ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

MATEMATICA

Ciclo 37

Settore Concorsuale: 01/A4 - FISICA MATEMATICA

Settore Scientifico Disciplinare: MAT/07 - FISICA MATEMATICA

INVESTIGATION OF BOLTZMANN-GIBBS LEARNING ENGINES: HIGH-DIMENSIONAL INFERENCE IN MEAN-FIELD THEORY AND OPTIMIZATION OF DEEP NETWORKS WITH FINITE RESOURCES

Presentata da: Gianluca Manzan

Coordinatore Dottorato

Giovanni Mongardi

Supervisore

Pierluigi Contucci

Co-supervisore

Daniele Tantari

Esame finale anno 2025

Borsa di dottorato del Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005), risorse FSE REACT-EU, Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione” e Azione IV.5 “Dottorati su tematiche Green.”-J35F21003200006

Abstract

This thesis investigates the inference properties of Boltzmann-Gibbs learning engines, focusing on their ability to learn from data and applying this understanding to optimization problems. The study focuses on high-dimensional inference in Restricted Boltzmann Machines (RBMs) and Hopfield networks. These models rely on inferring structures from data as a prerequisite to leverage their generative capabilities.

Central to this work is the teacher-student paradigm, where a teacher network generates datasets analyzed by a student network. Learning performance is evaluated under various conditions, including Bayes-optimal and mismatched regimes. In low-temperature settings, the student network effectively learns through memorization. Conversely, high-temperature datasets induce a modern signal retrieval (sR) phase, where the student aggregates partial information from noisy inputs. These findings, derived from mean-field spin glass theory, are extended to RBMs by analyzing the effects of different architectures, particularly through a parametric class of priors. Gaussian hidden units are shown to facilitate entry into the sR region, although the needed critical amount of data remains influenced by the dataset generation process.
Inference capabilities are further enhanced by introducing ferromagnetic coupling among replicated Hopfield networks. This coupling significantly expands the signal retrieval region beyond what can be achieved by modifying unit priors. Analytical results demonstrate that as the number of coupled student networks grows, the sensitivity of the learning region size diminishes, confirming that aligned replicas enhance learning efficiency without allowing for improbable weight regularizations.

Finally, the thesis explores energy-efficient training by optimizing the classification performance of deep neural networks (DNNs) under finite resource constraints, addressing a critical challenge in sustainable machine learning and satellite technology. By optimizing the distribution of neurons within hidden layers, the robustness of DNNs against radiation noise is improved. Insights drawn from Deep Boltzmann Machines (DBMs) inform the identification of optimal architectures based on thermodynamic parameters, such as form factors and inverse temperatures. Experimental validation across multiple datasets demonstrates measurable improvements in network robustness.

Together, these results provide a unified perspective on the learning performance of RBMs and related models. They also contribute to advancing sustainable technology applications by enabling robust neural network designs that perform efficiently under resource constraints.

Acknowledgements

The completion of my PhD research project—and this thesis, which aims to provide a detailed and comprehensive account of my work—has been made possible not only through my passion and perseverance but also thanks to the invaluable support of many people who stood by my side, both scientifically and personally.

I would like to begin by expressing my gratitude on the scientific front. My deepest thanks go to my supervisor, Prof. Pierluigi Contucci. From the very first colloquium to the conclusion of this challenging yet rewarding journey, his guidance has been indispensable. His teachings and advice have shown me how a scientist should approach research, and his extensive knowledge, experience, and patience have significantly contributed to the success of this endeavor. Prof. Contucci's ability to balance rigorous scientific expectations with a genuine respect for personal growth has left a profound impression on me. I am also especially grateful for his supervision during the initial stages of my PhD, as he coordinated the work of the project, laying the foundation for a successful outcome. In this regard, I would also like to acknowledge the contributions of Prof. Paolo Branchini and Sacha Cormenier. Prof. Branchini's boundless enthusiasm always inspired me. His energy for exploring new subjects was always shared during our collaborative work as well as his support and guidance. Sacha, with his unwavering dedication and collaborative spirit, played a key role in our shared project. Over time, our professional relationship grew into a true friendship, as we navigated numerous challenges together. I am also grateful to Prof. Federico Ricci Tersenghi, whose expertise significantly enhanced the reliability of certain results presented in this work, as well as to Prof. Emanuele Mingione and Dr. Diego Alberici for their invaluable discussions and support.
I am particularly grateful to Prof. Contucci for granting me the freedom to explore my own ideas and choose my collaborators, which ultimately led me to work with my co-supervisor, Prof. Daniele Tantari.

I owe a special debt of gratitude to Prof. Tantari. I sincerely thank him for taking on this role and for introducing me to the fascinating field of Restricted Boltzmann Machine models. Despite my initial rudimentary understanding of the subject, he patiently guided me through the rich intersection of machine learning and spin glasses. His dedication, trust, and encouragement have been a source of inspiration. I deeply appreciate the countless hours we spent discussing ideas, addressing doubts, and nurturing scientific curiosity—often in front of a blackboard. I am also profoundly grateful to him for connecting me with the research group in Madrid, where I spent nearly the final year of my PhD.

This brings me to Madrid—a city where I left a piece of my heart. A very special thanks goes to Professors Aurélien Decelle and Beatriz Seoane. They welcomed me warmly into their research group, providing invaluable scientific insights and introducing me to exciting topics that continue to shape my current investigations. Beyond being brilliant scientists, they are among the kindest and most supportive individuals I have had the privilege to meet. I often reflect on how their passion for science rekindled my own enthusiasm after a challenging period when my motivation had waned.

Their presence reminded me that it is the people—far more than the place itself—who leave the deepest impressions on our memories.

Among young researchers, I owe thanks to Filippo Zimmaro and Godwin Osabutey for their camaraderie and shared problem-solving sessions during long afternoons. I also owe sincere thanks to Francesco Camilli, whose guidance during the early stages of my PhD was essential. I met Francesco during his final year as a PhD student, and I was immediately struck by his determination and hard work. Watching him tirelessly strive toward ambitious goals inspired me to adopt the same mindset. His willingness to answer questions, share thoughtful advice, and devise creative solutions inspired me to face my own challenges with greater confidence.
I am equally grateful to Francesco Alemanno, who has been a constant source of both professional support and friendship throughout this journey. His advice and readiness to help with my doubts were invaluable, but what I cherish most is the friendship we shared outside of work. You reminded me of the importance of balance during this demanding period of life. Francesco, your warmth, generosity, and optimism have meant a lot to me.

My heartfelt thanks also go to Giovanni Catania, whose profound scientific knowledge and kindness made an indelible impression. Giovanni, thank you for being such a gracious host in Madrid, for helping me learn the language, and for the countless memorable moments we shared over beers and long conversations. I learned so much from you, not only academically, but also about the joy of building connections.
I also want to thank Nicolas Béreux, Alfonso Navas, Lorenzo Rosset, and all the friends from the Complutense. The time we spent together in Madrid remains one of the most cherished chapters of my life. I smile every time I think of it.

On a deeply personal note, I want to include a reminder to myself—as I reread this one day—that hardships are part of life. Even when everything seems to be going terribly wrong, one must hold onto hope and keep moving forward.

Midway through my first PhD year, I faced an unimaginable loss: the sudden passing of my father. The grief and pain that followed were indescribable and, although the intensity has softened over time, that feeling remains with me. Even now, as I write this part of the text, I remember how that summer unfolded. My motivation to work trembled, shaken by an overwhelming sense of sadness and anger. I felt lost and unable to continue on a path that had once filled me with purpose. It was only through the strength and unity of my family that I found the will to go on.

My mother, in particular, became the pillar that held everything together. Mom, I know how much you suffered during that time, even though you tried to shield us from it. I saw it in your quiet tears and in the way you continue to carry that sorrow today. You always put yourself last so that my sister and I could pursue our goals. I cannot thank you enough for the countless sacrifices you made. I only hope that, one day, I can repay even a fraction of the immense debt of gratitude I owe you.
To my sister, I am equally grateful. You were there for our mother when I could not be, offering her support and strength in countless ways. I've watched you grow through your own challenges with a determination that I deeply admire. Your achievements and resilience are a constant source of motivation for your older brother.
To you, Arianna, I am truly happy to have shared so many beautiful moments with you. Your presence has brought light and warmth to my life in ways that words can hardly express. I also want to thank you for being there for me during the most challenging periods of my life. Your unwavering support and words of affection have helped guide me through difficult times, and I will forever be grateful for that. I cherish the unique connection we share—our effortless chemistry, where words are unnecessary because we simply understand each other. With you, even the hardest days feel

warmer and brighter. I deeply admire your intuitive understanding of emotions and relationships—a rare and truly special quality that makes you someone remarkable. Your empathy and courage inspire me in ways I cannot fully express, and I feel profoundly grateful for the many meaningful moments we've shared together.

Lastly, I want to thank my father. I know it was your quiet encouragement and example that set me on this path, even though you never pressured me to follow any particular direction. You taught me the value of hard work, sacrifice, and standing up for what you believe in. Those lessons have shaped the person I am today and continue to guide me in everything I do. This thesis is also a tribute to your memory—a way of honoring everything you represented to me and the profound influence you had. I hope you would be happy for this result.

Publications and pre-prints

The thesis is built upon the papers, preprints and works in progress I have been developing on in the last three years. Particularly Chapter 4 is related to a novel work soon to be published.

A. Chapter 2 is based on the following paper:
Francesco Alemanno, Luca Camanzi, Gianluca Manzan, Daniele Tantari.
Hopfield model with planted patterns: a teacher-student self-supervised learning model
Applied Mathematics and Computation 458, 128253 (2023).
arXiv:https://arxiv.org/pdf/2304.13710.

B. Chapter 3 and 5 are based on the following pre-prints:

- Gianluca Manzan, Daniele Tantari
The effect of priors on Learning with Restricted Boltzmann Machines
arXiv:https://arxiv.org/pdf/2412.02623.
- Paolo Branchini, Pierluigi Contucci, Sacha Cormenier, Gianluca Manzan
Architectural Optimisation in Deep Neural Networks.
Tests of a theoretically inspired method.
npj Arificial Intelligence https://www.nature.com/npjai/.

C. Chapter 4 is an extension of the work A, in collaboration with
Giovanni Catania, Aurélien Decelle, Beatriz Seoane, Daniele Tantari.

Contents

Introduction

Deep neural networks (DNNs) are at the forefront of the technological revolution driven by Artificial Intelligence (AI) [1]. These networks have the ability to learn complex patterns and represent high-dimensional data, making them invaluable tools across a wide range of domains. Examples are image recognition, language processing, protein design [2]. Over the past decade, DNNs have firmly established themselves due to their exceptional performance in diverse applications, their adaptability to various datasets, and their suitability for both research and industrial purposes [3–5]. Their advancement as a powerful tool was further recognized by this year's Nobel Prize in Physics [6].

As with other revolutionary technologies, the initial phase of DNN adoption has focused primarily on achieving functionality, often without regard for the quality-to-cost ratio. Over time, however, resource constraints make optimization a critical concern. Historically, the pursuit of optimization has often driven significant scientific progress. For instance, during the 19^{th} century, Carnot's studies on the efficiency of heat engines, grounded in the second law of thermodynamics, became a cornerstone of the Industrial Revolution [7]. As the founding fathers of Deep Learning often claim we are nowadays in the same pre-scientific age, searching for the *thermodynamics of learning* [1, 8].

Central to this effort is the investigation of the bounds of the learning mechanism [9, 10], which forms the core of modern machine learning technologies. Attempts to understand these bounds have unveiled a lack of a solid underlying theoretical framework that explains remarkable success of DNN over the multitude of their applications.

In this context, substantial contributions to the theoretical foundations of neural networks have come from statistical mechanics, particularly the replica method [11]. This approach is widely regarded as a powerful—albeit not strictly rigorous—tool and will be extensively employed throughout this thesis. One of its most prominent applications is the analysis of the Hopfield model [12], one of the earliest artificial neural networks designed to explore associative memory mechanisms analogous to those observed in biological systems [13].

The Hopfield model represents memories as stable configurations of neurons, which can be recalled via Hebbian synaptic connections [14]. Using statistical mechanics, the recall process can be depicted in phase diagrams, which illustrate how many memories the network can reliably store before confusion states emerge. Since the advent of Hopfield's work, numerous extensions have been proposed to enhance the network's representational capacity [15–17], ultimately culminating in the modern learning mechanisms of Artificial Intelligence. A particularly relevant case is unsupervised learning, which aims to derive internal representations of datasets by mapping them onto probability distributions. The goal of these networks is often to generate typical configurations of the data, but they can also be tailored to specific tasks.

One well-known example of unsupervised models is the Deep Boltzmann Machine (DBM) [18, 19]. Under the statistical mechanics framework, a DBM can be modeled as a multi-species spin system [20], where the interactions are learned from the data. These networks consist of neurons arranged in successive layers, with interactions—referred to as weights—representing the synapses that connect the units in adjacent layers. From a generative perspective, DBMs have been already studied [18, 20, 21], but many questions remain open. For instance, generalizing Parisi's theory [22, 23] to non-convex structures such as DBMs is an ongoing challenge, as it is necessary to characterize the learning equilibrium states of these networks. Studying phase diagrams in these models can help answer critical questions, such as:

- How much data, sampled from the environment, is necessary for the machine to efficiently learn a representation of the same environment?
- How does the machine learn the structure of the dataset, and how is this encoded in the interactions?

While directly answering these questions for deep stratified networks is challenging, an insightful approach involves investigating Restricted Boltzmann Machines (RBMs). These are simpler architectures, composed of only two layers, which are typically stacked [18, 24, 25], thereby forming their proper deep counterpart. Despite their relative simplicity, RBMs are capable of learning rich internal representations [26, 27], making them a foundational step toward understanding DBMs. This thesis is situated precisely at this intersection. Our primary goal is to contribute to the theoretical understanding of RBMs inference. We approach this problem using the so-called teacher-student scenario [28–30], in which a neural network with a given architecture (*teacher*) defines the model of the environment (*signal*), while another network (*student*) has to learn some key properties about this environment by leveraging a dataset generated by the teacher. In this controlled setup, it is possible to rigorously investigate the data requirements for the student network to transition into a learning regime and to study how this phase is influenced by the student's architectural properties, such as unit priors or weights regularization. Initially, we apply this inference methodology to the Hopfield model, which can be interpreted as a particular case of a RBM representation. Subsequently, we extend our analysis to a broader class of RBM models. Finally, we focus on a specific DBM implementation, evaluating its performance under defined constraints.

More specifically, the first part of the thesis provides a brief yet rigorous and self-contained introduction to statistical inference and the foundational tools of statistical mechanics used throughout the work. To illustrate these concepts, we include concrete examples from well-known models, such as the Curie-Weiss and Hopfield models. This approach highlights the distinction between the direct problem (analyzing configurations sampled from the distribution of these spin systems) and the inverse problem (inferring system parameters from observed data), which constitutes the primary focus of the thesis.

In Chapter 2, we introduce the teacher-student scenario within the context of the Hopfield model and explore its parameter estimation theory. Here, the data represents the dual version of the connections in the original Hopfield model. In the direct perspective, the weights or patterns represent the saved memories within the network, that we aim to reproduce. Typically, these weights are independent and identically distributed (i.i.d.) random variables, drawn from a uniform binary distribution. Parisi's spin glass theory, allow us to depict the network's sampled configurations. Each sample may or may not resemble an embedded pattern, depending on key parameters such as the total number of initially stored patterns and the system's noise level (temperature) [31]. The dual inverse perspective, on the other hand, uses data as interactions, focusing on reconstructing weights (training) from observed data. We show that having complete prior knowledge about the generating process (Bayes-optimality [32, 33]) significantly simplifies the description of the learning phase in the dual Hopfield model. However, when the prior information is incomplete, the difference between classical memorization mechanisms [34] and modern machine learning approaches becomes apparent.

In Chapter 3, we extend our analysis to RBMs and examine the impact of architectural choices on their learning mechanisms. Specifically, we construct student networks with units (hidden, visible, patterns) whose priors interpolate between continuous (Gaussian) and binary distributions. This parametric approach allows us to model the effects of various architectural choices for both teacher and student RBMs on the *efficiency of learning*. Efficiency is quantified in terms of the dataset size required for learning and the level of regularization needed to enable the learning phase, referred to as signal retrieval (sR).

In Chapter 4, we address the limitations of the learning phase identified in earlier chapters, particularly the constraints on data availability and the role of temperature in expanding the sR phase. The breakthrough idea presented in [35] enhances the Perceptron's generalization performance by incorporating ferromagnetic coupling. Specifically, this approach reduces the onset of the hard phase [36], enabling standard thermal updating algorithms to more efficiently reach the signal solution and avoid being trapped in metastable states. Inspired by this, we apply the same collective learning approach to the Hopfield teacher-student scenario. By coupling a number of y Hopfield students, we investigate the sR phase of the new system, exploring if this collective learning strategy can reduce data requirements or increase learning temperature.

The final chapter focuses on the application of DBMs in supervised learning tasks, specifically in image classification. Here, we utilize a dataset of labeled images, divided into training and test sets. The training set is used to teach the network to distinguish between image classes, while the test set evaluates its performance. To adapt a DBM for this task, we add a final layer with neurons corresponding to the number of labels and use standard backpropagation to compute the weights across all layers, optimizing for prediction accuracy. This experiment serves as a practical entry point to the problem of resource optimization. As discussed at the beginning, optimizing neural network performance is critical in addressing the energy-intensive nature of modern AI systems.
The chapter extends this theme by exploring methods to achieve maximum performance under constrained resources. Using the number of neurons in hidden layers as a proxy for resource allocation, we investigate the optimal network topology for enhancing classification performance. This problem is particularly relevant for satellite applications, where embedding efficient architectures in hardware designed to operate in radiation-rich environments is critical [37, 38]. To meet these demands, we design a compact network optimized for integration into hardware that functions in ambient radiation environments. To assess the feasibility of these architectures, we introduce a robustness metric specifically tailored to quantify the resilience of the network's predictions. This metric is sensitive to environmental factors, such as interactions with energetic neutrons [39]. Guided by recent theoretical results [20, 40], which establish a relationship between neuron placement and network performance, we propose a neuron reconfiguration procedure. We apply this procedure to evaluate the behavior of the performance metrics and identify the most robust architecture across multiple datasets. We find that only a small number of thermodynamic parameters are sufficient to determine these optimal architectures.

Chapter 1

Preliminaries

This chapter introduces the fundamental concepts that will be utilized throughout the subsequent chapters of the manuscript. We begin by discussing the notions of direct and inverse problems, which form the core framework for understanding the dual nature of inference and generation in Restricted Boltzmann Machines (RBMs). Before diving into the analysis, we will first present some essential results from Statistical Mechanics that will be referenced in the main chapters. This will be done by initially exploring the Curie-Weiss model, followed by an examination of the generative properties of RBMs, starting with the influential Hopfield model [12]. The Hopfield model serves as one of the pioneering frameworks for understanding the mechanisms underlying associative memory [34, 41], providing key insights into energy-based models such as RBMs.

1.1 Statistical Inference

Inference is a crucial process that involves extracting meaningful information about a model of a specific environment from data. The environment is characterized by a distinct *signal*, also referred to as the *ground truth*, which, in our case, consists of a large set of N components. The goal of inference is to (at least partially) reconstruct this signal. In the following chapters, we will consider the case where both the size of the set and the amount of data, denoted by M, grow simultaneously, with N being very large. This regime is commonly referred to as *high-dimensional inference*. For our purposes, we require at least an extensive scaling of the number of observations with respect to the number of signal components, such that $M = \alpha N$, with $\alpha > 0$.
A key characteristic of real-world datasets is the inevitable presence of noise, which arises from unpredictable factors and affects the generative process that produces the observations. This necessitates a probabilistic approach to inference. Statistical inference [42] addresses this challenge by deducing properties of an underlying probabilistic model from noisy datasets. Among the available frameworks, the Bayesian approach is one of the most widely used, providing a structured methodology for integrating prior knowledge with observed data to make predictions about the signal.
We introduce the general random generative process with the following notation

$$\boldsymbol{x} \to \boldsymbol{Y} \, ,$$

where we consider $\boldsymbol{Y}$ as the data produced by some model, which is characterized by a set of paremeters $\boldsymbol{x}$. With the $\to$ symbol we are referring to the random process generating the data starting from the signal components. The parameters are known and fixed, while the (noisy) data is not, therefore it is modeled as a random variable. For this reason the study of this problem can be also recast as the investigation of the forward probability distribution $P(\boldsymbol{Y}|\boldsymbol{x})$. It involves the

description of the possible data configuration we could obtain starting from a particular instance of the parameters. Therefore we are trying to predict what data the model is going to generate.
The opposite case

$$\boldsymbol{y} \leftarrow \boldsymbol{X}\,,$$

is referred to as *inverse* problem. In this case the random process has already taken place and data are known. We would like to use the data to *infer* the parameters originating the process. Here $\boldsymbol{X}$ is modeled as a random variable and the investigation of its probability distribution $P(\boldsymbol{X}|\boldsymbol{y})$ derives the instance of the possible parameters responsable for data generation. The random generic process is modeled by the *likelihood* function

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{y}) := P(\boldsymbol{y}|\boldsymbol{x})\,.$$

The notation $P(\boldsymbol{y}|\boldsymbol{x})$ refers to the data formation step for given values of the unknown parameters. It has to be interpreted as a function of the parameters at given observations.
As one can already imagine the reconstruction process is based on a multitude of assumptions that can lead to an imperfect reconstruction of the signal. As an example one can ignore the nature of the model shaping the environment i.e. the form of the likelihood is incorrect, or having a partial a priori knowledge over the parameters $\boldsymbol{x}$, which will be encoded inside the prior term $P(\boldsymbol{x})$.

All the information (even if partial or incorrect) that we have on the data and on the process, are combined in the *Bayes formula*

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{\int dP(\boldsymbol{x}')P(\boldsymbol{y}'|\boldsymbol{x})} = \frac{P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{P(\boldsymbol{y})}\,. \tag{1.1}$$

Eq. (1.1) is known as *posterior* distribution and it contains all our *believes* about the parameters configuration, given the data. Another useful version of the Bayes formula is the following *chain rule*

$$P(\boldsymbol{x},\boldsymbol{y}) = P(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y}) = P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})\,. \tag{1.2}$$

Inference is the basis for machine learning problems. Typically, we want our models to uncover the underlying ground truth from the data itself, for this reason this method is always called *learning*. Once the machine learns from the data we can direct it towards a specific task, as an example data generation [43] or classification [44] . The machines used for performing inference are usually very large, in the sense they contains a huge number of parameters. This huge set can be tuned correctly only because of the presence of a huge number of observations, thus in a *high dimensional inference* setting. We will explore (chapters 2-3) the remarkable effectiveness of the learning mechanism when dealing with noisy datasets. In this context, we observe a significant improvement in signal retrieval, thanks to the availability of a large number of (noisy) examples.

Peculiar properties, deriving directly from the Bayes formula (1.1) are the *Nishimori identities*.

Proposition 1. *Let be (X,Y) a couple of random variables, with joint distribution $P(X,Y)$ and conditional distribution $P(X|Y)$. Take $k \geq 1$ i.i.d. random variables $X^{(1)}, X^{(2)}, \cdots, X^{(k)}$ with distribution $P(X|Y)$. Let us denote with the symbol $\mathbb{E}$ the expectation with respect to $P(X,Y)$, and with $\langle\cdot\rangle$ the one w.r.t. the product measure, then the following identity holds:*

$$\mathbb{E}\langle g\left(Y, X, X^{(2)}, X^{(3)}, \cdots X^{(k)}\right)\rangle = \mathbb{E}\langle g\left(Y, X^{(1)}, X^{(2)}, X^{(3)}, \cdots X^{(k)}\right)\rangle \tag{1.3}$$

Proof. from the chain rule $P(X,Y) = P(Y)P(X|Y)$ it is equivalent sampling the couple X,Y from the joint distribution $P(X,Y)$ or first samplping Y according to its marginal $P(Y)$ and then to draw X from the conditional one. Applying directly such information

$$\begin{aligned}
\mathbb{E}\langle g\left(Y,X,X^{(2)},X^{(3)},\cdots X^{(k)}\right)\rangle &= \mathbb{E}_{XY}\mathbb{E}_{X^{(2)}|Y}\cdots\mathbb{E}_{X^{(k)}|Y}g\left(Y,X,X^{(2)},X^{(3)},\cdots X^{(k)}\right)\\
&= \mathbb{E}_{Y}\mathbb{E}_{X|Y}\mathbb{E}_{X^{(2)}|Y}\cdots\mathbb{E}_{X^{(k)}|Y}g\left(Y,X,X^{(2)},X^{(3)},\cdots X^{(k)}\right)\\
&= \mathbb{E}_{Y}\mathbb{E}_{X^{(1)}|Y}\mathbb{E}_{X^{(2)}|Y}\cdots\mathbb{E}_{X^{(k)}|Y}g\left(Y,X^{(1)},X^{(2)},X^{(3)},\cdots X^{(k)}\right)\\
&= \mathbb{E}\langle g\left(Y,X^{(1)},X^{(2)},X^{(3)},\cdots X^{(k)}\right)\rangle\,.
\end{aligned}$$

□

This identity is a powerful tool and it can be used to carry on a huge amount of simplifications in studying high dimensional inference models [45]. When we are allowed to use such simplification it usually referred as *Bayesian optimal setting*. It is "optimal" since both priors and posterior are exactly known. This can be appreciated already from the proof of Eq. (1.3). Inside the proof we replaced X by $X^{(1)}$. This can be done only if we impose the correct likelihood function; then sampling $X^{(1)}$ from $P(X^{(1)}|Y)$ is equivalent to sampling it from $P(X|Y)$. Using another function the proof cannot be carried out. Instead, under the correct hypothesis, inside the average over all the random variables, the signal X can be replaced by a replica, drawned from the posterior.

Starting from Chapter 2 we will largely use this property. We are going to work in the mismatch prior setting showing that the inference problem can be mapped into a spin glass out of the Nishimori line, as so the previous identities will be obtained only with specific configurations.

The study of the posterior distribution (1.1) is then the most fundamental step when dealing with a inference problem. It is a way to understand whether data contain enough information so that the signal can be retrieved. At this level Statistical Mechanics comes into play. It helps us in the analysis of the posterior distribution, providing us with all the information to answer the previous question. We can interpret the posterior as a Boltzmann Gibbs distribution [46]

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})}{P(\boldsymbol{y})} = Z(\boldsymbol{y})^{-1}e^{-\beta\mathcal{H}(\boldsymbol{x};\boldsymbol{y})}\,, \tag{1.4}$$

where $\mathcal{H}(\boldsymbol{x};\boldsymbol{y})$ is called Hamiltonian of the system, or simply energy function, which encodes all the microscopic degrees of freedom of the considered system.

$$\begin{aligned}
\mathcal{H}(\boldsymbol{x};\boldsymbol{y}) &= -\log(P(\boldsymbol{x})) - \log(\mathcal{L}(\boldsymbol{x}|\boldsymbol{y}))\\
Z(\boldsymbol{y}) &= \sum_{\boldsymbol{x}} e^{-\mathcal{H}(\boldsymbol{x}|\boldsymbol{y})}\,.
\end{aligned}$$

In the language of disordered systems $\boldsymbol{x}$ plays the role of model's configurations and the data $\boldsymbol{y}$ that of quenched disordered interactions. In this notation the evidence is nothing but the partition function (Z) of the model.
This connection allows to interpret common estimators used in Bayesian inference under statistical machanics terms: for instance, the maximum-a-posteriori (MAP) estimator is equivalent to the ground state of the Hamiltonian in (1.4). Conversely, marginal probabilities computed on the posterior coincide with equilibrium expectation values over the Boltzmann measure.
To complete the Statistical mechanics connection one should add a fictitious temperature β

$$P_\beta(\boldsymbol{x}|\boldsymbol{y}) = Z(\boldsymbol{y},\beta)^{-1}e^{-\beta\mathcal{H}(\boldsymbol{x};\boldsymbol{y})} = Z(\boldsymbol{y},\beta)^{-1}\exp\left(\beta\log\mathcal{L}(\boldsymbol{x}|\boldsymbol{y}) + \beta\log P(\boldsymbol{x})\right)\,, \tag{1.5}$$

in this way the ground state of $\mathcal{H}$ can be obtained when $\beta \to \infty$. Typically, the prior distribution is factorized over the components $\boldsymbol{x}$, so that the second term in the exponential of (1.5) is interpreted

as a one-body interaction term, while all the remaining interacting parts are encoded in the likelihood function.
We will see, in the following, the deepest meaning of these quantities and the advantage of using a statistical mechanics perspective to investigate inference problems.

1.2 Statistical Mechanics formalism

Statistical mechanics aims to describe the thermodynamics of systems composed of a large number of components. It is now widely accepted that macroscopic physical phenomena are manifestations of underlying microscopic processes. Statistical mechanics serves as a bridge between these two levels, deriving the macroscopic physical laws of thermodynamics from a statistical average description of microscopic phenomena, as exemplified by the description of an ideal gas [46].
From a macroscopic perspective, such a system can be physically described using macroscopic observables such as pressure, volume, and temperature. From a microscopic perspective, the physical description of the same system must be based on the microscopic dynamics of its many components (molecules), each with its internal degrees of freedom. If the gas is allowed to relax over a sufficient amount of time, its macroscopic properties become fixed. However, at the microscopic level, there will be a vast number of states consistent with these fixed macroscopic constraints.
The microscopic system is then represented by a probability density function, which weights the collection of states compatible with the macroscopic constraints and generally depends on time. However, as said for the ideal gas, we are interested in determining macroscopic properties related to the equilibrium. After a sufficiently long time interval, observables become time-independent and therefore, the entire statistical mechanics analysis is restricted to cases where the probability density function of states does not depend explicitly on time. The complete information about the microscopic interactions between the microscopic degrees of freedom is encoded in the Hamiltonian $\mathcal{H}$, which can be generally expressed as a sum of k-body interactions:

$$\mathcal{H}(\boldsymbol{y}) = \sum_{k=1}^{N} \mathcal{H}^k(y_1, \cdots, y_N) \tag{1.6}$$

$$\mathcal{H}^k(\boldsymbol{y}) = \sum_{i_1, \cdots, i_k} J_{i_1, \cdots, i_k} \, y_{i_1}, \cdots, y_{i_k} \, , \tag{1.7}$$

Here, each y_i represents a generic degree of freedom. In this manuscript, we will mainly focus on systems with discrete degrees of freedom, with some finite set $\mathcal{X}$ representing the possible configurations of a single system's component y_i. For example, we will consider components that can be in two different states, $\mathcal{X} = \{-1, 1\}$. These degrees of freedom are often referred as *spins* and will be denoted by σ_i.
Typically, it is assumed that the system described by the Hamiltonian in Eq. (1.6) is in thermal equilibrium with an external temperature T. Under these conditions, the probability distribution governing the microscopic configurations in equilibrium is given by the Boltzmann distribution

$$P(\boldsymbol{y}) = Z^{-1} e^{-\beta \mathcal{H}(\boldsymbol{y})} \, , \tag{1.8}$$

where $\beta = 1/T$ is the inverse temperature. At zero temperature, the Boltzmann distribution (1.8) is concentrated on configurations that minimize the total energy.
The sign and nature of the couplings $J_{i_1, \cdots, i_k}$ determine the type of system being studied. For instance, if all couplings ($k \geq 2$) are positive the system is classified as ferromagnetic. Conversely,

if they are all negative, the system is an anti-ferromagnet. When the couplings include both positive and negative values, the system is categorized as a spin glass. Each coupling factor imposes a constraint on the spins it connects. Specifically, if $J_{i_1,\cdots,i_k} \geq 0$, the variables $y_1, \ldots, y_k$ must align (i.e., having the same sign) to minimize energy. Conversely, if $J_{i_1,\cdots,i_k} \leq 0$, the variables should anti-align to achieve energy minimization. A model is described as *unfrustrated* if a configuration that satisfies all constraints, exists. Otherwise, the model is considered *frustrated*, in which case, the ground state is defined as the configuration that violates the fewest amount of constraints.

The presence of frustration significantly alters the nature of phase transitions among these systems, distinguishing spin glasses (SG) from ferromagnets. In spin glasses, the interaction couplings are typically drawn from a specified distribution, allowing both positive and negative values. This variability in the couplings introduces *disorder*, a defining characteristic of spin glasses. Consequently, the energy landscape becomes dependent on the specific instance of the interaction parameters. To address this, the focus often shifts to studying the averaged behavior of such systems under disorder, aiming to reduce dependence on individual instances of the interactions. When the thermodynamic properties of the system exhibit minimal sample-to-sample fluctuations, the system is referred to as *self-averaging*, and its averaged description becomes a reliable representation. In the following sections, we provide specific examples to illustrate these concepts, beginning with the well-known Curie-Weiss ferromagnet and progressing to the spin-glass Hopfield model. We start by presenting a dynamical example that demonstrates how a system of spins evolves over time, eventually converging to states distributed according to the (time-independent) Boltzmann-Gibbs probability distribution. As a result, the system's properties are derived from the study of a probability distribution P over a finite configuration space Σ, where P represents the invariant (equilibrium) distribution of the dynamics, encapsulating all information about its long-term behavior.

Consider then a system consisting of N random spin variables, labeled $i = 1, \ldots, N$ with each spin σ_i taking values in $\mathcal{X} = \{-1, 1\}$. A configuration of the system is represented by the vector $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_N$, and the entire configuration space is $\Sigma = \mathcal{X}^N = \{-1, 1\}^N$. A simple stochastic dynamic for generating a configuration $\boldsymbol{\sigma}_t$ at time t can be described as follows:

- at $t = 0$ σ_i^0, $i = 1, \cdots, N$, are sampled independently and uniformly in $\{-1, 1\}$
- At each time step t the configuration $\boldsymbol{\sigma}^t$ is updated as follows:
- **Initialization**: At $t = 0$, the spins σ_i^0 are sampled independently and uniformly from $\{-1, 1\}$.
- **Evolution**: at each subsequent time step t, the configuration $\boldsymbol{\sigma}_t$ is updated according to the following procedure:

  1. **Select a Spin**: choose a single spin randomly from the set $i = 1, \ldots, N$.
  2. **Compute Magnetization**: calculate the direction similarity of the system, defined as the fraction of spins pointing up or down:

$$m^t = \frac{1}{N}\sum_i^N \sigma_i^t \quad m_i^t = m^t - \frac{\sigma_i^t}{N}$$

  3. **Flip Probability**: update the selected spin σ_i^t by flipping its direction with the following probability:

$$p_i^{\text{flip}} = \begin{cases} 1 & \text{if } m_i^t \sigma_i^t < 1 \\ \exp\left(-2\beta m_i^t\right) & \text{otherwise}, \end{cases} \tag{1.9}$$

> Here, $\beta \in \mathbb{R}^+$ is the inverse temperature, which takes into account external sources of randomness or errors in the system. This parameter reflects the level of stochasticity in the dynamics; with larger β corresponding to lower randomness and a stronger tendency of the system going towards its equilibrium configuration.

The update rule described above can be interpreted as a transition probability, $T_{\boldsymbol{\sigma}^t \to \boldsymbol{\sigma}^{(t+1)}}$, from the state $\boldsymbol{\sigma}^t$ to $\boldsymbol{\sigma}^{(t+1)}$. The sequence of configurations $\{\boldsymbol{\sigma}^t_{t\geq 0}\}$ forms a Markov chain [47] whose invariant distribution is the Boltzmann-Gibbs one

$$\begin{aligned} \mu_{\beta,N} &= Z^{-1}_{\beta,N} \exp\left(\tfrac{\beta}{N} \textstyle\sum_{i<j} \sigma_i \sigma_j\right) , \\ Z_{\beta,N} &= \textstyle\sum_{\boldsymbol{\sigma}\in\Sigma} \exp\left(\tfrac{\beta}{N} \textstyle\sum_{i<j} \sigma_i \sigma_j\right) . \end{aligned} \tag{1.10}$$

This distribution satisfies the detailed balance condition:

$$T_{\boldsymbol{\sigma}\to\boldsymbol{\sigma}'} \mu_{\beta,N}(\boldsymbol{\sigma}) = T_{\boldsymbol{\sigma}'\to\boldsymbol{\sigma}} \mu_{\beta,N}(\boldsymbol{\sigma}') .$$

This means that, after a transient period, the system thermalizes and becomes approximately stationary. As a consequence, any observable property of the system, which can be derived as a time average over the chain, can equivalently be reformulated as a static problem involving the invariant distribution.

The above equilibrium distribution is defined by the Hamiltonian $\mathcal{H}(\boldsymbol{\sigma}) = -\frac{1}{N}\sum_{i<j}\sigma_i\sigma_j$, as in Eq.(1.4). Then the Boltzmann-Gibbs distribution can be rewritten in the more compact way, as

$$P_\beta(\boldsymbol{\sigma}) \;\; = Z^{-1}_\beta \exp\left(-\beta\,\mathcal{H}(\boldsymbol{\sigma})\right) \tag{1.11}$$

with its corresponding partition function

$$Z_\beta \;\; = \textstyle\sum_{\boldsymbol{\sigma}\in\Sigma} \exp\left(-\beta\,\mathcal{H}(\boldsymbol{\sigma})\right) .$$

This formulation applies to a general function $\mathcal{H} : \Sigma \longrightarrow \mathbb{R}$ and describes any system whose interactions are encoded in the Hamiltonian $\mathcal{H}$.

The ultimate goal of Statistical Mechanics is to study the state of the system. The state is characterized by the expectation values of all possible observables

$$\langle \mathcal{O} \rangle = \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma})\mu_{\beta,N} = Z^{-1}_{\beta,N} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma}) e^{-\beta\mathcal{H}(\boldsymbol{\sigma})} ,$$

representing the outcomes of measurements. To compute the expectation value of any observable, it is necessary to determine the system's free energy,:

$$F_{\beta,N} := -\frac{1}{\beta} \log Z_{\beta,N} ,$$

which, as we will show in the following lines, serves as the generating function for the averaged observables. The link between observable measurements and the free energy involves techniques such as linear response theory, which provides a framework for understanding how the system responds to small perturbations

Theorem 1. ***(Linear response)***
Consider a perturbed Hamiltonian $\mathcal{H}(h, \boldsymbol{\sigma}) : [0, 1] \times \Sigma \to \mathbb{R}$ of the form

$$\mathcal{H}(h, \boldsymbol{\sigma}) = \mathcal{H}_0(\sigma) + h\mathcal{H}^1(\boldsymbol{\sigma}) \tag{1.12}$$

with $\mathcal{H}^1(\boldsymbol{\sigma}) = \mathcal{O}(\boldsymbol{\sigma})$ and indicate with

$$\langle \mathcal{O} \rangle_h = Z_{\beta,h}^{-1} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma}) e^{-\beta \mathcal{H}(h,\boldsymbol{\sigma})}$$

the average w.r.t. the Boltzmann-Gibbs distribution with the perturbed Hamiltonian at inverse temperature β. Then it holds

$$\frac{\partial}{\partial h} \log Z_{\beta,h}|_{h=0} = -\beta \langle \mathcal{H}^1 \rangle_0 \, .$$

Proof. By direct computation

$$\frac{\partial}{\partial h} \log Z_{\beta,h}|_{h=0} = Z_{\beta,h}^{-1} \sum_{\boldsymbol{\sigma}\in\Sigma} (-\beta \, \partial_h \mathcal{H}(h,\boldsymbol{\sigma}))$$

□

This result describes how we can generate the expectation of any observable using the free energy, which serves as a critical tool for identifying phase transitions. To compute the expected value of a generic observable $\mathcal{O}$, it is sufficient to perturb the original Hamiltonian by the observable itself, $\mathcal{H} \to \mathcal{H} + h\mathcal{O}$, with a sufficiently small external field h. Afterward, one can compute the free energy and its derivative with respect to h. Another important property of the free energy is captured by the following

Theorem 2. ***(Fluctuation-Dissipation)***
Consider the perturbed Hamiltonian $\mathcal{H}(h,\boldsymbol{\sigma})$ of (1.12) and the corresponding Boltzmann-Gibbs distribution at inverse temperature β. Then for any observable $\mathcal{O}$ it holds

$$\frac{\partial}{\partial h} \langle \mathcal{O} \rangle_h|_{h=0} = -\beta \langle \mathcal{O}; \mathcal{H}^1 \rangle_0 \, ,$$

where we indicte as

$$\langle \mathcal{O}; \mathcal{H}^1 \rangle_h = \langle \mathcal{O} \, \mathcal{H}^1 \rangle_h - \langle \mathcal{O} \rangle_h \langle \mathcal{H}^1 \rangle_h$$

the covariance between the observable and the perturbation, under the Boltzmann-Gibbs distribution.

Proof. Always by direct computation

$$\begin{aligned}
\frac{\partial}{\partial h} \langle \mathcal{O} \rangle_h &= Z_{\beta,h}^{-1} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma}) e^{-\beta \mathcal{H}(h,\boldsymbol{\sigma})} && (1.13)\\
&= -\beta Z_{\beta,h}^{-1} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma}) \mathcal{H}^1(\boldsymbol{\sigma}) e^{-\beta \mathcal{H}(h,\boldsymbol{\sigma})} + && (1.14)\\
&\quad - Z_{\beta,h}^{-2} \partial_h Z_{\beta,h} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{O}(\boldsymbol{\sigma}) e^{-\beta \mathcal{H}(h,\boldsymbol{\sigma})} && (1.15)\\
&= -\beta \left(\langle \mathcal{O} \, \mathcal{H}^1 \rangle_h - Z_{\beta,h}^{-1} \sum_{\boldsymbol{\sigma}\in\Sigma} \mathcal{H}^1(\boldsymbol{\sigma}) e^{-\beta \mathcal{H}(h,\boldsymbol{\sigma})} \, \langle \mathcal{O} \rangle_h \right) && (1.16)\\
&= \langle \mathcal{O}; \mathcal{H}^1 \rangle_h && (1.17)
\end{aligned}$$

which is concluded by the evaluation at $h = 0$. □

The Fluctuation-Dissipation Theorem connects two distinct quantities: the system's response to a perturbation and the fluctuations of the unperturbed system. To illustrate it, take $\mathcal{H}^1(\boldsymbol{\sigma}) = \mathcal{O}(\boldsymbol{\sigma})$

$$\frac{\partial}{\partial h}\langle \mathcal{O}\rangle_h|_{h=0} = -\beta\left(\langle(\mathcal{H}^1)^2\rangle_0 - \langle\mathcal{H}^1\rangle_0^2\right)$$

where the right-hand side is simply the variance of the observable $\mathcal{O}$ at equilibrium. Building on this, we can relate the Fluctuation-Dissipation Theorem to the linear response theory. Specifically, by differentiating the free energy $F_{\beta,N}$ with respect to h, we obtain

$$\frac{\partial^2}{\partial h^2}F_{\beta,h}|_{h=0} = -\beta\left(\langle(\mathcal{H}^1)^2\rangle_0 - \langle\mathcal{H}^1\rangle_0^2\right)\ ;$$

which shows that the variance (fluctuation) of an observable at equilibrium can be computed from the second derivative of the free energy and provides a connection between the equilibrium fluctuations and the response of the system to perturbations.
These results highlight the importance of the free energy, but its effective computation is often challenging. However, in the thermodynamic limit, where the number of components $N \to \infty$, the free energy density $f_\beta = F_{\beta,N}/N$ can be analyzed more efficiently. We will explore how this thermodynamic limit regime allows for powerful approximations.

