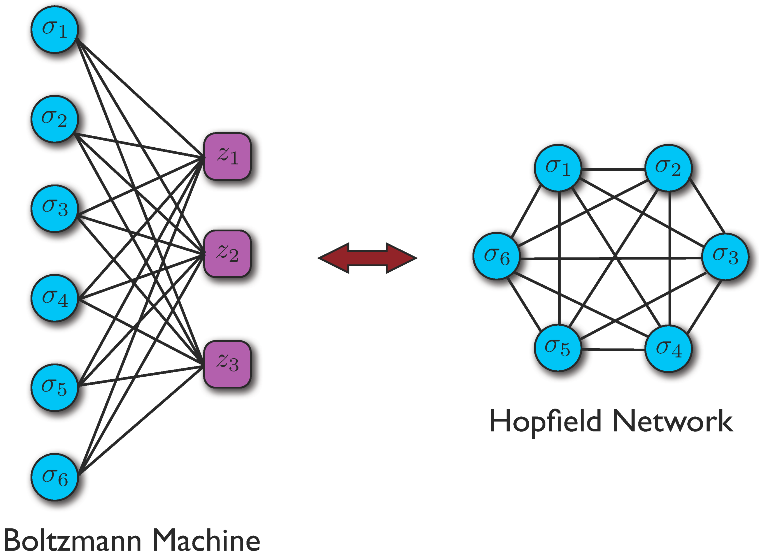

Figure 5.3: **Correspondence between a two-layer restricted Boltzmann machine and an Hopfield neural network.** Note, crucially, that the storage capacity in the latter, i.e. $\lambda \sim P/N$, matches the ratio among the size of the hidden layer over the visible one.

for those in the hidden layer and w_i^μ to label the links (or weights) between the neuron i in one layer and the neuron μ in the other layer. We can then write the cost function for the Boltzmann machine as

$$H_N(\sigma, z, w) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} w_i^\mu \sigma_i z_\mu .$$

This Hamiltonian represents, in the jargon of statistical mechanics, a *bipartite spin-glass*. In order to study the related phase diagram, we work out the statistical mechanics machinery, starting by writing the quenched pressure of the Boltzmann machine as

$$\alpha_N(\beta) = \frac{1}{N} \mathbb{E} \ln \sum_{\boldsymbol{\sigma}} \sum_{\boldsymbol{z}} \exp\left\{ \frac{\beta}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} w_i^\mu \sigma_i z_\mu \right\}.$$

Remarkably, as there are no links within each party, from a statistical mechanics perspective, these networks are simple to deal with because the sums are factorized. In particular, we can carry out the sum over $\boldsymbol{z}$ to get

$$\alpha_N(\beta) \sim \frac{1}{N} \ln \sum_{\boldsymbol{\sigma}} \exp\left\{ \frac{\beta}{2N} \sum_{i,j=1}^{N} \left(\sum_{\mu=1}^{P} w_i^\mu w_j^\mu \right) \sigma_i \sigma_j \right\},$$

hence the leading contribution of the Boltzmann machine is nothing but the Hopfield model. In fact, if we carefully look at the expressions above, we can observe that the role of a machine weight w_i^μ and the one of a neural network pattern ξ_i^μ, is exactly the same.

Remark 5.4. A crucial point in this equivalence is that the storage capacity of the Hopfield neural network, i.e. $\lambda = \lim_{N\to\infty} P/N$, mirrors the ratio among the sizes of the two layers in the Boltzmann machine (hidden over visible): as we know that pattern recognition can be accomplished by the Hopfield model just if the critical capacity at which it is loaded does not exceed the critical threshold, we would be tempted to use this phase transition also in the RBM framework. In this equivalence, a large λ (much more the critical threshold) implies that the size of the hidden layer is far too big, thus the network learning would typically suffer of overfitting (see Fig. 5.2).

5.5 Statistical equivalence of RBM and Hopfield networks

To deepen the last remark we study a *hybrid* two-layer Boltzmann Machine (HBM) as a network in which the activity of the neurons in the visible layer is Boolean, $\sigma_i = \pm 1,\ i \in (1, ..., N)$, while those in the hidden layer are continuous (analog). The synaptic connections between units in the two layers are fixed and symmetric, and are defined by the synaptic matrix ξ_i^μ.

The net input to neurons σ_i in the digital layer is the sum of the activities in the hidden one, weighted by the synaptic matrix, i.e. $\sum_\mu \xi_i^\mu z_\mu$. Analogously, for neurons z_μ in the latter is the sum of the activities in the visible layer, always weighted by the synaptic matrix, i.e. $\sum_i \xi_i^\mu \sigma_i$. Because of the different nature of the units in the two layers, also their dynamics will be different. Indeed, in the analog layer neurons evolve continuously in time, while in the digital layer the evolution takes place in discrete steps. In particular, the activity in the hidden layer is described by the stochastic differential equation

$$T\frac{dz_\mu}{dt} = -z_\mu(t) + \sum_i \xi_i^\mu \sigma_i + \sqrt{\frac{2T}{\beta}}\,\eta_\mu(t), \tag{5.31}$$

where η is a white gaussian noise with zero mean and covariance $\langle \eta_\mu(t)\eta_\nu(t')\rangle = \delta_{\mu\nu}\ \delta(t-t')$. Briefly speaking, the parameter T quantifies the timescale of the dynamics, and the parameter β determines the strength of the fluctuations. The first term in the right hand side is a leakage term, the second term is the input signal and the third term is a noise source. Noise acting on different hidden units is uncorrelated, so they evolve independently from each other. Eq. (5.31) describes an Ornstein-Uhlembeck diffusion process [131] and, for fixed values of $\boldsymbol{\sigma}$, the equilibrium distribution of z_μ is a Gaussian distribution centered around the input signal, which in mathematical terms is

$$\mathcal{P}(z_\mu|\sigma) = \sqrt{\frac{\beta}{2\pi}} \exp\Big\{ -\frac{\beta}{2}\Big(z_\mu - \sum_i \xi_i^\mu \sigma_i\Big)^2\Big\}. \tag{5.32}$$

In deriving this probability distribution, we are tacitly assuming that the activity of Boolean units $\boldsymbol{\sigma}$ must be constant, while in fact it depends on time. To make both features compatible, we should assume that the timescale of diffusion T is faster than the update rate of neurons in the visible layer. Therefore, a different equilibrium distribution for $\boldsymbol{z}$, characterized by different values of $\boldsymbol{\sigma}$, holds between each subsequent update of $\boldsymbol{\sigma}$. Since hidden units are independent, their joint distribution is simply the product of distributions of each individual neurons, i.e. $\mathcal{P}(\boldsymbol{z}|\boldsymbol{\sigma}) = \prod_{\mu=1}^{P} \mathcal{P}(z_\mu|\boldsymbol{\sigma})$. On the other side, the evolution in the visible layer follows a standard neural dynamics of Glauber type [11]. At a specified sequence of time intervals (much larger than T), the activity of units in the digital layer is updated randomly according to a probability depending on their input. In the same way, when updating the digital units $\boldsymbol{\sigma}$, the analog variables z are kept fixed, i.e. the update of digital units is instantaneous. Furthermore, also the activity of the σ_i is independent on other units, and the probability is a logistic function of its input, leading to

$$\mathcal{P}(\sigma_i|\boldsymbol{z}) = \frac{\exp[\beta\sigma_i \sum_\mu \xi_i^\mu z_\mu]}{\exp[\beta \sum_\mu \xi_i^\mu z_\mu] + \exp[-\beta \sum_\mu \xi_i^\mu z_\mu]}. \tag{5.33}$$

Each σ_i are independent on the other visible units (since there are no intra-layer connections), so that their joint distribution is again the product of individual distributions, i.e. $\mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{z}) = \prod_{i=1}^{N} \mathcal{P}(\sigma_i|\boldsymbol{z})$.

Once we have the conditional distributions of either layers at our disposal, by applying Bayes' theorem we can determine their joint distribution, $\mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z})$, together with the marginal distributions $\mathcal{P}(\boldsymbol{z})$ and $\mathcal{P}(\boldsymbol{\sigma})$ by the chain of equalities $\mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z}) = \mathcal{P}(\boldsymbol{z}|\boldsymbol{\sigma})\mathcal{P}(\boldsymbol{\sigma}) = \mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{z})\mathcal{P}(\boldsymbol{z})$. Using the fact that marginal distributions depend on single layer variables. The result is, for the joint distribution

$$\mathcal{P}(\boldsymbol{\sigma}, \boldsymbol{z}) \propto \exp\Big(-\frac{\beta}{2}\sum_\mu z_\mu^2 + \beta \sum_{i,\mu} \sigma_i \xi_i^\mu z_\mu\Big). \tag{5.34}$$

The marginal distribution describing the statistics of visible neurons is equal to

$$\mathcal{P}(\boldsymbol{\sigma}) \propto \exp\Big(\frac{\beta}{2}\sum_{i,j}\big(\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}\big)\sigma_i\sigma_j\Big). \tag{5.35}$$

A first inspection to last equality confirms that such a marginal distribution is *exactly equal* those describing the relaxation of neurons in Hopfield networks, where the synaptic weights of the Hopfield network are given by the expression in round brackets: the stored patterns of the Hopfield model corresponds to the synaptic weights of the HBM, described by the $\boldsymbol{\xi}$ variables and the number of patterns corresponds to the number P of hidden units.

In conclusion, HBM and Hopfield network share the same probability distribution, once the hidden variables are marginalized, meaning that both models are statistically equivalent. However, the analogy can be pushed further by linking retrieval in the Hopfield network to capability of HBM to learn a specific pattern of neural activation. The maximum number of patterns P that can be retrieved in a Hopfield network is known [11], and is $\simeq 0.14N$. In particular, the models should also share the same phase diagram. As a consequence, non-retrieval regions in Hopfield model (which, we recall, takes place for $P > 0.14N$ at low thermal noise) can be linked to overfitting issues in the HBM. This can be understood as follows: if the HBM has a huge large number P of hidden variables (with respect to the visible layers units), the model we are considering is too complex, provoking overfitting in learning the observed patterns. This behaviour causes the inability of the HBM to reproduce the statistics of the observed system. The correspondence between Hopfield network and HBM allows to predicts that the maximum number of hidden variables in the HBM is *precisely* $0.14N$. For a numerical check of the equivalence between HBMs and RBMs and the relation between non-retrieval phases of the former and overfitting issues in the latter, we refer to the work [21].

Remark 5.5. We would like to stress that this analogy does not hold by default nor with the standard Hopfield model, as the latter is equipped with digital patterns (and no contrastive divergence could be derived with those patterns) neither with the analog Hopfield model, as the latter is equipped with real-valued patterns (and no retrieval phase is allowed with those patterns): it is thanks to the study of the hybrid Hopfield model we studied at the end of the previous Chapter that we know that such an equivalence concretely works.

Part Three: Statistical Mechanics for Unlearning and Sleeping

Chapter 6

Beyond the standard paradigm: Unlearning for low storage

Once understood that the larger is the critical capacity of the Hopfield model for pattern recognition, the stronger are the inferential skills of its dual RBM (and consequently more resistent to overfitting), it is natural to ask if it is possible to increase the critical threshold larger than that of the standard reference[1]

To address this problem, in this and the next Chapters[2] we deal with an entirely novel problem. In particular, we will study a way to get rid off the exponential proliferation of the unwanted spurious states naturally generated when storing the P pure patterns. A remarkable point that Physicist and Mathematicians could appreciate is that we arrived at writing these two conclusive Chapters following intuitions that have been entirely "theoretically physically driven", and lie in the Hamilton-Jacobi approach to the statistical mechanics of complex system (which we largely exploited in this thesis).

Indeed, whatever the route, i.e. starting from the Hebbian prescription [61] as in the original Hopfield paper [65] or upon marginalizing over the hidden layer in Boltzmann Machine (hence, after learning via contrastive divergence) [21], unfortunately, we always end up in a network whose attractors are by far more than the stored patterns [11] (namely more than the solely *pure states* we would see retrieved by the network). The excess stock

[1] Further, there is a remarkable Theorem of Information Theory due to Elisabeth Gardner stating that, for networks with real and symmetric interactions, the maximal critical capacity is $\lambda_c = 1$, quite larger than the Hopfield value $\lambda_c \sim 0.14$.[53].

[2] The difference between these two Chapters, in a nutshell, is that in the present one we will present the extended theory we developed in the low-storage regime - and we related this to *unlearning in neural networks* - while in the final one we will tackle the whole high storage regime - and we will related that analysis to *sleeping in neural networks*.

of (local) minima is indeed constituted by *spurious states*, whose simplest example is a 3-pattern mixture defined as

$$\sigma_i = \text{sign}\left(\xi_i^1 + \xi_i^2 + \xi_i^3\right). \tag{6.1}$$

As we saw in Chapter 4.3, the Mattis overlap for such a state state with any of the three patterns is - in the large N limit - $m^\nu = N^{-1}\sum_i \xi_i^\nu \sigma_i = 0.5$ (for $\nu = (1,2,3)$). Hence, while smaller in amplitude than the Mattis overlap of a pure state (whose amplitude is one), it is still a metastable state: if orbiting in the surrounding, the network can be attracted by such spurious states and converge to them rather than to the pure ones. Unfortunately, as the patterns are added linearly to the memory kernel, there is an exponential (combinatorial) proliferation of these unwanted meta-stable states in the retrieval landscape of the Hopfield network. Hopfield himself suggested a procedure to prune - or remove - (a part of) them from his coupling matrix [66]. In a nutshell, Hopfield's idea is again transparent, elegant and brilliant: since there are sensibly much more spurious states (i.e. metastable minima) than pure states (global minima), we can prepare the system at random and make a quench (e.g. a search for minima with steepest descent rather than conjugate gradient or stochastic algorithms). In this way, the system will end up in a spurious state with high probability. Then, we can collect this equilibrium (spurious) configuration and subtract it from the memory kernel, via the pruning rule

$$J_{ij} = \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu - \langle \sigma_i \sigma_j \rangle_{spurious},$$

i.e. increasing its energy. We can do this iteratively and check that effectively the network becomes progressively cleared from these nasty attractors: this procedure is called *unlearning* [98, 97, 47, 79] and it has been linked to REM sleep [39] (offering a possible intriguing picture for its interpretation) due to the effectiveness of the random starting point setting for the quenching procedure in consolidating memories (phenomenon to be eventually correlated with the rapid eye movements in that part of sleep). This will be the focus of the present Chapter.

As a sideline, we also note that this investigation will also be of interest for Deep Learing (the quest to overcome the reductionistic Gaussian description of the statistical datasets discussed in the First Chapter of this thesis), whose main characteristic we quickly revise hereafter. Within the theoretical framework of *disordered statistical mechanics* [65, 11] for AI, Hopfield recently offered a connectionist perspective where the high skills of deep learning machines could possibly be framed [82]. The key idea is simple, and it is on many-body extensions of his celebrated pairwise model, and can be understood as follows. Suppose we want to retrieve one out of P random patterns ξ^μ stored in the network and we want to describe this property via a cost function $H(\boldsymbol{\sigma}|\boldsymbol{\xi})$ that resembles Hamiltonians in Physics (such that the minima of the Hamiltonian would match the patterns themselves [11, 93, 108]), the *simplest and most natural* guess would be summing all the squared scalar products between the neurons and the patterns, i.e. $H(\boldsymbol{\sigma}|\boldsymbol{\xi}) \propto -\sum_\mu^P(\boldsymbol{\sigma}\cdot\boldsymbol{\xi}^\mu)^2$. In the thermodynamic limit, patterns become orthogonal.[1] If the state vector $\boldsymbol{\sigma}$ is uncorrelated with all of them, each parenthesized term would be very small. On the other hand, if the state network $\boldsymbol{\sigma}$ is sufficiently correlated with one of the P patterns (meaning that we are in retrieval mode), then its contribution in the summation would be no longer negligible. The central point in this argument lies in the requirement of local convexity of the Hamiltonian. However, we stressed that the Hopfield model Hamiltonian is the simplest and most natural, but it is surely not the only possible. Indeed, all previous arguments could be generalized straightforwardly beyond the parabolic approximation coded by the pairwise interactions, for instance including (even) higher order contributions (the so-called *P-spin* terms). From a connectionist perspective [11, 89, 51], memories are stored in the connections, adding more and more P-spin contributions to the Hamiltonian adds more and more synapses where information can be filed. Furthermore, an intimate relation between Deep Learning architectures and P-spin models has been recently argued in [92] by means of renormalization group techniques, thus motivating the possibility to tackle modern AI problems by means of statistical mechanics tools [15, 16, 109].

Recently, we pursued the goal to give these two branches of AI - deep learning and unlearning - a unified point of view by allowing Hopfield model to include also P-spin interaction terms. In doing this, we shall approach the problem to give an expression for the free energy of the new model within the Hamilton-Jacobi framework we discussed so far. In order to frame the model in the most natural way, let us briefly recall the philosophy behind the method by starting again with Curie-Weiss model. We have previously shown that its free energy obeys a standard (i.e. *classical non-relativistic*) Hamilton-Jacobi PDE in the space of the coupling and external field (playing the role of time and space coordinates respectively). To infer statistical properties on the CW network, we can thus use this mechanical analogy and study a

[1] With the term *orthogonal* we mean that $\lim_{N\to\infty} N^{-1}\xi^\mu\xi^\nu = \lim_{N\to\infty} N^{-1}\sum_i^N \xi_i^\mu\xi_i^\nu = \delta_{\mu\nu}$. However, at finite N, this is a N-long sum of terms whose probability of being ±1 is one half: it is a random walk with zero mean and variance N, hence spurious correlations are expected to vanish $\propto 1/\sqrt{N}$.

fictitious particle of unitary mass classically moving in this $1+1$ spacetime (under the proper PDE derived from the statistical mechanical framework). Its properties, once translated back in the original setting, exactly recover all the results of the standard statistical paradigm.

Pushing further this mechanical analogy, we introduced a very natural generalization of the Hopfield Hamiltonian, which is simply its relativistic version.The first remarkable feature of such an extension is that, while the classical (Hopfield) model is described by a second-order monomial Hamiltonian, its relativistic version (if Taylor-expanded in the order parameters) is an infinite sum of P-spin terms. The second point is that such terms turn out to be exactly solely the even ones and with alternate signs, naturally suggesting a comprehensive picture for both Deep Learning and unlearning.

We conclude this introduction by stressing that - while numerical and heuristic explorations have been largely exploited in the Computer Science Literature in the past decade, our aim here is to frame the problem in a rigorous and well controllable analytical formulation. In this scenario, important contributions already appeared in the Mathematical and Theoretical Physics Literature (see for example [27, 20, 34, 29, 30, 32, 31, 126, 125, 102] and references therein). In this Chapter, we will deal with the (much more controllable) low storage regime for both the classical (i.e. original Hopfield) and relativistic models in the Hamilton-Jacobi framework, since for the latter technical difficulties in the analytical solution are still unsolved.

The Chapter is structured as follows. As a preliminary introduction, we set up the mathematical framework, i.e. the mechanical analogy for neural networks. Then, we show the analogy at the classical level and solve the original pairwise Hopfield model, re-obtaining all the well-known existing results. After that, we push further the analogy to include higher order (P-spin) contributions to the Hopfield cost function (such that all P-spin contributions can be resummed in the relativistic Hamiltonian for a free particle). Also in this case, within the Hamilton-Jacobi framework we obtain an exhaustive statistical picture. As a technical note, we give an explicit proof of the existence of the thermodynamic limit for the free energy and re-obtain the above picture from a purely statistical mechanical setting (i.e. by using the standard Guerra's interpolation technique). Finally, we perform extensive numerical numerical analysis of the capabilities of this extended model, in particular in reducing the attracting power of spurious configurations. This will be shown trough a one-to-one comparison among performances with the Hopfield model.

6.1 The Hamilton-Jacobi formalism (classical)

In this Section, we will briefly review the mechanical analogy for the classical (i.e. pairwise) Hopfield model. We will show that the free energy can be interpreted as the HJ action, therefore it obeys an Hamilton-Jacobi PDE. By means of the mechanical analogy, we re-derive the expression for the Hopfield free energy (in the low storage regime) in terms of the order parameters and the associated self-consistency equations.

In order for this Chapter to be self-contained and well-readable, we rewrite the following

Definition 6.1. The Hamiltonian of the Hopfield model equipped with N neurons σ_i, $i \in (1, ..., N)$ and P patterns ξ^μ, $\mu \in (1, ..., P)$ is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{\mu=1}^{P}\sum_{1\leq i<j\leq N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j, \tag{6.2}$$

where patterns bit are extracted i.i.d. with probability $\mathcal{P}(\xi_i^\mu = +1) = \mathcal{P}(\xi_i^\mu = -1) = 1/2$ for all $i = 1, \dots, N$ and $\mu = 1, \dots, P$.

Since all (i.e. both dynamical and slow-evolving) variables are Boolean, we can again include self-interaction, with an error becoming negligible in the thermodynamic limit. Therefore, the partition function is

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \exp\Big\{\frac{\beta}{2N}\sum_{\mu=1}^{P}\sum_{i,j=1}^{N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j\Big\}, \tag{6.3}$$

The ultimate goal is to find the expression for the pressure $\alpha(\beta) = -\beta f(\beta) = \lim_{N\to\infty} N^{-1}\log Z_N(\beta)$.As usual, once that the partition function is introduce, we can also define the Boltzmann factor $B_N(\beta, \boldsymbol{\sigma})$, the Boltzmann-Gibbs averages $\omega(\cdot)$ and $\langle\cdot\rangle = \mathbb{E}\omega(\cdot)$, where $\mathbb{E}$ is the usual average over the pattern realizations.

