SAPIENZA
UNIVERSITÀ DI ROMA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Dottorato in Matematica

Retrieval Capabilities of Multitasking and Hierarchical Neural Networks

Statistical mechanics of spontaneous parallel processing

Tesi di Dottorato

Relatori:
Prof. Francesco Guerra
Dott. Elena Agliari
Dott. Adriano Barra

Candidato:
Andrea Galluzzi

Anno Accademico 2014/2015

Summary

Scope of the present research is to develop novel mathematical tools in order to face the continuously growing need of modern theoretical approaches for a proper development of Artificial Intelligence.
Using Statistical Mechanics and Graph Theory languages and techniques, we will start this thesis by introducing the mean field Hopfield model as the *harmonic oscillator* in Neural Networks. This will set the reference framework in order to extend its capabilities: in our research , we succeed in formalizing for the first time neural networks able to spontaneous parallel processing (a step forward with respect to the original harmonic oscillator, where only sequential processing was allowed to emerge as a collective feature shared over the distributed memories across the net).
Indeed, the Hopfield model (together with the related Hebb's learning rule) provides a prototypical associative memory model that has attracted a great attention by the communities of Theoretical Physicists and Mathematicians mainly due to its natural formalization within the canonical setting of Statistical Mechanics (possibly beyond the adherence of its processing paths with those empirically found in biological information processing systems).
Through well controlled learning procedures, in this attractor networks it is possible to store and sequentially retrieve patterns of information. The retrieval of a stored pattern does coincide, mathematically, with the thermalization of the system in one of the several minima of the related free energy (each minimum corresponding to a pattern to be retrieved) such that, through the analogy between thermodynamical relaxation and selection of a distributed memory, we can adapt the mathematical tools (i.e. models and methods) originally developed for statistical mechanical treatments of spin glasses (other complex systems whose free energy landscape is rugged) to the analysis of neural networks, and, in this thesis, this is the route that we aim to contribute to pave, moving from serial to parallel information processing.
Indeed, properly modifying the structure of the memories -pattern's definitions- (in the pertinent phase space where the system dynamics takes place) or carefully diluting the network architecture (in the topological space where spins dialogue) we will build models of neural networks able to recall simultaneously multiple patterns of information. We will therefore analyze in details the mathematical structure of these networks and discuss the resulting properties.

The thesis is structured as follows:

In the first Chapter we briefly revise the Hopfield model: after an historical digression on the role of the so-called *mean-field* approximation in Physics

(and in particular in Statistical Mechanics), we will construct its related Hamiltonian in two novel ways (with respect to the original Hopfield proposal). More precisely, starting from the paradigmatic models for ferromagnets and for spin-glasses (i.e., the Curie-Weiss model and the Sherrington-Kirkpatrick model, respectively) we will show how to recover the Hopfield model and the underlying deep connections among these models.

The second Chapter is entirely dedicated to parallel processing networks and it is split into two main Sections, the former dealing with multitasking network, the latter dealing with hierarchical network.
We will start with purely mean field models, the so-called *multitasking associative networks* and we will perform an extended treatment of its capabilities and properties, mixing techniques stemming from Statistical Mechanics and Graph Theory (whose usage is more typical for Theoretical Phycisists and Mathematicians) with those of common usage in Robotics and Automation as Signal-to-Noise, stability analysis and other related operational approaches. After discussing as toy-examples the simultaneous retrieval of two or three patterns, we will explore the whole *low-storage behavior* of the network, that can be defined in a simple way as follows: consider a network built of by N binary spins (i.e. Ising spins), that we want to use to store and retrieve P patterns (i.e., N-length vectors of binary entries ± 1). Now, as we are interested in the network performances in the thermodynamic limit (i.e. sending $N \to \infty$ in order to deal with averages, rather than full probability distributions), we need to specify how P scales with N. If such a scaling is extensive, namely if $P \propto N$, we talk of *high storage* regime, while if the amount of patters scales sub-linearly in the number of spins (such that $\lim_{N\to\infty}(P/N) \to 0$), we talk of *low storage*.
At a first glance, the low storage regime looks as a pathological regime or a simplifying analysis avoiding the high storage, but, actually, this is not the case. The origin of this idea lies in the properties of the Hopfield network and, in particular, in the theory of Amit, Gutfreund and Sompolinsky who showed how to load that original network in order to let it work in the high storage regime. However, to understand that most modern variants of the Hopfield network can not handle extensive storage (i.e. $P \sim N$) it is enough a simple and heuristical consideration of Graph Theory: the Hopfield model is a fully connected mean-field network. This implies that, as the memory is distributed -namely it is shared over the synapses (i.e. the links connecting the spins and whose values can be both positive and negative tacitly locating neural networks in the larger bulk of spin glasses)- we can feed $O(N^2)$ synapses (i.e. links) with the information contained in the patterns to store. However let us now consider a minimal modification of the Hopfield model that makes it more biologically plausible: let us collapse the Hopfield network

on an Erdös-Rényi graph (instead of the original fully connected network). This has the advantage of avoiding the assumption that each neuron interacts with all the other neurons in the network, that is clearly biologically false, despite mathematically convenient. However, from an Artificial Intelligence perspective, the major difference between a random graph and a fully connected network resides in the number of links: N^1 for the former, N^2 for the latter. It is then evident that, as the amount of synapsis does no longer scale quadratically with the amount of neurons, the overall network performance can not remain unaltered. This is a general result when embedding associative networks on structured or biological interesting topologies (and it is a particularly severe limitation for Hebb learning rules, as those we will investigate in this work).
Once understood this theoretical bound to the maximal storage capacity of the variations on the Hopfield theme, we analyze in all details our multitasking extension: a key (and novel) assumption is the introduction of blank entries in pattern's definition, that is, pattern entries may assume values ± 1 (carrying information) or simply be blank (denoting lack of information). It is remarkable that this novel approach to dilution, that is seen as a must by Biologists, will play as the real core of parallel processing such that, making the network topology more adherent to biological demands, we will also obtain -as a result- that network's performances also match better those of biological neural networks.
Once explored exhaustively the multitasking network, we will try to face another fundamental and intrinsic limitation of the original Hopfield scenario: its mean-field nature. To overcome this obstacle -at least partially- we try to adapt the hierarchical ferromagnet, introduced by Dyson in the Literature almost four decades ago, implementing on its structure the Hebb rule for learning and inferring the resulting properties the network spontaneously shows.
Concretely, we introduce and investigate the statistical mechanics of hierarchical neural networks: in these systems, spins interact with a strength that is a (decreasing) function of a suitably introduced concept of *distance*, such that different levels (i.e. hierarchies) of degenerate-strength couplings immediately emerge.

First, we approach these systems à la Mattis, that is, by thinking at the Dyson model as a single-pattern hierarchical neural network, and, through this perspective, we discuss the stability of different retrievable states as predicted by the related (approximate) self-consistencies equation. The mathematical key argument here is properly reabsorbing fluctuations of the magnetization related to higher levels of the hierarchy into effective fields for the lower levels: remarkably, mixing Amit's ansatz technique (to select

candidate retrievable states) with the interpolation procedure (to solve for the free energy of these states) we show that (due to gauge symmetry) the Dyson model accomplishes both serial and parallel processing.
One step forward, we extend this scenario toward multiple stored patterns by implementing the Hebb prescription for learning within the couplings. This results in an Hopfield-like networks constrained on a hierarchical topology, for which, restricting to the low storage regime (where the number of patterns grows at most logarithmical with the amount of spins), we give an explicit expression of its mean field bound and of the related improved bound.
As a result of the present investigation, the hierarchical neural network (both for its underling topology, as well as for its emerging properties) is actually much closer to real biology with respect to neural network models previously developed.
Finally, our general considerations on the whole strategy exploited in this Ph.D. training period will be collected in the Conclusions of the thesis.

Papers published (or submitted) during my Ph.D. time

Published papers:

1. Title: *Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines.*
   Authors: Elena Agliari, Adriano Barra, Andrea De Antoni, Andrea Galluzzi
   Journal: **Neural Networks** Volume 38, Pages 52-63 (2013).

2. Title: *Multitasking associative networks.*
   Authors: Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, Francesco Moauro
   Journal: **Physical Review Letters** Volume 109 Numero 26 Pages 268101 (2012).

3. Title: *Parallel processing in immune networks.*
   Authors: Elena Agliari, Adriano Barra, Silvia Bartolucci, Andrea Galluzzi, Francesco Guerra, Francesco Moauro
   Journal: **Physical Review E** Volume 87 Numero 4 Pages 042701 (2013).

4. Title: *Multitasking attractor networks with neuronal threshold noise.*
   Authors: Elena Agliari, Adriano Barra, Andrea Galluzzi, Marco Isopi
   Journal: **Neural Networks** Volume 49, Pages 19-29, (2014).

5. Title: *Ferromagnetic models for cooperative behavior: Revisiting Universality in complex phenomena.*
   Authors: E. Agliari, A. Barra, A. Galluzzi, A. Pizzoferrato, D. Tantari,
   Proceeding: Proceedings of the Conference **Mathematics for Planet Heart** IndAM (2013).

6. Title: *Mean field bipartite spin models treated with mechanical techniques.*
   Authors: A. Barra, A. Galluzzi, F. Guerra, A. Pizzoferrato, D Tantari
   Journal: **The European Physical Journal B**, DOI: 10.1140/epjb/e2014-40952-4 (2013).

7. Title: *A walk in the statistical mechanical formulation of neural networks.*
   Authors: E. Agliari, A. Barra, A. Galluzzi, D. Tantari, F. Tavani

Proceeding: **The NCTA2014: Neural computation theory and application**, (2014).

8. Title: *Metastable states in the hierarchical Dyson model drive parallel processing in thehierarchical Hopfield network.*
   Authors: E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, F. Tavani
   Journal: **Journal of Physics A: Mathematical and Theoretical** 48 (1), 015001 (2015).

9. Title: *Retrieval capabilities of hierarchical networks: from Dyson to Hopfield.*
   Authors: E. Agliari, A. Barra, A. Galluzzi, F. Guerra D. Tantari, F. Tavani
   Journal: **Physical Review Letters**, Vol. 114, n. 2, pages 18923 (2015).

10. Title: *Hierarchical neural networks perform both serial and parallel processing*
    Authors: E. Agliari, A. Barra, A. Galluzzi, F. Guerra D. Tantari, F. Tavani
    Journal: **Neural Networks**. Volume 66, Pages 22-35 (2015).

11. Title: *Topological properties of hierarchical networks*
    Authors: E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, F. Tavani
    Journal: **Physical Review E**. Volume 91, Pages 068101 (2015).

12. Title: *Universality for Couplings Correlation in Mean Field Spin Glasses*
    Authors: A. Galluzzi, F. Guerra, D. Tantari
    Proceeding: **Theory and Applications in Mathematical Physics** (World Scientific) (2015).

Submitted papers

1. Title:
   Emerging heterogeneities in Italian customs.
   Authors:
   Elena Agliari, Adriano Barra, Andrea Galluzzi, Marco Alberto Javarone, Andrea Pizzoferrato, Daniele Tantari
   Journal:
   submitted to PLoS One

2. Title:
   Insights in Economical Complexity in Spain: the hidden boost of migrants in international tradings.
   Journal:
   Elena Agliari, Adriano Barra, Andrea Galluzzi, Francisco Requena-Silvente, Daniele Tantari
   Journal:
   submitted to Nature Palgrave Communications.

Contents

Chapter 1

Introduction

Neural networks are such a fascinating field of science that its development is the result of contributions and efforts from an incredibly large variety of scientists, ranging from *engineers* (mainly involved in electronics and robotics) [60, 70], *physicists* (mainly involved in statistical mechanics and stochastic processes) [6, 17], and *mathematicians* (mainly working in logics and graph theory) [5, 22] to *(neuro) biologists* [34, 63] and *(cognitive) psychologists* [13, 44].

Tracing the genesis and evolution of neural networks is very difficult, probably due to the broad meaning they have acquired along the years[1]; scientists closer to the robotics branch often refer to the W. McCulloch and W. Pitts model of perceptron [68] [2], or the F. Rosenblatt version [40], while researchers closer to the neurobiology branch adopt D. Hebb's work as a starting point [21]. On the other hand, scientists involved in statistical mechanics, that joined the community in relatively recent times, usually refer to the seminal paper by Hopfield [49] or to the celebrated work by Amit Gutfreund Sompolinky [18], where the statistical mechanics analysis of the Hopfield model is effectively carried out.

Whatever the reference framework, at least 30 years elapsed since neural networks entered in the theoretical physics research and much of the former results can now be re-obtained or re-framed in modern approaches, as we want to highlight in the present work. In particular, we show that toy models for paramagnetic-ferromagnetic transition [65] are natural proto-

[1] Seminal ideas regarding automation are already in the works of Lee during the XIIX century, if not even back to Descartes, while more modern ideas regarding *spontaneous cognition*, can be attributed to A. Turing [7] and J. Von Neumann [50] or to the join efforts of M. Minsky and S. Papert [58], just to cite a few.

[2] Note that the first "transistor", crucial to switch from analogical to digital processing, was developed only in 1948 [68].

types for the autonomous storage/retrieval of information patterns and play as operational amplifiers in electronics. Then, we move further analyzing the capabilities of glassy systems (ensembles of ferromagnets and antiferromagnets) in storing/retrieving extensive numbers of patterns so to recover the Hebb rule for learning [21] in two different ways (the former guided by ferromagnetic intuition, the latter guided by glassy counterpart), both far from the original route contained in his milestone *The Organization of Behavior*.

1.1 Statistical Mechanics

Hereafter we summarize the fundamental steps that led theoretical physicists towards artificial intelligence; despite this parenthesis may look rather distant from neural network scenarios, it actually allows us to outline and to historically justify the physicists perspective.

Statistical mechanics aroused in the last decades of the XIX century thanks to its founding fathers Ludwig Boltzmann, James Clarke Maxwell and Josiah Willard Gibbs [12]. Its "solely" scope (at that time) was to act as a theoretical ground of the already existing empirical thermodynamics, so to reconcile its noisy and irreversible behavior with a deterministic and time reversal microscopic dynamics. While trying to get rid of statistical mechanics in just a few words is almost meaningless, roughly speaking its functioning may be summarized via toy-examples as follows. Let us consider a very simple system, e.g. a perfect gas: its molecules obey a Newton-like microscopic dynamics (without friction -as we are at the molecular level- thus time-reversal as dissipative terms in differential equations capturing system's evolution are coupled to odd derivatives) and, instead of focusing on each particular trajectory for characterizing the state of the system, we define order parameters (e.g. the density) in terms of microscopic variables (the particles belonging to the gas). By averaging their evolution over suitably probability measures, and imposing on these averages energy minimization and entropy maximization, it is possible to infer the macroscopic behavior in agreement with thermodynamics, hence bringing together 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) [61], statistical mechanics was immediately recognized as a deep and powerful bridge linking microscopic dynamics of a system's constituents with (emergent) macroscopic properties shown by the system itself, as exemplified by the equation of state for *perfect gases* obtained by considering an Hamiltonian for a single particle accounting for the kinetic contribution

only [12].

One step forward beyond the perfect gas, Van der Waals and Maxwell in their pioneering works focused on *real gases* [52], where particle interactions were finally considered by introducing a non-zero potential in the microscopic Hamiltonian describing the system. This extension implied 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 by 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[3]), this *system* eventually changes its state from liquid to vapor (or solid, depending on parameter values); of course, the same applies generally to liquids.

The fact that the macroscopic behavior 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. 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 [45]) and not too much intricate than circuits that may arise from basic human creativity: remarkably simpler than expected and certainly trivial with respect to overall cerebral functionalities like learning or computation, thus the aptness of a *thermodynamic formulation* of neural interactions -to *reveal* possible emergent capabilities- was immediately pointed out, despite the route was not clear yet.

Interestingly, a big step forward to this goal was prompted by problems stemmed from condensed matter. In fact, theoretical physicists quickly realized that the purely kinetic Hamiltonian, introduced for perfect gases (or

[3]We chose the temperature here (as an example of tunable parameter) because in neural networks we will deal with white noise affecting the system. Analogously, in condensed matter, disorder is introduced by thermal noise, namely temperature. There is a deep similarity between them. In stochastic processes, prototype for white noise generators are random walkers, whose continuous limits are Gaussians, namely just the solutions of the Fourier equation for diffusion. However, the same celebrated equation holds for temperature spread too, indeed the latter is related to the amount of exchanged heat by the system under consideration, necessary for entropy's growth [52, 57]. Hence we have the first equivalence: white noise in neural networks mirrors thermal noise in structure of matter.

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 are from potentials. An ensemble of harmonic oscillators (mimicking atomic oscillations of the nuclei around their rest positions) was the first scenario for understanding condensed matter: however, as experimentally revealed by crystallography, nuclei are arranged according to regular lattices hence motivating mathematicians in study periodical structures to help physicists in this modeling, but merging statistical mechanics with lattice theories resulted soon in practically intractable models[4].

As a paradigmatic example, let us consider the one-dimensional Ising model, originally introduced to investigate magnetic properties of matter: the generic, out of N, nucleus labeled as i is schematically represented by a spin σ_i, which can assume only two values ($\sigma_i = -1$, spin down and $\sigma_i = +1$, spin up); nearest neighbor spins interact reciprocally through positive (i.e. ferromagnetic) interactions $J_{i,i+1} > 0$, hence the Hamiltonian of this system can be written as $H_N(\sigma) \propto -\sum_{1=1}^{N} J_{i,i+1}\sigma_i\sigma_{i+1} - h\sum_{1=1}^{N}\sigma_i$, where h tunes the external magnetic field and the minus sign in front of each term of the Hamiltonian ensures that spins try to align with the external field and to get parallel each other in order to fulfill the minimum energy principle.
Clearly, this model can trivially be extended to higher dimensions, however, due to prohibitive difficulties in facing the topological constraint of considering nearest neighbor interactions only, soon shortcuts were properly implemented to turn around this path. It is just due to an effective shortcut, namely the so called "mean field approximation", that statistical mechanics approached complex systems and, in particular, artificial intelligence.

1.2 The Role of Mean Field Limitations

As anticipated, the "mean field approximation" allows overcoming prohibitive technical difficulties owing to the underlying lattice structure. This consists in extending the sum on nearest neighbor couples (which are $\mathcal{O}(N)$) to include all possible couples in the system (which are $\mathcal{O}(N^2)$), properly rescaling the coupling ($J \to J/N$) in order to keep thermodynamical observable linearly extensive. If we consider a ferromagnet built of by N Ising spins

[4]For instance the famous Ising model [62], dated 1920 (and curiously invented by Lenz) whose properties are known in dimensions one and two, is still waiting for a solution in three dimensions.

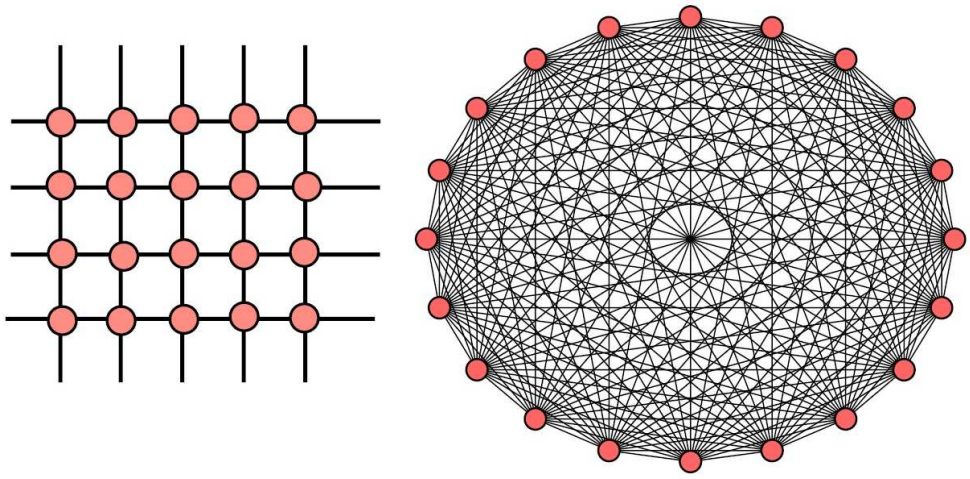

Figure 1.1: Example of regular lattice (left) and complete graph (right) with $N = 20$ nodes. In the former only nearest-neighbors are connected in such a way that the number of links scales linearly with N, while in the latter each node is connected with all the remaining $N - 1$ in such a way that the number of links scales quadratically with N.

$\sigma_i = \pm 1$ with $i \in (1, ..., N)$, we can then write

$$H_N(\sigma|J) = -\frac{1}{N}\sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j \sim -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sigma_i\sigma_j, \tag{1.1}$$

where in the last term we neglected the diagonal term ($i = j$) as it is irrelevant for large N. From a topological perspective the mean-field approximation equals to abandon the lattice structure in favor to a complete graph (see Fig.(1.2)). When the coupling matrix has only positive entries, e.g. $P(J_{ij}) = \delta(J_{ij} - J)$, this model is named Curie-Weiss model and acts as the simplest microscopic Hamiltonian able to describe the paramagnetic-ferromagnetic transitions experienced by materials when temperature is properly lowered. An external (magnetic) field h can be accounted for by adding in the Hamiltonian an extra term $\propto -h\sum_{i=1}^{N}\sigma_i$.

According to the principle of minimum energy, the two-body interaction appearing in the Hamiltonian in Eq.(1.1) tends to make spins parallel with each other and aligned with the external field if present. However, in the presence of noise (i.e. if temperature T is strictly positive), maximization of entropy must also be taken into account. When the noise level is much

higher than the average energy (roughly, if $T \gg J$), noise and entropy-driven disorder prevail and spins are not able to "feel" reciprocally; as a result, they flip randomly and the system behaves as a *paramagnet*. Conversely, if noise is not too loud, spins start to interact possibly giving rise to a phase transition; as a result the system globally rearranges its structure orientating all the spins in the same direction, which is the one selected by the external field if present, thus we have a *ferromagnet*.

In the early '70 a scission occurred in the statistical mechanics community: on the one side "pure physicists" saw mean-field approximation as a merely bound to bypass in order to have satisfactory pictures of the structure of matter and they succeeded in working out iterative procedures to embed statistical mechanics in (quasi)-three-dimensional reticula, yielding to the *renormalization group* developed by Kadanoff and Wilson in America [51] and Di-Castro and Jona-Lasinio in Europe [11]; this proliferative branch gave then rise to superconductivity, superfluidity [16] and many-body problems in condensed matter [48].
Conversely, from the other side, the mean-field approximation acted as a breach in the wall of complex systems: a thermodynamical investigation of phenomena occurring on general structures lacking Euclidean metrics (e.g. Erdös-Rényi graphs [8, 31], small-world graphs [19, 25], diluted, weighted graphs [33]) was then possible.

In general, as long as the summations run over all the indeces (hence mean-field is retained), rather complex coupling patterns can be solved (see e.g., the striking Parisi picture of mean-field glassy systems [59]) and this paved the strand to complex system analysis by statistical mechanics, whose investigation largely covers neural networks too.

1.3 Serial Processing

Hereafter we discuss how to approach neural networks from models mimicking ferromagnetic transition. In particular, we study the Curie-Weiss model and we show how it can store one pattern of information. Then, we notice that such a stored pattern has a very peculiar structure which is hardly *natural*, but we will overcome this (fake) flaw by introducing a gauge variant known as Mattis model. This scenario can be looked at as a primordial neural network. The successive step consists in extending, through elementary thoughts, this picture in order to include and store several patterns. In this way, we recover, via the first alternative route (w.r.t. the original one by Hebb), both the Hebb rule for synaptic plasticity and, as a corollary, the Hopfield model for neural networks too.

1.3.1 The Curie-Weiss Paradigm.

The statistical mechanical analysis of the Curie-Weiss model (CW) can be summarized as follows: Starting from a microscopic formulation of the system, i.e. N spins labeled as $i, j, ...$, their pairwise couplings $J_{ij} \equiv J$, and possibly an external field h, we derive an explicit expression for its (macroscopic) free energy $\alpha(\beta)$. The latter is the effective energy, namely the difference between the internal energy U, divided by the temperature $T = 1/\beta$, and the entropy S, namely $\alpha(\beta) = S - \beta U$, in fact, S is the "penalty" to be paid to the Second Principle for using U at noise level β. We can therefore link macroscopic free energy with microscopic dynamics via the fundamental relation

$$\alpha(\beta) = \lim_{N\to\infty} \frac{1}{N} \ln \sum_{\{\sigma\}}^{2^N} \exp\left[-\beta H_N(\sigma|J,h)\right], \tag{1.2}$$

where the sum is performed over the set $\{\sigma\}$ of all 2^N possible spin configurations, each weighted by the Boltzmann factor $\exp[-\beta H_N(\sigma|J,h)]$ that tests the likelihood of the related configuration. From expression (1.2), we can derive the whole thermodynamics and in particular phase-diagrams, that is, we are able to discern regions in the space of tunable parameters (e.g. temperature/noise level) where the system behaves as a paramagnet or as a ferromagnet.
Thermodynamical averages, denoted with the symbol $\langle . \rangle$, provide for a given observable the expected value, namely the value to be compared with measures in an experiment. For instance, for the magnetization $m(\sigma) \equiv \sum_{i=1}^{N} \sigma_i/N$ we have

$$\langle m(\beta) \rangle = \frac{\sum_{\{\sigma\}} m(\sigma) e^{-\beta H_N(\sigma|J)}}{\sum_{\{\sigma\}} e^{-\beta H_N(\sigma|J)}}. \tag{1.3}$$

When $\beta \to \infty$ the system is noiseless (zero temperature) hence spins feel reciprocally without errors and the system behaves ferromagnetically ($|\langle m \rangle| \to 1$), while when $\beta \to 0$ the system behaves completely random (infinite temperature), thus interactions can not be felt and the system is a paramagnet ($\langle m \rangle \to 0$). In between a phase transition happens.

In the Curie-Weiss model the magnetization works as *order parameter*: its thermodynamical average is zero when the system is in a paramagnetic (disordered) state ($\to \langle m \rangle = 0$), while it is different from zero in a ferromagnetic state (where it can be either positive or negative, depending on the sign of the external field). Dealing with order parameters allows us to avoid managing an extensive number of variables σ_i, which is practically impossible and, even more important, it is not strictly necessary.

Now, an explicit expression for the free energy in terms of $\langle m \rangle$ can be obtained carrying out summations in Eq.(1.2) and taking the *thermodynamic limit* $N \to \infty$ as

$$\alpha(\beta) = \ln 2 + \ln \cosh[\beta(J\langle m \rangle + h)] - \frac{\beta J}{2}\langle m \rangle^2. \tag{1.4}$$

In order to impose thermodynamical principles, i.e. energy minimization and entropy maximization, we need to find the extrema of this expression with respect to $\langle m \rangle$ requesting $\partial_{\langle m(\beta) \rangle}\alpha(\beta) = 0$. The resulting expression is called the *self-consistency* and it reads as

$$\partial_{\langle m \rangle}\alpha(\beta) = 0 \Rightarrow \langle m \rangle = \tanh[\beta(J\langle m \rangle + h)]. \tag{1.5}$$

This expression returns the average behavior of a spin in a magnetic field. In order to see that a phase transition between paramagnetic and ferromagnetic states actually exists, we can fix $h = 0$ (and pose $J = 1$ for simplicity) and expand the r.h.s. of Eq.(1.5) to get

$$\langle m \rangle \propto \pm\sqrt{\beta J - 1}. \tag{1.6}$$

Thus, while the noise level is higher than one ($\beta < \beta_c \equiv 1$ or $T > T_c \equiv 1$) the only solution is $\langle m \rangle = 0$, while, as far as the noise is lowered below its critical threshold β_c, two different-from-zero branches of solutions appear for the magnetization and the system becomes a ferromagnet (see Fig.(1.2)). The branch effectively chosen by the system usually depends on the sign of the external field or boundary fluctuations: $\langle m \rangle > 0$ for $h > 0$ and vice versa for $h < 0$.

Clearly, the lowest energy minima correspond to the two configurations with all spins aligned, either upwards ($\sigma_i = +1, \forall i$) or downwards ($\sigma_i = -1, \forall i$), these configurations being symmetric under spin-flip $\sigma_i \to -\sigma_i$. Therefore, the thermodynamics of the Curie-Weiss model is solved: energy minimization tends to align the spins (as the lowest energy states are the two ordered ones), however entropy maximization tends to randomize the spins (as the higher the entropy, the most disordered the states, with half spins up and half spins down): the interplay between the two principles is driven by the level of noise introduced in the system and this is in turn ruled by the tunable parameter $\beta \equiv 1/T$ as coded in the definition of free energy.

A crucial bridge between condensed matter and neural network could now be sighted: One could think at each spin as a basic neuron, retaining only its ability to spike such that $\sigma_i = +1$ and $\sigma_i = -1$ represent firing and quiescence, respectively, and associate to each equilibrium configuration of this spin system a *stored pattern* of information. The reward is that, in this

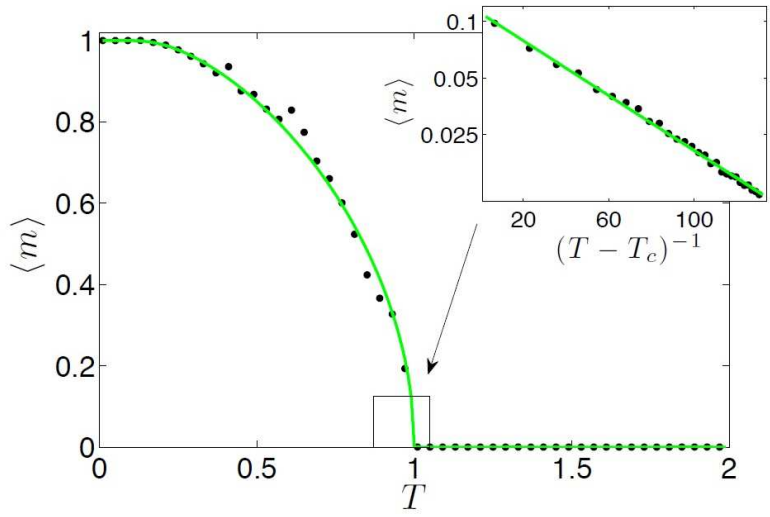

Figure 1.2: Average magnetization $\langle m \rangle$ versus temperature T for a Curie-Weiss model in the absence of field ($h = 0$). The critical temperature $T_c = 1$ separates a magnetized region ($|\langle m \rangle| > 0$, only one branch shown) from a non-magnetized region ($\langle m \rangle = 0$). The box zooms over the critical region (notice the logarithmic scale) and highlights the power-law behavior $m \sim (T - T_c)^{\beta}$, where $\beta = 1/2$ is also referred to as critical exponent (see also Eq.(1.6)). Data shown here (•) are obtained via Monte Carlo simulations for a system of $N = 10^5$ spins and compared with the theoretical curve (solid line).

way, the spontaneous (i.e. thermodynamical) tendency of the network to relax on free-energy minima can be related to the spontaneous retrieval of the stored pattern, such that the cognitive capability emerges as a natural consequence of physical principles.