1.2.1 Curie-Weiss model

Our first step is the application of the Statistical Mechanics formalism on the system which time evolution is described in Eq.(1.9). The model we are referring is called *Curie-Weiss model* (CW) [48] and it is depicted by the Hamiltonian

$$\mathcal{H}(\boldsymbol{\sigma}) = -\frac{1}{N}\sum_{i<j}\sigma_i\sigma_j - h\sum_i \sigma_i \tag{1.18}$$

where σ_i represents the spin at site i, N is the number of spins, and h is an external magnetic field, like the one in Thm(1- 2). This model serves as a prototypical example of a ferromagnetic mean-field model, since, as we are going to see, in the thermodynamic limit, its free energy density coincides with that of a system of identically independent spins. As described above, our goal is to obtain a complete description of the equilibrium properties of this system by determining the most probable configurations $\boldsymbol{\sigma} \in \Sigma$, sampled from the Boltzmann-Gibbs distribution $\mu_{\beta,N}$ (1.11), calculating the free energy at $N \to \infty$.
To compute the free energy density, we start with the partition function:

$$-\beta f_\beta = \lim_{N\to\infty}\frac{1}{N}Z_N(\beta)\,,$$
$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}\in\Sigma}\exp\left(\frac{\beta}{N}\sum_{i<j}\sigma_i\sigma_j + \beta h\sum_i\sigma_i\right)\,.$$

The Boltzmann factor can be rewritten in terms of the averaged alignment of the entire configuration, or simply the *magnetization*: $m = 1/N\sum_i\sigma_i$

$$Z_N(\beta) = \sum_{\boldsymbol{\sigma}\in\Sigma}\exp\left(\frac{N\beta}{2}m(\boldsymbol{\sigma})^2 + N\beta\, h\, m(\boldsymbol{\sigma}) - \frac{\beta}{2}\right)\,.$$

The last term, $-\beta/2$, is a constant additive factor that can be absorbed into the partition function, as it does not affect the Boltzmann distribution; moreover it is subleading in N, in the thermodynamic limit. Therefore, we redefine the partition function as follows

$$Z_N(\beta) = e^{-N\beta f(\beta)} = \sum_{\boldsymbol{\sigma}\in\Sigma} \exp\left(\frac{N\beta}{2} m(\boldsymbol{\sigma})^2 + N\beta\, h\, m(\boldsymbol{\sigma})\right) .$$

The free energy density in the thermodynamic limit can be derived using the following variational principle

$$-\beta f = \lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta) = \inf_N \frac{\log Z_N(\beta)}{N} = \sup_m \left[\log 2 + \log\cosh\left(\beta m\right) - \frac{1}{2}\beta m^2\right] . \tag{1.19}$$

This result comes from the fact that the partition function can be approximated by performing a Gaussian linearization

$$\mathbb{E}_g[e^{\lambda g}] = \int \mathcal{D}g\, e^{\lambda g} = e^{\frac{\lambda^2}{2}} , \tag{1.20}$$

where $\lambda \in \mathbb{R}$, $g \sim \mathcal{N}(0,1)$ and $\mathcal{D}g$ is the Gaussian measure. Applying the same steps to the partition function

$$\begin{aligned} Z_N(\beta) &= \sum_{\boldsymbol{\sigma}} e^{\frac{N\beta}{2} m(\boldsymbol{\sigma})^2 + N\beta h\ m(\boldsymbol{\sigma})} = \sum_{\boldsymbol{\sigma}} e^{\frac{1}{2}\left((\sqrt{\beta N} m(\boldsymbol{\sigma}))^2\right) + N\beta h\ m(\boldsymbol{\sigma})} \\ &= \sum_{\boldsymbol{\sigma}} \mathbb{E}_g\, e^{\sqrt{\beta N} m(\boldsymbol{\sigma}) g + N\beta h\ m(\boldsymbol{\sigma})} = \mathbb{E}_g \sum_{\boldsymbol{\sigma}} e^{(\sqrt{\frac{\beta}{N}} g + \beta\, h) \sum_i \sigma_i} \\ &= \int \frac{dg}{\sqrt{2\pi}} \left[2\cosh\left(\sqrt{\frac{\beta}{N}} g + \beta\, h\right)\right]^N . \end{aligned}$$

Using the change of variables $g\sqrt{\beta/N} = \beta m$, we have

$$Z_N(\beta) = \sqrt{\frac{\beta N}{2\pi}} \int dm\, \exp\left(N\left[\log 2 + \log\cosh(\beta m + \beta\, h) - \frac{\beta m^2}{2}\right]\right) . \tag{1.21}$$

Applying the Laplace or saddle point method to (1.21)

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta) &= \lim_{N\to\infty} \frac{1}{N} \log \int dm\, \exp\left(N\left[\log 2 + \log\cosh(\beta m + \beta\, h) - \frac{\beta m^2}{2}\right]\right) \\ &= \sup_m \left[\log 2 + \log\cosh(\beta m + \beta\, h) - \frac{\beta m^2}{2}\right] , \end{aligned}$$

in the second passage we can neglect the non extensive term $\mathcal{O}(\sqrt{N})$ and finally obtain the result stated in Eq.(1.19). The study of the variational principle defines different regimes for the possible configurations $\boldsymbol{\sigma}$ derived from (1.11), which are the equilibrium states of the dynamics (1.9). Calling

$$\hat{g}(\beta) = \sup_m g(m;\beta) = \sup_m \left[\log 2 + \log\cosh(\beta m + \beta h) - \frac{\beta m^2}{2}\right] , \tag{1.22}$$

the corresponding free energy density results as

$$f(\beta) = \inf_m \left[\frac{m^2}{2} - \frac{\log 2}{\beta} - \frac{1}{\beta}\log\cosh(\beta m + \beta h)\right] .$$

The transition behavior of the Curie-Weiss model can be resumed in the following

Proposition 2. *When $h = 0$, the critical inverse temperature $\beta_c = 1$ separates two sets of maximizers of the variational principles (1.22). If $\beta \leq \beta_c$ there exist a unique maximizer $\bar{m}(\beta) = 0$. For $\beta > \beta_c$ the maximum of $f(m;\beta)$ is reached at two symmetric points $\pm\bar{m}(\beta)$, with $\bar{m}(\beta) > 0$.*

Proof. $g(m;\beta)$ is an even function of m. We just restrict the proof when $m \geq 0$ and study the concavity of g only in this range of magnetizations. We have

$$\frac{1}{2m}\partial_m g(m;\beta)|_{h=0} = \partial^2_m g(m;\beta)|_{h=0} = \beta\left(\frac{\tanh(\beta m)}{m} - 1\right)$$

which is a decreasing function of m, thus f is a concave function of m^2. Consequently is has a unique maximum at $m = 0$, when its first derivative in $m = 0$ is negative

$$\partial^2_m g(m;\beta)|_{m=0} = \beta(\beta - 1) < 0\,,$$

giving $\beta < 1 = \beta_c$. Conversely it has a unique maximum at $m = \bar{m}(\beta)$. By simmetry the other maximum will be at $m = -\bar{m}(\beta)$. □

The values of the maximizers of g (which minimize the free enegy density f) can be obtained with the solution of the saddle point equation

$$m = \tanh(\beta m + \beta h)\,. \tag{1.23}$$

The numerical investigation of the solutions of Eq.(1.23), at $h = 0$, gives exactly $m = 0$ when $\beta < 1$ or $\pm\bar{m}(\beta) \neq 0$ otherwise. The two values of the magnetization signals the presence of a *phase transition* at $\beta_c = 1$:

- Paramagnetic (P) region: $m = 0$ and $\beta \leq \beta_c$. The highest probable configurations to be sampled $\boldsymbol{\sigma}$ are just random configuration.

- Ferromagnetic (F) region: $m = \pm\bar{m}(\beta)$ when $\beta > \beta_c$. The configurations $\boldsymbol{\sigma}$ are aligned configuration, which ordered is characterized by the value of $\bar{m}(\beta)$, solution of (1.23).

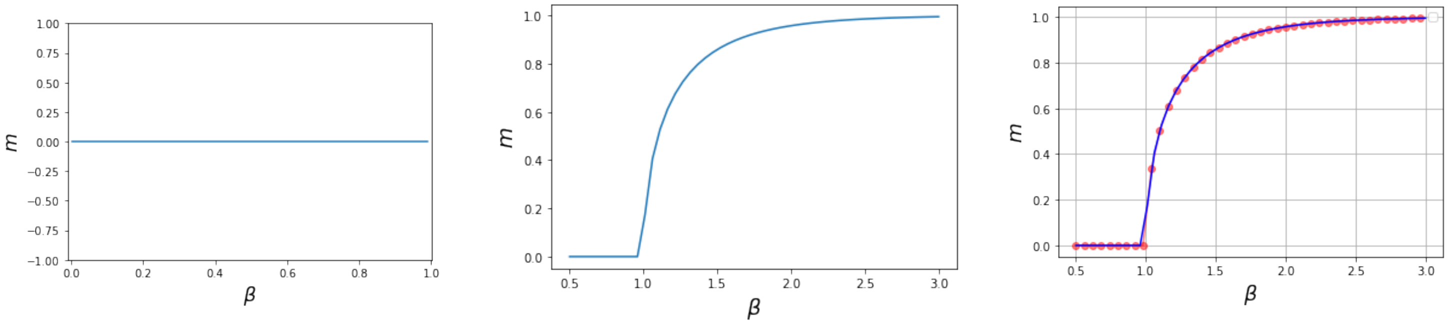


Figure 1.1: Magnetization of the CW model for different ranges of temperature. On the **Left**, at low temperature we are in a paramagnetic (P) phase. On the **Center** we see the P-F transition, while at the **Right** the agreement of with the magnetization obtained at the end of the dynamics (1.9). The data points are the one in red, the blue line is the transition obtained by the equilibrium analysis (1.23).

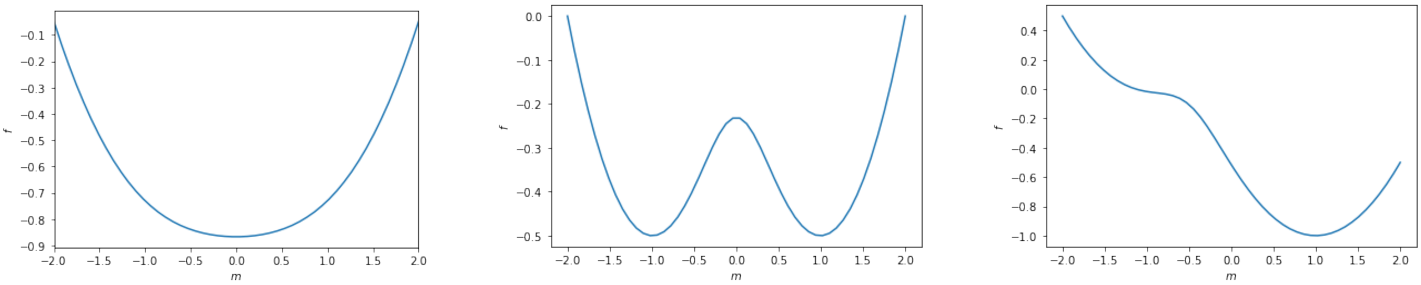


Figure 1.2: Free energy of the CW model. For high temperatures(**Left**) we have a unique maximum at $m = 0$. For $\beta > \beta_c = 1$ (**Center**) the free energy has two symmetric maxima. (**Right**) The presence of an external magnetic field select one of the two minimal configuration, when $\beta > \beta_c$. When $h > 0$ the positive magnetized configuration corresponds to the global minimum. The contrary happens when $h < 0$.

The minima of the free energy function for $\beta > \beta_c$ are often referred to as *pure states*. Unlike the typical Gibbs ones that represent an ensemble average over all configurations, pure states represent configurations, that singularly dominate the equilibrium behavior of the system. This behavior is illustrated in Fig. 1.2, where the free energy landscape transitions from having a single minimum $(m = 0)$ above the critical temperature to two minima $(m = \pm\bar{m})$ below it.

Mean field connection

The Curie-Weiss model is often described as a mean-field model. In this section we will see that it behaves as such in the thermodynamic limit $(N \to \infty)$

Proposition 3. *In the thermodynamic limit, the Curie-Weiss free energy density coincides with its mean-field approximation, i.e.*

$$\lim_{N\to\infty} \frac{1}{\beta N} \log Z_N(\beta) = \lim_{N\to\infty} \inf_{m\in[0,1]} \frac{G(P_m)}{N} \,,$$

where $G[.]$ is the Gibbs free energy and P_m is the product measure $P_m(\boldsymbol{\sigma}) = \prod_i p_m(\sigma_i)$, where p_m is a bernoully probability distribution with mean m.

Proof. First we give the definition of Gibbs free nergy:

$$G[P_m] = \langle H \rangle_{P_m} - \frac{1}{\beta} S[P_m]$$

where H is the internal energy

$$\begin{aligned} \langle H \rangle_{P_m} &= -\left\langle \frac{1}{N} \sum_{i<j}^{N} \sigma_i \sigma_j \right\rangle_{P_m} = -\frac{1}{N} \sum_{i<j} \langle \sigma_i \sigma_j \rangle_{P_m} = \\ &= -\frac{1}{N} \sum_{i<j} \langle \sigma_i \rangle_{P_m} \langle \sigma_j \rangle_{P_m} = -\frac{N(N-1)}{2N} m^2 \,. \end{aligned}$$

The second term S corresponds to the entropy

$$S[P_m] \;=\; -N \left[\frac{1+m}{2} \log\left(\frac{1+m}{2}\right) + \frac{1-m}{2} \log\left(\frac{1-m}{2}\right) \right] ,$$

therefore the functional to be optimized is

$$\frac{G[P_m]}{N} = -\frac{m^2}{2} + \frac{1}{\beta}\left[\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) + \frac{1-m}{2}\log\left(\frac{1-m}{2}\right)\right] + \mathcal{O}(1/N)\,.$$

Now, it is sufficient to compute the extremum of the mean-field approximation, and compare it with the solution of the variational formula (1.19). The extremum is given by

$$m = \frac{1}{2\beta}\log\left(\frac{1+m}{1-m}\right)\;;$$

using the identity $1/2\,\log\left(1+x/1-x\right) = \tanh^{-1}(x)$ the previous equation became exactly the optimal of the variational principle in (1.19): $m = \tanh(\beta m)$. □

This confirm that, in the thermodynamic limit, the CW distribution with zero external field, is well-approximated as a mixture distribution over the two pure states $m = \pm\bar{m}$, solution of (1.23). The probability distribution can be expressed as

$$P(\boldsymbol{\sigma}) = \frac{1}{2}(P_{\bar{m}}(\boldsymbol{\sigma}) + P_{\bar{m}}(\boldsymbol{\sigma})) \tag{1.24}$$

$$P_{\bar{m}}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} p_{\bar{m}}(\sigma_i) : p(\sigma_i)_{\bar{m}} = \frac{1+\bar{m}}{2}\delta_{\sigma_i,1} + \frac{1-\bar{m}}{2}\delta_{\sigma_i,-1} \tag{1.25}$$

When generating M independent samples from the CW distribution, it is possible to define a subset $V \subset \{1,\cdots,M\}$ s.t.

$$\begin{cases}\mu \in V \implies m^{\mu} = \bar{m}\\ \mu \notin V \implies m^{\mu} = -\bar{m}\end{cases} \tag{1.26}$$

This division highlights the coexistence of two distinct macroscopic behaviors in the ferromagnetic regime $(\beta > \beta_c)$.

Thermodynamic states

To complete the analysis of the CW model, we now give some results regarding the thermodynamic states, specifically the averages of the system's observables with respect to the Boltzmann-Gibbs distribution. These averages can be derived from the free energy through differentiation as in (1-1.27). We are interested in the limiting free energy density and in the interpretation of the observables in the same limit. Particularly we are going to use the linear part of the Hamiltonian (1.18) as perturbation $\mathcal{H}^1(\boldsymbol{\sigma}) = -\sum_i \sigma_i$, that will became our observable $\mathcal{O} = \mathcal{H}^1$.
Through the fluctuation dissipation relation

$$\begin{aligned}\langle\mathcal{O}\rangle &\underset{N\to\infty}{=} \frac{\partial}{\partial h}F|_{h=0}\,,\\ \langle m(\boldsymbol{\sigma})\rangle &\underset{N\to\infty}{=} \frac{\partial}{\partial h}f|_{h=0} = m\,,\end{aligned}$$

from which one can also derive $\langle m^2(\boldsymbol{\sigma})\rangle \underset{N\to\infty}{=} m^2$. Note that this result holds outside the line $h = 0, \beta \geq \beta_c$, where the averaged magnetization $\langle m(\boldsymbol{\sigma})\rangle \underset{N\to\infty}{=} 0$. This is because the limiting distribution of the magnetization has two symmetric peaks $\pm m$ asshown in (1.24-1.26).
For later purposes (Chapter 4) it is useful to obtain the scaling factor of the connected correlation

$C_{ij} = \frac{c_{ij}}{N} = \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$ of the CW model. To achieve this, we utilize again the fluctuation-dissipation relation

$$\begin{aligned}
\frac{\partial}{\partial h}\mathcal{O}|_{h=0} &= -\beta\left[\langle\left(\sum_i \sigma_i\right)^2\rangle - \langle\sum_i \sigma_i\rangle^2\right] , \\
\frac{\partial}{\partial h}\left(\sum_i \sigma_i\right)|_{h=0} &= \beta\left[\langle\left(\sum_i \sigma_i\right)^2\rangle - \langle\sum_i \sigma_i\rangle^2\right] .
\end{aligned} \tag{1.27}$$

Eq.(1.27) can be further developed. Decomposing the two squared ones one obtains

$$\begin{aligned}
\langle\left(\sum_i \sigma_i\right)^2\rangle - \langle\sum_i \sigma_i\rangle^2 &= N - \langle\sum_i \sigma_i\rangle^2 + \langle\sum_{i\neq j}\sigma_i\sigma_j\rangle - \sum_{i\neq j}\langle\sigma_i\rangle\langle\sigma_j\rangle \\
&= N\left(1-m^2\right) + N(N-1)C_{ij} .
\end{aligned}$$

It is possible to notice the l.h.s. is

$$\begin{aligned}
\frac{\partial}{\partial h}\left(\sum_i \sigma_i\right)|_{h=0} = N\frac{\partial\, m(\boldsymbol{\sigma})}{\partial h}|_{h=0} \\
= \beta N\left(1 - m^2 + (N-1)C_{ij}\right),
\end{aligned}$$

from which one obtain an expression of the correlation in function of the derivative of the magnetization

$$C_{ij} = \frac{\partial_h m(\boldsymbol{\sigma})|_0 - \beta(1-m^2)}{\beta(N-1)} \overset{N>>1}{\sim} \frac{\partial_h m(\boldsymbol{\sigma})|_0 - \beta(1-m^2)}{\beta N} . \tag{1.28}$$

In the thermodynamic limit we can use Eq.(1.23) to produce an explicit expression for $\partial_h m|_0$

$$\begin{aligned}
\frac{\partial m}{\partial h}|_{h=0} &= \beta(1-m^2)\left(\frac{\partial m}{\partial h}+1\right) , \\
\frac{\partial m}{\partial h}|_{h=0} &= \frac{\beta(1-m^2)}{1-\beta(1-m^2)} ;
\end{aligned}$$

Recollecting all pieces in (1.27), finally

$$C_{ij} = \frac{c_{ij}}{N} = \frac{1}{\beta N}\left(\frac{\beta(1-m^2)}{1-\beta(1-m^2)} - \beta(1-m^2)\right) = \frac{1}{N}\frac{(1-m^2)^2}{1-\beta(1-m^2)} . \tag{1.29}$$

1.3 Associative memory networks

The statistical mechanics approach to spin systems, particularly the study of spin glass models, has made significant contributions to the theoretical foundations and development of neural networks. Initially, these investigations aimed to model biological neurons [49], exploring intelligent behavior. This required a model that was both simple to analyze and complex enough to exhibit biologic macroscopic functionalities.
The focus of this chapter is on *memory recall*. We aim to recast this intricate process as an emergent property of large systems (thermodynamic limit), where individual components interact through simplified mechanisms compared to the complexities of biological systems. Rather than providing a detailed description of a biological brain, our objective is to provide a simple model capable of reproducing some of its fundamental properties. To achieve this, the basic units of the networks discussed here are conceptualized as *artificial neurons*. These neurons operate based on specific rules inspired by biological principles. One such rule is the Hebbian rule, which establishes a connection between stored memories and synaptic interactions among neurons [50].

A stored memory in the artificial brain is represented as a vector with N components, referred to as a pattern $\boldsymbol{\xi}^\mu$, where $\mu = 1, \cdots, p$ denotes the number of stored memories. The synaptic connections between neurons are

$$J_{ij} = \begin{cases} \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu & i \neq j\,, \\ 0 & i = j\,. \end{cases} \tag{1.30}$$

A neuron is formalized as a spin variable $\sigma_i \in \{-1, 1\}$, indicating its active or inactive state in response to external stimuli received from the other neurons. Then the dynamic process describing the state of a neuron is

$$\sigma_i(t+1) = \text{sign}\left(\sum_j J_{ij}\sigma_j(t) + \theta_i\right)\,, \quad \text{sign}(x) = \begin{cases} -1 & x \leq 0 \\ +1 & x > 0 \end{cases}\,. \tag{1.31}$$

In this framework, the i-th neuron receives signals from the others, quantified as $\sum_{j\neq i} J_{ij}\sigma_j(t)$. If the cumulative signal exceeds a certain threshold level (θ_i) the neuron i activates ($\sigma_i = 1$).

A pattern $\boldsymbol{\xi}^\mu$ is considered retrieved, if, after the dynamics stabilize, the network reaches a configuration $\boldsymbol{\sigma} = \boldsymbol{\xi}^\mu$. There is no reason to think the activation pattern of neurons $\boldsymbol{\xi}^\mu$ resemble real-world data, such as photographic images. This is due to the numerous biological processes involved in neural activation. Therefore the pattern components ξ_i^μ are modeled as identical independent distributed (i.i.d.) random variables; e.g. somewhere in the artificial brain some neurons are on $\sigma_i = 1,\ i = 1, \cdots, k$ and in the rest of the locations they are off.

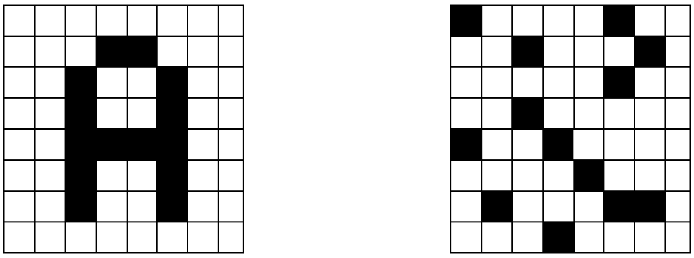


Figure 1.3: Difference between a realistic random configuration of a pattern (**right**) with respect to a simplified photographic letter configuration pattern (**left**).

In the Hopfield model [12] we consider uncorrelated pattern components ξ_i^μ as i.i.d Rademacher random variables. The successful retrieval of a saved pattern is evaluated through the similarity (overlap) between the configuration of the brain $\boldsymbol{\sigma}(t)$ and one pattern $\boldsymbol{\xi}^{\mu}$

$$m^{\mu}(t) = \frac{1}{N}\sum_{i} \sigma_i(t)\xi_i^{\mu} \,. \tag{1.32}$$

This overlap measure can also quantify the similarity between patterns.
When extracting the components of the two patterns, say ξ_i^1 and ξ_i^2, and compare them for each $i = 1, \cdots, N$, we are going to obtain as many matched components as the unmatched ones. This is due to their nature of uncorrelated i.i.d. random variables, which make them almost orthogonal:

$$\sum_i \xi_i^{\mu}\xi_i^{\nu} = \delta_{\mu\nu} + \mathcal{O}(1/\sqrt{N}) \,.$$

Using this condition it is possible to show that the patterns are stable configuration of the dynamics (1.31). For instance, starting from $\sigma_i(t) = \xi_i^1$

$$\begin{aligned} \sigma_i(t+1) &= \text{sign}\left(\sum_j w_{ij}\sigma_j(t)\right) \qquad (1.33)\\ &= \text{sign}\left(\frac{1}{N}\sum_{\mu}\sum_j \xi_i^{\mu}\xi_j^{\mu}\xi_j^1\right)\\ &= \text{sign}\left(\xi_i^1 + \frac{1}{N}\sum_{\mu\neq 1}\sum_j \xi_i^{\mu}\xi_j^{\mu}\xi_j^1\right)\\ &= \xi_i^1 \,\text{sign}\left(1 - a_i^1\right) \,. \end{aligned}$$

In the last step, we use the fact that ξ_i^1 is binary and called $a_i^1 = -\frac{1}{N}\sum_{\mu\neq 1}\sum_j \xi_i^1\xi_i^{\mu}\xi_j^{\mu}\xi_j^1$. If we want the first pattern to be stable we need $a_i^1 < 1$. Due to the random nature of the patterns components, a_i^1 behaves like a random walk of $N(p-1)$ steps and width $1/N$. This implies $a_i^1 < 1$, if $\sqrt{\frac{p-1}{N}} < 1$, i.e., when the number of patterns satisfy $p << N$, the set of saved memories are stable points of the presented dynamics.

Beyond stability, we aim to address *pattern completion* [51]. This involves extending the dynamics in Eq. (1.31) to a broader context, requiring the evolution from a random initial condition towards the memories.
The presented time evolution of spins, can be also interpreted as a zero-temperature process governed by the following Hamiltonian

$$\mathcal{H}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j \,, \tag{1.34}$$

here, the effective field acting on spin σ_i is $h_i^{eff} = \sum_j J_{ij}\sigma_j$. The process (1.31) tends to align the spins along the effective field direction at the next time step, leading towards decreasing energy states. If a one-to-one correspondence exists between stored patterns and energy minima, the network will converge to the nearest stored pattern, erasing noise from the initial condition. Moreover the rule (1.31) definitely determines the spin state at the next time step and doesn't consider for external perturbations that could deviates the dynamics. To account for external perturbations we extend the previous dynamics to a Glauber dynamics [14], with flipping rate $(1 + e^{-2\beta h^{eff}})^{-1}$. This process reduces to the original one when $\beta \to \infty$ and became completely random when $\beta \to 0$. As for the CW model, the invariant distribution resulting from the previously discussed dynamics is the Boltzmann-Gibbs distribution, with the Hamiltonian given in equation (1.34). If the study of the system's equilibrium properties reveals that the patterns correspond to equilibrium states, then the network dynamics will tend to converge toward the stored memories, even when the initial condition is far from equilibrium.

1.3.1 Hopfield Network

The Statistical mechanics approach to the Hopfield model starts from the energy Hamiltonian (1.34) that we report here, replacing the interactions with their explicit Hebb rule representation

$$H(\boldsymbol{\sigma}|\{\boldsymbol{\xi}^p\}) = -\frac{\beta}{2N}\sum_{\mu=1}^{p}\sum_{i=1}^{N}\sum_{j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\,. \tag{1.35}$$

Here each $\xi_i^\mu \sim \frac{1}{2}(\delta_1 + \delta_{-1})$ independently, making the Hopfield network a spin-glass model. We will first analyze how memory retrieval works with a finite number of patterns p [34]. This means that in the thermodynamic limit, the ratio $p/N \to 0$. This is exactly the condition needed in Eq.(1.33) to ensure the patterns as stable dynamical configuration.

The diagonal part of the interactions is not taken into account since it is not extensive in the size of the system N. The energy function is constructed by a sum of quadratic terms each one representing the contribution of each one of the patterns. As stated in the previous section we proceed with the analysis of the partition function

$$Z(\beta,\{\boldsymbol{\xi}^\mu\}) = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma};\{\boldsymbol{\xi}^p\})} = \sum_{\boldsymbol{\sigma}}\exp\left\{\frac{\beta}{2N}\sum_{\mu=1}^{p}\sum_{i=1}^{N}(\sigma_i\xi_i^\mu)^2\right\}. \tag{1.36}$$

By introducing a set of integration variables (one for each pattern) m^μ we can linearize the quadratic term in the exponential

$$\begin{aligned} Z(\beta,\{\boldsymbol{\xi}^p\}) &= \int\prod_\mu dm^\mu\sum_{\boldsymbol{\sigma}}\exp\left\{-\frac{\beta N}{2}\sum_\mu m_\mu^2 + \beta\sum_\mu m^\mu\sum_i\sigma_i\xi_i^\mu\right\} \\ &= \int\prod_\mu dm^\mu\exp\left\{-\frac{\beta N}{2}\sum_\mu m_\mu^2 + \sum_i\log\left(2\cosh\left(\beta\sum_\mu m^\mu\xi_i^\mu\right)\right)\right\}, \end{aligned} \tag{1.37}$$

where we used the gaussian integration (1.20) and omit the normalization constant, since it does not affect physical properties of the system. Given the large number of neurons in the brain ($\sim 10^{11}$) we are going to discuss phase transitions in the limit $N \to \infty$. As for the CW model, under this condition the integral can be evaluated by steepest descent, with free energy density

$$-\beta f = -\frac{\beta}{2}\boldsymbol{m}^2 + \frac{1}{N}\sum_i\log\left(2\cosh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi}_i)\right), \tag{1.38}$$

where $\boldsymbol{m} = (m^1,\cdots,m^p)$ and $\boldsymbol{\xi}_i = (\xi_i^1,\cdots,\xi_i^p)$. The extremization condition of the free energy gives the equation of state

$$\boldsymbol{m} = \frac{1}{N}\sum_i\boldsymbol{\xi}_i\tanh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi}_i)\,, \tag{1.39}$$

the parameter $\boldsymbol{m}$, appearing as order parameter in Eq.(1.38) is nothing but the limiting expected overlap similarity of Eq,(1.32). This connection can be revealed using the saddle point condition, starting from the first line of Eq.(1.37):

$$m^\mu = \frac{1}{N}\sum_i\sigma_i\xi_i^\mu\,. \tag{1.40}$$

In the limit of large N, the sum over i in (1.37) becomes equivalent to the average over the random components of the vector $\boldsymbol{\xi} = (\xi^1,...,\xi^p) = (\xi^1 = \pm1,...,\xi^p = \pm1)$, according to the self-averaging property [52]. This corresponds to the configurational average over disorder, which we denote in this case by $\mathbb{E}_{\boldsymbol{\xi}}[\cdot]$,

$$-\beta f = -\frac{\beta}{2}\boldsymbol{m}^2 + \mathbb{E}_{\boldsymbol{\xi}}\Big[\log\Big(2\cosh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi})\Big)\Big]\,, \tag{1.41}$$

$$\boldsymbol{m} = \mathbb{E}_{\boldsymbol{\xi}}\Big[\boldsymbol{\xi}\tanh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi})\Big]\,. \tag{1.42}$$

For the purpose of this thesis we restrict to the case of a single pattern retrieval (even if other solutions of (1.42) exist [14]). To be consistent with the choice made in (1.33) we assume the first pattern to be retrieved, i.e. $m_1 = m$, $m_2 = m_3 = \cdots = m_p = 0$. Then the equation of state (1.42) simplifies considerably

$$m = \mathbb{E}_{\xi}\Big[\xi\tanh(\beta\,m\cdot\xi)\Big] = \tanh(\beta\,m)\,, \tag{1.43}$$

which behaves exactly as the Curie-Weiss model of the previous sections, at $h = 0$. This means the presence of two symmetric energy minima $\pm\boldsymbol{\xi}^1$ under $T = 1$, leading a random initial condition towards the nearer minimum point, following the dynamics (1.31). Concluding, when the number of pattern p is finite, we have obtained *pattern completion*.

Many saved memories

It is possible to go beyond the finite number retrieval of memories. We break the assumption of finite embedded memories, considering an extensive number of patterns: $p/N \to \alpha$. When the scale α becomes too large the network becomes overloaded with information, leading to interference between stored patterns. As a result, the system may enter a disordered phase, known as the *spin glass phase*, even at low temperatures. In this state, the neural configuration becomes randomly frozen, unable to retrieve any specific memory.

In this case, the computation of the configurational average of the free energy density, relies on the calculation of the configurational average of the n^{th} power partition function, the so called *replica trick*:

$$-\beta f(\beta,\alpha) = \lim_{N\to\infty}\frac{1}{N}\,[\log Z]^{\boldsymbol{S}} = \lim_{\substack{N\to\infty\\ n\to 0}}\frac{\log[Z^n]^{\boldsymbol{S}}}{Nn}\,.$$

The thermodynamic states description of the Hopfield network is realized on a system of many replicas, idexed by $a = 1, \cdots, n$ which is described by a set of order parameters

$$m^a = Q(\boldsymbol{\sigma}^a, \boldsymbol{\xi})\,,$$
$$q^{ab} = Q(\boldsymbol{\sigma}^a, \boldsymbol{\sigma}^b)\,,$$

respectively the magnetization of one of the replicas with the chosen pattern to be retrieved and the (spin-glass) overlap between two different replicas. The presence of these two sets of order parameters divides the phase diagram of the Hopfield network as in Fig.1.4

- Paramagnetic (P) region: $m = q = 0$;
- Pattern retrieval (R) region: $m \neq 0$, $q > 0$;
- Spin Glass (SG) region: $m = 0$, $q > 0$.

The phase diagram of Fig.1.4 is obtained under the *Replica Symmetric* (RS) assumption, that simplify the number of order parameters up to two: $m^a = m,\ \forall a = 1, \cdots, n$ and $q^{ab} = q,\ \forall a, b =$

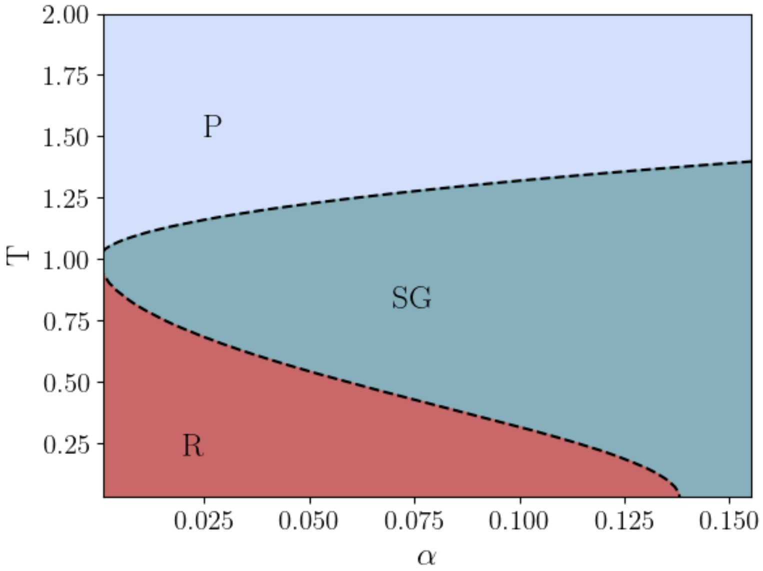


Figure 1.4: Hopfield model phase diagram when the number of patterns becomes extensive. It is possible to see three phases: paramagnetic (P) and spin-glass (SG), where the solution is random and an ordered phase (R), where it is possible to retrieve one pattern.

$1, \cdots, n$, giving the following free energy density and saddle point equations

$$-\beta f^{RS}(\beta,\alpha) = \text{Extr}\left[-\frac{\alpha}{2}\ln\left(1-\beta+\beta q\right)+\frac{\alpha}{2}\frac{\beta q}{1-\beta+\beta q}+\frac{\beta^2\alpha}{2}\hat{q}(q-1)+\right.$$
$$\left.-\frac{\beta}{2}m^2+\ln 2+\mathbb{E}_\sigma\int\mathcal{D}z\ln\cosh\beta\left(m\sigma+z\sqrt{\alpha\hat{q}}\right)\right],$$

$$\begin{aligned}
p =&\mathbb{E}_\sigma\int Dz\tanh\beta(p\sigma+z\sqrt{\alpha\hat{q}})\sigma\,,\\
q =&\mathbb{E}_\sigma\int Dz\tanh^2[\beta(p\sigma+z\sqrt{\alpha\hat{q}})]\,,\\
\hat{q} =&\frac{\alpha\beta^2 q}{(1-\beta+\beta q)^2}\,.
\end{aligned}$$

The RS solution says that for $\alpha \lesssim 0.138$ it is possible to recover the selected pattern. It is necessary to precise that the RS solution is unstable due to replica symmetry breaking (RSB) at very low temperatures. The RSB region (which requires the introduction of a collection of infinite order parameters) is small and the qualitative behavior of the order parameter m is well described by the RS solution. In conclusion retrieval, thus memorization, is possible up to a critical load, beyond which a spin-glass regime occurs where the system gets confused [41, 49].

1.3.2 Connection with the Restricted Boltzmann Machine

In the previous section, we explored how the Hopfield network effectively achieves pattern formation, successfully implementing associative memory. Interestingly, this model is a special case of a broader class of powerful energy-based models known as Restricted Boltzmann Machines (RBMs) [53]. Consequently, studying the Hopfield model serves as a milestone for understanding the generative processes underlying modern ML algorithms.

To illustrate the connection, we start with the Hamiltonian (1.35) in the extensive case and combine it with the definition of magnetization (1.40):

$$\mathcal{H}(\boldsymbol{\sigma}|\{\boldsymbol{\xi}^\mu\}) = -\sum_{\mu}^{p} N(m^\mu)^2 \,,$$

where m^μ denotes the magnetization. The Hamiltonian term within the partition function (1.36) can then be linearized by introducing an extensive number of normal random variables, leading to

$$Z(\beta, \{\boldsymbol{\xi}^\mu\}) = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\tau}} \exp\left(\beta \sum_{i}^{N}\sum_{\mu}^{p} \sigma_i \xi_i^\mu \tau^\mu\right).$$

From the ML perspective this model has a natural application in the context of *unsupervised* learning, where a machine is trained over a dataset (training set) of unlabeled examples to retrieve their probability distribution. To this task the Hopfield model belongs to the class RBMs [54, 55]. The first (visible) group of variables $\boldsymbol{\sigma}$, reproduces the data with components $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$. The second (hidden) group has the role of building an internal representation of the data structure using its units $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$.

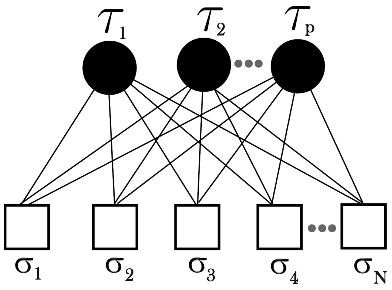


Figure 1.5: Bipartite representation of a RBM. The neurons are divided in two layers: hidden ($\boldsymbol{\tau}$) and visible ($\boldsymbol{\sigma}$). Only neurons of different layers are connected by synaptic interactions.

Together, they form a bipartite network structure, as illustrated in Fig.1.5, characterized by the parametric probability distribution

$$P_{\boldsymbol{w}}(\boldsymbol{s}) = z_{\boldsymbol{w}}^{-1} \mathbb{E}_{\boldsymbol{\tau}} \exp\left(\sum_{i,j}^{N,P} w_i^j \sigma_i \tau_j\right), \tag{1.44}$$

whose parameters $\boldsymbol{w} \in \mathbb{R}^{NP}$ need to be fit with the data. In statistical mechanics, RBMs are studied in the context of multi-species spin-glass models [56–63]. A notable advantage of the RBM framework is its flexibility in selecting different priors for the hidden nodes, enabling the introduction of higher-order interactions among the visible nodes [27]. For instance, instead of restricting hidden units to Gaussian distributions and data units to binary spins, one can generalize their set of priors. An example of such generalization involves choosing one of the RBM components $x \in \{\sigma, \tau, \xi\}$ from a mixture distribution

$$x = \sqrt{\Omega_x} g_x + \sqrt{1 - \Omega_x} \epsilon_x \,,$$

where $\Omega_x \in [0, 1]$, $g_x \sim N(0, 1)$ and ϵ_x is a Rademacher random variables taking values ±1. By applying the same replica symmetry (RS) analysis, used for the original Hopfield model, one can investigate the impact of these priors on the retrieval phase of RBMs, see Fig.1.6.

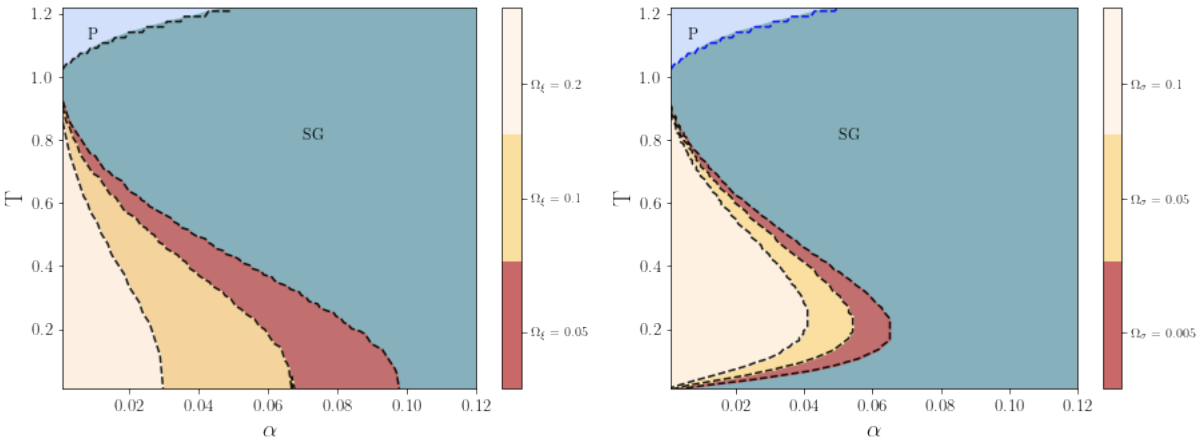


Figure 1.6: Examples of phase diagrams representing the retrieval transition in RBM models. The represented phases are the ones in Fig.1.4. **Left**: effect on the retrieval phase when the RBM's weights $\boldsymbol{\xi}$ have Gaussian tails, while the hidden units have $\Omega_\tau = 1$. The retrieval region shrinks the more weights display a Gaussian nature. **Right**: Same effect displayed when $\boldsymbol{\sigma}$ have Gaussian tails while having binary patterns and Gaussian hidden units. Here the effect is the same as in the left panel.

The computation leading to Fig. 1.6 can be derived from Appendix D by setting $\hat{\beta} = 0$, where the order parameter p plays the role of magnetization. Alternatively, these results can be found in [64, 65]. In the broader context of Boltzmann machines, the retrieval of patterns in associative neural networks exhibits a richer behavior. Specifically, retrieval remains feasible even at high load for any pattern distribution that interpolates between Boolean and Gaussian statistics.