We now turn on setting up the Hamilton-Jacobi framework. To do this, we have to introduce P spatial variables $x_\mu \in \mathbb{R}$, $\mu \in (1, ..., P)$ and a temporal variable $t \in \mathbb{R}^+$. Then, we can give the following

Definition 6.2. The generalized partition function in the Hamilton-Jacobi framework is defined as

$$Z_N(\beta; t, \boldsymbol{x}) \doteq Z_N(t, \boldsymbol{x}) = \sum_{\boldsymbol{\sigma}} \exp\Big\{-\frac{t}{2N}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j + \sum_{i=1}^{N}\sum_{\mu=1}^{P} x_\mu \xi_i^\mu \sigma_i\Big\}. \tag{6.4}$$

Of course, with respect to this partition function we can introduce the Boltzmann-Gibbs averages $\omega_{t,\boldsymbol{x}}(\cdot)$ and $\langle\cdot\rangle_{t\boldsymbol{x}}$ (if needed) in the same way as before. We highlight that the standard statistical mechanics framework is recovered by setting $t = -\beta$ and $\boldsymbol{x} = 0$. In the same way, the intensive pressure is introduced:

Definition 6.3. The intensive pressure $\alpha_N(t,\boldsymbol{x})$ of associated to the generalized partition function (6.4) is

$$\begin{aligned}\alpha_N(t,\boldsymbol{x}) &= \frac{1}{N}\log Z_N(t,\boldsymbol{x}) = \\ &= \frac{1}{N}\log\sum_{\boldsymbol{\sigma}}\exp\Big\{-\frac{tN}{2}\sum_{\mu=1}^{P}m_\mu^2 + N\sum_{\mu=1}^{P}x_\mu m_\mu\Big\},\end{aligned} \tag{6.5}$$

where

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i, \tag{6.6}$$

are the usual Mattis overlaps $\mu \in (1,...,P)$.

It is easy to check that the space-time derivatives of the interpolating free energy (6.5) read as

$$\begin{aligned}\frac{\partial\alpha_N(t,\boldsymbol{x})}{\partial t} &= -\frac{1}{2}\sum_{\mu=1}^{P}\omega_{t,\boldsymbol{x}}(m_\mu^2), \\ \frac{\partial\alpha_N(t,\boldsymbol{x})}{\partial x_\mu} &= \omega_{t,\boldsymbol{x}}(m_\mu).\end{aligned} \tag{6.7}$$

With these equalities, it is easy to prove the following

Proposition 6.1. *By construction, $\alpha_N(t,\boldsymbol{x})$ obeys the following (classical) Hamilton-Jacobi PDE and it plays as the action $S_N(t,\boldsymbol{x})$ in the mechanical analogy*

$$\frac{\partial\alpha_N}{\partial t} + \frac{1}{2}\left(\frac{\partial\alpha_N}{\partial x^\mu}\right)^2 + V_N(t,\boldsymbol{x}) = 0, \tag{6.8}$$

$$V_N(t,\boldsymbol{x}) = \frac{1}{2}\sum_{\mu=1}^{P}\left(\omega_{t,\boldsymbol{x}}(m_\mu^2) - \omega_{t,\boldsymbol{x}}(m_\mu)^2\right). \tag{6.9}$$

Remark 6.1. This partial differential equation describes, even at finite volume N, the motion of a classical (non-relativistic) particle, with unitary mass[1] in $P+1$ dimensions.

[1] Note that, according to equation (6.8), the classical momentum is $\omega_{t,\boldsymbol{x}}(m_\mu)$.

Remark 6.2. In the thermodynamic limit, away from critical point, the Mattis magnetizations self-average, i.e.

$$\lim_{N\to\infty}\sum_{\mu=1}^{P}(\omega_{t,\boldsymbol{x}}(m_\mu^2)-\omega_{t,\boldsymbol{x}}(m_\mu)^2)=0, \tag{6.10}$$

meaning that $\lim_{N\to\infty}V_N(t,\boldsymbol{x}) = V(t,\boldsymbol{x}) = 0$. Therefore, the Hamilton-Jacobi PDE equation reduces to the dynamics of a free particle with unitary mass in the $P+1$-dimensional space.

Remark 6.3. In the thermodynamic limit, the motion has space-time symmetries whose Noëther currents, derived respectively for the momentum conservation and for the energy conservation as

$$\lim_{N\to\infty}\sum_{\mu=1}^{P}(\omega_{t,\boldsymbol{x}}(m_\mu^2)-\omega_{t,\boldsymbol{x}}(m_\mu)^2) \;=\; 0 \tag{6.11}$$

$$\lim_{N\to\infty}\sum_{\mu=1}^{P}(\omega_{t,\boldsymbol{x}}(m_\mu^4)-\omega_{t,\boldsymbol{x}}(m_\mu^2)^2) \;=\; 0, \tag{6.12}$$

which are respectively the momentum and energy conservation. These equalities mirror the classical self-averaging properties in the statistical mechanical jargon.

By using 6.2 in the thermodynamic limit, the fictitious particle moves across Galilean trajectory, i.e. a straight line $\boldsymbol{x} = \boldsymbol{x}_0 + \omega_{t,\boldsymbol{x}}(\boldsymbol{m})\cdot(t-t_0)$ where $\boldsymbol{m}$ is the P-momentum of the particle.
Because of the mechanical analogy, the determination of an explicit expression of the free energy reduces to the explicit calculation of the action of the free motion. As initial conditions, we are free choose $t_0 = 0$ (which leaves to leave only with a one-body system), so that

$$\alpha(t,\boldsymbol{x}) = \alpha(0,\boldsymbol{x}_0) + \int_0^t dt'\mathcal{L}(t'), \tag{6.13}$$

where $\mathcal{L} = \frac{1}{2}\omega_{t,\boldsymbol{x}}(\boldsymbol{m})^2$ is the Lagrangian. Since it is practically the kinetic energy of the particle, and since the momentum is a constant of motion, the Lagrangian itself is a constant of motion as long as $V(t,\boldsymbol{x}) = 0$ (i.e. in the thermodynamic limit). Hence, the only calculations required are due to evaluate the (simple) Cauchy condition

$$\alpha(0,\boldsymbol{x}_0) = \log 2 + \mathbb{E}\log\cosh\boldsymbol{x}_0\cdot\boldsymbol{\xi}, \tag{6.14}$$

where of course $\boldsymbol{x}_0 = \boldsymbol{x}(t) - \omega_{t,\boldsymbol{x}}(\boldsymbol{m})t$. Now, since in the thermodynamic limit, the Mattis magnetizations are self-averaging quantities, they can be replaced directly with their equilibrium values. Then, we can simply drop the Boltzmann-Gibbs average and write simply $\boldsymbol{m}$ instead of $\omega_{t,\boldsymbol{x}}(\boldsymbol{m})$, always taking in mind that now we mean the values of Mattis magnetization at the equilibrium. Then, we have

Theorem 6.1. *The infinite volume limit of the Hopfield action* (6.5) *reads as*

$$\alpha(t,\boldsymbol{x}) = \log 2 + \mathbb{E}\log\cosh(\boldsymbol{x} - t\boldsymbol{m})\cdot\boldsymbol{\xi} + \frac{t}{2}\boldsymbol{m}^2. \tag{6.15}$$

Moreover, the (classical) Hopfield free energy is recovered by setting $t = -\beta$ *and* $\boldsymbol{x} = 0$*, therefore*

$$\alpha(\beta) = \alpha(-\beta, 0) = \log 2 + \mathbb{E}\log\cosh(\beta\boldsymbol{m}\cdot\boldsymbol{\xi}) - \frac{\beta}{2}\boldsymbol{m}^2. \tag{6.16}$$

Remark 6.4. By direct comparison, it is easy to note that the free energy associated to the intensive pressure (6.16) precisely equals the one of the standard Hopfield model in the low storage regime (therefore, also the same self-consistency equations directly follow).

6.2 The Hamilton-Jacobi formalism (relativistic)

So far, we have shown that Hamilton-Jacobi framework allows to establish a mechanical analogy which is a very powerful method to solve the thermodynamics of Hopfield neural networks in the low storage regime. However, its importance is not only computational, as we will show in this Section. Indeed, the analogy can now be used to carry out a very natural extension of the Hopfield cost-function. To this aim, we can notice that, as the free energy plays as an action, we can interprete the exponent in the Maxwell-Boltzmann weight as the product of the $P+1$ momentum with the $P+1$ position vector (i.e. $-tE+\boldsymbol{x}\cdot\boldsymbol{m}$ has precisely a covariant form). In other words, the underlying metric is not Euclidean, rather its Minkowskian version, as it happens in special relativity.

Since the Hopfield Hamiltonian is nothing but the (kinetic) energy associated to the fictitious particle, the mechanical analogy naturally suggests the extension of Hopfield model by its relativistic deformation, i.e.

$$-\frac{\boldsymbol{m}^2}{2} \to -\sqrt{1+\boldsymbol{m}^2}, \tag{6.17}$$

since $\boldsymbol{m}$ plays the role of momentum. Therefore, we can introduce the (relativistic) Hopfield model with the following

Definition 6.4. The Hamiltonian for the relativistic Hopfield model equipped with N neurons σ_i, $i \in (1, ..., N)$ and P patterns ξ^μ, $\mu \in (1, ..., P)$ is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -N\sqrt{1+\boldsymbol{m}^2}, \tag{6.18}$$

where the spin-dependence is implicit in the definition of the Mattis magnetizations

$$m_\mu = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i. \tag{6.19}$$

Remark 6.5. Notice that, Taylor-expanding the Hamiltonian (6.17) in the Mattis magnetizations (the convergence of Taylor series is of course garantued as long as $|\boldsymbol{m}| < 1$), we obtain an infinite list of many-body (P-spin) contributions (going in the direction suggested by Hopfield regarding Deep Learning [82]). Truncating the expansions at the next-to-leading order, we have

$$\begin{aligned} -\frac{H_N(\boldsymbol{\sigma}|\boldsymbol{\xi})}{N} &= 1 + \frac{1}{2N^2}\sum_{i,j=1}^{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \\ &- \frac{1}{8N^4}\sum_{i,j,k,l=1}^{N}\sum_{\mu,\nu=1}^{P} \xi_i^\mu \xi_j^\mu \xi_k^\nu \xi_l^\nu \sigma_i \sigma_j \sigma_k \sigma_l + \dots \end{aligned} \tag{6.20}$$

Further, we notice that the r.h.s. of Eq. (6.20) is an alternate-sign series, hence it will have both contributions in learning (those with the minus sign) and in unlearning (those with the plus sign [98, 97, 47]).

Once this Hamiltonian of the model is introduced and discussed, we can introduce partition function, Boltzmann factor and Boltzmann-Gibbs averages, free energy and mechanical analogy properly from the previous Sections, e.g.

Definition 6.5. The partition function of the Hopfield model is

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}} \exp\left\{\beta N \sqrt{1 + \frac{1}{N^2}\sum_{\mu=1}^{P}\sum_{i,j=1}^{N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j}\right\}, \tag{6.21}$$

with associated intensive pressure $\alpha(\beta) = -\beta f(\beta)$ in the thermodynamic limit

$$\alpha(\beta) = \lim_{N\to\infty} \alpha_N(\beta) = \lim_{N\to\infty} \frac{1}{N}\log Z_N(\beta). \tag{6.22}$$

In order to develop the model in the Hamilton-Jacobi framework, we need the following

Definition 6.6. The generalized partition function and the action of the relativistic Hopfield model, suitable for the mechanical analogy, read as

$$Z_N(\beta;t,\boldsymbol{x}) \doteq Z_N(t,\boldsymbol{x}) = \sum_{\boldsymbol{\sigma}} \exp\left\{N(-t\sqrt{1+\boldsymbol{m}^2}+\boldsymbol{x}\cdot\boldsymbol{m})\right\}, \tag{6.23}$$

$$\alpha_N(t,\boldsymbol{x}) = \frac{1}{N}\log Z_N(t,\boldsymbol{x}). \tag{6.24}$$

Remark 6.6. We stress that the entire partition function can be expressed in covariant form. Indeed, we can endow the $P+1$-dimensional space with a Minkowskian signature $(+,-,\dots,-)$, to which is associated a pseudo-Riemmanian metric tensor $\eta_{AB} = \mathrm{diag}(+,-,\dots,-)$ with $A,B = 0$ (time), $1,\dots P$ (space). Then, the exponent in the Boltzmann factor can be written as $-Nx_A p^A = -N\eta_{AB}x^A p^B$ in the Einstein sum notation, with $x^A = (t,\boldsymbol{x})$ and $p^A = (\sqrt{1+\boldsymbol{m}^2},\boldsymbol{m})$.

The expectation values $\omega_{t,\boldsymbol{x}}(\cdot)$ and $\langle\cdot\rangle_{t,\boldsymbol{x}}$ (if needed) are straightforwardly defined as always. Again, we stress that, by setting $t=-\beta$ and $\boldsymbol{x}=0$ we obtain the relativistic statistical mechanical framework we are interested in.

The next step is the computation of the computing the space-time derivatives of the free energy. The relevant ones are

$$\begin{aligned}
\frac{\partial\alpha_N(t,\boldsymbol{x})}{\partial t} &= -\omega_{t,\boldsymbol{x}}(\sqrt{1+\boldsymbol{m}^2}),\\
\frac{\partial\alpha_N(t,\boldsymbol{x})}{\partial x^\mu} &= \omega_{t,\boldsymbol{x}}(m^\mu),\\
\frac{\partial^2\alpha_N(t,\boldsymbol{x})}{\partial t^2} &= N(\omega_{t,\boldsymbol{x}}(1+\boldsymbol{m}^2) - \omega_{t,\boldsymbol{x}}(\sqrt{1+\boldsymbol{m}^2})^2),\\
\nabla^2_{\boldsymbol{x}}\alpha_N(t,\boldsymbol{x}) &= N(\omega_{t,\boldsymbol{x}}(\boldsymbol{m}^2) - \omega_{t,\boldsymbol{x}}(\boldsymbol{m})^2).
\end{aligned} \tag{6.25}$$

Then, we can state the following

Proposition 6.2. *By construction, the intensive pressure $\alpha_N(t,\boldsymbol{x})$ obeys the following (relativistic) Hamilton-Jacobi PDE*

$$\partial_t^2\,\alpha_N - \nabla^2_{\boldsymbol{x}}\alpha_N = N\left(1-(\partial_t\alpha_N)^2+(\nabla_{\boldsymbol{x}}\alpha_N)^2\right), \tag{6.26}$$

or in the manifestly covariant form

$$(\partial_A\alpha_N)^2 + \frac{1}{N}\Box\,\alpha_N = 1, \tag{6.27}$$

where $\Box$ is the D'Alambertian differential operator, i.e. $\Box = \partial_A\partial^A$ (still in Einstein notation).

Remark 6.7. By requiring that the derivatives of the intensive pressure are regular functions in x^μ and t, in the thermodynamic limit, a.s.[1] we have the simpler differential equation

$$(\partial_A \alpha)^2 = 1. \tag{6.28}$$

From the mechanical perspective, in the thermodynamic limit, the $P+1$-momentum of the particle reads as

$$p^A = -\frac{\partial \alpha}{\partial x_A} = (\omega_{t,\boldsymbol{x}}(\sqrt{1+\boldsymbol{m}^2}), \omega_{t,\boldsymbol{x}}(\boldsymbol{m})). \tag{6.29}$$

In terms of this momentum, the equation for the action (6.28) is nothing but the on-shell relation [86] relating energy, momentum and (unitary) mass. As in the classical case, the fictitious particle in this mechanical analogy moves on the straight lines $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{v}(t - t_0)$ for arbitrary $(t_0, \boldsymbol{x}_0)$. Now, the particle velocity $\boldsymbol{v}$ is related to the spatial momentum through the relation $\omega_{t,\boldsymbol{x}}(\boldsymbol{m}) = \gamma \boldsymbol{v}$, with γ being the Lorentz factor. By well-known relation, we can write

$$\boldsymbol{v} = \frac{\omega_{t,\boldsymbol{x}}(\boldsymbol{m})}{\sqrt{1+\omega_{t,\boldsymbol{x}}(\boldsymbol{m})}}. \tag{6.30}$$

Then, the Lorentz in terms of the (spatial) momentum is $\gamma = \sqrt{1+\omega_{t,\boldsymbol{x}}(\boldsymbol{m})^2}$. Summing all these observations together, we end up in the determination of an explicit expression for the relativistic free energy in terms of the Mattis magnetizations as it reduces to the calculation of the action of this free motion. As Cauchy conditions, we still choose $t_0 = 0$, such that

$$\begin{aligned} \alpha(t, \boldsymbol{x}) &= \alpha(0, \boldsymbol{x}_0) + \int_0^t dt' \mathcal{L}(t') = \\ &= \alpha(0, \boldsymbol{x}_0) - \frac{t}{\gamma} = \alpha(0, \boldsymbol{x}_0) - \frac{t}{\sqrt{1+\omega_{t,\boldsymbol{x}}(\boldsymbol{m})^2}}, \end{aligned} \tag{6.31}$$

since the Lagrangian $\mathcal{L} = -\gamma^{-1}$ is constant on classical trajectories. Thus, again we have

$$\alpha(0, \boldsymbol{x}_0) = \log 2 + \mathbb{E} \log \cosh \boldsymbol{x}_0 \cdot \boldsymbol{\xi}. \tag{6.32}$$

Noting that $\boldsymbol{x}_0 = \boldsymbol{x} - \boldsymbol{v}t$ with (6.30). Then, setting $t = -\beta$, $\boldsymbol{x} = 0$ (in order to re-obtain the statistical mechanical framework) and writing as $\boldsymbol{m}$ the thermodynamic value of the Mattis magnetizations (by virtue of their self-averaging properties), we can state the next

[1] *Almost surely* because when ergodicity breaks down a gradient's catastrophe prevents regularity even in the infinite volume limit [20].

Theorem 6.2. *The free energy density of the relativistic Hopfield network in the thermodynamic limit reads as*

$$\alpha(\beta) = \log 2 + \mathbb{E} \log\cosh\Big(\beta\, \boldsymbol{\xi} \cdot \frac{\boldsymbol{m}}{\sqrt{1+\boldsymbol{m}^2}}\Big) + \frac{\beta}{\sqrt{1+\boldsymbol{m}^2}}. \tag{6.33}$$

By virtue of the extremality condition, the order parameters satisfy the self-consistency equations

$$m_\mu = \mathbb{E}\, \xi^\mu \tanh\Big(\beta\, \boldsymbol{\xi} \cdot \frac{\boldsymbol{m}}{\sqrt{1+\langle\boldsymbol{m}\rangle^2}}\Big), \tag{6.34}$$

for each $\mu = 1, \dots, P$.

Remark 6.8. Since the intensive pressure is related to the Hamilton-Jacobi action in the mechanical analogy, the extremality condition for the free energy is nothing but the Least Action Principle.

Remark 6.9. We also notice that, if we take the low momentum limit $|\boldsymbol{m}| \ll 1$, we can expand the relativistic model at the lowest order

$$\begin{aligned} \frac{1}{\sqrt{1+\boldsymbol{m}^2}} &= 1 - \frac{\boldsymbol{m}^2}{2} + \mathcal{O}\left(\boldsymbol{m}^3\right), \\ \frac{\boldsymbol{m}}{\sqrt{1+\boldsymbol{m}^2}} &= \boldsymbol{m} + \mathcal{O}\left(\boldsymbol{m}^3\right), \end{aligned} \tag{6.35}$$

so recovering the classical Hopfield model and results.

6.3 The thermodynamic limit

In this Section, we shall move to prove the existence of thermodynamic limit of the intensive pressure $\alpha(\beta)$ for the relativistic Hopfield model. The method we will apply is based on a slight modification the Guerra-Toninelli scheme which was originally carried out for pairwise models. We therefore interpolate the pressure for the system with N neurons and two other non-interacting systems (consisting in N_1 and N_2 neurons respectively such that $N = N_1 + N_2$). In doing this, we will apply the Fekete Lemma [112] in order to show that the extensive free energy of the first system is strictly smaller than the sum of those pertaining to the two subsystems (i.e. free energy is sub-additive). In this case, the main adaptation will consist in a proof by reduction to absurd, assuming the free energy to be super-additive.Note that, in the proof, we will omit the β-dependence without loosing in generality.