1.3.2 From Curie-Weiss to Hopfield

Actually, the Hamiltonian (1.1) would encode for a rather poor model of neural network as it would account for only two stored patterns, corresponding to the two possible minima (that in turn would represent pathological network's behavior with all the neurons contemporarily completely firing of completely silenced), moreover, these ordered patterns, seen as information chains, have the lowest possible entropy and, for the Shannon-McMillan Theorem, in the

large N limit[5] they will never be observed.

This criticism can be easily overcome thanks to the Mattis-gauge, namely a re-definition of the spins via $\sigma_i \to \xi_i^1 \sigma_i$, where $\xi_i^1 = \pm 1$ are random entries extracted with equal probability:

$$P(\xi_i^\mu) = \frac{1}{2}\delta_{\xi_i^\mu - 1} + \frac{1}{2}\delta_{\xi_i^\mu + 1}, \tag{1.7}$$

and kept fixed in the network (in statistical mechanics these are called *quenched* variables to stress that they do not contribute to thermalization, a terminology reminiscent of metallurgy [59]). Fixing $J \equiv 1$ for simplicity, the Mattis Hamiltonian reads as

$$H_N^{Mattis}(\sigma|\xi) = -\frac{1}{2N}\sum_{i,j=1}^{N,N} \xi_i^1 \xi_j^1 \sigma_i \sigma_j - h \sum_{i=1}^{N} \xi_i^1 \sigma_i. \tag{1.8}$$

The Mattis magnetization is defined as $m_1 = \frac{1}{N}\sum_{i=1}^{N} \xi_i^1 \sigma_i$. To inspect its lowest energy minima, we perform a comparison with the CW model: in terms of the (standard) magnetization, the Curie-Weiss model reads as $H_N^{CW} \sim -(N/2)m^2 - Nhm$ and, analogously we can write $H_N^{Mattis}(\sigma|\xi)$ in terms of Mattis magnetization as $H_N^{Mattis} \sim -(N/2)m_1^2 - Nhm_1$. It is then evident that, in the low noise limit (namely where collective properties may emerge), as the minimum of free energy is achieved in the Curie-Weiss model for $\langle m \rangle \to \pm 1$, the same holds in the Mattis model for $\langle m_1 \rangle \to \pm 1$. However, this implies that now spins tend to align parallel (or antiparallel) to the vector ξ^1, hence if the latter is, say, $\xi^1 = (+1, -1, -1, -1, +1, +1)$ in a model with $N = 6$, the equilibrium configurations of the network will be $\sigma = (+1, -1, -1, -1, +1, +1)$ and $\sigma = (-1, +1, +1, +1, -1, -1)$, the latter due to the gauge symmetry $\sigma_i \to -\sigma_i$ enjoyed by the Hamiltonian. Thus, the network relaxes autonomously to a state where some of its neurons are firing while others are quiescent, according to the *stored pattern* ξ^1. Note that, as the entries of the vectors ξ are chosen randomly ± 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 now. On the one side, according to the self-consistency equation (1.5) $\langle m \rangle$ versus h displays the typical graded/sigmoidal

[5] The *thermodynamic limit* $N \to \infty$ is required for both mathematical convenience, e.g. it allows saddle-point/stationary-phase techniques, and in order to neglect observable fluctuations by a central limit theorem argument.

response of a charging neuron [45], and one would be tempted to call the spins σ neurons. On the other side, 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 (and provides the first derivation of the Hebbian prescription in this work): If we want a network able to cope with P patterns, the starting Hamiltonian should have simply the sum over these P previously stored[6] patterns, namely

$$H_N(\sigma|\xi) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\left(\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\right)\sigma_i\sigma_j, \tag{1.9}$$

where we neglect the external field ($h = 0$) for simplicity. As we will see in the next section, this Hamiltonian constitutes indeed the Hopfield model, namely the harmonic oscillator of neural networks, whose coupling matrix is called *Hebb matrix* as encodes the Hebb prescription for neural organization [17].

1.3.3 From Sherrington-Kirkpatrick 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 systems" realm. We propose that complex behaviors can be distinguished by simple behaviors as for the latter the number of free-energy minima of the system *does not scale* with the volume N, while for complex systems the number of free-energy minima *does scale* with the volume according to a proper function of N. For instance, the Curie-Weiss/Mattis model has two minima only, whatever N (even if $N \to \infty$), and it constitutes the paradigmatic example for a simple system. As a counterpart, the prototype of complex system is the Sherrington-Kirkpatrick model (SK), originally introduced in condensed matter to describe the peculiar behaviors exhibited by real glasses [6, 59]. This model has an amount of minima that scales $\propto \exp(cN)$ with $c \neq f(N)$, and its Hamiltonian reads as

$$H_N^{SK}(\sigma|J) = \frac{1}{\sqrt{N}}\sum_{i<j}^{N,N}J_{ij}\sigma_i\sigma_j, \tag{1.10}$$

[6] The part of neural network's theory we are analyzing is meant for spontaneous retrieval of already stored information -grouped into patterns (pragmatically vectors)-. Clearly it is assumed that the network has already overpass the learning stage.

where, crucially, coupling are Gaussian distributed[7] as $P(J_{ij}) \equiv \mathcal{N}[0,1]$. This implies that links can be either positive (hence favoring parallel spin configuration) as well as negative (hence favoring anti-parallel spin configuration), thus, in the large N limit, with large probability, spins will receive conflicting signals and 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. This mirrors a clear request also in electronics, namely the need for inverters (that once mixed with op-amps) result in flip-flops (crucial for information storage as we will see).

The mean-field statistical mechanics for the low-noise behavior of spin-glasses has been first described by Giorgio Parisi and it predicts a hierarchical organization of states and a relaxational dynamics spread over many timescales (for which we refer to specific textbooks [59]). Here we just need to know that their natural order parameter is no longer the magnetization (as these systems do not magnetize), but the *overlap* q_{ab}, as we are explaining. Spin glasses are balanced ensembles of ferromagnets and antiferromagnets (this can also be seen mathematically as $P(J)$ is symmetric around zero) and, as a result, $\langle m \rangle$ 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 because at large temperatures it is expected to be zero, as everything is uncorrelated, but at low temperature their overlap is strictly non-zero as spins freeze in disordered but correlated states. More precisely, given two "replicas" of the system, labeled as a and b, their overlap q_{ab} can be defined as the scalar product between the related spin configurations, namely as $q_{ab} = (1/N) \sum_i^N \sigma_i^a \sigma_i^b$[8], thus the mean-field spin glass has a completely random paramagnetic phase, with $\langle q \rangle \equiv 0$ and a "glassy phase" with $\langle q \rangle > 0$ split by a phase transition at $\beta_c = T_c = 1$.

The Sherrington-Kirkpatrick model displays a large number of minima as expected for a cognitive system, yet it is not suitable to act as a cognitive system because its states are too "disordered". We look for an Hamiltonian whose minima are not purely random like those in SK, as they must represent ordered stored patterns (hence like the CW ones), but the amount of these minima must be possibly extensive in the number of spins N (as in the SK and at contrary with CW), hence we need to retain a "ferromagnetic flavor" within a "glassy panorama": we need *something in between*.

[7] Couplings in spin-glasses are drawn once for all at the beginning and do not evolve with system's thermalization, namely they are *quenched* variables too.

[8] Note that, while in the Curie-Weiss model, where $P(J) = \delta(J-1)$, the order parameter was the first momentum of $P(m)$, in the Sherrington-Kirkpatrick model, where $P(J) = \mathcal{N}[0,1]$, the variance of $P(m)$ (which is roughly q_{ab}) is the good order parameter.

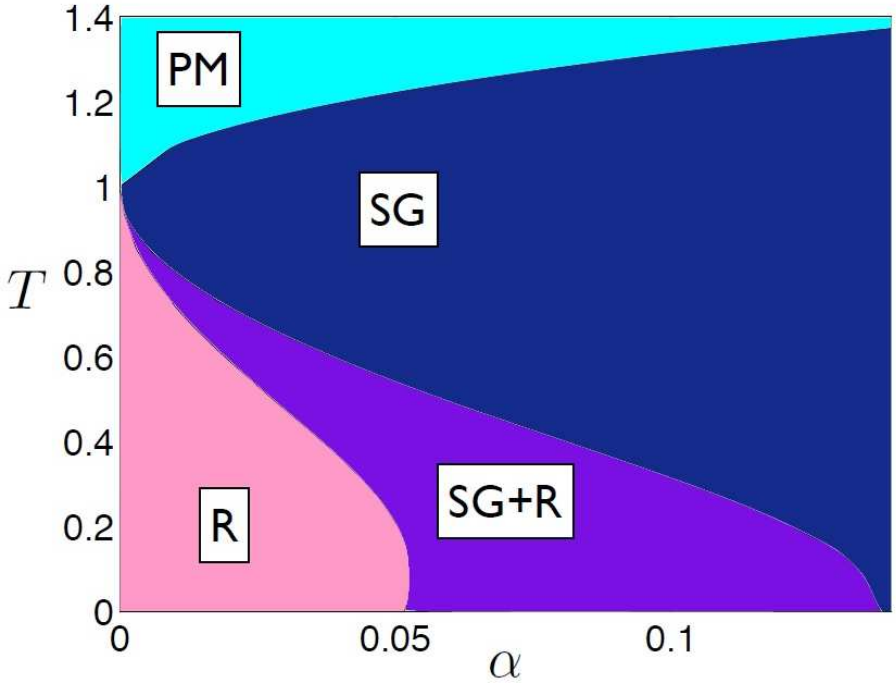

Figure 1.3: Phase diagram for the Hopfield model [17]. According to the parameter setting, the system behaves as a paramagnet (PM), as a spin-glass (SG), or as an associative neural network able to perform information retrieval (R). The region labeled (SG+R) is a coexistence region where the system is glassy but still able to retrieve.

Remarkably, the Hopfield model defined by the Hamiltonian (1.9) lies exactly in between a Curie-Weiss model and a Sherrington-Kirkpatrick model. Let us see why: When $P = 1$ the Hopfield model recovers the Mattis model, which is nothing but a gauge-transformed Curie-Weiss model. Conversely, when $P \to \infty$, $(1/\sqrt{N}) \sum_{\mu}^{P} \xi_i^{\mu} \xi_j^{\mu} \to \mathcal{N}[0, 1]$, by the standard central limit theorem, and the Hopfield model recovers the Sherrington-Kirkpatrick one. In between these two limits the system behaves as an associative network [4]. Such a crossover between CW (or Mattis) and SK models, requires for its investigation both the P Mattis magnetization $\langle m_{\mu} \rangle$, $\mu = (1, ..., P)$ (for quantifying retrieval of the whole stored patterns, that is the *vocabulary*), and the two-replica overlaps $\langle q_{ab} \rangle$ (to control the glassyness growth if the vocabulary gets enlarged), as well as a tunable parameter measuring the ratio between the stored patterns and the amount of available spins, namely $\alpha = \lim_{N \to \infty} P/N$, also referred to as *network capacity*.

As far as P scales sub-linearly with N, i.e. in the low storage regime

defined by $\alpha = 0$, the phase diagram is ruled by the noise level β only: for $\beta < \beta_c$ the system is a paramagnet, with $\langle m_\mu \rangle = 0$ and $\langle q_{ab} \rangle = 0$, while for $\beta > \beta_c$ the system performs as an attractor network, with $\langle m_\mu \rangle \neq 0$ for a given μ (selected by the external field) and $\langle q_{ab} \rangle = 0$. In this regime no dangerous glassy phase is lurking, yet the model is able to store only a tiny amount of patterns as the capacity is sub-linear with the network volume N. Conversely, when P scales linearly with N, i.e. in the high-storage regime defined by $\alpha > 0$, the phase diagram lives in the α, β plane (see Fig.(1.3)). When α is small enough the system is expected to behave similarly to $\alpha = 0$ hence as an associative network (with a particular Mattis magnetization positive but with also the two-replica overlap slightly positive as the glassy nature is intrinsic for $\alpha > 0$). For α large enough ($\alpha > \alpha_c(\beta), \alpha_c(\beta \to \infty) \sim 0.14$) however, the Hopfield model collapses on the Sherrington-Kirkpatrick model as expected, hence with the Mattis magnetizations brutally reduced to zero and the two-replica overlap close to one. The transition to the spin-glass phase is often called "blackout scenario" in neural network community. Making these predictions quantitative is a non-trivial task in statistical mechanics and, nowadays several techniques are available, among which we quote the replica-trick (originally used by the pioneers Amit-Gutfreund-Sompolinsky [18]), the martingale method (originally developed by Pastur, Sherbina and Tirozzi [53]) and the cavity field technique (recently developed by Guerra and some of us in [2]).

Chapter 2

Dilution in the Hebb Rules

The paradigm, introduced almost three decades ago by Amit, Gutfreund and Sompolinsky [17, 18], of analyzing neural networks through techniques stemmed from statistical mechanics of disordered systems (in particular the Replica Trick [59] for the Hopfield model [49]) has been so prolific that its applications have gone far beyond Artificial Intelligence and Robotics, overlapping Statistical Inference [9], System Biology [66], Financial Market planning [64], Theoretical Immunology [32] and much more.
As a result, research in this field is under continuous development, ranging from the diverse applications outlined above, to a deeper and deeper understanding of the core-theory behind. For the sake of reaching results closer to experimental neuroscience outcomes, scientists involved in the field tried to bypass the rather crude mean field description of a fully connected network of interacting spins, embedding them in diluted topologies as Erdös-Rényi graphs [46], small-worlds [67] or even finitely connected graphs [10]. The main point was showing robustness of the mean-field paradigm even in these diluted, and in some sense "closer to biology", versions and this was indeed successfully achieved (with the exception of too extreme degrees of dilution, where the associative capacities of the network trivially break down).

Recently, a mapping between Hopfield networks and Boltzmann machines [1] allowed the introduction of dilution into associative networks from a different perspective with respect to standard link removal à la Sompolinsky [46] or à la Coolen [10, 67]. In fact, while in their papers these authors perform dilution directly on the Hopfield network, through the equivalence with Boltzmann machine, one may perform link dilution on the Boltzmann machine and then map back the latter into the associative Hopfield-like network [30]. Remarkably, the resulting model still works as an associative performer, as the Hebbian structure is preserved, but its capabilities are quite different from the standard scenario. In particular, the resulting associative

network may still be fully-connected but the stored patterns of information display entries which, beyond coding information through digital values ± 1, can also be blank [27, 30]. In fact, any missing link in the bipartite Boltzmann machine corresponds to a blank entry in the related pattern of the associative network.
Now, while standard (i.e., performed directly on the Hopfield network) dilution does not change qualitatively the system performances, the behavior of the system resulting from hidden (i.e., performed on the underlying Boltzmann machine) dilution becomes "multitasking" because retrieval of a single pattern, say ξ^1, does not exhaust the whole spins, and the ones coupled with the blank entries of ξ^1 are free to align with ξ^2, whose entries will partially be blank as well, hence eliciting, in turn, the retrieval of ξ^3 and so on up to a parallel logarithmic (with respect to the volume of the network N) load of all the stored patterns.
As a consequence, by tuning the degree of dilution in the hidden Boltzmann network and the level of noise in the directed network, the system exhibits a very rich phase diagram, whose investigation is the subject of the present chapter.

Let us now move on and generalize the system described above in order to account for the existence of blank entries in the patterns ξ's. More precisely, we replace Eq.(1.7) by

$$P(\xi_i^\mu) = \frac{1-d}{2}\delta_{\xi_i^\mu -1} + \frac{1-d}{2}\delta_{\xi_i^\mu +1} + d\delta_{\xi_i^\mu}, \tag{2.1}$$

where d encodes the degree of "dilution" in pattern entries. Patterns are still assumed as quenched and, of course, the definitions of the Hamiltonian (1.9) and of the overlaps (1.3), with the Glauber dynamics provided by:

$$\sigma_i(t+\delta t) = \text{sign}[\tanh[\beta h_i(t)] + \eta_i(t)],$$

(where $\eta \in [-1, +1]$ is a random number and represent the stochasticity and h_i is the field acting on the i-th spin) still hold.

As discussed in [27, 30], this kind of extension has strong biological motivations and also yields highly non-trivial thermodynamic outcomes. In fact, the distribution in Eq.(1.7) necessarily implies that the retrieval of a unique pattern does employ all the available spins, so that no resources are left for further tasks. Conversely, with Eq.(2.1) the retrieval of one pattern still allows available spins (i.e., those corresponding to the blank entries of the retrieved pattern), which can be used to recall other patterns up to the exhaustion of all spins. The resulting network is therefore able to process several patterns simultaneously.

In particular, in the low-storage regime, it was shown both analytically (via density of states analysis) and numerically (via Monte Carlo simulations) [30], that the system evolves toward an equilibrium state where several patterns are simultaneously retrieved. In the noiseless limit $T = 0$ and for d not too large, the equilibrium state is characterized by a hierarchical overlap

$$\boldsymbol{m} = (1 - d)(1, d, d^2, ..., 0), \tag{2.2}$$

hereafter referred to as "parallel ansatz". On the other hand, in the presence of noise or for large degrees of dilution in pattern entries, this state ceases to be a stable solution for the system and different states, possibly spurious, emerge. In the following highlight the equilibrium states of this system as a function of the parameters d and T, and finally build a phase diagram; to this task we first develop a rigorous mathematical treatment for calculating the free energy of the model and then we obtain the self-consistencies constraining the phase-diagram; finally, we solve these equations both numerically and with a stability analysis. In this way we are able to draw the phase diagram, whose peculiarities lie in the stability of both even and odd mixture of spurious states (in proper regions of the parameters) and the formation of parallel spurious state. Both these results generalize the standard counterpart of classical Hopfield networks.
Findings are double-checked through Monte Carlo runs that are in agreement with the picture we obtained.

2.1 Notes About the Coupling Distribution

As it is immediate to check, each $\xi_i^\mu = 0$ in the i^{th} entry of the bit-string ξ^μ in the associative network, which ultimately affects the interaction matrix $\mathbf{J} = J_{ij}$. Of course, the larger the degree of dilution, the stronger the difference between such (random) coupling matrix and its Hopfield counterpart. This section is devoted to the investigation of the properties of the matrix $\mathbf{J}$.

Let us consider a set of N nodes labeled as $i = 1, ..., N$ and let us associate to each node a string of length P and built from the alphabet $\{-1, 0, 1\}$, meaning that the generic element ξ_i^μ, with $i \in [1, N]$ and $\mu \in [1, P]$, can equal either ± 1 or 0. For the network described by the Hamiltonian in Eq.(1.9), the interaction strength between two arbitrary nodes i and j is given by

$$J_{ij} = \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu. \tag{2.3}$$

For the following treatment it is more convenient not to normalize the coupling Jij, differently from the definition used in Eq.(1.9). Of course $J_{ij} \in$

$[-P, P]$. Equation (2.3) gives rise to a network of mutually and symmetrically interacting nodes, where a link between nodes i and j is drawn whenever they do interact directly ($J_{ij} \neq 0$), either imitatively ($J_{ij} > 0$) or anti-imitatively ($J_{ij} < 0$).

First, one can calculate the probability that two nodes (since they are arbitrary we will drop the indexes) in the network are linked together, namely

$$P_{\rm link}(d,P) = P(J \neq 0; d, P) = 1 - P(J = 0; d, P) = 1 - \sum_{k=0}^{P} P_{\rm sum-0}(k; d, P), \tag{2.4}$$

where $P_{\rm sum-0}(k; d, P)$ is the probability that two strings display (an even number) k of non-null matchings summing up to zero; otherwise stated, there exist exactly k values of μ such that $\xi_i^\mu \xi_j^\mu \neq 0$ and they are half positive and half negative. In particular, $P_{\rm sum-0}(0; d, P) = [d(2-d)]^P$, because this is the probability that, for any $\mu \in [1, P]$, at least one entry (either ξ_i^μ or ξ_j^μ or both) is equal to zero. More generally,

$$P_{\rm sum-0}(k; d, P) = \left(\frac{1-d}{2}\right)^{2k} [d(2-d)]^{P-k} \binom{P}{k} \left[2^k \binom{k}{k/2}\right], \tag{2.5}$$

where the first and the second factors in the r.h.s. require that k entries are non-zero and the remaining $P-k$ entries are zero; the third factor accounts for permutation between zero and non-zero entries, while the last term is the number of configurations leading to a null sum for non-null entries. Therefore, we have

$$P(J = 0; d, P) = [d(2-d)]^P \sum_{k=0}^{P} \left[\frac{(1-d)^2}{2d(2-d)}\right]^k \binom{P}{k} \binom{k}{k/2}, \tag{2.6}$$

whose plot is shown in Fig.(2.1). As for its asymptotic behavior, we distinguish the following cases (for simplicity we assume P finite and even):

$$P(J = 0; d, P) = 1 - P(1-d)^2 + \frac{3}{4} P(P-1)(1-d)^4 + \mathcal{O}(1-d)^6 \tag{2.7}$$

$$P(J = 0; d, P) = \frac{(-1)^{P/2}\sqrt{\pi}}{\Gamma(1/2 - P)\Gamma(1 + P/2)} (1 - 2P\, d) + \mathcal{O}(d^2)$$

$$\approx \frac{1 - 2P\, d}{4^{P/2}} \binom{P}{P/2} + \mathcal{O}(d^2). \tag{2.8}$$

The average number of nearest neighbors per node $\langle z \rangle_{d,P,H}$ follows immediately as $\langle z \rangle_{d,P,N} = N P_{\rm link}(d, P)$.

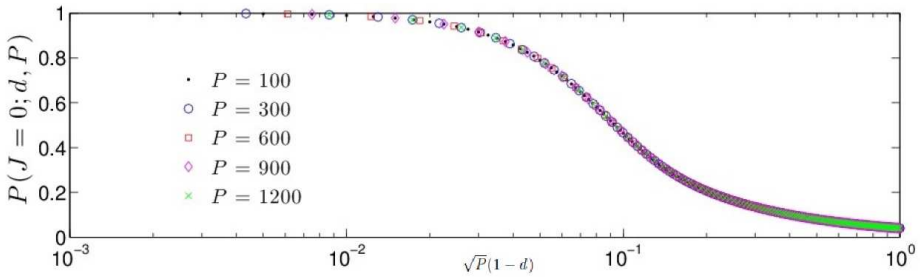

Figure 2.1: The probability $P(J=0;d;P)$ is plotted as a function of the dilution d and for different values of P, as shown by the legend. Notice the semilogarithmic scale and that dilution is rescaled by $\sqrt{p}$ so to highlight the common scaling of the distributions.

More generally, we can derive the coupling distribution $P(J;d,P)$, once having defined $P_{+1}(k)$, $P_{-1}(k)$ and $P_0(k)$, as the probability that, given two strings, they display k matches each equal to $+1$, -1 and 0, respectively, namely

$$P_{+1}(k;d) = P_{-1}(k;d) = \left[\frac{(1-d)^2}{2}\right]^k, \quad P_0(k;d) = [d(2-d)]^k. \tag{2.9}$$

Hence, we can write

$$\begin{aligned} P(J;d,P) &= \sum_{l=0}^{(P-J)/2} \frac{P_{+1}(l+J;d)P_{-1}(l;d)P_0(P-2l-J;d)P!}{l!(l+J)!(P-2l-J)!} \\ &\sim \mathcal{N}(0,\sigma_J(d,P)). \end{aligned} \tag{2.10}$$

The last asymptotic holds for large P; the null mean value $\langle J\rangle_{d,P} = 0$ is due to the symmetry characterizing $P(\xi_i^\mu)$, while the standard deviation is $\sigma_J = \sqrt{\langle J^2\rangle_{d,P}} = \sqrt{P}(1-d)$.

An explicit, exact expression for this probability can be written for a particular value of d, by exploiting Gauss's Hypergeometric Theorem [69], so that when $4x^2 = 1$, corresponding to $d = 1-\sqrt{2}/2 \approx 0.293$, we have

$$P(J;1-\sqrt{2}/2,P) = 4^{-P}\binom{2P}{P+J} \sim \frac{e^{-J^2/P}}{\sqrt{\pi P}}. \tag{2.11}$$

In the last passage we used the Stirling approximation assuming $P \pm J$ large, namely that the distribution is peaked on non-extreme values of J.

It is worth underlining that $P(J;d,P)$ does not depend on the size N. Indeed, patterns are drawn independently and randomly so that the coupling

J_{ij} may be regarded as the distance covered by a random walk of length B and endowed with a waiting probability $d(2-d)$. Hence, the end-to-end distance is distributed normally around zero and with variance (mean squared distance) which is given by the diffusion law, namely $\sim P$. The possibility of the walker to stop simply reduces the effective walk length to $[(1-d)(2-d)]P = (1-d)^2 P$ in agreement with results above.

2.1.1 Pattern dilution versus Topological dilution

Dilution on pattern entries does not necessarily yield to a topological dilution for the associative network, but, as we will see, can induce non-trivial cooperative effects. On the other hand, a topological dilution can be realized by directly cutting the edges on a standard Hopfield network. In this section we highlight the deep difference between these two kinds of dilution.

First, we recall that, according to a mean-field approach, the network is expected to display a giant component when the average link probability is larger than $1/N$. In the thermodynamic limit and assuming a large enough size P (stemming from either low, i.e. $P \sim \log N$, or high, i.e. $P \sim N$, storage regimes) to ensure the result in Eq.(2.10) to hold, for any finite value of d the emergent graph turns out to be always over-percolated. In fact, $P_{\text{link}}(d,P) = 1 - P(J=0;d,P) \sim 1 - 1/\sqrt{2\pi\sigma_J^2}$, so that it suffices that $\sigma_J > N/[\sqrt{2\pi}(N-1)] \to 1/\sqrt{2\pi}$ and this leads to $d < 1 - (2\pi P)^{-1/2} \to 1$.

On the other hand, when P is finite we can check the possible disconnection of the network by studying $P(J=0;d,P)$ from Eq.(2.7) and we get that $P_{\text{link}}(d,P) < 1/N$ for $d > 1 - 1/\sqrt{PN}$. Thus, in the thermodynamic limit, for any finite d, the graph is still overpercolated. Replacing $1/N$ with $(\log N)/N$, one also finds that the graph is even always connected.

Different scenarios may emerge if we take d properly approaching to 1 as N is increased [25].

Another kind of dilution can be realized by directly cutting edges in the resulting associative network, as for instance early investigated in the neural scenario by Sompolinsky on the Erdös-Rényi graph [17, 46] or more recently by Coolen and coworkers on small worlds and scale-free structures [10, 47].

Such different ways of performing dilution - either on links of the associative network (see [10, 17, 46, 47]) or on pattern entries (see Eq.(1.7)) - yield deeply different thermodynamic behaviors. To see this, let us consider the field insisting on each spin, namely for the generic i^{th} spin $h_i = \frac{1}{N}\sum_{i\neq j=1}^{N} J_{ij}\sigma_j$, and analyze its distribution $P(h|d)$ at zero noise level. When dilution is realized on links (d is the fraction of links cut), only an average fraction d of the H available spins participates to h, in such a way that both the peak and the span of the distribution decrease with d (Fig.(2.2), left).

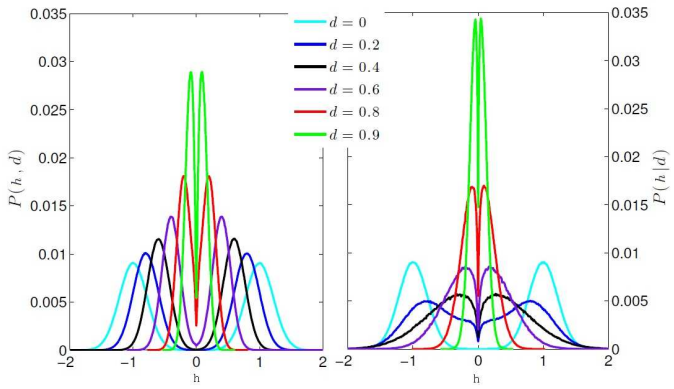

Figure 2.2: Left panel: Distribution of the field h acting on the spins with (Sompolinsky) dilution. Right panel: Distribution of the field h acting on the spins with (our) dilution.

Conversely, when dilution is realized on the single bit ξ_i^μ (d is the fraction of null entries in a pattern), as $d > 0$, $P(h|d)$ gets broader and peaked at smaller values of fields.

The latter effect is due to the fact that couplings are, on average, of smaller magnitude. As for the former effect, we notice that, at β, N and P fixed, when dilution is introduced in bit-strings, couplings are made *uniformly* weaker (this effect is analogous to a rise in the fast noise) so that the distribution of spin configurations, and consequently also $P(h|d)$, gets broader. At small values of dilution this effect dominates, while at larger values the overall reduction of coupling strengths prevails and fields get not only smaller but also more peaked (Fig.(2.2), right).