Chapter 2

Dual Hopfield model

Up to this point, we have provided a brief explanation of how Hopfield networks serve as paradigmatic models for *memory storage* and retrieval—an illustrative example of the direct process described in Sec. 1.1. This concept extends naturally to Restricted Boltzmann Machines (RBMs), which generalize the Hopfield model. In contrast, modern artificial intelligence systems primarily rely on the machine *learning* paradigm, rather than memorization.

This chapter is devoted to demonstrating that such learning mechanism, aligns precisely with the problem of self-supervised learning [66], where a machine is trained to infer the probability distribution of a dataset using a subset of unlabeled examples. [1]
We show that a teacher-student self-supervised learning problem can be formulated using Boltzmann machines as a suitable generalization of the Hopfield model with structured patterns. In this framework, the spin variables represent the machine's weights, and the patterns correspond to examples from the training set. The learning performance is analyzed by examining the phase diagram in terms of key parameters: the size of the training set, the noise in the dataset, and the inference temperature (which corresponds to weight regularization).
In scenarios where the dataset is small but informative, the machine can achieve learning through memorization. However, when the dataset is noisy, a critical threshold of extensive examples is required. Beyond this threshold, the limits of memory storage provide an opportunity for the emergence of a new learning regime where the system learns by generalization.

2.1 Introduction and Motivations

What is the maximum amount of patterns that can be stored by a neural network and efficiently retrieved? What is the minimum amount of examples needed for a neural network to understand the hidden structure of a noisy, high dimensional dataset? They seem two different questions, the former concerning the limit of a memorization mechanism, the latter involving the beginning of a learning process. In facts they are strictly related, being a common life experience, especially for science students, that learning comes to help when memory starts to fail.

Since the Hopfield seminal work, several generalizations have been investigated in relation to their critical storage capacity and retrieval capabilities. For example, super-linear capacity has been found by allowing multi-body interactions [67], parallel retrieval has been studied in relation to patterns sparsity [68–72] or hierarchical interactions [73–76] and non-universality has been shown

[1] The term self-supervised learning is often considered synonymous with the classical concept of *unsupervised* learning, as opposed to supervised learning, which requires labeled data.

with respect to more general patterns entries and unit priors [64, 77–81]. A specific attention has been given to model with intra pattern and among patterns correlation [82–85]. In this works the negative effects of the correlation structure on the system's capacity emerge and different non-Hebb rules are proposed to mitigate it, restoring the possibility of pattern retrieval. Other works on the effect of patterns correlation on the critical capacity are [15–17].
From this perspective it appears as a collective endeavour in favour of an artificial intelligence (AI) that works exclusively by machine memorization, in contrast to modern AI that is being dominated by machine learning. Recent models have started to consider patterns correlation as connected to the existence of a new regime which is more close to learning than memory retrieval. For example in [28, 29, 86] patterns are blurred copies of some prototypes while in [87–89] patterns are generated from a set of hidden features: in both cases there exists a regime where the Hopfield network can extract (learn) the underlying structure more than merely memorize the examples.

We start giving a generalization of the Hopfield model with planted correlated patterns, together with its natural interpretation in terms of the machine weights posterior distribution in a teacher-student self-supervised learning problem. Successively, the model is studied in three different regimes: in the Bayes optimal [33, 90] setting the machine can work by memorization if the dataset noise is relatively small while it can retrieve the signal by generalization when the training set is made of a sufficiently high number of weakly informative examples; beyond Bayes-optimality [91, 92] the learning regime depends on the relation between dataset noise and weight regularization (inference temperature). In the Appendix the proofs of the main results are given together with the derivation of the conjectures.

We consider a Hopfield model with N binary spins $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N) \in \{-1, 1\}^N$ and M quenched random binary patterns $\boldsymbol{\mathcal{S}} := \{\boldsymbol{s}^\mu\}_{\mu=1}^M = \{s_1^\mu, \dots, s_N^\mu\}_{\mu=1}^M \in \{-1, 1\}^{NM}$. Given a specific realization of the patterns and a planted configuration $\hat{\boldsymbol{\xi}} \in \{-1, 1\}^N$, we consider the Boltzmann-Gibbs distribution

$$\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\xi}}) = \hat{Z}^{-1}(\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\xi}}) \exp\left(\frac{\beta}{N}\sum_{\mu=1}^{M}\sum_{i<j}^{N} s_i^\mu s_j^\mu \xi_i \xi_j + \lambda \sum_{i=1}^{N} \hat{\xi}_i \xi_i\right), \tag{2.1}$$

where $\beta \geq 0$ is the inverse temperature, $\lambda \geq 0$ is the amplitude of an external field in the direction of the planted configuration and

$$Z(\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\xi}}) = \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta}{N}\sum_{\mu=1}^{M}\sum_{i<j}^{N} s_i^\mu s_j^\mu \xi_i \xi_j + \lambda \sum_{i=1}^{N} \hat{\xi}_i \xi_i\right) \tag{2.2}$$

is the model partition function. The term *planted pattern* or *planted configuration* is peculiar of the self-supervised problem. We are saying on top of our system we "plant" the solution $\hat{\boldsymbol{\xi}}$, which will be the signal to be retrieved by the student after interacting with the data. The solution will be corrupted by noise, in our case the inverse temperature $\hat{\beta}$ of the teacher. Consequently we assume the student patterns are independent but with a specific spatial correlation induced by the planted configuration i.e. they are distributed according to

$$P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) = \prod_{\mu=1}^{M} P(\boldsymbol{s}^\mu|\hat{\boldsymbol{\xi}}) = \prod_{\mu=1}^{M} z_\mu^{-1} \exp\left(\frac{\beta}{N}\sum_{i<j}^{N} \hat{\xi}_i \hat{\xi}_j s_i^\mu s_j^\mu + h^\mu \sum_{i=1}^{N} \hat{\xi}_i s_i^\mu\right), \tag{2.3}$$

in contrast with Sec.1.3, where patterns consist of i.i.d. Rademacher random variables. The partition function

$$z_\mu := \sum_{\boldsymbol{s}} \exp\left(\frac{\beta}{N}\sum_{i<j}^{N} \hat{\xi}_i \hat{\xi}_j s_i s_j + h^\mu \sum_{i=1}^{N} \hat{\xi}_i s_i^\mu\right) = \sum_{\boldsymbol{s}} \exp\left(\frac{\beta}{N}\sum_{i<j}^{N} s_i s_j + h^\mu \sum_{i=1}^{N} s_i^\mu\right), \tag{2.4}$$

is nothing but the partition function of the classical Curie-Weiss model at inverse temperature β and external field $h^\mu \in \mathbb{R}$. In the second equality we just use the Gauge transformation $s_i \to s_i \hat{\xi}_i$. Finally we assume the planted configuration $\hat{\boldsymbol{\xi}}$ is a quenched Rademacher random vector.
The model (2.1, 2.3), in particular in the limit $\lambda, \boldsymbol{h} \to 0$, is motivated by its natural application in the context of self-supervised learning with RBMs. In this setting, given an unknown probability distribution P^0 and a training set of M examples $\boldsymbol{\mathcal{S}} = \{\boldsymbol{s}^\mu \sim P^0\}_{\mu=1}^M$ drawn independently from P^0, the aim is to approximately learn P^0 with

$$P_{\boldsymbol{w}}(\boldsymbol{s}) = z_{\boldsymbol{w}}^{-1} \mathbb{E}_{\boldsymbol{\tau}} \exp\left(\sum_{i,j}^{N,P} w_i^j s_i \tau_j\right), \tag{2.5}$$

by tuning $\boldsymbol{w}$. The learning performance clearly depends on the structure of the data, i.e. P^0, the properties of the machine, i.e. $P_{\boldsymbol{w}}$, and the amount of data. A crucial research question is thus about the typical size of the training set necessary for the machine to efficiently learn, given its architecture and the structure of the data. To explore this question we consider a controlled scenario where the dataset $\boldsymbol{\mathcal{S}}$ is generated from a (teacher) machine $P_{\hat{\boldsymbol{w}}}$ and another (student) machine $P_{\boldsymbol{w}}$ is trained over $\boldsymbol{\mathcal{S}}$. In the case $P = 1$, choosing $\tau \sim \mathcal{N}(0,1)$ and $\hat{\boldsymbol{w}} = \sqrt{\beta/N}\hat{\boldsymbol{\xi}}$, we find that the probability distribution of the dataset is exactly as in Eq.(2.3), i.e. it has a spatial correlation induced by the planted configuration $\hat{\boldsymbol{\xi}}$. The parameter β^{-1} can be interpreted either as the amount of noise in the training set examples or as the same time the typical strength of the machine weights, i.e. weights regularization. In a Bayesian framework, the posterior distribution of the student's machine weights $\boldsymbol{w} = \sqrt{\beta/N}\boldsymbol{\xi}$ given the dataset reads as

$$\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}) = \frac{P(\boldsymbol{\xi}) \prod_{\mu=1}^M P(\boldsymbol{s}^\mu|\boldsymbol{\xi})}{P(\boldsymbol{\mathcal{S}})} = Z^{-1}(\boldsymbol{\mathcal{S}}) \exp\left(\frac{\beta}{N}\sum_{\mu=1}^M \sum_{i<j}^N s_i^\mu s_j^\mu \xi_i \xi_j\right), \tag{2.6}$$

which is exactly the Hopfield model of Eq.(2.1) in absence of fields. In the following we refer to the machine weights $\boldsymbol{\xi}$ as the student pattern to distinguish them from the teacher (planted) pattern $\hat{\boldsymbol{\xi}}$, also denoted as the *signal*. In the statistical inference framework of Sec.1.1, Eq.(2.3) define the so called *direct* Hopfield model describing the dataset, while Eq. (2.6) can be considered as the corresponding *inverse* model. Interestingly it is still a Hopfield model, *Dual* w.r.t. the direct model, where the spin variables correspond to the machine weights $\boldsymbol{\xi}$ while the training set's examples $\boldsymbol{\mathcal{S}}$ play the role of the patterns. We also refer to the dual patterns as planted disorder to enlighten that they are in turn drawn from a Hopfield model with a planted pattern, i.e. $\hat{\boldsymbol{\xi}}$. This teacher-student setting has been introduced in [65], originally inspired by [93] and recently studied in the case $P = 2$ for boolean RBMs [94]. It was also used in [30] to propose alternative posterior based methods for inverse problems in the case of structured dataset.

In this setting it is possible to quantify the learning performance of the student machine by measuring how much the student pattern $\boldsymbol{\xi}$ is close to the signal $\hat{\boldsymbol{\xi}}$, i.e. by computing $Q(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}})$, once introduced the overlap between two vectors $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2 \in \mathbb{R}^N$ as

$$Q(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) = \frac{\boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^2}{N}. \tag{2.7}$$

Similarly we can evaluate the amount of information contained in the data with the overlap $Q(\boldsymbol{s}^\mu, \hat{\boldsymbol{\xi}})$ between the examples and the teacher pattern. Finally we can characterize the memorization performance of the machine by introducing the overlap $Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})$ between the examples and the student pattern (sampled from the posterior (2.6)). We define the bracket $\hat{\langle . \rangle}$ as the expected value w.r.t the joint distribution of signal, training set and student pattern $(\hat{\boldsymbol{\xi}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\xi})$, i.e. for any function

$f : \{-1,1\}^N \times \{-1,1\}^{NM} \times \{-1,1\}^N \to \mathbb{R}$,

$$\langle f \rangle^{\hat{}} := 2^{-N} \sum_{\hat{\boldsymbol{\xi}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\xi}} f(\hat{\boldsymbol{\xi}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\xi}) \hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\xi}}) P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}). \tag{2.8}$$

It is interesting to note that using the Gauge transformation $\xi_i \to \xi_i \hat{\xi}_i$ and $s_i^\mu \to s_i^\mu \hat{\xi}_i$, we get for any bounded function f of the overlap that

$$\begin{aligned} \left\langle f(Q(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}})) \right\rangle^{\hat{}} &= 2^{-N} \sum_{\hat{\boldsymbol{\xi}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\xi}} f(Q(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}})) \hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\xi}}) P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) \\ &= \sum_{\boldsymbol{\mathcal{S}}, \boldsymbol{\xi}} f(Q(\boldsymbol{\xi}, \mathbf{1})) \hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \mathbf{1}) P(\boldsymbol{\mathcal{S}}|\mathbf{1}) = \langle f(Q(\boldsymbol{\xi}, \mathbf{1})) \rangle, \end{aligned} \tag{2.9}$$

where we have defined the bracket $\langle . \rangle$ as the expectation w.r.t. the joint distribution $\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \mathbf{1}) P(\boldsymbol{\mathcal{S}}|\mathbf{1})$ that does not depend on the signal $\hat{\boldsymbol{\xi}}$ (ferromagnetic gauge), i.e. for any function $g : \{-1,1\}^{NM} \times \{-1,1\}^N \to \mathbb{R}$,

$$\langle g \rangle := \sum_{\boldsymbol{\mathcal{S}}, \boldsymbol{\xi}} g(\boldsymbol{\mathcal{S}}, \boldsymbol{\xi}) \hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \mathbf{1}) P(\boldsymbol{\mathcal{S}}|\mathbf{1}). \tag{2.10}$$

Note that $\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}, \mathbf{1})$ is the Boltzmann-Gibbs distribution induced by the partition function

$$Z(\boldsymbol{\mathcal{S}}) = \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta}{N} \sum_{\mu=1}^{M} \sum_{i<j}^{N} s_i^\mu s_j^\mu \xi_i \xi_j + \lambda \sum_{i=1}^{N} \xi_i \right), \tag{2.11}$$

that is a Hopfield model in a field, whose patterns are drawn from

$$P(\boldsymbol{\mathcal{S}}|\mathbf{1}) = \prod_{\mu=1}^{M} z_\mu^{-1} \exp\left(\frac{\beta}{N} \sum_{i<j} s_i^\mu s_j^\mu + h^\mu \sum_{i=1}^{N} s_i^\mu \right), \tag{2.12}$$

i.e. the distribution of M independent Curie-Weiss models at inverse temperature β and external field $\boldsymbol{h}$. In the following we indicate with $\mathbb{E}_{\beta,\boldsymbol{h}}$ the expectation w.r.t. the pattern distribution (2.12). Eq. (2.9) means that the overlap with the signal can be interpreted as the magnetization of a Hopfield model at inverse temperature β with patterns $\boldsymbol{\mathcal{S}}$ extracted independently from a Curie-Weiss model at the same temperature. Analogously it holds

$$\langle f(Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})) \rangle^{\hat{}} = \langle f(Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})) \rangle \tag{2.13}$$

$$\left\langle f(Q(\boldsymbol{s}^\mu, \hat{\boldsymbol{\xi}})) \right\rangle^{\hat{}} = \langle f(Q(\boldsymbol{s}^\mu, \mathbf{1})) \rangle \tag{2.14}$$

i.e. the overlap between the examples and the signal corresponds, in the ferromagnetic gauge, to the magnetization of $\boldsymbol{\mathcal{S}}$. For this reason in the following we always consider, without loss of generality, that the patterns are drawn from (2.12), which does not depend on the signal $\hat{\boldsymbol{\xi}}$, keeping in mind that a non-zero magnetization in this model corresponds to a macroscopic alignment with the planted configuration $\hat{\boldsymbol{\xi}}$.

2.2 Learning by memorization from few highly informative examples

At low enough temperature we expect that the examples $\boldsymbol{\mathcal{S}}$ are polarized, i.e. in terms of the original variables they are correlated with the planted configuration $\hat{\boldsymbol{\xi}}$. At the same time, as long as the number of examples do not exceed the critical load of the student machine, the student pattern $\boldsymbol{\xi}$ will be aligned with one of them and therefore we expect it to be polarized as well. From the machine learning perspective, since each example carries a lot of information about the signal, the student machine can easily learn $\hat{\boldsymbol{\xi}}$ by memorization, even if $M = 1$.

We can formalize this result and precisely quantify the goodness of the learning performance in terms of the temperature β and the amount of data M by computing the system's free energy

$$f_N = -\frac{1}{\beta N}\mathbb{E}_{\beta,\boldsymbol{h}} \log Z(\boldsymbol{\mathcal{S}}) \tag{2.15}$$

where $Z(\boldsymbol{\mathcal{S}})$ is given by Eq. (2.11). In the thermodynamic limit we have the following

Theorem 3. *If $\beta \leq 1$, or $\beta > 1$ and $\lambda, h \in \mathbb{R} \setminus \{0\}$, it holds for any $\boldsymbol{\epsilon} \in \{-1, 1\}^M$*

$$-\beta f \ := \ \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\beta,h\boldsymbol{\epsilon}} \log Z(\boldsymbol{\mathcal{S}}) = \sup\nolimits_{\boldsymbol{p}\in\mathbb{R}^M} g(\boldsymbol{p}) \tag{2.16}$$

with

$$g(\boldsymbol{p}) = \log 2 - \frac{\beta \boldsymbol{p}^2}{2} + \langle \log\cosh\left(\beta \boldsymbol{s}\cdot\boldsymbol{p} + \lambda\right)\rangle_{\boldsymbol{s}}\,, \tag{2.17}$$

where we have defined the random vector $\boldsymbol{s} \in \{-1, 1\}^M$ whose entries are i.i.d. random variables with mean

$$m_0(\beta, h) := \mathrm{argmax}_{x\in\mathbb{R}} \left[\log 2 + \log\cosh(\beta x + h) - \frac{\beta x^2}{2}\right]. \tag{2.18}$$

The variational principle expressing the free energy corresponds to what one would expect for a Hopfield model at a low load of biased patterns and this is natural since the model for the dual patterns (2.3), as well as the Curie Weiss model (2.12), is mean-field in the thermodynamic limit, therefore spatial correlations vanish. The solution of the variational principle is a stationary point of $g(\boldsymbol{p})$, therefore a solution of

$$\boldsymbol{p} = \langle \boldsymbol{s}\tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p} + \lambda)\rangle_{\boldsymbol{s}} = \frac{\sum_{\boldsymbol{s}} \boldsymbol{s}\, e^{\beta m_0 \boldsymbol{1}\cdot\boldsymbol{s}} \tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p} + \lambda)}{(2\cosh(\beta m_0))^M}. \tag{2.19}$$

Note that Theorem 3 states that the limiting free energy doesn't depend on $\boldsymbol{\epsilon}$, therefore we can consider without loss of generality the case of a positive uniform external field $\boldsymbol{h} = h\boldsymbol{\epsilon} = h\boldsymbol{1}$ acting on the examples. In terms of the learning scenario this corresponds to saying that the student does not care whether the examples are aligned or anti-aligned with the planted configuration $\hat{\boldsymbol{\xi}}$. Note that this is not a symmetry by global spin flipping because each example field can have a different sign and thus a different alignment w.r.t $\hat{\boldsymbol{\xi}}$. This result is particularly useful because in the limit of zero external field it is well known that the Curie Weiss measure at low temperature is a mixture measure $P^{CW} = 1/2(P^{m_0} + P^{-m_0})$ and consequently the training set $\boldsymbol{\mathcal{S}}$ is in general composed of two clusters of examples with opposite global magnetization: this is in general a complication when dealing with inverse problems [30, 95, 96]. Nevertheless, using a Bayesian approach and thanks to the resulting Hebbian interaction, the posterior distribution works exactly as the examples were all aligned in the same direction.

From the solution of the free energy variational principle all the model order parameters can be derived according to the following

Proposition 4. *Assuming $\boldsymbol{\epsilon} = \mathbf{1}$ and given $\boldsymbol{p} \in \mathbb{R}^M$ the global maximizer of $g(\boldsymbol{p})$, then it holds*

$$\bullet\ \lim_{N\to\infty} \left\langle Q(\boldsymbol{s}^\mu, \hat{\boldsymbol{\xi}}) \right\rangle^{\hat{}} = \lim_{N\to\infty} \left\langle \frac{\boldsymbol{s}^\mu \cdot \mathbf{1}}{N} \right\rangle = m_0(\beta, h) \qquad \forall \mu = 1, \dots, M; \tag{2.20}$$

$$\bullet\ \lim_{N\to\infty} \langle Q(\boldsymbol{s}^\mu, \boldsymbol{\xi}) \rangle^{\hat{}} = \lim_{N\to\infty} \left\langle \frac{\boldsymbol{s}^\mu \cdot \boldsymbol{\xi}}{N} \right\rangle = p^\mu \qquad \forall \mu = 1, \dots, M; \tag{2.21}$$

$$\bullet\ \lim_{N\to\infty} \left\langle Q(\hat{\boldsymbol{\xi}}, \boldsymbol{\xi}) \right\rangle^{\hat{}} = \lim_{N\to\infty} \left\langle \frac{\mathbf{1} \cdot \boldsymbol{\xi}}{N} \right\rangle = m = \langle \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p} + \lambda) \rangle_{\boldsymbol{s}} . \tag{2.22}$$

Moreover the random variables $Q(\boldsymbol{s}^\mu, \hat{\boldsymbol{\xi}})$, $Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})$ and $Q(\hat{\boldsymbol{\xi}}, \boldsymbol{\xi})$ are self-averaging.

Since we are interested in the limit $h, \lambda \to 0$, we need to study the system of equations

$$\boldsymbol{p} = \langle \boldsymbol{s} \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p}) \rangle_{\boldsymbol{s}} = \frac{\sum_{\boldsymbol{s}} \boldsymbol{s}\, e^{\beta m_0(\beta) \mathbf{1} \cdot \boldsymbol{s}} \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p})}{(2 \cosh(\beta m_0(\beta)))^M}, \tag{2.23}$$

where $m_0(\beta) = m_0(\beta, 0^+)$. In the case $M = 1$ the equation (2.23) takes the form

$$p = \langle s \tanh(\beta s p) \rangle_s = \tanh(\beta p). \tag{2.24}$$

whose solutions are $p = \pm m_0(\beta)$ and from which

$$m = \langle \tanh(\beta s p) \rangle_s = \langle s \rangle_s \tanh(\beta p) = m_0 \tanh(\beta p) = \pm m_0(\beta)^2, \tag{2.25}$$

A similar ferromagnetic bifurcation occurs also for generic number of examples M. In fact when $\beta \leq 1$ the only solution for the example magnetization is $m_0 = 0$, therefore the average $\langle \cdots \rangle_{\boldsymbol{s}}$ becomes uniform over $\{-1, 1\}^M$. As a consequence, using $|\tanh(z)| < |z|$, it holds

$$\begin{aligned} \boldsymbol{p}^2 &= \langle (\boldsymbol{s} \cdot \boldsymbol{p}) \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p}) \rangle_{\boldsymbol{s}} \leq \langle |\boldsymbol{s} \cdot \boldsymbol{p}| | \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p})| \rangle_{\boldsymbol{s}} \\ &\leq \beta \left\langle (\boldsymbol{s} \cdot \boldsymbol{p})^2 \right\rangle_{\boldsymbol{s}} = \beta \sum_{\mu, \nu} p_\mu p_\nu \langle s_\mu s_\nu \rangle_{\boldsymbol{s}} = \beta \boldsymbol{p}^2. \end{aligned} \tag{2.26}$$

Hence, if $\beta < 1$ the only solution is $\boldsymbol{p} = 0$, from which $m = 0$. This means that at high temperature there is no information about the planted pattern in a dataset composed of a finite number of examples, simply because they are uncorrelated with the signal $\hat{\boldsymbol{\xi}}$. It is immediate to verify that $\boldsymbol{p} = \mathbf{0}$ is a solution at any temperature, however, when $\beta > 1$ it is unstable and other solutions with non zero overlap $\boldsymbol{p}$ appear. It is possible to characterize those solutions thanks to the following propositions.

Proposition 5. *The solutions of Eqs. (2.23) have equal components, i.e. $p^\mu = \bar{p}\ \forall \mu = 1, \dots, M$.*

Thanks to Proposition 5 it is sufficient to solve the one-dimensional equation

$$\bar{p} = \left\langle s_1 \tanh(\beta \bar{p} \sum_{\mu=1}^{M} s_\mu) \right\rangle_{\boldsymbol{s}} = \frac{\sum_{\boldsymbol{s}} s_1\, e^{\beta m_0 \sum_{\mu=1}^{M} s_\mu} \tanh(\beta \bar{p} \sum_{\mu=1}^{M} s_\mu)}{(2 \cosh(\beta m_0))^M}, \tag{2.27}$$

where $\bar{p}$ is the value of each component of $\boldsymbol{p}$. By studying Eq. (2.27) we have the following

Proposition 6. *The points $p^\mu = \bar{p} = \pm m_0(\beta)$, $\forall \mu = 1, \dots, M$, are solutions of eqs. (2.23).*

To check whether they are actually minimizers one should look at the free energy, obtaining the following

Proposition 7. *For $\beta > 1$, $\lambda = 0$ and $h = 0^+$, the maximum of g is attained in the two symmetric points $p^\mu = \pm m_0(\beta)$, $\forall \mu = 1, \dots, M$.*

The previous results show that when the temperature is low enough, i.e. $\beta > 1$, both the example magnetization m_0 and the overlap of the system with the examples $\bar{p}$ are different from zero. This means that the examples are all macroscopically aligned with the signal $\hat{\boldsymbol{\xi}}$ (i.e. they are largely informative) and that $\boldsymbol{\xi}$ is macroscopically aligned with all the examples. As a consequence the system must be macroscopically aligned with the signal (learning is possible and easy) and in fact in this regime $m \neq 0$. It is interesting to note that while $\bar{p}$ does not depend on M, i.e. the alignment with the examples only depends from the posterior temperature, the system magnetization m, i.e. the alignment with the signal, increases with the size of the dataset. In fact its value is simply obtained as

$$m = \langle \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p}) \rangle_{\boldsymbol{s}} = \frac{\sum_{\boldsymbol{s}} e^{\beta M m_0 m(\boldsymbol{s})} \tanh(\beta M \bar{p} m(\boldsymbol{s}))}{(2\cosh(\beta m_0))^M}, \tag{2.28}$$

where $m(\boldsymbol{s}) = M^{-1} \sum_{\mu=1}^{M} s^\mu$. We report in Figure 2.1 the value of the magnetization as a function of the temperature for different size of the dataset M. As M increases, it is evident that the magnetization tends to 1 for $\beta > 1$. The system's free energy

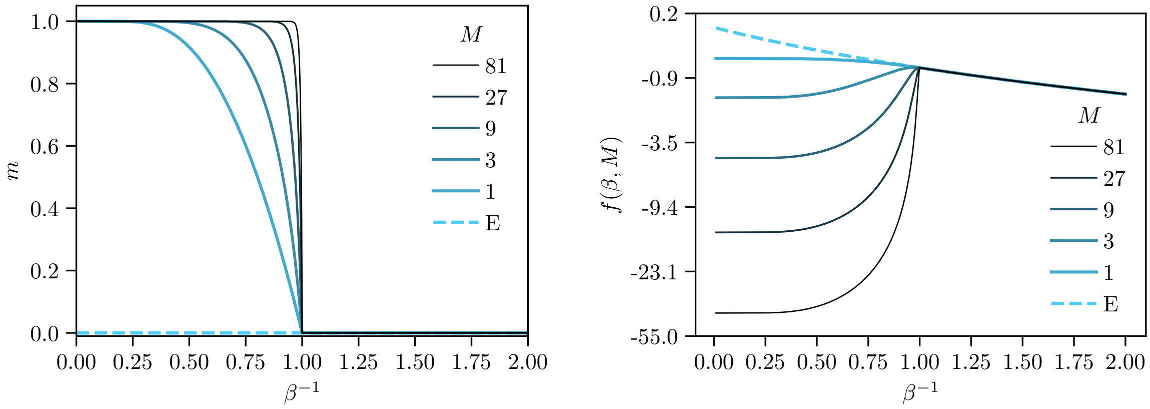


Figure 2.1: Learning performance with a finite M dataset size. *Left:* System's magnetization, i.e. the overlap between teacher and student pattern is evaluated as a function of the temperature β^{-1}. The overlap increases with the number of examples M as long as the system is below the critical temperature. *Right:* System's free energy as a function of β^{-1}. The free energy corresponding to solutions with $m > 0$ are painted in solid lines, while the ergodic (E) free energy, i.e. the one corresponding to $m = 0$, appears with a dashed line. As long as $\beta^{-1} < 1$, the global minimum of the free energy is the state where the machine can learn the original pattern.

$$f = \frac{M}{2}\bar{p}^2 - \frac{1}{\beta} \frac{\sum_{\boldsymbol{s}} e^{\beta M m_0 m(\boldsymbol{s})} \ln 2\cosh(\beta M \bar{p} m(\boldsymbol{s}))}{(2\cosh(\beta m_0))^M} \tag{2.29}$$

is also displayed to show the instability of the solution $\bar{p} = 0$ at low temperature.

2.3 Learning by generalization from many noisy examples

In the previous Section we have shown that when the examples are highly correlated with the signal, i.e. at $\beta > 1$, learning is possible and easy, with any finite number of examples $M > 0$.

Conversely when the examples are poorly correlated with the signal, i.e. $\beta < 1$, there is not enough information in the posterior distribution to retrieve the original pattern. In this Section we show that learning is possible also in the low correlation regime (high temperature) as long as we consider a larger dataset. In particular we consider the case in which the machine can leverage on an extensive number of examples, i.e.

$$\lim_{N\to\infty} M/N = \alpha > 0. \tag{2.30}$$

In this regime the free energy can only be derived exploiting the replica method under the replica symmetric approximation [14] from which one gets the following

Conjecture 1. *For $\lambda = 0$, $\boldsymbol{h} = 0$, $\beta < 1$ and $\alpha > 0$, the limiting free energy of the posterior distribution is*

$$-\beta f = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\beta,\boldsymbol{h}} Z(\boldsymbol{\mathcal{S}}) = \mathrm{Extr}_{m,\hat{m},q,\hat{q}}\, f(m,\hat{m},q,\hat{q}), \tag{2.31}$$

where

$$\begin{aligned} f(m,\hat{m},q,\hat{q}) = &-\frac{\alpha}{2}\left[\ln\big((1-\beta)(1-\beta+\beta q)\big) - \frac{\beta q}{(1-\beta)} + \frac{\beta^2(q^2-m^2)}{(1-\beta)(1-\beta+\beta q)}\right] \\ &+\frac{\hat{q}q}{2} - \hat{m}m - \frac{\hat{q}}{2} + \int D\mu(z)\ln 2\cosh(\hat{m}+z\sqrt{\hat{q}}). \end{aligned} \tag{2.32}$$

Thus the saddle point equations read as

$$m = \int D\mu(z)\tanh(\hat{m}+z\sqrt{\hat{q}}) \tag{2.33}$$

$$\hat{m} = \frac{\alpha\beta^2 m}{(1-\beta)(1-\beta+\beta q)} \tag{2.34}$$

$$q = \int D\mu(z)\tanh^2(\hat{m}+z\sqrt{\hat{q}}) \tag{2.35}$$

$$\hat{q} = \frac{\alpha\beta^3(m^2-q^2)}{(1-\beta)(1-\beta+\beta q)^2} + \frac{\alpha\beta^2 q}{(1-\beta)(1-\beta+\beta q)}. \tag{2.36}$$

The solution of the saddle point equations have a physical interpretation in terms of the model's order parameters according to the following

Conjecture 2. *Given (m,q) solutions of Eqs. (2.33,2.35) it holds*

$$m = \lim_{N\to\infty}\left\langle \frac{\mathbf{1}\cdot\boldsymbol{\xi}}{N}\right\rangle = \lim_{N\to\infty}\left\langle Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\right\rangle^{\hat{}} \tag{2.37}$$

$$q = \lim_{N\to\infty}\left\langle Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)\right\rangle = \lim_{N\to\infty}\left\langle Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)\right\rangle^{\hat{}}, \tag{2.38}$$

where $\boldsymbol{\xi}^1,\boldsymbol{\xi}^2$ are two replicas of the systems, i.e. two independent configurations sampled from the posterior distribution (2.1) with the same data $\boldsymbol{\mathcal{S}}$.

The details about the derivation of Conjectures 1 and 2 are provided in Appendix B. As in Theorem 3 the free energy is given as the solution of a variational principle in terms of the model' s order parameters: in that case the overlap with the examples $p^\mu \sim Q(\boldsymbol{s}^\mu,\boldsymbol{\xi})$. In this case, because of the high temperature ($\beta < 1$) the system never aligns with any specific examples ($\boldsymbol{p} = 0$) and this overlap does not play a role in the free energy principle. Nevertheless, the signal (toward the teacher pattern) carried by an extensive number of examples can become macroscopic and could bring to non zero system's magnetization and system's overlap (m and q) that in fact emerge as the two natural order parameters. Eqs. (2.33) and (2.35) are similar to those of the standard Hopfield model in Sec.1.3.1, with a random gaussian field of a different variance, still proportional to the load α. Moreover the signal term $\hat{m}$ doesn't point towards the examples but towards the teacher pattern and it is proportional to α, thus showing the beneficial effect of the training set size.

It is important to recall that we are studying the problem in which the student machine is exactly as the teacher one (same architecture) and also the temperatures are the same. Therefore by construction the model satisfies the Nishimori conditions, in particular

$$\langle Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)\rangle^{\hat{}} = \left\langle Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\right\rangle^{\hat{}}. \tag{2.39}$$

Infact we have at $\lambda = 0$ and $\boldsymbol{h} = 0$ that

$$\begin{aligned}\left\langle Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\right\rangle^{\hat{}} &= 2^{-N}\sum_{\hat{\boldsymbol{\xi}},\boldsymbol{\mathcal{S}},\boldsymbol{\xi}} Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}})P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) \\ &= \sum_{\hat{\boldsymbol{\xi}},\boldsymbol{\mathcal{S}},\boldsymbol{\xi}} Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}})P(\hat{\boldsymbol{\xi}}|\boldsymbol{\mathcal{S}})P(\boldsymbol{\mathcal{S}}) \\ &= 2^{-N}\sum_{\boldsymbol{\mathcal{S}},\boldsymbol{\xi}^1,\boldsymbol{\xi}^2} Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)P(\boldsymbol{\xi}^1|\boldsymbol{\mathcal{S}})P(\boldsymbol{\xi}^2|\boldsymbol{\mathcal{S}})P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) =: \langle Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2)\rangle^{\hat{}}.\end{aligned} \tag{2.40}$$

For this reason, according to Conjecture 2, we expect that the solution of the saddle point equations satisfies $m = q$ and at the same time, see Eqs. (2.34,2.36) $\hat{m} = \hat{q}$. It is easy to show that this is in fact a solution of Eqs. (2.33-2.36) by using the identity

$$\int D\mu(z)\tanh(\hat{m} + z\sqrt{\hat{m}}) = \int D\mu(z)\tanh^2(\hat{m} + z\sqrt{\hat{m}}). \tag{2.41}$$

We checked numerically this is a stable solution. This condition indicates the absence of a spin-glass region ($m = 0$, $q > 0$). Analogously it is easy to show [14] that overlap and magnetization have the same distribution. The expected self-averaging of the system's magnetization motivates the belief that the model is replica symmetric and that conjectures 1 and 2 therefore hold [25, 57]. Figs. 2.2 show the value of the magnetization, i.e. the learning performance, as a function of β and α. It is evident the occurrence of a second order phase transition from a paramagnetic region where the only solution is $m = q = 0$ to a ferromagnetic region where $m = q > 0$ and learning is feasible. The phase transition occurs at a critical temperature $\beta_c(\alpha)$ if we fix the size of the dataset α or equivalently at a critical size $\alpha_c(\beta)$ for a given level of the temperature. The critical line can be obtained analytically by studying the reduced equation

$$m = \int D\mu(z)\tanh(\hat{m}(m) + z\sqrt{\hat{m}(m)}) \tag{2.42}$$

where

$$\hat{m}(m) = \frac{\alpha\beta^2 m}{(1-\beta)(1-\beta+\beta m)}. \tag{2.43}$$

By expanding for small values of the magnetization we get

$$m = \frac{\alpha\beta^2}{(1-\beta)^2}m + o(m) \tag{2.44}$$

from which the bifurcation must be at $\alpha\beta^2/(1-\beta)^2 = 1$, i.e. at

$$\beta_c^{-1}(\alpha) := 1 + \sqrt{\alpha} \qquad \alpha_c(\beta) := \frac{1-\beta^2}{\beta^2}. \tag{2.45}$$

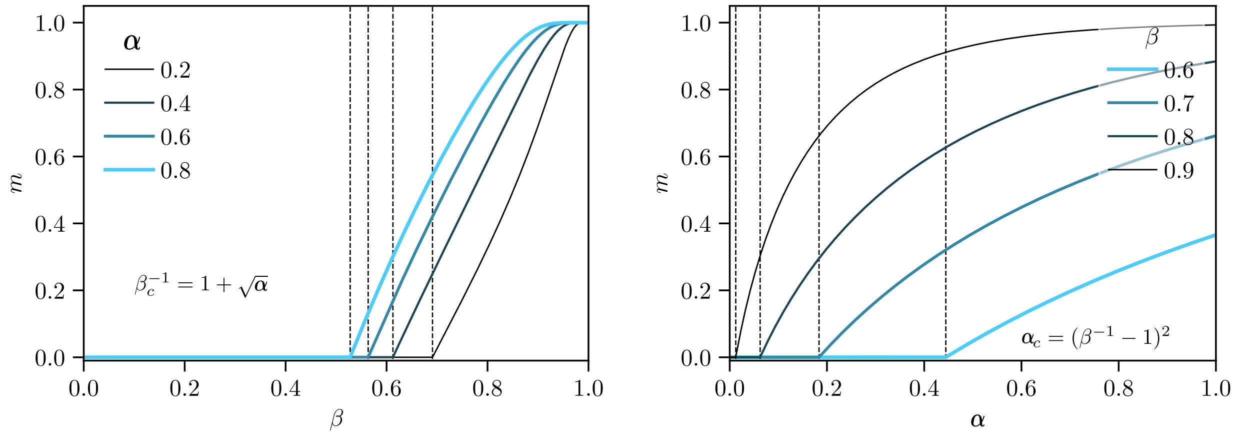


Figure 2.2: Learning performance with a noisy ($\beta < 1$) but extensive ($M = \alpha N$) dataset on the *Nishimori* line. *Left*: the system's magnetisation m is shown as a function of the inverse temperature β and for different dataset size α. *Right*: the magnetisation m is shown as a function of α and different inverse temperatures β. The inferred pattern's quality displays a second order phase transition. Moreover it increases with α and decreases with the dataset noise β^{-1}.

As expected the critical size is an increasing function of the data temperature β^{-1}, thus of the data correlation with the signal. What is interesting is that, despite we are in the regime in which each example s^μ does not share macroscopic correlation with the signal ($m_0(\beta) = 0$), still the machine is able to retrieve it $m > 0$ as soon as the dataset is sufficiently large. It means that the dataset contains enough information but divided in many (an extensive number of examples) small (poorly correlated examples) pieces.

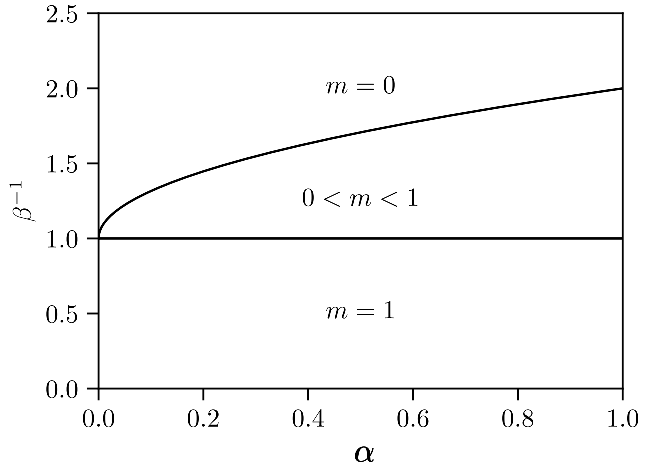


Figure 2.3: Phase diagram of the model on the *Nishimori* line. For $\beta^{-1} > 1 + \sqrt{\alpha}$, the student machine is in the paramagnetic phase with $m = 0$, where learning is impossible. Conversely it enters a learning phase where it can infer the original pattern by generalization from a sea of corrupted examples that the teacher provides. For $\beta^{-1} < 1$ each example is highly informative and the learning performance is optimal ($m = 1$).

In Fig. 2.3 the phase diagram is shown. At low temperature we know from the previous section that learning is always possible, in particular a perfect retrieval of the teacher's pattern ($m = q = 1$) is achieved when $M \to \infty$. At high temperature, i.e. poorly informative dataset, a paramagnetic region where learning is not possible is separated from a ferromagnetic region where the signal inference is still possible leveraging on a sufficiently large dataset.

It is interesting to note that the critical line coincides with the paramagnetic to spin-glass transition line in the standard Hopfield model: the two systems becomes frozen in the same moment, when the signal from the patterns become macroscopic and prevails w.r.t. the temperature noise. In the standard Hopfield model different patterns led to different signals because they were independent and unbiased. For this reason the system got confused by increasing their number (the network load) and enters a spinglass regime. In our model each example carries a vanishing but non zero bias toward the signal, thus this bias becomes macroscopic by increasing extensively the size of the dataset and the system enters a ferromagnetic, ordered phase, where learning is possible.

2.4 Inference temperature vs dataset noise

The assumption that the student's machine is exactly equal to the teacher one is not realistic. The interesting research question is in fact related to the representation performance of a particular learning machine in relation to different possible data structures. To this aim, in this section we investigate one possible miss-matching between data (i.e. teacher machine) and student machine: the one related to the use of an inference temperature which differs from the real generating temperature of the data. We therefore assume that the training set is generated at an inverse temperature $\hat{\beta}$, i.e.

$$P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) = \prod_{\mu=1}^{M} z^{-1} \exp\left(\frac{\hat{\beta}}{N}\sum_{i<j}^{N} \hat{\xi}_i\hat{\xi}_j s_i^\mu s_j^\mu\right), \tag{2.46}$$

while the student patterns are still sampled at an inverse temperature β as in Eq. (2.1), which represents the posterior distribution of the learning problem with a miss-matched prior. It represents the more realistic situation in which the dataset noise $\hat{\beta}^{-1}$ is unknown and the machine is trained with a different weights regularization β.