Let us start by introducing the relative Mattis magnetization of the two subsystems as

$$m_1^\mu = \frac{1}{N_1}\sum_{i=1}^{N_1}\xi_i^\mu\sigma_i, \quad m_2^\mu = \frac{1}{N_2}\sum_{i=1}^{N_2}\xi_i^\mu\sigma_i,$$

so that the global order parameter can be expressed as

$$\begin{aligned} m^\mu &= \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i = \frac{1}{N}\Big(\sum_{i=1}^{N_1}\xi_i^\mu\sigma_i + \sum_{j=1}^{N_2}\xi_j^\mu\sigma_j\Big) = \\ &= \rho_1\boldsymbol{m}_1 + \rho_2\boldsymbol{m}_2, \end{aligned}$$

with the relative densities $\rho_i = N_i/N$ for $i = 1, 2$. Analogously to what we did in the CW and the SK cases, let us introduce an interpolating parameter $t \in [0,1]$ and make the following

Definition 6.7. The interpolating pressure $\alpha_N(t)$ is defined as follows

$$\begin{aligned} \alpha_N(t) = \frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big\{tN\sqrt{1+\boldsymbol{m}^2} \\ + (1-t)\Big(N_1\sqrt{1+\boldsymbol{m}_1^2} + N_2\sqrt{1+\boldsymbol{m}_2^2}\Big)\Big\}. \end{aligned} \tag{6.36}$$

Remark 6.10. Of course, the two limits recover the two different cases. Indeed,

$$\alpha_N(1) = \alpha_N, \tag{6.37}$$

$$\alpha_N(0) = \rho_1\alpha_{N_1} + \rho_2\alpha_{N_2}. \tag{6.38}$$

In the following, we will suppress the β-dependence of the intensive pressure. The original model with N interacting neurons is therefore recovered as

$$\alpha_N(1) = \alpha_N(0) + \int_0^1 ds[\partial_t\alpha_N(t)]_{t=s}. \tag{6.39}$$

Since the integral operator is monotonous (i.e. it respects inequalities), it is sufficient to prove that the derivative of the interpolating free energy w.r.t. t has a negative semi-definite sign. The introduction of the interpolating intensive pressure (and therefore the generalized partition function $Z_N(t)$), it is straightforward to introduce the Boltzmann-Gibbs average $\omega_t(\cdot)$. Then, it is easy to prove the following equality

$$\frac{\partial\alpha_N}{\partial t} = \omega_t\Big(\sqrt{1+\boldsymbol{m}^2} - \rho_1\sqrt{1+\boldsymbol{m}_1^2} - \rho_2\sqrt{1+\boldsymbol{m}_2^2}\Big). \tag{6.40}$$

Now, since $N = N_1 + N_2$, we trivially have $\rho_2 = 1 - \rho_1$. Therefore

$$\begin{aligned}\sqrt{1+\boldsymbol{m}^2} - \rho_1\sqrt{1+\boldsymbol{m}_1^2} - \rho_2\sqrt{1+\boldsymbol{m}_2^2} = \\ = \sqrt{1+\left(\rho_1\boldsymbol{m}_1 + \left(1-\rho_1\right)\boldsymbol{m}_2\right)^2} - \rho_1\sqrt{1+\boldsymbol{m}_1^2} - \left(1-\rho_1\right)\sqrt{1+\boldsymbol{m}_2^2}.\end{aligned} \tag{6.41}$$

Now, let us suppose for a moment that this is a positive quantity. Then, with simple algebraic manipulations, we have

$$2\rho_1\left(1-\rho_1\right) > 2\rho_1\left(1-\rho_1\right)\left(\sqrt{\left(1+\boldsymbol{m}_1^2\right)\left(1+\boldsymbol{m}_2^2\right)} - \boldsymbol{m}_1\boldsymbol{m}_2\right),$$

Since $2\rho_1\left(1-\rho_1\right) > 0$, then the only possibility to hold is that the quantity in rounded brackets on the r.h.s. is less than 1, i.e.

$$\sqrt{\left(1+\boldsymbol{m}_1^2\right)\left(1+\boldsymbol{m}_2^2\right)} < 1 + \boldsymbol{m}_1\boldsymbol{m}_2.$$

The l.h.s. is always non-negative (since, in the worst case, $\boldsymbol{m}_1\boldsymbol{m}_2 = -1$, which directly leads to an absurd). Therefore, taking the square of this equality and then with simple manipulations, we find

$$\left(\boldsymbol{m}_2^2 - \boldsymbol{m}_1^2\right)^2 < 0, \tag{6.42}$$

which is impossible, therefore proving the next

Proposition 6.3. *The t-derivative of the interpolating free energy (6.36) is semi-definite negative, namely*

$$\sqrt{1+\boldsymbol{m}^2} - \rho_1\sqrt{1+\boldsymbol{m}_1^2} - \rho_2\sqrt{1+\boldsymbol{m}_2^2} \leq 0. \tag{6.43}$$

Remark 6.11. We also highlight that the function $f : x \mapsto \sqrt{1+x^2}$ is convex, which is, for $\lambda \in [0,1]$

$$f\left(\lambda x_1 + \left(1-\lambda\right)x_2\right) \leq \lambda f\left(x_1\right) + \left(1-\lambda\right)f\left(x_2\right).$$

Therefore, identifying

$$\lambda = \rho_1, \qquad x_1 = \boldsymbol{m}_1, \qquad x_2 = \boldsymbol{m}_2,$$

then it follows immediately that

$$\sqrt{1+\boldsymbol{m}^2} \leq \rho_1\sqrt{1+\boldsymbol{m}_1^2} + \rho_2\sqrt{1+\boldsymbol{m}_2^2}.$$

By using Proposition (6.3), we have

$$\alpha_N\left(1\right)-\alpha_N\left(0\right)=\int_0^1 ds[\partial_t\alpha_N(t)]_{t=s}\leq 0,$$

or more transparently

$$N\alpha_N\leq N_1\alpha_{N_1}+N_2\alpha_{N_2}. \tag{6.44}$$

Then, by applying the Fekete lemma, the following

Theorem 6.3. *The infinite volume limit of the intensive pressure defined by the relativistic Hopfield cost function in the low storage regime exists and it equals its infimum, that is*

$$\exists\lim_{N\to\infty}\alpha_N(t=1)=\inf_{N\in\mathbb{N}}\left\{\alpha_N(t=1)\right\}=\alpha.$$

is proven. A similar procedure can be carried out for all $\beta\in\mathbb{R}^+$.

6.4 Guerra's interpolating scheme

To end the analytical treatment of relativistic Hopfield model in the low storage, we will confirm the expression for the intensive pressure in terms of the Mattis overlaps by a standard statistical mechanics route. In doing this, we will use again Guerra's interpolation scheme, whose philosophy is by now crystal clear. Then, by re-introducing the interpolation parameter $t\in[0,1]$, we introduce the following

Definition 6.8. The interpolating free energy is

$$\begin{aligned}\alpha_N(\beta;t)\doteq\alpha_N(t)=&\\ =\frac{1}{N}\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\exp\Big(t\beta N\sqrt{1+\boldsymbol{m}^2}+(1-t)\beta\boldsymbol{\psi}\cdot\boldsymbol{m}\Big),&\end{aligned} \tag{6.45}$$

where $\boldsymbol{\psi}=(\psi^1,\dots\psi^P)$ are tunable auxiliary fields.

Needless to say, from the interpolating partition function it straightforwardly follows the definition of the Boltzmann-Gibbs averages $\omega_t(\cdot)$. As always, we will use the sum rule in order to reduce the problem of finding the intensive pressure (in the thermodynamic limit) to a one-body system computation and the t-derivative correction. It is easy to show that, for the former, we easily get

$$\alpha(t=0)=\log 2+\mathbb{E}\log\cosh\Big(\beta\sum_{\mu=1}^P\psi^\mu\langle m_\mu\rangle\Big). \tag{6.46}$$

On the other side, the computation of the t-derivative leads to

$$\frac{\partial \alpha}{\partial t} = \beta\omega_t\Big(\sqrt{1+\omega_t(\boldsymbol{m})^2} - \sum_{\mu=1}^{P}\psi^\mu\omega_t(\boldsymbol{m})\Big), \tag{6.47}$$

In general, the evaluation of this quantity is an hard task. However, in the thermodynamic limit, calling $\bar{m}_\mu$ the equilibrium values of the μ-th Mattis overlap (i.e. $\lim_{N\to\infty} P(\boldsymbol{m}) = \delta(\boldsymbol{m}-\bar{\boldsymbol{m}})$), by requiring for the self-averaging of the order parameters *and* the energy to hold almost surely [17, 24, 59], which is

$$\lim_{N\to\infty}\sum_{\mu=1}^{P}\omega_t((m_\mu-\bar{m}_\mu)^2) = 0, \tag{6.48}$$

$$\lim_{N\to\infty}\omega_t((\sqrt{1+\boldsymbol{m}^2}-\sqrt{1+\bar{\boldsymbol{m}}^2})^2) = 0, \tag{6.49}$$

we obtain

$$\omega_t\Big(\sqrt{1+\boldsymbol{m^2}} - \frac{\boldsymbol{m}\cdot\bar{\boldsymbol{m}}}{\sqrt{1+\bar{\boldsymbol{m}}^2}} + \frac{1}{\sqrt{1+\bar{\boldsymbol{m}}^2}}\Big) = 0.$$

Comparing the above equation with the r.h.s. of (6.47) and choosing $\psi^\mu = \frac{\bar{m}_\mu}{\sqrt{1+\bar{\boldsymbol{m}}^2}}$, we can write

$$\frac{\partial\alpha(t)}{\partial t} - \frac{\beta}{\sqrt{1+\bar{\boldsymbol{m}}^2}} = 0. \tag{6.50}$$

By merging (6.50) with the Cauchy condition (6.46), we are finally able to state the next

Theorem 6.4. *The infinite volume limit of the intensive pressure defined by the relativistic Hopfield cost function in the low storage regime in terms of the Mattis magnetizations reads as*

$$\alpha(\beta) = \log 2 + \mathbb{E}\log\cosh\Big(\beta\boldsymbol{\xi}\cdot\frac{\bar{\boldsymbol{m}}}{\sqrt{1+\bar{\boldsymbol{m}}^2}}\Big) + \frac{\beta}{\sqrt{1+\bar{\boldsymbol{m}}^2}}, \tag{6.51}$$

and the related self-consistency equations for the Mattis magnetizations are

$$\bar{m}_\mu = \mathbb{E}\,\xi^\mu\tanh\Big(\beta\boldsymbol{\xi}\cdot\frac{\bar{\boldsymbol{m}}}{\sqrt{1+\bar{\boldsymbol{m}}^2}}\Big). \tag{6.52}$$

Remark 6.12. This results are in perfect agreement with our findings with the previously exploited Hamilton-Jacobi framework, recalling that, in those equations, $\boldsymbol{m}$ stands for the equilibrium values of the Mattis overlaps.

We would like to conclude this Section with the analysis of the critical behavior of the system. Taking, without loss of generality, only ξ^1 as the candidate pattern to be retrieved, it is convenient to consider the following

Definition 6.9. The rescaled Mattis overlap associated to the pattern ξ^1 is

$$\hat{m}_1 = \sqrt{N}(m_1 - \bar{m}_1), \tag{6.53}$$

where, as usual, $m_1 = N^{-1}\sum_{i=1}^N \xi_i^1 \sigma_i$ while $\bar{m}_1$ is its thermodynamic limit, namely $\lim_{N\to\infty} \omega_t(m_1^2) \to \bar{m}_1$.

Notice that $\omega_t(\hat{m}_1^2)$ scales as N times the variance of $\omega_t(m_1^2)$. The analysis of critical behaviour of the relatistic Hopfield model is then performed by studying $\omega_t(\hat{m}_1^2)$ as a function of β. What we are looking for are those values β_c for which the fluctuations of the order parameters (with respect to its mean value $\bar{\boldsymbol{m}}$) diverge.

To this goal, we again exploit the interpolation scheme (6.45), and then set $t = 1$ to find $\omega(\hat{m}_1^2)$. In the same spirit of what we did for the intensive pressure, we write

$$\omega(\hat{m}_1^2) = \omega_{t=0}(\hat{m}_1^2) + \int_0^1 ds[\partial_t \omega_t(\hat{m}_1^2)]_{t=s}. \tag{6.54}$$

We start by evaluating the expectation value at $t = 0$. To do this, it is useful to approach the critical line from the high noise region. The advantage in such a procedure lies in the fact that, in this region, we can take benefit of CLT-like arguments to assume the probability distribution of the $\hat{m}_1$ is a Gaussian.[1] Then, the Cauchy condition can be easily evaluated as $\omega_{t=0}(\hat{m}_1^2)$. Indeed, we have

$$\begin{aligned}\omega_{t=0}(\hat{m}_1^2) &= \lim_{N\to\infty} \omega_{t=0}(N(m_1 - \bar{m}_1)^2) = \\ &= \lim_{N\to\infty} [1 + (N-1)\bar{m}_1^2 + N\bar{m}_1^2 - 2N\bar{m}_1^2] = \\ &= 1 - \bar{m}_1^2.\end{aligned} \tag{6.55}$$

Finally, in the ergodic region, we have trivially $\bar{m}_1 = 0$, so proving the previous statement. We must now face the t-derivative: to this task it is useful to state the next For the t-derivative of the expectation value, it will be useful the following

[1] In this way, we can use Wick theorem $\mathbb{E}(zf(z)) = \mathbb{E}(z^2)\mathbb{E}(\partial_z f(z))$, with z Gaussian random variable.

Proposition 6.4. *Recalling that $\psi_\mu = \bar{m}_\mu/(\sqrt{1+\bar{\boldsymbol{m}}^2})$, let F be a smooth function of the Mattis overlaps. Then, above and close to the critical point (where the Mattis overlaps are zero or infinitesimal), the following streaming equation holds:*

$$\frac{d}{dt}\omega_t(F) \sim \frac{\beta}{2}(\omega_t(F\hat{\boldsymbol{m}}^2) - \omega_t(F)\omega_t(\hat{\boldsymbol{m}}^2)). \tag{6.56}$$

To prove this Proposition, let us first note that

$$\begin{aligned}\frac{d}{dt}\omega_t(F) = \beta N\Big(&\omega_t(\sqrt{1+\boldsymbol{m}^2}) - \sum_{\mu=1}^{P}\frac{\bar{m}_\mu}{\sqrt{1+\bar{\boldsymbol{m}}^2}}\omega_t(Fm_\mu) \\ &- \omega_t(F)\omega_t(\sqrt{1+\boldsymbol{m}^2}) + \sum_{\mu=1}^{P}\frac{\bar{m}_\mu}{\sqrt{1+\bar{\boldsymbol{m}}^2}}\omega_t(F)\omega_t(m_\mu)\Big).\end{aligned} \tag{6.57}$$

Now, since we are approaching the critical line from the high noise region (where the Mattis overlap changes continuously from zero to a non-vanishing value), we can expand $\sqrt{1+x^2} \sim 1 + x^2/2$ and $1/(\sqrt{1+x^2}) \sim 1 - x^2/2$ in the above equation. Finally, adding and subtracting twice $(\beta N/2)\omega_t(F)\hat{m}^2$ we get the proof of the above statement. Then, with the above results we get the following Cauchy problem for the variance of the rescaled Mattis overlap:

$$\frac{d}{dt}\omega_t(\hat{m}_1^2) = \beta\omega_t(\hat{m}_1^2)^2, \tag{6.58}$$

$$\omega_{t=0}(\hat{m}_1^2) = 1. \tag{6.59}$$

By solving the Cauchy problem, we can easily state the following

Theorem 6.5. *The centered and rescaled fluctuations of the Mattis overlap $\omega_t(m_1)$ associated to the retrieved pattern ξ^1 behaves as*

$$\omega(\hat{m}_1^2) = \omega_{t=1}(\hat{m}_1^2) = \frac{1}{1-\beta}, \tag{6.60}$$

above the critical line. Clearly, the ergodic region is limited to $\beta < \beta_c = 1$. In the low-noise (i.e. non-ergodic) region, the Mattis overlap may assume non-vanishing values.

6.5 Numerical Simulations

We would like to conclude this Chapter by presenting numerical simulations for the relativistic Hopfield model, always comparing our outcomes to the classical Hopfield model reference. More precisely, we performed extensive Monte Carlo (MC) simulations to check the retrieval capabilities of both models. As a preliminary check, we compared the numerical results with the theoretical predictions for the self-consistency equations. We also performed a numerical study aiming in describing how the spurious attraction basins are reduced in the relativistic Hopfield model w.r.t to its classical counterpart. In the first part of this Section, we will give a brief description of numerical methods we adopted in our MC simulations, then we move to the presentation of the results and their discussion.

6.5.1 Stochastic neural dynamics

The first step is to point the neural dynamics we adopted in our MC simulations. Our implementation follows the standard Glauber dynamics

$$\begin{aligned} \sigma_i(t+1) &= \operatorname{sign}\left[\tanh\left(\beta h_i(\boldsymbol{\sigma}(t))\right) + \eta_i(t)\right] && (6.61)\\ h_i(\boldsymbol{\sigma}(t)) &= \boldsymbol{\xi}\cdot\frac{\boldsymbol{m}}{\sqrt{1+\boldsymbol{m}^2}}, && (6.62)\end{aligned}$$

where η_i are random variables uniformly sampled from $[-1,+1]$. Of course, the update rule is formally unchanged w.r.t to the classical Hofpield model: the only difference between the relativistic Hopfield model. The parameter β clearly tunes the noise level in the network, triggering the amplitude of the hyperbolic tangent (the signal term). In particular, for $\beta \to \infty$ (which is the zero-temperature limit), the hyperbolic tangent becomes a ± 1 step function, meaning that stochasticity is carried out from the update rule and the dynamics is deterministic. On the contrary, for $\beta \to 0$, the hyperbolic tangent returns zero for each value of the signal, and the dynamics becomes fully random.

We stress also that the entire stochatistic neural dynamics could be framed in probabilistic form, by noting that Glauber dynamics can be redefined as

$$\begin{aligned} \mathcal{P}_{t+1}(\boldsymbol{\sigma}) &= \sum_{\sigma'} W[\boldsymbol{\sigma'}\to\boldsymbol{\sigma}]\mathcal{P}_t(\boldsymbol{\sigma'}), && (6.63)\\ W[\boldsymbol{\sigma'}\to\boldsymbol{\sigma}] &= \prod_{i=1}^{N}\frac{e^{\sigma_i h_i(\boldsymbol{\sigma}(t))}}{2\cosh\left(\beta h_i(\boldsymbol{\sigma}(t))\right)}. && (6.64)\end{aligned}$$

Here, $\mathcal{P}_t(\boldsymbol{\sigma})$ is the probability to find (at time t) the network in the state $\boldsymbol{\sigma}$, while $W[\boldsymbol{\sigma'}\to\boldsymbol{\sigma}]$ is the transition probability from the state $\boldsymbol{\sigma'}$ to $\boldsymbol{\sigma}$.

Because of the symmetry of couplings, Detailed Balance holds, ensuring that there exists a stationary probability distribution at $t \to \infty$ such that

$$\mathcal{P}_\infty(\boldsymbol{\sigma})W[\boldsymbol{\sigma}' \to \boldsymbol{\sigma}] = \mathcal{P}_\infty(\boldsymbol{\sigma}')W[\boldsymbol{\sigma} \to \boldsymbol{\sigma}'], \tag{6.65}$$

Since the probability distribution $\mathcal{P}_\infty$ should have the maximum entropy Gibbs-expression $\mathcal{P}_\infty \propto \exp(-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}))$, it is possible to evaluate the neuron-flip probability as

$$\mathcal{P}_t(\sigma_i \to \sigma_i') = \frac{1}{1 + e^{\beta[H(\boldsymbol{\sigma}|\boldsymbol{\xi}) - H(\boldsymbol{\sigma}'|\boldsymbol{\xi})]}},$$

which is nothing but the acceptance criterion of the Glauber algorithm. Then, the MC simulations are carried out with the following method. At each evolution step:

1. Select at random a neuron in the network, and compute the difference $\Delta H(\boldsymbol{\sigma}|\boldsymbol{\xi})$ in the energy after its spin-flip. If $\Delta H(\boldsymbol{\sigma}|\boldsymbol{\xi}) < 0$ (i.e. the flip is convenient), the move is accepted with probability

$$\exp(\beta \Delta H(\boldsymbol{\sigma}|\boldsymbol{\xi}))/[1 - \exp(\Delta H(\boldsymbol{\sigma}|\boldsymbol{\xi}))],$$

   otherwise is rejected (Glauber criterion). Otherwise, if $\Delta H(\boldsymbol{\sigma}|\boldsymbol{\xi}) > 0$ (i.e. the flip is not convenient), the move is rejected;

2. Iterate the rule for a number of evolution steps which is intensive in the network size.

6.5.2 Comparison between theory and MC runs

In this Section, we will check the theoretical predictions and the results of MC simulations carried out as described above: all the simulation have been carried on our group computing cluster, equipped with 12 CPU Intel 3.2 Ghz that survey 512 GPU Nvidia for High Performance Parallel Processing [77, 96] .

To make the comparison exhaustive, we will also include the (classical) Hopfield model in this analysis. From the theoretical side, we solve the self-consistency equations at fixed network size N both for pure and spurious states. In Fig. 6.1 we show the behavior of the Mattis order parameter as a function of the temperature $T = \beta^{-1}$ by comparing numerical solutions of the self-consistency solutions (dashed lines) and MC simulations (data points). The upper plot refers to the pairwise (i.e. classical) Hopfield model (6.7), while the lower one is its relativistic counterpart (6.18).

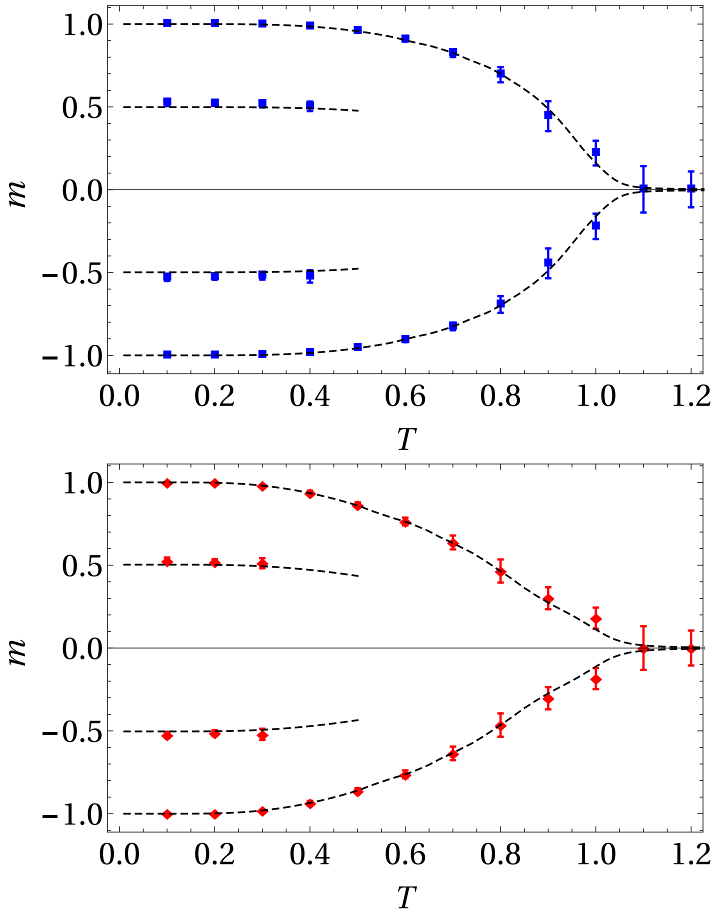

Figure 6.1: **Comparisons between MC simulations and self-consistency solutions.** The black dashed lines are the numerical solutions of the self-consistency equations (6.34), while blue squares and red diamonds are respectively the MC results for classical and relativistic Hopfield models. The network parameters are fixed to $N = 1000$ with $P = 3$ (orthogonal) stored patterns. The data points are the averages over 20 different pattern realizations, for each of which we sampled 20 different stochastic evolution starting with 20 random initial conditions (i.e. 8000 runs at any given noise level T).