2.2 Statistical Mechanics Analysis

We now solve the general model described by the Hamiltonian (1.9), with patterns diluted according to (2.1), in the low storage regime $P \sim \log N$, such that the limit $\alpha = \lim_{N\to\infty} P/N = 0$ holds[1]

As standard in disordered statistical mechanics, we introduce three types of average for an observable $o(\sigma, \xi)$:

[1]Results outlined within this scaling can be extended with little effort to the whole region $P \sim N^\gamma$, with $\gamma < 1$, such that the constraint $\alpha = 0$ is preserved, as realized in the Willshaw model [20] concerning neural sparse coding.
Note further that there is a deep similarity with the Potts model with pairwise interaction [41].

i. the Boltzmann average $\omega(o) = \sum_{\{\sigma\}} o(\sigma,\xi)\exp[-\beta\mathcal{H}(\sigma;\xi)]/Z_{N,P}(\beta,d)$, where

$$Z_{N,P}(\beta,d) = \sum_{\{\sigma\}} \exp\left[-\beta H_N(\sigma,\xi)\right]$$

is called "partition function",
ii. the average $\mathbb{E}$ performed over the quenched disordered couplings ξ,
iii. the global expectation $\mathbb{E}\omega(o)$ defined by the brackets $\langle o\rangle_\xi$.
Given these definitions, for the average energy of the system E we can write $E \equiv \lim_{N\to\infty}(\langle H_N(\sigma,\xi)\rangle/N)$.
Also, we are interested in finding an explicit expression for the order parameters of the model, namely the averaged P Mattis magnetizations

$$\langle m^\mu\rangle = \lim_{N\to\infty} \mathbb{E}\omega\left(\frac{1}{N}\sum_{j=1}^{N}\xi_j^\mu\sigma_j\right). \tag{2.12}$$

To this task we need to introduce the statistical pressure

$$\alpha(\beta,d) = \lim_{N\to\infty}\frac{1}{N}\ln(Z_{N,P}(\beta,d)),$$

which is immediately related to the free energy per site $f(\beta,d)$ by the relation $f(\beta,d) = -\alpha(\beta,d)/\beta$ because, by maximizing $\alpha(\beta,d)$ with respect to the P magnetizations $\langle m^\mu\rangle$, we get exactly the self consistence equations for these order parameters, whose solutions will give us a picture of the phase diagram.

In the past decades, scientists involved in disordered statistical mechanics investigations, even beyond Artificial Intelligence, paved several strands for solving this kind of problems, and nowadays a plethora of techniques is available. We extend early ideas of Guerra, on the line developed in [43], consisting in modeling disordered statistical mechanics through dynamical system theory and in particular, here, we are going to proceed as follows:
Our statistical-mechanics problem is mapped into a diffusive problem embedded in a P-dimensional space and with given, known, boundaries. We solve the diffusive problem via standard Green-propagator technique, and then we will map back the obtained solutions in terms of their original statistical mechanics meaning.
To this task, let us introduce and consider a generalized Boltzmann factor $B_N(\mathbf{x},t)$ depending on $P+1$ parameters $\mathbf{x},t$ (which we think of as *generalized P-dimensional Euclidean space* and *time*)

$$B_N(\mathbf{x},t;\xi,\sigma) = \exp\left(\frac{t}{2N}\sum_{i\neq j}^{N}\sigma_i\sigma_j\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu + \sum_{\mu=1}^{P}x_\mu\sum_{j=1}^{N}\xi_j^\mu\sigma_j\right), \tag{2.13}$$

and the generalized statistical pressure

$$\alpha_N(\mathbf{x},t) = \frac{1}{N}\ln\left[\sum_{\{\sigma\}} B_N(\mathbf{x},t;\xi,\sigma)\right]. \tag{2.14}$$

Notice that, for proper values of $\mathbf{x},t$, namely $\mathbf{x}=0$ and $t=\beta$, classical statistical mechanics is recovered as

$$\alpha(\beta,d) = \lim_{N\to\infty}\alpha_N(\mathbf{x}=0,t=\beta) = \lim_{N\to\infty}\frac{1}{N}\ln\left[\sum_{\{\sigma\}} B_N(\mathbf{x}=0,t=\beta;\xi,\sigma)\right].$$

In the same way, the average $\langle\cdot\rangle_{(\mathbf{x},t)}$ will be denoted by $\langle\cdot\rangle$, wherever evaluated in the sense of statistical mechanics, namely

$$\langle o\rangle_{(\mathbf{x},t)} = \frac{\sum_{\{\sigma\}} o(\sigma,\xi)B_N(\mathbf{x},t;\xi,\sigma)}{\sum_{\{\sigma\}} B_N(\mathbf{x},t;\xi,\sigma)}, \tag{2.15}$$

$$\langle o\rangle = \frac{\sum_{\{\sigma\}} o(\sigma,\xi)\exp[-\beta H(\sigma,\xi)]}{\sum_{\{\sigma\}}\exp[-\beta H(\sigma,\xi)]} = \langle o\rangle_{(\mathbf{x}=0,t=\beta)}. \tag{2.16}$$

It is immediate to see that the following equations hold:

$$\begin{aligned}\partial_t\alpha_N(\mathbf{x},t) &= \tfrac{1}{2}\textstyle\sum_\mu\langle m_\mu^2\rangle_{(\mathbf{x},t)},\\ \partial_{x_\mu}\alpha_N(\mathbf{x},t) &= \langle m_\mu\rangle_{(\mathbf{x},t)},\end{aligned} \tag{2.17}$$

and, defining a vector $\Gamma_N(\mathbf{x},t)$ of elements $\Gamma_N^\mu(\mathbf{x},t)\equiv-\partial_{x_\mu}\alpha_N(\mathbf{x},t)$, by construction $\Gamma_N^\mu(\mathbf{x},t)$ obeys the following equation:

$$\partial_t\Gamma_N^\mu(\mathbf{x},t) + \sum_{\nu=1}^{P}\Gamma_N^\nu(\mathbf{x},t)[\partial_{x_\nu}\Gamma_N^\mu(\mathbf{x},t)] = \frac{1}{2N}\sum_{\nu=1}^{P}\partial_{x_\nu^2}^2\Gamma_N^\mu(\mathbf{x},t), \tag{2.18}$$

which happens to be in the form of a Burgers' equation for the vector $\Gamma_N(\mathbf{x},t)$ with a kinematic viscosity $(2N)^{-1}$. As it is well-known, the Burger equation can be mapped into a P-dimensional diffusive problem using the Cole-Hopf transformation [43] as follow:

$$\psi_N(\mathbf{x},t) = \exp\left[-N\int dx_\mu\Gamma_N^\mu(\mathbf{x},t)\right] = \exp[N\alpha_N(\mathbf{x},t)], \tag{2.19}$$

and its t and x streaming read off as

$$\begin{aligned}\partial_t\psi_N(\mathbf{x},t) &= N(\partial_t\alpha_N(\mathbf{x},t))\psi(\mathbf{x},t),\\ \partial_{x_\mu}\psi_N(\mathbf{x},t) &= N(\partial_{x_\mu}\alpha_N(\mathbf{x},t))\psi(\mathbf{x},t),\end{aligned} \tag{2.20}$$

in such a way that

$$\partial^2_{x_\mu x_\nu}\psi_N(\mathbf{x},t) = N\psi_N(\mathbf{x},t)\left\{\partial^2_{x_\mu x_\nu}\alpha_N(\mathbf{x},t) + N[\partial_{x_\mu}\alpha_N(\mathbf{x},t)][\partial_{x_\nu}\alpha_N(\mathbf{x},t)]\right\}. \tag{2.21}$$

Now, from equations (2.20), (2.21) we get

$$\partial_t\psi_N(\mathbf{x},t) - \frac{1}{2N}\sum_\mu\left[\partial^2_{x_\mu^2}\psi_N(\mathbf{x},t)\right] = 0. \tag{2.22}$$

Therefore, we established a reformulation of the problem of calculating the thermodynamic potential $\alpha(\beta,d)$ over the equilibrium configuration of the order parameters for an attractors network model in terms of a diffusion equation for the function $\psi_N(\mathbf{x},t)$, namely the Cole-Hopf transform of the Mattis magnetizations, with a diffusion coefficient $D=(2N)^{-1}$, that is

$$\begin{aligned}\partial_t\psi_N(\mathbf{x},t) - D\nabla^2\psi_N(\mathbf{x},t) = 0,\\ \psi_N(\mathbf{x},0) = \sum_{\{\sigma\}}\exp\left(\sum_\mu x_\mu\sum_j\xi_j^\mu\sigma_j\right).\end{aligned} \tag{2.23}$$

We solve this Cauchy problem (2.23) through standard techniques: first, we map the diffusive equation in the Fourier space, then we calculate the Green propagator for the homogenous configuration, and finally we will inverse-transform the solution.
Let us consider the Fourier transform:

$$\begin{aligned}\widetilde{\psi}_N(\mathbf{k},t) &= \textstyle\int_{\mathbb{R}^P} d^Px\exp\big(-i\sum_\mu x_\mu k_\mu\big)\psi_N(\mathbf{x},t),\\ \psi_N(\mathbf{x},t) &= \textstyle\frac{1}{(2\pi)^P}\int_{\mathbb{R}^P} d^Pk\exp\big(i\sum_\mu x_\mu k_\mu\big)\widetilde{\psi}_N(\mathbf{k},t),\end{aligned} \tag{2.24}$$

and the related Green problem:

$$\partial_t\widetilde{G}(\mathbf{k},t) + Dk^2\widetilde{G}(\mathbf{k},t) = \delta(t), \tag{2.25}$$

where $\widetilde{G}(\mathbf{k},t)$ is the Green propagator in the k-space, which can be decomposed as

$$\widetilde{G}(\mathbf{k},t) = \widetilde{G}_R(\mathbf{k},t) + \widetilde{G}_S(\mathbf{k},t), \tag{2.26}$$

being $\widetilde{G}_R(\mathbf{k},t)$ the general solution of the homogeneous problem and $\widetilde{G}_S(\mathbf{k},t)$ a particular solution of the non-homogeneous problem. Hence, the full solution will be

$$\psi_N(\mathbf{x},t) = \int_{\mathbb{R}^P} d^Px' G_R(\mathbf{x}-\mathbf{x}',t)\psi_N(\mathbf{x}',0), \tag{2.27}$$

where the function $\widetilde{G}_R(\mathbf{k},t)$ fulfills

$$\begin{aligned}&\partial_t\widetilde{G}_R(\mathbf{k},t)-Dk^2\widetilde{G}_R(\mathbf{k},t)=0,\\&\widetilde{G}_R(\mathbf{k},0)=1,\end{aligned}\tag{2.28}$$

hence

$$\begin{aligned}&\widetilde{G}(\mathbf{k},t)=\exp(-Dk^2t),\\&G(\mathbf{x},t)=\tfrac{1}{(2\sqrt{\pi Dt})^P}\exp(\tfrac{-\mathbf{x}^2}{4Dt}).\end{aligned}\tag{2.29}$$

Therefore, we get

$$\psi_N(\mathbf{x},t)=\left(\frac{N}{2\pi t}\right)^{\frac{P}{2}}\int\left(\prod_{\mu=1}^{P}dx'_\mu\right)\exp\left[-N\Phi(\mathbf{x}',\mathbf{x},t)\right],\tag{2.30}$$

$$\Phi(\mathbf{x}',\mathbf{x},t)=\frac{\sum_\mu^P(x_\mu-x'_\mu)^2}{2t}-\ln 2-\frac{1}{N}\sum_{j=1}^{N}\ln\left[\cosh\left(\sum_{\mu=1}^{P}x'_\mu\xi_j^\mu\right)\right]\tag{2.31}$$

and

$$\alpha_N(\mathbf{x},t)=\frac{1}{N}\ln\left[\psi_N(\mathbf{x},t)\right].\tag{2.32}$$

We can solve now the saddle-point equation

$$\alpha(\mathbf{x},t)=\lim_{N\to\infty}\alpha_N(\mathbf{x},t)=\mathrm{Extr}\{\Phi\},\tag{2.33}$$

where we neglected $\mathcal{O}(N^{-1})$ terms, as we performed the thermodynamic limit. Finally, by replacing $t=\beta$ and $\mathbf{x}=0$ and $x'_\nu=\beta\langle m_\nu\rangle$ (hence the original statistical mechanics framework), we obtain the following expressions for the statistical pressure

$$\alpha(\beta,d)=\frac{\beta}{2}\sum_\mu\langle m_\mu\rangle^2-\ln(2)-\left\langle\ln\left[\cosh\left(\beta\sum_{\mu=1}\langle m_\mu\rangle\xi^\mu\right)\right]\right\rangle_\xi,\tag{2.34}$$

whose extremization offers immediately the P desired self-consistency equations for all the $\langle m_\nu\rangle$,

$$\langle m_\nu\rangle=\left\langle\xi^\nu\tanh\left(\beta\sum_{\mu=1}\xi^\mu\langle m_\mu\rangle\right)\right\rangle_\xi\qquad\forall\mu\in[1,P],\tag{2.35}$$

where with the index ξ we emphasized once more that the disorder average over the quenched patterns is performed as well.

Of course, the self-consistence equations (2.35) recover those obtained in [30] via different analytical techniques, where they were also shown to yield to the parallel ansatz (2.2), which, in turn, can be formally written as

$$\sigma_i = \xi_i^1 + \sum_{\nu=2}^{P} \xi_i^\nu \prod_{\mu=1}^{\nu-1} \delta(\xi_i^\mu), \tag{2.36}$$

and it will be referred to as $\sigma^{(P)}$.

The parallel ansatz (2.2) can be understood rather intuitively. To fix ideas let us assume zero noise level and that one pattern, say $\mu = 1$, is perfectly retrieved. This means that the related average magnetization is $m_1 = (1 - d)$, while a fraction d of spins is still available and they can arrange to retrieve a further pattern, say $\mu = 2$. Again, not all of them can match non-null entries in pattern ξ^2 and the related average magnetization is $m_2 = d(1 - d)$. Proceeding in the same way, for all spins, we get the parallel state. Notice that, the number K of patterns which are, at least partially, retrieved does not necessarily equal P. In fact, due to discreteness, it must be $d^{K-1}(1 - d) \leq 1/N$, namely at least one spin must be aligned with ξ^K, and this implies $K \lesssim \log N$.

Such a hierarchical, *parallel*, fashion for alignment, providing an overall energy (see Eq.(1.9))

$$E^{(\mathrm{P})} = -N \sum_{k=1}^{P} [(1 - d)d^{k-1}]^2 + P = -N \frac{(1 - d^{2P})(1 - d)}{1 + d} + P, \tag{2.37}$$

is more optimal than a *uniform* alignment of spins amongst the available patterns, as this case would yield $m_k = (1 - d)/P$ for any k and an overall energy

$$E^{(\mathrm{U})} = -N \sum_{k=1}^{P} \left(\frac{1 - d}{P} \right)^2 + P = -\frac{(1 - d)^2 N}{P} + P, \tag{2.38}$$

being $(1 - d^{2+2P}) > (1 - d^2)/P$.
On the other hand, as we will see in Sec. 2.3, when $d > d_c \approx 1/2$, the state (2.2) is no longer stable and spurious states do emerge.

Before proceeding, it is worth stressing that, although the parallel state (2.2) displays non-zero overlap with several patterns, it is deeply different, and must not be confused with, a spurious state in standard Hopfield networks. In fact, in the former case, at least one pattern is completely retrieved, while in spurious states, the overlap with each memory pattern involved is only partial.

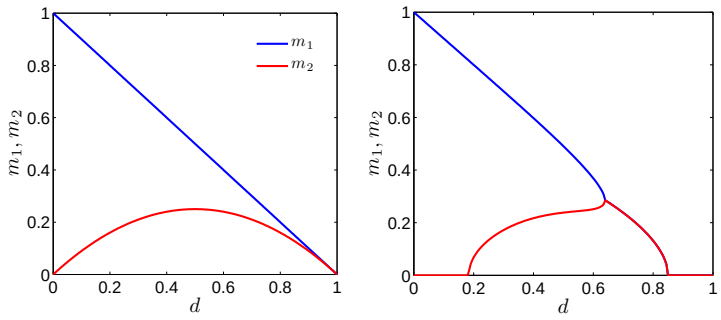

Figure 2.3: Behavior of the two Mattis magnetizations m_1 and m_2 versus d at two (small) noise levels, namely $\beta^{-1} = 10^{-4}$ (left panel) and $\beta^{-1} = 0.05$ (right panel).

Moreover, in standard Hopfield networks, spurious states are somehow undesirable because they provide corrupted information with respect to the best retrieval achievable where one, and only one, pattern is exactly retrieved. Conversely, in our model, the retrieval of more-than-one pattern is unavoidable (for finite d and $\beta \to \infty$) and the quality of retrieval may be excellent (perfect) in the case of patterns poorly (not) overlapping.
Finally, and most importantly, for $\beta \to \infty$ and in a wide region of dilution, the parallel state $\sigma^{(P)}$ corresponds to a global minimum for the energy. This is not the case for an arbitrary mixture of states.

2.2.1 The case $P = 2$

The self-consistencies encoded into Eq.(2.35) for the simplest case $P = 2$ are

$$m_1(\beta, d) = d(1-d)\tanh(\beta m_1) + \frac{(1-d)^2}{2}\big[\tanh[\beta(m_1+m_2)] + \tanh[\beta(m_1-m_2)]\big], \tag{2.39}$$

$$m_2(\beta, d) = d(1-d)\tanh(\beta m_2) + \frac{(1-d)^2}{2}\big[\tanh[\beta(m_1+m_2)] - \tanh[\beta(m_1-m_2)]\big]. \tag{2.40}$$

The solution of these equations (m Vs d) for different values of β is reported in Fig.(2.3). In the low (fast) noise limit ($\beta \to \infty$), when no dilution is present ($d = 0$) the second magnetization m_2 disappears and the first magnetization m_1 approaches the value 1 as expected because the Hopfield model is recovered. As dilution is increased, m_1 decreases linearly, while m_2 displays a parabolic profile with peak at $d = 0.5$. In the presence of (fast) noise, m_2 starts growing for higher values of dilution because (as will be cleared by the

signal-to-noise analysis of the next section) the signal[2] insisting on the latter, which is proportional to $d(1-d)$, must be higher than the noise level in order to be effective. Also notice that, from intermediate dilution onwards, m_1 and m_2 collapse and the related curves converge at a "bifurcation" point.
Let us now deepen these results, first from a more intuitive point of view, and later from a more rigorous one.

In the zero (fast) noise limit, let us fix ξ^1 as the pattern corresponding to the maximum overlap with the magnetic configuration, so that the expected Mattis magnetization is $\langle m_1 \rangle = (1-d)$. The remaining $N\,d$ "free" spins will seek for patterns to align with, namely displaying non-null entries in correspondence with the null entries of ξ^1. Actually, due to dilution, one expects that the second best-matching pattern only engages $N\,d(1-d)$ spins, while the remaining $N\,d^2$ will match other patterns; in general, the k-th best-matching pattern is expected to engage $N\,d^{k-1}(1-d)$.

Such a hierarchical fashion for alignment is more optimal than a uniform alignment of spins amongst the available patterns which would yield $m_k = d/B$ for any k and an overall energy $-N/2\sum_k (d/P)^2 = -(d^2 N)/(2P)$. Indeed, the hierarchical solution is the one that minimizes the energy (recall that the magnetization are summed quadratically) as well as the most likely from a combinatorics point of view, providing an overall energy $-N/2\sum_k [(1-d)d^k]^2 = -N(1-d^{2+2P})(1-d)/[2(1+d)]$.

Therefore, the system is able to perform the "parallel retrieval" of K patterns, whose magnetizations are $m_\mu = (1/N)\sum_{1=1} \xi_i^\mu h_i$, that is $\langle m_1 \rangle = (1-d)$, $\langle m_2 \rangle = d(1-d)$, ..., $\langle m_K \rangle = d^K(1-d)$. It is easy to see that it must be $d^{K+1} = 0$. Hence, for any finite value of d, an infinite number of patterns can in principle be retrieved, i.e. $d^K \to 0$, for $K \to \infty$. More accurately, taking into account the discreteness of the system, we have that the last pattern to be retrieved will match only one spin, which yields $N\,d^K(1-d) = 1$, from which $K = [\log N + \log(1-d)]/\log(1/d) \sim \log N$. In the low storage regime, with P finite or scaling logarithmically with N, the retrieval of all patterns can, in principle, always be accomplished.
When noise is also introduced, we have that for the i-th pattern to be retrieved the field felt by spins has to be larger than the noise level, that is $[d(1-d)^i] > \beta^{-1}$, if this condition is not fulfilled the field is confused with the noise and the pattern can not be retrieved.

In the case of large degree of dilution, i.e. d close to 1, patterns are so sparse that not all the N spins can be matched; assuming that patterns get orthogonal, only a fraction $P(1-d)/N$ ($= \alpha(1-d)$ or $= \alpha \log N(1-d)/N$ in low and high storage regime, respectively) of spins is aligned with a given

[2]We use the term "fields" for the forces acting on h_i and "channels" for those on m_μ.

pattern, the remaining are free and their mean value is zero. In this condition the emergent graph is also disconnected.

Beyond constraints on d, probably the most striking feature displayed by m_1, m_2 is the bifurcation occurring at intermediate values of dilution (see Fig.(2.3)). In order to understand this phenomenon we can divide spins into four sets: $\mathcal{S}_1$, which contains spins i corresponding to zero entries in both patterns ($\xi_i^1 = \xi_i^2 = 0$), therefore behaving paramagnetically; $\mathcal{S}_2$, which includes spins seeing only one pattern ($|\xi_i^1| \neq |\xi_i^2|$);

$\mathcal{S}_3$, which contains spins corresponding to two parallel, non-null entries ($\xi_i^1 = \xi_i^2 \neq 0$), thus being the most stable; $\mathcal{S}_4$, which includes spins i corresponding to two parallel, non-null entries ($\xi_i^1 = -\xi_i^2 \neq 0$), hence intrinsically frustrated.
The cardinality of these sets are: $|\mathcal{S}_1| = d^2$, $|\mathcal{S}_2| = 2d(1-d)$, $|\mathcal{S}_3| = (1-d)^2/2$, and $|\mathcal{S}_4| = (1-d)^2/2$. Now, the most prone spin to align with the related patterns are those in $\mathcal{S}_3$ and in $\mathcal{S}_2$, and this requires $(1-d) < \beta^{-1}$ for the field to get effective. As d is further reduced, m_1 and m_2 grow paired, due to the symmetry of the sets $\mathcal{S}_2$ and $\mathcal{S}_3$. The growth proceeds paired until the magnetizations get the value $m_1 = m_2 = (1-d)^2/2 + d(1-d)$, where the two contributes come from spins aligned with both patterns and with the unique pattern they see, respectively. From this dilution onwards frustrated spins also start to align so that one magnetization necessarily prevails over the other. This explanation can be extended to any finite B and, in general, the number of sets turns out to be $P + 1 + \sum_{k=0}^{P} \lfloor \frac{P-k}{2} \rfloor$.

Now we want to quantify these bifurcation points, and to this task let us call

$$x = \langle m_1 \rangle - \langle m_2 \rangle. \tag{2.41}$$

We use Eqs. (2.39) and (2.40) and expand for small values of x

$$x = d(1-d)[\tanh(\beta\langle m_1\rangle) - \tanh(\beta\langle m_2\rangle)] + (1-d)^2 \tanh\left(\beta\langle m_1\rangle - \langle m_2\rangle\right) \tag{2.42}$$

where

$$\begin{aligned} &d(1-d)\left[\tanh\left(\beta\langle m_1\rangle\right) - \tanh\left(\beta\langle m_2\rangle\right)\right] \sim \\ &d(1-d)\left[\tanh(\beta\langle m_1\rangle) - \tanh(\beta\langle m_2\rangle) + \frac{\beta x}{\cosh^2(\beta\langle m_1\rangle)}\right], \end{aligned} \tag{2.43}$$

and

$$(1-d)^2 \tanh(\beta\langle m_1\rangle - \langle m_2\rangle) \sim (1-d)^2 \beta x + O(x^3). \tag{2.44}$$

Thus, the leading term is

$$x \sim \left[\frac{d(1-d)\beta}{\cosh^2(\beta\langle m_1\rangle)} + \beta(1-d)^2\right] x. \tag{2.45}$$

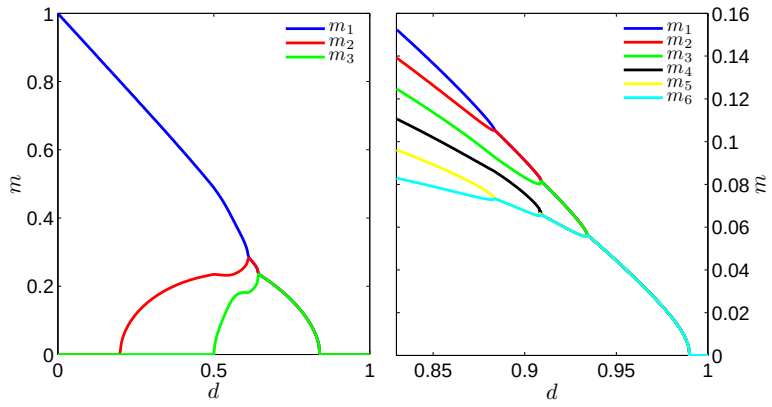

Figure 2.4: Parallel retrieval of three (left panel) and of six (right panel) patterns. Behavior of the two Mattis magnetization versus d at noise level $\beta^{-1} = 0.05$.

The critical value of β corresponding to the bifurcation point is defined as

$$\beta_c^{bif} = \frac{1}{(1-d)^2\left[1 + \frac{(1-d)}{d}\frac{1}{\cosh^2(\beta_c^{bif} m_1)}\right]}. \tag{2.46}$$

This mechanism can be easily generalized to the case of multiple patterns.

We move now to analyze the critical noise level at which the magnetizations disappear and the network dynamics becomes ergodic, still in this test-case of two patterns: Expanding expressions (2.40) we find

$$\begin{aligned}
&\langle m_2\rangle \sim d(1-d)[\beta\langle m_2\rangle] + \frac{(1-d)^2}{2}[\beta\langle m_1\rangle + \beta\langle m_2\rangle + \\
&+ \frac{\beta^3}{3}(\langle m_1\rangle^3 + \langle m_2\rangle^3 + 3\langle m_1\rangle^2\langle m_2\rangle + 3\langle m_1\rangle\langle m_2\rangle^2)] + \\
&+ d(1-d)\frac{\beta^3}{3}\langle m_2\rangle^3 - \frac{(1-d)^2}{2}[\beta\langle m_1\rangle - \beta\langle m_2\rangle + \frac{\beta^3}{3}(\langle m_1\rangle^3 + \\
&- \langle m_2\rangle^3 - 3\langle m_1\rangle^2\langle m_2\rangle + 3\langle m_1\rangle\langle m_2\rangle^2)],
\end{aligned} \tag{2.47}$$

such that we can write

$$\langle m_2\rangle \sim (1-d)\beta\langle m_2\rangle + \mathcal{O}(\langle m_2\rangle^3). \tag{2.48}$$

Therefore the critical noise level turns out to be

$$\beta_c = \frac{1}{1-d}. \tag{2.49}$$

This calculation can easily be generalized to several patterns, too.

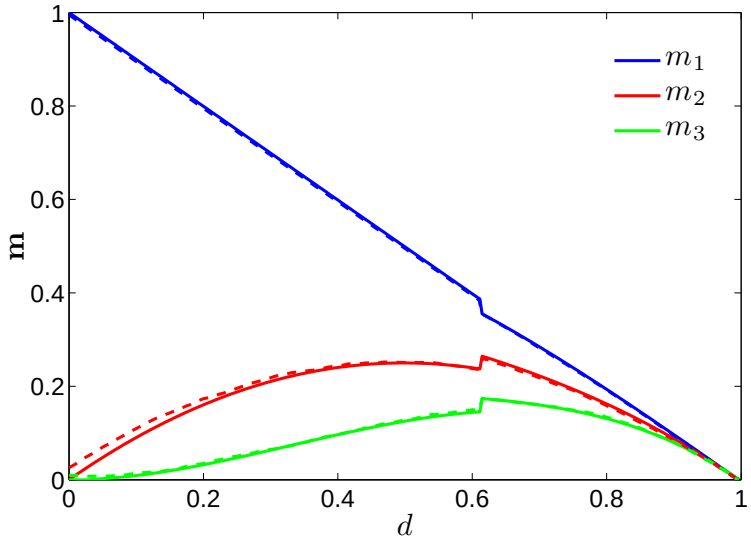

Figure 2.5: Parallel retrieval of three strategies. Behavior of three Mattis magnetization versus d in the slow (fast) noise limit (i.e. $\beta^{-1} = 10^{-4}$). Continuous lines correspond to numerical solution of Eqs. (2.50)-(2.52), while dashed lines correspond to Monte Carlo simulations.