When M is finite, it is possible to proof a result analogous of Theorem 3. At $\lambda = 0$, in the ferromagnetic gauge, the following holds

$$-\beta f = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\hat{\beta},\mathbf{0}^+} \log Z(\boldsymbol{\mathcal{S}}) = \sup_{\boldsymbol{p}\in\mathbb{R}^M} \hat{g}(\boldsymbol{p}). \tag{2.47}$$

The limiting free energy density trial function is given by

$$\hat{g}(\boldsymbol{p}) := \log 2 - \frac{\beta \boldsymbol{p}^2}{2} + \left\langle \log\cosh\left(\beta \boldsymbol{s}\cdot\boldsymbol{p}\right)\right\rangle_{\boldsymbol{s},\hat{\beta}}, \tag{2.48}$$

and the conditions necessary for applying a generalized saddle point approximation, can be derived directly from Lemmas 1-3, which are fully detailed in Appendix A. The difference in the derivation of (2.47) w.r.t. (2.16) is that, now, the random vector $\boldsymbol{s} \in \{-1,1\}^M$ has i.i.d. entries with mean $m_0(\hat{\beta})$, which depends on the generating temperature $\hat{\beta}$. By extremizing Eq. (2.48), $\boldsymbol{p}$ has to be a solution of

$$\boldsymbol{p} = \left\langle \boldsymbol{s}\tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p})\right\rangle_{\boldsymbol{s},\hat{\beta}} = \frac{\sum_{\boldsymbol{s}} \boldsymbol{s}\, e^{\hat{\beta} m_0 \mathbf{1}\cdot\boldsymbol{s}} \tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p})}{\left(2\cosh(\hat{\beta} m_0)\right)^M}, \tag{2.49}$$

from which, the learning performance can be derived as

$$\lim_{N\to\infty} \left\langle Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi})\right\rangle^{\hat{}} = m = \left\langle \tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p})\right\rangle_{\boldsymbol{s},\hat{\beta}} = \frac{\sum_{\boldsymbol{s}} e^{\hat{\beta} m_0 \mathbf{1}\cdot\boldsymbol{s}} \tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p})}{\left(2\cosh(\hat{\beta} m_0)\right)^M}. \tag{2.50}$$

Similarly to Eq. (2.26) it holds the following

Proposition 8. *As long as*

$$\beta^{-1} > 1 + (M-1)m_0^2(\hat{\beta}) \tag{2.51}$$

the only solution of Eqs. (2.49) is $\boldsymbol{p} = 0$*. As a consequence from Eq. (2.50)* $\boldsymbol{m} = 0$.

Proof. It is sufficient to see that

$$\begin{aligned} \boldsymbol{p}^2 &= \langle \boldsymbol{p} \cdot \boldsymbol{s} \tanh(\beta \boldsymbol{s} \cdot \boldsymbol{p}) \rangle_{\boldsymbol{s},\hat{\beta}} \leq \beta \left\langle (\boldsymbol{p} \cdot \boldsymbol{s})^2 \right\rangle_{\boldsymbol{s},\hat{\beta}} = \beta \sum_{\mu,\nu} p^\mu p^\nu \langle s_\mu s_\nu \rangle_{\boldsymbol{s},\hat{\beta}} \\ &= \beta \sum_{\mu,\nu} p^\mu p^\nu (\delta_{\mu\nu} + (1-\delta_{\mu\nu}) m_0^2(\hat{\beta})) = \beta(1 - m_0^2(\hat{\beta}))\boldsymbol{p}^2 + \beta m_0^2(\hat{\beta})(\sum_\mu p^\mu)^2 \\ &\leq \beta(1 + (M-1)m_0^2(\hat{\beta}))\boldsymbol{p}^2. \end{aligned} \tag{2.52}$$

□

As soon as the inference temperature drops below the threshold provided by Proposition 8 other solutions of Eqs. (2.49) appear according to the following

Proposition 9. *As long as* $\beta^{-1} < 1 + (M-1)m_0^2(\hat{\beta})$ *the global maximum of* $\hat{g}(\boldsymbol{p})$ *is attained far from* 0*. Moreover there exist solutions of Eqs. (2.49) of the form* $p^\mu = \pm\bar{p}$*,* $\forall \mu = 1, \dots, M$ *where* $\bar{p} > 0$ *is unique.*

Proof. It is sufficient to study the reduced equation

$$p = \left\langle s_1 \tanh\left(\beta p \sum_\mu s_\mu\right)\right\rangle_{\boldsymbol{s},\hat{\beta}} := f(p; \beta, \hat{\beta}). \tag{2.53}$$

The function f is odd and bounded in $(-1, 1)$. Moreover its derivative in $p = 0$ is

$$\frac{\partial f}{\partial p}|_{p=0} = \beta \left\langle s_1 \sum_\mu s_\mu \right\rangle_{\boldsymbol{s},\hat{\beta}} = \beta(1 + (M-1)m_0^2(\hat{\beta})). \tag{2.54}$$

As soon as this derivative becomes larger than 1 the function must intersect the bisector far from the origin in at least two symmetric points $\pm\bar{p}$, $\bar{p} > 0$. At the same time, the unique maximum of $\hat{g}(\boldsymbol{p})$ restricted to $p^\mu = p, \forall \mu = 1, \dots, M$, is attained in $\bar{p} > 0$, see the proof of Proposition 7. This proves the uniqueness of $\bar{p}$ and the instability of $\boldsymbol{p} = 0$. □

Note that Proposition 9 does not prove that the maximum of $\hat{g}(\boldsymbol{p})$ is in $\boldsymbol{p} = \pm\bar{p}\mathbf{1}$ because there could exist other solutions of Eqs. (2.49) that are not homogeneous. For example as long as $\hat{\beta} < 1$ (thus $m_0(\hat{\beta}) = 0$) and $\beta > 1$, there exist solutions in which the system is aligned with a single example (pure states), i.e. $p^\mu = \bar{p}_1 \neq 0$, $p^\nu = 0 \;\forall \nu \neq \mu$. In fact it is sufficient to fix $\bar{p}_1$ as the solution of

$$p = \langle s_1 \tanh(\beta p s_1) \rangle_{\boldsymbol{s},0}, \tag{2.55}$$

i.e. $\bar{p}_1 = \pm m_0(\beta)$. Analogously there could exist solutions in which the system is homogeneously aligned with a subset $E_k \subset \{1, \dots, M\}$, $|E_k| = k$, of the examples (mixed states), i.e. $p^\mu = \bar{p}_k \neq 0$ $\forall \mu \in E_k$, $p^\nu = 0 \;\forall \notin E_k$: in this case $\bar{p}_k$ has to be the solution of

$$p = \left\langle s_1 \tanh(\beta p \sum_{\mu=1}^k s_k) \right\rangle_{\boldsymbol{s},0}. \tag{2.56}$$

However, if the value of the inference temperature is not too low with respect to the dataset noise, then Proposition 6 can be generalized according to the following

Proposition 10. *As long as*

$$\beta^{-1} > 1 - m_0^2(\hat{\beta}) \tag{2.57}$$

the solutions of Eqs. (2.49) have equal components.

Proof. Following the proof of Proposition 6 it can be proved that for $\mu \neq \nu$ it holds

$$|p^\mu - p^\nu| \leq \beta(1 - m_0^2(\hat{\beta}))|p^\mu - p^\nu|. \tag{2.58}$$

Therefore as long as $\beta(1 - m_0^2(\hat{\beta})) < 1$ the only solution is homogeneous. □

In the region between the instability condition of Proposition 9 and the homogeneity condition of Proposition 10, i.e.

$$1 - m_0^2(\hat{\beta}) < \beta^{-1} < 1 + (M-1)m_0^2(\hat{\beta}), \tag{2.59}$$

which is non empty only if $\hat{\beta} > 1$, the global maximum of $\hat{g}(\boldsymbol{p})$ is attained in the two symmetric point $\boldsymbol{p} = \pm\bar{p} \neq 0$ and consequently the system is magnetized, i.e. it is aligned with the signal since

$$m = \pm \left\langle \tanh(\beta\bar{p}\sum_\mu s_\mu) \right\rangle_{\boldsymbol{s},\hat{\beta}} \neq 0. \tag{2.60}$$

Conversely if $\hat{\beta} < 1$ the value of the system's magnetization given by Eq. (2.50), i.e. the learning performance, is always zero because $m_0(\hat{\beta}) = 0$ independently from the value of $\boldsymbol{p}$. The phase diagram of the model, in terms of the value of m and $\boldsymbol{p}$ is shown in Fig. 2.4, where four different regions appear:

- Paramagnetic (P) region: $\boldsymbol{p} = \boldsymbol{0}$ and $m = 0$;
- Signal retrieval (sR) region: $\boldsymbol{p} = \bar{p}\boldsymbol{1}$, $m > 0$;
- Example retrieval (eR) region: $p \neq 0$, $m = 0$;
- Mixed retrieval (mR) region: $\boldsymbol{p} \neq \boldsymbol{0}$, $m > 0$.

In the paramagnetic region the inference temperature is too high and the machine neither stores the examples $\boldsymbol{p} = \boldsymbol{0}$ nor can learn the signal $m = 0$. In the eR region the inference temperature is low enough to allow the storage of the examples $\boldsymbol{p} \neq \boldsymbol{0}$ but they are not enough informative to allow signal learning. The stability of the different $\boldsymbol{p} \neq 0$ solutions can be investigated exactly as in the case of the standard Hopfield model (see [49]) but every one of them leads to a poor learning performance $m = 0$. Conversely in the sR region the dataset noise is low and the stored examples informative: in this region the machine can learn the signal by example memorization. Moreover, since $\boldsymbol{p} = \bar{p}\boldsymbol{1}$ in this region, it seems that the machine is working by uniformly using all the examples in a kind of early attempt of learning by generalization. Interestingly the machine can work efficiently even at temperatures β^{-1} higher then the dataset noise $\hat{\beta}^{-1}$. However the learning performance m increases monotonically by lowering the inference temperature. Finally in the mR region, different stable solutions for $\boldsymbol{p} \neq 0$ coexists, everyone leading to a positive global magnetization m, where typically the globally stable one is non-homogeneous, with $\max|p_i| > \bar{p}$. It seems to suggest that in this regime the machine prefers to learn the signal mainly leveraging

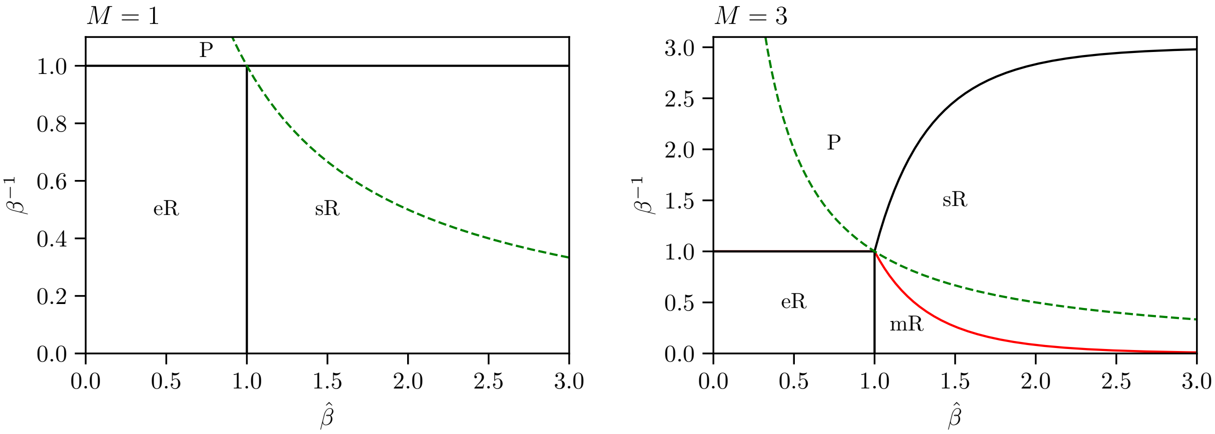


Figure 2.4: Phase diagram of the model in the case of *mismatched* setting and finite M ($M = 1$ on the left, $M = 3$ on the right) in terms of the dataset information $\hat{\beta}$ and the inference temperature β^{-1}. According to the values of m and $\boldsymbol{p}$ solutions of Eqs. (2.50,2.49) four different regimes appear: in the paramagnetic (P) regime $m = 0$, $\boldsymbol{p} = 0$; in the example retrieval (eR) regime $\boldsymbol{p} \neq \boldsymbol{0}$ but $m = 0$; in the signal retrieval (sR) regime $\boldsymbol{p} = \bar{p}\boldsymbol{1}$ is homogeneous and $m > 0$; in the mixed retrieval (mR) regime it is $\boldsymbol{p} \neq \boldsymbol{0}$ and $m > 0$. Only in the sR and mR regimes the machine can learn the original signal and the learning performance monotonically increases with $\hat{\beta}$. The Nishimori line $\hat{\beta} = \beta$ is shown in green.

on the high amount of information carried by a single (or few) examples, typical behaviour of a learning by memorization.

As in the previous section we expect that increasing the size of the dataset proportionally to the size of the system, i.e. $M = \alpha N$, it could be possible to retrieve the original pattern even when the examples are particularly noisy, i.e. their generating temperature is higher than 1. In this regime it is possible to generalize the Conjecture 1 and obtain the replica symmetric approximation of the limiting free energy in terms of the two temperatures β and $\hat{\beta}$ as

$$-\beta f^{RS} = \mathrm{Extr}_{m,\hat{m},p,\hat{p},q,\hat{q}}\, \hat{f}(p,q,m,\hat{p},\hat{q},\hat{m}), \tag{2.61}$$

where

$$\begin{aligned}\hat{f}(p,q,m,\hat{p},\hat{q},\hat{m}) = {}& \ln 2 - \frac{\alpha}{2}\ln\Big((1-\hat{\beta})(1-\beta+\beta q)\Big) + \frac{\alpha}{2}\frac{\beta(1-\hat{\beta})q+\hat{\beta}\beta m^2}{(1-\hat{\beta})(1-\beta+\beta q)} \\ & + \frac{\hat{q}q}{2} - \hat{m}m - \frac{\hat{q}}{2} - \hat{p}p + \frac{\beta}{2}p^2 + \left\langle \int \mathcal{D}z \ln\cosh\Big(\hat{p}s + z\sqrt{\hat{q}} + \hat{m}\Big)\right\rangle_s .\end{aligned} \tag{2.62}$$

The RS saddle point equations read as

$$m = \int D\mu(z)\left\langle \tanh(\beta p s + \hat{m} + z\sqrt{\hat{q}})\right\rangle_s \tag{2.63}$$

$$q = \int D\mu(z)\left\langle \tanh^2(\beta p s + \hat{m} + z\sqrt{\hat{q}})\right\rangle_s \tag{2.64}$$

$$p = \int D\mu(z)\left\langle s\tanh(\beta p s + \hat{m} + z\sqrt{\hat{q}})\right\rangle_s \tag{2.65}$$

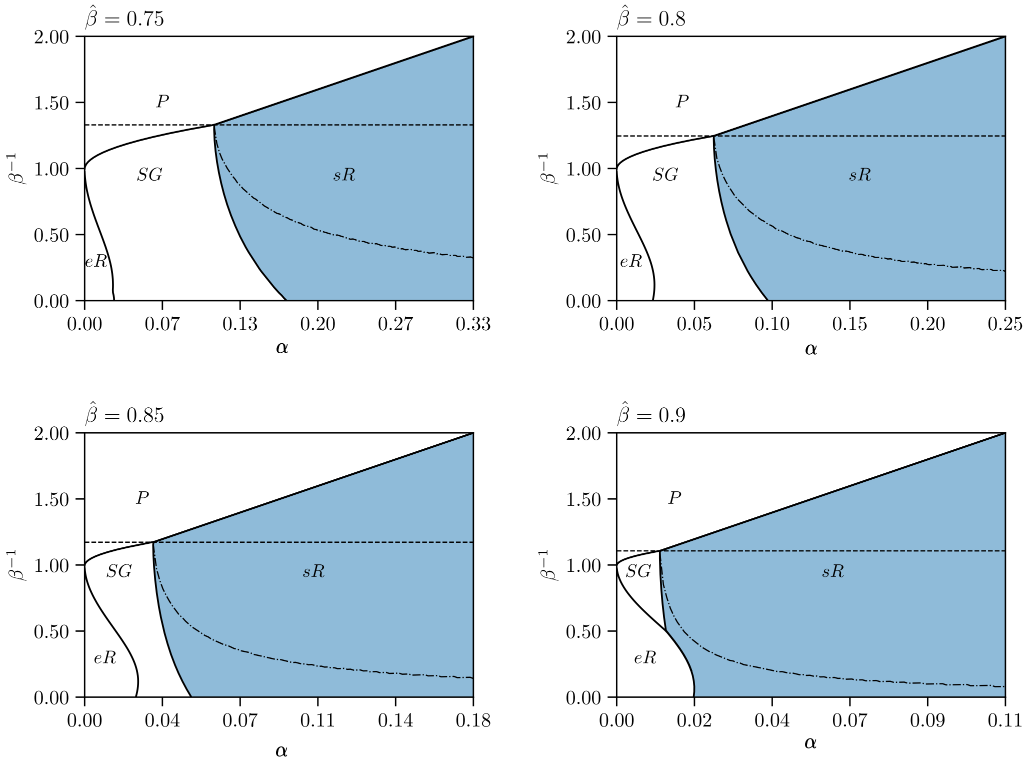


Figure 2.5: Phase diagram of the model in the mismatched setting, where $\hat{\beta} \neq \beta$, and extensive dataset $M = \alpha N$. In the paramagnetic (P) and spin glass (SG) regions learning is impossible. For higher values of the dataset size, the machine enters a signal retrieval region (sR) where it learns by generalizations. In this region the learning performance m has a maximum (dot-dash line) for a specific value of the inference temperature. In particular if β^{-1} gets too low the machine enters the example retrieval (eR) region where it is forced to work by memorization when this approach is inefficient for learning. The dotted line is the Nishimori condition $\beta = \hat{\beta}$.

where s is an auxiliary Rademacher random variable with symmetric distribution and

$$\begin{aligned}\hat{m} &= \frac{\alpha\hat{\beta}\beta m}{(1-\hat{\beta})(1-\beta+\beta q)} \\ \hat{q} &= \frac{\alpha\hat{\beta}\beta^2 m^2 + \alpha\beta^2 q(1-\hat{\beta})}{(1-\hat{\beta})(1-\beta+\beta q)^2}.\end{aligned} \tag{2.66}$$

The order parameter p has to be interpreted as the overlap between the student pattern and the examples, i.e.

$$p = \lim_{N\to\infty} \langle Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})\rangle = \lim_{N\to\infty} \widehat{\langle Q(\boldsymbol{s}^\mu, \boldsymbol{\xi})\rangle}\,. \tag{2.67}$$

Equations (2.63,2.64,2.65) reduce to those of Conjecture 1 when $\hat{\beta} = \beta$ and $p = 0$. In fact in that case, since $\hat{\beta} = \beta < 1$, the system is never aligned with any example. Conversely if $\beta \neq \hat{\beta}$, even if $\hat{\beta} < 1$, the inference temperature β^{-1} could be in principle low enough to allow example retrieval. It is important to stress however that, since the examples are only weakly correlated with the signal, this situation would prevent the system to be aligned with the original pattern. In Figure 2.5 the phase diagram of the model is shown for different values of the generating temperature $\hat{\beta}$. Four different regions appear, depending on the properties of the globally stable solution of Eqs. (2.63,2.64,2.65):

- Paramagnetic (P) region: $m = q = p = 0$;

- Signal retrieval (sR) region: $m \neq 0$, $q > 0$, $p = 0$;
- Example retrieval (eR) region: $p \neq 0$, $q > 0$, $m = 0$;
- Spin Glass (SG) region: $m = p = 0$, $q > 0$.

Only in the sR phase student and teacher patterns are correlated and the learning performance is positive. In all other phases the student patterns is uncorrelated with the signal, being a random guess (P region), aligned with a noisy example (eR) or aligned with a spurious low energy state (SG region). From the high temperature (P) phase to the low temperature (SG or sR) phases a second order phase transition occurs when the system's overlap q detaches from zero. The transition line can be obtained by expanding eq. (2.63) or eq. (2.64) depending on the magnetization m behavior. For small values of both m and q (P to sR), from eq. (2.63) we get

$$m = \frac{\alpha\hat{\beta}\beta}{(1-\beta)(1-\hat{\beta})} m + O(mq, m^2) \tag{2.68}$$

that gives the instability condition

$$\frac{\alpha\hat{\beta}\beta}{(1-\beta)(1-\hat{\beta})} = 1 \quad \Longrightarrow \quad \beta^{-1} = 1 + \alpha\frac{\hat{\beta}}{1-\hat{\beta}}. \tag{2.69}$$

On the other hand if $m = 0$ across the transition (P to SG), by expanding Eq. (2.64) for small values of q we get

$$q = \frac{\alpha\beta^2}{(1-\beta)^2} q + O(q^2) \tag{2.70}$$

that gives the usual instability condition

$$\frac{\alpha\beta^2}{(1-\beta)^2} = 1 \quad \Longrightarrow \quad \beta^{-1} = 1 + \sqrt{\alpha} \tag{2.71}$$

Therefore, starting from the paramagnetic region and decreasing the inference temperature a phase transition occurs as soon as one of the two instability condition is satisfied, i.e. at

$$\beta^{-1}(\alpha; \hat{\beta}) := \max\left\{1 + \sqrt{\alpha}; 1 + \alpha\frac{\hat{\beta}}{1-\hat{\beta}}\right\}. \tag{2.72}$$

Interestingly the two lines cross exactly at the point $(\beta, \alpha) = (\hat{\beta}, (1-\hat{\beta})^2/\hat{\beta}^2))$, in agreement with the usual property of the Nishimori line of crossing a triple critical point. For smaller values of α the transition is towards a SG regime, while for higher values of α the transition is towards a sR region, where learning is easy. The other two transitions, from sR to SG (second order) and from SG to eR (first order) can be found numerically and shown in Figure 2.5. The phase diagram shows a non monotone behavior of the learning performance in terms of the inference temperature: if β^{-1} is too high the learning performance is that of a random guess (P phase), if β^{-1} is too low the learning performance can deteriorate because of the emergence of low energy configurations that are uncorrelated with the signal (they can be either correlated with the examples, eR region, or completely uncorrelated with both signal and examples, SG region). In particular the eR phase identifies a regime in which the machine is forced (through β) to work by memorization ($p > 0$) in a situation where this approach is highly inefficient for learning. By increasing α the memory storage limit of the machine becomes beneficial for the occurrence of a region where learning is possible by generalization. Interestingly the phase diagram of Fig. 2.5 seems qualitatively similar to that of Fig. 2.4 where in both cases the x-axis measures the amount of information contained in the dataset.

Finally note that Eqs. (2.63)-(2.65) becomes exactly those of the classicl Hopfield model if we force $m = 0$. This means that a purely SG solution always exists below the inference temperature $1+\sqrt{\alpha}$, which is only locally stable inside the sR region. In this case a Monte-Carlo simulation, performed at a low temperature, can remain trapped in the locally stable spin-glass state. Fortunately, this occurrence can be avoided by using Simulated Annealing and lowering the temperature very slowly; this is possible because the critical temperature for signal retrieval is higher than $1 + \sqrt{\alpha}$. A similar strategy that leverages on the hierarchy between temperatures of ergodicity breaking is investigated in [91].

Appendix A

Proofs

In this section the proofs of the main results are provided, together with some technical results, in the form of lemmas and propositions, needed for the proofs. Since it will be used many times in the rest of the section we recall a standard result about the Fourier decomposition of a function of boolean (± 1) variables. Given $\Lambda = \{1, \dots, M\}$, any function $f : \{-1, 1\}^M \to \mathbb{R}$ can be decomposed as

$$f(\boldsymbol{s}) = \sum_{X \subset \mathcal{P}(\Lambda)} \langle f, s_X \rangle \, s_X, \tag{A.1}$$

where $s_X = \prod_{\mu \in X} s_\mu$ and $\langle f, g \rangle = 2^{-M} \sum_{\boldsymbol{s}} f(\boldsymbol{s}) g(\boldsymbol{s}) = \langle fg \rangle_{\boldsymbol{s},0}$.

Proof. of Theorem 3

By using gaussian linearization we can write the partition function as

$$\begin{aligned} Z(\boldsymbol{\mathcal{S}}) &:= \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta}{2N} \sum_{\mu=1}^{M} \sum_{i,j=1}^{N} s_i^\mu s_j^\mu \xi_i \xi_j + \lambda \sum_{i=1}^{N} \xi_i\right) \\ &= \sum_{\boldsymbol{\xi}} \int D\mu(\boldsymbol{z}) \exp\left(\sqrt{\frac{\beta}{N}} \sum_{\mu=1}^{M} \sum_{i=1}^{N} s_i^\mu \xi_i z^\mu + \lambda \sum_{i=1}^{N} \xi_i\right). \end{aligned} \tag{A.2}$$

where $\mathbb{R}^M \ni \boldsymbol{z} \sim \mathcal{N}(0, \mathbb{I})$. By making a change of variables $p^\mu = z^\mu / \sqrt{N\beta}$ we have

$$\begin{aligned} Z(\boldsymbol{\mathcal{S}}) &\propto \sum_{\boldsymbol{\xi}} \int d\boldsymbol{p} \exp\left(-\beta N \boldsymbol{p}^2/2 + \sum_{i=1}^{N} (\beta \boldsymbol{s_i} \cdot \boldsymbol{p} + \lambda) \xi_i\right) \\ &= \int d\boldsymbol{p} \; \exp\left(N \left(\log 2 - \frac{\beta}{2} \boldsymbol{p}^2 + \frac{1}{N} \sum_{i=1}^{N} \log\cosh(\beta \boldsymbol{s}_i \cdot \boldsymbol{p} + \lambda)\right)\right) \\ &= \int d\boldsymbol{p} \; e^{N g_N(\boldsymbol{p}, \boldsymbol{\mathcal{S}})} \end{aligned} \tag{A.3}$$

We recall that the random vector of the examples $\boldsymbol{\mathcal{S}}$ is drawn from Eq. (2.12): using Lemma 1 on the function $F : \{-1, 1\}^M \to \mathbb{R}$, $F(\boldsymbol{s}) = \log\cosh(\beta \boldsymbol{s} \cdot \boldsymbol{p} + \lambda)$, it holds

$$\lim_{N \to \infty} \mathbb{E}_{\beta, h\boldsymbol{\epsilon}} \, g_N(\boldsymbol{p}, \boldsymbol{\mathcal{S}}) = g(\boldsymbol{p}) := \log 2 - \frac{\beta \boldsymbol{p}^2}{2} + \langle \log\cosh(\beta \boldsymbol{s} \cdot \boldsymbol{p} + \lambda) \rangle_{\boldsymbol{s}, \boldsymbol{\epsilon}}, \tag{A.4}$$

where $\langle . \rangle_{\boldsymbol{s}, \boldsymbol{\epsilon}}$ denotes the mean-field expectation w.r.t. the random vector $\boldsymbol{s} \in \{-1, 1\}^M$, whose entries are independent with mean $\langle s^\mu \rangle_{\boldsymbol{s}, \boldsymbol{\epsilon}} = m_0(\beta, \epsilon^\mu h)$ and $m_0(\beta, h)$ is the unique solution of

$m_0 = \tanh(\beta m_0 + h)$. For any compact set $K \subset \mathbb{R}$, using Lemma 2 on $F : K \times \{-1,1\}^M \to \mathbb{R}$, $F(\boldsymbol{p}, \boldsymbol{s}) = \log\cosh(\beta \boldsymbol{s} \cdot \boldsymbol{p} + \lambda)$, it holds that

$$g_N(\boldsymbol{p}, \boldsymbol{\mathcal{S}}) \xrightarrow{p} g(\boldsymbol{p}), \tag{A.5}$$

uniformly in K, meaning that $g_N(\boldsymbol{p}, \boldsymbol{\mathcal{S}})$ is self-averaging uniformly in any compact. Thanks to Lemma 3, g_N and g satisfy the conditions necessary for a generalized saddle point approximation, i.e. Proposition 11 and Proposition 12, so that

$$-\beta f = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}_{\beta,h\boldsymbol{\epsilon}} \int_{\mathbb{R}^M} d\boldsymbol{p}\, e^{N g_N(\boldsymbol{p};\boldsymbol{\mathcal{S}})} = \sup_{\mathbb{R}^M} g(\boldsymbol{p}). \tag{A.6}$$

It is easy to show that $-\beta f$ does not depend on ϵ by using the transformations $s^\mu \to \epsilon^\mu s^\mu$ and $p^\mu \to \epsilon^\mu p^\mu$. This concludes the proof if one chooses $\epsilon = \mathbf{1}$.

□

Lemma 1. *Let $\boldsymbol{s}_i \in \{-1,1\}^M$ the i-th marginal of $\boldsymbol{\mathcal{S}} = \{\boldsymbol{s}_i\}_{i=1}^N \in \{-1,1\}^{NM}$, distributed according to (2.12). For any function $F : \{-1,1\}^M \to \mathbb{R}$ it holds*

$$\lim_{N\to\infty} \mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\, F(\boldsymbol{s}_i) = \langle F(\boldsymbol{s})\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\,, \quad \forall i = 1,\dots,N, \tag{A.7}$$

where $\langle . \rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}$ denotes the expectation w.r.t. the random vector $\boldsymbol{s} \in \{-1,1\}^M$ whose entries are independent with mean $\langle s^\mu \rangle_{\boldsymbol{s},\boldsymbol{\epsilon}} = m_0(\beta, \epsilon^\mu h)$ and $m_0(\beta,h)$ is the unique solution of $m_0 = \tanh(\beta m_0 + h)$.

Proof. By Fourier decomposition F can be written as

$$F(\boldsymbol{s}) = \sum_{X\subset\mathcal{P}(\Lambda)} c_X \prod_{\mu\in X} s^\mu. \tag{A.8}$$

As a consequence, for any factorized probability $\pi(\boldsymbol{s}) = \prod_{\mu=1}^M \pi_\mu(s^\mu)$ it holds

$$\langle F(\boldsymbol{s})\rangle_\pi = \sum_{X\subset\mathcal{P}(\Lambda)} c_X \left\langle \prod_{\mu\in X} s^\mu \right\rangle_\pi = \sum_{X\subset\mathcal{P}(\Lambda)} c_X \prod_{\mu\in X} \langle s^\mu\rangle_{\pi_\mu}. \tag{A.9}$$

Therefore, since the Fourier decomposition has a finite number of terms, it is

$$\begin{aligned} \lim_{N\to\infty} \mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\, F(\boldsymbol{s}_i) &= \sum_{X\subset\mathcal{P}(\Lambda)} c_X \prod_{\mu\in X} \lim_{N\to\infty} \mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\, s_i^\mu \\ &= \sum_{X\subset\mathcal{P}(\Lambda)} c_X \prod_{\mu\in X} \langle s^\mu\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}} = \langle F(\boldsymbol{s})\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\,, \end{aligned} \tag{A.10}$$

where in the second line we have used that in the Curie-Weiss model at $\beta > 0$ and $h > 0$

$$\lim_{N\to\infty} \mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\, s_i^\mu = m_0(\beta, \epsilon^\mu h) = \langle s^\mu\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}} \quad \forall i = 1,\dots,N. \tag{A.11}$$

□

Lemma 2. *Let $\boldsymbol{s}_i \in \{-1,1\}^M$ the i-th marginal of $\boldsymbol{\mathcal{S}} = \{\boldsymbol{s}_i\}_{i=1}^N \in \{-1,1\}^{NM}$, distributed according to (2.12). Given a compact set $K \subset \mathbb{R}^M$ and a bounded function $F : K \times \{-1,1\}^M \to \mathbb{R}$, it holds, uniformly in K, that*

$$\frac{1}{N}\sum_{i=1}^N F(\boldsymbol{p}, \boldsymbol{s}_i) \xrightarrow{p} \langle F(\boldsymbol{p}, \boldsymbol{s})\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\,.$$

Proof. Given the set

$$A_{N,\epsilon}=\left\{\boldsymbol{\mathcal{S}}:\sup_{\boldsymbol{p}\in K}\left|\frac{1}{N}\sum_{i=1}^{N}F(\boldsymbol{p},\boldsymbol{s}_i)-\langle F(\boldsymbol{p},\boldsymbol{s})\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\right|>\epsilon\right\},\tag{A.12}$$

we need to show that, $\forall\epsilon>0$, $\lim_{N\to\infty}\mathbb{P}(A_{N,\epsilon})=0$. To this aim let's consider the Fourier decomposition of $F(\boldsymbol{p},\boldsymbol{s})$ as

$$F(\boldsymbol{p},\boldsymbol{s})=\sum_{X\in\mathcal{P}(\Lambda)}c_X(\boldsymbol{p})\prod_{\mu\in X}s^{\mu}\tag{A.13}$$

and define the set

$$B_{N,\epsilon}=\left\{\boldsymbol{\mathcal{S}}:\sup_{X\in\mathcal{P}(\Lambda)}\left|\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}-\left\langle\prod_{\mu\in X}s^{\mu}\right\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\right|>\frac{\epsilon}{L_F}\right\},\tag{A.14}$$

where $L_F=\sum_{X\in\mathcal{P}(\Lambda)}\sup_{\boldsymbol{p}\in K}|c_X(\boldsymbol{p})|<\infty$. According to this definition note that within the set $B^c_{N,\epsilon}$ it holds

$$\sup_{p\in K}\left|\frac{1}{N}\sum_{i=1}^{N}F(\boldsymbol{p},\boldsymbol{s}_i)-\langle F(\boldsymbol{p},\boldsymbol{s})\rangle_{\boldsymbol{s},\epsilon}\right|\;\leq\;\sup_{p\in K}\sum_{X\in\mathcal{P}(\Lambda)}|c_X(\boldsymbol{p})|\left|\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}-\left\langle\prod_{\mu\in X}s^{\mu}\right\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\right|\leq\epsilon.$$

Therefore $B^c_{N\epsilon}\subseteq A^c_{N\epsilon}$ and thus $A_{N,\epsilon}\subseteq B_{N,\epsilon}$. As a consequence

$$\begin{aligned}\mathbb{P}(A_{N,\epsilon})\;&\leq\;\mathbb{P}(B_{N,\epsilon})=\mathbb{P}\left(\bigcup_{X\in\mathcal{P}(\Lambda)}\left\{\boldsymbol{\mathcal{S}}:\left|\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}-\left\langle\prod_{\mu\in X}s^{\mu}\right\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\right|>\frac{\epsilon}{L_F}\right\}\right)\\&\leq\;\sum_{X\in\mathcal{P}(\Lambda)}\mathbb{P}\left(\left\{\boldsymbol{\mathcal{S}}:\left|\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}-\left\langle\prod_{\mu\in X}s^{\mu}\right\rangle_{\boldsymbol{s},\epsilon}\right|>\frac{\epsilon}{L_F}\right\}\right)\\&\leq\;\sum_{X\in\mathcal{P}(\Lambda)}(L_F/\epsilon)^2\,\mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}-\left\langle\prod_{\mu\in X}s^{\mu}\right\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}\right)^2\end{aligned}\tag{A.15}$$

that goes to zero in the thermodynamic limit since

$$\mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{\mu\in X}s_i^{\mu}\right)^2=\frac{1}{N^2}\sum_{i\neq j}\prod_{\mu\in X}\mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\left(s_i^{\mu}s_j^{\mu}\right)+\frac{1}{N}\tag{A.16}$$

and using the factorization property of the Curie-Weiss model that

$$\lim_{N\to\infty}\mathbb{E}_{\beta,h\boldsymbol{\epsilon}}\left(s_i^{\mu}s_j^{\mu}\right)=m_0^2(\beta,\epsilon^{\mu}h)=\langle s^{\mu}\rangle^2_{\boldsymbol{s},\boldsymbol{\epsilon}}\quad\forall i,j=1,\dots,N,\;i\neq j.\tag{A.17}$$

□

Lemma 3. *The functions $g_N:\mathbb{R}^M\times\{-1,1\}^{NM}\to\mathbb{R}$ and $g:\mathbb{R}^M\to\mathbb{R}$ defined in eqs. (A.3) and (A.4) satisfy*

$$|g_N(\boldsymbol{p},\boldsymbol{\mathcal{S}})-g(\boldsymbol{p})|\leq 2\beta\sum_{\mu=1}^{M}|p_{\mu}|\,,\qquad\forall\boldsymbol{p}\in\mathbb{R}^M,\boldsymbol{\mathcal{S}}\in\{-1,+1\}^{NM}.\tag{A.18}$$

As a consequence it holds

- $\sup_N \sup_{K\times\{-1,1\}^{NM}} |g_N| < \infty$ *for any compact* $K \subset \mathbb{R}^M$*;*
- $\exists\, C_1 < \infty$*:* $\sup_{\mathbb{R}^M\times\Sigma_N} g_N < C_1$ *;*
- $\exists\, C_2 < \infty$*:* $\int_{\mathbb{R}^M} e^{g_N(\boldsymbol{p},s)} d\boldsymbol{p} < C_2$*;*
- $\exists\, K \subset \mathbb{R}^M, \delta > 0$*:* $g_N(\boldsymbol{p}, s) - \sup_K g < -\delta, \quad \forall(\boldsymbol{p}, s) \in K^c \times \Sigma_N$.

Proof. It holds,

$$\begin{aligned}
|g_N(\boldsymbol{p},\boldsymbol{\mathcal{S}}) - g(\boldsymbol{p})| &= \left|\frac{1}{N}\sum_{i=1}^N \log\cosh\left(\beta\boldsymbol{s}_i\cdot\boldsymbol{p}+\lambda\right) - \langle\log\cosh\left(\beta\boldsymbol{s}\cdot\boldsymbol{p}+\lambda\right)\rangle_{\boldsymbol{s},\epsilon}\right| \\
&= \left|\frac{1}{N}\sum_{i=1}^N\left\langle\log\frac{\cosh\left(\beta\boldsymbol{s}_i\cdot\boldsymbol{p}+\lambda\right)}{\cosh\left(\beta\boldsymbol{\tau}\cdot\boldsymbol{p}+\lambda\right)}\right\rangle_{\boldsymbol{\tau},\epsilon}\right| = \left|\frac{1}{N}\sum_{i=1}^N\left\langle\int_{\beta\boldsymbol{\tau}\cdot\boldsymbol{p}+\lambda}^{\beta\boldsymbol{s}_i\cdot\boldsymbol{p}+\lambda}\tanh(x)dx\right\rangle_{\boldsymbol{\tau},\epsilon}\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N\left|\langle\beta\boldsymbol{s}_i\cdot\boldsymbol{p}-\beta\boldsymbol{\tau}\cdot\boldsymbol{p}\rangle_{\boldsymbol{\tau},\boldsymbol{\epsilon}}\right| \leq 2\beta\sum_{\mu=1}^M|p_\mu|.
\end{aligned}$$

As a consequence of the previous relation g_N is bounded uniformly by

$$g_N(\boldsymbol{p},\boldsymbol{\mathcal{S}}) \leq g(\boldsymbol{p})+|g_N(\boldsymbol{p},\boldsymbol{\mathcal{S}}) - g(\boldsymbol{p})| \leq \log 2-\frac{\beta\boldsymbol{p}^2}{2}+\langle\log\cosh(\beta\boldsymbol{s}\cdot\boldsymbol{p}+\lambda)\rangle_{\boldsymbol{s},\boldsymbol{\epsilon}}+2\beta\sum_{\mu=1}^M|p_\mu| := \hat{g}(\boldsymbol{p}). \tag{A.19}$$

Since $\hat{g}$ is continuous, it is bounded in any compact set $K \subset \mathbb{R}^M$. Moreover it is easy to see that $\sup_{\mathbb{R}^M}\hat{g} < \infty$ and that $\int_{\mathbb{R}^M} e^{g(\boldsymbol{p})}d\boldsymbol{p} < \infty$ since it has gaussian tails. Moreover since $\hat{g}$ goes to $-\infty$ at infinity it always exists a sufficiently large ball B_{r_δ} of radius r_δ such that $\hat{g}(\boldsymbol{p}) - \sup\hat{g} < -\delta$ for $\boldsymbol{p} \in B^c_{r_\delta}$. Therefore the properties for g_N are proved uniformly in N. □

Proposition 11. *Let* $K \subset \mathbb{R}^M$ *a compact set,* μ_N *a probability distribution over a finite set* Σ_N*,* $F_N : K \times \Sigma_N \to \mathbb{R}$ *a sequence of bounded functions such that* $\sup_N \sup_{K\times\Sigma_N}|F_N| < \infty$ *and* $F : K \to \mathbb{R}$ *bounded such that*

- $F_N \xrightarrow{p} F$ *uniformly in* K*.*
- $\lim_{N\to\infty}\frac{1}{N}\log\int_K e^{NF} = \sup_K F$*;*

Then it holds

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)} = \sup_K F. \tag{A.20}$$

Proof. Let's define $C = \sup_N\sup_{K\times\Sigma_N}|F_N| < \infty$ and the set

$$A_{N,\epsilon} = \left\{s \in \Sigma_N : \sup_{p\in K}|F_N(\boldsymbol{p},s) - F(\boldsymbol{p}| > \epsilon\right\}. \tag{A.21}$$

By assumptions it holds, $\forall\epsilon > 0$, that $\mu_N(A_{N,\epsilon}) \to 0$. Therefore it holds, $\forall\epsilon > 0$, that

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)} = 0,$$

where $\mathbb{I}_{\mathbb{A}}$ is the indicator function of the set A, since, calling $|K|$ the Lebesgue measure of K,

$$\left|\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}\right| \leq \mu_N(A_{N,\epsilon})\left|C+\log(|K|)/N\right|. \tag{A.22}$$

Thanks to these results we just need to evaluate the expected value over $A^c_{N,\epsilon}$ that has full probability in the limit. Inside $A^c_{N,\epsilon}$ we can substitute F_N with F at the exponent with an error ϵ i.e.

$$\left|\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}-\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF(\boldsymbol{p})}\right|<\epsilon. \tag{A.23}$$

Since $\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}}=\mu_N(A^c_{N,\epsilon})\to 1$ and using the standard Laplace approximation, it holds

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF(\boldsymbol{p})}=\sup_K F, \tag{A.24}$$

therefore, $\forall\epsilon>0$,

$$\left|\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}-\sup_K F\right|=\left|\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}-\sup_K F\right|<\epsilon. \tag{A.25}$$

□

Proposition 12. *Let $K\subset\mathbb{R}^M$ a compact set, μ_N a probability distribution over a finite set Σ_N, $F_N:\mathbb{R}^M\times\Sigma_N\to\mathbb{R}$ and $F:\mathbb{R}^M\to\mathbb{R}$ satisfying the conditions of Proposition 11 when restricted on the set K. Assuming moreover that*

- $\exists\, C_1<\infty$: $\sup_{\mathbb{R}^M\times\Sigma_N}F_N<C_1$;
- $\exists\, C_2<\infty$: $\int_{\mathbb{R}^M}e^{F_N(\boldsymbol{p},s)}d\boldsymbol{p}<C_2$;
- $\exists\,\delta>0$: $F_N(\boldsymbol{p},s)-\sup_K F<-\delta,\quad\forall(\boldsymbol{p},s)\in K^c\times\Sigma_N$;

then it holds

$$\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\mu_N}\log\int_{\mathbb{R}^M}d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}=\sup_K F. \tag{A.26}$$

Proof. Since

$$\frac{1}{N}\mathbb{E}_{\mu_N}\log\int_{\mathbb{R}^M}d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}=\frac{1}{N}\mathbb{E}_{\mu_N}\log\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}+\frac{1}{N}\mathbb{E}_{\mu_N}\log\left(1+\frac{\int_{K^c}d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}}{\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}}\right) \tag{A.27}$$

it is sufficient to apply Proposition 11 for the first term and showing that the second term is vanishing in the limit. Defining the set $A_{N,\epsilon}$ as in the proof of Proposition 11, with $\epsilon<\delta$ and denoting $F^*=\sup_K F$ and $C_0=\sup_N\sup_{K\times\Sigma_N}|F_N|$, it holds

$$\frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A_{N,\epsilon}}\log\left(1+\frac{\int_{K^c}d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}}{\int_K d\boldsymbol{p}e^{NF_N(\boldsymbol{p},s)}}\right)\leq\mu_N(A_{N,\epsilon})\frac{1}{N}\log\left(1+\frac{e^{(N-1)C_1}C_2}{|K|e^{-NC_0}}\right)\overset{N\to\infty}{\longrightarrow}0, \tag{A.28}$$

since $\mu_N(A_{N,\epsilon}) \to 0$. Moreover, choosing $\epsilon_2 < \delta - \epsilon$, it holds $\forall N > N_{\epsilon_2}$

$$\begin{aligned} & \frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}} \log\left(1+\frac{\int_{K^c} d\boldsymbol{p} e^{NF_N(\boldsymbol{p},s)}}{\int_K d\boldsymbol{p} e^{NF_N(\boldsymbol{p},s)}}\right) \leq \frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}} \frac{\int_{K^c} d\boldsymbol{p} e^{NF_N(\boldsymbol{p},s)}}{\int_K d\boldsymbol{p} e^{NF_N(\boldsymbol{p},s)}} \\ \leq \; & \frac{1}{N}\mathbb{E}_{\mu_N}\mathbb{I}_{A^c_{N,\epsilon}} \frac{e^{(N-1)F^*}\int_{K^c} d\boldsymbol{p} e^{(N-1)(F_N(\boldsymbol{p},s)-F^*)+F_N(\boldsymbol{p},s)}}{\int_K d\boldsymbol{p} e^{-N|F_N(\boldsymbol{p},s)-F(\boldsymbol{p})|+NF(\boldsymbol{p})}} \leq \frac{1}{N}\frac{e^{(N-1)F^*-\delta(N-1)}C_2}{e^{-\epsilon N}\int_K d\boldsymbol{p} e^{NF(\boldsymbol{p})}} \\ \leq \; & \frac{1}{N}\frac{e^{(N-1)F^*-\delta(N-1)}C_2}{e^{-\epsilon N}e^{NF^*-N\epsilon_2}} = Ce^{-N(\delta-\epsilon-\epsilon_1)} \overset{N\to\infty}{\longrightarrow} 0, \end{aligned}$$

where we have used that

$$\left|\frac{1}{N}\log\int_K d\boldsymbol{p} e^{NF(\boldsymbol{p})} - F^*\right| < \epsilon_2. \tag{A.29}$$

□

Proof. of Proposition 4
To find the value of the magnetization (third point) we use that, $\forall N \in \mathbb{N}$,

$$\left\langle \frac{\boldsymbol{1}\cdot\boldsymbol{\xi}}{N} \right\rangle = -\beta\partial_\lambda\, f_N. \tag{A.30}$$

Thanks to the concavity of f_N in λ, we can exchange the thermodynamic limit with the derivative obtaining

$$m = \partial_\lambda(-\beta f) = \langle \tanh(\beta \boldsymbol{s}\cdot\boldsymbol{p}+\lambda)\rangle_{\boldsymbol{s}}\,.$$

Moreover

$$\mathrm{Var}(\frac{\boldsymbol{1}\cdot\boldsymbol{\xi}}{N}) = -\beta N^{-1}\,\partial^2_\lambda f_N \tag{A.31}$$

and therefore the magnetization has to be self-averaging in the thermodynamic limit. Analogous arguments, based on the response of the free energy to linear external perturbations [47] can be used for the first two points that are just generalizations of classical results about the Curie Weiss model [97].