Despite redundant, we stress that two branches are due to the underlying spin-flip symmetry: both the pattern ξ^1 and its symmetric partner $-\xi^1$ are attractors for the neural dynamics. Moreover, we highlight the presence of the spurious states (evidenced by the two segments of magnetization's values $m_1 \sim \pm 0.5$) for noise level $T \leq 0.45$. We also stress that, since we are work-

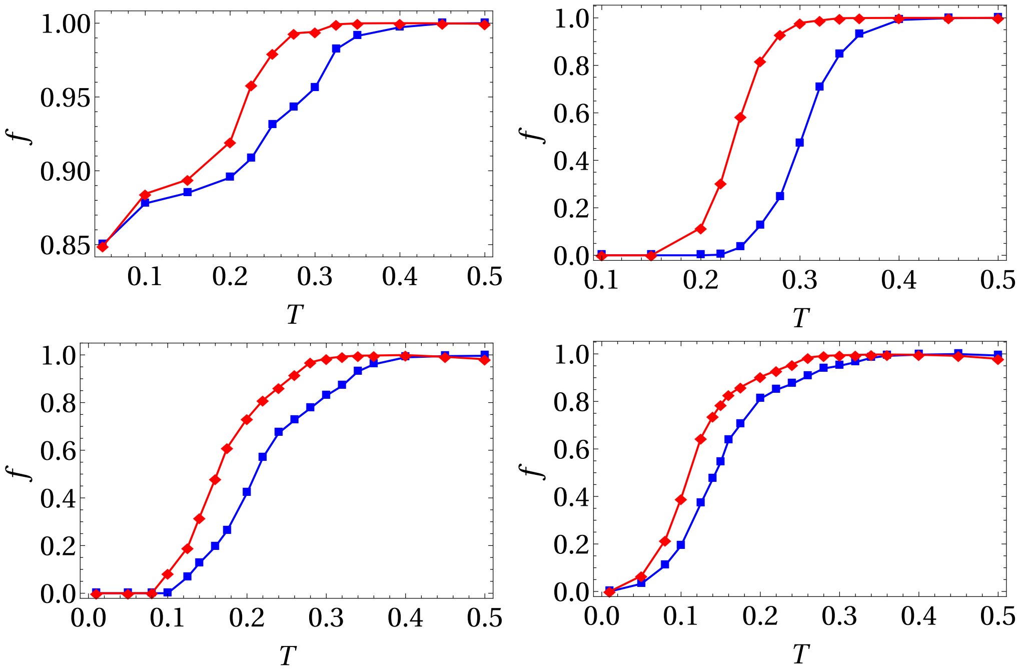

Figure 6.2: **Attraction basins analysis for classical and relativistic Hopfield models.** The plots show the retrieval frequency for classical (blue) and relativistic (red) Hopfield models with $N = 1000$ for random (upper left with $P = 3$) and spurious (with $P = 3, 5, 7$) initial conditions.

ing at finite network size N, there is no phase transition in the statistical mechanics sence. However, the variances of the data-points spread at the bifurcation point (i.e. at $T \sim 1$), therefore signaling that a typical second order phase transition takes place in the thermodynamic limit $N \to \infty$, which is of course confirmed by our theoretical analysis.

6.5.3 Depth of the attractors and energy gaps

This final Section is devoted to a numerical analysis of attraction basins of spurious states. To this aim, we compare the classical and relativistic Hopfield networks (with $N = 1000$) and check the retrieval performances of both models. Therefore, we prepared the system in specific configurations (to be discussed below) and then let the systems evolves with standard stochastic neural dynamics (for different values of the thermal noise T) towards the equilibrium configuration. After the system as relaxed on a stable configuration, we then measure the retrieval frequency f, which is the fraction of evolutions ending in pure state configurations as a function of the noise level T. The analysis can be performed in three different ways, as we shall summarize.

- **Attractors from random initial conditions**

At first, we prepared the system in a fully random initial configuration for a network with $P = 3$ stored random patterns, then we let it thermalize at a given noise level T with the Markov dynamics (6.63). We collect the final state of the relaxation process (for both for the classical and the relativistic models) and then compute the retrieval frequency f. We performed 20 different stochastic evolutions for 20 different initial random conditions and 20 pattern realizations. Results are shown in Fig. 6.2, upper left plot. It emerges that the relativistic model slightly improves the retrieval performances of original Hopfield network, as it can be understood by the fact that the relativistic (red) curve is always above the pairwise counterpart. Then, the relativistic model shows increased performances that its classical counterpart (as long as the spurious states are locally stable).

- **Attractors from spurious initial conditions**

In order to better understand how spurious attraction basins are downsized in the relativistic Hopfield model, we performed the following analysis. We prepared the system sharply within a spurious state (the 3-patterns mixture for $P = 3$, see Eq. (6.1)) and let it thermalize at a given noise level β^{-1} according (6.63). The statistics in this case is composed by 40 different stochastic evolutions for each of the 40 pattern realizations. The results for the retrieval frequency f are reported in the remaining plots of Fig. 6.2 for $P = 3$ (upper right plot), $P = 5$ and $P = 7$ (respectively lower left and right plots). We stress that we analyzed the noise range $T \in [0, 0.5]$, in which spurious states are dynamically stable attractors. For higher noise levels, spurious states are not stable, so that there is generally no more reward in the relativistic extension.

By inspecting the plots, also in this case it is clear that relativistic Hopfield model systematically outperforms w.r.t. the original pairwise one.

- **Attractors from noisy spurious configurations**

In our last analysis (which is indeed inspired by the works of the Gardner on the estimation of depth and stability of the basins of attractions of pure and spurious states [53, 52]), we proceed as follows. We aligned the network (with $P = 3$ random patterns) in the spurious configuration. Then, we randomly spin-flip a percentage d of the neurons (i.e. for e.g. $d = 0$, no random spin-flip at all is performed, while for e.g. $d = 50\%$, one half of the spins are flipped). In other words, we are preparing the system into a known state (e.g. the spurious state), and then we reshuffle it by *kicking* randomly a percentage d of its neurons. Then, we check if the networks returns (or the escapes) from the initial attractor as a function of d. Also in this case, we realized 40 different stochastic evolutions for each of the 40 pattern realizations. Results are shown in Fig. 6.3 for $T = 0.2$ and 0.3, focusing on a 3-mixture spurious state (in order to quantify the pruning capabilities of the relativistic model w.r.t the pairwise Hopfield one).It is clear that, at mild noise level ($T = 0.2$, left plot), spurious states already becomes unstable in the relativistic extension also for very low spin-flips percentage d (to be compared with the classical case, in which they are still stable). For larger values of $d \sim 0.35$, both models improve the retrieval frequency (but the relativistic one always outperforms). Finally, for moderate noise levels ($T = 0.3$, right plot) the relativistic model always escapes from the spurious state (regardless of d), making it clear that the associated attraction basin is strongly corrupted by the unlearning contributions in the Hopfield Hamiltonian (6.18).

In conclusion, relativistic Hopfield model has improved retrieval performances w.r.t. the pairwise counterpart. A reasonable argument supporting these results lies in the fact that energy barrier between the spurious states and the maxima surrounding them are indeed downsized. Indeed, we checked

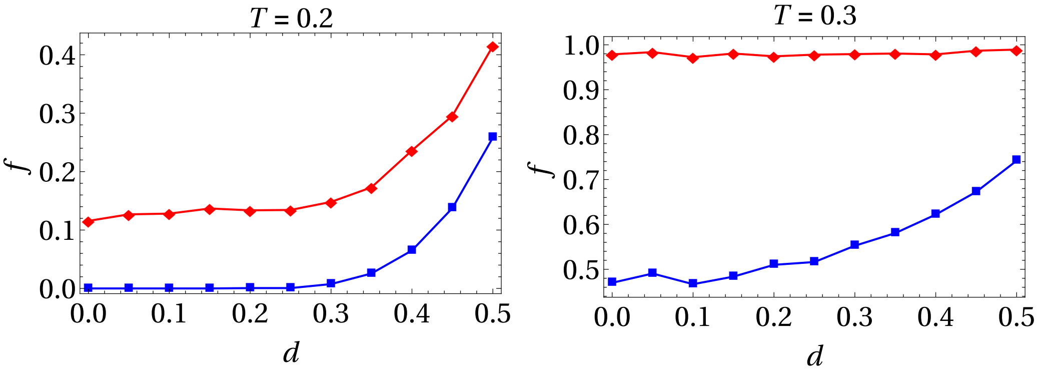

Figure 6.3: **Stability of retrieval performances for noisy spurious inputs.** Results for retrieval frequency as a function of the initial spin-flip fraction d for $T = 0.2$ (left) and 0.3 (right) with $0 \leq d \leq 0.5$. The network parameters are fixed to $N = 1000$ and $P = 3$ stored random patterns.

this idea by preparing the network in a spurious pattern and then performing a noiseless random walk (i.e. not of Glauber type). In other words, at each evolution step, a spin σ_i is selected and (if it is not already aligned to the i-th entry of a given pattern), it is flipped. The procedure is performed until the network is completely aligned to that pattern. In such a walk, an energy barrier has to be crossed. Collecting the energy gaps and averaging on 1600 different samples (i.e. 40 pattern realizations and 40 different stochastic evolutions for each one of them), we find that these energy barriers are strongly downsize up to $\Delta E_{relativ}/\Delta E_{classic} \sim 0.75$. This confirms our hypotesis, also suggesting how to go beyond the standard pairwise Hopfield model in order to account network pruning.

Chapter 7

Beyond the standard paradigm: Sleeping for high storage

In the last Chapter, we started to see how more sophisticated neural network models can be carried out in order to overcome the intrinsic limitations of Hopfield paradigm. In our case, this was achieved by adding an infinite series of P-spin contributions, showing how this leads to an enhancement of retrieval performances (at least in the low storage limit). This can also be understood by simply looking at the critical storage capacity. As remarked by Krotov and Hopfield in [82], P-spin neural network models allow for a critical threshold which grows more than linearly with the network size (namely, assuming n-body coupling between the neurons, the critical capacity roughly grows as N^{n-1}). However, also keeping the model to present pairwise interactions, it is still possible to enhance the critical capacity by saturating the Gardner upper bound ($\lambda_c = 1$). Indeed, we aforementioned the so-called *unlearning*, whose core idea is to suitably modify the interactions and make spurious configurations less stable. Even if we implemented it in the alternating signs series (i.e. the Taylor expansion of relativistic Hopfield Hamiltonian), unlearning was originally developed for the pairwise model. Much progress have been made in the last two decades (see for instance [35, 66, 98, 97, 47, 79, 134, 40, 44, 87, 88, 90, 100, 106, 105] and references therein) about unlearning in the standard Hopfield model, however a comprehensive picture through statistical mechanics was not still carried out systematically and rigorously: this will be the goal of this final Chapter.

We will keep the paradigmatic Hopfield model as standard reference and implement it by including reinforcement and remotion features in such a way that they are able to work simultaneously during the network *sleep* (as inspired by real sleeping and dreaming mechanisms in mammal brains). Strongly oversimplifying, a sleeping session can be split in two different modes, i.e. *rapid eye movement* (REM) and *slow wave* (SW) sleeps. While the former yields to erasure of unnecessary memories, the latter achieves in consolidating of the important ones [14, 107, 123]. In the Literature on Artificial Intelligence, reinforcement (of pure states) and remotion (of spurious states) are addressed in separate ways (see e.g. [66, 135, 73, 124]). In this Chapter, we review our recent results reported in [50, 2], in which we proposed a unified framework for synaptic plasticity accounting both for reinforcement and remotion. As we will show, the combination of these two features leads to an enlargement of region (in the space of the parameters), but most remarkably to the disappearance of the spin glass phase.

Before going further on the subject, let us now deepen the context of our work. Since the Hopfield model threshold $\lambda_c \sim 0.14$ is far away from Gardner's upper bound $\lambda_c = 1$, scientists tried to improve its retrieval performances by implementing some extensions and variations on theme (e.g. keeping the network out of equilibrium [38, 42] or allowing the network to process multiple tasks simultaneously [9, 8, 7]). A crucial inspiration came with Crick and Mitchinson's paper [39], in which Authors argued that the REM phase of sleep is associated to a reverse learning mechanism removing irrelevant information (in order to save memory and avoid overloading catastrophes), see also [14, 107, 123] for empirical evidences and [47, 68, 18, 98, 97, 104, 135] for theoretical investigations. Another interesting point in making a link between AI and sleeping (and in particular REM phase) is that generally dreams are not entirely uncorrelated with learned (i.e. experienced) informations, as well as (as we discussed above) spurious attractors in the Hopfield model shows unavoidably implies short-length correlations with pure memories.

A first step towards the implementation of sleeping in AI is due to Hopfield himself (together with Feinstein and Palmer [66]). As we already discussed in this thesis, the key observation lies in the fact that Hopfield model fails to retrieve stored information when the number of spurious states are exponentially more abundant than the number of pure states (regardless their depth in the free energy landscape). This means that, making a quench from $T \to \infty$ to $T = 0$, the system will end more likely in configurations which are uncorrelated with the stored patterns (i.e. spurious states). By sampling a number of these final configurations, one can quantify the two-point correlation functions $\langle \sigma_i \sigma_j \rangle_{\text{exp}}$ for each i and j. The recipe they proposed to face this problem is to update the synaptic matrix according to an *inverse* Hebbian learning, so that uncorrelated configurations are increase in energy and become less stable. In mathematical terms, Hopfield's proposal consists in the update algorithm

$$J_{ij} \to J_{ij} - \frac{\epsilon}{N}\langle \sigma_i \sigma_j \rangle_{\text{exp}} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu - \frac{\epsilon}{N}\langle \sigma_i \sigma_j \rangle_{\text{exp}}, \tag{7.1}$$

with ϵ is a tunable (but small) parameter, the so-called *unlearning strength*, and the subscript exp again means that we are dealing with quantity which are averaged on an experimental sample. This rule should be re-iterated in order to clean the minima landscape from unwanted attractors. In particular, we stress that the minus sign in (7.1) is central, since it has to increase the energetic value relative to such spurious configurations. Hopfield's rule is only one between the infinite possible algorithmic choice. Indeed, many unlearning scenario have been already proposed in the Literature, all sharing the same core idea. Among these rules, an interesting choice is due to Plakhov and Semenov [104]. Their scheme consists in replacing the pure pairwise correlations between spins with correlations between internal fields, namely

$$J_{ij} \to J_{ij} - \epsilon \langle h_i h_j \rangle_{\text{exp}}. \tag{7.2}$$

The interesting point (and the main advantage) w.r.t to Hopfield rule is that, with a suitable choice of the unlearning strength, this algorithm is ensured to converge (up to scaling factors) to the projector (or pseudo-inverse) matrix

$$J_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{P} \xi_i^\mu (C^{-1})_{\mu,\nu} \xi_j^\nu, \tag{7.3}$$

where

$$C_{\mu,\nu} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu, \tag{7.4}$$

is the 2-point correlation matrix between the patterns. Remarkably, the model (7.3) naturally emerges when requiring that stored patterns are dynamically stable [80, 75], regardless if they are random and uncorrelated or encode information of a structured dataset (for a nice discussion, see also [11]). Notably, the model (7.3) shows a storage capacity reaches $\alpha_c = 1$. From the point of view of unlearning *à la* Plakhov&Semenov, the statistical mechanics of the continuous-time limit (i.e. $\epsilon \sim dt$) of (7.3), which is realized by the coupling matrix[1]

$$J_{ij}(t) = \frac{1}{N} \sum_{\mu,\nu=1}^{P} \xi_i^\mu (\mathbb{I} + tC)^{-1}_{\mu,\nu} \xi_j^\nu, \tag{7.5}$$

where $\mathbb{I}$ is the identity matrix and $t \in \mathbb{R}^+$ is the sleeping time, was studied by Dotskenko and coworkers [47, 46]. When realizing the phase diagram of the model, it turns out that the maximal storage capacity increases as t gets larger and larger, approaching the Gardner upper bound.[2] However, in the large t limit (for which the storage capacity is maximal), the coupling matrix identically vanishes. This signals that unlearning scenario (7.2) also affects pure memories, and in the infinite sleeping time limit, all stored information is destroyed. Indeed, the retrieval region is stretched toward higher values in α with respect to the Hopfield reference, but at the same time it is also confined to smaller values of T, ultimately disappearing when $t \to \infty$.

In the rest of the Chapter, we will review our model, which is forced to interpolate between Hopfield and [103, 75] model, and show how this is well-defined and thermodynamical stable (meaning that retrieval region is still present in the large sleeping time limit) with a critical capacity saturating the Gardner upper bound. A note on methodologies is in order here. As we usually did throughout this thesis, once introduced the model, we first solve it by means of the replica trick technique, then we confirm our findings by following the Guerra-Toninelli interpolation schemes. In particular, the latter makes it possible to deal also with the analysis of order parameter fluctuations, therefore giving a rigorous basis to our results.

Since the plan is now clear, we are now in position to introduce the model according to the following:

Definition 7.1. Consider a network composed by N neurons $\{\sigma_i\}_{i=1,\dots,N}$, with $\sigma_i \in \{-1, +1\}$ $\forall i$, and P Boolean patterns ξ^μ, with $\mu = 1, \dots, P$. For each sleeping time $t \in \mathbb{R}^+$, the *reinforcement&removal* model is described by

[1] While quite marginal in AI, we stress that such a learning rule is *non-local*, since the coupling between neurons i and j now depends on pattern entries related to all the neurons in the system (and this is a criticality from the biological point of view). A local algorithm which is known to converge towards the projector matrix is the called Adeline learning rule (see [76, 135] for an overview.)

[2] Actually, the critical threshold found by [47] is approximately 1.07. This is not to be meant as a violation of Garner's bound: the overflow is due to the underlying replica-symmetry approximation.

the Hamiltonian[1]

$$H_{N,P}(\boldsymbol{\sigma}|\boldsymbol{\xi},t) = -\frac{1}{2N}\sum_{i,j=1}^{N}\sum_{\mu,\nu=1}^{P}\xi_i^\mu\xi_j^\nu\left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu}\sigma_i\sigma_j, \tag{7.6}$$

where the P patterns are i.i.d. extracted according to the probability

$$P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = \frac{1}{2}, \quad \forall i = 1,\dots,N, \forall\mu = 1,\dots,P,$$

and the correlation matrix is defined as

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

Remark 7.1. Note that the interpretation of t as the sleep extent is clear: for $t=0$ the system reduces to the standard Hopfield model, while for $t\to\infty$ the system approaches the pseudo-inverse matrix model.

Remark 7.2. The "temporal variable" t within an (equilibrium) statistical mechanical theory may look weird. However, it should be noticed that the time-scale for a sleeping session is much longer than that characterizing neuronal dynamics. This is also reasonable from a biological perspective, since neural dynamics takes place with frequencies of the order of $O(10^2)$ Hz (i.e. the typical spiking time, considering also the absolute refractory period of a biological neuron).

Once the Hamiltonian is introduced, we can define the basic thermodynamical quantities, namely

Definition 7.2. The partition function of the *reinforcement&removal* model (7.6) is

$$\begin{aligned} Z_N(\beta,t) &= \sum_{\boldsymbol{\sigma}} e^{-\beta H_{N,P}(\boldsymbol{\sigma}|\boldsymbol{\xi},t)} = \\ &= \sum_{\boldsymbol{\sigma}}\exp\Big\{\frac{\beta}{2N}\sum_{i,j=1}^{N}\sum_{\mu,\nu=1}^{P}\xi_i^\mu\xi_j^\nu\left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu}\sigma_i\sigma_j\Big\}. \end{aligned} \tag{7.7}$$

The associated infinite volume limit of the intensive free energy is defined as

$$f(\beta,\lambda,t) = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}\log Z_N(\boldsymbol{\sigma}|\boldsymbol{\xi},t). \tag{7.8}$$

Again, we shall denote with $\omega(\cdot)$ and $\langle\cdot\rangle$ the (t-dependent) Boltzmann-Gibbs averages (where the last one can be intended also on the replicated system).

[1] A note on the notation: the denominator $1/(\mathbb{I}+tC)$ is intended as the inverse matrix $(\mathbb{I}+tC)^{-1}$.

7.1 The replica trick

We now directly move to the replica trick resolution of the system under the assumption of replica symmetry. Using the standard approach, we write the large N free-energy as

$$f(\beta,\lambda,t) = -\lim_{N\to\infty}\frac{1}{\beta N}\mathbb{E}'\log Z_N(\boldsymbol{\sigma}|\boldsymbol{\xi},t) = -\lim_{\substack{n\to 0\\ N\to\infty}}\frac{\mathbb{E}'Z_N(\boldsymbol{\sigma}|\boldsymbol{\xi},t)^n-1}{\beta n N}. \quad (7.9)$$

As we did for the standard Hopfield model, we will assume that the only candidate for retrieval is the pattern ξ^1, while ξ^μ for $\mu\geq 2$ contribute to the slow noise. Therefore, also in this case $\mathbb{E}'$ is the average over the $P-1$ not-retrieved patterns. The replicated partition function can be put in Gaussian form as

$$\mathbb{E}'Z_N(\boldsymbol{\sigma}|\boldsymbol{\xi},t)^n = \mathbb{E}\mathcal{C}\sum_{\boldsymbol{\sigma}^{(1)}}\cdots\sum_{\boldsymbol{\sigma}^{(n)}}\int\Big(\prod_{\mu a}d\mu(z_\mu^{(a)})\Big)\Big(\prod_{ia}d\mu(\phi_i^{(a)})\Big)\cdot$$
$$\cdot\exp\Big(\sqrt{\frac{\beta}{N}(t+1)}\sum_{\mu i a}z_\mu^{(a)}\xi_i^\mu\sigma_i^{(a)} + i\sqrt{\frac{t}{N}}\sum_{\mu i a}z_\mu^{(a)}\xi_i^\mu\phi_i^{(a)}\Big), \quad (7.10)$$

where $\mathcal{P}(z_\mu^a)=\mathcal{P}(\phi_i^a)=\mathcal{N}(0,1)$ and $\mathcal{C}$ is a (ξ-dependent) normalization constant coming from the double Gaussian integration (its contribution to the free energy is trivial since it is constant, so we will omit it). The only difference of our model w.r.t. to the system studied in [47] is that here we have $\beta(1+t)$ (instead of β). As we will see, this factor is crucial to keep stable the thermodynamics of the system, since the critical temperature at zero load $\lambda=0$ is kept fixed at $\beta_c=1$ as t is tuned. Before to go further, we have to define the order parameters.

Definition 7.3. Besides the usual Mattis overlap

$$m_\mu = \frac{1}{N}\sum_i \xi_i^\mu\sigma_i, \quad (7.11)$$

we introduce the generalized overlaps

$$Q_{ab} = \frac{1}{N}\sum_i\Big(\sigma_i^{(a)} + i\sqrt{\frac{t}{\beta(1+t)}}\phi_i^{(a)}\Big)\Big(\sigma_i^{(b)} + i\sqrt{\frac{t}{\beta(1+t)}}\phi_i^{(b)}\Big). \quad (7.12)$$

Remark 7.3. Such a definition of the overlap could be weird, since it measure the overlaps complex linear combinations of different variables typologies. However, as we will see such a trick allows to strongly simplify the computations, and *a fortiori* we will see that its thermodynamic value is indeed real, thus making everything well-defined.