2.2.2 The case $P = 3$

When three patterns are considered, the related self-consistent equations that constraint the system to parallel processing are the following (we skip the brackets $\langle . \rangle$ for the sake of clearness):

$$
\begin{aligned}
m1 = {} & d^2(1-d)\tanh[\beta m1] - (1/4)d(1-d)^2\tanh[\beta(-m1-m2)] + \\
& + (1/4)d(1-d)^2\tanh[\beta(m1-m2)] - (1/4)d(1-d)^2\tanh[(-m1+m2)] + \\
& + (1/4)d(1-d)^2\tanh[\beta(m1+m2)] - (1/4)d(1-d)^2\tanh[\beta(-m1-m3)] + \\
& - (1/4)d(1-d)^2\tanh[\beta(m1-m3)] - (1/8)(1-d)^3\tanh[\beta(-m1-m2-m3)] + \\
& + (1/8)(1-d)^3\tanh[\beta(m1-m2-m3)] - (1/8)(1-d)^3\tanh[\beta(-m1+m2-m3)] + \\
& + (1/8)(1-d)^3\tanh[\beta(m1+m2-m3)] - (1/4)d(1-d)^2\tanh[\beta(-m1+m3)] + \\
& + (1/4)d(1-d)^2\tanh[\beta(m1+m3)] - (1/8)(1-d)^3\tanh[\beta(-m1-m2+m3)] + \\
& + (1/8)(1-d)^3\tanh[\beta(m1-m2+m3)] - (1/8)(1-d)^3\tanh[\beta(-m1+m2+m3)] + \\
& + (1/8)(1-d)^3\tanh[\beta(m1+m2+m3)]]
\end{aligned}
\tag{2.50}
$$

$$
\begin{aligned}
m2 = &-(1/4)d(1-d)^2\tanh[\beta(-m1-m2)] - (1/4)d(1-d)^2\tanh[\beta(m1-m2)]+\\
&+d^2(1-d)\tanh[\beta m2] + (1/4)d(1-d)^2\tanh[\beta(-m1+m2)]+\\
&+(1/4)d(1-d)^2\tanh[\beta(m1+m2)] - (1/4)d(1-d)^2\tanh[\beta(-m2-m3)]+\\
&-(1/8)(1-d)^3\tanh[\beta(-m1-m2-m3)] - (1/8)(1-d)^3\tanh[\beta(m1-m2-m3)]+\\
&+(1/4)d(1-d)^2\tanh[\beta(m2-m3)] + (1/8)(1-d)^3\tanh[\beta(-m1+m2-m3)]+\\
&+(1/8)(1-d)^3\tanh[\beta(m1+m2-m3)] - (1/4)d(1-d)^2\tanh[\beta(-m2+m3)]+\\
&-(1/8)(1-d)^3\tanh[\beta(-m1-m2+m3)] - (1/8)(1-d)^3\tanh[\beta(m1-m2+m3)]+\\
&+(1/4)d(1-d)^2\tanh[\beta(m2+m3)] + (1/8)(1-d)^3\tanh[\beta(-m1+m2+m3)]+\\
&+(1/8)(1-d)^3\tanh[\beta(m1+m2+m3)]]
\end{aligned}
\tag{2.51}
$$

$$
\begin{aligned}
m3 = &-(1/4)d(1-d)^2\tanh[\beta(-m1-m3)] - (1/4)d(1-d)^2\tanh[\beta(m1-m3)]+\\
&-(1/4)d(1-d)^2\tanh[\beta(-m2-m3)] - (1/8)(1-d)^3\tanh[\beta(-m1-m2-m3)]-\\
&-(1/8)(1-d)^3\tanh[\beta(m1-m2-m3)] - (1/4)d(1-d)^2\tanh[\beta(m2-m3)]-\\
&-(1/8)(1-d)^3\tanh[\beta(-m1+m2-m3)] - (1/8)(1-d)^3\tanh[\beta(m1+m2-m3)]+\\
&+d^2(1-d)\tanh[\beta m3] + (1/4)d(1-d)^2\tanh[\beta(-m1+m3)]+\\
&+(1/4)d(1-d)^2\tanh[\beta(m1+m3)] + (1/4)d(1-d)^2\tanh[\beta(-m2+m3)]+\\
&+(1/8)(1-d)^3\tanh[\beta(-m1-m2+m3)] + (1/8)(1-d)^3\tanh[\beta(m1-m2+m3)]+\\
&+(1/4)d(1-d)^2\tanh[\beta(m2+m3)] + (1/8)(1-d)^3\tanh[\beta(-m1+m2+m3)]+\\
&+(1/8)(1-d)^3\tanh[\beta(m1+m2+m3)]].
\end{aligned}
\tag{2.52}
$$

Recalling the picture explained in the previous subsection, the magnetizations m_1, m_2 and m_3 again grow together until all spins corresponding to equal non-null entries and to single non-null entries are aligned. Then spins which are aligned only with two patterns out of three start to feel the field and get aligned hence breaking the symmetry. At this point, say m_1 and m_2, still grow while m_3 decreases. The next symmetry-breaking occurs when all spins corresponding to equal non-null entries $\xi^1 = \xi^2$ get aligned. From this point onward one magnetization prevails against the other. The same process applies, mutatis mutandis, for larger number of patterns (see Fig.2.4).

The last subtlety to be investigated is given by the small discontinuities in the behavior of the magnetizations (see for instance Fig.2.5). To explain this feature, let us consider the set of patterns $\xi_1, \xi_2, ..., \xi_P$ and assume the zero fast noise limit ($\beta \to \infty$) for the sake of simplicity, so that we can take $|m^k| = (1-d)d^{k-1}$, for $k = 1, ..., P$ as (absolute) Mattis magnetizations. The field insisting on the arbitrary spin σ_i can be written as

$$
h_i = \frac{1}{N}\sum_{j\neq i}^{N} J_{ij}\sigma_j = \sum_{\mu=1}^{P}\xi_i^\mu m^\mu - \frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i \approx \sum_{\mu=1}^{P}\xi_i^\mu m^\mu, \tag{2.53}
$$

where in the last passage we dropped the second sum as it is vanishing in the thermodynamic limit. Now, let us consider the spin h_1, which, again without loss of generality can be thought of as aligned with the first pattern and equal to $+1$. The field insisting on this lymphocyte is $h_1 = (1-d)[1+$

$\sum_{\mu=2}^{B} \epsilon(1,\mu) d^{\mu-1}]$, where $\epsilon(1,\mu) = \mathrm{sign}(\xi_1^\mu, m^\mu)$. We notice that, in general, h_1 is not positive definite so that the occurrence of the condition $h_1 < 0$ would lead to the spin flip $h_1 = 1 \to h_1 = -1$ and, consequently, to $m_1 < (1-d)$. In order to understand this effect we focus on $\epsilon(1,\mu)$. By assumption, $m_1 = (1-d)$ and $h_1 = \xi_1^1$, so that the first entry of pattern $\mu = 1$ effectively contributes to the related magnetization m_1. As for the following magnetizations $m_{\mu>2}$, effective contributes can arise only from entries $\xi_j^{\mu>2}$ corresponding to null entries in ξ_j^1. Otherwise stated, there is no correlation between ξ_1^μ and m^μ for $\mu > 1$ (in fact, $\epsilon(1,\mu)$ is zero on average), and one can count the pattern configurations leading to $h_1 < 0$ applying combinatorics.

Seeking for clarity, we consider the following explicit cases:
- The probability that the first entries of all patterns $\mu > 1$ are misaligned with respect to the related magnetizations is $[(1-d)/2]^{P-1}$, hence giving a field $h_1 = (1-d)[1 - \sum_{\mu>2} d^{\mu-1}] = 1 - 2d + d^{P+1}$. Such a field turns out to be negative in the interval $a_1 < d < 1$, where $a_1 \to 1/2$ for $P \to \infty$.
- The probability that the first entries of all patterns $\mu > 1$ but one, say ξ^l, are misaligned and that $\xi_1^l = 0$ is $d[(1-d)/2]^{P-2}$, and this would lead to $h_1(l) = (1-d) - d(1-d^P) + (1-d)d^{l-1}$, which is negative for $a_2 < d < 1$, where $a_2 \to 1/2$ for $P \to \infty$; of course $h_1(l)$ is growing with l.
- The probability that the first entries of all patterns $\mu > 1$ but one, say ξ^l, are misaligned and that $\xi_1^l = 1$, is $d[(1-d)/2]^{P-1}$ and this configuration yields $h_1(l) = (1-d) - d(1-d^P) + 2(1-d)d^{l-1}$. For instance, when $l = 2$ and $P \gg 1$, the field is negative for $d > 1/\sqrt{2}$; when $l = 3$ the field is negative for $d > a_3$, where $a_3 \approx 0.648$.
Summarizing, in the zero noise limit $\beta \to \infty$ for any given dilution d, the probability that $m_1 < (1-d)$ can be written as a sum over pattern configurations leading to $h_1 < 0$. For instance, for $P = 3$, only one out of the 3^{B-1} possible configurations, i.e. $\mathrm{sign}(\xi_2^\mu, m^\mu) = \mathrm{sgn}(\xi_3^\mu, m^\mu) = -1$, can yield a spin-flip: the corresponding field is $h_1 = (1-d)(1-d-d^2)$, which is negative for $d > (\sqrt{5}-1)/2 \approx 0.62$ (see Fig.(2.5)). Therefore, for that value of dilution onwards, m_1 is reduced with respect to the optimal value $(1-d)$. The extent of the loss is a fraction $1/9$ of the total, namely ≈ 0.34 (see Fig.(2.5)).

Notice that while the change reduces m_1, other magnetizations are favored by the spin-flip and undergo a proportional increment. Also, the occurrence of a magnetization reduction with respect to the optimal value is more likely for the highest magnetization m_1, because fields insisting on spins contributing to m_1 are the most complex, being the sum of $P-1$ terms. The same discussion can be applied in turns to m_2: now the number of terms which sum up to give the field insisting on the $(1-d)d$ spins which contribute effectively to m_2 is $P-2$, so that there are far less configurations able to yield a negative field. Consequently, a loss in m_2 is less likely. Therefore, as long as the

number of patterns allows readjustments in the value of magnetizations with respect to those expected, the arbitrary m_k may display complex corrections (possibly occurring at slightly different values of d) due to the combination of several simple corrections, each corresponding to the readjustment affecting the previous magnetizations $m_{\mu<k}$ (see Fig.(2.5)).

2.2.3 Signal to noise ratio

As usually done in the neural network context [17], we couple the statistical mechanics inspection to signal-to-noise analysis. Aim of this procedure is trying to confirm the "parallel ansatz" we implicitly made by studying the stability of the basins of attractions (whose fixed points are the learned strategies) created in the hierarchical fashion we prescribed. We recall that the model we are investigating describes a low storage of information in the associative network so that no slow noise is induced by the underlying spin glass, i.e. $\alpha = 0$. Nonetheless, we study the signal to noise ratio in the zero fast noise limit ($\beta \to \infty$) as a problem formulated in general terms of α, d; then, we take the limit $\alpha \to 0$ to get estimate about the stability of the basins of attractions (where the presence of fast noise can possibly produce fluctuations).

Without loss of generality, we assume that the network is retrieving the first pattern. This means that spins are aligned with the non-null entries in the first bit-string ξ^1, while the remaining spins explore the other patterns. Thus, for the generic spin σ_i we can write

$$\sigma_i = \xi_i^1 + \sum_{\nu=2}^{P} \xi_i^\nu \prod_{\mu=1}^{\nu-1} \delta(\xi_i^\mu). \tag{2.54}$$

Accordingly, the local field acting on the i^{th} lymphocyte can be written as

$$h_i = \frac{1}{N} \sum_{j \neq i}^{N} \sum_{\mu}^{P} \xi_i^\mu \xi_j^\mu \left[\xi_j^1 + \sum_{\nu=2}^{P} \xi_j^\nu \prod_{\mu=1}^{\nu-1} \delta(\xi_j^\mu) \right]. \tag{2.55}$$

- In the reference case $P = 1$, like for the pure states of the Hopfield network, we set
$$\sigma_i = \xi_i^1 + \delta(\xi_i^1) k_i, \tag{2.56}$$
where k_i is a random variable uniformly distributed on the values ± 1 added to ensure that there are no nulls entries in the state of the network. Hence we find
$$\langle h_i \sigma_i \rangle_\xi = \langle signal + noise \rangle_\xi = \langle signal \rangle_\xi \tag{2.57}$$

being $\langle noises \rangle_\xi = 0$, and so for large N we have

$$\langle signal \rangle_\xi = \frac{N-1}{N}(1-d) = (1-d), \tag{2.58}$$

while

$$\langle (noises)^2 \rangle_\xi = \frac{P-1}{N}(1-d)^2 = \alpha(1-d)^2. \tag{2.59}$$

- In the test case of two patterns retrieved, $P = 2$, we set:

$$\sigma_i = \xi_i^1 + \delta(\xi_i^1)[\xi_i^2 + \delta(\xi_i^2)k_i]. \tag{2.60}$$

Now, we need to distinguish between the various possible configurations:

  - $\forall i$ such that $\xi_i^1 \neq 0, \xi_i^2 = 0$ and so that $\sigma_i = \xi_i^1 \neq 0$ for large value of N

$$\langle signal \rangle_\xi = (1-d), \quad \langle noises \rangle_\xi = 0, \tag{2.61}$$

$$\langle (noises)^2 \rangle_\xi = \frac{(N-1)(P-2)}{N^2}(1-d)^2 = \alpha(1-d)^2. \tag{2.62}$$

  - $\forall i$ such that $\xi_i^1 \neq 0, \xi_i^2 \neq 0$ and so that $\sigma_i = \xi_i^1 \neq 0$ if $\xi_i^1 = \xi_i^2$

$$\langle signal \rangle_\xi = 2(1-d) - (1-d)^2, \quad \langle noises \rangle_\xi = 0, \tag{2.63}$$

if $\xi_i^1 = -\xi_i^2$

$$\langle signal \rangle_\xi = (1-d)^2, \quad \langle noises \rangle_\xi = 0. \tag{2.64}$$

and in both cases

$$\langle (noises)^2 \rangle_\xi = \frac{(N-1)(P-1)}{N^2}(1-d)^3 + \frac{(N-1)(P-2)}{N^2}d(1-d)^2 = \alpha(1-d)^2. \tag{2.65}$$

  - $\forall i$ such that $\xi_i^1 = 0, \xi_i^2 \neq 0$ and so that $\sigma_i = \xi_i^2 \neq 0$

$$\langle signal \rangle_\xi = d(d-1), \quad \langle noises \rangle_\xi = 0, \tag{2.66}$$

$$\langle (noises)^2 \rangle_\xi = \frac{(N-1)(P-1)}{N^2}(1-d)^3 + \frac{(N-1)(P-2)}{N^2}(1-d)^2 d = \alpha(1-d)^2. \tag{2.67}$$

Therefore, in the regime of low storage of strategies we are exploring ($\alpha = 0$), the retrieval is stable, states are well defined and the amplitude of the signal on the first channel is order $(1-d)$ while on the second is of order $d(1-d)$, in perfect agreement with both the statistical mechanics analysis and Monte Carlo simulations.

Once proved that these parallel states exist, it would be interesting trying to understand deeper their structure in the configurational space. To this task let us fix a pattern ξ_i^1, with $i = 1, ..., N$, and a dilution d, in such a way that $N\,d$ of ξ^1 entries are expected to be null and the remaining $N(1-d)$ are expected to be half equal to $+1$ and half equal to -1. The number of spins configurations displaying maximum overlap with ξ^1 corresponds to the degeneracy induced by null entries, namely 2^{Nd}; all these configurations lay in an energy minimum because their Mattis magnetization is maximum (actually the same holds for the symmetrical configurations due to the gauge symmetry of the model).

Let us now generalize this discussion by introducing the number of configurations $n(m,d)$ whose overlap with the given pattern displays m misalignments in such a way that $n(m,d)$ is given not only by the degeneracy induced by null entries, but also by the degeneracy induced by the choice of m entries out of $N(1-d)$ which have to be mismatched. It is easy to see that $n(m,d) = 2^{Nd}\binom{N(1-d)}{m}$. Interestingly, for such configurations the signal felt by a spin i can be written as $h_i = \xi_i^1[N((1-d)) - 2m]$ and the effect of the correction due to the m misalignments might be vanishing in the presence of a sufficiently large level of noise, so that the system is not restricted to the 2^{Nd} configurations corresponding to the minimum energy, but it can also explore all the configurations $n(m,d)$.

Therefore, we can count the number of configurations $\tilde{n}(x,d)$ exhibiting a number of misalignments, with respect to ξ^1, up to a given threshold x; in the presence of noise such configurations are all accessible, namely they all lay in the same "deep" minimum. Indeed, we can write $\tilde{n}(x,d) = \sum_{m=0}^{x} n(m,d)$; of course, for $x = N(1-d)$ we recover $\tilde{n}(x,d) = 2^N$. Moreover, when $x = N(1-d)/2$, we can exploit the identity $\sum_{k=0}^{i}\binom{2i}{k} = 1/2[4^i + \binom{2i}{i}]$, and assuming without loss of generality $N(1-d)$ to be even we get

$$\tilde{n}(N(1-d)/2,d) = \sum_{m=0}^{x} n(m,d) = \frac{2^{Nd}}{2}\left[2^{N(1-d)} + \binom{N(1-d)}{N(1-d)/2}\right] \approx$$
$$\frac{2^N}{2}\left[1 + \sqrt{\frac{2}{\pi N(1-d)}}\right], \tag{2.68}$$

where in the last passage we used the Stirling approximation given that

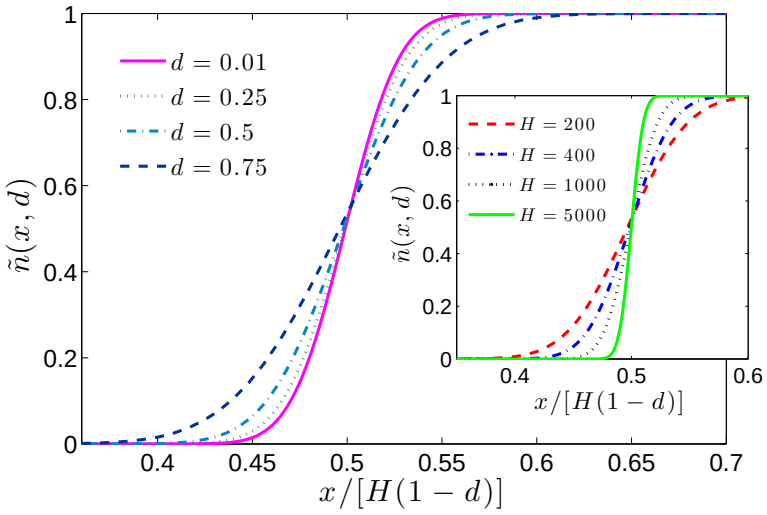

Figure 2.6: Normalized number of accessible configurations $\tilde{n}(x, d)$ as a function of x and d for a system made up of N spins. The critical line $x_c = (1-d)$, corresponds to the emergence of a giant component.

$N((1-d)) \gg 1$. Then, we have $\tilde{n}(N(1-d)/2, d) \gtrsim 1/2$, and similar calculations can be drawn for smaller thresholds, e.g., $\tilde{n}(N(1-d)/2-1, d) \lesssim 1/2$.

As shown in Fig.(2.6), once d is fixed, when x is small only a microscopic fraction $\tilde{n}(x, d)/2^N$ of configuration is accessible (in the thermodynamic limit this fraction is vanishing), while by increasing the tolerance x, more and more configuration get accessible and correspondently their fraction gets macroscopic. From a different perspective, each configuration can be looked at as a node of a graph and those accessible are connected together. The link probability is then related to x and when x is large enough a "giant component" made up of all accessible configurations emerges. This is a percolation process in the space of configurations. Indeed, similarly to what happens in canonical percolation processes, the curves representing the giant component relevant to different sizes N intersect at around $1/2$, which distinguishes the percolation threshold x_c. According to Eq.(2.68) we can write $x_c \approx N(1-d)/2$.

Interestingly, when a giant component emerges retrieval is no longer meaningful because the system may retrieve essentially anything and this corresponds to the critical line (in the d, β plane) where all the magnetizations simultaneously disappear.

2.3 The Emergence of Spurious States

In Sec. 2.2, we explained why we expect the parallel state (2.36) to occur, exploiting the fact that each pattern tends to align as many spins among those still available. Actually, this intuitive approach yields the correct picture for $T = 0$ (no fast noise) and not-too-large d, while when either T or the degree of dilution are large enough, the system can relax to a state where only one pattern is retrieved or fall into a spurious state where several patterns are partially retrieved, but none exactly. For instance, when patterns are sparse, none of them can generate an attraction basin strong enough to align all available spins, in such a way that stationary, mixture states can emerge.

Let us start from the noiseless case and consider the state (2.36) corresponding to the parallel ansatz (2.2): we notice that, on average, there exists a fraction $2[(1-d)/2]^P$ of spins σ_i corresponding to the entries $\xi_i^1 = 1, \xi_i^k = -1, \forall k \in [1, P]$ (and analogously for the "gauged" case $\xi_i^1 = -1, \xi_i^k = +1$) and expected to be aligned with the first entry ξ_i^1, in such a way that the overall field insisting on each of them is $h_i = m_1 - m_2 - m_3 - .... - m_P$. Of course, such spins are the most unstable, and, at zero noise level, they flip whenever h_i happens to be negative, that is, when $m_1 < \sum_{k=2}^{P} m_k$. Exploiting the ansatz $m_k = d^{k-1}(1-d)$, this can be written as

$$h_i = (1-d)\left[1 - \frac{d - d^P}{1-d}\right] = 1 - 2d + d^P, \tag{2.69}$$

which becomes negative for a value of dilution $d_c(P)$, which converges exponentially from above to $1/2$ as P gets large. From this point onwards, the first pattern is no longer completely retrieved and the system fails to parallel retrieve (according to the definition in Eq.(2.36)). Therefore, when $d \geq d_c(P)$, genuine spurious states emerge and the system relaxes to states which correspond to mixture of $p \leq P$ patterns, but none of them is completely retrieved (at least up to extreme values of dilution). As we will see in Sec. 2.4.4, the transition at $d_c(P)$ is first order.

Moreover, from Eq.(2.69) we find that the case $P = 2$ has no solution in the range $d \in [0, 1]$, meaning that the parallel-retrieval state is always a stable solution in the zero noise limit; on the other hand, $d_c(3) \approx 0.62$, $d_c(4) \approx 0.54$ and so on.

Such phenomenology concerns relatively large degrees of dilution, yet, the presence of noise can also destabilize the true parallel-retrieval state (2.2) in the regime of small degrees of dilution. In fact, we expect that the spins aligned according to the k-th pattern associated to a magnetization $m_k = d^{k-1}(1-d)$ will loose stability at noise levels $T > d^{k-1}(1-d)$. In particular, at $T > d(1-d)$, only one pattern will be retrieved and the pure

state is somehow recovered. As we will see in Sec. 2.4.4, such estimates are correct for small d.

Typical spurious states emerging in standard associative networks are the so-called symmetric mixtures of $p \leq P$ states, which can be described as

$$\sigma_i = \text{sign}\left(\sum_{\mu=1}^{p} \xi_i^\mu\right), \tag{2.70}$$

and it will be referred to as $\sigma^{(S)}$. We anticipate that the symmetric mixture turns out to emerge also in the diluted model under investigation.
Now, in the standard Hopfield model, odd mixtures of p patterns, are metastable, i.e. their energies are higher than those of the pure patterns, and, moreover, the smaller p and the more energetically favorable the mixture. On the other hand, even mixtures of p patterns are unstable (they are saddle-points of the energy).

More precisely, at the critical temperature of the standard Hopfield model, namely at $T_c = 1$, all the symmetric spurious states become extrema in the free-energy landscape. They are either minima, maxima, or saddle-points. As $T < 0.461$ spurious states become successively stable. First, the symmetric three mixtures become stable and begin to attract. As the temperature is lowered further, more and more of the symmetric odd mixtures become attractors. Lower mixtures become stable at higher temperature. The pure pattern attractors remain the absolute minima in the landscape all the way down to $T = 0$. They always have the largest basins of attraction.

The instability of even mixtures is often associated to the fact that, for a macroscopic fraction of spins, $\sigma^{(S)}$ is not defined due to the ambiguity of the sign. For instance, when $p = 2$, $\sum_{\mu=1}^{p} \xi_i^\mu$ occurs to be null for half of the spins and the related values are defined stochastically according to the distribution

$$P(\sigma_i) = \frac{1}{2}(\delta_{\sigma_i+1} + \delta_{\sigma_i+1}). \tag{2.71}$$

However, as we will show in Sec. 2.4.3, this is not the case for this diluted model as it displays wide regions in the parameter space (d, T) where even and/or odd symmetric mixtures are stable.

As we will see in Sec. 2.4.3, the symmetric mixture $\sigma^{(S)}$ can become unstable and relax to a different spurious state which is a "hybrid" state between the symmetric mixture $\sigma^{(S)}$ and the parallel state $\sigma^{(P)}$.

To begin and fix ideas, let us set $P = 3$ and start from the state $\sigma_i = \text{sign}(\xi_i^1 + \xi_i^2 + \xi_i^3)$. In the presence of dilution the argument $\xi_i^1 + \xi_i^2 + \xi_i^3$ can be zero and in that situation one can adopt the following hierarchical rule: take $\sigma_i = \xi_i^1$ provided that $\xi_i^1 \neq 0$; otherwise, if $\xi_i^1 = 0$, then take $\sigma_i = \xi_i^2$

provided that $\xi_i^2 \neq 0$; otherwise, if also $\xi_i^2 = 0$, then take $\sigma_i = \xi_i^3$ provided that $\xi_i^3 \neq 0$; otherwise, if also $\xi_i^3 \neq 0$, then put $\sigma_i = \pm 1$ with probability $1/2$. In this way we can built a state, generally defined for any P, and, being $\Xi = \sum_\mu \xi_i^\mu$, it can written as

$$\sigma_i = (1 - \delta_{\Xi,0})\text{sign}(\Xi) + \delta_{\Xi,0}[\xi_i^1 + \delta_{\xi_i^1,0}\xi_i^2 + \delta_{\xi_i^1,0}\delta_{\xi_i^2,0}\xi_i^3 + ...], \tag{2.72}$$

which will be referred to as $\sigma^{(H)}$.

The related average Mattis magnetizations can be calculated as the sum of one contribution m_0 (the same for any μ) deriving from the spins corresponding to non ambiguous sign function (i.e., $\Xi \neq 0$), and another contribution accounting for hierarchical corrections (i.e., $\Xi = 0$). Let us focus on the first term:

$$\begin{aligned} m_0 &= \langle \xi^\mu \text{sign}(\Xi) \rangle_\xi & (2.73)\\ &= \frac{1-d}{2} \left\langle \text{sign}(1 + \sum_{\nu \neq \mu}^{P} \xi^\mu) - \text{sign}(-1 + \sum_{\nu \neq \mu}^{P} \xi^\mu) \right\rangle_\xi & (2.74)\\ &= (1-d) \left[\mathcal{P}(\sum_{\nu \neq \mu}^{P} \xi^\nu < 1) - \mathcal{P}(\sum_{\nu \neq \mu}^{P} \xi^\nu > 1) \right], & (2.75) \end{aligned}$$

where, in the last step, we exploited the implicit symmetry in pattern entries and $\mathcal{P}(\sum_{\nu \neq \mu}^{P} \xi^\nu \gtrless 1)$ represents the probability that the specified inequality is verified over the distribution (2.1). The latter quantity can also be looked at as the probability for a symmetric random walk with holding probability d to be at distance $\gtrless 1$ from its origin after a time span $P - 1$. Hence, we get

$$m_0 = (1-d)[\mathcal{P}(0 \to 0, P-1) + \mathcal{P}(0 \to 1, P-1)], \tag{2.76}$$

where $\mathcal{P}(x_0 \to x, t)$ is the probability for a symmetric random walk with stopping probability d to move from site x_0 to site x in t steps, namely

$$\mathcal{P}(x_0 \to x, t) = \sum_{s=0}^{t-(x-x_0)} \frac{t!}{s! \left(\frac{t-s-(x-x_0)}{2}\right)! \left(\frac{t-s+(x-x_0)}{2}\right)!} d^s \left(\frac{1-d}{2}\right)^{t-s}. \tag{2.77}$$

The second contribution to the magnetization is $(1-d)\sum_{k=1}^{P-1} \mathcal{P}(0 \to 1, P-k) d^{k-1}$.

Finally, by summing the two contributions we find the following expres-

sions for $P = 3$

$$m_1 = \frac{1}{2}(1 + d - 3d^2 + d^3), \tag{2.78}$$

$$m_2 = \frac{1}{2}(1 - d)(1 + d^2), \tag{2.79}$$

$$m_3 = \frac{1}{2}(1 - 3d + 5d^2 - 3d^3), \tag{2.80}$$

and for $P = 5$

$$m_1 = \frac{1}{8}(3 + 9d - 42d^2 + 74d^3 - 65d^4 + 21d^5), \tag{2.81}$$

$$m_2 = \frac{1}{8}(1 - d)(3 + 6d^2 - d^4), \tag{2.82}$$

$$m_3 = \frac{1}{8}(1 - d)(3 - 4d + 18d^2 - 20d^3 + 11d^4), \tag{2.83}$$

$$m_4 = \frac{1}{8}(1 - d)(3 - 4d + 18d^2 - 28d^3 + 19d^4), \tag{2.84}$$

$$m_5 = \frac{1}{8}(1 - d)(3 - 4d + 18d^2 - 36d^3 + 27d^4). \tag{2.85}$$

The expressions for arbitrary P can be analogously calculated exactly and some examples are shown in Fig.(2.7).

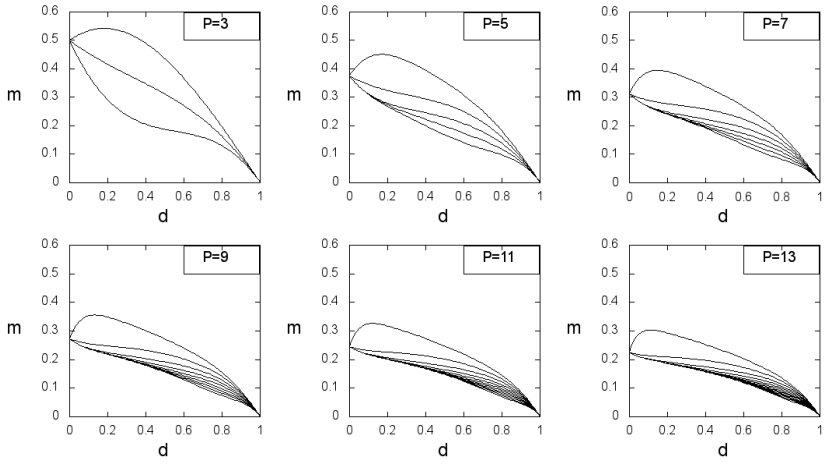

Figure 2.7: Mattis magnetizations $\mathbf{m}$ versus dilution d, according to the analytical expression derived in Sec. 2.3. Each panel refers to a different value of P, as specified.

We expect σ^H to become globally stable in the region of very large dilutions ($d > d_H(P)$); intuitively, dilution must be large enough to make magnetizations rather close to each other in such a way that the least signalled

spins corresponding to $(-,-,...,-,+,+,...,+)$ (overall $(P-1)/2$ negative entries and $(P+1)/2$ positive entries) are stable. This means $\sum_{1=1}^{N}(1-\delta_{\Xi,0})\text{sign}(\Xi)\xi_i^\mu/N > \sum_{k=1}^{(P-1)/2} h_k(P+1)/(P-k)$, where $h_k = 2\sum_l[(1-d)/2]^{2l}d^{P-2l}(P-k)!/[l!(l-1)!(P-k-2l+1)!]$ and P is odd. This condition is fulfilled for values of dilution larger than $d_H(P)$, which converges to 1 as P gets larger, hence, in order to tackle this limit, dilution must become a function of the system size $d \to d(N)$. In this case the network itself becomes diluted as well and different techniques are required; this will not be discussed in this manuscript.