□

Proof. of Proposition 5
Applying the Fourier decomposition to the function $f(\boldsymbol{s};\boldsymbol{p}) = \boldsymbol{s}\tanh(\beta\boldsymbol{s}\cdot\boldsymbol{p})$ it holds

$$p^\mu \;=\; \langle f^\mu(\boldsymbol{s};\boldsymbol{p})\rangle_{\boldsymbol{s}} = A^\mu(\boldsymbol{p}) + \sum_{\nu\neq\mu} A^\nu(\boldsymbol{p})\,\langle s_\mu s_\nu\rangle_{\boldsymbol{s}} = A^\mu(\boldsymbol{p}) + m_0^2(\beta)\sum_{\nu\neq\mu} A^\nu(\boldsymbol{p}) \tag{A.32}$$

where

$$A^\mu(\boldsymbol{p}) = \langle \tanh(\beta p^\mu + \beta\sum_{\nu\neq\mu} p^\nu s_\nu)\rangle_{\boldsymbol{s},0} \tag{A.33}$$

In fact it is easy to check by symmetry that

$$\begin{aligned} \langle s_\mu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), 1\rangle \;&=\; A^\mu(\boldsymbol{p}) \\ \langle s_\mu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), s_\nu\rangle \;&=\; \left\langle \tanh(\beta(p^\nu s_\mu + p^\mu s_\nu + \sum_{k\neq\mu,\nu} p^k s_k s_\mu s_\nu)) \right\rangle_{\boldsymbol{s},0} = 0 \quad \forall\nu\in\Lambda, \nu\neq\mu \\ \langle s_\mu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), s_\mu s_\nu\rangle \;&=\; \langle s_\nu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), 1\rangle = A^\nu(\boldsymbol{p}) \;\; \forall\nu\in\Lambda, \nu\neq\mu \\ \langle s_\mu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), s_k s_\nu\rangle \;&=\; \left\langle \tanh(\beta(p^\mu s_k s_\nu + \sum_{l\neq\mu} p^l s_l s_\mu s_k s_\nu)) \right\rangle_{\boldsymbol{s},0} = 0 \quad \forall k,\nu\in\Lambda, k,\nu\neq\mu \\ \langle s_\mu \tanh(\beta(\boldsymbol{s}\cdot\boldsymbol{p})), s_X\rangle \;&=\; 0 \;\; \forall X\subset\mathcal{P}(\Lambda), |X|>2. \end{aligned}$$

Using eq. (A.32) it holds $\forall \mu, \nu \in \Lambda, \mu \neq \nu$, that

$$p^\mu - p^\nu = (1 - m_0^2(\beta))\,(A^\mu(\boldsymbol{p}) - A^\nu(\boldsymbol{p}))\,. \tag{A.34}$$

and by direct computation we have that

$$\begin{aligned}
A^\mu(\boldsymbol{p}) - A^\nu(\boldsymbol{p}) &= \left\langle \tanh(\beta p^\mu + \beta \sum_{k\neq\mu} p^k s_k) \right\rangle_{\boldsymbol{s},0} - \left\langle \tanh(\beta p^\nu + \beta \sum_{l\neq\nu} p^l s_l) \right\rangle_{\boldsymbol{s},0} \\
&= \frac{1}{2}\left(\left\langle \tanh(\beta p^\mu + \beta p^\nu + \beta \sum_{k\neq\mu,\nu} p^k s_k) \right\rangle_{\boldsymbol{s},0} + \left\langle \tanh(\beta p^\mu - \beta p^\nu + \beta \sum_{k\neq\mu,\nu} p^k s_k) \right\rangle_{\boldsymbol{s},0} \right) \\
&- \frac{1}{2}\left(\left\langle \tanh(\beta p^\nu + \beta p^\mu + \beta \sum_{k\neq\mu,\nu} p^k s_k) \right\rangle_{\boldsymbol{s},0} + \left\langle \tanh(\beta p^\nu - \beta p^\mu + \beta \sum_{k\neq\mu,\nu} p^k s_k) \right\rangle_{\boldsymbol{s},0} \right) \\
&= \left\langle \tanh(\beta(p^\mu - p^\nu) + \beta \sum_{k\neq\mu,\nu} p^k s_k) \right\rangle_{\boldsymbol{s},0} .
\end{aligned}$$

Thus $p^\mu - p^\nu$ satisfies an equation of the form

$$p^\mu - p^\nu = (1 - m_0^2)\,\langle \tanh(\beta(p^\mu - p^\nu) + \beta Z^{\boldsymbol{p}}) \rangle_{Z^{\boldsymbol{p}}} \tag{A.35}$$

with $Z^{\boldsymbol{p}} = \sum_{k\neq\mu,\nu} p^k s_k$ a random noise. The function $A\,\langle \tanh(Bx + Z^{\boldsymbol{p}}) \rangle_{Z^{\boldsymbol{p}}}$ is odd and for $x \geq 0$ it is always under the line ABx. In fact, for any vector $\boldsymbol{p}$ and $\lambda \in (0,1)$,

$$\begin{aligned}
\frac{d}{d\lambda} A\,\langle \tanh(Bx + \lambda \boldsymbol{p}\cdot\boldsymbol{s}) \rangle_{\boldsymbol{s},0} &= A\left\langle (\boldsymbol{p}\cdot\boldsymbol{s})\left(1 - \tanh^2(Bx + \lambda\boldsymbol{p}\cdot\boldsymbol{s})\right) \right\rangle_{\boldsymbol{s},0} \\
&= -A\left\langle (\boldsymbol{p}\cdot\boldsymbol{s}) \tanh^2(Bx + \lambda\boldsymbol{p}\cdot\boldsymbol{s}) \right\rangle_{\boldsymbol{s},0} \\
&= -\frac{A}{2}\left\langle (\boldsymbol{p}\cdot\boldsymbol{s}) \left(\tanh^2(Bx + \lambda\boldsymbol{p}\cdot\boldsymbol{s}) - \tanh^2(Bx - \lambda\boldsymbol{p}\cdot\boldsymbol{s})\right) \right\rangle_{\boldsymbol{s},0} \\
&\leq 0
\end{aligned} \tag{A.36}$$

since $\tanh^2(x+y) - \tanh^2(x-y) \geq 0, \forall x, y \geq 0$. As a consequence it holds $\forall \boldsymbol{p}$ and $x \geq 0$

$$A\,\langle \tanh(Bx + \boldsymbol{p}\cdot\boldsymbol{s}) \rangle_{\boldsymbol{s},0} \leq A \tanh(Bx) \leq ABx. \tag{A.37}$$

Therefore we have that

$$|p^\mu - p^\nu| \leq \beta(1 - m_0^2(\beta))\,|p^\mu - p^\nu|\,. \tag{A.38}$$

From the theory of the Curie Weiss model it holds that $\beta(1 - m_0^2(\beta)) < 1$ at any temperature. In fact if $\beta < 1$ it is $m_0(\beta) = 0$ and $\beta(1 - m_0^2(\beta)) = \beta < 1$. On the contrary if $\beta > 1$, $\tanh(\beta x)$ intersects the bisector from above at the point $x = m_0(\beta) > 0$, thus with

$$1 > \frac{d}{dx}\tanh(\beta x)|_{m_0} = \beta(1 - \tanh^2(\beta m_0)) = \beta(1 - m_0^2). \tag{A.39}$$

Therefore the only solution of (A.38) is $|p^\mu - p^\nu| = 0$. □

Proof. of Proposition 6
We have to solve

$$\bar{p} = \sum_{\boldsymbol{s}} s_1 \frac{e^{\beta m_0 \sum_{\mu=1}^M s^\mu}}{(2\cosh(\beta m_0))^M} \tanh(\beta\bar{p}\sum_{\mu=1}^M s^\mu). \tag{A.40}$$

By symmetry ($\boldsymbol{s} \to -\boldsymbol{s}$) this is equivalent to

$$\bar{p} = \sum_{\boldsymbol{s}} s_1 \frac{\cosh(\beta m_0 \sum_{\mu=1}^{M} s^\mu)}{(2\cosh(\beta m_0))^M} \tanh(\beta \bar{p} \sum_{\mu=1}^{M} s^\mu). \tag{A.41}$$

Evaluating the rhs in the point $\bar{p} = m_0$ and denoting $z = 2\cosh(\beta m_0)$ we have

$$\begin{aligned}
&\sum_{\boldsymbol{s}} s_1 \frac{\sinh(\beta m_0 \sum_{\mu=1}^{M} s^\mu)}{(2\cosh(\beta m_0))^M} = z^{-M} \sum_{\boldsymbol{s}} s_1 \sinh(\beta m_0 \sum_{\mu=1}^{M-1} s^\mu + \beta m_0 s^M) \\
= \ & z^{-M} \sum_{s^M} \sum_{s^1,\dots,s^{M-1}} s_1 \left[\sinh(\beta m_0 \sum_{\mu=1}^{M-1} s^\mu) \cosh(\beta m_0 s^M) + \cosh(\beta m_0 \sum_{\mu=1}^{M-1} s^\mu) \sinh(\beta m_0 s^M) \right] \\
= \ & z^{-M} \sum_{s^M} \sum_{s^1,\dots,s^{M-1}} s_1 \sinh(\beta m_0 \sum_{\mu=1}^{M-1} s^\mu) \cosh(\beta m_0 s^M) = \sum_{s^1,\dots,s^{M-1}} s_1 \frac{\sinh(\beta m_0 \sum_{\mu=1}^{M-1} s^\mu)}{(2\cosh(\beta m_0))^{M-1}} \\
= \ & \dots = \sum_{s^1} s_1 \frac{\sinh(\beta m_0 s^1)}{2\cosh(\beta m_0)} = \tanh(\beta m_0).
\end{aligned}$$

□

Proof. of Proposition 7
Thanks to Proposition 5 it is sufficient to evaluate the free energy along the line $p^\mu = \bar{p}$, where

$$f(\bar{p}) = \frac{M}{2}\bar{p}^2 - \frac{1}{\beta} \left\langle \log 2\cosh(\beta \bar{p} \sum_{\mu=1}^{M} s_\mu) \right\rangle_{\boldsymbol{s}}. \tag{A.42}$$

Since $f(\bar{p})$ is an even function we can just study the branch $\bar{p} \geq 0$. As a function of $\bar{p}^2$ the free energy is convex since

$$\frac{df}{d\bar{p}^2} = \frac{f'(\bar{p})}{2\bar{p}} = \frac{M}{2} \left(1 - \frac{\left\langle s_1 \tanh(\beta \bar{p} \sum_{\mu=1}^{M} s_\mu) \right\rangle_{\boldsymbol{s}}}{\bar{p}} \right) \tag{A.43}$$

is an increasing function, therefore the position of the minimum depends on the sign of the derivative in zero. Since for $\beta > 1$

$$\frac{df}{d\bar{p}^2}|_{\bar{p}=0} = \frac{M}{2} \left(1 - \beta - \beta(M-1)m_0^2 \right) < \frac{M}{2}(1-\beta) < 0, \tag{A.44}$$

the minimum is attained away from zero. □

Appendix B

Replica computation of the RS conjecture

We define the model partition function as

$$Z(\boldsymbol{S}) := \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta}{N}\sum_{\mu=1}^{M}\sum_{i<j} s_i^\mu s_j^\mu \xi_i \xi_j\right) = \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta}{2N}\sum_{\mu=1}^{M}\sum_{i,j} s_i^\mu s_j^\mu \xi_i \xi_j - M\frac{\beta}{2}\right). \quad \text{(B.1)}$$

We assume that the system could be aligned with a subset $\ell_1 = \mathcal{O}(1)$ of examples, extracted by the Curie-Weiss at inverse temperature $\hat{\beta}$ and we measure the corresponding overlaps with

$$p^\mu(\boldsymbol{\xi}) = \frac{1}{N}\sum_{i=1}^{N} s_i^\mu \xi_i \qquad \mu = 1, ..., \ell_1 \ . \quad \text{(B.2)}$$

to get

$$Z(\boldsymbol{S}) = \sum_{\boldsymbol{\xi}} \exp\left(\frac{\beta N}{2}\sum_{\mu=1}^{\ell_1}(p^\mu(\xi))^2 + \frac{\beta}{2N}\sum_{\mu=\ell_1+1}^{M}\sum_{i,j} s_i^\mu s_j^\mu \xi_i \xi_j\right) \ . \quad \text{(B.3)}$$

In the present analysis we focus on the case when $\ell_1 = 1$ for the sake of brevity. Our aim is to compute the disorder average free energy density:

$$-\beta f(\beta, \hat{\beta}, \alpha) = \lim_{N\to\infty} \frac{1}{N}\left[\ln Z\right]^{\boldsymbol{S}} \ , \quad \text{(B.4)}$$

here $[\cdots]^{\boldsymbol{S}}$ corresponds to the average versus disorder, given by the dual patterns. It is possible to rewrite the previous expression by exploiting the standard replica trick as

$$-\beta f(\beta, \hat{\beta}, \alpha) = \lim_{\substack{N\to\infty \\ n\to 0}} \frac{\ln[Z^n]^{\boldsymbol{S}}}{Nn} \quad \text{(B.5)}$$

where

$$[Z^n]^{\boldsymbol{S}} = \sum_{\boldsymbol{S}} P_{\hat{\beta}}^{CW}(\boldsymbol{S})\ Z^n(\boldsymbol{S}) = \sum_{\boldsymbol{S}} \prod_{\mu=1}^{M} \frac{1}{z(\hat{\beta})}\ e^{\hat{\beta}/2N\sum_{i,j} s_i^\mu s_j^\mu}\ Z^n(\boldsymbol{S}) \,,$$

and

$$Z^n(\boldsymbol{S}) = \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n} \exp\left(\frac{\beta N}{2}\sum_{a=1}^{n}(p^a(\xi))^2 + \frac{\beta}{2N}\sum_{a=1}^{n}\sum_{\mu=2}^{M}\sum_{i,j} s_i^\mu s_j^\mu \xi_i^a \xi_j^a\right) \ .$$

The summation over $\boldsymbol{\mathcal{S}}$, is intended over $(\boldsymbol{s}^1, \dots, \boldsymbol{s}^M)$. One can subdivide $[Z^n]^{\boldsymbol{\mathcal{S}}}$ in two parts: the one related to the first aligned example (say $\boldsymbol{s}^{\mathbf{1}}$) and the second related to the others. These two pieces will be evaluated separately

$$\begin{aligned}
[Z^n]^{\boldsymbol{\mathcal{S}}} = & \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n} \sum_{\boldsymbol{s}} \frac{1}{z(\hat{\beta})} \exp\left(\frac{\hat{\beta}}{2N} \sum_{i,j} s_i s_j + \frac{\beta N}{2} \sum_a (p^a(\xi))^2 \right) \times \\
& \times \sum_{\boldsymbol{\mathcal{S}}} \prod_{\mu=2}^{M} \frac{1}{z(\hat{\beta})} \exp\left(\frac{\hat{\beta}}{2N} \sum_{i,j} s_i^\mu s_j^\mu + \frac{\beta}{2N} \sum_a \sum_{i,j} s_i^\mu s_j^\mu \xi_i^a \xi_j^a \right) , \\
= & \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n} \sum_{\boldsymbol{s}} \frac{1}{z(\hat{\beta})} \exp\left(\frac{\hat{\beta} N}{2} m_0^2(s) + \frac{\beta N}{2} \sum_a (p^a(\xi))^2 \right) \times \\
& \times \left(\frac{2^N}{z(\hat{\beta})} \overline{e^{\hat{\beta}/2N \sum_{i,j} s_i s_j + \beta/2N \sum_a \sum_{i,j} s_i s_j \xi_i^a \xi_j^a}}^{\boldsymbol{s}} \right)^{M-1} ,
\end{aligned}$$

where in the first line we simply drop out the index for the first example and we introduced the magnetization as $m_0(\boldsymbol{s}) = 1/N \sum_i^N s_i$, while in the second line we denoted by $\overline{\cdots}^{\,\boldsymbol{s}}$ the average w.r.t a single example. The second term can be computed by introducing Gaussian variables $(z^1, \dots, z^n)$ and z to linearize the exponent as

$$\int \prod_{a=1}^{n} Dz^a Dz \; \overline{e^{\sqrt{\beta/N} \sum_i \sum_a z^a \xi_i^a s_i + \sqrt{\hat{\beta}/N} \sum_i z s_i}}^{\boldsymbol{s}} = \int \prod_{a=1}^{n} Dz^a Dz \; e^{\sum_i \ln\cosh(\sqrt{\beta/N} \sum_a z^a \xi_i^a + \sqrt{\hat{\beta}/N} z)}$$

$$\approx \int \prod_{a=1}^{n} Dz^a Dz \; e^{\beta/2\left[\sum_{a\neq b} z_a z_b q_{ab} + \sum_a z_a^2\right] + \hat{\beta}/2 z^2 + \sqrt{\hat{\beta}\beta} z \sum_a z_a m_a} =: \det\big(\boldsymbol{\Xi}(\boldsymbol{q}, \boldsymbol{m})\big)^{-1/2} , \qquad \text{(B.6)}$$

where in the last line we have expanded the $\ln\cosh(x)$ and we have introduced the quantities

$$q_{ab}(\boldsymbol{\xi}) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^a \xi_i^b , \quad m_a(\boldsymbol{\xi}) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^a . \qquad \text{(B.7)}$$

The averaged partition function thus becomes

$$[Z^n]^{\boldsymbol{\mathcal{S}}} = \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n} \sum_{\boldsymbol{s}} e^{\frac{\hat{\beta} N}{2} m_0(\boldsymbol{s})^2 + \frac{\beta N}{2} \sum_a (p^a(\boldsymbol{\xi}))^2} \frac{2^{N(M-1)}}{z(\hat{\beta})^M} \det\big(\boldsymbol{\Xi}(\boldsymbol{q}, \boldsymbol{m})\big)^{(1-M)/2}. \qquad \text{(B.8)}$$

If we denote by $\mathcal{D}(m_0)$ the density of states for the $\boldsymbol{s}$ configuration

$$\mathcal{D}(m_0) = \sum_{\boldsymbol{s}} \delta\left(m_0 - \frac{1}{N} \sum_i s_i \right) \propto \sum_{\boldsymbol{s}} \int d\hat{m}_0 \exp -N\hat{m}_0 \left(m_0 - \frac{1}{N} \sum_i s_i \right)$$

and with $\mathcal{D}(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p})$ the one related to the states of the student machine

$$\begin{aligned}\mathcal{D}(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p}) &= \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n}\prod_{a<b}\delta\left(q_{ab}-\frac{1}{N}\sum_i\xi_i^a\xi_i^b\right)\prod_a\delta\left(m_a-\frac{1}{N}\sum_i\xi_i^a\right)\prod_a\delta\left(p_a-\frac{1}{N}\sum_i s_i\xi_i^a\right)\\ &\propto \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n}\int\prod_{a<b}d\hat{q}_{ab}\ \exp\left(-N\sum_{a<b}\hat{q}_{ab}\left(q_{ab}-\frac{1}{N}\sum_i\xi_i^a\xi_i^b\right)\right)\times\\ &\times\int\prod_a d\hat{m}_a\ \exp\left(-N\sum_a\hat{m}_a\left(m_a-\frac{1}{N}\sum_i\xi_i^a\right)\right)\times\\ &\times\int\prod_a d\hat{p}_a\ \exp\left(-N\sum_a\hat{p}_a\left(p_a-\frac{1}{N}\sum_i s_i\xi_i^a\right)\right)\end{aligned}$$

the averaged partition function becomes

$$[Z^n]^{\boldsymbol{S}}\propto\int dm_0\,\mathcal{D}(m_0)\int d\boldsymbol{q}d\boldsymbol{m}d\boldsymbol{p}\,\mathcal{D}(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p})\,e^{\frac{\hat{\beta}N}{2}m_0^2+\frac{\beta N}{2}\sum_a(p^a)^2}\,\frac{2^{N(M-1)}}{z(\hat{\beta})^M}\det\big(\Xi(\boldsymbol{q},\boldsymbol{m})\big)^{(1-M)/2},$$

and then by saddle point approximation the free energy density results as

$$-\beta f\approx\lim_{\substack{N\to\infty\\ n\to 0}}\frac{1}{Nn}\ln\left(e^{N\,\mathrm{Extr}\,f(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p},\boldsymbol{m_0};\hat{\boldsymbol{q}},\hat{\boldsymbol{m}},\hat{\boldsymbol{p}},\hat{\boldsymbol{m}}_{\boldsymbol{0}})}\right)\approx\lim_{n\to 0}\frac{1}{n}\,\mathrm{Extr}\,f(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p},\boldsymbol{m_0};\hat{\boldsymbol{q}},\hat{\boldsymbol{m}},\hat{\boldsymbol{p}},\hat{\boldsymbol{m}}_{\boldsymbol{0}})\tag{B.9}$$

where

$$\begin{aligned}f(\boldsymbol{q},\boldsymbol{m},\boldsymbol{p},\boldsymbol{m_0};\hat{\boldsymbol{q}},\hat{\boldsymbol{m}},\hat{\boldsymbol{p}},\hat{\boldsymbol{m}}_{\boldsymbol{0}}) = {}&\frac{\beta}{2}\sum_a(p^a)^2+\frac{\hat{\beta}}{2}m_0^2+(M-1)\ln 2-Mf_{CW}-\frac{\alpha}{2}\ln\det\big(\Xi(\boldsymbol{q},\boldsymbol{m})\big)\\ &-\hat{m}_0m_0-\sum_a\hat{p}^ap^a-\sum_{a<b}\hat{q}^{ab}q^{ab}-\sum_a\hat{m}^am^a\\ &+\ln\sum_s e^{\hat{m}_0 s}\sum_{\xi^1,\dots,\xi^n}\exp\left(\sum_a\hat{p}^a\,s\,\xi_a+\sum_{a<b}\hat{q}^{ab}\xi_a\xi_b+\sum_a\hat{m}^a\xi_a\right).\end{aligned}$$

The Replica symmetric (RS) ansatz assumes $q_{ab} = q, m_a = m, p_a = p,\ \forall a, b = 1, \ldots, n$. Under the RS ansatz the term $\ln\det\big(\Xi(\boldsymbol{q}, \boldsymbol{m})\big)$ can be computed as

$$
\begin{aligned}
\det\big(\Xi(q,m)\big)^{-1/2} &:= \int \prod_{a=1}^{n} Dz_a Dz \exp\left(\frac{\beta q}{2}\sum_{a\neq b} z_a z_b + \frac{\beta}{2}\sum_a z_a^2 + \frac{\hat\beta z^2}{2} + \sqrt{\hat\beta\beta} m \sum_a z_a z\right) \\
&= \int \prod_{a=1}^{n} Dz_a \exp\left(\frac{\beta q}{2}\sum_{a\neq b} z_a z_b + \frac{\beta}{2}\sum_a z_a^2\right) \times \\
&\qquad\qquad \times \int \frac{dz}{\sqrt{2\pi}} \exp\left(-(1-\hat\beta)\frac{z^2}{2} + \sqrt{\hat\beta\beta} m \sum_a z_a z\right) \\
&= \int \prod_{a=1}^{n} Dz_a \exp\left(\frac{\beta q}{2}\sum_{a\neq b} z_a z_b + \frac{\beta}{2}\sum_a z_a^2\right) (1-\hat\beta)^{-\frac{1}{2}} \times \\
&\qquad\qquad \times \exp\left(\frac{\hat\beta\beta m^2}{2(1-\hat\beta)}\big(\sum_{a\neq b} z_a z_b + \sum_a z_a^2\big)\right) \\
&= (1-\hat\beta)^{-\frac{1}{2}} \det\left(\mathbb{I} - \beta\boldsymbol{q} - \frac{\hat\beta\beta m^2}{2(1-\hat\beta)}\mathbf{1}\right)^{-\frac{1}{2}} ,
\end{aligned}
$$

where we introduced the matrix $\boldsymbol{q} = (1-q)\boldsymbol{I} + q\mathbf{1}$. A matrix of the form $A\boldsymbol{I} + B\mathbf{1}$ has eigenvalues $\lambda_1 = A + nB$ with multiplicity 1 and $\lambda_2 = A$ with multiplicity $n-1$. Hence $\ln\det\big(\Xi(\boldsymbol{q}, \boldsymbol{m})\big)$ can be expressed in the $n \to 0$ limit as

$$
\begin{aligned}
\ln\det\big(\Xi(\boldsymbol{q}, \boldsymbol{m})\big) &= \ln\left((1-\hat\beta)\det\left(\mathbb{I} - \beta\boldsymbol{q} - \frac{\hat\beta\beta m^2}{2(1-\hat\beta)}\mathbf{1}\right)\right) \\
&= \ln\Big[(1-\hat\beta)(1-\beta+\beta q) - n((1-\hat\beta)\beta q + \hat\beta\beta m^2)\Big] + (n-1)\ln\Big[(1-\hat\beta)(1-\beta+\beta q)\Big] \\
&\approx n\left[\ln\Big((1-\hat\beta)(1-\beta+\beta q)\Big) - \frac{\beta(1-\hat\beta)q + \hat\beta\beta m^2}{(1-\hat\beta)(1-\beta+\beta q)}\right] .
\end{aligned}
$$

Finally the term related to the density of states becomes under the RS ansatz

$$
\begin{aligned}
&\sum_s e^{\hat m_0 s - n\frac{\hat q}{2}} \int \mathcal{D}z \left[\sum_\xi \exp\Big(\xi(\hat p\ s\ + \sqrt{\hat q} z + \hat m)\Big)\right]^n \\
&= 2\mathbb{E}_s\, e^{\hat m_0 s - n\frac{\hat q}{2}} \int \mathcal{D}z \left[2\mathbb{E}_\xi \exp\Big(\xi(\hat p\ s\ + \sqrt{\hat q} z + \hat m)\Big)\right]^n , \\
&\approx e^{-n\frac{\hat q}{2}} \left(\ 2\mathbb{E}_s\, e^{\hat m_0 s} + n\, 2\mathbb{E}_s\, e^{\hat m_0 s} \int \mathcal{D}z\ \ln 2\mathbb{E}_\xi\, e^{\xi(\hat p\ s\ + \sqrt{\hat q} z + \hat m)}\right) \\
&\approx e^{-n\frac{\hat q}{2}}\ 2\mathbb{E}_s e^{\hat m_0 s} \left(1 + \frac{n}{\mathbb{E}_s e^{\hat m_0 s}}\mathbb{E}_s\, e^{\hat m_0 s} \int \mathcal{D}z\ \ln 2\mathbb{E}_\xi\, e^{\xi(\hat p\ s\ + \sqrt{\hat q} z + \hat m)}\right) .
\end{aligned}
$$

Putting all together we define

$$
-\beta f^{RS}(\beta, \hat\beta, \gamma) = \lim_{n\to 0} \frac{1}{n} \operatorname{Extr} f^{RS}(p, q, m; \hat p, \hat q, \hat m) . \tag{B.10}
$$

where

$$
\begin{aligned}
f^{RS}(p,q,m;\hat{p},\hat{q},\hat{m}) = &\ln 2 + \ln\cosh\hat{m}_0 + \frac{\hat{\beta}}{2}m_0^2 + (M-1)\ln 2 - Mf_{CW} - \hat{m}_0 m_0 + n\frac{\beta}{2}p^2 - n\hat{p}p + \\
&+ \frac{n}{2}\hat{q}q - n\hat{m}m - n\frac{\alpha}{2}\ln\Big((1-\hat{\beta})(1-\beta+\beta q)\Big) + n\frac{\alpha}{2}\frac{\beta(1-\hat{\beta})q + \hat{\beta}\beta m^2}{(1-\hat{\beta})(1-\beta+\beta q)} + \\
&- n\frac{\hat{q}}{2} + n\ln 2 + \frac{n}{\mathbb{E}_s e^{\hat{m}_0 s}}\,\mathbb{E}_s e^{\hat{m}_0 s}\int \mathcal{D}z \ln\cosh\Big(\hat{p}s + z\sqrt{\hat{q}} + \hat{m}\Big) + \mathcal{O}(n^2)\,.
\end{aligned}
$$

Since $\hat{\beta} < 1$, by extremizing with respect to $(\hat{m}_0, m_0)$ we obtain $m_0 = 0$ and therefore

$$
\begin{aligned}
-\beta f^{RS}(\beta,\hat{\beta},\alpha) = \operatorname{Extr}\Bigg[&- \frac{\alpha}{2}\ln\Big((1-\hat{\beta})(1-\beta+\beta q)\Big) + \frac{\alpha}{2}\frac{\beta(1-\hat{\beta})q + \hat{\beta}\beta m^2}{(1-\hat{\beta})(1-\beta+\beta q)} + \frac{1}{2}\hat{q}q + \\
&- \hat{m}m - \frac{\hat{q}}{2} - \hat{p}p + \frac{\beta}{2}p^2 + \ln 2 + \mathbb{E}_s \int \mathcal{D}z \ln\cosh\Big(\hat{p}s + z\sqrt{\hat{q}} + \hat{m}\Big)\Bigg].
\end{aligned}
$$

The equations for the saddle point are

$$
\begin{aligned}
m &= \mathbb{E}_s \int Dz \tanh(\hat{m} + z\sqrt{\hat{q}} + \hat{p}s) \\
\hat{m} &= \frac{\alpha\hat{\beta}\beta m}{(1-\hat{\beta})(1-\beta+\beta q)} \\
q &= \mathbb{E}_s \int Dz \tanh^2(\hat{m} + z\sqrt{\hat{q}} + \hat{p}s) \\
\hat{q} &= \frac{\alpha\hat{\beta}\beta^2(m^2)}{(1-\hat{\beta})(1-\beta+\beta q)^2} + \frac{\alpha\beta^2 q}{(1-\beta+\beta q)^2} \\
p &= \mathbb{E}_s \int Dz \tanh(\hat{m} + z\sqrt{\hat{q}} + \hat{p}s)s \\
\hat{p} &= \beta p
\end{aligned}
$$

where integration by parts was used to calculate q. In particular on the Nishimori line, where $\hat{\beta} = \beta < 1$ and $p = 0$, it is

$$
-\beta f^{RS}(\beta,\alpha) = \operatorname{Extr}\left[-\frac{\alpha}{2}\ln\det\Xi\Big|_{\hat{\beta}=\beta} + \frac{1}{2}\hat{q}q - \hat{m}m - \frac{\hat{q}}{2} + \ln 2 + \int Dz \ln\cosh(\hat{m} + z\sqrt{\hat{q}})\right] \tag{B.11}
$$

and

$$
\begin{aligned}
m &= \int Dz \tanh(\hat{m} + z\sqrt{\hat{q}}) \\
\hat{m} &= \frac{\alpha\beta^2 m}{(1-\beta)(1-\beta+\beta q)} \\
q &= \int Dz \tanh^2(\hat{m} + z\sqrt{\hat{q}}) \\
\hat{q} &= \frac{\alpha\beta^3(m^2-q^2)}{(1-\beta)(1-\beta+\beta q)^2} + \frac{\alpha\beta^2 q}{(1-\beta)(1-\beta+\beta q)}\,.
\end{aligned}
$$

Chapter 3

The effect of priors on Learning with Restricted Boltzmann Machines

In this chapter, we build upon the results of the Dual Hopfield network concerning the critical dataset size required for effective learning. Specifically, we expand the teacher-student framework to consider scenarios where a student RBM learns from examples generated by a teacher RBM. This analysis focuses on the impact of unit priors on learning efficiency.
We explore a parametric class of priors that interpolate between continuous (Gaussian) and binary variables, as introduced at the end of chapter 1. This framework allows us to model various architectural designs for both the teacher and student RBMs.

By examining the phase diagram of the posterior distribution under both Bayes-optimal and mismatched conditions, we reveal the existence of a triple point that determines the critical dataset size necessary for learning through generalization. This critical size is significantly influenced by the properties of the teacher RBM—and consequently, the structure of the data—but is independent of the properties of the student RBM.
However, a careful selection of the student RBM's priors can still enhance training efficiency. Specifically, it can extend the signal retrieval region, thereby improving learning outcomes.

Introduction and motivations

Restricted Boltzmann Machines (RBMs) [98–100] are a type of generative stochastic neural network (NN) that can learn a probability distribution over its set of inputs. Their ability to learn rich internal representations [101–103] makes them a fundamental building block in deep learning architectures [21, 104, 105].

We recall RBMs have a visible layer of N units $\boldsymbol{s} = \{s_i\}_{1\leq i\leq N}$, a hidden layer of P units $\boldsymbol{\tau} = \{\tau_\mu\}_{1\leq\mu\leq P}$ and a set of internal connections $\boldsymbol{\xi} = \{\xi_i^\mu\}_{1\leq i\leq N, 1\leq\mu\leq P}$. Given $\boldsymbol{\xi}$, the joint distribution of the visible and hidden layers has the Gibbs structure

$$P\left(\boldsymbol{s},\boldsymbol{\tau}\middle|\boldsymbol{\xi}\right) = z^{-1}\left(\boldsymbol{\xi}\right) P\left(\boldsymbol{s}\right) P\left(\boldsymbol{\tau}\right) \exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu s_i \tau_\mu\right), \tag{3.1}$$

where $P\left(\boldsymbol{s}\right)$ and $P\left(\boldsymbol{\tau}\right)$ are priors on the visible and hidden layers, respectively, β is the inverse temperature modulating the weights intensity and $z\left(\boldsymbol{\xi}\right)$ is the partition function normalizing the distribution. RBMs with given weigths can generate data $\boldsymbol{s}$ by sampling the marginal distribution

$$P\left(\boldsymbol{s}\middle|\boldsymbol{\xi}\right) = z^{-1}\left(\boldsymbol{\xi}\right)\psi\left(\boldsymbol{s};\boldsymbol{\xi}\right) = z^{-1}\left(\boldsymbol{\xi}\right)P(\boldsymbol{s})\mathbb{E}_{\tau}\left[\exp\left(\sqrt{\frac{\beta}{N}}\sum_{i,\mu}^{N,P}\xi_i^{\mu}s_i\tau_{\mu}\right)\right] \tag{3.2}$$

where $\mathbb{E}_{\tau}$ is the expectation over the hidden units. Marginal Gibbs distributions of the form (3.2) are also known as generalized Hopfield networks [55, 64, 72–74, 78]. RBMs with generic priors, i.e. the generalized Hopfield networks, have been extensively studied in the statistical mechanics literature, for their properties as models of associative memory [34, 41] and their challenging connection with the Parisi theory of spin glasses [47, 56, 106–109]. These studies only address the *direct problem* where the statistical properties of the generative model are investigated.

In a machine learning context the weights are learned [110, 111] from a dataset of examples, whose distribution the RBM must reconstruct. Therefore the study of the *inverse problem* of weights optimization is fundamental for a theoretical understanding of the RBM learning mechanisms. As seen in Chapter 2, a very useful approach is the *teacher-student* setting where a *student* RBM is trained with data produced by another *teacher* RBM [65, 94, 112]. Such studies are crucial to isolate individual characteristics of data and machine architecture in a controlled environment and explain their effects on the NN training. For instance in the case of the Dual Hopfield we saw the effects of the inference temperature in relation with the dataset's size and noise. Other examples are [113] where the choice of the hidden layer's size in relation to the number of patterns in the data and their correlation is studied; or [114] where the role of possible multi-body interactions is analyzed in the context of the so called Dense Hopfield models.

3.1 The RBM Teacher-Student framework

Consider the inference problem of a student machine (S-RBM) trained over a dataset generated by a teacher machine (T-RBM). The T-RBM model is built upon P quenched patterns $\hat{\boldsymbol{\xi}} \sim P(\hat{\boldsymbol{\xi}})$. A dataset of M examples $\boldsymbol{\mathcal{S}} := \{\boldsymbol{s}^a\}_{a=1}^M = \{s_1^a, \ldots, s_N^a\}_{a=1}^M$ is generated by drawing independent samples from the T-RBM distribution

$$P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) = \prod_{a=1}^{M} P(\boldsymbol{s}^a|\hat{\boldsymbol{\xi}}) = \prod_{a=1}^{M} z^{-1}(\hat{\boldsymbol{\xi}})P(\boldsymbol{s}^a)\,\mathbb{E}_{\hat{\tau}}\exp\left(\sqrt{\frac{\hat{\beta}}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}s_i^a\hat{\xi}_i^{\mu}\hat{\tau}_{\mu}\right) . \tag{3.3}$$

and provided to the student. The student is trained over $\boldsymbol{\mathcal{S}}$ to recover its structure, i.e. the teacher patterns. In this process the student patterns $\boldsymbol{\xi}$ are optimized. In a Bayesian framework, they are sampled from the posterior distribution

$$\begin{aligned} P(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}) &= \frac{P(\boldsymbol{\xi})\prod_{a=1}^{M}P(\boldsymbol{s}^a|\boldsymbol{\xi})}{P(\boldsymbol{\mathcal{S}})} = Z^{-1}(\boldsymbol{\mathcal{S}})P(\boldsymbol{\xi})\prod_{a=1}^{M}z^{-1}(\boldsymbol{\xi})\,\mathbb{E}_{\tau}\exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}s_i^a\xi_i^{\mu}\tau_{\mu}\right) \\ &= Z^{-1}(\boldsymbol{\mathcal{S}})z^{-M}(\boldsymbol{\xi})\prod_{\mu=1}^{P}P(\boldsymbol{\xi}^{\mu})\mathbb{E}_{\tau}\exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{a=1}^{M}\xi_i^{\mu}s_i^a\tau^a\right) \end{aligned} \tag{3.4}$$

with partition function

$$Z(\boldsymbol{\mathcal{S}}) = \mathbb{E}_{\boldsymbol{\xi}}z^{-M}(\boldsymbol{\xi})\prod_{\mu=1}^{P}\mathbb{E}_{\tau}\exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{a=1}^{M}\xi_i^{\mu}s_i^a\tau^a\right) . \tag{3.5}$$

In general, the student is unaware of the properties of the T-RBM, therefore we assume $\hat{\beta} \neq \beta$, $P(\hat{\boldsymbol{\xi}}) \neq P(\boldsymbol{\xi})$, $P(\hat{\boldsymbol{\tau}}) \neq P(\boldsymbol{\tau})$. Without the term $z^{-M}(\boldsymbol{\xi})$, the posterior distribution (3.4) would correspond to P independent generalized Hopfield models, one for each student pattern $\boldsymbol{\xi}^{\mu}$, where the examples $\{\boldsymbol{s}^a\}_{a=1}^{M}$ act as dual patterns [30, 65, 114]. In [113] it is shown that, for teacher patterns that are uncorrelated, the interaction term $z(\boldsymbol{\xi})^{-M}$ has the only effect of enforcing orthogonality between student patterns. Furthermore, the learning performance of a S-RBM with P hidden units learning P teacher patterns is equivalent to that of P separate S-RBM with a single hidden unit, each learning one pattern. For this reason, in the following we assume $P = 1$, see Fig. 3.1.