The replica trick computations for such a model follow the same procedure as the usual Hopfield model, but - because of their complexity - we will only give the intermediate result (however, the interested reader could consult our original work [50]). The expression for the replicated partition function is therefore

$$\mathbb{E}' Z_N(\boldsymbol{\sigma}|\boldsymbol{\xi}, t)^n = \int d\mu(\boldsymbol{m}, \boldsymbol{Q}, \boldsymbol{P}) \exp(-NA[\boldsymbol{m}, \boldsymbol{Q}, \boldsymbol{P}]), \tag{7.13}$$

where

$$\begin{aligned} &A[\boldsymbol{m}, \boldsymbol{Q}, \boldsymbol{P}] = \\ &= \frac{\beta}{2(1+t)} \sum_a (m_1^{(a)})^2 + \frac{\lambda\beta^2}{2} \sum_{ab} P_{ab} Q_{ab} + \frac{\lambda}{2} \log\det\left[\mathbb{I} - \beta(1+t)\boldsymbol{Q}\right] \\ &- \mathbb{E} \log \sum_{\boldsymbol{\sigma}} \int \Big(\prod_a d\mu(\phi^{(a)})\Big) \exp\Big[\beta \sum_a m_1^{(a)} \xi^1 \Big(\sigma^{(a)} + i\sqrt{\tfrac{t}{\beta(1+t)}}\phi^{(a)}\Big) \\ &+ \frac{\lambda\beta^2}{2} \sum_{ab} P_{ab} \Big(\sigma^{(a)} + i\sqrt{\tfrac{t}{\beta(1+t)}}\phi^{(a)}\Big)\Big(\sigma^{(b)} + i\sqrt{\tfrac{t}{\beta(1+t)}}\phi^{(b)}\Big)\Big]. \end{aligned} \tag{7.14}$$

Here, λ is the storage load, m_1^α is the Mattis overlap (of the α-th replica) associated to the pattern ξ^1 to be retrieved and $\boldsymbol{P}$ is the conjugated overlap matrix (entering because of the Fourier representation of $\boldsymbol{Q}$ Dirac deltas). The previous problem has the form of Laplace integral, so it can be evaluated with saddle point method. In this limit, of course, the intensive free energy of the model is finally realized as

$$f(\beta, \lambda, t) = \lim_{n\to 0} A[\boldsymbol{m}, \boldsymbol{Q}, \boldsymbol{P}]. \tag{7.15}$$

At this point, we cannot go further without imposing some structure of the overlap matrices. Then, as in the Hopfield case, we will adopt the RS *Ansatz*

$$m_1^{(a)} = m \quad \forall a, \tag{7.16a}$$

$$Q_{ab} = Q\delta_{ab} + q(1-\delta_{ab}), \tag{7.16b}$$

$$P_{ab} = P\delta_{ab} + p(1-\delta_{ab}). \tag{7.16c}$$

We stress that the diagonal overlap P should not be confused with the number of stored patterns (which is indeed implicit in the definition of the storage capacity λ).

Remark 7.4. Here, we would like to stress two points. First of all, since the generalized overlap involves combinations of different variables, it is not ensured that diagonal entries in $\boldsymbol{Q}$ will be equal to 1, so we have to introduce the diagonal generalized overlap Q. Furthermore, *a priori* it is not ensured that the diagonal *conjugate* one can be consistently set to zero. Ultimately, we have to deal with *five* different order parameters.

After straightforward computations, we can finally state the next

Proposition 1. *The infinite volume limit of the replica-symmetric free energy for the model (7.6), expressed in terms of the order parameters m and q, reads as*

$$\begin{aligned}
&f_{RS}(\beta,\lambda,t) = \\
&= \frac{m^2}{2(1+t)}\Big(1+\frac{t}{\Delta}\Big) + \frac{(1+t)(\Delta-1)}{2t}Q + \frac{\lambda\beta}{2}p(Q-q) \\
&+ \frac{\lambda}{2\beta}\Big(\log[1-\beta(1+t)(Q-q)] - \frac{q\beta(1+t)}{1-\beta(1+t)(Q-q)}\Big) + \frac{(1+t)(1-\Delta)}{2t\Delta} \\
&+ \frac{\log\Delta}{2\beta} + \frac{\lambda pt}{2(1+t)\Delta} - \frac{1}{\beta}\int_{-\infty}^{+\infty} d\mu(z)\log 2\cosh\Big[\frac{\beta}{\Delta}(m+\sqrt{\lambda p}z)\Big].
\end{aligned} \tag{7.17}$$

where $\Delta = 1+\lambda\beta t(1+t)^{-1}(P-p)$. The associated self-consistency equations are therefore

$$m = \frac{1+t}{\Delta+t}\int_{-\infty}^{+\infty} d\mu(z)\tanh\Big[\frac{\beta}{\Delta}(m+\sqrt{\lambda p}z)\Big], \tag{7.18a}$$

$$p = \frac{q(1+t)^2}{[1-\beta(1+t)(Q-q)]^2}, \tag{7.18b}$$

$$\Delta = 1+\frac{\lambda t}{1-\beta(1+t)(Q-q)}, \tag{7.18c}$$

$$q = Q + \frac{t}{\beta(1+t)\Delta} - \frac{1}{\Delta^2}\int_{-\infty}^{+\infty} d\mu(z)\cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\lambda p}z)\Big], \tag{7.18d}$$

$$Q\Delta^2 = 1 - \frac{t\Delta}{\beta(1+t)} + \frac{\lambda pt^2}{(1+t)^2} - \frac{m^2t(t+2\Delta)}{(1+t)^2} \tag{7.18e}$$

$$- \frac{2\lambda\beta pt}{(1+t)\Delta}\int_{-\infty}^{+\infty} d\mu(z)\cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\lambda p}z)\Big]. \tag{7.18f}$$

Remark 7.5. Notice that, in the limit $t \to 0$, both the free energy (7.17) and the self-consistency equations (7.18) reduces to the Amit-Gutfreund-Sompolinsky ones [13], as they should.

7.1.1 Remotion *or* Reinforcement: a separate analysis

Before turning on the solutions of the self-consistency equations and the realization of the phase diagram, in this Section we would like to justify why our model account both for reinforcement and remotion. In particular, in the generalized kernel appearing in 7.6, the denominator (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 memories. To this aim, we separate the whole Hamiltonian (7.6) in two different models by considering separately the numerator (reinforcement) and the denominator (remotion) in the generalized kernel:

$$H_N^{(1)} = -\frac{1}{2}\sum_{\mu}\sum_{ij}\xi_i^\mu\xi_j^\mu(1+t)\sigma_i\sigma_j, \tag{7.19a}$$

$$H_N^{(2)} = -\frac{1}{2}\sum_{\mu\nu}\sum_{ij}\xi_i^\mu\xi_j^\nu(\mathbb{I}+tC)^{-1}_{\mu,\nu}\sigma_i\sigma_j. \tag{7.19b}$$

Let us analyze these two models separately:

- For the former (which, in our claim, is the responsible of reinforcement effect), it is formally equivalent to the Hopfield model, but with a rescaled thermal noise $\tilde{\beta} = \beta(1 + t)$. As a consequence, the zero-capacity critical temperature is precisely $\tilde{T}_c = \tilde{\beta}_c^{-1} = 1$, which implies $T_c = (1 + t)$. See Figure 7.1 (left panel).

- The latter model is precisely (7.5), whose statistical mechanics has been carried out in [47] (in the standard replica-symmetric regime)Remarkably, the $\lambda = 0$ critical temperature for breaking the retrieval functionality of the model is $T_c = (1 + t)^{-1}$. See Figure 7.1 (right panel).

By comparing both the critical temperatures, it is reasonable to expect that, in the full model (7.6), the $\lambda = 0$ critical temperature could be fixed at $T_c = 1$, therefore saving the whole retrieval region. As we will show, this turns out to be true.

With these ideas in mind, it is also reasonable to expect that - in the full model (7.6) - the mashing effect of unlearning can be compensated by the rescaling of the thermal noise, therefore giving an optimal balance between the Reinforcement and the Removal features. The evaluation of the phase diagram for our model is presented in the next Section.

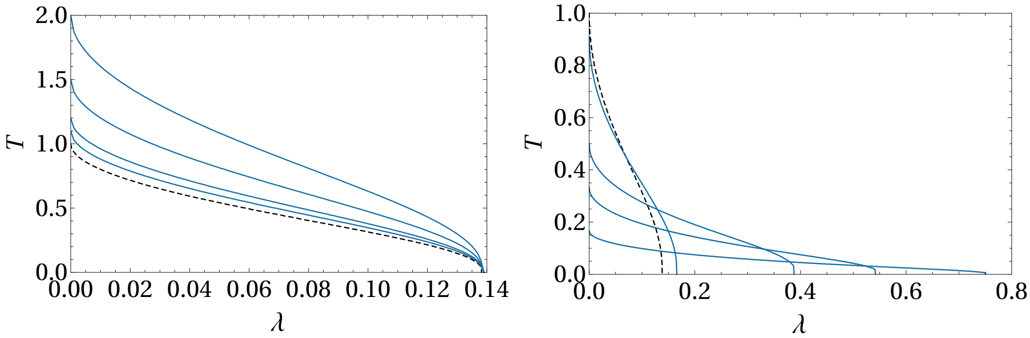

Figure 7.1: **Reinforcing and unlearning models.** Left: the plot shows the retrieval regions for the reinforcing model $H^{(1)}$ for $t = 0$ (Hopfield), 0.1, 0.2, 0.5 and 1. The critical temperature in the zero-capacity limit is $T_c = (1 + t)$ and this trivial shift in the critical temperature is the solely novelty of this model. Right: the plot shows the retrieval regions for the Dotsenko model as also discussed in [47]. The critical temperature grows with t, by the critical temperature in the zero-capacity limit decreases as $T_c = (1 + t)^{-1}$, so that the retrieval regions are mashed on the horizontal axes.

7.2 Guerra's interpolating scheme

In this Section, we will re-derive the above results from the point of view of Guerra's interpolation scheme, therefore giving a rigorous basis to our findings. In order to simplify our computations, we will express the Hamiltonian in a convenient form with the following

Definition 7.4. The Hamiltonian in the Gaussian representation (7.10) of (7.6) can be written in the form

$$H_{N,P}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{\xi}) = \frac{a}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} z_\mu \xi_i^\mu k_i, \tag{7.20}$$

where the *multi-spin* k_i stands for the complex linear combination $\sigma_i + b\phi_i$, with

$$a = \sqrt{\beta(t+1)}, \quad b = i\sqrt{\frac{t}{\beta(t+1)}}. \tag{7.21}$$

Remark 7.6. We stress again that, for the sake of mathematical convenience, as deepened when inspecting the hybrid Hopfield network, we take solely the pattern candidate for retrieval (i.e. the *signal*) to be Boolean, while all the remaining ones (acting as *slow noise* on the retrieval) are chosen as

Gaussian. Although neural networks, in general, do not exhibit the universality properties of spin glasses [54], this is no longer true if we confine our focus solely to the structure of the slow noise generated by patterns.[1]

Here, we are interested in the expression of the intensive pressure $\alpha(\beta, \lambda)$ in the high-storage regime $P = \lambda N$ in the thermodynamic limit. To do this, we have to introduce the following order parameters.

Definition 7.5. The natural order parameters for the neural network model (7.6) are the overlaps Q_{ab} and P_{ab} between the ks and the zs variables of the replicated system. In mathematical terms:

$$Q_{ab} = \frac{1}{N}\sum_{i=1}^{N} k_i^{(a)} k_i^{(b)}, \tag{7.22}$$

$$P_{ab} = \frac{1}{P}\sum_{\mu\geq 2} z_\mu^{(a)} z_\mu^{(b)}, \tag{7.23}$$

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} \xi_i^1 k_i. \tag{7.24}$$

Remark 7.7. We stress that, in the definition of the overlaps, we tacitly assume that only the first pattern ξ^1 is the candidate to be retrieved. Therefore, in the practical computations, we separate again between the signal and noise terms. This justify the definition of the overlap P in order to include only z_μ variables with $\mu \geq 2$.

Remark 7.8. In the replica symmetric (RS) regime, order parameters do not fluctuate in the thermodynamic limit[2], i.e.

$$q_{ab} \overset{RS}{\to} Q\delta_{ab} + q(1-\delta_{ab}), \tag{7.25}$$

$$p_{ab} \overset{RS}{\to} P\delta_{ab} + p(1-\delta_{ab}), \tag{7.26}$$

$$m_1 \overset{RS}{\to} m, \tag{7.27}$$

In order to set up the Guerra's interpolation framework, we need to introduce the generalized intensive pressure. This is the content of the following

[1] As extensively discussed in [23, 22], by varying the nature of the neurons as well as of the pattern entries, for instance ranging from Boolean (Ising) to standard Gaussians, the retrieval performances of the network vary sensibly and, in some limits, are entirely lost: in this sense neural networks do not share *universality* with standard spin-glasses.

[2] This request is of course consistent with the replica-symmetric ansatz when approaching the problem via the replica trick [37, 50].

Definition 7.6. Given the interpolating parameter $s \in [0,1]$, the auxiliary fields (which are i.i.d Gaussian variables) $\{\eta_i\}_{i\in(1,\dots,N)}$, $\{\lambda_\mu\}_{\mu\in(2,\dots,P)}$ and the tunable scalars C_1, C_2, C_3, C_4, C_5 (to be set *a posteriori*), the generalized intensive pressure is

$$
\begin{aligned}
\alpha_N(s) = \frac{1}{N}\,\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\int d\mu\,(z,\phi)\exp\Big[&\sqrt{s}\frac{a}{\sqrt{N}}\sum_{i,\mu\geq 2} z_\mu\xi_i^\mu k_i \\
&+\sqrt{s}\frac{a}{\sqrt{N}}\sum_i z_1\xi_i^1 k_i + \sqrt{1-s}\Big(C_1\sum_i^N \eta_i k_i + C_2\sum_{\mu\geq 2}\lambda_\mu z_\mu\Big) \\
&+\frac{1-s}{2}\Big(C_3\sum_{\mu\geq 2} z_\mu^2 + C_4\sum_i k_i^2 + C_5 a\sum_i \xi_i^1 k_i\Big)\Big].
\end{aligned}
\tag{7.28}
$$

The quenched average $\mathbb{E}$ is performed over non-recalled patterns (contributing to the noise) and the auxiliary fields.

Remark 7.9. As usual, the $s = 1$ choice recovers the original model, namely $\alpha(\beta,\lambda,t) = \lim_{N\to\infty}\alpha_N(s=1)$, while for $s \to 0$ it is a more tractable a one-body problem.

Also in this case, we denote with $\omega_s(\cdot)$ and $\Omega_s(\cdot)$ the Boltzmann-Gibbs averages (respectively of the system and of its replicated version) implicitly defined by the above intensive pressure, and $\langle\cdot\rangle = \mathbb{E}\Omega_s(\cdot)$.

Proposition 7.1. *The infinite volume limit of the quenched pressure related to the model (7.6) can be obtained by using the Fundamental Theorem of Calculus as*

$$
\alpha(\beta,\lambda,t) \equiv \lim_{N\to\infty}\alpha_N(s=1) = \lim_{N\to\infty}\Big(\alpha_N(s=0) + \int_0^1 ds\partial_s\alpha_N(s)\Big). \tag{7.29}
$$

Let us start with the computation of the derivative (by:

$$
\begin{aligned}
\frac{d\alpha_N(s)}{ds} =& \frac{1}{2N}\,\mathbb{E}\Big[\frac{a}{\sqrt{sN}}\sum_{i,\mu\geq 2}\xi_i^\mu\omega_s(z_\mu k_i) - \frac{1}{\sqrt{1-s}}\Big(C_1\sum_i \eta_i\omega_s(k_i) \\
&+ C_2\sum_{\mu\geq 2}\lambda_\mu\omega_s(z_\mu)\Big) + \frac{a}{\sqrt{sN}}\sum_i \xi_i^1\omega_s(z_1 k_i) - C_3\sum_{\mu\geq 2}\Omega_s(z_\mu^2) \\
&- C_4\sum_i \omega_s(k_i^2) - C_5 a\sum_i \omega_s(\xi_i^1 k_i)\Big].
\end{aligned}
\tag{7.30}
$$

We can proceed further by using Wick's Theorem on z^1 and the auxiliary fields, we have

$$\begin{aligned}\frac{d\alpha_N(s)}{ds} =&\frac{1}{2N}\,\mathbb{E}\Big[\frac{a^2}{N}\sum_{i,\mu\geq 2}\Big(\omega_s(z_\mu^2k_i^2)-\omega_s(z_\mu k_i)^2\Big)+\frac{a^2}{N}\omega_s((\sum_i \xi_i^1 k_i)^2)\\ &-C_1^2\sum_i\Big(\omega_s(k_i^2)-\omega_s(k_i)^2\Big)-C_2^2\sum_{\mu\geq 2}\Big(\omega_s(z_\mu^2)-\omega_s(z_\mu)^2\Big)\\ &-C_3\sum_{\mu\geq 2}\omega_s(z_\mu^2)-C_4\sum_i\omega_s(k_i^2)-C_5 a\sum_i\omega_s(\xi_i^1k_i)\Big].\end{aligned} \tag{7.31}$$

We can now directly introduce the order parameters (7.24) (after considering the replicated system) in order to get

$$\begin{aligned}\frac{d\alpha_N}{ds} =&\frac{1}{2}\,\mathbb{E}\Big[a^2\lambda\Omega_s(Q_{11}P_{11})+a^2\Omega_s(m_1^2)-a^2\lambda\Omega_s(Q_{12}P_{12})-C_1^2\Omega_s(Q_{11})\\ &+C_1^2\Omega_s(Q_{12})-C_2^2\lambda\Omega_s(P_{11})+C_2^2\lambda\Omega_s(P_{12})-\lambda C_3\Omega_s(P_{11})\\ &-C_4\Omega_s(Q_{11})-aC_5\Omega_s(m_1)\Big].\end{aligned} \tag{7.32}$$

Now, fixing the scalars $C_{1,..,5}$ as

$$\begin{aligned}&C_1^2=a^2\lambda p,\quad C_2^2=a^2q,\quad C_3=a^2(Q-q),\\ &C_4=a^2\lambda(P-p),\quad C_5=2ma,\end{aligned} \tag{7.33}$$

we can recast the streaming $\partial_s\alpha_N(s)$ in the form

$$\begin{aligned}\frac{d\alpha_N}{ds}=&\frac{1}{2}\,\mathbb{E}\Big[a^2\lambda\Omega_s((q_{11}-Q)(p_{11}-P))+a^2\Omega_s((m_1-m)^2)\\ &-a^2\lambda\Omega_s((q_{12}-q)(p_{12}-p))\Big]+\frac{\lambda a^2}{2}(qp-QP)-\frac{a^2}{2}m^2.\end{aligned} \tag{7.34}$$

Remark 7.10. Requiring replica symmetry, the evaluation of the s-integral in Eq. (7.29) is trivial, since the r.h.s. of Eq. (7.34) reduces to

$$\partial_s\alpha_N(s)=\frac{\lambda a^2}{2}(qp-QP)-\frac{a^2}{2}m^2 \tag{7.35}$$

that does not depend on s any longer.

For the one-body contribution, we have

$$\begin{aligned}\alpha_N(s=0)=\frac{1}{N}\,\mathbb{E}\log\sum_{\boldsymbol{\sigma}}\int d\mu\,(z,\phi)\exp\Big[C_1\sum_i\eta_ik_i+\frac{C_4}{2}\sum_i k_i^2\\ +\frac{C_5a}{2}\sum_i\xi_i^1k_i+C_2\sum_{\mu\geq 2}\lambda_\mu z_\mu+\frac{C_3}{2}\sum_{\mu\geq 2}z_\mu^2\Big].\end{aligned} \tag{7.36}$$

The computation of the one-body contribution is straightforward (but somewhat cumbersome), so we only give the final result:

$$\begin{aligned}\alpha_N(s=0) &= -\frac{\lambda}{2}\log(1-C_3) - \frac{1}{2}\log(1-C_4b^2) + \frac{\lambda}{2}\frac{C_2^2}{1-C_3} + \frac{C_4}{2} \\ &\quad + b^2\frac{C_1^2+C_4^2+\frac{C_5^2a^2}{4}}{1-C_4b^2} + \mathbb{E}\log 2\cosh\left[\frac{C_1\eta+\frac{C_5a}{2}}{1-C_4b^2}\right].\end{aligned} \tag{7.37}$$

Then, putting everything together, recalling the choice for the parameters $C_1, ..., C_5$ as prescribed in the relations 7.33and performing the trivial rescaling of the overlaps as

$$P \to \frac{\beta^2}{a^2}P, \quad p \to \frac{\beta^2}{a^2}p, \quad m \to \frac{\beta}{a^2}m, \tag{7.38}$$

after some trivial manipulation we arrive at the following

Theorem 7.1. *The thermodynamic limit of the intensive pressure in the replica symmetric regime of neural network model defined in Eq. (7.6) is*

$$\begin{aligned}\alpha_{RS}(\beta,\lambda,t) &= -\frac{\beta m^2}{2(1+t)}\Big(1+\frac{t}{\Delta}\Big) - \frac{(1+t)(\Delta-1)}{2t}\beta Q - \frac{\lambda\beta^2}{2}p(Q-q) \\ &\quad - \frac{\lambda}{2}\Big(\log[1-\beta(1+t)(Q-q)] + \frac{q\beta(1+t)}{1-\beta(1+t)(Q-q)}\Big) \\ &\quad - \frac{(1+t)(1-\Delta)\beta}{2t\Delta} - \frac{\log\Delta}{2} - \frac{\lambda\beta pt}{2(1+t)\Delta} \\ &\quad + \int_{-\infty}^{+\infty} d\mu(\eta)\log 2\cosh\Big[\frac{\beta}{\Delta}(m+\sqrt{\lambda p}\eta)\Big],\end{aligned} \tag{7.39}$$

where again $\Delta = 1 + \lambda\beta t(t+1)^{-1}(P-p)$.

Remark 7.11. By a direct comparison, we see that this expression of the intensive pressure $\alpha(\beta,\lambda,t) = -\beta f(\beta,\lambda,t)$ leads to the same intensive free energy (7.17), therefore leading to the same self-consistency equations, commuting Proposition 1 into Theorem 7.1.

7.2.1 Analysis of the overlap fluctuations and ergodicity breaking

Before turning on the numerical resolution of the self-consistency equations (7.18), we would like to conclude the theoretical analysis of the modelin order to determine the critical behaviour of the system and the ergodicity breaking. To address this point, we study the behaviour of the overlap fluctuations, which we suitably center around their thermodynamic values and properly rescale (in order to allow them to diverge when the system approaches the critical line). This is possible since they are meromorphic functions, and their poles identify the evolution of the critical surface $\beta_c(\lambda, t)$ (if any).