2.4 Stability Analysis

The set of solutions for self-consistent equations (2.35) describes states whose stability may vary strongly. In fact, provided the network has reached them, in the noiseless limit (of whatever kind) it would persist in those states. However, the equations do not contain any information about whether the solutions will be stable against small perturbations, that is to say if the system will indeed really thermalize on these states or will fall apart more or less quickly. In order to evaluate their stability we need to check the second derivative of the free-energy [17]. More precisely, we further need to build up the so called "stability matrix" $\mathbf{A}$ with elements

$$A^{\mu\nu} = \frac{\partial^2 f_\beta(m)}{\partial m^\mu \partial m^\nu}. \tag{2.86}$$

Then, we evaluate and diagonalize $\mathbf{A}$ at a point $\tilde{\mathbf{m}}$, representing a particular solution of the self-consistence equations (2.35), in order to determine whether $\tilde{\mathbf{m}}$ is stable or not. Being $\{E_\mu\}_{\mu=1,...,P}$, the set of related eigenvalues, $\tilde{\mathbf{m}}$ is stable whenever all of them are positive.

Now, from Eq.(2.34) and (2.86), remembering that $\alpha(\beta,d) = -\beta f(\beta,d)$, we find straightforwardly

$$A^{\mu\nu} = [1-\beta(1-d)]\delta^{\mu\nu} + \beta Q^{\mu\nu}, \tag{2.87}$$

where

$$Q^{\mu\nu} = \langle \xi^\mu \xi^\nu \tanh^2(\beta \sum_\mu m^\mu \xi^\mu)\rangle_\xi. \tag{2.88}$$

Of course when $d=0$ we recover $A^{\mu\nu} = (1-\beta)\delta^{\mu\nu} + \langle \xi^\mu\xi^\nu \tanh^2(\beta\sum_\mu m^\mu\xi^\mu)\rangle_\xi$, namely the result known for the standard Hopfield model.

We now consider several states, known to be solutions of self-consistence equations (2.35) and check their stability. In this way we will find constraints in the region (T,d) where those states are stable and then we will build up the phase diagram.

2.4.1 Paramagnetic State

Let us start with the paramagnetic state, which is described by

$$m^\mu = 0 \, \forall \mu \tag{2.89}$$

this state trivially fulfills Eq.(2.35).

By replacing this expression in Eq.(2.87) and in Eq.(2.88) we find

$$A^{\mu\nu} = \delta_{\mu\nu}[1 - \beta(1-d)]. \tag{2.90}$$

Therefore, in this case, $\mathbf{A}$ is diagonal and its eigenvalues are directly $E_\mu = A^{\mu\mu} = 1 - \beta(1-d), \forall \nu \in [1, P]$. We can conclude the paramagnetic state exists and is stable in the region $1 - \beta(1-d) > 0$, that is (remembering that $T = \beta^{-1}$)

$$\text{PM} \quad \text{stability} \Rightarrow T > 1 - d. \tag{2.91}$$

This region is highlighted in Fig.(2.8).

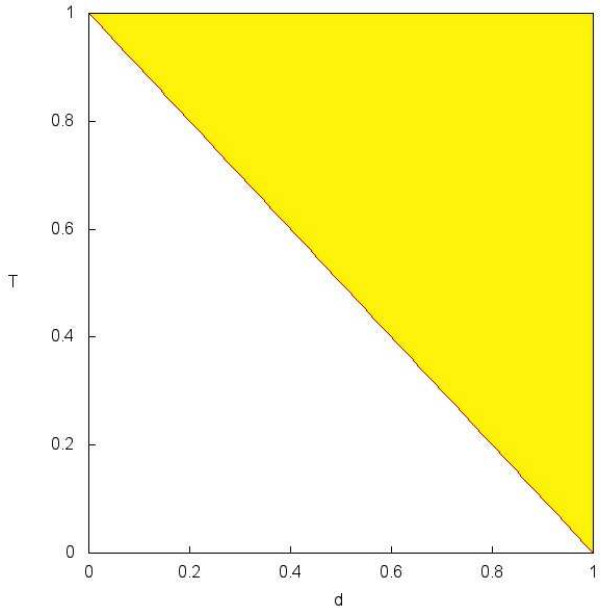

Figure 2.8: (Color on line) In the parameter space (T, d) we highlighted the region where the paramagnetic state exists and is stable. As proved in Sec. 2.4.1, this region includes points fulfilling $T < 1 - d$; notice that this result is independent of P.

2.4.2 Pure State

Let us now consider the pure state, that is any of the P configurations

$$m^\mu = \delta_{\mu\nu}, \tag{2.92}$$

m being the extent of the overlap, which, in general, depends on d and on T. The related self-consistence equations are

$$\begin{aligned} m^\mu &= (1-d)\tanh(\beta m^\mu), & (2.93)\\ m^{\nu\neq\mu} &= 0. & (2.94) \end{aligned}$$

The first equation has solution in the whole half-plane $T > 1-d$, and this ensures that, in the same region, the pure-state exists. In order to check its stability, we calculate the stability matrix finding

$$\begin{aligned} A^{\mu\nu} &= 0 \vee \mu \neq \nu & (2.95)\\ A^{\mu\mu} &= 1-\beta(1-d)[1-\tanh^2(\beta m^\mu)] & (2.96)\\ A^{\nu\nu} &= 1-\beta(1-d)[1-(1-d)\tanh^2(\beta m^\mu)]. & (2.97) \end{aligned}$$

Therefore $\mathbf{A}$ is diagonal and the eigenvalues are $E_\mu = A^{\mu\mu}$ and $E_\nu = A^{\nu\nu}$. Notice that these eigenvalues do not depend on P and that $E_\mu \geq E_\nu$, so that the analysis can be restricted on E_ν. Requiring the positivity for E_ν, we get the region in the plane (T, d), where the pure state is stable; such a region is shown in Fig.(2.9). We stress that this result is universal with respect to P (in the low-storage regime).

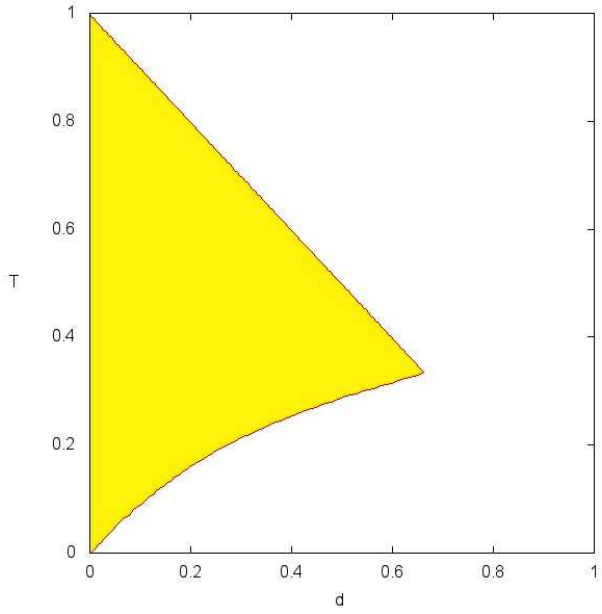

Figure 2.9: In the parameter space (T, d) we highlighted the region where the pure state exists and is stable. This result was found by numerically solving the self-consistence equation Eq.(2.35) and the inequality $E_\nu > 0$, where E_ν is the smallest eigenvalues of the stability matrix $\mathbf{A}$ (see Eq.(2.97)); notice that this result is independent of P.

2.4.3 Symmetric State

A symmetric mixture of states corresponds to configurations leading to

$$m^\mu = m(d,T)\,\forall \mu \in [1,p] m^\mu = 0\,\forall \mu \in [p+1,P] \tag{2.98}$$

where $p \leq P$ order parameters are equivalent and non null, while the remaining $P-p$ are vanishing.

Let us start with the case $p = P = 3$, yielding $m = m(d,T)(1,1,1)$. In this special case the three self-consistence equations collapse on

$$\begin{aligned} m(d,T) = 2\left(\frac{1-d}{2}\right)^3 \left[\tanh^2(3\beta m) + \tanh^2(\beta m)\right] + \\ + d\left(\frac{1-d}{2}\right)^2 \tanh^2(2\beta m) + 2\left(\frac{1-d}{2}\right) d^2 \tanh^2(\beta m) \end{aligned} \tag{2.99}$$

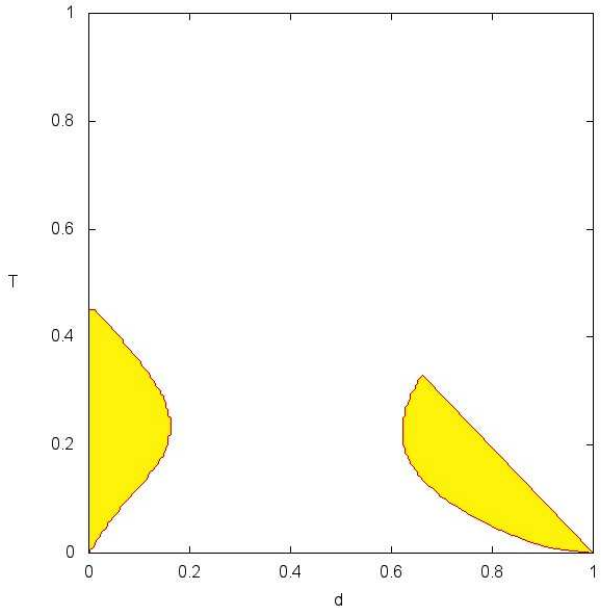

Figure 2.10: In the parameter space (T, d) we highlighted the region where the symmetric state $\sigma^{(S)}$, for the special case $p = P = 3$, exists and is stable. Notice that two disconnected regions emerge: the one corresponding to lower values of dilution derives from the fact that p is odd, while the one corresponding to larger values of dilution from the fact that $p = P$.

and the matrix $\mathbf{A}$ reads as

$$\begin{pmatrix} a & b & b \\ b & a & b \\ b & b & a \end{pmatrix} \tag{2.100}$$

a and b being parameters related to m, d and β. More precisely, the eigenvalues of $\mathbf{A}$ are $(a + 2b, a - b, a - b)$, which can be written as

$$a - b = 1 - \beta(1 - d) + 2\beta \Big\{ \tanh^2(2\beta m) d \left(\frac{1-d}{2}\right)^2 + \tag{2.101}$$

$$+ \tanh^2(\beta m) \left[\frac{d^2(1-d)}{2} + \right. \tag{2.102}$$

$$\left. + 4 \left(\frac{1-d}{2}\right)^3 \right] \Big\},$$

$$a + 2b = 1 - \beta(1-d) + 2\beta\Big\{ \tanh^2(3\beta m) 3 \left(\frac{1-d}{2}\right)^3 + \tag{2.103}$$

$$+ \tanh^2(\beta m)\left[\frac{d^2(1-d)}{2} + \left(\frac{1-d}{2}\right)^3\right]\Big\} + \tag{2.104}$$

$$+ 8d\beta \tanh^2(2\beta m)\left(\frac{1-d}{2}\right)^2 .$$

The conditions for the existence and the stability of the symmetric, odd mixture with $p = P = 3$, yield a system of equations which was solved numerically and the region were such conditions are all fulfilled is shown in Fig.(2.4.3). Notice that the region is actually made up of two disconnected parts, each displaying peculiar features, as explained later.
This result is robust with respect to P, being P odd and $p = P$.

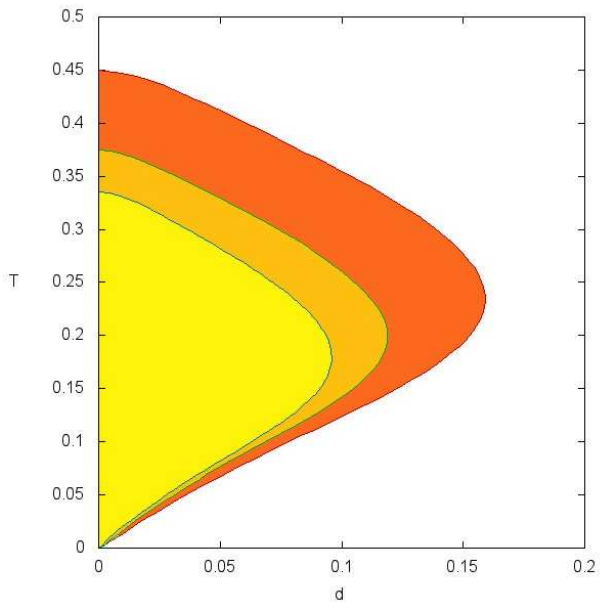

Figure 2.11: In this plot we focused on the region of the parameter space (T, d), where odd symmetric spurious state exist and are stable. In particular, we chose $P = 7$ and we considered any possible odd mixture, i.e. $p = 3$, $p = 5$ and $p = 7$; each value of p is represented by a different curve. Notice that the smaller p and the wider the region, analogously to the standard Hopfield model.

We can further generalize the analysis by considering $P > p$, still being

p odd. In this case we get the following stability matrix

$$\begin{pmatrix} a & b & b & 0 \\ b & a & b & 0 \\ b & b & a & 0 \\ 0 & 0 & 0 & c \end{pmatrix} \tag{2.105}$$

with eigenvalues $(a-b, a-b, a+2b, c)$, where

$$\begin{aligned} c &= 1-\beta(1-d) \\ &\times \Big\{1-2\Big[\left(\frac{1-d}{2}\right)^3[\tanh^2(3m)+3\tanh^2(m)]+d\left(\frac{1-d}{2}\right)^2 \\ &\times 3\tanh^2(2m)+3\frac{1-d}{2}d^2\tanh^2(m)\Big] \\ &\times \Big[1-2\left(\frac{1-d}{2}\right)^3[\tanh^2(3\beta m)+3\tanh^2(\beta m)] \\ &+ 3d\left(\frac{1-d}{2}\right)^2\tanh^2(2\beta m)+3\frac{1-d}{2}d^2\tanh^2(\beta m)\Big]\Big\} \end{aligned} \tag{2.106}$$

has degeneracy $P-p$.

Such states ($p<P$, p odd) are stable only at small d. This is due to the fact that the eigenvalue c occurs only when $p<P$ and it reads as ($\mu>p$):

$$\begin{aligned} A^{\mu\mu} &= [1-\beta(1-d)]+\beta\langle(\xi^\mu)^2\rangle_\xi\langle\tanh^2[\beta m\sum_\nu^p\xi^\nu]\rangle_\xi \\ &= [1-\beta(1-d)]+\beta(1-d)\langle\tanh^2[\beta m\sum_\nu^p\xi^\nu]\rangle_\xi. \end{aligned} \tag{2.107}$$

Thus, one can see that the r.h.s term contains factors $(1-d)$ at least of second order in such a way that when d is close to 1, i.e. for high dilution, and $T<1-d$, such term becomes negative. On the other hand, in the case $\mu\leq p$, we get

$$A^{\mu\mu}=[1-\beta(1-d)]+\beta\langle(\xi^\mu)^2\tanh^2[\beta m\sum_{\nu=1}^p\xi^\nu]\rangle_\xi$$

and therefore the r.h.s term contains even first order term $(1-d)$, which are comparable with $\beta(1-d)$.

Moreover, we find that the p-component, odd symmetric state exists and is stable in a region of the space (T,d), which gets smaller and smaller as p

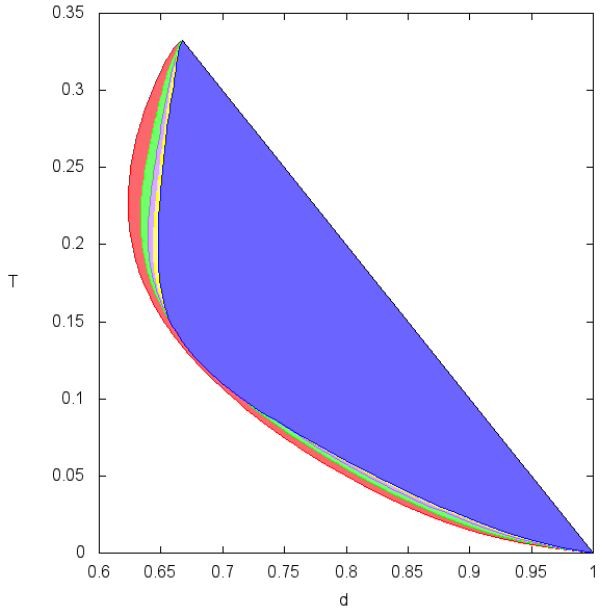

Figure 2.12: In this plot we focused on the region of the parameter space (T, d), where symmetric spurious state with $p = P$ exist and are stable. In particular, we chose $P = 7$ and we considered any possible mixture, i.e. $p = 3$, $p = 4$, $p = 5$, $p = 6$ and $p = 7$; each value of p is represented by a different curve. Notice that the smaller p and the wider the region, yet the region tends to an "asymptotic shape".

grows (see Fig.(2.11)). The emergence of such states can be seen as a feature of robustness of the standard Hopfield model with respect to dilution.

Finally, the case $P = p$ always admits a region of existence and stability in the regime of high dilution. The latter region is independent of the parity and depends slightly on P (see Fig.(2.12)). The emergence of such states is due to the failure of hierarchical retrieval, namely uniformity prevails.

2.4.4 Parallel State

The parallel-retrieval state can be looked at as the extension to arbitrary values of d of the pure state holding for the special case $d = 0$. We recall that in the noiseless limit the parallel-retrieval state can be described as

$$m^{\mu} = (1 - d)d^{\mu-1}. \tag{2.108}$$

In this case the stability matrix is diagonal with terms:

$$A^{\mu\mu} = 1-\beta(1-d)+\beta\langle(\xi^{\mu})^2\tanh^2[\beta(1-d)(\xi^1+d\xi^2+...+d^P\xi^P)]\rangle, \quad (2.109)$$

and, consistently, taking the limit $\beta\to\infty$, we get the simplified form

$$A^{\mu\mu} = \lim_{\beta\to\infty} = 1-\beta(1-d)+\beta\langle(\xi^{\mu})^2(1-\delta[(\xi^1+d\xi^2+...+d^P\xi^P)])\rangle. \quad (2.110)$$

Now, the third term in the r.h.s. is either $\beta\langle(\xi^{\mu})^2\rangle = \beta(1-d)$ (when the polynomial of order P is zero) or 0; the latter case would trivially yield $A^{\mu\mu} < 0$. Therefore, in the limit $\beta\to\infty$ the stability of the parallel-retrieval state is constrained by the smallest real root $\in [0,1]$ of the polynomial $\xi^1 + d\xi^2 + ... + d^P\xi^P$ with $\xi^i = 1, 0, -1$. This corresponds to $\xi^1 = 1$ and $\xi^i = -1, \forall i > 1$, under gauge symmetry and returns the same result found, from a more empirical point of view, in Sec. 2.3. More precisely, the critical dilution converges exponentially to 1/2 as P grows.

In particular, for $P = 3$ we find that the parallel-retrieval state exists and is stable in the interval $d \in (0, \frac{\sqrt{5}-1}{2}) \simeq (0, 0.618)$. The point $d_c(3) = \frac{\sqrt{5}-1}{2}$ corresponds to the unique real root in $(0,1)$.

When noise is introduced, the critical dilution d_c, separating the parallel-retrieval state from spurious states, is shifted towards larger values, as suggested by Eq.(2.109). On the opposite side, namely in the regime of small dilution, the parallel state is progressively depleted and, as the temperature is increased, magnetizations vanish, starting from m_P, and proceeding up to m_2. One can distinguish a set of temperatures $T_P(d) < T_{P-1}(d) < ... < T_2(d) < T_1(d)$, such that when $T > T_k(d)$, all magnetizations $m_i, \forall i \leq k$ are null on average. Hence, above $T_2(d)$ the pure state retrieval is recovered, while above $T_1(d) = 1-d$ the paramagnetic state emerges.

In Fig.(2.13) we highlight the region of the parameter space (T, d) where such parallel states exist and are stable. This was obtained numerically for the case $P = 5$; for larger values of P the region is slightly restricted to account for the shift in d_c.

2.5 Monte Carlo Simulations

In this Section we discuss details on Monte Carlo simulations.
All the simulations were performed on a system Ubuntu Linux with Intel Core I7, 3.2Ghz, 12 CPU, Nvidia-Fermi technology, 12 Gb RAM and OpenMP libraries. The simulations were carried out sequentially according to the following algorithm:

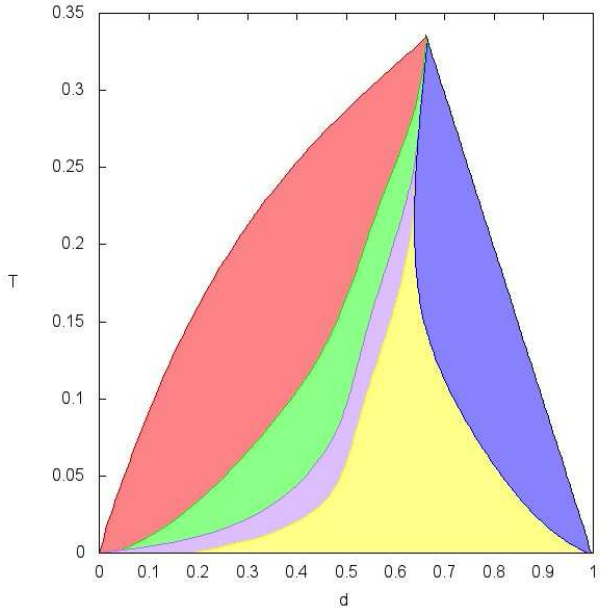

Figure 2.13: In this plot we focused on the region of the parameter space (T, d), where parallel retrieval states exist and are stable. In particular, we chose $P = 5$ and we considered any possible state with $k = 2$, $k = 3$, $k = 4$ and $k = 5$ non-null magnetization.

1. Building and storing of the coupling matrix.

   First, we generate P patterns according to the distribution ($d = 0$):

   $$P(\xi_i^\mu) = \frac{1-d}{2}\delta_{(\xi_i^\mu - 1)} + \frac{1-d}{2}\delta_{(\xi_i^\mu + 1)} + d\delta_{(\xi_i^\mu)}, \tag{2.111}$$

   then, we build a char-matrix $J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$ with entries ranging $\in [0, 2P+1]$ and acting as key pointing to another hash-matrix $\tilde{J}_{ij}$ where the $N(N-1)/2$ real numbers accounting for the Hebb interactions are stored. If the amount of patterns do not exceed $P = 256$, i.e. one byte, it is then possible to account for 10^5 spins with no need of swapping on hard disk (which would sensibly affect the performance of the simulation). This condition is fulfilled for the low storage regime we are interested in.

2. Initialize the network status.

   We checked the two standard approaches: The first is to initialize the

network in a (assumed) fixed point of the dynamics, namely

$$\sigma_i = \xi_i^1 \quad \forall i \in [1, ..., N], \tag{2.112}$$

and check its evolution: This gives information on the structure of the basins of attraction of the minima as we vary the dilution (see Point 5).

The second approach is to initialize the network randomly: We set $\sigma_i = 1$ with probability 0.5 and $\sigma_i = -1$ otherwise. This is a standard procedure to follow the relaxation to a fixed point with no initial assumption and gives information on the structure of the basins of attraction of the minima at fixed dilution.

3. Evolution dynamics

   The spin status evolves according to a standard (random and sequential) Glauber dynamics for Ising-like systems [17]: At each time interval, the spins state is updated according to its input signals, where the probability of the unit's activity is equal to a rectified value of the input (logit transfer function), i.e.

$$Pr[\sigma_i(t) = \pm 1] = \frac{1}{1 + \exp[\mp 2\beta \sum_j J_{ij}\sigma_j]}. \tag{2.113}$$

   The field-updating process is managed by a linked list whose parsing is parallelized through OpenMP.

4. Convergence of the simulation.

   Due to the peculiar structure of the fields induced by pattern dilution, the field insisting on a given spin may be zero and the related spin would flip indefinitely. To avoid this pathological situation we skip the updating of these "paramagnetic" spins and focus on the remaining ones: In the zero noise limit convergence is almost immediate, such that when the whole ensemble of spins remains unchanged for the whole N-length of the update cycle, dynamics is stopped and the resulting P pattern overlaps are printed on a file.

   Relaxation at non-zero noise is checked through the linked list (see next step): The pointer of each spin that is aligned with its own field is stored, the ones of spins with no net fields are removed from the linked list, while all the other spins, mismatched to their own fields, are added into the linked list.

5. Making the P patterns sparser.

   There can be two deeply different ways of increasing dilution. The former is a Bernoullian approach and essentially if one starts from a dilution $d = 0.45$ toward a dilution $d = 0.5$ (just as a concrete example) may forget the starting information and generate a random pattern with on average one half of zero entries; the latter is a Markovian dilution by which one needs to start from the previous coupling matrix (and patterns) diluted at $d = 0.45$ and increases dilution on that structure.
   Dilution is tuned at steps of 0.01, ranging from $d = 0$ to $d = 1$.
   We take as the state of the network the last equilibrium state, then go to point (3).

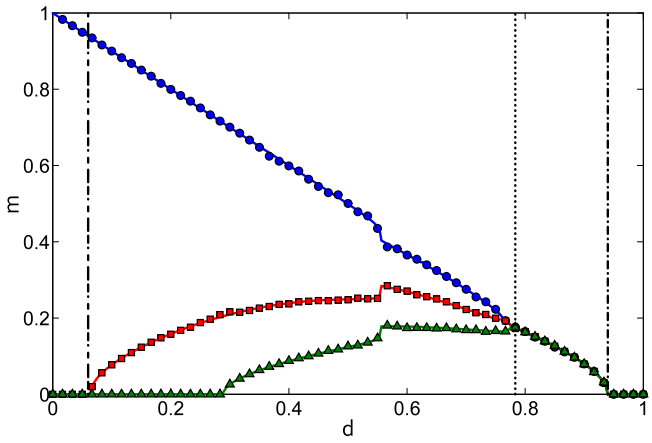

Figure 2.14: Data from Monte Carlo simulations (symbols) and analytical predictions (solid lines) obtained for a system with $P = 3$ patterns and set at a temperature $T = 0.06$ are compared. Simulations are performed on a set of 10^5 spins. The dashed line at $d \approx 0.06$ marks the boundary of the pure state regime; the dotted line at $d \approx 0.78$ marks the onset of the symmetric phase; the semi-dashed line at $d \approx 0.94$ marks the onset of the paramgnetic phase.

Through Markovian dilution, we can follow the evolution of the pure Hopfield attractors while tuning d. In general, the results obtained via numerical simulations are in perfect agreement with the theory: This point is not surprising, as, due to the load storage regime, $\lim_{N\to\infty} P/N = 0$, hence replica symmetry is never broken and our solution is the real solution of the model (no approximations have been made).

Chapter 3

Hierarchical Structures

In the last decade some steps forward toward *more realistic* systems have been achieved merging statistical mechanics [42,59,65] and graph theory [17,19]. In particular, mathematical methodologies were developed to deal with spin systems embedded in random graphs, where the ideal, full homogeneity among spins is lost [23, 24]. Thus, networks of spins arranged according to Erdös-Rényi [26], small-world [25], or scale-free [47] topologies were addressed, yet finite-dimensional networks were still out of debate.

Focusing on neural networks, it should be noted that, beyond the difficulty of treating non-trivial topologies for spin architectures, one has also to cope with the complexity of their coupling pattern, meant to encode the Hebbian learning rule. The emerging statistical mechanics is much trickier than that for ferromagnets; indeed neural networks can behave either as ferromagnets or as spin-glasses, according to the parameter settings: their phase space is split into several disconnected pure states, each coding for a particular stored pattern, so to interpret the thermalization of the system within a particular energy valley as the spontaneous retrieval of the stored pattern associated to that valley. However in the high-storage limit, where the amount of patterns scales linearly with the number of spins, neural networks approach pure spin-glasses (loosing retrieval capabilities at the blackout catastrophe [17]) and, as a simple Central Limit argument shows [4], when the amount of patterns diverge faster that the amount of spins they become purely spin glasses. For the sake of exhaustiveness we also stress that, even in the retrieval region, neural networks are *exactly* linear combinations of two-party spin glasses [2,3]: due to the combination of such difficulties, neural networks on a finite dimensional topology have not been extensively investigated so far.

However, very recently, a non-mean-field model, where a topological distance among spins can be defined and couplings can be accordingly rescaled, turned out to be, to some extent, treatable also for complex systems such

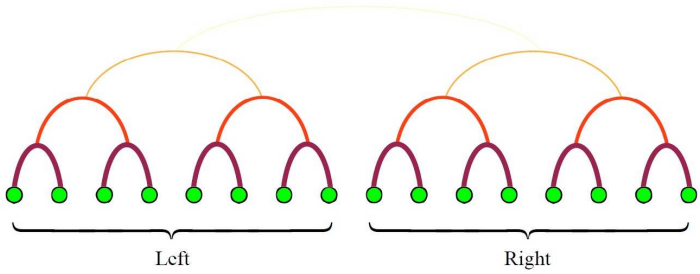

Figure 3.1: Schematic representation of the hierarchical topology, that underlies the system under study: green spots represent nodes where spins live, while different colors and thickness for the links mimic different intensities in their mutual interactions: the brighter and thinner the link, the smaller the related coupling.

as spin-glasses [15, 56]. More precisely, spins are arranged according to a hierarchical architecture as shown in Fig.(3.1): each pair of nearest-neighbor spins form a "dimer" connected with the strongest coupling, then spins belonging to nearest "dimers" interact each other with a weaker coupling and so on recursively. In particular, the Sherrington-Kirkpatrick model for spin-glasses defined on the hierarchical topology has been investigated in [55]: despite a full analytic formulation of its solution still lacks, renormalization techniques, [14, 56], rigorous bounds on its free-energies [54] and extensive numerics [38, 39] can be achieved nowadays and they give extremely sharps hints on the thermodynamic behavior of systems defined on these peculiar topologies.