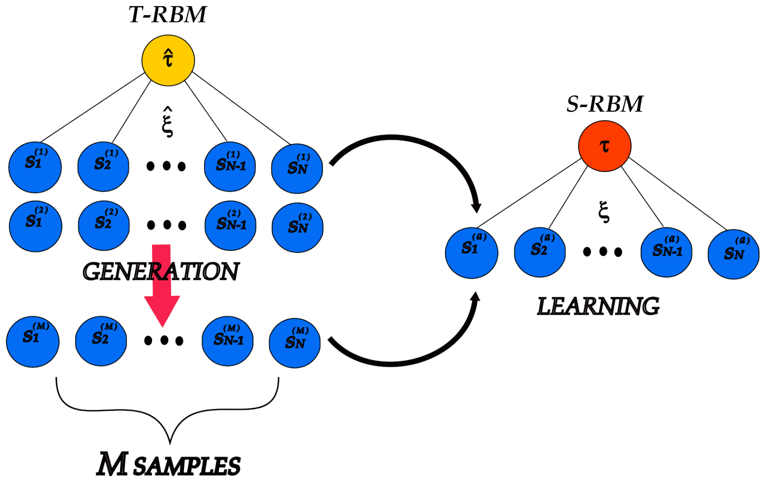


Figure 3.1: Inverse teacher-student problem. On the left, the T-RBM generates the data, following the interactions of the planted signal $\hat{\boldsymbol{\xi}}$. On the right, a representation of the S-RBM, which tries to align its own weight vector $\boldsymbol{\xi}$ towards $\hat{\boldsymbol{\xi}}$ using information extracted from the dataset $\boldsymbol{\mathcal{S}}$.

The *efficiency* of the learning mechanism depends on how effectively the student can recover the ground truth encoded in the teacher's patterns. The overlap

$$Q(\hat{\boldsymbol{\xi}}, \boldsymbol{\xi}) = \frac{1}{N} \sum_{i=1}^{N} \hat{\xi}_i \xi_i, \tag{3.6}$$

serves as a reliable measure of learning performance, as it quantifies the similarity between the student's inferred patterns and the teacher's true patterns. Since we are considering a generic RBM as teacher machine, we cannot use the gauge transformation (2.14) and consequently a dataset drawn from a CW. The learning *efficiency* is well known to depend on the data-to-neuron ratio, α, the noise level in the examples, $\hat{T} = \hat{\beta}^{-1}$, and the parameter β, often referred to as the student's inverse inference temperature, which reflects the typical magnitude of the student's weights and acts as a form of regularization.

Our goal is to analyze different scenarios, choosing RBMs with different priors for both teacher and student to observe the effects of these hyper-parameters on the learning efficiency. To this aim, we choose priors from a parametric family of distributions. For each label $x \in \{\hat{\xi}, \hat{\tau}, \xi, \tau, s\}$, consider the random variable

$$\sqrt{\Omega_x} g_x + \sqrt{1 - \Omega_x} \epsilon_x \,. \tag{3.7}$$

where $\Omega_x \in [0, 1]$, $g_x \sim N(0, 1)$ and ϵ_x is a Rademacher random variables taking values ± 1. We denote its probability distribution with P_{Ω_x}, which interpolates between a standard gaussian and a binary distributions, having zero mean and unitary variance for any choice of Ω_x. We assume the entries of the teacher pattern $\{\hat{\xi}_i\}_{i=1}^{N}$ are drawn independently from $P_{\Omega_{\hat{\xi}}}$. Similarly $\hat{\tau} \sim P_{\Omega_{\hat{\tau}}}$,

$\tau \sim P_{\Omega_\tau}$, $\xi_i \sim P_{\Omega_\xi}$ and $s_i^\mu \sim P_{\Omega_s}$. We indicate with $\boldsymbol{\Omega} = \{\Omega_x\}_{x\in\{\hat{\xi},\tau,\xi,\tau,s\}}$ the set of prior hyperparameters.

We leverage techniques from statistical mechanics to calculate the expected value of the overlap (3.6) in the limit of large dimension and large dataset $N, M \to \infty$, $M/N = \alpha$ as a function of $\boldsymbol{\Omega}$. Specifically, we obtain it as a byproduct of the limiting quenched free energy

$$-\beta f(\boldsymbol{\Omega};\alpha,\hat{\beta},\beta) = \lim_{M,N\to\infty} \frac{1}{N}\mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S}} \log\left[Z(\boldsymbol{S})\right], \tag{3.8}$$

where $\mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S}}$ is the expected value w.r.t. the joint distribution of the teacher pattern and generated dataset. In fact, we will show in Section 3.2 that Eq. (3.8) can be expressed as the result of a variational principle w.r.t. a set of order parameters.

3.2 Free energy and saddle point equations

The quenched free energy can be computed exploiting the replica method in the replica symmetry (RS) approximation (see the Appendix C) and reads as

$$-\beta f^{RS}(\boldsymbol{\Omega};\alpha,\hat{\beta},\beta) = -\beta\,\mathrm{Extr}_{\Lambda,\Lambda_\tau,\hat{\Lambda},\hat{\Lambda}_\tau}\,\hat{f}(\Lambda,\Lambda_\tau,\hat{\Lambda},\hat{\Lambda}_\tau)\,, \tag{3.9}$$

where we introduced the sets of parameters $\Lambda = \{p,m,q,d\}$ and $\Lambda_\tau = \{p_\tau,m_\tau,q_\tau,d_\tau\}$, together with their conjugates $\hat{\Lambda}$ and $\hat{\Lambda}_\tau$. The function $\hat{f}$ to be extremized is defined as

$$\begin{aligned}
-\beta\hat{f}(\Lambda,\Lambda_\tau,\hat{\Lambda},\hat{\Lambda}_\tau) = {}& \frac{1}{2}\hat{q}q + \alpha\frac{1}{2}\hat{q}_\tau q_\tau - \hat{m}m - \alpha\hat{m}_\tau m_\tau - \hat{d}d - \alpha\hat{d}_\tau d_\tau + \alpha^2\frac{\beta}{2}\Omega_\xi p_\tau^2 + \frac{\beta}{2}\Omega_\tau p^2 + \\
&+ \frac{\alpha}{\mathbb{E}_{\hat{\tau}}e^{\hat{d}_{\hat{\tau}}\hat{\tau}^2}}\left\langle \mathbb{E}_{\hat{\tau}}e^{\hat{d}_{\hat{\tau}}\hat{\tau}^2}\mathbb{E}_s \log\mathbb{E}_\tau \exp\left\{-(\frac{\hat{q}_\tau}{2}-\hat{d}_\tau)\tau^2 + \tau(\hat{m}_\tau\hat{\tau} + \sqrt{\hat{q}_\tau}z + \hat{p}_\tau s)\right\}\right\rangle_z + \\
&+ \frac{1}{\mathbb{E}_{\hat{\xi}}e^{\hat{d}_{\hat{\xi}}\hat{\xi}^2}}\left\langle \mathbb{E}_{\hat{\xi}}e^{\hat{d}_{\hat{\xi}}\hat{\xi}^2}\mathbb{E}_s \log\mathbb{E}_\xi \exp\left\{-(\frac{\hat{q}}{2}-\hat{d})\xi^2 + \xi(\hat{m}\hat{\xi} + \sqrt{\hat{q}}z + \hat{p}s)\right\}\right\rangle_z + \\
&- \frac{\beta\alpha}{2}qq_\tau + \alpha\sqrt{\hat{\beta}\beta}mm_\tau - \alpha p_\tau\hat{p}_\tau - p\hat{p} + \frac{\beta\alpha}{2}dd_\tau - \alpha\log\mathbb{E}_\tau e^{\frac{\beta}{2}d\tau^2}\,,
\end{aligned} \tag{3.10}$$

where $\langle . \rangle_z$ denotes the expectation w.r.t. a standard gaussian random variable $z \sim \mathcal{N}(0,1)$. Extremization gives

$$\hat{p}_\tau = \beta\alpha\,\Omega_\xi\,p_\tau, \quad \hat{m}_\tau = \sqrt{\beta\hat{\beta}}m, \quad \hat{q}_\tau = \beta q, \quad \hat{d}_\tau = \beta\frac{d}{2}, \tag{3.11}$$

$$\hat{p} = \beta\Omega_\tau p, \quad \hat{m} = \alpha\sqrt{\beta\hat{\beta}}m_\tau, \quad \hat{q} = \alpha\beta q_\tau, \quad \hat{d} = \frac{\alpha\beta}{2}\left(d_\tau - \langle\tau^2\rangle_\tau\right), \tag{3.12}$$

for the conjugated parameters where we define $\langle\tau^2\rangle_\tau$ as follows $\langle\tau^2\rangle_\tau = \mathbb{E}_\tau[\tau^2 e^{\frac{\beta}{2}d\tau^2}]/\mathbb{E}_\tau[e^{\frac{\beta}{2}d\tau^2}]) = (1-\beta\,d\,\Omega_\tau^2)/(1-\beta\,d\,\Omega_\tau)^2$. The order parameters have to be solutions of the following saddle point equations:

$$p = \left\langle s \langle \xi \rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}}, \qquad p_\tau = \left\langle s \langle \tau \rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}}, \tag{3.13}$$

$$m = \left\langle \hat{\xi} \langle \xi \rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}}, \qquad m_\tau = \left\langle \hat{\tau} \langle \tau \rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}}, \tag{3.14}$$

$$q = \left\langle \langle \xi \rangle^2_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}}, \qquad q_\tau = \left\langle \langle \tau \rangle^2_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}}, \tag{3.15}$$

$$d = \left\langle \langle \xi^2 \rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}}, \qquad d_\tau = \left\langle \langle \tau^2 \rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}}. \tag{3.16}$$

The averages $\langle . \rangle_{\hat{\xi}}$ and $\langle . \rangle_s$ are respectively w.r.t. $P_{\Omega_{\hat{\xi}}}(\hat{\xi})$ and $P_{\Omega_s}(s)$, while $\langle . \rangle_{\hat{\tau}}$ is over $P_{\Omega_{\hat{\tau}}}(\hat{\tau})e^{\frac{\hat{\beta}}{2}\hat{\tau}^2}$. Finally $\langle . \rangle_{\xi|z,s,\hat{\xi}}$ and $\langle . \rangle_{\tau|z,s,\hat{\tau}}$ stand for the expectation over respectively

$$P_{\Omega_\xi}(\xi) e^{-\frac{\alpha}{2}\beta\left(q_\tau - d_\tau + \langle \tau^2 \rangle_\tau\right)\xi^2 + \xi\left(\alpha\sqrt{\hat{\beta}\beta} m_\tau \hat{\xi} + \sqrt{\alpha\beta q_\tau} z + \beta\Omega_\tau p s\right)}, \tag{3.17}$$

$$P_{\Omega_\tau}(\tau) e^{-\frac{\beta}{2}(q-d)\tau^2 + \tau\left(\sqrt{\hat{\beta}\beta} m \hat{\tau} + \sqrt{\beta q} z + \beta\alpha\Omega p_\tau s\right)}. \tag{3.18}$$

We recall the optimal order parameters p, m, q, d solving Eqs. (3.13-3.16) have to be interpreted as expected overlaps. First of all, m is the limiting expected overlap between the teacher pattern $\hat{\boldsymbol{\xi}}$ and the student pattern $\boldsymbol{\xi}$, i.e.

$$m = \lim_{N,M\to\infty} \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}} \left[Q(\hat{\boldsymbol{\xi}}, \boldsymbol{\xi}) \right] \tag{3.19}$$

where $\mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}}$ is the expectation w.r.t. the joint distribution $P(\hat{\boldsymbol{\xi}})P(\boldsymbol{S}|\hat{\boldsymbol{\xi}})P(\boldsymbol{\xi}|\boldsymbol{S})$ of teacher patterns, dataset and student patterns. Similarly p is the limiting expected overlap between the student pattern $\boldsymbol{\xi}$ and an example $\boldsymbol{s}^a$, i.e.

$$p = \lim_{N,M\to\infty} \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}} \left[Q(\boldsymbol{\xi}, \boldsymbol{s}^a) \right]. \tag{3.20}$$

Then q is the limiting expected overlap between any two student patterns $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ from two independent posterior samples, i.e.

$$q = \lim_{N,M\to\infty} \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}^1\times\boldsymbol{\xi}^2} \left[Q(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2) \right] = \lim_{N,M\to\infty} \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S}} \left[Q(\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{S}}\left[\boldsymbol{\xi}\right], \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{S}}\left[\boldsymbol{\xi}\right]) \right] \tag{3.21}$$

Finally d is the limiting expected self-overlap i.e.

$$d = \lim_{N,M\to\infty} \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}} \left[Q(\boldsymbol{\xi}, \boldsymbol{\xi}) \right]. \tag{3.22}$$

The RS saddle-point Eqs.(3.13-3.16) can be solved by numerical iteration for any values of the hyper-parameters $\hat{\beta}$, β, α and $\boldsymbol{\Omega}$. We expect the RS solution to be exact when the student is fully informed about the teacher's prior and hyperparameters and matches them with its own, i.e. $\beta = \hat{\beta}$, $\Omega_{\hat{\xi}} = \Omega_\xi$ and $\Omega_{\hat{\tau}} = \Omega_\tau$. This is the Nishimori line for this version of RBM teacher-student problem. Outside of this regime, i.e. in the mismatched setting, replica symmetry breaking corrections are expected at low temperature. It is important to remark on the role of $z(\boldsymbol{\xi})^{-M}$ in Eq. (3.5). This term is responsible for the emergence of the quadratic term $-\frac{\alpha\beta}{2}\xi^2\langle\tau^2\rangle_\tau$ within the distribution (3.17). This mechanism enables the automatic regularization of the self-overlap, consistently yielding $d = 1$ as the solution to Eq. (3.16), as will be shown throughout the following sections, and in the Appendix C.

3.3 Exploring arbitrary priors: the Bayesian Optimal setting

With the set of equations (3.13-3.16), we can gain deeper insights into what is happening in the student learning process. In this Section we analyze the ideal situation of the Bayesian optimality setting [45], in which the student has access to all the information about the generating process and therefore it is able to mimic the teacher architecture and hyper-parameters to have a better chance of recovering the ground truth $\hat{\boldsymbol{\xi}}$. In this regime we therefore assume $\hat{\beta} = \beta$, $\Omega_{\hat{\xi}} = \Omega_\xi$ and $\Omega_{\hat{\tau}} = \Omega_\tau$. At low temperature all the examples have macroscopic alignment with the signal and the student can easily learn it perfectly. Conversely, at a sufficiently high generating temperature $\hat{\beta}$ the examples have vanishing overlap with the teacher pattern. At the same time, the high inference temperature $\beta = \hat{\beta}$ prevents the student pattern to have macroscopic alignment with a single example, i.e. $p = 0$. The saddle point equations (3.13-3.16) reduces to

$$m = \left\langle \hat{\xi} \langle \xi \rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad m_\tau = \left\langle \hat{\tau} \langle \tau \rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} \tag{3.23}$$

$$q = \left\langle \langle \xi \rangle^2_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad q_\tau = \left\langle \langle \tau \rangle^2_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} \tag{3.24}$$

$$d = \left\langle \langle \xi^2 \rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad d_\tau = \left\langle \langle \tau^2 \rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} . \tag{3.25}$$

We can further reduce the number of equations thanks to the Nishimori identities [32, 33, 115]. Infact, in the Bayes optimal setting it holds

$$\mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}^1\times\ldots\times\boldsymbol{\xi}^k} f(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^k) = \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}^2\times\ldots\times\boldsymbol{\xi}^k} f(\hat{\boldsymbol{\xi}},\boldsymbol{\xi}^2,\ldots,\boldsymbol{\xi}^k), \tag{3.26}$$

for any regular function f and with $(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^k)$ being k independent samples from the posterior distribution. In particular we can apply the identity (3.26) to the overlap obtaining that

$$\mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}^1\times\boldsymbol{\xi}^2} Q(\boldsymbol{\xi}^1,\boldsymbol{\xi}^2) = \mathbb{E}_{\hat{\boldsymbol{\xi}},\boldsymbol{S},\boldsymbol{\xi}} Q(\hat{\boldsymbol{\xi}},\boldsymbol{\xi}). \tag{3.27}$$

As a consequence, recalling Eqs. (3.19,3.21), it must be $m = q$ and therefore it is sufficient to solve

$$m = \left\langle \xi \langle \xi \rangle_{\xi|z,s} \right\rangle_{z,s} \qquad m_\tau = \left\langle \tau \langle \tau \rangle_{\tau|z,s} \right\rangle_{z,s} \tag{3.28}$$

$$d = \left\langle \langle \xi^2 \rangle_{\xi|z,s} \right\rangle_{z,s} \qquad d_\tau = \left\langle \langle \tau^2 \rangle_{\tau|z,s} \right\rangle_{z,s} , \tag{3.29}$$

where the averages are now recast in the form

$$P_{\Omega_\xi}(\xi) e^{-\frac{\alpha}{2}\beta\left(m_\tau - d_\tau + \langle \tau^2 \rangle_\tau\right)\xi^2 + \xi\sqrt{\alpha\beta m_\tau} z} , \tag{3.30}$$

$$P_{\Omega_\tau}(\tau) e^{-\frac{\beta}{2}(m-d)\tau^2 + \tau\sqrt{\beta m} z} . \tag{3.31}$$

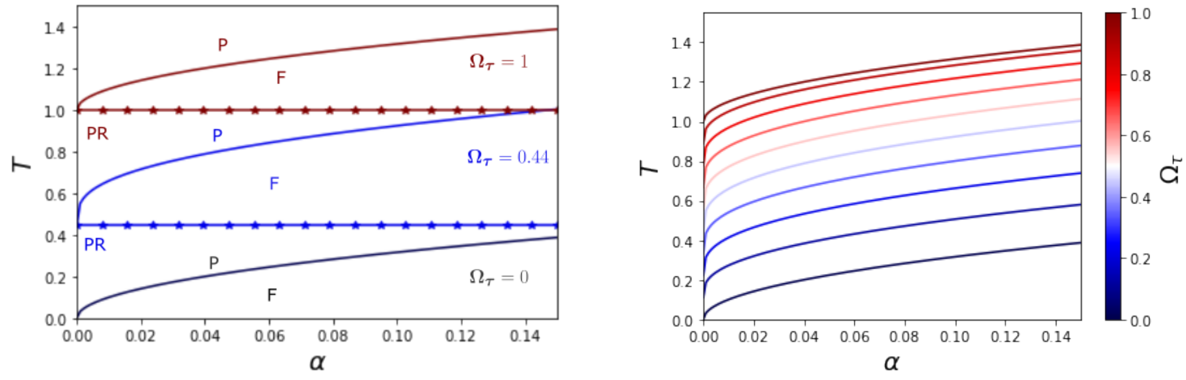


Figure 3.2: Retrieval phase transition lines in the case of the Bayesian optimal scenario. Below the transition curve it is possible to recover the planted signal $\hat{\boldsymbol{\xi}}$. **Left**: P-F transition line for $\Omega_\tau \in \{0, 0.44, 1\}$. For each choice of the τ prior the perfect retrieval (PF) region starts below the star marked line, i.e. $T < \Omega_\tau$. **Right**: Different P-F lines (without the perfect retrieval region) for all the possible values of Ω_τ. Above each colored line (taken singularly) the student is in a paramagnetic phase, below in a ferromagnetic regime.

In Fig. (3.2) the learning phase diagram in the (α, T) plane is presented as a function of $\mathbf{\Omega}$. One can observe the three aforementioned different regimes:

- the *Paramagnetic* (P) phase, where the order parameter $m = q$ vanishes;
- the *Ferromagnetic* (F) phase, where $m > 0$.
- the *Perfect Retrieval* (PR) phase, where $m = 1$.

In the ferromagnetic phase the student RBM is capable of learning (at least partially) the teacher pattern. Conversely this ability is lost in the paramagnetic phase. The ferromagnetic phase emerges even at high temperature once the dataset size reaches a critical threshold, i.e. $\alpha \geq \alpha_c(T, \mathbf{\Omega})$. In fact, at high generating temperature, each example provides only a vanishing amount of information about the teacher pattern. Therefore an extensively large number of examples are needed for the student to retrieve the signal. The corresponding P-F phase transition is thus the onset of learning. Descending from high temperatures, where $m = m_\tau = 0$, the P-F transition line can be obtained by expanding in $m \sim 0$. First of all, note that at $m = m_\tau = 0$ Eqs. (3.28, 3.29) give

$$d_\tau = \langle \tau^2 \rangle_0 = \mathbb{E}_\tau[\tau^2 e^{\frac{\beta}{2} d\,\tau^2}]/\mathbb{E}_\tau[e^{\frac{\beta}{2} d\,\tau^2}]\,, \qquad \tau \sim P_{\Omega_\tau} \tag{3.32}$$

$$d = \langle \xi^2 \rangle_0 = \mathbb{E}_\xi[\xi^2 e^{\frac{\alpha\beta}{2}(d_\tau - \langle \tau^2 \rangle_\tau)\,\xi^2}]/\mathbb{E}_\xi[e^{\frac{\alpha\beta}{2}(d_\tau - \langle \tau^2 \rangle_\tau)\,\xi^2}]\,, \qquad \xi \sim P_{\Omega_\xi} \tag{3.33}$$

where $\langle \cdot \rangle_0$ denotes the expectations w.r.t. the distributions of Eqs. (3.30) or (3.31) at zero effective field. Note that, since $d_\tau = \langle \tau^2 \rangle_\tau$ it is $e^{\frac{\alpha\beta}{2}(d_\tau - \langle \tau^2 \rangle_\tau)\,\xi^2} = 1$ and therefore $d = 1$. This is expected because, thanks to the Nishimori identities, it must be

$$\mathbb{E}_{\hat{\boldsymbol{\xi}}, \boldsymbol{S}, \boldsymbol{\xi}} f(\boldsymbol{\xi}) = \mathbb{E}_{\hat{\boldsymbol{\xi}}, \boldsymbol{S}, \boldsymbol{\xi}} f(\hat{\boldsymbol{\xi}}) = \mathbb{E}_{\hat{\boldsymbol{\xi}}} f(\hat{\boldsymbol{\xi}}). \tag{3.34}$$

Therefore the statistics of the student pattern matches those of the teacher's: in particular the self-overlap must be equal to that of the teacher pattern, being one by definition. By expanding the effective distributions in Eqs. (3.30, 3.31) as

$$P_{\Omega_\xi}(\xi) \left(1 + \sqrt{\alpha\beta m_\tau} z\, \xi + O(q_\tau)\right) \tag{3.35}$$

$$P_{\Omega_\tau}(\tau) \left(1 + \sqrt{\beta m} z\, \tau + O(q)\right) e^{\frac{\beta}{2}\tau^2} \tag{3.36}$$

one obtains

$$m \;=\; \left\langle \xi \langle \xi \rangle_{\xi|z,s} \right\rangle_{z,s} = \alpha\beta m_\tau + o(m_\tau) \tag{3.37}$$

$$m_\tau \;=\; \left\langle \tau \langle \tau \rangle_{\tau|z,s} \right\rangle_{z,s} = \beta m \langle \tau^2 \rangle_0^2 + o(m)\,. \tag{3.38}$$

As a consequence the transition line turns out to be

$$1 = \alpha\beta^2 d_\tau^2\,. \tag{3.39}$$

Eqs. (3.32) and (3.39) can be solved numerically obtaining all the P-F transition curves in Fig.(3.2). Note that the transition $T_c(\mathbf{\Omega}, \alpha)$ is only function of Ω_τ and that it is an increasing function of Ω_τ and α. In particular, in the low load limit, $\alpha \to 0$, the transition temperature is $T_c(\mathbf{\Omega}, 0) = \Omega_\tau$. At lower temperatures, specifically when $T \leq T_c(\mathbf{\Omega}, 0)$ the Perfect Retrieval regime appears. This is the region where the examples exhibit a finite overlap with the teacher pattern, then the student can perfectly retrieve the signal, i.e. $m = 1$, for any extensive dataset $\alpha > 0$.

3.4 Exploring arbitrary priors: Mismatched Setting

As in Sec.2.4, a more realistic scenario is when the student is not aware of the model underlying the data. This means that the inference temperature is different from the dataset noise, i.e. $\beta \neq \hat{\beta}$ and that the hyperparameters of the S-RBM do not match the ones of the T-RBM, i.e. $\Omega_\xi \neq \Omega_{\hat{\xi}}$ and $\Omega_\tau \neq \Omega_{\hat{\tau}}$. In this part, we focus on the last scenario, as the effect of various generating temperatures was studied previously. As in Chapter 2 the phase diagrams show four different regimes:

- the *Paramagnetic* (P) region, where $m = q = p = 0$;
- the *Signal Retrieval* (sR) region, where $m \neq 0$, $q > 0$, $p = 0$;
- the *Example retrieval* (eR) region, where $p \neq 0$, $q > 0$, $m = 0$;
- the *Spin Glass* (SG) region, where $m = p = 0$, $q > 0$.

In the eR phase the student is not capable of learning the teacher pattern because it is aligned with one example that shares vanishing overlap with the signal. In the SG phase, the student retrieves a spurious pattern unrelated to both the teacher and any examples. The sR phase is the only region where learning is possible, with phase transitions marking the boundaries between different learning regimes that depend on $\mathbf{\Omega}$.

Transition to the Spin Glass Phase

At very high temperature ($\beta \sim 0$) the distributions (3.17) and (3.18) have no external effective fields and therefore $\langle\xi\rangle = \langle\tau\rangle = 0$. In this regime the student is just making a random guess. All the relevant order parameters $p = m = q = 0$ together with their hidden counterparts $p_\tau = m_\tau = q_\tau = 0$, i.e. the system is in the paramagnetic phase. As in the Bayesian Optimal setting, lowering the temperature may start a spin glass transition, with nonzero overlaps q and q_τ; while the other parameterns remain zero. Assuming this transition is continuous, we can linearise Eqs. (3.15) for small q. Expanding first Eqs. (3.17) and (3.18) we get that ξ and τ in Eqs. (3.15) have to be averaged out over respectively

$$P_{\Omega_\xi}(\xi)\left(1+\sqrt{\alpha\beta q_\tau}\, z\xi + O(q_\tau)\right) e^{\frac{\alpha}{2}\beta(d_\tau - \langle\tau^2\rangle_\tau)\xi^2} \tag{3.40}$$

$$P_{\Omega_\tau}(\tau)\left(1+\sqrt{\beta q}\, z\tau + O(q)\right) e^{\frac{\beta}{2} d\tau^2}\,. \tag{3.41}$$

First of all it is

$$d_\tau \;=\; \langle\tau^2\rangle_0 = \mathbb{E}_\tau[\tau^2 e^{\frac{\beta}{2}d\tau^2}]/\mathbb{E}_\tau[e^{\frac{\beta}{2}d\tau^2}], \qquad \tau \sim P_{\Omega_\tau} \tag{3.42}$$

$$d \;=\; \langle\xi^2\rangle_0 = \mathbb{E}_\xi[\xi^2 e^{\frac{\alpha\beta}{2}(d_\tau - \langle\tau^2\rangle_\tau)\xi^2}]/\mathbb{E}_\xi[e^{\frac{\alpha}{2}\beta(d_\tau - \langle\tau^2\rangle_\tau)\xi^2}], \quad \xi \sim P_{\Omega_\xi}\,, \tag{3.43}$$

where $\langle\cdot\rangle_0$ denotes the expectations w.r.t. the distributions of Eqs. (3.17) or (3.18) at zero effective field (i.e. $m = m_\tau = q = q_\tau = p = p_\tau = 0$). Therefore also in this case it is $d = 1$ at the transition. Moreover we get

$$q \;=\; \left\langle \langle\xi\rangle^2_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} = \alpha\beta q_\tau \langle\xi^2\rangle_0^2 + o(q_\tau) = \alpha\beta q_\tau + o(q_\tau) \tag{3.44}$$

$$q_\tau \;=\; \left\langle \langle\tau\rangle^2_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} = \beta q \langle\tau^2\rangle_0^2 + o(q) = \beta q d_\tau^2 + o(q), \tag{3.45}$$

From Eqs. (3.44-3.45) a second order phase transition to the spin glass phase may occur at

$$1 = \alpha\beta^2 d_\tau^2 \tag{3.46}$$

that can be solved to obtain the critical temperature $T_c(\alpha)$ or equivalently the critical size $\alpha_c(T)$. Note that there is no dependence of $\hat{\beta}$, meaning that the spin glass transition only depends on a (too low) inference temperature of the student. As expected [65, 114] the P-SG transition (3.46) corresponds to the P-F transition in the Nishimori regime (3.39). Moreover it also corresponds to the P-SG transition of a generalized Hopfield model with random patterns at inverse temperature β and load α. Indeed, the expression (3.46) matches the one found in [64]. For example fixing $\Omega_\xi = \Omega_\tau = 0$ one gets the bipartite SK P-SG transition $T_c(\alpha) = \sqrt{\alpha}$, while using $\Omega_\xi = 0$ and $\Omega_\tau = 1$ it is possible to recover the P-SG transition line of the Hopfield model $T_c(\alpha) = 1 + \sqrt{\alpha}$.

Transition to the Retrieval Phase

When the size of the dataset is sufficiently large, lowering the temperature may lead to a continuous transition towards a signal retrieval (sR) region. The transition line can be obtained by expanding Eq. (3.14) for small m and m_τ, keeping $p = p_\tau = q = q_\tau = 0$. As previously done, first of all Eqs. (3.17-3.18) can be expanded as

$$P_{\Omega_\xi}(\xi)\left(1 + \alpha\sqrt{\beta\hat{\beta}}m_\tau\hat{\xi}\xi + O(m_\tau)\right)e^{\frac{\alpha}{2}\beta(\,d_\tau - \langle\tau^2\rangle_\tau\,)\xi^2} \tag{3.47}$$

$$P_{\Omega_\tau}(\tau)\left(1 + \sqrt{\beta\hat{\beta}}m\tau\hat{\tau} + O(m)\right)e^{\frac{\beta}{2}d\tau^2}, \tag{3.48}$$

from which it holds

$$m \;=\; \left\langle\hat{\xi}\langle\xi\rangle_{\xi|z,s,\hat{\xi}}\right\rangle_{z,s,\hat{\xi}} = \alpha\sqrt{\hat{\beta}\beta}m_\tau\langle\hat{\xi}^2\rangle_{\hat{\xi}}\langle\xi^2\rangle_0 + o(m_\tau) = \alpha\sqrt{\hat{\beta}\beta}\,m_\tau\,d\,\langle\hat{\xi}\rangle_{\hat{\xi}} + \mathcal{O}(m_\tau) \tag{3.49}$$

$$m_\tau \;=\; \left\langle\hat{\tau}\langle\tau\rangle_{\tau|z,s,\hat{\tau}}\right\rangle_{z,s,\hat{\tau}} = \sqrt{\hat{\beta}\beta}m\langle\hat{\tau}^2\rangle_{\hat{\tau}}\langle\tau^2\rangle_0 + o(m) = \sqrt{\hat{\beta}\beta}\,m\,d_\tau\langle\hat{\tau}^2\rangle_{\hat{\tau}} + \mathcal{O}(m), \tag{3.50}$$

where again $\langle\cdot\rangle_0$ denotes the expectations w.r.t. the distributions of Eqs. (3.17) or (3.18) at zero effective field. In Eq.(3.49) we can use $d_{\hat{\xi}} := \langle\hat{\xi}^2\rangle_{\hat{\xi}} = 1$ which derives from the analysis of the teacher partition function (C.34). Again the self-overlaps d and d_τ are solutions of Eq.(3.42-3.43), which implies $d = 1$. Recalling that $\langle.\rangle_{\hat{\tau}}$ is the average over the tilted distribution $P_{\Omega_{\hat{\tau}}}(\hat{\tau})e^{\frac{\hat{\beta}}{2}\hat{\tau}^2}$, it holds

$$d_{\hat{\tau}} \;:=\; \langle\hat{\tau}^2\rangle_{\hat{\tau}} = \frac{1 - \hat{\beta}\Omega_{\hat{\tau}}^2}{(1 - \hat{\beta}\Omega_{\hat{\tau}})^2}\,, \tag{3.51}$$

thus, from Eqs. (3.49-3.50), a second order phase transition to the sR phase may occur at

$$1 = \alpha\hat{\beta}\beta d_\tau d_{\hat{\tau}}\,. \tag{3.52}$$

For example when both the T-RBM and S-RBM have only binary units the critical temperature reads as $T_c(\alpha) = \alpha\hat{\beta}$ while in the case of the Dual Hopfield model, i.e. $\Omega_\tau = \Omega_{\hat{\tau}} = 1$, $T_c(\alpha) = 1 + \alpha\hat{\beta}$. Since $d = 1$ and d_τ doesn't depend of α, the P-sR critical temperature $T_c(\alpha)$ grows linearly with the dataset's size α. Conversely, from Eq. (3.46), the critical temperature of the P-SG transition grows only as $O(\sqrt{\alpha})$. This means that for larger values of α the P-sR phase transition occurs first. On the contrary for small α the spin glass phase emerges at a higher temperature.

A triple point exists when the two transition lines (3.46) and (3.52) cross, i.e. at the solution of

$$\begin{cases} 1 = \alpha\beta^2 d_\tau^2 \\ 1 = \alpha\beta\hat{\beta}d_\tau d_{\hat{\tau}}, \end{cases} \tag{3.53}$$

giving

$$P_{triple} = (\alpha_c, T_c) = \left(\frac{1}{\hat{\beta}^2 d_{\hat{\tau}}^2} \, , \, \frac{d_\tau}{\hat{\beta} d_{\hat{\tau}}} \right) . \tag{3.54}$$

As expected the Nishimori line $\beta = \hat{\beta}$ crosses the triple point when the student and teacher priors match, i.e. $d_\tau = d_{\hat{\tau}}$. It is interesting to note that the α of the triple point only depends on the properties of the data, i.e. of the teacher, while its temperature may also depend on the student's priors. The critical temperature at the triple point, T_c, is shown in Fig. 3.3 to be an increasing function of Ω_τ. Explicitly, T_c increases with Ω_τ, but decreases with $\Omega_{\hat{\tau}}$. Conversely, the critical value α_c decreases as both $\hat{T}$ and $\Omega_{\hat{\tau}}$ increase.

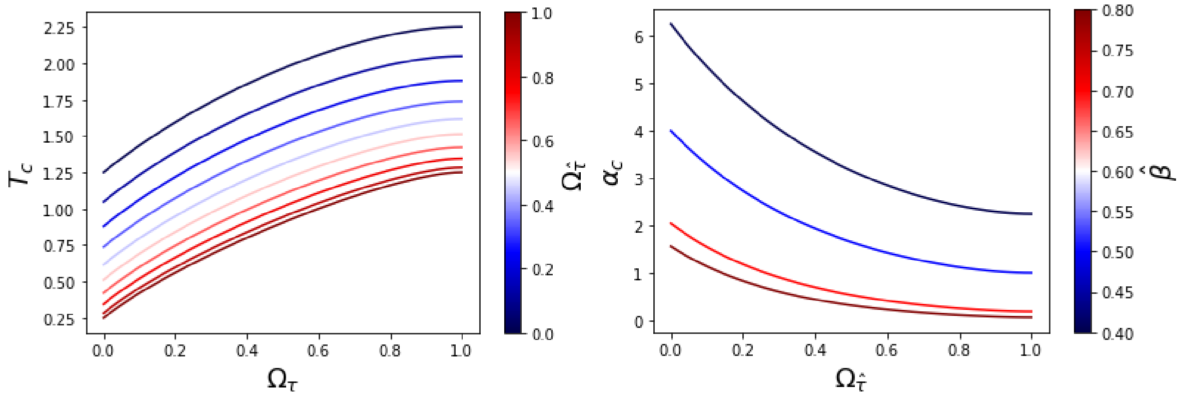


Figure 3.3: Triple point's coordinates as a function of the prior parameters. **Left**: Critical temperature for different choices of the the hidden units prior of both S-RBM and T-RBM. T_c is an increasing funciton of Ω_τ, but decreases with $\Omega_{\hat{\tau}}$. **Right**: Critical size as function of the relevant teacher variables. It is a decreasing function of both $\Omega_{\hat{\tau}}$ and $\hat{T} = 1/\hat{\beta}$.

These properties can be observed from the phase diagrams obtained by different choices of the priors, which are presented in Figs. (3.4-3.8). All the depicted transition lines from the SG region to the sR or the eR regions can be found numerically by solving the saddle point Eqs. (3.13-3.16).

In all the cases the triple point appears to be the point of the sR region with smaller α. In other words the α of the triple point corresponds to the minimum size of the dataset for which learning is possible, at least for suitable choice of the inference temperature. For this reason we refer to it as the model's α_c. For $\alpha \geq \alpha_c$ there exists an optimal inference temperature where the learning performance, i.e. the overlap with the teacher pattern, is maximal. Notably if the inference temperature is set too low one can eventually exit the sR region and enters a glassy phase. However, for sufficiently large values of α, the sR region extends down to zero temperature.

Finally, the example retrieval region may emerge at low inference temperature when α is very close to zero. This phase corresponds to the pattern retrieval region in the generalized Hopfield models [64] under the critical capacity. In this region the student pattern tends directly to be aligned with one of the examples instead of learning their archetype.

The role of the S-RBM priors

By fixing the properties of the T-RBM, i.e. of the generating process, it is possible to explore the effects of the S-RBM priors on the learning performance. In the following discussion the examples are assumed to have binary entries, i.e. $\Omega_s = 0$. In the following section, as well as in the subsequent sections, the phase diagrams presented illustrate the existing phases.

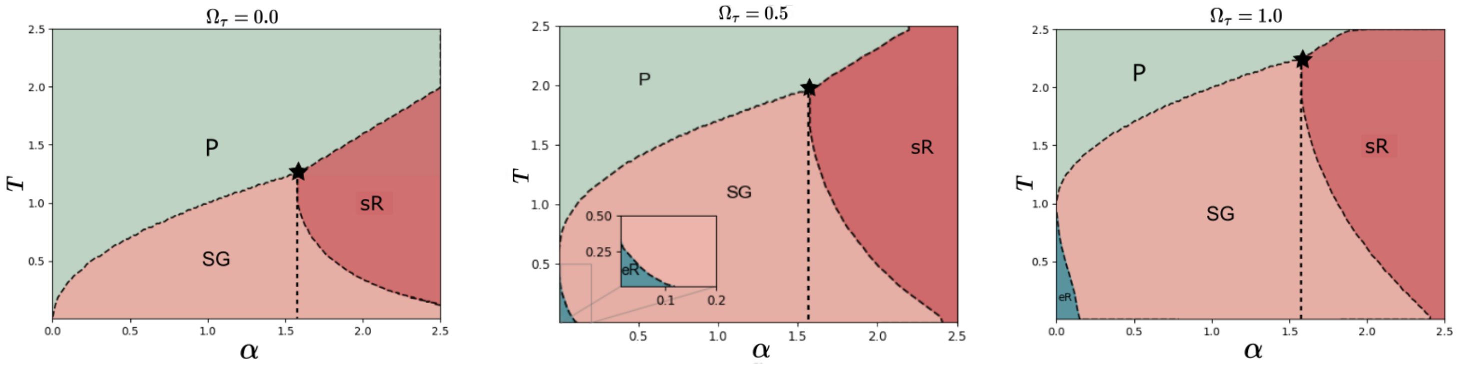


Figure 3.4: Phase diagram of different S-RBM configurations, each with a different choice of hidden unit prior Ω_τ. All Students have a binary prior for ξ, i.e., $\Omega_\xi = 0$. The teacher is generating ($\hat{\beta} = 0.8$) binary data ($\Omega_s = 0$), and its architecture is fixed by the choice $\Omega_{\hat{\xi}} = \Omega_{\hat{\tau}} = 0$. The black star represents the critical point (α_c, T_c), while the vertical dashed line shows that the position $\alpha_c \simeq 1.55$ is the same for all the images. The eR phase emerges when the τ prior has a gaussian tail.

In Fig. (3.4) the phase diagrams of different S-RBMs with binary patterns ($\Omega_\xi = 0$) are shown in the case of a binary T-RBM, i.e. $\Omega_{\hat{\xi}} = \Omega_{\hat{\tau}} = 0$. One can observe the appearance of the eR phase at low temperature as soon as the student hidden unit starts to have a Gaussian tail, i.e. $\Omega_\tau > 0$. The critical temperature at $\alpha = 0$ is exactly Ω_τ. In the limit $\Omega_\tau \to 1$ this phase extends up to the critical capacity of the Hopfield model $\alpha = 0.14$. As expected the critical size α_c of the triple point doesn't change with the student priors while its temperature increases with Ω_τ. As a consequence the area of the *sR* phase expands in the direction of higher temperatures. This is typically an advantage for MC algorithms that samples from the posterior starting from high temperature like simulated annealing [116, 117].

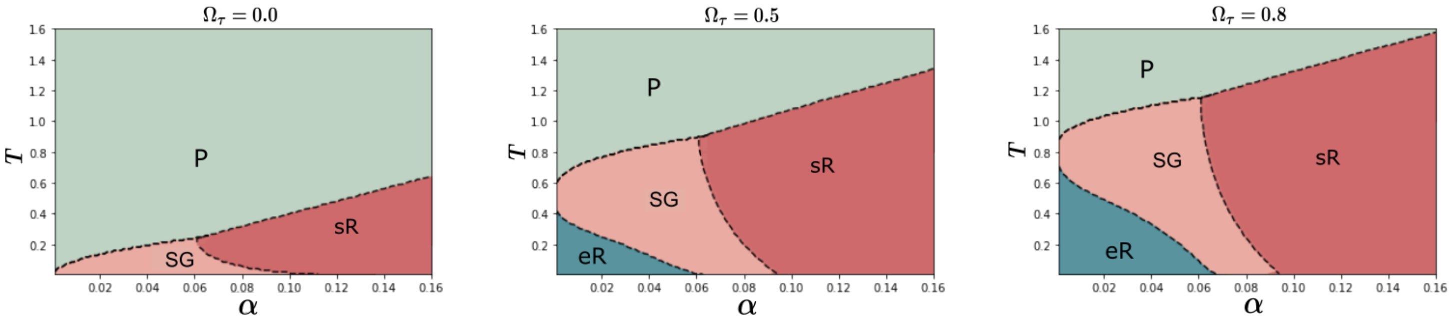


Figure 3.5: Phase diagrams showing the effect of student's hidden unit prior when the teacher generates ($\hat{\beta} = 0.8$) the dataset ($\Omega_s = 0$) using a Hopfield model. The student pattern is chosen with $\Omega_\xi = 0$. The critical size needed to enter the inference regime is unaltered by the choice of Ω_τ and its value is $\alpha_c \simeq 0.06$, while the temperature T_c increases.

In Fig. (3.5), we observe similar behavior for a dataset generated by a Hopfield like T-RBM, i.e. $\Omega_{\hat{\tau}} = 1$, $\Omega_{\hat{\xi}} = 0$. The different generation procedures significantly affect the number of data points required for the student to transition into the sR regime: as prescribed by Eq.(3.54), α_c is smaller than that of Fig.(3.4). One can in principle investigate the effect of changing Ω_ξ, however we have verified that the phase diagram is not affected by that prior since it only enters in the self-overlap, and $d = 1$ is always solution of Eqs. (3.13-3.16). Interestingly, the benefit of using Gaussian units—where at any given temperature, the sR region occurs at smaller α—remains consistent regardless of the dataset properties. Specifically, it is not always advantageous for the student to align their hyperparameters with those of the teacher.

The role of the T-RBM priors

By fixing the S-RBM architecture, one can instead investigate the impact of the data structure, specifically the T-RBM priors, on learning performance. For now, it is assumed that the examples consist of binary entries. From Eq. (3.54), an effect on both the minimum amount of data required to enter the sR phase and the retrieval temperature is expected.