To this aim, we will use the generalized Guerra's interpolation scheme (see Eq. (7.28)) and using a sum rule (perfectly analogous to the one for the intensive pressure in the thermodynamic limit, see Eq. (7.29)). In this way, we are able to evaluate the evolution of the order parameter correlators from $s = 0$ (where their evaluation is simple) and propagate it up to $s = 1$. Therefore, for the correlation function of a generic thermodynamical observable O, we need to evaluate the Cauchy condition $\langle O(s = 0)\rangle$ and the derivative $\partial_s\langle O(s)\rangle$. In contrast with the case of the intensive free energy, where we imposed replica symmetry, here we impose ergodic behaviour (since we want to trace the ergodicity breaking critical line). In other words, we assume that the system is approaching this boundary from the high fast-noise limit (where the expectation values of the overlaps can be consistently set to zero in order to simplify the computations). The first step is therefore to introduce the centered and rescaled overlaps, as stated in the next

Definition 7.7. The centered and rescaled overlap fluctuations θ_{lm} and ρ_{lm} are introduced as

$$\begin{aligned}\theta_{lm} &= \sqrt{N}\big[Q_{lm} - \delta_{lm}Q - (1-\delta_{lm})q\big] &(7.40)\\ \rho_{lm} &= \sqrt{P}\big[P_{lm} - \delta_{lm}P - (1-\delta_{lm})p\big]. &(7.41)\end{aligned}$$

Remark 7.12. Of course, in this analysis the signal is absent, thus there is no need to introduce a rescaled Mattis order parameter. Here, we only consider the boundary between the ergodic region and the spin-glass phase.

In the next definition, we will introduce a generalized r-replicated pressure. In order to make notation more compact, we will denote the r-replicated system variables simply with the subscript R.

Definition 7.8. Given an observable O (which is a smooth function of neurons of the r-replicated system) and a source fields J, the $r-$replicated interpolating pressure $\mathcal{A}^r_J(s)$ is

$$\begin{aligned}\mathcal{A}^r_J(s) = \mathbb{E}\log\sum_{\sigma_R}\int d\mu\,(z_R,\phi_R)\exp\Big[&\sqrt{s}\frac{a}{\sqrt{N}}\sum_{l=1}^{r}\sum_{i,\mu} z_\mu^{(l)}\xi_i^\mu k_i^{(l)} + J\hat{O}\\ &+\sqrt{1-s}\Big(C_1\sum_{l=1}^{r}\sum_{i}\eta_i k_i^{(l)} + C_2\sum_{l=1}^{r}\sum_{\mu}\lambda_\mu z_\mu^{(l)}\Big)\\ &+\frac{1-s}{2}\Big(C_3\sum_{l=1}^{r}\sum_{\mu}(z_\mu^{(l)})^2 + C_4\sum_{l=1}^{r}\sum_{i}(k_i^{(l)})^2\Big)\Big].\end{aligned} \tag{7.42}$$

where $C_{1,2,3,4}$ are the same given in the previous section (see Eq. (7.33)).

Remark 7.13. Of course, here there is no C_5, since the signal term is absent.

By construction, the derivative of the r-replicated pressure with respect to the external source fields J are

$$\langle O(s)\rangle_s = \frac{\partial\mathcal{A}^r_J(s)}{\partial J}\Big|_{J=0}, \qquad \partial_s\langle O(s)\rangle_s = \frac{\partial(\partial_s\mathcal{A}^r_J)}{\partial J}\Big|_{J=0}. \tag{7.43}$$

In order to evaluate the fluctuations of O, we need to evaluate first $\partial_s\mathcal{A}^r_J$. By standard computations, we get

$$\partial_s\mathcal{A}^r_J = \frac{1}{2}\sqrt{\lambda}\beta(1+t)\sum_{l,m=1}^{r}\Big[\langle g_{l,m}\rangle_s - \langle g_{l,m+r}\rangle_s\Big], \qquad g_{l,m} = \theta_{l,m}\rho_{l,m}. \tag{7.44}$$

Then, using (7.43) and performing the same rescaling we did in the previous section, namely

$$(P,p) \to \frac{\beta^2}{a^2}(P,p), \tag{7.45}$$

it can be proved the following

Proposition 7.2. *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:*

$$\begin{aligned}d_\tau\langle O\rangle_s = &\frac{1}{2}\sum_{a,b}^{r}\langle O\cdot g_{a,b}\rangle_s - r\sum_{a=1}^{r}\langle O\cdot g_{a,r+1}\rangle_s\\ &+\frac{r(r+1)}{2}\langle O\cdot g_{r+1,r+2}\rangle_s - \frac{r}{2}\langle O\cdot g_{r+1,r+1}\rangle_s,\end{aligned} \tag{7.46}$$

where d_τ is the derivative

$$d_\tau = \frac{1}{\beta(1+t)\sqrt{\alpha}}\frac{d}{ds}. \tag{7.47}$$

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):

$$\begin{array}{lll}\langle\theta_{12}^2\rangle_s = A(s), & \langle\theta_{12}\theta_{13}\rangle_s = B(s), & \langle\theta_{12}\theta_{34}\rangle_s = C(s),\\ \langle\theta_{12}\rho_{12}\rangle_s = D(s), & \langle\theta_{12}\rho_{13}\rangle_s = E(s), & \langle\theta_{12}\rho_{34}\rangle_s = F(s),\\ \langle\rho_{12}^2\rangle_s = G(s), & \langle\rho_{12}\rho_{13}\rangle_s = H(s), & \langle\rho_{12}\rho_{34}\rangle_s = I(s),\\ \langle\theta_{11}^2\rangle_s = J(s), & \langle\theta_{11}\rho_{11}\rangle_s = K(s), & \langle\rho_{11}^2\rangle_s = L(s),\\ \langle\theta_{11}\theta_{12}\rangle_s = M(s), & \langle\theta_{11}\rho_{12}\rangle_s = N(s), & \langle\rho_{11}\theta_{12}\rangle_s = O(s),\\ \langle\rho_{11}\rho_{12}\rangle_s = P(s), & \langle\theta_{11}\rho_{22}\rangle_s = Q(s), & \langle\theta_{11}\theta_{22}\rangle_s = R(s).\\ \langle\rho_{11}\rho_{22}\rangle_s = S(s), & & \end{array} \tag{7.48}$$

Since we intend to approach the critical line for ergodicity breaking *from above* [26]), we can treat $\theta_{a,b}, \rho_{a,b}$ as Gaussian variables with zero mean (so that we can apply Wick-Isserlis theorem in the averages). Analogously, we can also treat both the k_i and z_μ as zero mean random variables (i.e. all averages of uncoupled fields vanish). It can be easily shown that this procedure considerably simplifies the evaluation of the critical lineThus, we have only to deal with the quantities

$$\begin{array}{lll}\langle\theta_{12}^2\rangle_s = A(s), & \langle\theta_{12}\rho_{12}\rangle_s = D(s), & \langle\rho_{12}^2\rangle_s = G(s),\\ \langle\theta_{11}^2\rangle_s = J(s), & \langle\theta_{11}\rho_{11}\rangle_s = K(s), & \langle\rho_{11}^2\rangle_s = L(s),\\ \langle\theta_{11}\rho_{22}\rangle_s = Q(s), & \langle\theta_{11}\theta_{22}\rangle_s = R(s), & \langle\rho_{11}\rho_{22}\rangle_s = S(s).\end{array} \tag{7.49}$$

By using Eq. (7.46) to the above quantities, we get the differential equations

$$d_\tau A = 2AD, \tag{7.50}$$

$$d_\tau D = D^2 + AG, \tag{7.51}$$

$$d_\tau G = 2GD. \tag{7.52}$$

Moreover, we can reduce the number of differential equations by suitably combining A and G, since we can easily see that

$$d_\tau \log\frac{A}{G} = 0 \implies A(\tau) = r^2 G(\tau), \quad r^2 = \frac{A(0)}{G(0)}. \tag{7.53}$$

Then, the problem is reduced to the two coupled differential equations

$$d_\tau D = D^2 + r^2 G^2, \tag{7.54}$$

$$d_\tau G = 2GD. \tag{7.55}$$

Finally, introducing the quantity $Y(\tau) = D(\tau) + rG(\tau)$, we end with

$$d_\tau Y \;=\; Y^2, \tag{7.56}$$

whose solution is trivially

$$Y(\tau) = \frac{Y_0}{1-\tau Y_0}, \quad Y_0 = D(0) + \sqrt{A(0)G(0)}. \tag{7.57}$$

The remaining part of the problem is therefore to evaluate the correlations at $s = 0$, namely the Cauchy conditions in Eq. (7.57). To do this, we introduce a one-body generating function for the momenta of z, k obtained by setting $s = 0$ and $r = 1$ in Eq. (7.42), and including source fields (j_i, J_μ) which couples respectively to (k_i, z_μ). Again, since we are approaching the critical line from the ergodic region, we can consistently set $m, p, q = 0$ in the coefficients (7.33). The result is the generating function

$$\begin{aligned} F(j,J) = \log\sum_\sigma \int d\mu\,(z,\phi)\exp\Big[\sum_i j_i k_i + \sum_\mu J_\mu z_\mu + \frac{a^2 Q}{2}\sum_\mu z_\mu^2 \\ + \frac{1-\Delta}{2b^2}\sum_i k_i^2\Big]. \end{aligned} \tag{7.58}$$

This quantity is very easy to handle with. Moreover, showing only the relevant terms in j and J, we have

$$F(j,J) = \frac{b^2\Delta + 1}{2\Delta^2}\sum_i j_i^2 + \frac{1}{2(1-a^2Q)}\sum_\mu J_\mu^2 + O(j^3). \tag{7.59}$$

Then, by simply using the definitions (7.49), we can evaluate all the observable at $s = 0$ simply as derivatives of $F(j, J)$. Therefore, we arrive at the results

$$\begin{aligned} D(0) &= \sqrt{NP}\,\big(\partial_j F\big)^2\big(\partial_J F\big)^2\Big|_{j,J=0} = 0, \\ A(0) &= \big(\partial_j^2 F\big)^2\Big|_{j,J=0} = \Big[\frac{\beta(1+t) - t\Delta}{\beta(1+t)\Delta^2}\Big]^2 = Q^2, \\ G(0) &= \big(\partial_J^2 F\big)^2\Big|_{j,J=0} = (1-\beta(1+t)Q)^{-2}. \end{aligned} \tag{7.60}$$

Putting these findings in (7.57), we get

$$Y(\tau) = \frac{Q}{1-\beta(1+t)Q - \tau Q}. \tag{7.61}$$

Finally, evaluating $Y(\tau)$ for $\tau = \beta(1+t)\sqrt{\lambda}s$, $s = 1$, we have the result

$$Y(s=1) = \frac{Q}{1-\beta(1+t)Q(1+\sqrt{\lambda})}, \tag{7.62}$$

where

$$\begin{aligned} Q\Delta^2 &= 1 - \frac{t\Delta}{\beta(1+t)}, \\ \Delta &= 1 + \frac{\lambda t}{1-\beta(1+t)Q}. \end{aligned} \tag{7.63}$$

are the relevant self-consistency equations in the ergodic region. Since we are interested in determining the critical temperature for ergodicity breaking, which is characterized by the fact that fluctuations (in this case Y) grow arbitrarily large, we can find the conditions for which the denominator in Eq. (7.62) is zero. When doing this, we reach the result which is resumed in the following

Theorem 7.2. *The ergodic region of the model defined by the cost function (7.6) is delimited (i.e. $\beta < \beta_c$) by the following critical surface in the (β, λ, t) space of the model parameters:*

$$\beta_c = \frac{1}{1+t}\Big[\frac{\Delta^2}{1+\sqrt{\lambda}} + t\Delta\Big] \quad \textit{with} \quad \Delta = 1 + \sqrt{\lambda}(1+\sqrt{\lambda})t. \tag{7.64}$$

Remark 7.14. At vanishing sleep extent $t = 0$, where the model recovers the original Hopfield's scenario, the critical surface correctly collapses over the Amit-Gutfreund-Sompolinsky critical line $\beta_c = (1+\sqrt{\lambda})^{-1}$. In the opposite limit $t \to \infty$, the ergodic region collapses on the axis $T = 0$. This have a profound implications, since both the ergodic region (together with the retrieval one, as we will see in a moment) - as sleeping time flows - *phagocytes* the spin-glass phase.[1] This means that spurious states are entirely suppressed with a proper rest, allowing the network to achieve a perfect retrieval, as suggested in the pioneering study by Kanter and Sompolinsky [75].

[1] We stress that the ergodic line does not affect the retrieval region, they simply *fade* one into the other. This is due to the fact 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. Therefore, the retrieval breaking critical line (which is of first order) is not detected by a second order inspection as that needed for criticality.

7.3 Analysis of the replica symmetric solution

Once derived the self-consistency equations (7.18) for the model (7.6) and analysed the critical behaviour of the system (by inspecting its ergodicity breaking character), we now turn on the complete solutions of the model, in order to finally depict the phase diagram as the sleeping time flows.

7.3.1 Zero-temperature (noise-less) critical capacity

A preliminary, but interesting, analysis to do is the determination of the critical capacity at $T = 0$ as the sleeping time t is tuned. This is central for checking the retrieval capacities of the model (7.6). To this aim, it is convenient to introduce the parameter $c \equiv \beta(Q - q)$, which satisfies the equation

$$c = \frac{\beta}{\Delta^2}\int_{-\infty}^{+\infty} d\mu(z)\cosh^{-2}\left[\frac{\beta}{\Delta}(m + \sqrt{\lambda p}z)\right] - \frac{t}{(1+t)\Delta}. \tag{7.65}$$

By using the self-consistency equations, it is easy to check that c is finite and, consequently, $q \to Q$ as $T \to 0$. Since the hyperbolic tangent in (7.18a) tends to the error function, after some rearrangement, we can express the $T = 0$ limit of self-consistency equations as

$$\begin{aligned}
m &= \frac{1+t}{\Delta+t}\mathrm{erf}\left(\frac{m}{\sqrt{2\lambda p}}\right),\\
p &= \frac{Q(1+t)^2}{[1-(1+t)c]^2},\\
\Delta &= 1 + \frac{\lambda t}{1-(1+t)c},\\
c &= \frac{1}{\Delta}\sqrt{\frac{2}{\pi\lambda p}}\exp\left(-\frac{m^2}{2\lambda p}\right) - \frac{t}{\Delta(1+t)},\\
Q\Delta^2 &= 1 + \frac{\lambda p t^2}{(1+t)^2} - \frac{m^2 t(t+2\Delta)}{(1+t)^2} - \frac{2\lambda t}{1+t}\sqrt{\frac{2}{\pi\lambda p}}\exp\left(-\frac{m^2}{2\lambda p}\right).
\end{aligned}$$

In this limit, it is possible to eliminate the parameter Q from the self-consistency equations, therefore proving the following

Proposition 7.3. *The $T = 0$ limit of the self-consistency equations can be*

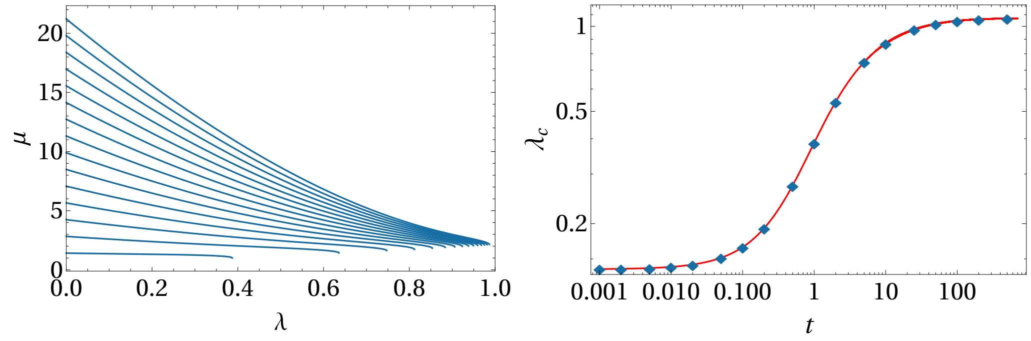

Figure 7.2: **Zero-temperature analysis of the critical capacity.** Left panel: numerical solutions for μ of the self-consistency equations in the zero temperature limit (7.67) for several unlearning times: $t = 1, 3, \dots, 29$. Right panel: temporal dependence of the critical capacity at zero temperature. The blue dots represent the storage capacity above which the only possible solution has $\mu = 0$. The red curve is the fit given by $y = x/(x + a)$, with $a = 2.84 \pm 0.01$ obtained by first normalizing data in $[0, 1]$, namely $\lambda_c \to [\lambda_c - \min(\lambda_c)]/[\max(\lambda_c) - \min(\lambda_c)]$.

resumed as

$$\mu = \frac{\Pi}{\sqrt{2}}\frac{1+t}{\Delta + t}\operatorname{erf}\left(\frac{\mu}{\sqrt{\lambda}}\right), \tag{7.67a}$$

$$\Delta = 1 + \frac{\lambda t}{1-(1+t)c}, \tag{7.67b}$$

$$c = \frac{\Pi}{\Delta}\sqrt{\frac{2}{\pi\lambda}}\exp\left(-\frac{\mu^2}{\lambda}\right) - \frac{t}{\Delta(t+1)}, \tag{7.67c}$$

$$\Delta^2[1-(1+t)c]^2 = \Pi^2(1+t)^2 + \lambda t^2 - 2\mu^2 t(t+2\Delta) \tag{7.67d}$$

$$- 2\lambda t(1+t)\Pi\sqrt{\frac{2}{\pi\lambda}}\exp\left(-\frac{\mu^2}{\lambda}\right), \tag{7.67e}$$

where $\mu = m(2p)^{-1/2}$, $\Pi = p^{-1/2}$.

Despite their intricate character, these self-consistency equations can be easily solved numerically. What we are interested in is the function $\lambda_c(t)$ above which the only possible solutions have $\mu = 0$ (which means $m = 0$, meaning that the system is no longer working in the retrieval mode). In the left plot of Fig. 7.2, we report such solutions for various sleep extent t. The end points of each curve separate the λ axis in the regions with respectively $\mu \neq 0$ and $\mu = 0$, therefore identifying the critical capacity for each fixed t

value. Then, in the right plot of Fig. 7.2, we report the critical capacity λ_c as a function of the sleep extent t.

To conclude this Section, we notice that the $t \to 0$ limit precisely recovers Hopfield model critical capacity $\lambda_c(t = 0) \sim 0.138$. In the opposite limit $t \to \infty$, we reach the upper bound $\lambda_c \sim 1.07$ (in agreement with [47]). Interestingly, the critical capacity displays a log-sigmoidal growth in t, suggesting that the time scale for unlearning t is intrinsically logarithmic. Also for relatively small values of t we can reach a critical threshold λ_c close to 1 (for instance, $\lambda_c(t = 1) \approx 0.4$ and $\lambda_c(t = 5) \approx 0.8$). Further increasing t, the improvement turns out to be slower, meaning that we should wait more time to get appreciable results.

7.3.2 Replica symmetric phase diagram

Once we checked that the critical capacity (at $T = 0$) grows as the sleeping time flows (therefore proving that dreaming effectively improves the retrieval performances of Hopfield model) we are finally interested in solve the self-consistency equations of model (7.6) in order to depict the phase diagram in the parameter space (β, λ, t). In particular, we are also interested in the asymptotic limit $t \to \infty$, where the model effectively should approach (by previous consideration) a stable behaviour. An example of solution for the order parameters and the free energy as functions of the thermal noise $T = \beta^{-1}$ at various storage capacity λ $(= 0, 0.05, 0.2, 0.5)$ and for large t (= 1000, which is far from the fast increase in the storage capacity, meaning that all of the sleeping effects are indeed present) is depicted in Fig. 7.3.

We solved the self-consistency equations (7.18) for various values of the sleeping time t, then we performed the following separate analysis.