Remarkably, as we are going to show, when implementing the Hebb prescription for learning on these hierarchical networks, an impressive phase diagram, much richer than the mean-field counterpart, emerges. More precisely, spins turn out to be able to orchestrate both serial processing (namely sharp and extensive retrieval of a pattern of information), as well as parallel processing (namely retrieval of different patterns simultaneously).

The remaining of the chapter is structured as follows: in the next subsections we provide a streamlined description of mean-field serial and parallel processors, and we introduce the hierarchical scenario. Then, we split in three sections our findings according to the methods exploited for investigation: statistical mechanics, signal-to-noise technique and extensive numerical simulations. All these approaches consistently converge to the scenario outlined above. Seeking for clarity and completeness, each technique is first applied to a ferromagnetic hierarchical mode (which can be thought of as

a trivial one-pattern neural network and acts as a test-case) and then for a low-storage hierarchical Hopfield model.

3.1 The Network on a Hierarchical Topology.

We now start our investigation of a neural network embedded in the hierarchical topology depicted in Fig.(3.1). As mentioned, two main difficulties are interplaying: the complexity of the emergent energy landscape (essentially due to frustration in the coupling pattern) and the non-mean-field nature of the model (essentially due to the inhomogeneity of the network architecture). It is therefore safer to proceed by steps discussing first the hierarchical ferromagnet (hence retaining only the second difficulty), known as Dyson hierarchical model (DHM). Then, via the Mattis gauge we reach a Mattis hierarchical model (MHN) and finally we extend to the Hopfield hierarchical model (HHM).

The Dyson hierarchical model [37] is a system made of N binary (Ising) spins $S_i = \pm 1$, $i = 1, ..., N$ in mutual interaction and built recursively in such a way that the system at the $(k+1)$-th iteration contains $N = 2^{k+1}$ spins and is obtained by taking two replicas of the system at the k-th iteration (each made of 2^k spins) and connecting all possible couples with overall $\binom{N}{2}$ couplings equal to $-J/2^{\sigma(k+1)}$, J and σ being real scalars tuning the interaction strength: the former acts uniformly over the network, the latter triggers the decay with the "distance" among spins. The resulting Hamiltonian can be written recursively as

$$H_{k+1}^{\text{Dyson}}(S|J,\sigma) = H_k^{\text{Dyson}}(S_1|J,\sigma) + H_k^{\text{Dyson}}(S_2|J,\sigma) - \frac{J}{2^{2\sigma(k+1)}} \sum_{i<j}^{2^{k+1}} S_i S_j, \quad (3.1)$$

where $\mathbf{S}_1 = \{S_i\}_{i=1}^{2^k}$ and $\mathbf{S}_2 = \{S_j\}_{i=2^k+1}^{2^{k+1}}$, while $H_0^{\text{Dyson}} \equiv 0$.

Before proceeding it is worth stressing that the parameters J and σ are bounded as $J > 0$ and $\sigma \in (\frac{1}{2}, 1)$: the former trivially arises from the ferromagnetic nature of the model which makes neighboring spin to "imitate" each other, while the latter can be understood by noticing that for $\sigma > 1$ the interaction energy goes to zero in the thermodynamic limit[1], while for $\sigma < \frac{1}{2}$ the interaction energy is no longer linearly-additive implying thermodynamic

[1]The sum $\sum_{i<j}^{2^{k+1}}$ brings a contribution scaling like $2^{2(k+1)} \sim N^2$, while the pre-factor scales as $2^{-2\sigma(k+1)} \sim N^{-2\sigma}$, thus, when $\sigma > 1$ the internal energy (the thermodynamical expectation of the Hamiltonian normalized over the system size) is overall vanishing in the thermodynamic limit $k \to \infty$.

instability[2]. Moreover, this model is intrinsically *non-mean-field* because a notion of metrics, or distance, has been implicitly introduced: two nodes are said to be at distance d if they get first connected at the d-th iteration. In general, calling d_{ij} the *distance* between the spins $i,\ j$, (thus $d_{ij} = 1, ..., k+1$), we can associate to each couple a distant-dependent coupling J_{ij} and rewrite Eq.(3.1) in a more familiar form as

$$H_{k+1}^{\text{Dyson}}(S|J,\sigma) = -\sum_{i<j} J_{ij} S_i S_j, \tag{3.2}$$

where

$$J_{ij} = \sum_{l=d_{ij}}^{k+1} \frac{J}{2^{2\sigma l}} = J\frac{4^{\sigma - d_{ij}\sigma} - 4^{-k\sigma-\sigma}}{4^\sigma - 1}. \tag{3.3}$$

The next step is to gauge the spins *à* la Mattis, namely, once extracted quenched values for the pattern entries $(\xi_i^\mu)_{\mu=1}$ from the distribution

$$P(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1), \tag{3.4}$$

we replace S_i with $\xi^1 S_i$. This results in the following hierarchical Mattis model

$$H_{k+1}^{\text{Mattis}}(S|J,\sigma) = -\sum_{i<j} J_{ij}\xi_i^1\xi_j^1 S_i S_j. \tag{3.5}$$

Finally, summing over p patterns, we obtain the Hopfield hierarchical model (HHM) that reads as (for $J = 1$)

$$\begin{aligned} H_{k+1}^{\text{Hopfield}}(S|\xi,\sigma) &= H_k^{\text{Hopfield}}(S_1|\xi,\sigma) + H_k^{\text{Hopfield}}(S_2|\xi,\sigma) \\ &- \frac{1}{2}\frac{1}{2^{2\sigma(k+1)}}\sum_{\mu=1}^{p}\sum_{i,j=1}^{2^{k+1}} \xi_i^\mu\xi_j^\mu S_i S_j, \end{aligned} \tag{3.6}$$

with $H_0^{\text{Hopfield}} \equiv 0$ and σ still within the previous bounds, i.e. $\sigma \in (\frac{1}{2}, 1)$. As anticipated, here we restrict the analysis to low storage limit only: recalling $N = 2^{k+1}$, we can fix p finite at first so to move straightforwardly from the DHM to the HHM (as the notion of distance is preserved) and, posing

$$J_{ij} = \frac{4^{\sigma - d_{ij}\sigma} - 4^{-k\sigma-\sigma}}{4^\sigma - 1}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu, \tag{3.7}$$

[2]The sum $\sum_{i<j}^{2^{k+1}}$ brings a contribution scaling like $2^{2(k+1)} \sim N^2$, while the pre-factor scales as $2^{-2\sigma(k+1)} \sim N^{-2\sigma}$, thus, when $\sigma < \frac{1}{2}$ the intensive energy is overall divergent in the thermodynamic limit $k \to \infty$.

we can write equivalently the Hamiltonian 3.6 in the more compact form

$$H_{k+1}^{\text{Hopfield}}(S|\xi,\sigma) = -\sum_{i<j}^{2^{k+1}} J_{ij} S_i S_j. \tag{3.8}$$

Thus in the HHM the Hebbian prescription is coupled with a function of the spin's distance.

3.2 Insights From Statistical Mechanics

Here we summarize findings that can be achieved by suitably extending interpolation techniques [35,36] beyond the mean-field paradigm: it is important to stress once more that, as this strand gives only (not-mean-field) bounds on the free energy (and not the full solution), the self-consistencies that result are not the true self-consistencies of the model.

3.2.1 Free Energies in the Dyson Model

As the Hamiltonian $H_{k+1}(S|J,\sigma)$ is given (see Eq.(3.1)) and the noise level $\beta^{-1} = T$ (where T stands for *noise* for historical reasons) introduced, it is possible to define the partition function $Z_{k+1}(\beta, J, \sigma)$ at finite volume $k+1$ as

$$Z_{k+1}(\beta, J, \sigma) = \sum_{\{S\}} \exp\left[-\beta H_{k+1}(S|J,\sigma)\right], \tag{3.9}$$

and the related free energy $\alpha_{k+1}(\beta, J, \sigma)$, namely the intensive logarithm of the partition function, as

$$\alpha_{k+1}(\beta, J, \sigma) = \frac{1}{2^{k+1}} \log \sum_{\{S\}} \exp\left[-\beta H_{k+1}(\vec{S}) + h \sum_{i=1}^{2^{k+1}} S_i\right], \tag{3.10}$$

where the sum runs over all possible $2^{2^{k+1}}$ spin configurations. Note that the usual free energy f is related to α by $f(\beta) = -\beta\alpha(\beta)$, hence we will find thermodynamic equilibria checking the maxima of $\alpha(\beta)$ and not the minima. We are interested in an explicit expression of the infinite volume limit of the intensive free energy, defined as

$$\alpha(\beta, J, \sigma) = \lim_{k\to\infty} \alpha_{k+1}(\beta, J, \sigma), \tag{3.11}$$

in terms of suitably introduced magnetizations m, that act as order parameters for the theory. In fact, as the free energy is just the difference between

the internal energy E of the system (i.e. the mean-value of the Hamiltonian) weighted by β, and the entropy S, namely $\alpha(\beta, J, \sigma) = -\beta E(\beta, J, \sigma) + S(\beta, J, \sigma)$, extremization of the free-energy over the order parameters equals to imposing thermodynamic prescriptions (i.e. minimum energy and maximum entropy principles) and therefore allows us to get a description of the thermodynamic equilibria of the system in terms of the self-consistencies for these m's.
To this task we introduce the global magnetization m, defined as the limit $m = \lim_{k\to\infty} m_{k+1}$ where

$$m_{k+1} = \frac{1}{2^{k+1}} \sum_{i=1}^{2^{k+1}} S_i, \tag{3.12}$$

and, recursively and with a little abuse of notation, level by level (over k levels) the k magnetizations $\vec{m}_a, ..., \vec{m}_k$, as the same $k \to \infty$ limit of the following quantities (we write explicitly only the two upper magnetizations related to the two main clusters *left* and *right* -see Fig.(3.1):

$$m_k^1 = \frac{1}{2^k} \sum_{i=1}^{2^k} S_i, \quad m_k^2 = \frac{1}{2^k} \sum_{i=2^k+1}^{2^{k+1}} S_i, \tag{3.13}$$

and so on. The *thermodynamical averages* are denoted by the brackets $\langle\cdot\rangle$ such that, e.g. for the observable $m_{k+1}(\beta, J, \sigma)$, we can write

$$\langle m_{k+1}(\beta, J, \sigma)\rangle = \frac{\sum_{\{\sigma\}} m_{k+1} e^{-\beta H_{k+1}(\vec{S}|J,\sigma)}}{Z_{k+1}(\beta, J, \sigma)}, \tag{3.14}$$

and clearly $\langle m(\beta, J, \sigma)\rangle = \lim_{k\to\infty}\langle m_{k+1}(\beta, J, \sigma)\rangle$.
Starting with the pure ferromagnetic case, which mirrors here the serial retrieval of a single pattern in the Hopfield counterpart, its free energy can be bounded as (see also [54])

$$\alpha(h, \beta, J, \sigma) \geq \sup_m \{\log 2 + \log\cosh\Big[h + \beta m J(C_{2\sigma-1} + \tag{3.15}$$

$$- C_{2\sigma})\Big] - \frac{\beta J}{2}(C_{2\sigma-1} - C_{2\sigma})m^2\}, \tag{3.16}$$

where

$$C_{2\sigma} = \frac{1}{2^{2\sigma} - 1}, \tag{3.17}$$

$$C_{2\sigma-1} = \frac{1}{2^{2\sigma+1} - 1}. \tag{3.18}$$

Now let us suppose that, instead of a global ordering, the system can be effectively split in two parts (the two largest communities called *left* and *right* in Fig.(3.1)), with two different magnetizations $m_{left} = m_1$ and $m_{right} = m_2$; we also assume $m_{left} = -m_{right}$. Through the interpolative route we approach a bound for the free energy related to such a mixed state. We stress the fact that the upper link, connecting the two communities with opposite magnetization, remains and it gives a contribute m in the system as (see also [28])

$$\begin{aligned}
&\alpha_{k+1} \geq \\
&\frac{1}{2}\log\cosh\left\{h+\beta J\left[m(2^{(k+1)(1-2\sigma)})+m_1\left(\sum_{l=1}^{k}2^{l(1-2\sigma)}-\sum_{l=1}^{k+1}2^{-2l\sigma}\right)\right]\right\} \\
&+\frac{1}{2}\log\cosh\left\{h+\beta J\left[m(2^{(k+1)(1-2\sigma)})+m_2\left(\sum_{l=1}^{k}2^{l(1-2\sigma)}-\sum_{l=1}^{k+1}2^{-2l\sigma}\right)\right]\right\} \\
&-\frac{\beta J}{2}\left[\left(\sum_{l=1}^{k}2^{l(1-2\sigma)}-\sum_{l=1}^{k+1}2^{-2l\sigma}\right)\left(\frac{m_1^2+m_2^2}{2}\right)-2^{(k+1)(1-2\sigma)}m^2\right] \\
&+\log 2. \qquad (3.19)
\end{aligned}$$

Notice that, thanks to the gauge simmetry $S_i \to -S_i$, the state considered mirrors the parallel retrieval of two patterns in the Hopfield counterpart.

Identifying $m_1 = m_2 = m$ we recover the previous bound as expected, and, quite remarkably, in the thermodynamic limit the two free energies assume the same values, thus serial and parallel retrieval are both equally accomplished by the network. Imposing thermodynamic stability we obtain the following self-consistencies

$$m_{1,2} = \tanh(h + \beta J m_{1,2}(C_{2\sigma-1} - C_{2\sigma})), \qquad (3.20)$$

whose behavior is depicted in Fig.(3.2).

3.2.2 Serial/Parallel Retrieval in Hopfield Hierarchical Model

Guided by the ferromagnetic model just described, we now turn to the hierarchical Hopfield model (HHM) and start its analysis from a statistical mechanical perspective, namely we infer the thermodynamic behavior of a

system described by the following recursive Hamiltonian

$$\begin{aligned} H_{k+1}^{Hopfield}(S|\xi,\sigma) &= H_k^{Hopfield}(S_1|\xi,\sigma) + H_k^{Hopfield}(S_2|\xi,\sigma) \\ &- \frac{1}{2}\frac{1}{2^{2\sigma(k+1)}}\sum_{\mu=1}^{p}\sum_{i,j=1}^{2^{k+1}}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j. \end{aligned} \tag{3.21}$$

To this task, we introduce suitably p Mattis magnetizations (or Mattis overlaps), over the whole system, as

$$m^\mu = \frac{1}{2^{k+1}}\sum_{i=1}^{2^{k+1}}\xi_i^\mu S_i,\ \ \mu\in[1,p]. \tag{3.22}$$

Even in this context, the definition above can account for the state of inner clusters by the sum over the (pertinent) spins. For instance, focusing on the two larger communities we have the $2p$ Mattis magnetizations

$$m_{left}^\mu = \frac{1}{2^k}\sum_{i=1}^{2^k}\xi_i^\mu S_i, \qquad m_{right}^\mu = \frac{1}{2^k}\sum_{i=2^k+1}^{2^{k+1}}\xi_i^\mu S_i, \tag{3.23}$$

with $\mu\in[1,p]$. Again, we will not enter in the mathematical details concerning non-mean-field bounds for the model free energy (as they can be found in [28]), while we streamline directly the physical results.
Still mirroring the previous section, we are interested in obtaining a bound limiting the free energy of the HHM, the latter being defined as the $k\to\infty$ limit of α_{k+1}, whose expression reads

$$\alpha_{k+1}(\beta,\{h_\mu\},\sigma) = \frac{1}{2^{k+1}}\log\sum_{\{S\}}\exp\left[-\beta H_{k+1}(\vec{S}) + \sum_{\mu=1}^{p}h^\mu\sum_{i=1}^{2^{k+1}}S_i\right], \tag{3.24}$$

where we accounted also for p external stimuli h^μ.
The non-mean field bound for serial processing free energy reads as

$$\begin{aligned} \alpha(\beta,\{h^\mu\},p) &\geq \sup_m[\log 2 + \Big\langle \log\cosh\Big(\sum_{\mu=1}^{p}\Big[h^\mu + \beta m^\mu(C_{2\sigma-1}-C_{2\sigma})\Big]\xi^\mu\Big)\Big\rangle_\xi \\ &- \frac{\beta}{2}\sum_{\mu=1}^{p}\langle(m^\mu)^2\rangle_\xi(C_{2\sigma-1}-C_{2\sigma})], \end{aligned} \tag{3.25}$$

with optimal order parameters fulfilling

$$\langle m^\mu\rangle_\xi = \langle\xi^\mu\tanh[\beta\sum_{\nu=1}^{p}[h^\nu + (C_{2\sigma-1}-C_{2\sigma})m^\nu]\,\xi^\nu]\rangle_\xi,$$

and whose critical noise is $\beta_c^{NMF} = C_{2\sigma-1} - C_{2\sigma}$, where the index NMF stresses that the estimate was obtained through a non mean field bound of the free energy.
Of course we can assume again that the two different families of Mattis magnetizations ($\{m_{1,2}^{\mu}\}_{\mu=1}^{p}$) (those playing for the two inner blocks of spins *left* and *right* lying under the $k+1$-th level) behave independently as the higher links connecting them go to zero quickly for $k \to \infty$ and we can start the interpolative machine: following this way we generalize the serial processing analysis to a two-pattern parallel retrieval analysis, which results in the following bound for the related free energy:

$$\alpha(\beta, \{h_\mu\}, p) \geq \sup_{\{m_{1,2}^{\mu}\}} [\log 2 + \frac{1}{2}\Big\langle \log\cosh\Big\{\sum_{\mu=1}^{p}\Big[h^\mu + \beta m_1^\mu\Big(\sum_{l=1}^{k} 2^{l(1-2\sigma)}$$
$$-\sum_{l=1}^{k} 2^{l(-2\sigma)}\Big) + \beta m^\mu 2^{(k+1)(1-2\sigma)}\Big]\xi^\mu\Big\}\Big\rangle_\xi + \frac{1}{2}\Big\langle \log\cosh\Big\{\sum_{\mu=1}^{p}\Big[h^\mu + \beta m_2^\mu$$
$$\times[\sum_{l=1}^{k} 2^{l(1-2\sigma)} - \sum_{l=1}^{k} 2^{l(-2\sigma)}] + \beta m^\mu 2^{(k+1)(1-2\sigma)}\Big]\xi^\mu\Big\}\Big\rangle_\xi - \frac{\beta}{2}\Big[\sum_{l=1}^{k} 2^{l(1-2\sigma)}$$
$$-\sum_{l=1}^{k} 2^{l(-2\sigma)}\Big]\cdot\sum_{\mu=1}^{p}\frac{\langle (m_1^\mu)^2\rangle_\xi + \langle (m_2^\mu)^2\rangle_\xi{}^2}{2} - \frac{\beta}{2}2^{(k+1)(1-2\sigma)}\sum_{\mu=1}^{p}\langle (m^\mu)^2\rangle_\xi,$$

Here we do not investigate further the parallel retrieval of larger ensembles of patterns, as the way to proceed is identical to the outlined one, but we simply notice that, if we want the system to handle M patterns, hence we assume it effectively splits M times into sub-clusters until the $k+1-M$ level, then the procedure keeps on working as long as

$$\lim_{k\to\infty}\sum_{l=k+1-M}^{k+1} 2^{l(1-2\sigma)}\sum_{\mu=1}^{p} m_l^\mu = 0. \tag{3.26}$$

Since the magnetizations are bounded, in the worst case we have

$$\begin{aligned}\sum_{l=k+1-M}^{k+1} 2^{l(1-2\sigma)}\sum_{\mu=1}^{p} m_l^\mu &\leq p\sum_{l=k+1-M}^{k+1} 2^{l(1-2\sigma)} \\ &\leq p\sum_{l=k+1-M}^{\infty} 2^{l(1-2\sigma)} \propto 2^{(1-2\sigma)(k+1-M)}p. \end{aligned}\tag{3.27}$$

If we want the system to handle up to p patterns, we need p different blocks of spins and then $M = \log(p)$.

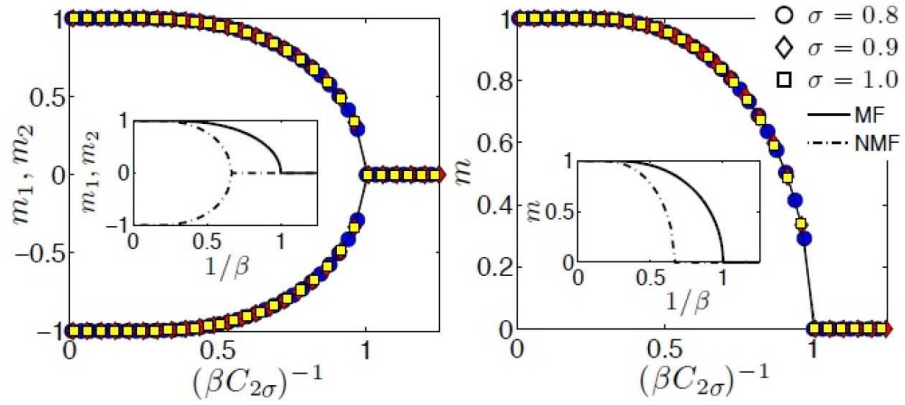

Figure 3.2: Main plots: numerical solutions of the non-mean-field self-consistent equations for the parallel state (left panel) and for the pure state (right panel) of the Dyson model (see Eq.(3.20)) obtained for different values of σ (as explained by the legend) and plotted versus a rescaled noise. Note that by rescaling the noise the dependence on σ is lost and all curves are collapsed. Insets: comparison between the numerical solutions of the non-mean-field self-consistent equations (dashed line) and of the mean-field self-consistent equations (solid line) as a function of the noise and for fixed $\sigma = 1$ (see Eq.(3.20)). Notice that for the Hopfield hierarchical model, numerical solutions for the Mattis magnetizations pertaining to the pure and to the mixed states are the same.

3.3 Insights From Signal-to-Noise Techniques

Results from statistical mechanics gave stringent hints on the network's behavior, however they act as bounds only.
This requires further inspection via other techniques: the first route we exploit is signal-to-noise. Through the latter, beyond generally confirming the predictions obtained via the first path, we obtain sharper statements regarding the evolution of the Mattis order parameters. These two approaches are complementary: while statistical mechanics describes the system with $N \to \infty$ and $\beta < \infty$, with the signal-to-noise technique we inspect the regime $N < \infty$ and $\beta \to \infty$.

3.3.1 A Glance at the Fields in the Dyson Network

Plan of this Section is to look at the dynamically stable configurations of the spins, that is to say, we investigate the configurations (global and lo-

cal minima) that imply each spin S_i to be aligned with its corresponding field $h_i(S)$, i.e. $S_i\, h_i(S) > 0, \forall i$. This approach basically corresponds to a negligible-noise statistical mechanical analysis but it is mathematically much more tractable.
We can rearrange the Dyson Hamiltonian in a useful form for such an investigation as follows

$$H_{k+1}^{\text{Dyson}}(\{S_1...S_{2^{k+1}}\}) = -\frac{J}{2}\sum_{\mu=1}^{k+1}\sum_{i=1}^{2^{k+1}} S_i \left[\sum_{l=\mu}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\right] \sum_{\{j\}:d_{ij}=\mu} S_j, \qquad (3.28)$$

thus, highlighting the field h_i insisting on the spin S_i we can write

$$H_{k+1}^{\text{Dyson}}(\{S_1...S_{2^{k+1}}\}) = -\sum_{i=1}^{2^{k+1}} S_i h_i(S), \qquad (3.29)$$

$$h_i(S) = J\sum_{\mu=1}^{k+1}\left[\sum_{l=\mu}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\right] \sum_{\{j\}:d_{ij}=\mu} S_j. \qquad (3.30)$$

While Glauber dynamics will be discussed in Sec. 4 (dedicated to numerics), we just notice here that the microscopic law governing the evolution of the system can be defined as a stochastic alignment to local field $h_i(S)$.

$$S_i(t+\delta t) = \text{sign}\left\{\tanh\left[\beta h_i\left(S\left(t\right)\right)\right] + \eta_i(t)\right\},$$

where the stochasticity lies in the independent random numbers $\eta_i(t)$, uniformly distributed over the interval $[-1,1]$ and tuned by β. The latter continues to rule the noise level even dynamically as it amplifies, or suppresses, the smoothness of the hyperbolic tangent; in particular, in the noiseless limit $\beta \to \infty$ we get

$$S_i(t+\delta t) = \text{sign}\left[h_i\left(S(t)\right)\right]. \qquad (3.31)$$

This is crucial for checking the stability of a state as, if $S_i h_i(S) > 0\ \forall\ i\ \in [1,N]$, the configuration $\{S\}$ is dynamically stable (at least for $\beta \to \infty$, as in the presence of noise there is a β-dependent probability to fluctuate away).

We keep the previous ensemble of non-independent order parameters m_i^n defined in detail as

$$m_i^n(S) = \frac{1}{2^n}\sum_{j=2^n\times i-(2^n-1)}^{2^n\times i} S_j \quad \text{with } i = 1,2,...,2^{k+1-n} \quad \text{and } n = 0,1,2,...k+1, \qquad (3.32)$$

namely

$$\begin{cases} m_i^0 = S_i & \text{with } i = 1, 2, .., 2^{k+1}, \\ m_i^1 = \frac{1}{2}\sum_{j=2i-1}^{2i} S_j & \text{with } i = 1, 2, .., 2^k \ \rightarrow m_1^1 = \frac{1}{2}\sum_{j=1}^{2} S_j, \\ m_i^2 = \frac{1}{2^2}\sum_{j=2^2 i-(2^2-1)}^{2^2 i} S_j & \text{with } i = 1, 2, .., 2^{k-1} \ \rightarrow m_1^2 = \frac{1}{4}\sum_{j=1}^{4} S_j, \\ ..... & \\ m_1^{k+1} = \frac{1}{2^{k+1}}\sum_{j=1}^{2^{k+1}} S_j. & \end{cases}$$

From Eq.(3.29), we get the following fundamental expression for the fields

$$h_i(S) = \left[J \sum_{\mu=1}^{k+1} \left(\sum_{l=\mu}^{k+1} \frac{1}{2^{2\sigma}} \right)^l \right] 2^{\mu-1} m_{f(\mu,i)}^{\mu-1}, \tag{3.33}$$

where we used the relation $m_{f(\mu,i)}^{\mu-1} = \sum_{\{j\}:d_{ij}=\mu} S_j$. Thus the order parameters $m_{f(\mu,i)}^{\mu-1}$ represent the magnetizations assumed by spins that lie at distance μ from S_i. Note that the function $f(\mu, i)$ can be estimated through the floor function $\lfloor \cdot \rfloor$ (e.g., $\lfloor 3.14 \rfloor = 3$) as

$$f(\mu, i) = \left\lfloor \frac{i + (2^{\mu-1} - 1)}{2^{\mu-1}} \right\rfloor + (-1)^{(\lfloor \frac{i+(2^{\mu-1}-1)}{2^{\mu-1}} \rfloor + 1)}.$$

Finally, we notice that the largest value allowed for a field -away from the boundary value $\sigma = 1/2$- for large k approaches a plateau (whose boundaries -in the (k, σ) plane- are important for finite-size-scaling during numerical analysis), hence we can easily check the right field normalization

$$Q(\sigma, k+1) = \sum_{\mu=1}^{k+1} J(\mu, k+1, \sigma) 2^{\mu-1} =$$
$$= J \frac{2^{-2(k+1)\sigma} \left(2^{2(k+2)\sigma} - 2^{k+2\sigma+2} + 2^{k+2} + 4^\sigma - 2 \right)}{-3 \times 4^\sigma + 16^\sigma + 2}, \tag{3.34}$$

as $Q(\sigma, k)$ represents the largest value allowed by a field.
Note that in the thermodynamic limit

$$\lim_{k\to\infty} Q(\sigma, k) = Q(\sigma) = J \frac{2^{2\sigma}}{-3 \times 4^\sigma + 4^{2\sigma} + 2}, \tag{3.35}$$

that is Q is always bounded whenever $\sigma > \frac{1}{2}$.

3.3.2 Metastabilities in the Dyson Network: Noiseless Case.

We can now proceed to the stability analysis explaining in details a few test cases that show how to proceed for any other case of further interest:
[*a*] the global ferromagnetic state, i.e. $S_i = +1$, $i \in (1, ..., 2^{k+1})$.
[*b*] the parallel/mixed state, i.e. the first half of spins up and the second half down, thus $S_i = +1$, $i \in (1, ..., 2^k)$ and $S_i = -1$, $i \in (2^k + 1, ..., 2^{k+1})$.
[*c*] the dimer, i.e. $S_1 = S_2 = +1$ while $S_i = -1$ for all $i \neq (1, 2)$.
[*d*] the square, i.e. $S_1 = S_2 = S_3 = S_4 = +1$ while $S_i = -1$ for all $i > 4$.

Let us go through each case analysis separately:

- [*a*] The global ferromagnetic state $S_i = +1 \ \ \forall i \in [1, 2^{k+1}] \ \ \Rightarrow \ \ m_i^n(S) = 1 \ \forall i, n$ has fields

$$h_i(S) \ = J \frac{4^{-(k+1)\sigma} \left[2^{2(k+2)\sigma} - 2^{k+2+2\sigma} + 2^{k+2} + 4^\sigma - 2\right]}{-3 \times 4^\sigma + 16^\sigma + 2}, \tag{3.36}$$

$$h_i(S) \ > 0 \ \forall k, \sigma \in (1/2, 1). \tag{3.37}$$

  Thus, the configuration $S_i = +1 \ \ \forall i \in [1, 2^{k+1}]$ is stable in the noiseless limit $\forall \sigma \in [\frac{1}{2}, 1]$. In the thermodynamic limit $k \to \infty$ we have

$$h_i(S) = J \frac{4^\sigma}{-3 \times 4^\sigma + 16^\sigma + 2}.$$

  To address network's behaviour in the presence of noise, fixing $J = 1$ without loss of generality, we can look at the solution of the following equation

$$\tanh(\beta h_i(S)) \simeq 1 \Rightarrow \tanh\left(\beta \frac{4^\sigma}{-3 \times 4^\sigma + 16^\sigma + 2}\right) \simeq 1. \tag{3.38}$$

  This allows to find the curve $\beta_c^{\text{no errors}}(\sigma)$ versus σ (shown in Fig.(3.3)). In fact, we know that, at the time $t+\delta t$, the system obeys the dynamics

$$S_i(t + \delta t) = \text{sign}(\tanh(\beta h_i(S)) + \eta_i),$$

  where η_i is a random variable, whose value is uniformly distributed in $[-1, 1]$. Imposing $\tanh(\beta h_i) \simeq 1$ we ask that $|h_i| \gg 1$, so the sign of the right hand side member of the equation is positive, thus the sign of S_i at the time $t + \delta t$ is the same of the field h_i at the time t. Then, fixed σ, for every $\beta > \beta_c^{\text{no errors}}(\sigma)$ the state $S_i = +1 \ \ \forall i \in [1, 2^{k+1}]$ is stable without errors.