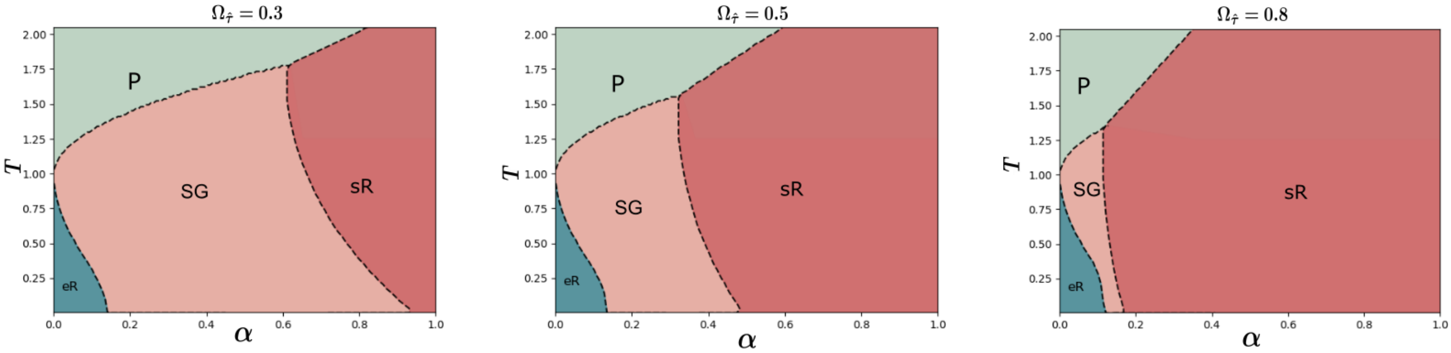


Figure 3.6: Effect of the teacher architecture ($\hat\beta = 0.8$) on the student's phase diagram. The signal's prior is defined by $\Omega_{\hat\xi} = 0$. The analyzed student is working with a Hopfield architecture: $\Omega_\xi = 0, \Omega_\tau = 1$. The more $\hat\tau$ is gaussian the more easily the student retrieves the pattern $\hat{\boldsymbol{\xi}}$, in particular P_{triple} decreases in both its coordinates.

Fig. (3.6) shows the effect of changing $\Omega_{\hat\tau}$. It is possible to appreciate the fact that the information present in the data increases notably the more the $\hat\tau$ prior becomes Gaussian. This can be seen by the reduced amount of generated samples we need to enter the sR regime (a similar effect of increasing $\hat\beta$ in Chapter 2). On the other hand, the critical temperature of the student at the triple point, T_c, is lowered. Nevertheless, if we fix an inference temperature, the student enters the sR regime more easily as $\Omega_{\hat\tau} \to 1$.

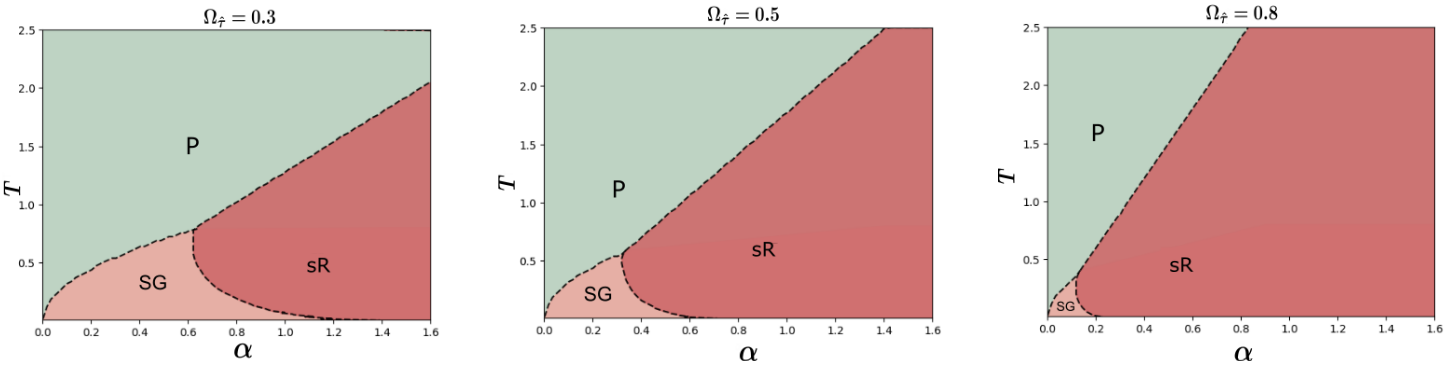


Figure 3.7: Effect of the teacher ($\hat\beta = 0.8$) architecture on a binary student's phase diagram. Since $\Omega_\tau = 0$ we do not have example memorization. As in Fig.(3.5) when $\Omega_{\hat\tau} \to 1$ the S-RBM retrieval region becomes larger.

Same effect is shown in Fig. (3.7) where the student priors are binary and the teacher prior changes. We checked that the phase diagrams are not affected by $\hat{\boldsymbol{\xi}}$, as expected from the transition line expressions.

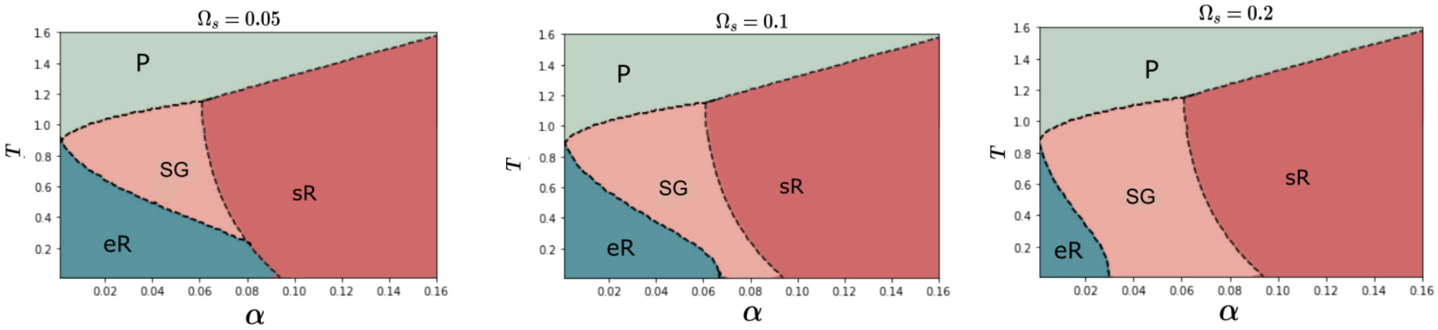


Figure 3.8: Phase diagram of the teacher-student problem ($\hat{\beta} = 0.8$) with $\mathbf{\Omega} = \{0, 1, 0, 0.9, \Omega_s\}$, $\Omega_s \in \{0.005, 0.1, 0.2\}$. The effect of the manipulation of the data affects only the eR phase. The more the data have gaussian nature, the more the example memorization region is compressed.

Finally in Fig. (3.8) we explore the role of the dataset unit prior Ω_s. The different values of Ω_s do not change the triple point P_c. As one can observe the only effect is the shift of the eR-SG transition line. This is related to the *memorization* capability of the S-RBM and is in agreement with the phase diagrams in Fig.(1.6).

Spherical regularization

In the previous sections we have seen that the term $z(\boldsymbol{\xi})^{-M}$ in the posterior plays the role of a regularization enforcing the Nishimori identity $d = 1$. Without that regularization the model would be ill defined in the case of an architecture with both Ω_ξ and Ω_τ different from zero. This is a typical problem with fully Gaussian disordered systems at low temperature [60, 63, 64, 77].

From a practical point of view is more convenient to replace that term directly with a spherical constraint $\delta\left(N - \sum_{i=1}^{N} \xi_i^2\right)$. The computation leading to the set of saddle point Eqs. (3.13-3.16) can be exactly followed with the only addition of a lagrange multiplier ω in the gaussian weight for the ξ variable, i.e. its effective distribution reads as

$$P_{\Omega_\xi}(\xi) e^{-\frac{\alpha\beta}{2}(\alpha\beta q_\tau - d_\tau + \omega)\xi^2 + \xi\left(\alpha\sqrt{\hat{\beta}}\beta m_\tau \hat{\xi} + \sqrt{\alpha\beta q_\tau} z + \beta\Omega_\tau p s\right)}, \tag{3.55}$$

where ω has to be fixed to obtain $d = 1$. Making a comparison with Eq.(3.17) one can see that the correct choice is $\omega = \langle\tau^2\rangle_\tau$. From the computational point of view it is more convenient to adjust the Lagrange multiplier during the solution of the saddle point equations. In Fig.(3.9) it is shown the phase diagram where the priors of both the weights and the student hidden variables have Gaussian tails. The weights have been regularized with a spherical constraint. One can compare these results with the right panel of Fig.(3.4), where $\Omega_\xi = 0$. It is possible to see the presence of the *eR* phase, due to $\Omega_\tau = 1$, that tends to quickly disappear as $\Omega_\xi \to 1$, in agreement with [64]. It is also observed a bending of the *sR* region at low temperature that however is not particularly significant since that is the region where RSB corrections are expected. Finally, as expected, the triple point position doesn't change.

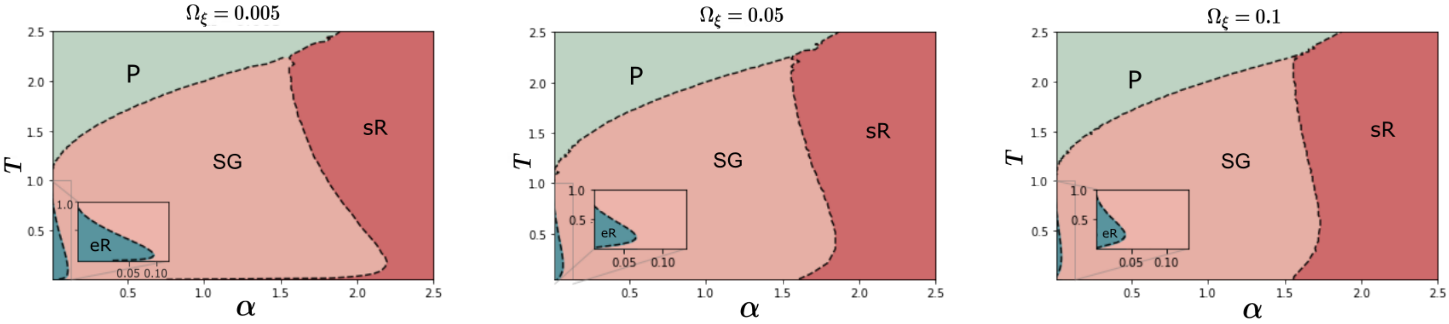


Figure 3.9: Spherical regularization effect on the retrieval capability of the model (case with $\Omega_\tau = 1$). Using binary data from a binary T-RBM ($\hat{\beta} = 0.8$), it is possible to see that α_c and T_c do not move. The only effect from the changes of Ω_ξ is the compression of the eR phase.

3.5 Discussion

We investigate the learning performance of Restricted Boltzmann Machines (RBMs) trained on a noisy dataset in a teacher-student setting. We considered RBMs with different unit and weight priors and analyzed the generalization capability in both the Bayes optimal scenario and in a more realistic mismatched case. We proved the existence of a well-defined critical dataset size below which learning is impossible, characterizing a triple point in the phase diagram where paramagnetic, spin glass, and signal retrieval phases intersect.

We show that Gaussian hidden units help the machine to more easily enter the signal retrieval region, even though the critical size only depends on the properties of the dataset, i.e., the teacher RBM. At the same time, certain choices of priors favor the emergence of an example retrieval region where the machine learns by memorization without generalizing well. Notably, the posterior distribution approach for training includes automatic self-regularization of the weights, preventing parameter explosions and other typical convergence issues.

Appendix C

Derivation of the RS equations

In this section we discuss the computation of the averaged partition function (3.5) in the mismatch case with one single student pattern. The posterior of the student's (single) weight is

$$P(\boldsymbol{\xi}|\boldsymbol{S}) = Z^{-1}(\boldsymbol{S}) z^{-M}(\boldsymbol{\xi}) P(\boldsymbol{\xi}) \mathbb{E}_{\boldsymbol{\tau}} \exp\left(\sqrt{\frac{\beta}{N}} \sum_{i=1}^{N} \sum_{a=1}^{M} \xi_i s_i^a \tau^a\right) \tag{C.1}$$

then its partition function $Z(\boldsymbol{S})$ can be written in the following form

$$\begin{aligned} Z(\boldsymbol{S}) &= \mathbb{E}_{\boldsymbol{\xi}} z^{-M}(\boldsymbol{\xi}) \mathbb{E}_{\boldsymbol{\tau}} \exp\left(\sqrt{\frac{\beta}{N}} \sum_{i=1}^{N} \sum_{a=1}^{M} \xi_i s_i^a \tau^a\right), & \text{(C.2)} \\ z(\boldsymbol{\xi})^M &= \sum_{\boldsymbol{S}} \mathbb{E}_{\boldsymbol{\tau}} \exp\left(\sqrt{\frac{\beta}{N}} \sum_{i=1}^{N} \sum_{a=1}^{M} s_i^a \tau^a \xi_i\right). & \text{(C.3)} \end{aligned}$$

Differently from the teacher, which generates the data at inverse temperature $\hat{\beta} < 1$, the student's temperature range is $T \in [0, \infty]$. Due to this possibility, when $T < 1$, $\boldsymbol{\xi}$ has the chance to be aligned with one of the data, say the first one $\boldsymbol{s}^{(1)}$. The alignment can be in principle valid also for the extensive hidden unit vector $\boldsymbol{\tau}$. It is measured with the two Mattis magnetizations

$$\begin{aligned} p &= \frac{1}{N} \sum_{i=1}^{N} \xi_i s_i^{(1)}, & \text{(C.4)} \\ p_\tau &= \frac{1}{M} \sum_{a=1}^{M} \tau^a s_{(1)}^a, & \text{(C.5)} \end{aligned}$$

which allow to rewrite the previous partition function (discarding irrelevant non extensive terms) as

$$Z(\boldsymbol{S}) = \mathbb{E}_{\boldsymbol{\xi}} z^{-M}(\boldsymbol{\xi}) \mathbb{E}_{\boldsymbol{\tau}} \exp\left(M\sqrt{\frac{\beta}{N}} p_\tau(\tau)\, \xi_{(1)} + N\sqrt{\frac{\beta}{N}} p(\xi)\, \tau^{(1)} + \sqrt{\frac{\beta}{N}} \sum_{i=2}^{N} \sum_{a=2}^{M} s_i^a \tau^a \xi_i\right).$$

In spin-glass models with planted disorder we introduce n independent replicas, then take the limit $N \to \infty$

$$Z^n(\boldsymbol{S}) = \mathbb{E}_{\{\boldsymbol{\xi}^\gamma\}} z(\{\boldsymbol{\xi}^\gamma\})^{-M} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \exp\left(M\sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} p_\tau^\gamma(\tau) \xi_{(1)}^\gamma + N\sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} p^\gamma(\xi) \tau_{(1)}^\gamma + \sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} \sum_{i=2}^{N} \sum_{a=2}^{M} s_i^a \tau^{a,\gamma} \xi_i^\gamma\right). \tag{C.6}$$

The shorthand notation $\{\boldsymbol{\xi}^\gamma\}$ stands for the average over all the replicas indexed by γ: $\{\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^n\}$, same is done for $\{\boldsymbol{\tau}^\gamma\}$. Then the averaged partition function over the quenched disorder induced by the dataset is

$$[Z^n]^{\boldsymbol{\mathcal{S}}} = \sum_{\boldsymbol{\mathcal{S}}} P_{\hat{\beta}}(\boldsymbol{\mathcal{S}})\, Z^n(\boldsymbol{\mathcal{S}}) = \sum_{\boldsymbol{\mathcal{S}}} \mathbb{E}_{\hat{\boldsymbol{\xi}}} \mathbb{E}_{\hat{\boldsymbol{\tau}}}\, z(\hat{\boldsymbol{\xi}})^{-M}\, e^{\sqrt{\hat{\beta}/N} \sum_{ia} s_i^a \hat{\xi}_i \hat{\tau}_a}\, Z^n(\boldsymbol{\mathcal{S}})\,, \tag{C.7}$$

$$z(\hat{\boldsymbol{\xi}})^M = \sum_{\boldsymbol{\mathcal{S}}} \mathbb{E}_{\hat{\boldsymbol{\tau}}} \exp\left(\sqrt{\frac{\hat{\beta}}{N}} \sum_{i=1}^N \sum_{a=1}^M s_i^a \hat{\tau}^a \hat{\xi}_i\right). \tag{C.8}$$

The presence of both $z(\hat{\boldsymbol{\xi}})$ and $z(\boldsymbol{\xi})$ make it diffucukt to compute $[Z^n]^{\boldsymbol{\mathcal{S}}}$ as a straightforward extremization problem. A possible solution is the application of the saddle point method to both (C.8) and (C.3). We can write them as an exponential function to be extremized over a set of parameters $\boldsymbol{p}_T$ and $\boldsymbol{p}_S$. Each set will depend respectively on $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$

$$z(\hat{\boldsymbol{\xi}})^M \approx e^{-N\, \mathrm{Extr}_{\boldsymbol{p}_T}\, \hat{\beta} f_T(\boldsymbol{p}_T(\hat{\boldsymbol{\xi}}))}\,, \qquad z(\boldsymbol{\xi})^M \approx e^{-N\, \mathrm{Extr}_{\boldsymbol{p}_S}\, \beta f_S(\boldsymbol{p}_S(\boldsymbol{\xi}))}\,, \tag{C.9}$$

which, after the extremization, could be inserted in the exponential of (C.7).

Teacher normalization

In order to evaluate Eq.(C.8) as a variational principle we start computing $\sum_{\boldsymbol{\mathcal{S}}}$ writing each $\boldsymbol{s}^a \in \boldsymbol{\mathcal{S}}$ in the interpolating form $\sqrt{\Omega_s}\boldsymbol{g}_s^a + \sqrt{\delta_s}\boldsymbol{\epsilon}_s^a$ for each $a = 1, \cdots, M$, obtaining

$$z(\hat{\boldsymbol{\xi}})^M = \int \prod_a [\mathcal{D}\boldsymbol{g}^a]\, \mathbb{E}_{\hat{\boldsymbol{\tau}}} \exp\left\{\sqrt{\frac{\hat{\beta}}{N}} \sum_{i=1}^N \sum_{a=1}^M \sqrt{\Omega_s}\, g_{s,i}^a \hat{\tau}^a \hat{\xi}_i\right\} \prod_{i,a} 2\cosh\left\{\sqrt{\frac{\hat{\beta}}{N}} \sqrt{\delta_s} \hat{\tau}^a \hat{\xi}_i\right\} \tag{C.10}$$

$$\approx \mathbb{E}_{\hat{\boldsymbol{\tau}}} \exp\left\{\frac{\hat{\beta}}{2N} \sum_{i=1}^N \sum_{a=1}^M \left(\hat{\tau}_a \hat{\xi}_i\right)^2\right\}, \tag{C.11}$$

where, in the first passage we already average over the set $\{\boldsymbol{\epsilon}_s^a\}_{a=1,\cdots,M}$, while in the last one, we exploit $\log\cosh(x) \approx \frac{x^2}{2}$ and $\Omega_s + \delta_s = 1$. It is possible to use the Fourier representation of the Dirac delta function to enforce the relation

$$\delta\left(d_{\hat{\tau}} - \frac{1}{M}\sum_a \hat{\tau}_a^2\right) = \frac{1}{2\pi}\int d\hat{d}_{\hat{\tau}}\, e^{-M\hat{d}_{\hat{\tau}}\left(d_{\hat{\tau}} - \frac{1}{M}\sum_a \hat{\tau}_a^2\right)}\,, \tag{C.12}$$

where we refer at $d_{\hat{\tau}}$ as the self-overlap for the teacher hidden variables. Using the following expression for the density of states of the $\hat{\tau}$ configurations

$$\mathcal{D}(\hat{d}_{\hat{\tau}}, d_{\hat{\tau}}) = \sum_{\hat{\boldsymbol{\tau}}} \delta\left(d_{\hat{\tau}} - \tfrac{1}{M}\sum_a \hat{\tau}_a^2\right) \propto \sum_{\hat{\boldsymbol{\tau}}} \int d\hat{d}_{\hat{\tau}}\, e^{-M\hat{d}_{\hat{\tau}}\left(d_{\hat{\tau}} - \frac{1}{M}\sum_a \hat{\tau}_a^2\right)}\,, \tag{C.13}$$

we can rewrite (C.11)

$$z(\hat{\boldsymbol{\xi}})^M = \int d d_{\hat{\tau}}\, \mathcal{D}(\hat{d}_{\hat{\tau}}, d_{\hat{\tau}}) \exp\left\{\frac{\alpha\hat{\beta}}{2} d_{\hat{\tau}} \sum_i \hat{\xi}_i^2\right\} \tag{C.14}$$

$$\approx e^{-N\hat{\beta}\, \mathrm{Extr}_{\boldsymbol{p}_T}\, f(\boldsymbol{p}_T)}\,, \tag{C.15}$$

which is nothing but the variational principle in (C.9) with $\boldsymbol{p}_T = \{\hat{d}_{\hat{\tau}}, d_{\hat{\tau}}\}$. Using $d_{\hat{\xi}}(\hat{\boldsymbol{\xi}}) = \frac{1}{N}\sum_i \hat{\xi}_i^2$ as well, for the self overlap of the teacher pattern, the function to be extremized in (C.15) is

$$-\hat{\beta} f(\boldsymbol{p}_T) = \alpha \left(-d_{\hat{\tau}}\hat{d}_{\hat{\tau}} + \frac{\hat{\beta}}{2} d_{\hat{\xi}}(\hat{\boldsymbol{\xi}})\hat{d}_{\hat{\tau}} + \log \mathbb{E}_{\hat{\tau}} e^{\hat{d}_{\hat{\tau}}\hat{\tau}^2} \right) \tag{C.16}$$

reaching its minimum at

$$\hat{d}_{\hat{\tau}} = \frac{\hat{\beta}}{2} d_{\hat{\xi}}(\hat{\boldsymbol{\xi}}) \,, \qquad\qquad d_{\hat{\tau}} = \langle \hat{\tau}^2 \rangle_{\hat{\tau}} = \frac{\mathbb{E}_{\hat{\tau}} \hat{\tau}^2 e^{\frac{\hat{\beta}}{2} d_{\hat{\xi}}(\hat{\boldsymbol{\xi}})\hat{\tau}^2}}{\mathbb{E}_{\hat{\tau}} e^{\frac{\hat{\beta}}{2} d_{\hat{\xi}}(\hat{\boldsymbol{\xi}})\hat{\tau}^2}} \,. \tag{C.17}$$

In view of the above, the averaged partition function (C.7) can be approximated as

$$[Z^n]^{\boldsymbol{\mathcal{S}}} \;=\; \sum_{\boldsymbol{\mathcal{S}}} \mathbb{E}_{\hat{\boldsymbol{\xi}}} \mathbb{E}_{\hat{\boldsymbol{\tau}}} \, e^{N\hat{\beta}\,\mathrm{Extr}_{\boldsymbol{p}_T} f(\boldsymbol{p}_T)} \, e^{\sqrt{\hat{\beta}/N}\sum_{ia} s_i^a \hat{\xi}_i \hat{\tau}_a} \; Z^n(\boldsymbol{\mathcal{S}}) \tag{C.18}$$

Student normalization

The same problem of the teacher normalization term affects Eq.(C.6), due to the presence of $z(\{\boldsymbol{\xi}^\gamma\})$. As already stated in Eqs.(C.9) we want to write it in a variational principle manner. The steps to do that are the same of the one used for (C.10-C.17). We start computing the sum over the examples using their interpolated form, then we impose the self overlap with the Dirac delta

$$\begin{aligned} z(\{\boldsymbol{\xi}^\gamma\})^M \;&=\; \int \prod_{a,\gamma} [\mathcal{D}\boldsymbol{g}^{a\gamma}] \; \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \exp\left\{ \sqrt{\frac{\beta}{N}} \sum_{\gamma,i,a} \sqrt{\Omega_s}\, g_{s,i}^{a\gamma} \tau_a^\gamma \xi_i^\gamma \right\} \prod_{i,a,\gamma} 2\cosh\left\{ \sqrt{\frac{\beta}{N}} \sqrt{\delta_s} \tau_a^\gamma \xi_i^\gamma \right\} && \text{(C.19)} \\ &\approx\; \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \; \exp\left\{ \frac{\beta}{2N} \sum_{\gamma=1}^{n} \sum_{i=1}^{N} \sum_{a=1}^{M} (\tau_a^\gamma \xi_i^\gamma)^2 \right\} \;, && \text{(C.20)} \\ &\approx\; \int \prod_\gamma dd_\tau^\gamma \, \mathcal{D}(\hat{\boldsymbol{d}}_{\boldsymbol{\tau}}, \boldsymbol{d}_{\boldsymbol{\tau}}) \exp\left\{ \frac{\alpha\beta}{2} \sum_\gamma d_\tau^\gamma d^\gamma(\boldsymbol{\xi}) \right\} && \text{(C.21)} \\ &\approx\; e^{-N\beta\,\mathrm{Extr}_{\boldsymbol{p}_S} f(\boldsymbol{p}_S)} \,, && \text{(C.22)} \end{aligned}$$

with density of states and self overlap for every student replica as

$$\mathcal{D}(\hat{\boldsymbol{d}}_{\boldsymbol{\tau}}, \boldsymbol{d}_{\boldsymbol{\tau}}) \;=\; \sum_{\{\boldsymbol{\tau}^\gamma\}} \delta\left(\sum_\gamma d_\tau^\gamma - \frac{1}{M} \sum_{a,\gamma} (\tau_a^\gamma)^2 \right) \propto \sum_{\{\boldsymbol{\tau}^\gamma\}} \int \prod_\gamma d\hat{d}_\tau^\gamma \, e^{-M\sum_\gamma \hat{d}_\tau^\gamma \left(d_\tau^\gamma - \frac{1}{M}\sum_a (\tau_a^\gamma)^2 \right)} \,,$$

$$d^\gamma(\boldsymbol{\xi}) = \frac{1}{N} \sum_i (\xi_i^\gamma)^2 \,.$$

Finally we can rewrite Eq.(C.6) as

$$Z^n(\boldsymbol{\mathcal{S}}) \;=\; \mathbb{E}_{\{\boldsymbol{\xi}^\gamma\}} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \exp\left(M\sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} p_\tau^\gamma(\tau) \xi_{(1)}^\gamma + N\sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} p^\gamma(\xi) \tau_{(1)}^\gamma + \sqrt{\frac{\beta}{N}} \sum_{\gamma=1}^{n} \sum_{i=2}^{N} \sum_{a=2}^{M} s_i^a \tau^{a,\gamma} \xi_i^\gamma + N\beta\,\mathrm{Extr}_{\boldsymbol{p}_S}\right.$$

The function $f_S(\boldsymbol{p}_S)$ together with the values of $\boldsymbol{p}_S$ that minimize it are

$$-\beta f_S(\boldsymbol{p}_S) \;=\; -\alpha \sum_\gamma d_\tau^\gamma \hat{d}_\tau^\gamma + \frac{\alpha\beta}{2} \sum_\gamma d^\gamma(\boldsymbol{\xi}) d_\tau^\gamma + \frac{1}{N} \log \mathbb{E}_{\{\tau_a^\gamma\}} e^{\sum_{\gamma,a} \hat{d}_\tau^\gamma \sum_a (\tau_a^\gamma)^2} \tag{C.23}$$

$$\hat{d}_\tau^\gamma = \frac{\beta}{2} d^\gamma(\boldsymbol{\xi}) \,, \qquad d_{\hat{\tau}}^\gamma = \langle (\tau^\gamma)^2 \rangle_\tau = \frac{\mathbb{E}_{\tau^\gamma} (\tau^\gamma)^2 \, e^{\frac{\beta}{2} d^\gamma(\boldsymbol{\xi})(\tau^\gamma)^2}}{\mathbb{E}_{\tau^\gamma} e^{\frac{\beta}{2} d^\gamma(\boldsymbol{\xi})(\tau^\gamma)^2}} \,. \tag{C.24}$$

RS free energy ans saddle point equations

Based on the preceding sections, we can rewrite the averaged partition function as

$$[Z^n]^{\boldsymbol{\mathcal{S}}} = \sum_{\{\xi^\gamma_{(1)}\}} \mathbb{E}_{\{\tau^\gamma_{(1)}\}} \sum_{\boldsymbol{s},\,\tilde{\boldsymbol{s}}} \exp\left\{ M\sqrt{\frac{\beta}{N}} p^\gamma_\tau(\tau)\xi^\gamma_{(1)} + N\sqrt{\frac{\beta}{N}} p(\xi)^\gamma \tau^\gamma_{(1)} \right\} \times \tag{C.25}$$

$$\times \mathbb{E}_{\hat{\boldsymbol{\xi}}} \mathbb{E}_{\hat{\boldsymbol{\tau}}} \sum_{\boldsymbol{\mathcal{S}}} \exp\left\{ N\hat{\beta}\, \mathrm{Extr}_{\boldsymbol{p}_T}\, f_T(\boldsymbol{p}_T) + \sqrt{\frac{\hat{\beta}}{N}} \sum_{i=2}^{N} \sum_{a=2}^{M} s^a_i \hat{\tau}^a \hat{\xi}_i \right\} \times \tag{C.26}$$

$$\times \sum_{\{\boldsymbol{\xi}^\gamma\}} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}}\ \exp\left\{ N\beta\, \mathrm{Extr}_{\boldsymbol{p}_S}\, f_S(\boldsymbol{p}_S) + \sqrt{\frac{\beta}{N}} \sum_{i=2}^{N} \sum_{a=2}^{M} \sum_{\gamma=1}^{n} s^a_i \tau^{a,\gamma} \xi^\gamma_i \right\}\ . \tag{C.27}$$

By incorporating the variational principle approximation (C.9) into (C.6) and (C.7), we account for both $z(\hat{\boldsymbol{\xi}})^{-M}$ and $z(\{\boldsymbol{\xi}^\gamma\})^{-M}$. Additionally we isolate the average over the first example as described in (C.4, C.5), denoting $\boldsymbol{s} = (s^{(1)}_1, \cdots, s^{(1)}_N)$ and $\tilde{\boldsymbol{s}} = (s^1_{(1)}, \cdots, s^M_{(1)})$. Starting with the first term (C.25) of the quenched free energy, this term can be transformed by constraining the overlap with the examples p^γ_τ and p^γ

$$\int \prod_\gamma dp^\gamma \prod_\gamma dp^\gamma_\tau \mathcal{D}(\boldsymbol{p}, \boldsymbol{p}_\tau) \sum_{\{\xi^\gamma_{(1)}\}} \mathbb{E}_{\{\tau^\gamma_{(1)}\}} \exp\left\{ M\sqrt{\frac{\beta}{N}} \sum_\gamma p^\gamma_\tau \xi^\gamma_{(1)} + N\sqrt{\frac{\beta}{N}} \sum_\gamma p^\gamma \tau^\gamma_{(1)} \right\}$$

$$= \int \prod_\gamma dp^\gamma \prod_\gamma dp^\gamma_\tau \mathcal{D}(\boldsymbol{p}, \boldsymbol{p}_\tau) \exp\left\{ \frac{N\beta}{2} \alpha^2 \Omega \sum_\gamma (p^\gamma_\tau)^2 + \frac{N\beta}{2} \Omega_\tau \sum_\gamma (p^\gamma)^2 \right\}, \tag{C.28}$$

where we introduced the notation for the density of states for the configurations $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$

$$\begin{aligned} \mathcal{D}(\boldsymbol{p}, \boldsymbol{p}_\tau) &= \sum_{\boldsymbol{s},\,\tilde{\boldsymbol{s}}} \prod_\gamma \delta\left(p^\gamma - \frac{1}{N} \sum_i s_i \xi_i \right) \prod_\gamma \delta\left(p^\gamma_\tau - \frac{1}{N} \sum_a \tilde{s}^a \tau^a \right) \\ &\propto \sum_{\boldsymbol{s},\,\tilde{\boldsymbol{s}}} \int \prod_\gamma d\hat{p}^\gamma \prod_\gamma d\hat{p}^\gamma_\tau \exp\left(-N \sum_\gamma \hat{p}_\gamma \left(p_\gamma - \frac{1}{N} \sum_i s_i \xi^\gamma_i \right) \right) - \left(M \sum_\gamma \hat{p}^\gamma_\tau \left(p^\gamma_\tau - \frac{1}{M} \sum_a \tilde{s}_a \tau^\gamma_a \right) \right)\ . \end{aligned}$$

Next, we turn our attention to the second and third terms, (C.26) and (C.27), which emphasize the data-dependent components. Since these terms are already linear in s^a_i we can already compute the average over the examples

$$\begin{aligned} \sum_{\boldsymbol{\mathcal{S}}} \sum_{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^n} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \mathbb{E}_{\hat{\boldsymbol{\xi}}}\, \mathbb{E}_{\hat{\boldsymbol{\tau}}} e^{\sum_{i,a} s^a_i \left(\sqrt{\frac{\hat{\beta}}{N}} \hat{\tau}_a \hat{\xi}_i + \sqrt{\frac{\beta}{N}} \sum_\gamma \tau^\gamma_a \xi^\gamma_i \right)} &= \sum_{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^n} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}} \mathbb{E}_{\hat{\boldsymbol{\xi}}}\, \mathbb{E}_{\hat{\boldsymbol{\tau}}} e^{\sum_{i,a} \ln\cosh(\left(\sqrt{\frac{\hat{\beta}}{N}} \hat{\tau}_a \hat{\xi}_i + \sqrt{\frac{\beta}{N}} \sum_\gamma \tau^\gamma_a \xi^\gamma_i \right)} \\ \approx \sum_{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^n} \mathbb{E}_{\{\boldsymbol{\tau}^\gamma\}}\ \mathbb{E}_{\hat{\boldsymbol{\xi}}}\ \mathbb{E}_{\hat{\boldsymbol{\tau}}}\ e^{\frac{M\hat{\beta}}{2} d_{\hat{\tau}} d_{\hat{\xi}} + M\beta \sum_{\gamma<\gamma'} q^{\gamma\gamma'}_\tau q^{\gamma\gamma'} + \frac{M\beta}{2} \sum_\gamma d^\gamma_\tau d^\gamma + M\sqrt{\hat{\beta}\beta} \sum_\gamma m^\gamma_\tau m^\gamma}\ , \end{aligned} \tag{C.29}$$

in the last line we expanded the $\log\cosh$ and introduced the order parameters

$$
\begin{aligned}
m^{\gamma} &= \sum_i \frac{\hat{\xi}_i \xi_i}{N}, & m^{\gamma}_{\tau} &= \sum_a \frac{\hat{\tau}_a \tau_a}{M}\,, \\
d^{\gamma} &= \sum_i \frac{(\xi^{\gamma}_i)^2}{N} & d^{\gamma}_{\tau} &= \sum_a \frac{(\tau^{\gamma}_a)^2}{M}\,, \\
q^{\gamma\gamma'} &= \sum_i \frac{\xi^{\gamma}_i \xi^{\gamma'}_i}{N} & q^{\gamma\gamma'}_{\tau} &= \sum_a \frac{\tau^{\gamma}_a \tau^{\gamma'}_a}{M}\,, \\
d_{\hat{\xi}} &= \sum_i \frac{\hat{\xi}^2_i}{N} & d_{\hat{\tau}} &= \sum_a \frac{\hat{\tau}^2_a}{M}\,.
\end{aligned}
$$

Consequently, we can fix the values of these parameters, as previously done in (C.28), and incorporate the corresponding densities of states $\mathcal{D}(\mathbf{\Lambda}, \mathbf{\Lambda}_\tau) := \mathcal{D}(\boldsymbol{m}, \boldsymbol{q}, \boldsymbol{d})\,\mathcal{D}(\boldsymbol{m}_\tau, \boldsymbol{q}_\tau, \boldsymbol{d}_\tau)\,\mathcal{D}(d_{\hat{\xi}}, d_{\hat{\tau}})$. Both the densities affect the extremized functions f_T and f_S, as their parameters depend respectively on $d_{\hat{\xi}}$ and d^{γ}. For brevity of notation we defined the sets of order parameters: $\mathbf{\Lambda} := \{\boldsymbol{p}, \boldsymbol{m}, \boldsymbol{q}, \boldsymbol{d}, d_{\hat{\xi}}\}$ and $\mathbf{\Lambda}_\tau := \{\boldsymbol{p}_\tau, \boldsymbol{m}_\tau, \boldsymbol{q}_\tau, \boldsymbol{d}_\tau, d_{\hat{\tau}}\}$ and their conjugated ones $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Lambda}}_\tau$, obtaining the following form of the averaged partition function

$$
\begin{aligned}
[Z^n]^{\boldsymbol{\mathcal{S}}} = \int d\,\mathbf{\Lambda} d\,\mathbf{\Lambda}_\tau \mathcal{D}(\mathbf{\Lambda}, \mathbf{\Lambda}_\tau) \exp\Big\{ &\frac{M\hat{\beta}}{2} d_{\hat{\tau}} d_{\xi} + M\beta \sum_{\gamma<\gamma'} q^{\gamma\gamma'}_{\tau} q^{\gamma\gamma'} + \frac{N\beta}{2}\alpha^2 \Omega \sum_{\gamma} (p^{\gamma}_{\tau})^2 + \frac{N\beta}{2}\Omega_\tau \sum_{\gamma} (p^{\gamma})^2 + \\
&+\frac{M\beta}{2}\sum_{\gamma} d^{\gamma}_{\tau} d^{\gamma} + M\sqrt{\hat{\beta}\beta}\sum_{\gamma} m^{\gamma}_{\tau} m^{\gamma} + N\hat{\beta}\,\mathrm{Extr}_{\boldsymbol{p}_T}\, f_T(\boldsymbol{p}_T) + N\beta\,\mathrm{Extr}_{\boldsymbol{p}_S}\, f_S(\boldsymbol{p}_S) \Big\}\,.
\end{aligned}
\tag{C.30}
$$

At this point we use the Laplace method on the averaged partition function (C.30). Remembering the replica trick we have

$$
-\beta f(\beta, \hat{\beta}, \alpha) = \lim_{N\to\infty} \frac{1}{N}\,[\ln Z]^{\boldsymbol{\mathcal{S}}} = \lim_{\substack{N\to\infty \\ n\to 0}} \frac{\ln[Z^n]^{\boldsymbol{\mathcal{S}}}}{Nn}
$$

$$
-\beta f \approx \lim_{\substack{N\to\infty \\ n\to 0}} \frac{1}{Nn} \ln\left(e^{N\,\mathrm{Extr}_{\substack{\mathbf{\Lambda},\mathbf{\Lambda}_\tau \\ \hat{\mathbf{\Lambda}},\hat{\mathbf{\Lambda}}_\tau}} \hat{f}(\mathbf{\Lambda},\mathbf{\Lambda}_\tau,\hat{\mathbf{\Lambda}},\hat{\mathbf{\Lambda}}_\tau)} \right) \approx \lim_{n\to 0} \frac{1}{n}\,\mathrm{Extr}_{\substack{\mathbf{\Lambda},\mathbf{\Lambda}_\tau \\ \hat{\mathbf{\Lambda}},\hat{\mathbf{\Lambda}}_\tau}}\, \hat{f}(\mathbf{\Lambda}, \mathbf{\Lambda}_\tau, \hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}_\tau) \tag{C.31}
$$

$$
\begin{aligned}
-\beta\hat{f}(\boldsymbol{\Lambda},\boldsymbol{\Lambda}_\tau,\hat{\boldsymbol{\Lambda}},\hat{\boldsymbol{\Lambda}}_\tau) ={}& \frac{\alpha\hat{\beta}}{2}d_{\hat{\tau}}d_{\hat{\xi}} + \alpha\beta\sum_{\gamma<\gamma'} q_\tau^{\gamma\gamma'}q^{\gamma\gamma'} + \frac{\beta}{2}\alpha^2\Omega_\xi\sum_\gamma (p_\tau^\gamma)^2 + \frac{\beta}{2}\Omega_\tau\sum_\gamma (p^\gamma)^2 + \\
&+ \frac{\alpha\beta}{2}\sum_\gamma d_\tau^\gamma d^\gamma + \alpha\sqrt{\hat{\beta}\beta}\sum_\gamma m_\tau^\gamma m^\gamma + \hat{\beta}\,\mathrm{Extr}_{\boldsymbol{p}_T}\, f_T(\boldsymbol{p}_T) + \\
&- \hat{d}_{\hat{\xi}}d_{\hat{\xi}} - \sum_\gamma \hat{p}^\gamma p^\gamma - \sum_{\gamma<\gamma'}\hat{q}^{\gamma\gamma'}q^{\gamma\gamma'} - \sum_\gamma \hat{m}^\gamma m^\gamma + \beta\,\mathrm{Extr}_{\boldsymbol{p}_S}\, f_S(\boldsymbol{p}_S) + \\
&- \alpha\hat{d}_{\hat{\tau}}d_{\hat{\tau}} - \alpha\sum_\gamma \hat{p}_\tau^\gamma p_\tau^\gamma - \alpha\sum_{\gamma<\gamma'}\hat{q}_\tau^{\gamma\gamma'}q_\tau^{\gamma\gamma'} - \alpha\sum_\gamma \hat{m}_\tau^\gamma m_\tau^\gamma + \\
&+ \alpha\log\mathbb{E}_{\hat{\tau}}\mathbb{E}_{\tilde{s}}\mathbb{E}_{\{\tau^\gamma\}}\exp\left(\hat{d}_{\hat{\tau}}\,\hat{\tau}^2 + \frac{1}{2}\sum_{\gamma\gamma'}\hat{q}_\tau^{\gamma\gamma'}\tau^\gamma\tau^{\gamma'} - \frac{1}{2}\sum_\gamma \hat{q}_\tau^{\gamma\gamma}(\tau^\gamma)^2 + \sum_\gamma \hat{m}_\tau^\gamma\tau^\gamma\hat{\tau} + \right. \\
&\qquad\qquad\qquad\qquad \left. + \sum_\gamma \hat{d}_\tau^\gamma(\tau^\gamma)^2 + \sum_\gamma \tilde{s}p_\tau^\gamma\tau^\gamma\right) + \\
&+ \log\mathbb{E}_{\hat{\xi}}\mathbb{E}_s\mathbb{E}_{\{\xi^\gamma\}}\exp\left(\hat{d}_{\hat{\xi}}\,\hat{\xi}^2 + \frac{1}{2}\sum_{\gamma\gamma'}\hat{q}^{\gamma\gamma'}\xi^\gamma\xi^{\gamma'} - \frac{1}{2}\sum_\gamma \hat{q}^{\gamma\gamma}(\xi^\gamma)^2 + \sum_\gamma \hat{m}^\gamma\xi^\gamma\hat{\xi} + \right. \\
&\qquad\qquad\qquad\qquad \left. \sum_\gamma \hat{d}^\gamma(\xi^\gamma)^2 + \sum_\gamma s\,\hat{p}^\gamma\xi^\gamma\right),
\end{aligned}
\tag{C.32}
$$

where, in the last two lines, we just collect the linearities in a, i inside the exponentials of $\mathcal{D}(\boldsymbol{\Lambda},\boldsymbol{\Lambda}_\tau)$ and substitute them with, respectively, N and M times the averages on one representative. As an example

$$
\begin{aligned}
\log\mathbb{E}_{\hat{\boldsymbol{\xi}}}\mathbb{E}_{\boldsymbol{s}}\mathbb{E}_{\{\boldsymbol{\xi}^\gamma\}}\exp&\left(\hat{d}_{\hat{\xi}}\sum_i \hat{\xi}_i^2 + \frac{1}{2}\sum_{\gamma\gamma'}\hat{q}^{\gamma\gamma'}\sum_i \xi_i^\gamma\xi_i^{\gamma'} - \frac{1}{2}\sum_\gamma \hat{q}^{\gamma\gamma}\sum_i(\xi_i^\gamma)^2 + \right. \\
&\qquad \left. + \sum_\gamma \hat{m}^\gamma\sum_i \xi_i^\gamma\hat{\xi}_i + \sum_\gamma \hat{d}^\gamma\sum_i(\xi_i^\gamma)^2 + \sum_{i\gamma} s_i\,\hat{p}^\gamma\xi_i^\gamma\right) =
\end{aligned}
$$

$$
= \log\left[\mathbb{E}_{\hat{\xi}}\mathbb{E}_s\mathbb{E}_{\{\xi^\gamma\}}\exp\left(\hat{d}_{\hat{\xi}}\hat{\xi}^2 + \frac{1}{2}\sum_{\gamma\gamma'}\hat{q}^{\gamma\gamma'}\xi^\gamma\xi^{\gamma'} - \frac{1}{2}\sum_\gamma \hat{q}^{\gamma\gamma}(\xi^\gamma)^2 + \sum_\gamma \hat{m}^\gamma\xi^\gamma\hat{\xi} + \sum_\gamma \hat{d}^\gamma(\xi^\gamma)^2 + \sum_\gamma s\,\hat{p}^\gamma\xi^\gamma\right)\right]^N,
$$

then, using the properties of the $\log$ function, one can express it as the extensive term of the last line of (C.32).