- *Spin glass versus mixed retrieval regions.* In this part of the analysis, we look for the transition between the retrieval phase and the spin glass region, in order to determine the critical curve $T_c(\lambda)$ beyond which the solution has $m = 0$. The situation we find is formally similar to the original Hopfield model: in the low storage regime, the replica symmetric free energy is continuous everywhere and differentiable *almost* everywhere (except for the critical point $T_c = 1$, where a second-order phase transition takes place). For higher values of the capacity $\lambda > 0$, the phase transition is of the first kind taking place at the critical temperature $T_c(\lambda)$. The upper left plot in Fig. 7.3 shows an example of this Mattis overlap behaviour. Then, collecting the points (λ, T) for various values of t where this phase transition takes place, we are able

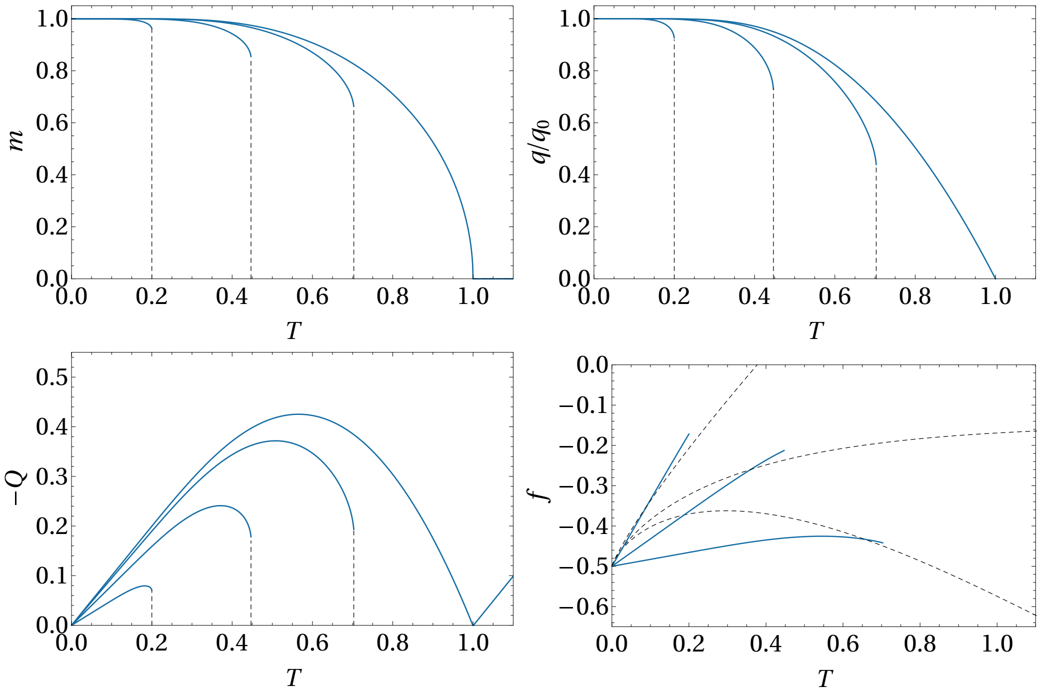

Figure 7.3: **Retrieval state solution for the order parameters and free energy at $t = 1000$.** First row: on the left, the plot shows the Mattis magnetization m as a function of the temperature for various storage capacity values ($\lambda = 0$, 0.05, 0.2 and 0.5, going from the right to the left). The vertical dotted lines indicates the jump discontinuity identifying the critical temperature $T_c(\alpha)$ separating 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 for pure states (i.e. $T < T_c(\alpha)$). Second row: on the left, the plot shows the solution for the diagonal overlap $-Q$ in the retrieval region for $\lambda = 0$, 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 ($\lambda = 0.05$, 0.2 and 0.5, going from the bottom to the top) for both the retrieval (blue solid lines) and spin-glass (i.e. $m = 0$, black dashed lines) states.

to determine how retrieval region evolves in function of the sleep extent t. Such results have been collected in Figure 7.4 for $t = 0$ (the Hopfield scenario, denoted by the black dashed curve) $0.1, 1, 1000$ (blue lines, respectively from the left to the right). In agreement with our previous results for $T = 0$, it clearly emerges that the critical storage capacity

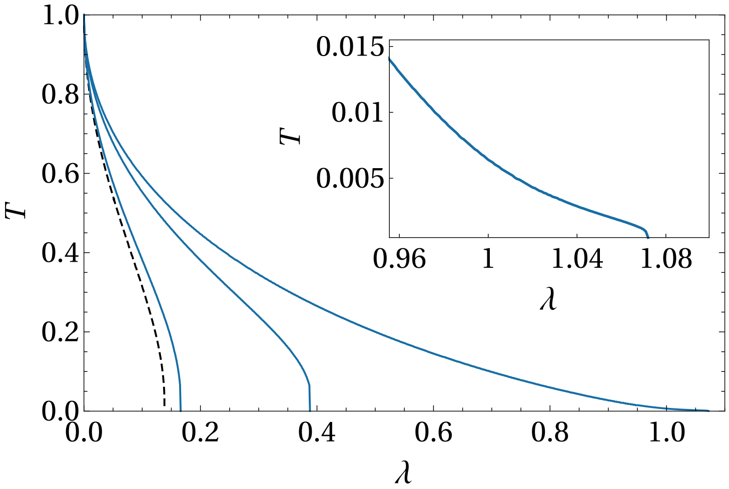

Figure 7.4: **Critical line for the transition between retrieval and spin-glass phases for various values of the unlearning time.** From the left to the right: $t = 0$ (Hopfield, black dashed line), 0.1, 1 and 1000. The inner plot on the top-right corner shows the tail of the critical curve for $t = 1000$.

λ effectively increases with the sleeping session, with the zero-capacity critical temperature $T_c(\lambda = 0)$ being stable to 1.[1] Thus, the whole retrieval region gets enlarged as the sleeping time t flows.

- *Mixed versus pure retrieval regions.* In this region, the pure states are global minima for the free energy. In order to identify the associated boundary, we solve the self-consistency equations (with fixed λ and T) for both retrieval ($m \neq 0$) and spin-glass ($m = 0$) solutions and compare the their free-energies. The lower right plot in Fig. 7.3 shows the behaviour of free energy at $t = 1000$ for both these solutions for various storage capacity $\lambda = 0.05, 0.2, 0.5$. Here, the solid blue lines correspond to retrieval solutions, while the black dashed ones are the spin glass solution counterparts. The intersection point between the corresponding curves identifies the critical temperature $T_R(\alpha)$ below which the pure states (globally) minimize the free-energy. The resulting boundary (together with the retrieval breaking critical line for $t = 1000$) is depicted in Fig. Thus, also the *pure* retrieval regions gets considerably enlarged while the network is sleeping. 7.5.

Once that the retrieval region (and the retrieval breaking transition line) is completely determined by solving the self-consistency equations, we can

[1] Actually, as we already remarked the asymptotic value of the critical capacity is $\lambda_c \sim 1.07$, in agreement with [47]. This is due to the replica symmetry assumption.

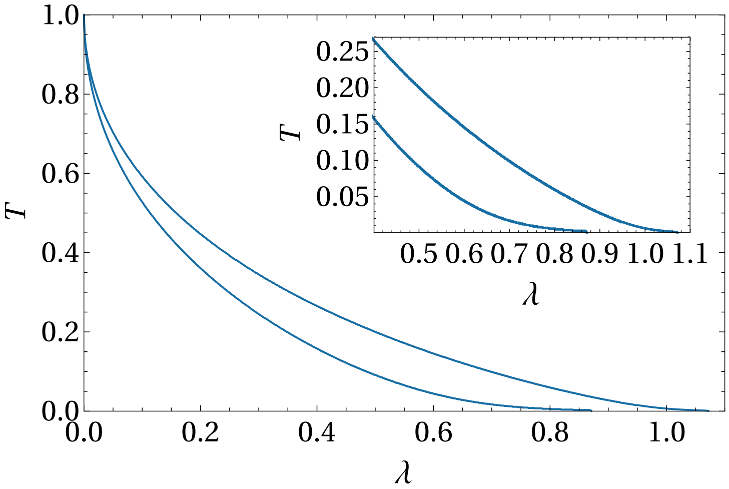

Figure 7.5: **Phase diagram in the large unlearning time limit (t = 1000).** The two curves trace the boundary of the maximal retrieval regions where patterns are global free energy minima (inner boundary) or local free energy minima (outer boundary). The inner plot on the top-right corner shows the tails of both the critical curves. We stress that, as already pointed out in [46], the extension of the retrieval region in the low-temperature regime up to $\lambda_c \sim 1.07$ is just a chimera of the replica symmetric approximation, while in the true RSB phase $\lambda_c \to 1$, according to Gardner's theory [52].

join all of our results (i.e. with ergodicity breaking critical line) and specify how the phase diagram evolves as the sleep extent t flows. The results are depicted in Fig. (7.6). The phase diagram is depicted for different choices of t: from left to right, $t = 0, 0.1, 1, 1000$. The remarkable aspect of the model is that, as t grows, the retrieval region (blue) and the ergodic region (yellow) get wider and wider, at the cost of the spin-glass region (red). For sufficiently long sleep extent, the latter progressively shrinks and collapse as $t \to \infty$. We also stress that the ergodicity breaking critical line changes its concavity.

7.4 Numerical results

We would like to conclude this Chapter with a numerical analysis concerning some aspects of the model (7.6). In particular, we will check that our replica symmetric ansatz is reasonable, by comparing the theoretical predictions with Monte Carlo (MC) simulations (where no assumptions are made). Then, we want to analyze the field distributions h_i and the robustness of the attraction basins of the pure minima.

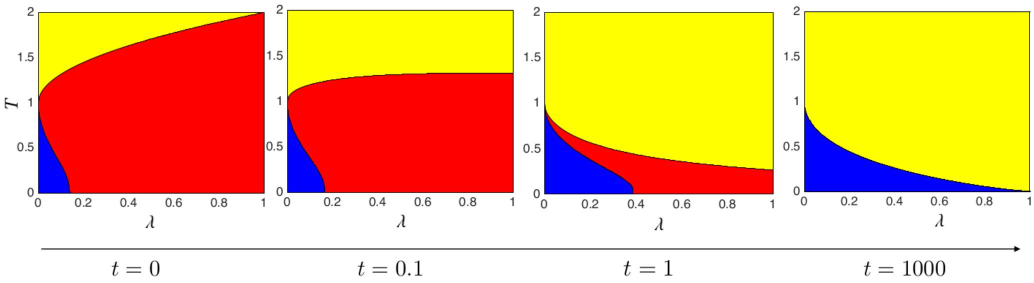

Figure 7.6: **Phase diagram evolution with the sleeping time.** The phase diagram of the model (7.6) is depicted for different values of t (from left to right, $t = 0, 0.1, 1, 1000$). As the sleeping extent grows, the retrieval region (blue) and the ergodic region (yellow) get wider invading the spin-glass region (red), which progressively shrinks up to collapse as $t \to \infty$.

7.4.1 Checking the Replica Symmetric assumption

In order to check the goodness of the replica symmetric solution of our model, we performed extensive Monte Carlo simulations mimicking the evolution of a finite-size network made of N neurons and P patterns. More precisely, for a given realization of the patterns ξ_i^μ, $T = \beta^{-1}$ and sleeping time t, we prepared the system near a randomly extracted pure state and let it evolve with sequential Glauber dynamics. Once the equilibrium state is reached,[1] we measure the thermal average of the Mattis overlap m_1. We performed these simulations for M realizations of the patterns for each different choice of the (N, P, β, t) parameters. A sample of results is shown in Fig. 7.7. We notice that, as t increases, the Mattis magnetization m_1 corresponding to the retrieved pattern vanishes at large values of T and λ. Remarkably, these results are also quantitatively consistent with those presented in Fig. 7.4. This check strongly corroborates the analytical findings. Further, in Fig. 7.8, we also report a finite size scaling of the MC for some values of t and λ and compare them to the theoretical predictions. Finite-size effects tend to overestimate the magnetization at temperatures just above the critical one. However, they are strongly downsized as the sleeping time flows.

7.4.2 Fields distributions in retrieved states

The next step, as standard in numerical approach to neural networks model, is to study the probability distribution of the internal fields in re-

[1] This can be checked by evaluating the stability of observables and the width of their fluctuations

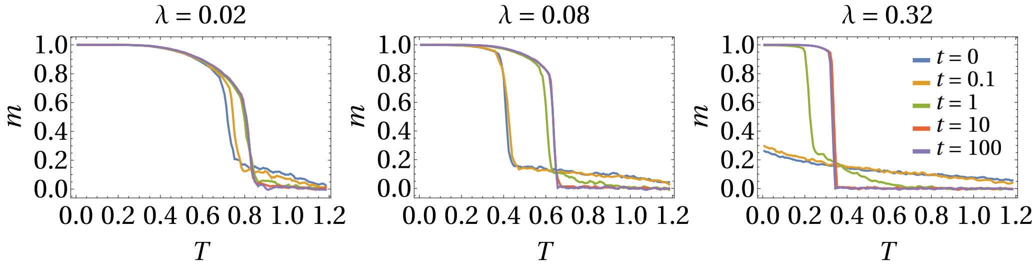

Figure 7.7: **Results from Monte Carlo simulations.** These panels report the results from Monte Carlo simulations for different choices of the parameters (P, β, t) and fixing $N = 5000$ and $M = 10$. From the left to the right, $\lambda = 0.0, 0.08, 0.32$. Also, we considered $t = 0, 0.1, 1, 10, 100$, which are depicted in different colors.

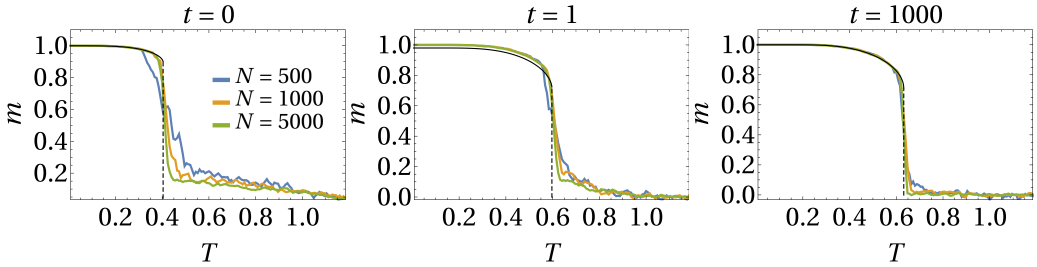

Figure 7.8: **Finite size scaling.** Average values for the Mattis magnetization m corresponding to the retrieved pattern ξ^1 obtained from numerical simulations with $\lambda = 0.08$ and $M = 10$. We consider different sizes ($N = 500, 1000, 5000$), and compare them to the theoretical solution of Eq. (7.18) in the thermodynamic limit (black curves). Each panel correspond to a different choice of t ($t = 0, 1, 1000$).

trieval states. To do this, we again perform extensive MC simulations at fixed network size N and for various sleep extents t ($= 0, 1, 2$). Since we want to examine the effects of reinforcement and remotion in the retrieval regime, we have to work with a storage capacity for which retrieval is certainly feasible (namely where pure states dominate the free energy landscape) for each t. Our choice of the parameters is $N = 1000$ and $P = 50$, with a ratio P/N well below the theoretical (Hopfield) critical threshold. We start the simulations from random initial configurations and simple check that the dynamics ends in a retrieval state. The dynamics is performed with standard Glauber dynamics

$$\sigma_i(\tau + 1) = \text{sign}[h_i(\tau)], \tag{7.68}$$

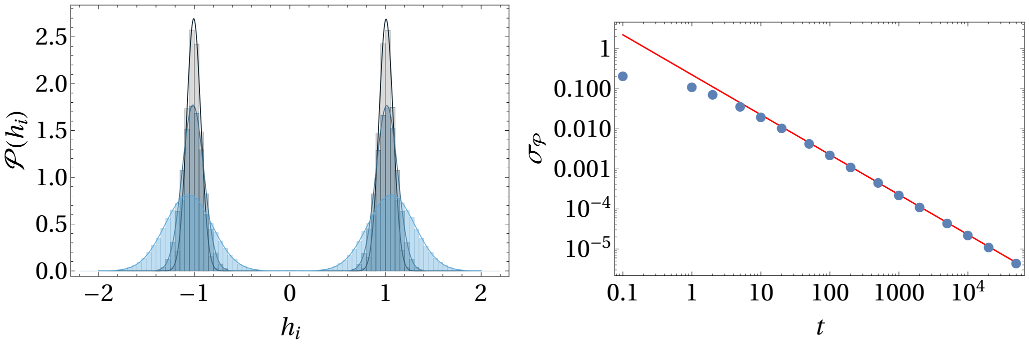

Figure 7.9: **Internal fields probability densities for various unlearning time**. On the left, the plot shows the numerical results (histograms) of the Monte Carlo simulations for the internal fields configuration and the comparison with best-fitting Gaussian distributions (smooth curves). The values of the unlearning time here considered are $t = 0$ (standard Hopfield case, in light blue), $t = 1$ (dark blue) and $t = 2$ (light gray). The statistics used in numerical simulations consists in 20 different stochastic evolutions (with different random initial conditions) and 20 different realizations of the stored patterns. On the right, the plot shows the standard deviation of the (best-fitting) Gaussian distribution of the internal fields configuration as a function of the unlearning time obtained by the previously described MC simulations. The results are again average on 20 different stochastic evolutions (with different random initial conditions) and 20 different realizations of the stored patterns for each unlearning time choice. The fit returns a power-law scaling as $\sigma(t) \sim 0.224 \cdot t^{-0.998}$.

where now the internal fields are computed as

$$h_i = \frac{1}{N} \sum_{j=1}^{N} \sum_{\mu\nu} \xi_i^{\mu} \xi_j^{\nu} (1+t)(1+tC)^{-1}_{\mu\nu} \sigma_j. \tag{7.69}$$

From the internal field configurations, we estimated numerically the probability density function $\mathcal{P}(h)$ (represented by histograms in the left plot of Fig. 7.9) and compared it to a standard Gaussian distribution. Remarkably, the fields distribution $\mathcal{P}(h)$ become more narrow as the sleeping time flows. Indeed, the standard deviation $\sigma_{\mathcal{P}(h)}$ scales as a power law in t, i.e. $\sigma_{\mathcal{P}(h)} \sim 1/t$, resulting from the fit in the right plot of Fig. 7.9 (the red curve). Thus, sleeping regularizes the internal field distributions, as can be seen by inspecting the plots in Fig. 7.9.

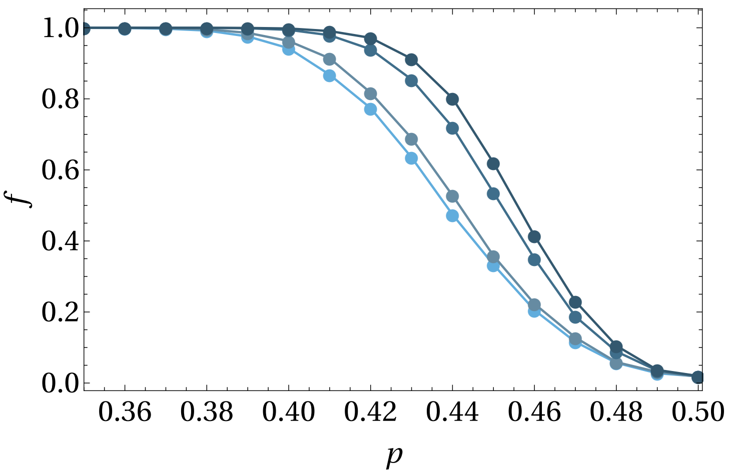

Figure 7.10: **Analysis of attraction basins.** The plots shows the retrieval frequency as a function of the spin-flip probability for $t = 0$, 0.1, 1 and 1000 (from the left to the right). These results are obtained with 200 different stochastic evolutions for each of the 200 pattern realizations.

7.4.3 Retrieval frequency for noisy inputs

Finally, it is also instructive to investigate the evolution of attraction basins of pure attractors. To do this, we proceed in a way similar to what we did for the relativistic Hopfield model. The first difference is that, here, we prepare the network in one of the pure states (say ξ^1), then introduce some noise p (meaning that each spin is flipped with probability p) and consider this configuration as initial condition for the network dynamics. The second difference is that the MC simulations are performed at zero thermal noise $T = 0$. Thus, we let the system evolve toward the equilibrium and measure the retrieval frequency f (i.e. the fractions of MC evolutions ending in a pure state) as a function of the noise p in the input. Also in this case, the network parameters are fixed to $N = 1000$ and $P = 50$. The results are plotted in Fig. 7.10. It clearly emerges that, as time t flows, dreaming has the effect of enlarging the attraction basins of pure memories. This is in agreement with the observation that - increasing the sleep extent - the retrieval region becomes larger (w.r.t. the Hopfield reference).

Conclusions

In this Section, we summarize the journey taken along this thesis. Recalling that our aim is to offer an extensive survey of the statistical mechanical approach to neural networks and machine learning, we first introduced a set of tools, namely the variational extremization procedures based on the Thermodynamical Principles (even in their statistical inference interpretation) and those related the Mechanical Principles (thanks to the mechanical analogy). Next, we used these tools to address the two limiting physical scenarios of interest for Artificial Intelligence: the mean-field ferromagnet (i.e. the Curie-Weiss model, CW) and the mean field spin glass (i.e. the Sherrington-Kirkpatrick model, SK). Then, we introduced the Hopfield model, that is a classical model used to mimic associative memory, and we showed that it recovers the CW and the SK models in the limit of, respectively, just one and too many stored patterns.[1] One step forward, the Hopfield model is studied in details: we discussed its original version with Boolean patterns, its real-valued extension (whose pattern entries are sampled from i.i.d. Gaussians) in which no retrieval region is present, and an hybrid version, accounting the storage of mixed information (namely whose pattern entries can be both analog or digital), the latter sharing the same phase diagram of the standard Hopfield model. This observation is crucial, especially under the analogy between the archetype for statistical learning - namely, the Restricted Boltzmann Machine - and the Hopfield model; in fact, as we showed, the two models share the same Gibbs probability distribution, thus suggesting a unified picture where learnt features from training in Boltzmann learning become retrievable patterns in Hopfield network. Such an analogy, however, requires that the bulk of patterns have to be real-valued such that stochastic gradient descent (and its variation) can be applied during the training of the network (thus ultimately motivating the interest for the hybrid neural network). It is important to remark that, in such an equivalence between Hopfield networks and Boltzmann machines, it emerges that the capacity

[1] Note that these extrema, in machine learning, mirror in turn the two extrema of under-fitting and over-fitting regimes.

for the former (i.e. $\lambda = \lim_{N\to\infty} P/N$) mirrors the ratio between the sizes of the hidden layer (i.e. P) over the visible layer (i.e. N), thus we have found a direct connection between the transition from retrieval to spin glass region in machine retrieval and the transition from a good statistical inference toward an overfitting regime in machine learning. This clearly implies that networks with the possible largest critical capacity are also those less prone to overfitting, and this motivates our next investigations to improve the Hopfield's critical capacity (that, we recall, is $\lambda_c \sim 0.14$, quite far from the maximal capacity for symmetric networks, i.e. $\lambda_c = 1$, as achieved by our extension). The main reason for a small critical capacity in the Hopfield paradigm is that the underlying Hebbian learning yields to a proliferation of spurious mixtures (that occupy huge volumes in the free energy landscape that, if removed, would allow free room for further pure pattern storage). Therefore, we started to exploit unlearning techniques, trying to get rid off these spurious states. Remarkably, for a thesis in Theoretical Physics, it is mandatory to note that we have been entirely driven by the mechanical analogy toward the first working generalization of the Hopfield network, namely its relativistic expression (but we will discuss later on the methodologies).

The mechanical analogy acted as a first guide, but - for technical reasons - it was not mature enough to tackle the high-storage regime of associative neural networks, hence we generalized the model offering a totally novel perspective: we proposed a *daily routine* for associative neural networks where the network Hebbian-learns during the *awake state* (thus behaving as a standard Hopfield model), then, during its *sleep state*, optimizing information storage, it consolidates pure patterns and removes spurious ones. This procedure forces the synaptic matrix to collapse to the projector one (ultimately approaching the Kanter-Sompolinksy model). This procedure keeps the learning Hebbian-like but, by taking advantage of a (properly stylized) sleep phase, still reaches the maximal critical capacity (for symmetric interactions).