- [*b*] The parallel/mixed state $S_j = +1 \quad S_i = -1 \quad \forall j \in [1, 2^k] \quad \forall i \in [2^k + 1, 2^{k+1}]$ has fields

$$\begin{aligned} \Rightarrow h_j(S) &= J\frac{4^{-(k+1)\sigma}\left(2^{2(k+2)\sigma} + 2^{k+1+2\sigma} - 2^{k+1+4\sigma} + 4^\sigma - 2\right)}{-3 \times 4^\sigma + 16^\sigma + 2} \\ &= -h_i(S) > 0 \ \forall \ k+1 \geq 2, \end{aligned} \tag{3.39}$$

$$\Rightarrow \lim_{k\to\infty} h_j(S) = J\frac{1}{2^{1-2\sigma} + 4^\sigma - 3}, \tag{3.40}$$

thus the configuration $S_j = +1 \quad S_i = -1 \quad \forall j \in [1, 2^k] \quad \forall i \in [2^k + 1, 2^{k+1}]$ is stable in the noiseless limit $\forall\ k+1 > 2,\ \sigma \in (1/2, 1)$. Using the same arguments of the previous case, fixing $J = 1$ without loss of generality, to infer network's behaviour in the presence of the noise we can look at the solution of the following equation

$$\tanh(\beta h_i(S)) \simeq 1 \Rightarrow \tanh\left(\beta\frac{1}{2^{1-2\sigma} + 4^\sigma - 3}\right) \simeq 1. \tag{3.41}$$

This allows to find the curve $\beta_c^{\text{no-errors}}(\sigma)$ versus σ (see Fig.(3.3)). Then, fixed σ, for every $\beta > \beta_c^{\text{no-errors}}(\sigma)$ the state $S_j = 1 \quad S_i = -1 \quad \forall j \in [1, 2^k] \quad \forall i \in [1 + 2^k, 2^{k+1}]$ is stable without errors. So we can see how, in the thermodynamic limit, the state with all spins aligned $S_j = +1 \quad \forall j \in [1, 2^{k+1}]$ and the state with half spins pointing upwards and half pointing downwards $S_j = +1 \quad \forall j \in [1, 2^k] \quad S_i = -1 \quad \forall i \in [1 + 2^k, 2^{k+1}]$ are both robust. For an arbitrary finite value of k it is possible to solve numerically Eq.(3.41) to get an estimate for $\beta_c^{\text{no-errors}}(\sigma)$ versus σ: in Figure 3.3 $\beta_c^{\text{no-errors}}(\sigma)$ is plotted for the state $S_j = +1 \quad S_i = -1 \quad \forall j \in [1, 2^k] \quad \forall i \in [1+2^k, 2^{k+1}]$ and the state $S_i = +1 \quad \forall i \in [1, 2^{k+1}]$.

- [*c*] The dimer $S_j = +1 \quad S_i = -1 \quad \forall j \in [1, 2] \quad \forall i \in [3, 2^{k+1}]$ has fields

$$h_1(S) = h_2(S) = \frac{2^{-2\sigma(k+1)}(2^{2\sigma(k+2)} + 2^{k+2+2\sigma} - 4^{1+(k+1)\sigma} - 2^{k+2} - 3 \times 4^\sigma + 6)}{(-3 \times 4^\sigma + 16^\sigma + 2)},$$

$$\lim_{k\to\infty} h_1(S) = \lim_{k\to\infty} h_2(S) = 2 \cdot \frac{4^\sigma - 4}{-3 \times 4^\sigma + 16^\sigma + 2} < 0 \ \forall \sigma \in (1/2, 1).$$

Therefore, the configuration $S_j = +1 \quad S_i = -1 \quad \forall j \in [1, 2] \quad \forall i \in [3, 2^{k+1}]$, in the thermodynamic limit, is unstable $\forall\ \sigma \in (1/2, 1)$.

- [d] The square $S_j = 1 \;\; S_i = -1 \;\; \forall j \in [1,4] \;\; \forall i \in [5, 2^{k+1}]$ has fields

$$\begin{aligned} h_j(S,k) &= -\frac{2^{1-2(k+1)\sigma}\left(-2^{k+1+2\sigma} + 2^{2k\sigma+1} + 2^{k+1} + 2^{2\sigma+1} - 4\right)}{-3\times 4^{\sigma} + 16^{\sigma} + 2} \\ &- \frac{-3\times 4^{-(k+1)\sigma} + 2^{1-2\sigma} + 1}{1 - 4^{\sigma}}, \end{aligned} \tag{3.42}$$

$$\begin{aligned} h_j(S,k+1) &= \frac{\left(2^{2(k+3)\sigma} - 2^{k+2+2\sigma} + 2^{k+2+4\sigma} - 2^{2(k+1)\sigma+3}\right)}{(-3\times 4^{\sigma} + 16^{\sigma} + 2)/(2^{-2(k+2)\sigma})} \\ &+ \frac{(+7\times 2^{2\sigma+1} - 7\times 16^{\sigma})}{(-3\times 4^{\sigma} + 16^{\sigma} + 2)/(2^{-2(k+2)\sigma})} \end{aligned} \tag{3.43}$$

thus

$$\lim_{k\to\infty} h_j(S) = \frac{4^{-\sigma}\left(16^{\sigma} - 8\right)}{-3\times 4^{\sigma} + 16^{\sigma} + 2} = \begin{cases} > 0, & \text{if } \sigma > \frac{3}{4} \\ < 0, & \text{if } \sigma < \frac{3}{4} \end{cases}.$$

Therefore, the configuration $S_j = +1 \;\; S_i = -1 \;\; \forall j \in [1,4] \;\; \forall i \in [5, 2^{k+1}]$ in the limit $(k \to \infty)$ for $T = 0$ is stable $\forall\; \sigma \in (\frac{3}{4}, 1)$

It is worth noticing that beyond the extensive meta-stable states (e.g. the parallel/mixed one) already suggested by the statistical mechanical route, stability analysis predicts that tighley connected modules (e.g. octangon, esadecagon, ...) with spins anti-aligned with respect to the bulk get dynamically stable in the thermodynamic limit: these *motifs* in turn are able to process small amount of information and an analysis of their capabilities can be found in [23, 24], and their robusting is due to their intrinsic loopy structure.

3.3.3 Signal Analysis for the Hopfield Hierarchical Model

Let us now consider the Hopfield hierarchical model (see Eq.(3.21)). As we are interested in obtaining an explicit prescription for the fields experienced by the spins, we can rewrite its Hamiltonian in terms of neural distance d_{ij} as

$$H_{k+1}(S|\xi,\sigma) = \sum_{i<j} S_i S_j \left[\sum_{l=d_{ij}}^{k+1} \left(\frac{-1}{2^{2\sigma l}}\right)\right] \sum_{\mu=1}^{p} \xi_i^{\mu}\xi_j^{\mu} \tag{3.44}$$

or inverting the order of the sums

$$H_{k+1}(S|\xi,\sigma) = -\sum_{\mu=1}^{p}\sum_{i=1}^{2^{k+1}} S_i \left[\sum_{l=\mu}^{k+1} \left(\frac{1}{2^{2\sigma}}\right)^{l}\right] \sum_{\{j\}:d_{ij}=\mu} S_j \sum_{\nu=1}^{p} \xi_i^{\nu}\xi_j^{\nu},$$

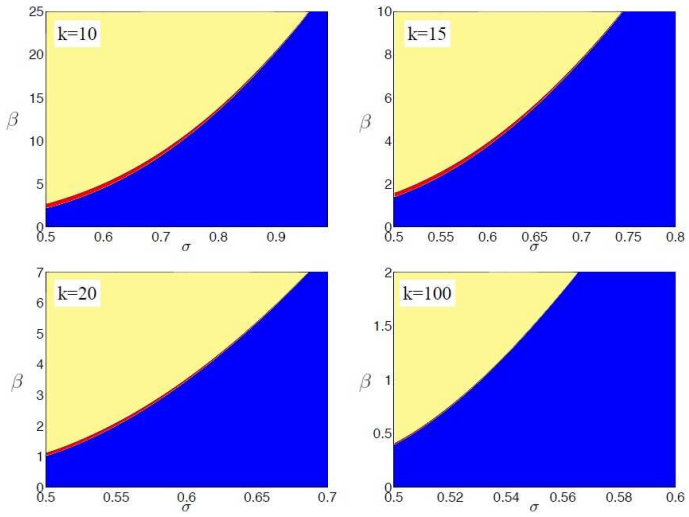

Figure 3.3: Phase diagram for the perfect retrieval accomplished by a pure state ($S_i = +1 \ \forall i = 1, ..., 2^{k+1}$) and parallel state ($S_i = +1 \ \forall i = 1, ..., 2^k$ and $S_i = -1 \ \forall i = 2^k + 1, ..., 2^{k+1}$). The line separating different regions corresponds to numerical solution of $\beta_c^{\text{no errors}}[\sigma]$ versus σ, obtained from (3.38) and (3.41) for different values of k (10, 15, 20, 100 respectively). In yellow, the area where both the pure and parallel states are perfectly retrieved, while in blue the area where none of them is retrieved. The red line represents the area where only the pure state is stable: this region vanishes as k gets larger (namely in the thermodynamic limit), hence confirming that the pure and the mixed state are both global minima.

such that, paying attention to the fields we can write

$$H_{k+1}(S|\xi,\sigma) = -\sum_{i=1}^{2^{k+1}} S_i h_i(S), \tag{3.45}$$

$$h_i(S) = \sum_{\mu=1}^{p} \left[\sum_{l=\mu}^{k+1} \left(\frac{1}{2^{2\sigma}} \right)^l \right] \sum_{\{j\}:d_{ij}=\mu} S_j \sum_{\nu=1}^{p} \xi_i^\nu \xi_j^\nu. \tag{3.46}$$

Mirroring the analysis carried on for the Dyson model, we introduce an ensemble of non-independent Mattis-like order parameters as

$$m_i^{\mu,n}(S) = \frac{1}{2^n} \sum_{j=i\times 2^n-(2^n-1)}^{i\times 2^n} S_j \xi_j^\mu \quad \text{with } i = 1, 2, ..., 2^{k+1-n}, \quad n = 0, 1, 2, ..., k+1 \tag{3.47}$$

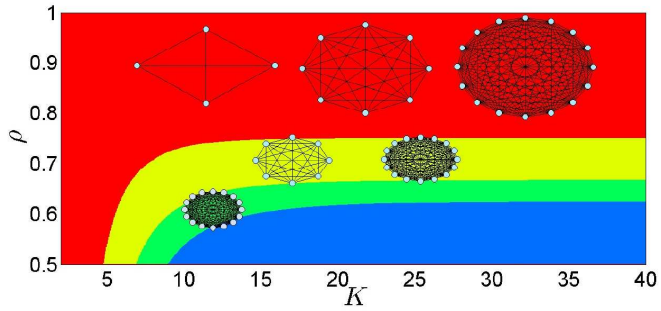

Figure 3.4: Stability and instability zones for various configurations in the plane (σ,k) when $\beta \to 0$, obtained by solving the inequality $S_i h_i(\sigma, k, [\mathbf{S}]) > 0$. In particular in the figure, the square represents the configuration $S_i = +1$ $\forall i \in [1,4]$ and $S_i = -1$ $\forall i \in [5, 2^{k+1}]$, the octagon the configuration $S_i = +1$ $\forall i \in [1,8]$ and $S_i = -1$ $\forall i \in [9, 2^{k+1}]$, and the esadecagon the configurations $S_i = +1$ $\forall i \in [1,16]$ and $S_i = -1$ $\forall i \in [17, 2^{k+1}]$. In red we can see the region where all of them are stable, in yellow the region where only the octagon and the esadecagon are stable, in green the region where only the esadecagon is stable, while in blue none of these reticular animals is stable.

so that

$$\begin{cases} m_i^{\mu,0} = S_i \xi_i^\mu & \text{with } i = 1, 2, .., 2^{k+1} \\ m_i^{\mu,1} = \frac{1}{2} \sum_{j=2i-1}^{2i} S_j \xi_j^\mu & \text{with } i = 1, 2, .., 2^k \ \to m_1^{\mu,1} = \frac{1}{2} \sum_{j=1}^{2} S_j \xi_j^\mu \\ m_i^{\mu 2n} = \frac{1}{2^2} \sum_{j=2^2 i-(2^2-1)}^{2^2 i} S_j \xi_j^\mu & \text{with } i = 1, 2, .., 2^{k-1} \ \to m_1^{\mu,2} = \frac{1}{4} \sum_{j=1}^{4} S_j \xi_j^\mu \\ ..... & \\ m_1^{\mu,k+1} = \frac{1}{2^{k+1}} \sum_{j=1}^{2^{k+1}} S_j \xi_j^\mu. & \end{cases}$$

As we saw for the Dyson case, this allows writing the fields as

$$h_i(S) = \sum_{\nu=1}^{p} \xi_i^\nu \sum_{d=1}^{k+1} \Big[\sum_{l=d}^{k+1} (\frac{1}{2^{2\sigma}})^l \Big] 2^{d-1} m_{f(d,i)}^{\nu,d-1} = \sum_{\nu=1}^{p} \xi_i^\nu \sum_{d=1}^{k+1} J(d, k+1, \sigma) 2^{d-1} m_{f(d,i)}^{\nu,d-1},$$

where

$$J(d, k+1, \sigma) 2^{\mu-1} = \frac{4^{\sigma - d\sigma} - 4^{-k\sigma - \sigma}}{4^\sigma - 1} 2^{d-1}. \tag{3.48}$$

The microscopic evolution of the system is defined as a stochastic alignment to local field $h_i(S)$:

$$S_i(t + \delta t) = \text{sign}\{\tanh[\beta h_i(S(t))] + \eta_i(t)\}, \tag{3.49}$$

where the stochasticity lies in the independent random numbers $\eta_i(t)$ uniformly drawn over the interval $[-1,1]$. In the noiseless limit $\beta \to \infty$ we have

$$S_i(t+\delta t) = \text{sign}[h_i(S(t))] \tag{3.50}$$

and so if $S_i h_i(S) > 0 \ \forall \ i \ \in \ [1,N]$, the configuration $[S]$ is dynamically stable (see Fig.(3.4)).

3.3.4 Signal to Noise Analysis for Serial Retrieval

Using equations (3.45) and (3.47) and posing $S_i = \xi_i^\mu$ in order to check the robustness of the serial pure-state retrieval (of the test pattern μ), we can write

$$\begin{aligned} \xi_i^\mu h_i(S) &= \xi_i^\mu \sum_{\nu=1}^{p} \xi_i^\nu \sum_{d=1}^{k+1} J(d,k+1,\sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu, \qquad (3.51) \\ &= \sum_{d=1}^{k+1} J(d,k+1,\sigma) 2^{d-1} + \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \sum_{d=1}^{k+1} J(d,k+1,\sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu. \end{aligned}$$

We can decompose the previous equation into two contributions, a stochastic noisy term $R(\xi)$ and a deterministic signal I as

$$\xi_i^\mu h_i(S) = I + R(\xi) \tag{3.52}$$

The signal term I is positive because

$$I = \sum_{d=1}^{k+1} J(d,k+1,\sigma) 2^{d-1} \geq 0, \tag{3.53}$$

while the noise $R(\xi)$ has null average (the latter being denoted by standard brackets), namely

$$\begin{aligned} R(\xi) &= \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \sum_{d=1}^{k+1} J(d,k+1,\sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu, \qquad (3.54) \\ \langle R(\xi) \rangle_\xi &= 0. \qquad (3.55) \end{aligned}$$

Thus, in order to see the regions of the tunable parameters $\sigma, k+1$ where the signal prevails over the noise and the network accomplishes retrieval, we need to calculate the second moment of the noise over the distribution of quenched variables ξ so to compare the signal amplitudes of I and $|\sqrt{\langle R^2(\xi) \rangle_\xi}|$:

$$\langle R^2(\xi)\rangle_\xi = \Big\langle \Big[\sum_{\nu\neq\mu}^{p}\xi_i^\nu \sum_{d=1}^{k+1} J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\Big]\times$$
$$\times\Big[\sum_{\eta\neq\mu}^{p}\xi_i^\eta \sum_{d=1}^{k+1} J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\eta\xi_j^\mu\Big]\Big\rangle_\xi. \qquad (3.56)$$

Neglecting off-diagonal terms (as they have null average), we get the following expressions for $\langle R^2(\xi)\rangle_\xi$:

$$\begin{aligned}\langle R^2(\xi)\rangle_\xi &= \Big\langle \sum_{\nu\neq\mu}^{p}(\xi_i^\nu)^2\Big(\sum_{d=1}^{k+1} J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\Big)^2\Big\rangle_\xi = \qquad (3.57)\\ &= \left\langle \sum_{\nu\neq\mu}^{p}\left(\sum_{d=1}^{k+1}(\frac{4^{\sigma-d\sigma}-4^{-(k+1)\sigma}}{4^\sigma-1})\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\right)^2\right\rangle_\xi,\end{aligned}$$

where we used $(\xi_i^\nu)^2 = 1\ \forall i,\nu$. Once again, as the ξ's are symmetrically distributed, only even order terms give contributions, thus we can safely neglect off-diagonal terms and write again

$$\begin{aligned}\langle R^2(\xi)\rangle_\xi &= (p-1)\sum_{d=1}^{k+1}\left\langle\left[\left(\frac{4^{\sigma-d\sigma}-4^{-k\sigma-\sigma}}{4^\sigma-1}\right)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\right]^2\right\rangle_\xi, \qquad (3.58)\\ &= (p-1)\sum_{d=1}^{k+1}\left(\frac{4^{\sigma-d\sigma}-4^{-k\sigma-\sigma}}{4^\sigma-1}\right)^2\langle\sum_{j:d_{ij}=d,\ k:d_{ik}=d}\sum\xi_j^\nu\xi_j^\mu\xi_k^\nu\xi_k^\mu\rangle_\xi.\end{aligned}$$

Therefore

$$\langle R^2(\xi)\rangle_\xi = (p-1)\sum_{d=1}^{k+1} J(d,\sigma,k+1)^2 2^{d-1}. \qquad (3.59)$$

Exploiting the approximation $\langle|x|\rangle \sim |\sqrt{\langle x^2\rangle}|$, we can simplify the previous expression into

$$\langle|R(\xi)|\rangle \sim \sqrt{\langle R^2(\xi)\rangle_\xi} = \sqrt{(p-1)\sum_{d=1}^{k+1} J(d,\sigma,k+1)^2 2^{d-1}}, \qquad (3.60)$$

where we consider the positive branch of the serial retrievl only. We are now ready to check the stability of the pure retrieval: as long as

$$I > \sqrt{\langle R^2(\xi)\rangle_\xi} \Rightarrow \xi_i^\mu h_i(S) = I + R(\xi) > 0, \qquad (3.61)$$

the pure state is stable. Hence we need to calculate explicitly

$$\sqrt{\langle R^2(\xi)\rangle_\xi} = \sqrt{\frac{(p-1)16^{-k\sigma}}{(4^\sigma-2)\left(4^\sigma-1\right)^2(16^\sigma-2)}}\cdot\sqrt{\Psi_1+\Psi_2},$$

where

$$\begin{aligned}\Psi_1 &= (4^\sigma-2)4^{2(k+1)\sigma} - 3\times 2^{k+2\sigma+1},\\ \Psi_2 &= 2^{k+6\sigma+1} - (16^\sigma-2)2^{2(k+1)\sigma+1} + 2^{k+2} - 64^\sigma + 2^{2\sigma+1} + 2^{4\sigma+1} - 4.\end{aligned}$$

The expression for the signal is much simpler, resulting in

$$I = \frac{4^{-(k+1)\sigma}\left(-2^{k+2\sigma+2} + 4^{(k+2)\sigma} + 2^{k+2} + 4^\sigma - 2\right)}{-3\times 4^\sigma + 16^\sigma + 2}. \tag{3.62}$$

Imposing $I = \sqrt{\langle R^2(\xi)\rangle_\xi}$ and solving for the variable p, we find the critical load allowed by the network, namely the function $P_c(\sigma,k)$, whose behavior is shown in Fig.3.5:

$$I = \sqrt{\langle R^2(\xi)\rangle_\xi} \Rightarrow P_c(\sigma,k). \tag{3.63}$$

Now, imposing the relation

$$P_c(\sigma,k) = k$$

and solving numerically with respect to σ, we can plot the maximum value $\sigma_{\max}(k)$ that the variable σ can reach such that the storage $P=k$ produces retrievable patterns, as shown in Figure 3.5.
In the thermodynamic limit we get

$$I - \sqrt{\langle R^2(\xi)\rangle} = \frac{2^{2\sigma}}{-3\times 4^\sigma + 16^\sigma + 2} - \frac{\sqrt{(p-1)}2^{2\sigma}}{\sqrt{\left(4^\sigma-1\right)(16^\sigma-2)}}, \tag{3.64}$$

$$P_c(\sigma) = \frac{\left(4^\sigma-1\right)(16^\sigma-2)}{\left(-3\times 4^\sigma + 16^\sigma + 2\right)^2} + 1. \tag{3.65}$$

3.3.5 Signal to Noise Analysis for Parallel Retrieval

Fixing $S_i = \xi_i^\mu\ \forall i \in [1,2^k]$ and $S_i = \xi_i^\gamma\ \forall i \in [1+2^k, 2^{k+1}]$ for $\mu \neq \gamma$, namely selecting μ and γ as test patterns to retrieve, we set the system in condition to handle contemporarily two patterns, the former managed by the first half of the spins, the latter by the second half. The robustness of this state is addressed hereafter following the same prescription outlined so far. Namely, being

$$S_i h_i(S) = S_i \sum_{\nu=1}^{p}\xi_i^\nu \sum_{d=1}^{k+1} J(d,k+1,\sigma) \sum_{j:d_{ij}=d}\xi_j^\nu S_j, \tag{3.66}$$

if $i \in [1, 2^k]$ we have

$$S_i h_i(S) = \xi_i^\mu \sum_{\nu=1}^{p} \xi_i^\nu \Bigg(\sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu + J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \Bigg), \tag{3.67}$$

while if $i \in [2^k + 1, 2^{k+1}]$, the same equation still holds provided we replace μ with γ and γ with μ, hence hereafter we shall consider only one of the two cases as they are symmetrical.

Again, we can decompose the above expression in the sum of a constant, positive term -that plays as the signal- $I > 0$, and a stochastic term for the noise $R(\xi)$, namely we can write

$$S_i h_i(S) = I + R(\xi), \tag{3.68}$$

$$I = \sum_{d=1}^{k} \Bigg(J(d, k+1, \sigma) 2^{d-1} \Bigg),$$

$$R(\xi) = J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\mu \xi_j^\gamma + \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \Big(\sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu + J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \Big).$$

In order to get a manageable expression for the noise, it is convenient to reshuffle $R(\xi)$ distinguishing four terms such that

$$R(\xi) = a + b + c + d, \tag{3.69}$$

where

$$a = J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\mu \xi_j^\gamma, \tag{3.70}$$

$$b = \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu, \tag{3.71}$$

$$c = \xi_i^\mu \sum_{\substack{\nu \neq \mu \\ \nu \neq \gamma}}^{p} \xi_i^\nu J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma, \tag{3.72}$$

$$d = \xi_i^\mu \xi_i^\gamma J(k+1, k+1, \sigma) 2^k. \tag{3.73}$$

As $\mu \neq \gamma$, we have that $\langle R(\xi)\rangle_\xi = 0$, while $\langle R^2(\xi)\rangle_\xi$ turns out to be

$$\langle R^2(\xi)\rangle_\xi = \langle a^2 + b^2 + c^2 + d^2 + 2(ab + ac + ad + bc + bd + cd)\rangle_\xi. \quad (3.74)$$

Let us consider these terms separately: skipping lenghty, yet straightforward calculations, we obtain the following expressions

$$\begin{aligned} \langle a^2\rangle_\xi &= \left\langle J^2(k+1,k+1,\sigma) \sum_{j:d_{ij}=k+1} \sum_{n:d_{in}=k+1} \xi_j^\mu \xi_j^\gamma \xi_n^\mu \xi_n^\gamma \right\rangle_\xi \\ &= J^2(k+1,k+1,\sigma) \times 2^k. \end{aligned} \quad (3.75)$$

$$\begin{aligned} \langle b^2\rangle_\xi &= \left\langle \left(\xi_i^\mu \sum_{\nu\neq\mu}^{p} \xi_i^\nu \sum_{d=1}^{k} J(d,k+1,\sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu \right)^2 \right\rangle_\xi \\ &= (p-1) \sum_{d=1}^{k} J^2(d,k+1,\sigma) 2^{d-1}. \end{aligned} \quad (3.76)$$

$$\begin{aligned} \langle c^2\rangle_\xi &= \left\langle \left(\xi_i^\mu \sum_{\nu\neq\mu\&\nu\neq\gamma}^{p} \xi_i^\nu J(k+1,k+1,\sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \right)^2 \right\rangle_\xi \\ &= (p-2) J^2(k+1,k+1,\sigma) 2^k. \end{aligned} \quad (3.77)$$

$$\langle d^2\rangle_\xi = \left\langle \left(\xi_i^\mu \xi_i^\gamma J(k+1,k+1,\sigma) 2^k \right)^2 \right\rangle_\xi = J^2(k+1,k+1,\sigma) 2^{2k}, \quad (3.78)$$

and, since a and b and, analogously, b and c, are defined over different blocks of spins, clearly

$$\langle 2ab\rangle_\xi = 0, \quad (3.79)$$
$$\langle 2bc\rangle_\xi = 0, \quad (3.80)$$
$$\langle 2bd\rangle_\xi = 0. \quad (3.81)$$

As a result, rearranging terms opportunely we finally obtain

$$\begin{aligned} \langle R^2(\xi)\rangle_\xi &= 4^{-2k\sigma} \Big(\frac{\left[4^k \left(4^\sigma - 1\right)^2 + 2^k \left(4^\sigma - 1\right)^2 + 2^k (p-2) \left(4^\sigma - 1\right)^2\right]}{\left(4^\sigma - 1\right)^2} \\ &+ (2((-3 \times 2^{k+2\sigma+1} + 2^{k+6\sigma+1} + 2^{k+2} + 2^{2\sigma+1} + 2^{4\sigma+1} - \\ &+ (4^\sigma - 2) 4^{2(k+1)\sigma} - (16^\sigma - 2) 2^{2(k+1)\sigma+1}) + \\ &- 64^\sigma)(p-1))((4^\sigma - 2)(16^\sigma - 2))^{-1} \Big), \end{aligned}$$

while the signal term reads as

$$I = \frac{2^{-2k\sigma-1}\left(-2^{k+2\sigma} - 2^{k+4\sigma} + 2^{2(k+1)\sigma+1} + 2^{k+1} + 2^{2\sigma+1} - 4\right)}{-3 \times 4^{\sigma} + 16^{\sigma} + 2}. \tag{3.82}$$

Imposing $I = \sqrt{\langle R^2(\xi)\rangle_\xi}$, and solving with respect to the variable p we can outline the function $P_c(\sigma, k+1)$ that returns the maximum allowed load the network may afford accomplishing parallel retrieval and whose behavior is shown in Fig.(3.5):

$$I = \sqrt{\langle R^2(\xi)\rangle_\xi} \Rightarrow P_c(\sigma, k+1). \tag{3.83}$$

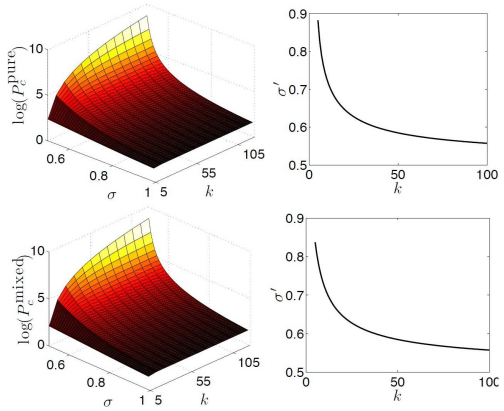

Figure 3.5: Upper panel (serial retrieval): On the left we show the maximum value of storable patterns P_c as a function of k and of σ (as results from Eq.(3.64)) for the pure state in order to have signal's amplitude greater than the noise (i.e. retrieval). Note the logarithmic scale for P_c highlighting its wide range of variability. On the right we show the maximum value of the neural interaction decay rate $\sigma'(k)$ versus k allowed to the couplings under the storage constraint $k = p$ and the pure state perfect retrieval constraint, in the $\beta \to \infty$ limit.
Lower panel (parallel retrieval): On the left there is the maximum value of storable patterns P_c as a function of k and of σ (as results from Eq.(3.85)) for th parallel state in order to have signal's amplitude greater than the noise (i.e. retrieval). Note the logarithmic scale for P_c highlighting its wide range of variability. On the right there is the maximum value of the neural interaction decay rate $\sigma'(k)$ versus k allowed to the couplings under the storage constraint $k = p$ and the parallel state perfect retrieval constraint, in the $\beta \to \infty$ limit.