Finally, as a last point of investigation, we find that, as long as the network is awake, ergodicity is bounded by the Amit-Gutfreund-Sompolinsky critical line (as it should), but sleeping destroys spin-glass states by extending both the retrieval and the ergodic region: after an entire sleeping session the solely surviving regions are retrieval and ergodic ones. Clearly, this allows the network to achieve the *perfect retrieval regime* (where the number of storable patterns exactly equals the number of neurons the network is built of). Summarizing all these findings, it is our opinion that we should enlarge the initial definition of *cognition* that we gave, splitting it between *learning* and *retrieval*, in order to account - as a true cognitive phase - also *sleeping*. Indeed, the latter, as we modeled it in the last Chapter, suggest a new bridge between a perfectly working machine retrieval model and a new tripartite Restricted Boltzmann Machine (see Fig. 7.11), whose inferential features constitute a very appealing open problem which we intend to address in our future works.

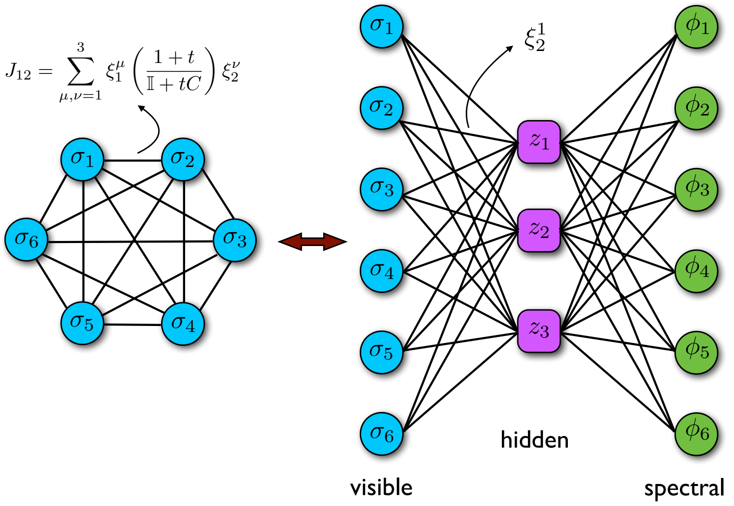

Figure 7.11: Stylized representation of the generalized Hopfield network (left) and its dual generalized (restricted) Boltzmann machine (right), namely the three-partite spin-glass under study: in machine learning jargon these parties are called *layers* and, here, they are respectively the visible, hidden and spectral layers. Note further that, as it should, when the network has not slept yet, the above duality reduces to the standard picture of Hopfield networks and restricted Boltzmann machines, see Chapter 5 and [3, 21, 41].

Moving on the techniques, we paid particular care when trying to present the exposition with some mathematical rigour. As the concepts behind the various models considered and the techniques used for their investigation are several and some of them somehow tricky, we tried to preserve the same *narrative scheme* whenever possible. All the models are at first introduced (i.e. they are all defined via their cost function, or Hamiltonian) and equipped with their related statistical mechanical package of definitions and tools (among whose, of primary importance the free energy - or pressure - whose analysis allows to paint the phase diagram of these models). Then, when possible, we systematically proved the existence of the infinite volume limit for this crucial function,[1] then we moved to search for its explicit

[1] Unfortunately, this has been possible for several but not all the models discussed

expression in terms of the natural order parameters related to the models always with the same approaches: at first an heuristic one (typically the replica trick), then with the one-parameter Guerra's interpolation technique and also with two-parameters Guerra's interpolation technique, namely with the mechanical analogy (by which we use the Hamilton-Jacobi scheme to solve for the free energy of these models, the latter playing as a mechanical action). Such an analogy has been the mathematical guide to overcome the actual state of the art regarding optimization storage in AI and, at the same time, it tacitly suggested a powerful physical extension of the standard paradigm, that turned out to be extremely fruitful.

Clearly the journey in the AI world is far from being over (nor it has been exhaustive in the thesis obviously): for instance, there is still a long way toward the comprehension of the learning skills of these dreaming neural networks (a missing point that is entirely to be investigated). We hope we will have the possibility to keep on working in the Academia in order to address this point and that Theoretical Physics will keep on leading the rigorous foundations of AI.

in the thesis. The high storage regime of associative neural network still escapes such a strong control.

Bibliography

[1] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cognitive Sci. **9**.1:147-169, (1985).

[2] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, *Dreaming neural networks: rigorous results*, J. Stat. Phys., in press (2019).

[3] 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).

[4] E. Agliari, A. Barra, B. Tirozzi, *Free energies of Boltzmann Machines: self-averaging, annealed and replica symmetric approximations in the thermodynamic limit*, J. Stat., in press.

[5] E. Agliari, et al., *Hierarchical neural networks perform both serial and parallel processing*, Neural Networks **66**, 22-35, (2015).

[6] E. Agliari, et al, *Immune networks: multitasking capabilities near saturation*, J. Phys. A: Math. & Theor. **46**(41):415003, (2013)

[7] E. Agliari, et al., *Multitasking associative networks*, Phys. Rev. Lett. **109**, 268101, (2012).

[8] E. Agliari, et al., *Multitasking attractor networks with neuronal threshold noises*, Neural Networks **49**, 19, (2013).

[9] E. Agliari, et al., *Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines*, Neural Networks **38**, 52, (2013).

[10] M. Aizenman, P. Contucci, *On the stability of the quenched state in mean-field spin-glass models*, J. Stat. Phys. **92**(5-6):765, (1998).

[11] D.J. Amit, *Modeling brain functions*, Cambridge Univ. Press (1989).

[12] D.J. Amit, H. Gutfreund, H. Sompolinsky, *Spin-glass models of neural networks*, Phys. Rev. A **32**.2:1007, (1985).

[13] D.J. 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).

[14] T. Andrillon, D. Pressnitzer, D. Leger, S. Kouider, *Formation and suppression of acoustic memories during human sleep*, Nature Comm. **8**, 179, (2018).

[15] E. Barkai, D. Hansel, I. Kanter, *Statistical mechanics of a multilayered neural network*, Phys. Rev. Lett., **65**, 2312 (1990).

[16] E. Barkai, I. Kanter, *Storage capacity of a multilayer neural network with binary weights*, Europhys. Lett., **14**, 107 (1991).

[17] A. Barra, *The mean field Ising model trough interpolating techniques*, J. Stat. Phys. **132**(5):787, (2008).

[18] A. Barra, M. Beccaria, A. Fachechi, *A new mechanical approach to handle generalized Hopfield neural networks*, Neural Networks **106**:205-222, (2018).

[19] A. Barra, A. Di Biasio, F. Guerra, *Replica symmetry breaking in mean field spin glasses trough Hamilton-Jacobi technique*, JSTAT P09006, (2010).

[20] A. Barra, et al., *On quantum and relativistic mechanical analogues in mean field spin models*, Proc. Royal Soc. A (London) **470**:20140589, (2014).

[21] A. Barra, et al., *On the equivalence among Hopfield neural networks and restricted Boltzman machines*, Neural Networks **34**, 1-9, (2012).

[22] A. Barra, et al., *Phase Diagram of Restricted Boltzmann Machines & Generalized Hopfield Models*, Phys. Rev. E **97**, 022310, (2018).

[23] A. Barra, et al., *Phase transitions of Restricted Boltzmann Machines with generic priors*, Phys. Rev. E **96**, 042156, (2017).

[24] A. Barra, G. Genovese, F. Guerra, *Equilibrium statistical mechanics of bipartite spin systems*, J. Phys. A (Math. & Theor.) **44**.24:245002, (2011).

[25] A. Barra, G. Genovese, F. Guerra, *The replica symmetric approximation of the analogical neural network*, J. Stat. Phys. **140**.4:784, (2010).

[26] A. Barra, F. Guerra, *About the ergodic regime of the analogical Hopfield neural network*, J. Math. Phys. **49**, 125217, (2008)

[27] A. Barra, F. Guerra, G. Genovese, D. Tantari, *How glassy are neural networks?*, JSTAT P07009, (2012).

[28] R.J. Baxter, *Exactly solved models in statistical mechanics*, Courier Dover Publ., (2007).

[29] A. Bovier, V. Gayrard, *Hopfield models as generalized random mean field models*, Mathematical aspects of spin glasses and neural networks, Birkhauser Press, Boston, (1998).

[30] A. Bovier, V. Gayrard, P. Picco, *Gibbs states of the Hopfield model in the regime of perfect memory*, Prob. Theor. Rel. Fields **100**.3:329-363, (1994).

[31] A. Bovier, V. Gayrard, P. Picco, *Gibbs states of the Hopfield model with extensively many patterns*, J. Stat. Phys. **79**.1: 395-414, (1995).

[32] A. Bovier, V. Gayrard, *The retrieval phase of the Hopfield model: a rigorous analysis of the overlap distribution*, Prob. Theor. Rel. Fields **107**.1:61-98, (1997).

[33] P. Carmona, Y. Hu, *Universality in Sherrington–Kirkpatrick's spin glass model*, Ann. Henri Poincarè **42**, 2, (2006).

[34] A.M. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. Le Cun, *The loss surfaces of multilayer networks*, Artificial Intelligence and Statistics, (2015).

[35] C.A. Christos, *Investigation of the Crick-Mithinson reverse-learning dream sleep hypothesis in a dynamical setting*, Neural Net. **9**(3):427-434, (1996).

[36] S. Cocco, R. Monasson, *Adaptive cluster expansion for inferring Boltzmann machines with noisy data*, Phys. Rev. Lett. **106**.9: 090601, (2011).

[37] A.C.C. Coolen, R. Kuhn, P. Sollich, *Theory of neural information processing systems*, Oxford Press (2005).

[38] A.C.C. Coolen, D. Sherrington, *Dynamics of fully connected attractor neural networks near saturation*, Phys. Rev. Lett. **71**(23):3886, (1993).

[39] F. Crick, G. Mitchinson, *The function of dream sleep*, Nature **304**, 111, (1983).

[40] P. Dayan, B.W. Balleine, *Reward, motivation, and reinforcement learning*, Neuron **36**.2:285-298, (2002).

[41] A. Decelle, G. Fissore, C. Furtlehner, *Spectral Learning of Restricted Boltzmann Machines*, arXiv preprint arXiv:1708.02917, (2017).

[42] B. Derrida, E. Gardner, A. Zippelius, *An exactly solvable asymmetric neural network model*, Europhys. Lett. **4**.2: 167, (1987).

[43] C. Di Castro, R. Raimondi, *Statistical Mechanics and Applications in Condensed Matter*, Cambridge University Press (2015).

[44] S. Diekelmann, J. Born, *The memory function of sleep*, Nature Rev. Neuroscience **11**(2):114, (2010).

[45] V. Dotsenko, *An introduction to the theory of spin glasses and neural networks*, World Scientific, (1995).

[46] V. Dotsenko, B. Tirozzi, *Replica symmetry breaking in neural networks with modified pseudo-inverse interactions*, J. Phys. A **24**:5163-5180, (1991).

[47] V. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, *Statistical mechanics of Hopfield-like neural networks with modified interactions*, J. Phys. A **24**, 2419, (1991).

[48] R. Ellis, *Entropy, large deviations, and statistical mechanics*, Taylor & Francis press (2005).

[49] A. Engel, C. Van den Broeck, *Statistical mechanics of learning*, Cambridge University Press (2001).

[50] A. Fachechi, E. Agliari, A. Barra, *Dreaming Neural Networks: forgetting spuirous memories and reinforcing pure ones*, Neural Networks, in press (2019).

[51] J.A. Fodor, Z.W. Pylyshyn, *Connectionism and cognitive architecture: A critical analysis*, Cognition **28**(1):3-71, (1988).

[52] E. Gardner, *Maximum storage capacity in neural networks*, Europhys. Lett. **4**(4):481, (1987).

[53] E. Gardner, *The space of interactions in neural network models*, J. Phys. A **21**, 257, (1998).

[54] G. Genovese, *Universality in bipartite mean field spin glasses*, J. Math. Phys. **53**(12):123304, (2012).

[55] G. Genovese, et al., *A mechanical approach to mean field spin models*, J. Math. Phys. **50**(5), 053303, (2009).

[56] S. Ghirlanda, F. Guerra, *General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity*, J. Phys. A **31**(46):9149, (1998).

[57] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, M.I.T. press (2017).

[58] F. Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Comm. Math. Phys. **233**(1):1-12, (2003).

[59] F. Guerra, *Sum rules for the free energy in the mean field spin glass model*, Fields Inst. Comm. **30**:161, (2001).

[60] F. Guerra, F.L. Toninelli, *The thermodynamic limit in mean field spin glass models*, Comm. Math. Phys. **230**(1):71-79, (2002).

[61] D.O. Hebb, *The organization of behavior: A neuropsychological theory*, Psychology Press, (1949).

[62] J. Hertz, R. Palmer, *Introduction to the theory of neural networks*, Lecture Notes, (1991).

[63] G. Hinton, R.R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks*, Science, **313**(5786), 504-507 (2006).

[64] G. Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*, available at http://learning.cs.toronto.edu, (2010).

[65] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the national academy of sciences 79.8 (1982): 2554-2558.

[66] J.J. Hopfield, D.I. Feinstein, R.G. Palmer, *Unlearning has a stabilizing effect in collective memories*, Nature Lett. **304**, 280158, (1983).

[67] J.J. Hopfield, D.W. Tank, *Neural computation of decisions in optimization problems*, Biol. Cybern. **52**(3):141-152, (1985).

[68] J.A. Horas, P.M. Pasinetti, *On the unlearning procedure yielding a high-performance associative memory neural network*, J. Phys. A **31**, L463-L471, (1998).

[69] H. Huang, *Reconstructing the Hopfield network as an inverse Ising problem*, Phys. Rev. E **81**.3:036104, (2010).

[70] H. Huang, K.Y. Michael Wong, and Y. Kabashima, *Entropy landscape of solutions in the binary perceptron problem*, J. Phys. A **46**, 375002, (2013).

[71] H. Huang, T. Toyoizumi, *Advanced mean-field theory of the restricted Boltzmann machine*, Phys. Rev. E **91**.5:050101, (2015).

[72] E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106**.4:620, (1957).

[73] L.P. Kaelbling, M.L. Littman, A.W. Moore, *Reinforcement learning: A survey*, J. Artif. Intel. Res. **4**:237-285, (1996).

[74] H.J. Kappen, F. Rodriguez, *Efficient learning in Boltzmann machines using linear response theory*, Neural Comput. **10**.5: 1137-1156, (1998).

[75] I. Kanter, H. Sompolinsky, *Associative recall of memory without errors*, Phys. Rev. A **35**.1:380, (1987).

[76] W. Kinzel, M. Opper, *Dynamics of learning*, in: E. Domany, J.L. van Hemmen, K. Schulten (Eds.) *Models of neural networks*, Springer, Berlin, 149-172 (1991).

[77] D. Kirk, *NVIDIA CUDA software and GPU parallel computing architecture*, ISMM.**7**, 103, (2007).

[78] C. Kittel, *Elementary statistical physics*, Courier Dover Publications, (2004).

[79] D. Kleinfeld, D.B. Pendergraft, *Unlearning increases the storage capacity of content addressable memories*, Biophys. J. **51**, 47-53, (1987).

[80] T.O. Kohonen, Self-organization and Associative Memory, Springer, Berlin (1984).

[81] D. Krotov, J.J. Hopfield, *Dense associative memory is robust to adversarial inputs*, arXiv:1701.00939, (2017).

[82] D. Krotov, J.J. Hopfield, *Dense associative memory for pattern recognition*, Adv. Neur. Inform. Proc. Sys., 1172-1180, (2016).

[83] Y. Le Cun, Y. Bengio, G. Hinton, *Deep learning*, Nature **521**:436-444, (2015).

[84] N. Le Roux, Y. Bengio, *Representational power of restricted Boltzmann machines and deep belief networks*, Neural computation **20**.6:1631-1649, (2008).

[85] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).

[86] F. Mandl, G. Shaw, *Quantum field theory*, John Wiley & Sons (Manchester), (2010).

[87] P. Maquet, *The role of sleep in learning and memory*, Science **294**.5544:1048, (2001).

[88] E. Marinari, *Forgetting Memories and their Attractiveness*, arXiv:1805.12368, (2018).

[89] M. McCloskey, N.J. Cohen, *Catastrophic interference in connectionist networks: The sequential learning problem*, Psychol. Learn. & Motiv. **24**:109-165, (1989).

[90] J.L. McGaugh, *Memory - a century of consolidation*, Science **287**.5451:248-251, (2000).

[91] W.S. McCulloch, W.Pitts, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[92] P. Mehta, D.J. Schwab, *An exact mapping between the variational renormalization group and deep learning*, preprint, arXiv:1410.3831, (2014).

[93] M. Mezard, G. Parisi, M.A. Virasoro, *Spin glass theory and beyond*, World Scientific (1985).

[94] M.L. Minsky, S.A. Papert, *Perceptrons - Expanded Edition*, MIT press, Boston, MA, (1987).

[95] Y. Miyashita, H.S. Chang, *Neuronal correlate of pictorial short-term memory in the primate temporal cortex*, Nature **331**.6151:68, (1988).

[96] J. Nickolls, et al., *Scalable parallel programming with CUDA*, Queue **6**(2):40-53, (2008).

[97] K. Nokura, *Paramagnetic unlearning in neural network models*, Phys. Rev. E **54**(5):5571, (1996).

[98] K. Nokura, *Spin glass states of the anti-Hopfield model*, J. Phys. A **31**, 7447, (1998).

[99] K-S. Oh, J. Keechul, *GPU implementation of neural networks*, Pattern Recognition **37**.6:1311, (2004).

[100] G. Parisi, *A memory which forgets*, J. Phys. A **19**, L617, (1986).

[101] L. Pastur, M. Shcherbina, B. Tirozzi, *On the replica symmetric equations for the Hopfield model*, J. Math. Phys. **40**(8): 3930, (1999).

[102] 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).

[103] L. Personnaz, I. Guyon, G. Dreyfus, *Information storage and retrieval in spin-glass like neural networks*, J. Phys. Lett. **46**, L-359:365, (1985).

[104] A.Y. Plakhov, *The converging unlearning algorithm for the Hopfield neural network: optimal strategy*, IEEE Int. Conf. on Pattern Recognition Vol. 2-Conference B: Computer Vision & Image Processing (1994).

[105] A.Y. Plakhov, S.A. Semenov, I.B. Shuvalova, *Convergent unlearning algorithm for the Hopfield neural network*, IEE Comp. Soc. Press. **2**(95), 30, (1995).

[106] A.Y. Plakhov, S.A. Semenov, *The modified unlearning procedure for enhancing storage capacity in Hopfield network*, IEEE Trans. 242, (1992).

[107] B. Rasch, J. Born, *About sleep's role in memory*, Physiol. Rev. **93**:681-766, (2013).

[108] E.T. Rolls, A. Treves, *Neural Networks and Brain Function*, Oxford University Press, (1998).

[109] M. Rosen-Zvi, A. Engel, I. Kanter, *Multilayer neural networks with extensively many hidden units*, Phys. Rev. Lett. **87**, 078101 (2001).

[110] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychol. Rev. **65**(6):386, (1958).

[111] D. Ruelle, *Small Random Perturbations of Dynamical Systems and the Definition of Attractors*, Commun. Math. Phys. **82**, 137-151 (1981).

[112] D. Ruelle, *Statistical mechanics: Rigorous results*, World Scientific (1999).

[113] R. Salakhutdinov, G. Hinton, *Deep boltzmann machines*, Artificial Intelligence and Statistics (2009).

[114] R. Salakhutdinov, H. Larochelle, *Efficient learning of deep Boltzmann machines*, Proc. thirteenth int. conf. on artificial intelligence and statistics, 693, 2010.

[115] J. Schmidhuber, *Deep learning in neural networks: An overview*, Neural networks **61**: 85-117, (2015).

[116] E. Schneidman, M.J. Berry II, R. Segev, M. Bialek, *Weak pairwise correlations imply strongly correlated network states in a neural population* Nature **440**(7087):1007, (2006).

[117] H.S. Seung, H. Sompolinsky, N. Tishby, *Statistical mechanics of learning from examples*, Phys. Rev. A **45**(8):6056, (1992).

[118] C.E. Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal, **27**: 379–423, 623–656, (1948).

[119] D. Sherrington, S. Kirkpatrick, *Solvable model of a spin-glass*, Phys. Rev. Lett. **35**(26):1792, (1975).

[120] N. Srivastava, R. Salakhutdinov, *Multimodal learning with deep boltzmann machines*, Adv. Neural Inform. Proc. Sys. , 2222, (2012).

[121] H. Steffan, R. Kühn, *Replica symmetry breaking in attractor neural network models*, Zeitschrift für Physik B Condensed Matter, **95**(2), 249–260 (1994).

[122] 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).

[123] R. Stickgold, J.A. Hobson, R. Fosse, M. Fosse, *Sleep, Learning and Dreams: Off-line Memory Reprocessing*, Science **294**, 1052, (2001).

[124] R.S. Sutton, A.G. Barto, *Reinforcement learning: An introduction*, MIT press, (1998).

[125] M. Talagrand, *Exponential inequalities and convergence of moments in the replica-symmetric regime of the Hopfield model*, Ann. of Prob. 1393, (2000).

[126] M. Talagrand, *Rigorous results for the Hopfield model with many patterns*, Prob. Theor. Rel. Fiel. **110**(2):177, (1998).

[127] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models*, Springer Science & Business Media, (2003).

[128] R.F. Thompson, *The neurobiology of learning and memory*, Science **233**.4767:941-947, (1986).

[129] E.W. Tramel, A. Dremeau, F. Krzakala, *Approximate message passing with restricted Boltzmann machine priors*, JSTAT 073401, (2016).

[130] J. Tubiana, R. Monasson, *Emergence of compositional representations in restricted Boltzmann machines*, Phys. Rev. Lett. **118**(13), 138301 (2017).

[131] H.C. Tuckwell, *Introduction to theoretical neurobiology*, Cambridge University Press (2005).

[132] A.M. Turing, *Computing machinery and intelligence* Mind, 433–460, (1950).

[133] J. Von Neumann, *The general and logical theory of automata*, Cerebral mechanisms in behavior, 1–41,(1951).

[134] J.L. Van Hemmen, L.B. Ioffe, R. Kühn, M. Vaas, *Increasing the efficiency of a neural network through unlearning*, Physica A: Statistical Mechanics and its Applications, **163**, 386-392 (1990).

[135] S. Wimbauer, J. Leo van Hemmen, *Hebbian unlearning*, Analysis of Dynamical and Cognitive Systems, Springer, Berlin, 1995.

[136] L. Zdeborova, F. Krzakala, *Statistical physics of inference: thresholds and algorithms*, Advances in Physics, **65**, 453-552, (2016).