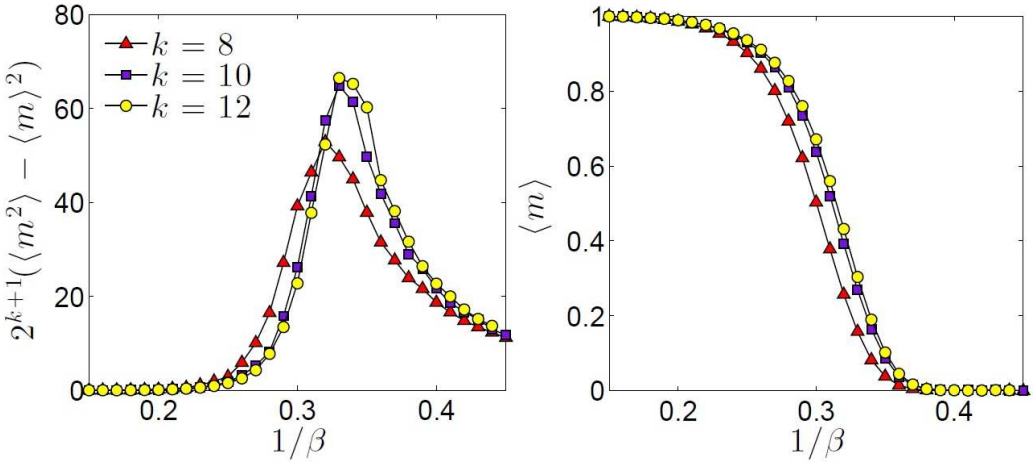

Figure 3.6: Starting from the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$ results of the simulations for DHM for $\sigma = 0.99$ and $N = 2^{k+1}$, $k+1 = 8, 10, 12$ are plotted. In the left panel, the rescaled magnetic susceptibility $2^{k+1}(\langle m^2\rangle - \langle m\rangle^2)$ is plotted vs β (one over the noise). In the right panel the magnetization $\langle m\rangle = \langle \frac{1}{N}\sum_{i=1}^{N} S_i\rangle$ is plotted vs β (one over the noise).

3.4 Insights from Numerical Simulations

Using the same machines described in the previous section 2.5. Aim of this Section is to present results from extensive numerical simulations to check the stability of parallel processing over the finite-size effects that is not captured by statistical mechanics or that can be hidden in the signal-to-noise analysis . Further this allows checking that the asymptotic behavior (in the volume) of the network is in agreement with previous findings.
All the simulations were carried out using the same machines described in the previous section 2.5 and according to the following algorithm.

1. Building the matrix coupling, pattern storage.
   Once extracted randomly from a uniform prior over ± 1 p patterns of length $k+1$, and defined the distance between two spins i and j as d_{ij} we build the matrix $\mathbf{J}$, for the HHM, as

$$J_{ij} = \frac{4^{\sigma - d_{ij}\sigma} - 4^{-(k+1)\sigma}}{4^{\sigma} - 1} \sum_{\mu=1}^{p} \xi_i^{\mu}\xi_j^{\mu}, \text{ for } i = 1, \cdots 2^{k+1},\ j = 1, \cdots, 2^{k+1}, \tag{3.84}$$

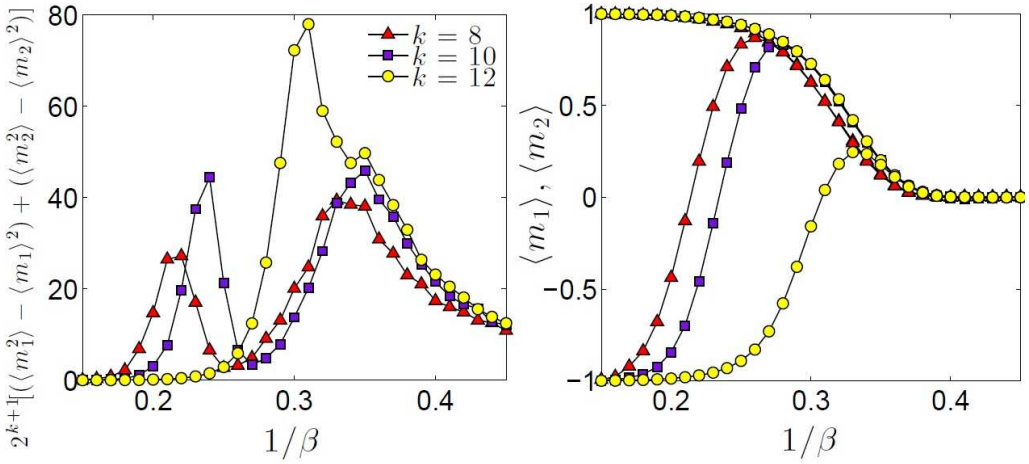

Figure 3.7: Starting from the state $S_i = +1, S_j = -1\ \forall i \in [1, 2^k]$ and $\forall j \in [2^k + 1, 2^{k+1}]$ results of the simulations for DHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. In the left panel, the rescaled magnetic susceptibility $2^{k+1}[(\langle m_1^2\rangle - \langle m_1\rangle^2) + (\langle m_1^2\rangle - \langle m_1\rangle^2]$ is plotted vs β (i.e. one over the noise) for $k+1 = 8, 10, 12$. In the right panel, the magnetizations $\langle m_1\rangle = \langle \frac{1}{2^k}\sum_{i=1}^{2^k} S_i\rangle$ and $\langle m_2\rangle = \langle \frac{1}{2^k}\sum_{i=1+2^k}^{2^{k+1}} S_i\rangle$ are plotted vs β (i.e. one over the noise) for $k+1 = 8, 10, 12$.

while for the DHM we use the form:

$$J_{ij} = \frac{4^{\sigma - d_{ij}\sigma} - 4^{-(k+1)\sigma}}{4^{\sigma} - 1}, \text{ for } i = 1, \cdots 2^{k+1} \text{ and } j = 1, \cdots, 2^{k+1}, \tag{3.85}$$

where $k+1$ is the number of levels of the hierarchical construction of the network, and $\sigma \in (\frac{1}{2}, 1]$.

2. Initialize the network.
   We used different initializations to test the stability of the resulting stationary configuration:

   -Pure retrieval: We initialize the network in an assumed fixed point of the dynamics, namely $S_i = \xi_i^{\mu}$ with $i = 1, ...2^{k+1}$ and $\mu = 1$ for the HHM, while $S_i = +1$ with $i = 1, ...2^{k+1}$ in the DHM case, and we check the equilibrium as reported in Fig[3.6].

   -Parallel retrieval: Since we study the multitasking features shown by this hierarchical network, we can also assign different types of initial conditions with respect to the pure state, e.g.

i) For the DHM, starting from the lowest energy level (after the standard one $S_i = 1\ \forall i$) we chose $S_i = +1$ for $i = 1, ..., 2^k$ and $S_i = -1$ for $i = 2^k + 1, ..., 2^{k+1}$ (viceversa is the same, and we check the equilibrium as reported in Fig[3.7]);

ii) For the HHM, looking for multitasking features, we set in the case $p = 2$, we set $S_i = \xi_i^1$ for $i = 1, ..., 2^k$ and $S_i = \xi_i^2$ $i = 2^k + 1, ..., 2^{k+1}$(Fig[3.10]); In the case $p = 4$ we set $S_i = \xi_i^\mu$ $\forall i \in \left[1 + \frac{(\mu-1)N}{4}, \frac{\mu N}{4}\right]$ and $\mu \in [1, 4]$(Fig[3.9])

In this way, we have two or four communities (sharing the same size) building the network with a different order parameter.

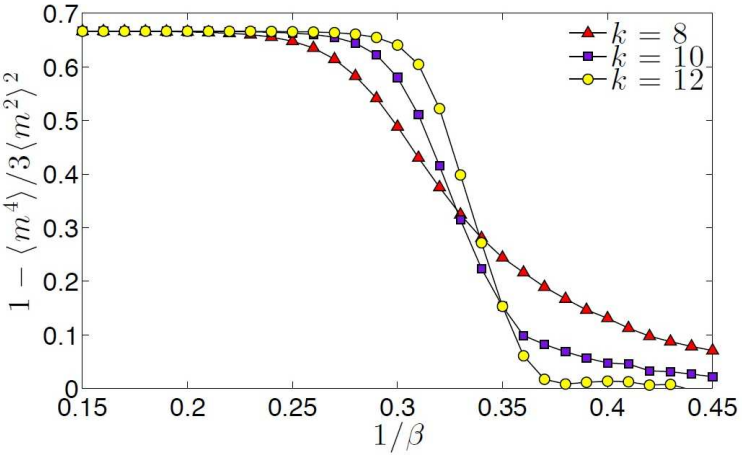

Figure 3.8: Starting from the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$ with $\sigma = 0.99$ for the DHM and $k+1 = 8, 10, 12$. Binder cumulant $1 - \frac{\langle m^4 \rangle}{3\langle m^2 \rangle^2}$ versus noise $\frac{1}{\beta}$ for $k + 1 = 8, 10, 12$. Plotting the binder cumulant for different values of $k + 1$ permits to find the critical noise of this state.

3. Evolution: Glauber dynamics.
The evolution of the spins follows a standard random asynchronous dynamics [5] and the state of the network is updated according to the field acting on the spins at every step of iteration, that is,

$$S_i(t + 1) = \text{sign}\{\tanh[\beta h_i(S(t)] + \eta(t)\}, \text{ for } \beta = T^{-1}$$

where $\eta(t)$ is the noise introduced as a random uniform contribution over the real interval $[-1, 1]$ in every step.
For each noise the stationary mean values of the order parameters

have been measured mediating over $O(10^3)$ different realizations. For the HHM the average of the order parameters is performed over the quenched variables. For DHM, to better highlight the stability of the parallel configuration, $S_i = +1$ for $i = 1, ..., 2^k$, $S_i = -1$ for $i = 2^k + 1, ..., 2^{k+1}$ and to break the Gauge invariance, during half of the relaxation period to equilibrium a small positive field is applied to the system.

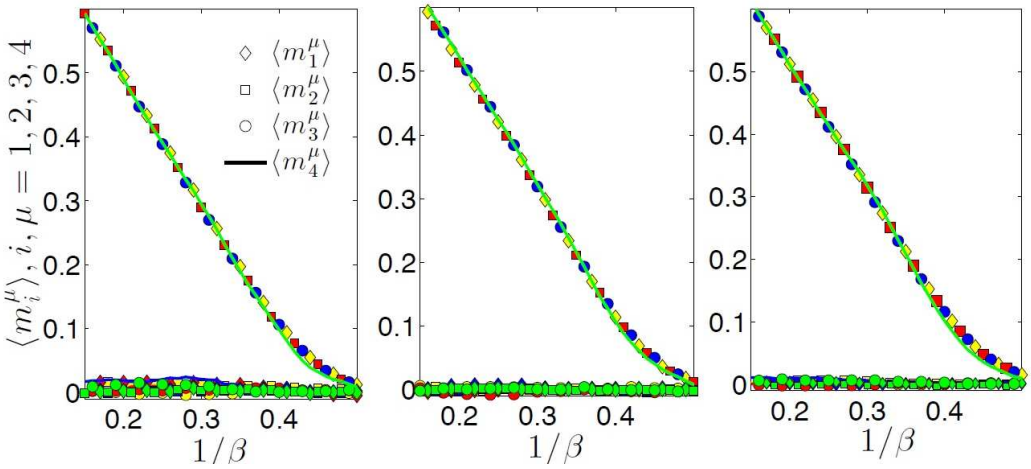

Figure 3.9: . Starting from the state $S_i = \xi_i^1, S_j = \xi_j^2, S_n = \xi_n^3, S_l = \xi_l^4$ $\forall i \in [1, 2^{k-1}], \forall j \in [2^{k-1}+1, 2^k], \forall n \in [2^k+1, \frac{3}{2}2^k], \forall l \in [\frac{3}{2}2^k+1, 2^{k+1}]$ results of the simulations for HHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. The Mattis order parameters $\langle m_i^\mu \rangle = \langle \frac{1}{2^{k-2}} \sum_{j=1+(i-1)2^{k-2}}^{i2^{k-2}} S_j \xi_j^\mu \rangle$ for $i, \mu \in [1, 4]$ are plotted vs noise,from left we have $k + 1 = 8, 10, 12$. Same colors correspond to the same pattern μ, while same symbols correspond to the same index i.

4. Results.
   It is worth noting that -at difference with paradigmatic prototypes for phase transitions (i.e. the celebrated Curie-Weiss model), as we can see from figures [3.6, 3.7, 3.8], in these models we studied here the critical noise level approaches its asymptotic value (obtained by analytical arguments in the thermodynamic limit) from above (i.e. from higher values of βs). This happens because the intensities of couplings are increasing functions (clearly upper limited) of the size of the system. As can be inferred from fig[3.7] (where we present results regarding simulations for the DHM at $\sigma = 0.99$, $k + 1 = 8, 10, 12$ $[S_i = +1, S_j = -1$ $\forall i \in [1, 2^k]$ and $\forall j \in [2^k + 1, 2^{k+1}]])$, the stability of the parallel configuration (in the low noise region) is confirmed and, as expected

from theoretical arguments, the noise region in which this configuration is stable increases with the size of the system up to coincide with that of the pure state. Also in the HHM case (figures [3.9, 3.10]) the stability of parallel configurations is verified (in the low noise region) for system's configurations shared by the two and four communities.

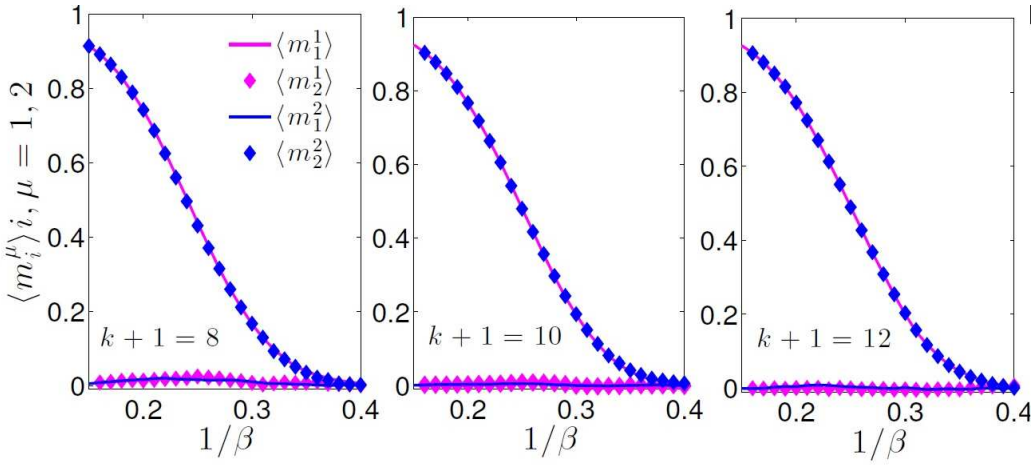

Figure 3.10: . Starting from the state $S_i = \xi_i^1, S_j = \xi_j^2 \ \forall i \in [1, 2^k], \forall j \in [2^k + 1, 2^{k+1}]$ results of the simulations for HHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. The Mattis order parameters $\langle m_i^\mu \rangle = \langle \frac{1}{2^{k-2}} \sum_{j=1+(i-1)2^{k-2}}^{i2^{k-2}} S_j \xi_j^\mu \rangle$ for $i, \mu \in [1, 2]$ are plotted vs noise,from left we have $k + 1 = 8, 10, 12$.

Chapter 4

Discussion

The comprehension of biological complexity is one of the main goals of this century research: the route to pave is long and scattered over countless branches. Focusing to neural networks, we notice that the deep difficulties in the statistical mechanics treatment prohibitive constraints beyond the mean field approximation (where each notion of distance or metrics for a space where to embed spins is lost),implied that their theory has been largely developed without investigating the crucial degree of freedom of neural distance. However, research is nowadays capable of investigations towards more realistic and/or better performing models: indeed, while the mean-field scenario, mainly represented by the Hopfield network as for retrieval and by the Boltzmann machines as for learning, has been so far understood (not completely at the rigorous level but at least largely), investigation of the non-mean-field counterpart is only at the beginning.

In this thesis we explored the retrieval capabilities of the multitasking associative network introduced the first time in [30] at the low storage level, and we tackled the problem of studying information processing (retrieval only) on hierarchical topologies introduced in [29], where spins interact with an Hebbian strength (or simply ferromagnetically in their simplest implementation, namely the Dyson model) that decays with their reciprocal distance.

In Chapter 2, we introduced a system characterized by (quenched) patterns which display a fraction d of null entries: interestingly, by paying the price of reducing the amount of information stored within each pattern (by a fraction d), we get a system able to retrieve several patterns at the same time. At zero noise level ($T = 0$), and for a relatively low degrees of dilution, the system converges to an equilibrium state characterized by overlap $\mathbf{m} = ((1-d), (1-d)d, ..., (1-d)d^k, (1-d)d^{P-1})$, where P is the number of stored patterns. Although this state displays non-null overlap with several patterns, it does not represent a spurious state, as can be seen by noticing, for

instance, that this state allows the complete retrieval of at least one pattern. However, through a careful inspection, we proved that there are regions in the (T, d) plane where genuine spurious states occur, hence the clear picture of the phase diagram that we offered becomes a fundamental issue in order to make the model ready for practical implementations.
A remarkable difference with respect to standard (serial processing) neural networks lies in the stability of mixture states: both even and odd mixtures are stable, which -within the world of spurious states - was a somewhat desired, and expected, result as there is neither a biological reason, nor a prescription from robotics, to weight differently odd and even mixtures (whose difference in terms of physical symmetries translates in the gauge invariance of the standard Hopfield model, that is explicitly broken within our framework due to the partial blankness of the pattern entries). Another expected feature, which we confirmed, is the emergence of parallel spurious states beyond standard ones. From classical neural network theory this is the natural generalization when moving from serial to parallel processing.

Beyond these somehow attended results, the phase diagram of the model is still very rich and composed by several not-overlapping regions where the retrieval states are deeply differently structured: beyond the paramagnetic state and the pure state, the system is able to achieve both a hierarchical organization of pattern retrievals (for intermediate values of dilution) and a completely symmetric parallel state (for high values of dilution), which act as the basis for the outlined mixtures when raising the noise level above thresholds whose value depends on the load P of the network.
These findings have been obtained developing a new strategy for computing the free energy of the model from which, imposing thermodynamic principles (i.e. extremizing the latter over the order parameters of the model), self-consistency has been obtained: the whole procedure has been based on techniques stemmed from partial differential equation theory. In particular, the key idea is showing that the noise-derivatives of the statistical pressure obey Burgers' equations, which can be solved through the Cole-Hopf transformation. The latter maps the evolution of the free energy over the noise into a diffusion problem which can be addressed through standard Green integration in momenta space and then mapped back in the original framework.

In chapter 3, we studied a Hebbian neural network, where spins are arranged according to a hierarchical architecture such that their couplings scale with their reciprocal distance. While a full statistical mechanical treatment is not yet achievable, stringent bounds for its free energy -intrinsically of non-mean-field nature- are still available and allows getting a picture of the network capabilities by far richer than the corresponding mean-field coun-

terpart (the Hopfield model within the low storage regime). Indeed, these networks are able to retrieve one pattern at a time accomplishing an extensive reorganization of the whole network state -mirroring serial processing as in standard Hopfield networks- but they are also able to switch to multitasking behavior handling multiple patterns at once -without falling into spurious states-, hence performing as parallel processors.
Remarkably, as far as the low storage regime is concerned, the underlying (weighted) topology -crucial for parallel processing- returns a phase space that shares similarities with the multitasking associative networks [30].
However, as theorems that definitively confirm this scenario are not fully available yet, to give robustness to the statistical mechanics predictions, we performed a signal-to-noise analysis checking whether those states -candidate by the first approach to mimic parallel retrieval- are indeed stable beyond the pure state related to serial processing and, remarkably, we found wide regions of the tunable parameters (strength of the interaction decay σ and noise level β) where indeed those states are extremely robust.
Clearly, as standard in thermodynamics, nothing is for free and even for this richness of behaviors there is a price to pay: as anticipated in the Summary of this thesis, emergent multitasking features in not-mean-field models require a substantial drop in network's capacity thus implying a new balance required by associative networks beyond the mean-field scenario.
While a satisfactory picture beyond such a mean-field paradigm is still far, we do hope that this work may act as one of the first steps in this direction.

Bibliography

[1] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, On the equivalence of hopfield networks and boltzmann machines, (Neural Networks, (2012)).

[2] A. Barra, G. Genovese, F. Guerra, The Replica Symmetric Approximation of the Analogical Neural Network, (J. Stat. Phys. 140, 784-796, (2010)).

[3] A. Barra, G. Genovese, F. Guerra, Equilibrium statistical mechanics of bipartite spin systems, (J. Phys. A 44, 245002, (2012)).

[4] A. Barra, G. Genovese, F. Guerra, D. Tantari, How glassy are neural networks, (J. Stat. Mech. 07, 07009, (2012)).

[5] A. C. C. Coolen, R. Kuhn, P. Sollich, Theory of neural information processing systems, (Oxford University Press, (2005)).

[6] A. K. Hertz John and R. Palmer, Introduction to the theory of neural networks, (Lecture Notes, (1991)).

[7] A. M. Turing ,Computing machinery and intelligence, (Mind. p. 433-460, (1950)).

[8] B. Bollobas, Modern graph theory, (Springer, volume 184, (1998)).

[9] B. Cheng and D. M. Titterington, Neural networks: A review from a statistical perspective, (Statistical Science, 9(1), 2-30, (1994)).

[10] B. Wemmenhove and A.C.C. Coolen, Finite connectivity attractor neural networks, (Journal of Physics A Mathematical and Theoretical, 36(9617), (2003)).

[11] C. Di Castro and G. Jona-Lasinio, On the microscopic foundation of scaling laws, (Phys. Lett, 29:322-323, (1969)).

[12] C. Kittel, Elementary statistical physics, (Courier Dover Publications, (2004)).

[13] C. Martindale, Cognitive psychology: A neural network approach, (Thomson Brooks/Cole, (1991)).

[14] C. Monthus, T. Garel, Dynamical barriers in the Dyson hierarchical model via real space renormalization, (J. Stat. Mech. P02023, (2013)).

[15] C. Monthus, T. Garel, Scaling of the largest dynamical barrier in the one-dimensional long-range Ising spin-glass, (Phys. Rev. B 89, 014408,(2014)).

[16] C. P. Bean, Magnetization of hard superconductors, (Physical Review Letters, 8:250. (1962)).

[17] D. J. Amit, Modeling brain function, (Cambridge University Press New York, (1989)).

[18] D. J. Amit, Hanoch Gutfreund, H. Sompolinsky,(Phys. Rev. Lett. 55, 1530, (1985)).

[19] D. J. Watts, S. H. Strogatz, Collective dynamics of smallworld networks, (Nature 393.6684: 440-442, (1998)).

[20] D. J. Willshaw and C. von der Malsburg, How patterned neural connections can be set up by self-organization, (Proc. R. Soc. Lond. B, 194:431-445, (1976)).

[21] D. O. Hebb, The Organization of Behavior (Lawrence Erlabaum Associates Publishers, (1949)).

[22] D. Saad, (On-line learning in neural networks, v. 17, Cambridge University Press, (2009)).

[23] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, D. Tantari, Immune networks: multitasking capabilities near saturation, (J. Phys. A 46(41), 415003, (2013)).

[24] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, D. Tantari, Immune networks: multi-tasking capabilities at medium load, (J. Phys. A 46(33), 335101, (2013)).

[25] E. Agliari, A. Barra, A Hebbian approach to complex network generation, (Europhys. Lett. 94, 10002,(2011)).

[26] E. Agliari, A. Barra, Criticality in diluted ferromagnet, (J. Stat. Mech. 10, 10003, (2008)).

[27] E. Agliari, A. Barra, A. De Antoni, and A. Galluzzi, Parallel retrievalof correlated patterns: From Hopfield networks to Boltzmann machines, (Neural Networks, 38:52-63, (2012)).

[28] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari and F. Tavani, Metastable states in the hierarchical Dyson model drive parallel processing in the hierarchical Hopfield network, (Journal of Physics A: Mathematical and Theoretical, Vol. 48, N. 1, (2014))

[29] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, and F. Tavani, Retrieval capabilities of hierarchical networks: from Dyson to Hopfield,(Phys. Rev. Lett. 114, 028103, (2015)).

[30] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, F. Moauro, Multitasking Associative Networks, (Physical Review Letters, 109:268101, (2012)).

[31] E. Agliari, A. Barra, F. Camboni, Criticality indiluted ferromagnets, (Journal of Statistical Mechanics: Theory and Experiment, 10003, (2008)).

[32] E. Agliari, A. Barra, F. Guerra, F. Moauro. A thermodynamic perspective of immune capabilities, (J. Theor. Biol., 287:48-63, (2011)).

[33] E. Agliari, L. Asti, A. Barra, R. Burioni, G. Uguzzoni, Analogue neural networks on correlated random graphs, (Journal of Physics A, 45:365001, (2012)).

[34] E. T. Roll, A.Treves, Neural networks and brain function, (1998).

[35] F. Guerra, Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model, (Comm. Math. Phys. 233, 1-12, (2003)).

[36] F. Guerra, F.L. Toninelli, The thermodynamic limit in mean fieeld spin glass models,(Comm. Math. Phys. 230(1), 71-79, (2002)).

[37] F. J. Dyson, Existence of a Phase-Transition in a One-Dimensional Ising Ferromagnet, (Comm. Math. Phys. 12, 91-107, (1969)).

[38] F. L. Metz, L. Leuzzi, G. Parisi, The renormalization flow of the hierarchical Anderson model at weak disorder, (Phys. Rev. B 89, 064201, (2014)).

[39] F. L. Metz, L. Leuzzi, G. Parisi, V. Sacksteder, Transition between localized and extended states in the hierarchical Anderson model, (Phys. Rev. B 88, 045103, (2013)).

[40] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, (Psychological review, 65(6), 386, (1958)).

[41] F. Yueh Wu, The Potts model, (Review of Modern Physics, 54:235, (1982)).

[42] G. Gallavotti, S. Miracle-Sole', Statistical mechanics of lattice systems,(Comm. Math. Phys. 5(5):317-323, (1967)).

[43] G. Genovese and A. Barra, A mechanical approach to mean Field models, (Journal of Mathematical Physics, 50:053303, (2009)).

[44] G. W. Domhoff ,Neural networks, cognitive development and content analysis, (American Psychological Association, (2003)).

[45] H. C. Tuckwell,Introduction to theoretical neurobiology, (Cambridge University Press, (2005)).

[46] H. Sompolinsky, Neural networks with non-linear synapses and a static noise, (Physical Review A, 34:2571, (1986)).

[47] I. Perez-Castillo, B. Wemmenhove, J.P.L. Hatchett, A.C.C. Coolen, N.S. Skantzos, T. Nikoletopoulos, Analytic solution of attractor neural networks on scale free graphs, (J. Phys. A 37, 8789-8799,(2004)).

[48] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Theory of superconductivity, (Physical Review, 108:1175, (1957)).

[49] J. J. Hopfield, Neural Network and Physical System With Emergent Collettive Computational Abilities, (Proceedings of the National Academy of Sciences USA, 52, 2552-2558, (1982)).

[50] J. Von Neumann, The general and logical theory of automata, (Cerebral mechanisms in behavior, p. 1-41, (1951)).

[51] K. G. Wilson, Renormalization group and critical phenomena, (Physical Review B, 4, 3174, (1971)).

[52] L. E. Reichl and I. Prigogine, A modern course in statistical physics, (University of Texas press, volume 71, Austin, (1980)).

[53] L. Pastur, M. Shcherbina and B. Tirozzi, The replica-symmetric solution without replica trick for the hopfield model, (Journal of Statistical Physics, 74:1161-1183, (1994)).

[54] M. Castellana, A. Barra, F. Guerra, Free-energy bounds for hierarchical spin models, (J. Stat. Phys. 155, 211, (2014)).

[55] M. Castellana, A. Decelle, S. Franz, M. Mezard, G. Parisi, The Hierarchical Random Energy Model, (Phys. Rev. Lett. 104, 127206, (2010)).

[56] M. Castellana, G. Parisi, A renormalization group computation of the critical exponents of hierarchical spin glasses, (Phys. Rev. E 83, 041134, (2011)).

[57] M. Kac, Random walk and the theory of brownian motion, (American Mathematical Monthly, p. 369-391, (1947)).

[58] M. L. Minsky and S. A. Papert, Perceptrons - Expanded Edition, (MIT press, Boston, MA. (1987)).

[59] M. Mezard, G. Parisi, M.A. Virasoro, Spin glass theory and beyond,(World Scientific, Singapore, Lect. Notes Phys.9, (1987)).

[60] M. T. Hagan, H. B. Demuth, M. H. and Beale, Neural network design, (Pws Pub., Boston,(1996)).

[61] P. Castiglione, M. Falcioni, A. Lesne and A. Vulpiani, Chaos and coarse graining in statistical mechanics, (Cambridge University Press, (2012)).

[62] R. J. Baxter, Exactly solved models in statistical mechanics, (Courier Dover Publications, (2007))

[63] R. M. Harris-Warrick,Dynamic biological networks (MIT press, (1992)).

[64] R. R. Trippi and E. Turban, Neural Networks in Finance and Investing:Using Artificial Intelligence to Improve Real World Performance, (McGraw-Hill, Inc. New York, NY, USA, (1992)).

[65] R. S. Ellis, Entropy, large deviations and statistical mechanics, (Springer-Verlag, (1985)).

[66] S. Nolfi and D. Floreano, Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines, (2000).

[67] T. Nikoletopoulos, A.C.C. Coolen, I. Pèerez-Castillo, N.S. Skantzos, J.P.L.Hatchett, and B Wemmenhove, Replicated transfer matrix analysis of ising spin models on 'small world' lattices, (Journal of Physics A: Mathematical and General, 37:6455, (2004)).

[68] W. McCulloc, W. Pitts, A logiacal Calculus of the Idea Immanent in Nervous Activity, (Bulletin of Mathematical Biophysics, 5, 115-133, (1943)).

[69] W. N. Bailey, Gauss Theorem in Generalised Hypergeometric Series, (Cambridge University Press,UK, (1935)).

[70] W. T. Miller, P. J. Werbos, and R. S. Sutton, Neural networks for control, (MIT press, (1995)).