From now on we assume the Replica symmetric (RS) ansatz for all the order parameters: $q^{\gamma\gamma'} = q, m^\gamma = m, p^\gamma = p, d^\gamma = d,\ \forall\gamma,\gamma' = 1,\dots,n$, same is valid also for their τ version and their conjugated. The last two terms of the partition function can be ulteriorly worked out

$$
\begin{aligned}
&\mathbb{E}_{\hat\tau} e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s}\, \mathbb{E}_{\{\tau^\gamma\}} \exp\left(\frac{1}{2}\hat q_\tau \left(\sum_\gamma \tau^\gamma\right)^2 - \frac{1}{2}\hat q_\tau \sum_\gamma (\tau^\gamma)^2 + \hat m_\tau \sum_\gamma \tau^\gamma \hat\tau + \hat d_\tau \sum_\gamma (\tau^\gamma)^2 + \tilde s \hat p_\tau \sum_\gamma \tau^\gamma\right) = \\
=& \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \left(\mathbb{E}_\tau \exp\left(-(\frac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right)\right)^n \right\rangle_z \\
=& \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \exp\left\{ n \log \left[\mathbb{E}_\tau \exp\left(-(\frac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right)\right]\right\} \right\rangle_z \\
\simeq& \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \left\{ 1 + n \log \left[\mathbb{E}_\tau \exp\left(-(\frac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right)\right]\right\} \right\rangle_z \\
=& \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} + n \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \log \left[\mathbb{E}_\tau \exp\left(-(\frac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right)\right] \right\rangle_z \\
=& \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \left\{ 1 + \frac{n}{\mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2}} \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \log \left[\mathbb{E}_\tau \exp\left(-(\frac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right)\right] \right\rangle_z \right\} ,
\end{aligned}
$$

where we use the expansion of the exponential for small n. Reconstructing the term inside the free energy and substituting $\log(1+x) \simeq x$ for small x

$$
\begin{aligned}
&\alpha \log \left(\mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \left\{ 1 + \tfrac{n}{\mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2}} \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \log \mathbb{E}_\tau \exp\left(-(\tfrac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right) \right\rangle_z \right\}\right) \simeq \\
&\quad \simeq\ \alpha \log \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} + \tfrac{\alpha n}{\mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2}} \left\langle \mathbb{E}_{\hat\tau}\, e^{\hat d_{\hat\tau} \hat\tau^2} \mathbb{E}_{\tilde s} \log \mathbb{E}_\tau \exp\left(-(\tfrac{\hat q_\tau}{2} - \hat d_\tau)\tau^2 + \tau(\hat m_\tau \hat\tau + \sqrt{\hat q_\tau} z + \tilde s\, \hat p_\tau)\right) \right\rangle_z ,
\end{aligned}
$$

the same procedure holds for

$$
\begin{aligned}
&\log \mathbb{E}_{\hat\xi} e^{\hat d_{\hat\xi} \hat\xi^2} \mathbb{E}_s\, \mathbb{E}_{\{\xi^\gamma\}} \exp\left(\frac{\hat q}{2}\left(\sum_\gamma \xi^\gamma\right)^2 - \frac{\hat q}{2} \sum_\gamma (\xi^\gamma)^2 + \hat m \sum_\gamma \xi^\gamma \hat\xi + \hat d \sum_\gamma (\xi^\gamma)^2 + s\, \hat p \sum_\gamma \xi^\gamma\right) \simeq \\
&\simeq \log \mathbb{E}_{\hat\xi} e^{\hat d_{\hat\xi}\, \hat\xi^2} + \tfrac{n}{\mathbb{E}_{\hat\xi}\, e^{\hat d_{\hat\xi}\, \hat\xi^2}} \left\langle \mathbb{E}_{\hat\xi}\, e^{\hat d_{\hat\xi}\, \hat\xi^2} \mathbb{E}_s \log \mathbb{E}_\xi \exp\left(-(\tfrac{\hat q}{2} - \hat d)\xi^2 + \xi(\hat m\, \hat\xi + \sqrt{\hat q}\, z + s\, \hat p)\right) \right\rangle_z .
\end{aligned}
$$

Collecting the previous two expansions, and imposing the RS ansatz on the other terms of the free energy (C.31), the final extremization can be rewritten as

$$
-\beta f^{RS}(\beta, \hat\beta, \alpha) \approx \lim_{n \to 0} \frac{1}{n} \operatorname{Extr}_{\substack{\Lambda, \hat\Lambda \\ \Lambda_\tau, \hat\Lambda_\tau}} f^{RS}(\Lambda, \hat\Lambda, \Lambda_\tau, \hat\Lambda_\tau) ,
$$

where now Λ is the RS reduced set of the pattern-related order parameters $\Lambda = \{p, q, m, d\}$, while $\hat\Lambda$ is the conjugated one ($\Lambda_\tau, \hat\Lambda_\tau$ are the same for the hidden units). The free energy can be divided into two parts according to its power of n

$$
f^{RS}(\Lambda, \hat\Lambda, \Lambda_\tau, \hat\Lambda_\tau) = f_0^{RS} + n f_1^{RS} + \mathcal{O}(n^2) ,
$$

$$
\begin{aligned}
f_0^{RS} =& \frac{\alpha\hat{\beta}}{2} d_{\hat{\tau}} d_{\hat{\xi}} + \hat{\beta}\,\mathrm{Extr}_{\boldsymbol{p}_T}\, f_T(\boldsymbol{p}_T) + \alpha \log \mathbb{E}_{\hat{\tau}} e^{\hat{d}_{\hat{\tau}}(\hat{\tau})^2} + \log \mathbb{E}_{\hat{\xi}} e^{\hat{d}_{\hat{\xi}}(\hat{\xi})^2} - \alpha \hat{d}_{\hat{\tau}} d_{\hat{\tau}} - \hat{d}_{\hat{\xi}} d_{\hat{\xi}}\ , \\
f_1^{RS} =& \frac{1}{2}\hat{q}q + \alpha\frac{1}{2}\hat{q}_\tau q_\tau - \hat{m}m - \alpha\hat{m}_\tau m_\tau - \hat{d}d - \alpha\hat{d}_\tau d_\tau + \alpha^2\frac{\beta}{2}\Omega_\xi p_\tau^2 + \frac{\beta}{2}\Omega_\tau p^2 + \\
&+ \frac{\alpha}{\mathbb{E}_{\hat{\tau}} e^{\hat{d}_{\hat{\tau}}\hat{\tau}^2}} \left\langle \mathbb{E}_{\hat{\tau}} e^{\hat{d}_{\hat{\tau}}\hat{\tau}^2} \mathbb{E}_s \log \mathbb{E}_\tau \exp\left\{ -(\frac{\hat{q}_\tau}{2} - \hat{d}_\tau)\tau^2 + \tau(\hat{m}_\tau\hat{\tau} + \sqrt{\hat{q}_\tau} z + \hat{p}_\tau s) \right\} \right\rangle_z + \\
&+ \frac{1}{\mathbb{E}_{\hat{\xi}} e^{\hat{d}_{\hat{\xi}}\hat{\xi}^2}} \left\langle \mathbb{E}_{\hat{\xi}} e^{\hat{d}_{\hat{\xi}}\hat{\xi}^2} \mathbb{E}_s \log \mathbb{E}_\xi \exp\left\{ -(\frac{\hat{q}}{2} - \hat{d})\xi^2 + \xi(\hat{m}\hat{\xi} + \sqrt{\hat{q}} z + \hat{p}s) \right\} \right\rangle_z + \\
&- \frac{\beta\alpha}{2} q q_\tau + \alpha\sqrt{\hat{\beta}\beta}\, m m_\tau - \alpha p_\tau \hat{p}_\tau - p\hat{p} + \frac{\beta\alpha}{2} d d_\tau + \beta\,\mathrm{Extr}_{\boldsymbol{p}_S}\, f_S(\boldsymbol{p}_S)\,.
\end{aligned}
\tag{C.33}
$$

The extremization of f_0^{RS} involves only the teacher order parameters. The extremizers should be s.t. $\mathrm{Extr}\, f_0^{RS} = 0$ in order to avoid divergences when $n \to 0$. Indeed, remembering Eqs.(C.17), the parameters of order zero in n satisfy the following

$$
\begin{aligned}
\hat{d}_{\hat{\xi}} = 0\,, &\quad d_{\hat{\xi}} = 1\,, \\
\hat{d}_{\hat{\tau}} = \frac{\hat{\beta}}{2}\,, &\quad d_{\hat{\tau}} = \langle \hat{\tau}^2 \rangle_{\hat{\tau}}\,,
\end{aligned}
\tag{C.34}
$$

which implies $f_0^{RS} = 0$ and ensure the student partition function is not diverging. Using Eqs.(C.24) and updating f_1^{RS} with (C.34) we obtain the values of Λ and Λ_τ that extremize f_1^{RS}

$$
\hat{p}_\tau = \beta\alpha\Omega_\xi p_\tau\,, \quad \hat{m}_\tau = \sqrt{\beta\hat{\beta}}m\,, \quad \hat{q}_\tau = \beta q\,, \quad \hat{d}_\tau = \beta\frac{d}{2}\,, \tag{C.35}
$$

$$
\hat{p} = \beta\Omega_\tau p\,, \quad \hat{m} = \alpha\sqrt{\beta\hat{\beta}}m_\tau\,, \quad \hat{q} = \alpha\beta q_\tau\,, \quad \hat{d} = \alpha\beta\left(\frac{d_\tau}{2} - \langle\tau^2\rangle_\tau\right)\,, \tag{C.36}
$$

$$
p = \left\langle s\langle\xi\rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad p_\tau = \left\langle s\langle\tau\rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} \tag{C.37}
$$

$$
m = \left\langle \hat{\xi}\langle\xi\rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad m_\tau = \left\langle \hat{\tau}\langle\tau\rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} \tag{C.38}
$$

$$
q = \left\langle \langle\xi^2\rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad q_\tau = \left\langle \langle\tau\rangle^2_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}} \tag{C.39}
$$

$$
d = \left\langle \langle\xi^2\rangle_{\xi|z,s,\hat{\xi}} \right\rangle_{z,s,\hat{\xi}} \qquad d_\tau = \left\langle \langle\tau^2\rangle_{\tau|z,s,\hat{\tau}} \right\rangle_{z,s,\hat{\tau}}\,. \tag{C.40}
$$

The internal distributions that average ξ and τ are respectively

$$
P(\xi|z,s,\hat{\xi}) = P_\xi(\xi)\exp\left\{ -\left(\frac{\hat{q}}{2} - \hat{d}\right)\xi^2 + \xi\left(\hat{m}\hat{\xi} + \sqrt{\hat{q}}z + \hat{p}s\right) \right\}\,, \tag{C.41}
$$

$$
P(\tau|z,s,\hat{\tau}) = P_\tau(\tau)\exp\left\{ -\left(\frac{\hat{q}_\tau}{2} - \hat{d}_\tau\right)\tau^2 + \tau\left(\hat{m}_\tau\hat{\tau} + \sqrt{\hat{q}_\tau}z + \hat{p}_\tau s\right) \right\}\,, \tag{C.42}
$$

while the outher distributions $P_{\hat{\xi}}(\hat{\xi})$ and $P_{\hat{\tau}}(\hat{\tau})e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}$ depend from the choice of the planted configuration and z is a Normal random variable.

Appendix D

RS equations in the interpolating case

So far, Eqs. (C.37-C.40) are general and valid for any arbitrary choice of the priors for all variables. In this section we derive explicitly the saddle point equations, replacing each of the random variables with their prior interpolation (3.7). The expressions of the distributions (C.41) and (C.42) can be recast as follows:

$$Z_\xi{}^{-1} \sum_{\epsilon} e^{-\frac{\xi^2}{2\gamma} + \xi(\phi\epsilon + h)} \,,$$

$$Z_\tau{}^{-1} \sum_{\epsilon_\tau} e^{-\frac{\tau^2}{2\gamma_\tau} + \xi(\phi_\tau \epsilon_\tau + h_\tau)} \,,$$

where we use the shorthand notation

$$\begin{aligned} \gamma = \frac{\Omega_\xi}{1 - \alpha\beta\Omega_\xi\left(d_\tau - q_\tau - \langle\tau^2\rangle_\tau\right)}, \qquad \gamma_\tau &= \frac{\Omega_\tau}{1 - \beta\Omega_\tau\left(d - q\right)}\,, \\ \phi = \frac{\sqrt{\delta}}{\Omega_\xi}\,, \qquad \phi_\tau &= \frac{\sqrt{\delta_\tau}}{\Omega_\tau}\,, \\ h = \hat{m}\hat{\xi} + \sqrt{\hat{q}}z + \hat{p}s\,, \qquad h_\tau &= \hat{m}_\tau\hat{\tau} + \sqrt{\hat{q}_\tau}z + \hat{p}_\tau s\,. \end{aligned}$$

The averages in the mean field equations (C.37-C.40) become:

$$\begin{aligned} \langle\xi\rangle_{\xi|z,s,\hat{\xi}} =& \partial_h \ln Z_\xi = \frac{\sum_\epsilon e^{\frac{\gamma}{2}(\phi\epsilon+h)^2}\gamma\left(\phi\epsilon + h\right)}{\sum_\epsilon e^{\frac{\gamma}{2}(\phi\epsilon+h)^2}} = \gamma\, h + \gamma\phi \tanh\left(\gamma\phi\, h\right) \\ \langle\xi^2\rangle_{\xi|z,s,\hat{\xi}} =& \partial_h^2 \ln Z_\xi + \langle\xi\rangle^2_{\xi|z,s,\hat{\xi}} = \gamma + \gamma^2(h^2 + \phi^2) + 2\gamma^2\phi\, h \tanh\left(\gamma\phi\, h\right)\,, \end{aligned}$$

the same is also valid for the corresponding τ counterparts. We can also proceed with the evaluation of the averages on the planted teacher and the example randomness, using their interpolating decomposition

$$\mathbb{E}_{s,\hat{\xi}} f(s,\hat{\xi}) = \mathbb{E}_{g_{s,\hat{\xi}}}\mathbb{E}_{\epsilon_{s,\hat{\xi}}} f(\sqrt{\Omega_s}\, g_s + \sqrt{\delta_s}\epsilon_s, \sqrt{\Omega_{\hat{\xi}}}\, g_{\hat{\xi}} + \sqrt{\delta_{\hat{\xi}}}\, \epsilon_{\hat{\xi}})\,.$$

This leads to the final form

$$p =\gamma\beta\Omega_\tau p + \gamma^2\phi^2\Omega_s\beta\Omega_\tau p\left(1 - \bar{q}\right) + \gamma\phi\sqrt{\delta_s}\bar{m}_s, \tag{D.1}$$

$$m =\gamma\alpha\sqrt{\hat{\beta}\beta}m_\tau + \gamma^2\phi^2\Omega_{\hat{\xi}}\,\alpha\sqrt{\hat{\beta}\beta}\, m_\tau\left(1 - \bar{q}\right) + \gamma\phi\sqrt{\delta_{\hat{\xi}}}\,\bar{m}_* \tag{D.2}$$

$$\begin{aligned} q =& \gamma^2\left(\left(\alpha\sqrt{\hat{\beta}\beta}\, m_\tau\right)^2 + \alpha\beta q_\tau + (\beta\Omega_\tau p)^2\right) + \gamma^2\phi^2\bar{q} + \\ &+ 2\gamma^2\phi\left(\beta\Omega_\tau p\sqrt{\delta_s}\bar{m}_s + \alpha\sqrt{\hat{\beta}\beta}q_\tau\sqrt{\delta_{\hat{\xi}}}\,\bar{m}_*\right) + \\ &+ 2\gamma^3\phi^2\left((\beta\Omega_\tau p)^2\Omega_s + (\alpha\sqrt{\hat{\beta}\beta}\, m_\tau)^2\Omega_{\hat{\xi}} + \alpha\beta q_\tau\right)\left(1 - \bar{q}\right) \end{aligned} \tag{D.3}$$

$$d =\gamma + \gamma^2\phi^2\left(1 - \bar{q}\right) + q\,. \tag{D.4}$$

Here we can distinguish the effective magnetization related to the signal and the example, along with an effective overlap

$$\begin{aligned}\bar{m}_* &= \left\langle \mathbb{E}_s \tanh\left(\gamma\phi\left(\alpha m_\tau\sqrt{\hat{\beta}\beta\delta_{\hat{\xi}}} + \sigma z + \beta\Omega_\tau p s\right)\right)\right\rangle_z \\ \bar{m}_s &= \left\langle \mathbb{E}_s\epsilon_s \tanh\left(\gamma\phi\left(\alpha m_\tau\sqrt{\hat{\beta}\beta\delta_{\hat{\xi}}} + \sigma z + \beta\Omega_\tau p s\right)\right)\right\rangle_z \\ \bar{q} &= \left\langle \mathbb{E}_s \tanh^2\left(\gamma\phi\left(\alpha m_\tau\sqrt{\hat{\beta}\beta\delta_{\hat{\xi}}} + \sigma z + \beta\Omega_\tau p s\right)\right)\right\rangle_z .\end{aligned} \tag{D.5}$$

where $\sigma^2 = \left(\alpha q_\tau\sqrt{\hat{\beta}\beta\Omega_{\hat{\xi}}}\right)^2 + \alpha\beta q_\tau$. The same can be done for the τ side. This turns out not to be a completely symmetric case, due to the presence of the Gaussian factor inside the average over $\hat{\tau}$

$$\mathbb{E}_{\hat{\tau}}\, e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2} f(\hat{\tau}) = \mathbb{E}_{\hat{g}_\tau}\mathbb{E}_{\hat{\epsilon}_\tau}\, e^{(\sqrt{\Omega_{\hat{\tau}}}\, g_{\hat{\tau}} + \sqrt{\delta_{\hat{\tau}}}\,\epsilon_{\hat{\tau}})^2} f(\sqrt{\Omega_{\hat{\tau}}}\, g_{\hat{\tau}} + \sqrt{\delta_{\hat{\tau}}}\,\epsilon_{\hat{\tau}}) ,$$

this difference is due to the presence of the self spherical constraint (C.10) in the teacher model. The hidden set of equations are

$$p_\tau = \gamma_\tau\beta\alpha\Omega_\xi p_\tau + \gamma_\tau^2\phi_\tau^2\Omega_s\beta\alpha\Omega_\xi\, p_\tau\,(1-\bar{q}_\tau) + \gamma_\tau\phi_\tau\sqrt{\delta_s}\bar{m}_{\tau,s} \tag{D.6}$$

$$m_\tau = \gamma_\tau\sqrt{\hat{\beta}\beta}\, m\left(\frac{\delta_{\hat{\tau}} + \Omega_{\hat{\tau}}(1-\hat{\beta}\Omega_{\hat{\tau}})}{(1-\hat{\beta}\Omega_{\hat{\tau}})^2}\right) + \frac{\sqrt{\delta_{\hat{\tau}}}}{\left(1-\hat{\beta}\Omega_{\hat{\tau}}\right)}\gamma_\tau\phi_\tau\bar{m}_{\tau,*} \tag{D.7}$$

$$+\frac{\gamma_\tau^2\phi_\tau^2}{(1-\hat{\beta}\Omega_{\hat{\tau}})}\sqrt{\hat{\beta}\beta}\, m\,\Omega_{\hat{\tau}}\,(1-\bar{q}_\tau) \tag{D.8}$$

$$q_\tau = \gamma_\tau^2\left[\beta q + (\beta\alpha\Omega_\xi\, p_\tau)^2 + (\sqrt{\hat{\beta}\beta}\, m)^2\left(\frac{\delta_{\hat{\tau}} + \Omega_{\hat{\tau}}(1-\hat{\beta}\Omega_{\hat{\tau}})}{(1-\hat{\beta}\Omega_{\hat{\tau}})^2}\right)\right] + \gamma_\tau^2\phi_\tau^2\bar{q}_\tau + \tag{D.9}$$

$$+2\gamma_\tau^2\phi_\tau\left(\beta\alpha\Omega_\xi p_\tau\sqrt{\delta_s}\bar{m}_{\tau,s} + \frac{\sqrt{\hat{\beta}\beta}\, m\sqrt{\delta_{\hat{\tau}}}}{(1-\hat{\beta}\Omega_{\hat{\tau}})}\bar{m}_{\tau,*}\right)$$

$$+2\gamma_\tau^3\phi_\tau^2\left(\Omega_s(\beta\alpha\Omega_\xi p_\tau)^2 + \beta q + \frac{(\sqrt{\hat{\beta}\beta}\, m)^2\Omega_{\hat{\tau}}}{(1-\hat{\beta}\Omega_{\hat{\tau}})}\right)(1-\bar{q}_\tau)$$

$$d_\tau = \gamma_\tau^2\phi_\tau^2\,(1-\bar{q}_\tau) + \gamma_\tau + q_\tau\ . \tag{D.10}$$

The effective magnetizations and overlap in the τ version are

$$\begin{aligned}\bar{m}_{\tau,*} &= \frac{1}{\mathbb{E}_{\hat{\tau}}\, e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}}\left\langle \mathbb{E}_s\,\mathbb{E}_{\hat{\tau}}\, e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}\epsilon_{\hat{\tau}} \tanh\left(\gamma_\tau\phi_\tau\left(\sqrt{\hat{\beta}\,\beta}\, m\hat{\tau} + \sqrt{\beta q}z + \beta\alpha\Omega_\xi p_\tau s\right)\right)\right\rangle_z \\ \bar{m}_{\tau,s} &= \frac{1}{\mathbb{E}_{\hat{\tau}} e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}}\left\langle \mathbb{E}_s\,\mathbb{E}_{\hat{\tau}} e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}\epsilon_{\tau,s} \tanh\left(\gamma_\tau\phi_\tau\left(\sqrt{\hat{\beta}\,\beta}\, m\hat{\tau} + \sqrt{\beta q}z + \beta_2\alpha\Omega_\xi p_\tau s\right)\right)\right\rangle_z \\ \bar{q} &= \frac{1}{\mathbb{E}_{\hat{\tau}} e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2}}\left\langle \mathbb{E}_s\,\mathbb{E}_{\hat{\tau}} e^{\frac{\hat{\beta}}{2}(\hat{\tau})^2} \tanh^2\left(\gamma_\tau\phi_\tau\left(\sqrt{\hat{\beta}\,\beta}\, m\hat{\tau} + \sqrt{\beta q}z + \beta\alpha\Omega_\xi p_\tau s\right)\right)\right\rangle_z .\end{aligned} \tag{D.11}$$

Numerical solver and complexity

The self-consistent equations (D.1-D.4) are solved using a relaxed fixed-point iteration scheme. At each step, the difference between the current and next estimates, described by the values of the parameters p, q, m, is computed, and the process continues until this difference falls below a suitably small threshold. The employed relaxation updates the order parameters as follows:

$$\Lambda^{(t+1)} = \Lambda^{(t)} + \boldsymbol{w}^{(t)}\boldsymbol{\Delta}(\Lambda^{(t)})\,, \tag{D.12}$$

where the $\boldsymbol{w}^{(t)}$ is a vector of relaxation weights. These weights can be constant across the entire parameter set Λ, or dynamically adjusted to better prevent oscillations or divergence, as proposed in e.g. [98]. Convergence is declared when the total residual—defined as the sum of absolute differences across all parameters—falls below a fixed threshold ϵ.

Eqs. (D.1-D.4) involve the numerical evaluation of the effective magnetizations and overlap defined in (D.5, D.11), whose computational complexity depends on the chosen interpolation parameters Ω_x, $x \in \{\hat{\xi}, \hat{\tau}, \xi, \tau, s\}$. In the most general case, where $\Omega_x \in (0,1)$, the order parameter x exhibits both binary and Gaussian behavior, requiring three integrals to be computed. For example, in the case of (D.11), this results in a computational complexity of $\mathcal{O}(n^3)$. However, this can be reduced to $\mathcal{O}(n^2)$, by restricting the search to solutions with $p_\tau = 0$.

A similar issue arises with the effective magnetizations in (D.5). The first of these equations takes the form

$$\bar{m}_* = \left\langle \mathbb{E}_s \, \mathbb{E}_{\hat{\xi}} \, \epsilon_{\hat{\xi}} \tanh\left(\gamma\phi\left(\alpha\sqrt{\hat{\beta}\,\beta}\, m_\tau \hat{\xi} + \alpha\sqrt{\beta q_\tau} z + \beta\Omega_\tau p s\right)\right)\right\rangle_z , \tag{D.13}$$

which also exhibits $\mathcal{O}(n^3)$ complexity. Unlike the τ-equations, we are also interested in solutions where $p \neq 0$; nevertheless, the reduction to $\mathcal{O}(n^2)$ remains valid. Indeed, solutions with $p \neq 0$ (the pattern ξ aligned with a single example s), imply $m = 0$, and consequently $m_\tau = 0$. Conversely, if $m \neq 0$, then $p = 0$, again leading to the same $\mathcal{O}(n^2)$ complexity. Moreover, in equation (D.13), there is no exponential factor of the form $e^{\frac{\hat{\beta}}{2}(\hat{\xi})^2}$, inside $\mathbb{E}_{\hat{\xi}}$. In the absence of such problematic mixing terms, a Gaussian convolution can be applied, yielding the simplified form of Eq.(D.5), and further reducing the complexity to $\mathcal{O}(n)$. Since both sets of effective magnetizations must be computed during each iteration, the overall computational complexity is $\mathcal{O}(n^2)$.

Additional remarks

The methodology described above can be specialized to the case of the Dual Hopfield model presented in Chapter 2, particularly for solving Eqs.(2.63-2.65), which can also be addressed using the iterative scheme of Eq.(D.12). Since the Hopfield network corresponds to a specific instance of the RBM-satisfying $\Omega_{\hat{\xi}} = \Omega_\xi = 0$ and $\Omega_{\hat{\tau}} = \Omega_\tau = 1$-the computational complexity reduces to $\mathcal{O}(n)$, as only a single Gaussian expectation needs to be computed.

A similar argument applies to the model discussed in Chapter 4, where the replica method is used to solve the self-consistency equations (E.26-E.28) describing a system of interacting Hopfield student networks. In this case, the evaluation of the equations involves two irreducible Gaussian integrals, resulting in an overall computational complexity of $\mathcal{O}(n^2)$.

Chapter 4

Inference with coupled Hopfield students

In the previous chapters we saw the learning capabilities of different RBMs, firstly we introduce the Dual Hopfield model ($\Omega_\tau = 1$) and then we investigate the effect of different units prior on the signal retirieval phase (sR). From there we observed that the inference temperature is affected by the architecture, but also that the minimum amount of data (α) necessary to enter in a learning phase is an exclusive property of the teacher ($\hat{\beta}, \Omega_{\hat{\tau}}$). We would like to discover a way to break through the barriers of the learning phase by either lowering the amount of data needed to learn the signal or increasing the learning temperature, thereby expanding the Signal Retrieval (sR) phase area. In recent years, researchers have introduced a theoretical framework that uses coupled replicas [118–121] to understand the effectiveness of training algorithms. This concept has provided convincing evidence [91] that coupling replicas can help in identifying favourable local minima, e.g. in the coupled perceptron in [35]. For these reasons we try to apply collective learning to the Hopfield Teacher-Student scenario of Chapter 2. We couple a number of y Hopfield Students and study the signal retrieval phase of the new system. Since we are interested only in the sR phase we search for the solutions with the example retrieval (eR) parameter $p = 0$.

4.1 Introduction and motivations

We consider a variation of the teacher-student scenario from Chapter 2. Instead of using a single Hopfield network to learn the teacher pattern $\hat{\boldsymbol{\xi}}$, we introduce a system of ferromagnetically coupled, replicated Hopfield networks. This system consists of y replicas, each with its own individual pattern $\boldsymbol{\xi}^u$, where $u = 1, \dots, y$. The learning process is driven by the same teacher, used to generate an extensive number of data $\boldsymbol{\mathcal{S}} = \{\boldsymbol{s}^\mu\}_{\mu=1,\cdots,M}$, as in Chapter 2:

$$P(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\xi}}) = \prod_{\mu=1}^{M} P(\boldsymbol{s}^\mu|\hat{\boldsymbol{\xi}}) = \prod_{\mu=1}^{M} z_\mu^{-1} \exp\left(\frac{\beta}{N}\sum_{i<j}^{N} \hat{\xi}_i\hat{\xi}_j s_i^\mu s_j^\mu\right). \tag{4.1}$$

The Hamiltonian of the new students model is given by

$$H(\{\boldsymbol{\xi}^u\}) = \sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{M} s_i^\mu s_j^\mu \xi_i^u \xi_j^v + \frac{\gamma}{y}\sum_{i<j}\sum_{u<v}\xi_i^u\xi_i^v \tag{4.2}$$

and the corresponding posterior

$$\hat{P}(\boldsymbol{\xi}|\boldsymbol{\mathcal{S}}) = Z^{-1}(\boldsymbol{\mathcal{S}}) \exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{M} s_i^\mu s_j^\mu \xi_i^u \xi_j^v + \frac{\gamma}{y}\sum_{i<j}\sum_{u<v}\xi_i^u\xi_i^v\right). \tag{4.3}$$

Notice that at $\gamma = 0$ we have a sum of independent Hopfield models with the same disorder, that correspond exactly to the Dual Hopfield model. To perform a sanity check, we aim to recover the expected result in the limit $\gamma \to 0$. It should be noted that, in this specific scenario, the use of an incorrect likelihood function places us outside the Bayes-optimal regime. As stated in Chapter 2, by applying the gauge transformation $\xi_i \to \xi_i \hat{\xi}_i$ and $s_i^\mu \to s_i^\mu \hat{\xi}_i$, we can replace the alignment towards $\hat{\boldsymbol{\xi}}$ with the one along $\mathbf{1}$ and the disorder average can be substituted with samples drawn from a CW model, as described by:

$$P(\boldsymbol{\mathcal{S}}|\mathbf{1}) = \prod_{\mu=1}^{M} z_\mu^{-1} \exp\left(\frac{\beta}{N}\sum_{i<j} s_i^\mu s_j^\mu\right). \tag{4.4}$$

In alignment with the analyses of the previous chapters, our focus is on computing the partition function and the quenched free energy density for this model. Specifically, we aim to determine:

$$Z(\boldsymbol{\mathcal{S}}) \;=\; \mathbb{E}_{\{\boldsymbol{\xi}^u\}} \exp\left(\sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{M} s_i^\mu s_j^\mu \xi_i^u \xi_j^v + \frac{\gamma}{y}\sum_{i<j}\sum_{u<v} \xi_i^u \xi_i^v\right) \tag{4.5}$$

$$-\beta f \;=\; \lim_{N\to\infty} \frac{1}{Ny}[\log Z(\boldsymbol{\mathcal{S}})]_{\boldsymbol{\mathcal{S}}} \tag{4.6}$$

where, $\alpha = M/N$ represents the network load. As done previously, we assume that the teacher operates at $\hat{\beta} < 1$ (in the figures of this chapter will be always fixed at $\hat{\beta} = 0.8$), meaning that each example provides no macroscopic useful information about the signal $\hat{\boldsymbol{\xi}}$. We expect that, as the patterns become extensive, the collection of students will enter a retrieval regime. However, the impact of the new parameters, γ and y, is not straightforward and needs to be investigated.
Before going on, a bit of notation. Neuron indeces are denoted with $i, j \in \{1, \dots, N\}$; replica indeces with $u, v \in \{1, \dots, y\}$. Then, we will introduce "fake" replicas (to compute the quenched free energy); they will be labelled with the usual letters $a, b \in \{1, \dots, n\}$.

4.2 Free energy and saddle point equation

The quenched free energy will be computed using a version of the replica symmetry (RS) approximation. The key challenge in this computation arises from the treatment of coupled students. These "real" replicas share the same quenched disorder as the "fake" replicas introduced via the replica trick, while also interacting through explicit pairwise couplings. The analysis follows the mean-field theory approach, introducing a set of order parameters that are evaluated using the saddle-point method in the thermodynamic limit. These parameters, first introduced in Chapter 2 and 3, are defined as:

- Magnetization for each student, towards the signal

$$m_u^a \;=\; \lim_{N,M\to\infty} \langle Q(\boldsymbol{\xi}^{a,u}, \mathbf{1})\rangle\,, \tag{4.7}$$

- Overlap with the examples

$$p_u^{a,\mu} \;=\; \lim_{N,M\to\infty} \langle Q(\boldsymbol{\xi}^{a,u}, \boldsymbol{s}^{\mu})\rangle\,, \tag{4.8}$$

- Pairwise overlap between two students u, v in replicas a, b

$$q_{uv}^{ab} = \lim_{N,M\to\infty} \langle Q(\boldsymbol{\xi}^{a,u}, \boldsymbol{\xi}^{b,v}) \rangle \, . \tag{4.9}$$

The simplest RS ansatz assumes permutation symmetry in the replica space, i.e., $q_{uv}^{ab} = q \, , \forall a, b$. However this was in the case when the "fake" replicas contains non-interacting degrees of freedom. In the present case this assumption fails. For a more accurate description we assume permutation symmetry over the replica space $q_{uv}^{ab} = q \, , \forall a \neq b$. While diagonal blocks ($a = b$) represent the average correlation between students in the same replica. Here, due to the choice of a uniform interaction, we assume a uniform overlap as well, for different students ($u \neq v$), within the same replica a. Under this fashion the diagonal blocks in the full replica space, take the following form $q_{uv}^{aa} = \delta_{uv} + t(1 - \delta_{uv}) \, , \forall a$. Due to the distinct roles of replicas (a, b) and students (u, v), two values of the overlap are distinguished. For the other order parameters we assume: $m_u^a \to m$, while all $p_u^a \to p$, where we suppose all the students align (at low temperatures) with the same example $\boldsymbol{s}^{(1)}$. However for the purposes of the sR analysis we turn off the example alignment, i.e. $p = 0$. Under these assumptions, the free energy can be expressed as a variational principle :

$$-\beta f^{RS}(\alpha, \hat{\beta}, \beta, y, \gamma) = -\beta \, \mathrm{Extr}_{\Lambda, \hat{\Lambda}} \, \hat{f}(\Lambda, \hat{\Lambda}) \, , \tag{4.10}$$

where we introduced the sets of parameters $\Lambda = \{m, q, t\}$, together with the conjugates $\hat{\Lambda}$. To proceed, we define the following quantities:

- $\Delta_T = 1 - \hat{\beta}$, derived from the teacher.
- $\Delta = 1 - \beta - \beta t$, reflecting the impact of student-to-student correlations in the diagonal blocks of the "fake" replica space.
- $\Delta_y = (1 - \beta + \beta q + (y - 1)\beta(q - t))$, accounting for the off-diagonal interactions weighted by the coupling parameter y.

The function $\hat{f}$ to be extremized is defined as

$$\begin{aligned} -\beta \hat{f}(\Lambda, \hat{\Lambda}) = {} & \ln 2 + \ln\cosh(\hat{m}_0) + \frac{\hat{\beta}}{2} m_0^2 + (M-1)\ln 2 - M f_{CW}(\beta_1) - \alpha \hat{m}_0 m_0 + \\ & - n\frac{y(y-1)}{2} \hat{t} t + y^2 \frac{n}{2} \hat{q} q - n y \hat{m} m + \frac{\alpha}{2} n \frac{y\left(\beta q + \hat{\beta}\beta m^2/\Delta_T\right)}{\Delta_y} + \\ & - \frac{\alpha}{2} n \left\{ (y-1)\ln[\Delta_T \, \Delta] + \ln\left(\Delta_T \, \Delta_y\right) \right\} + \\ & + n \left\langle \mathbb{E}_\sigma \ln \left[2\mathbb{E}_{\xi^u} \exp\left\{ \sum_u \xi^u \left(\hat{m} + z\sqrt{\hat{q}} \right) + \frac{1}{2}\left(\hat{t} - \hat{q} + \frac{\gamma}{y} \right) \sum_{u \neq v} \xi^u \xi^v \right\} \right] \right\rangle_z \, , \end{aligned} \tag{4.11}$$

where $\langle . \rangle_z$ denotes the expectation w.r.t. a standard gaussian random variable $z \sim \mathcal{N}(0, 1)$. This framework allows us to analyze the interplay between the mean alignment m, the replica overlap q, and the student-to-student interaction term t. By solving the extremization condition we obtain the following set of equations

$$\hat{m} = \alpha \frac{\hat{\beta}\beta m}{\Delta_T \, \Delta_y} \, , \tag{4.12}$$

$$\hat{q} = \alpha\beta^2 \left(\frac{q + \hat{\beta} m^2/\Delta_T}{\Delta_y^2} \right) \, , \tag{4.13}$$

$$\hat{t} = \frac{\alpha\beta^2(t - q)}{\Delta \, \Delta_y} + \hat{q} \tag{4.14}$$

for the conjugated parameters. The order parameters can be written as the following expectations

$$m \quad = \left\langle \frac{1}{y}\langle \textstyle\sum_u \xi^u\rangle_{\{\xi^u\}|z,s}\right\rangle_{z,s}, \tag{4.15}$$

$$q \quad = \left\langle \frac{1}{y^2}\langle \textstyle\sum_u \xi^u\rangle^2_{\{\xi^u\}|z,s}\right\rangle_{z,s}, \tag{4.16}$$

$$t \; = \left\langle \frac{1}{y(y-1)}\langle \textstyle\sum_{u<v} \xi^u\xi^v\rangle_{\{\xi^u\}|z,s}\right\rangle_{z,s}. \tag{4.17}$$

In the above expressions $\langle.\rangle_{\xi|z,s,\hat{\xi}}$ stands for the expectation over

$$P(\{\xi^u\})\,\mathbb{E}_{\{\xi^u\}}\exp\left\{\sum_u \xi^u\Big(\hat{m}+z\sqrt{\hat{q}}\Big)+\frac{1}{2}\sum_{u\neq v}\Big(\frac{\gamma}{y}+\hat{t}-\hat{q}\Big)\xi^u\xi^v\right\}, \tag{4.18}$$

where $P(\{\xi^u\})$ stands for the product measure of the binary distribution for all the students' patterns. It is possible to appreciate that the internal average is nothing but a Curie-Weiss model with couplings $J = \frac{\gamma}{y}+\hat{t}-\hat{q}$ and external field $h = \hat{m}+z\sqrt{\hat{q}}$. With the sets (4.12-4.14) and (4.15-4.17), we can determine the phase behavior of the system under different temperature and dataset size regimes.

4.3 Retrieval effect of the coupled students

As we already said in the introduction we focus exclusively on the sR and SG neighboring phases because the parameter p signals the presence of spurious memorization inference states, when $\hat{\beta} < 1$.

The magnetization toward the signal is illustrated in Fig. 4.1. Increasing both y and γ has a similar effect: the inference temperature of the system rises as the values of these parameters increase. Interestingly, we observed that the learning performance reaches a maximum, similar to what is depicted in Fig. 2.5. This behavior is further demonstrated in the accompanying images.

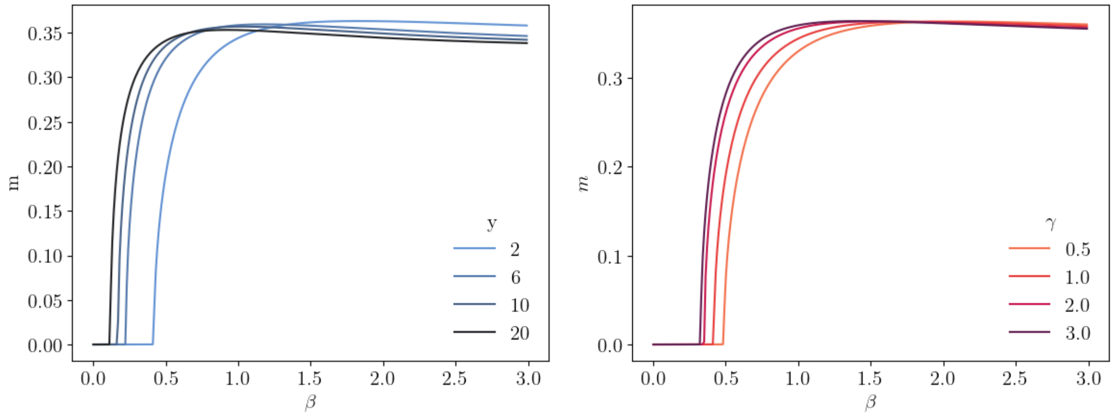


Figure 4.1: **Left**: Magnetization of the coupled Hopfield system for different values of the inverse students' temperature β. The amount of data given to the system are $\alpha = 0.12$. A transition occurs while lowering the temperature. The step towards a magnetized phase moves towards higher temperature the more the system size y increases. **Right**: Magnetization of the coupled Hopfield system for different values of the coupling strenght γ. Also here the amount of data is $\alpha = 0.12$ and we can see a transition towards the magnetized phase is happening for lower temperature. It is possible to appreciate the same affect as for y, i.e. the more the system is coupled the lower is the regularization in order to enter the retrieval phase.

The transition points shown in the previous figure can also be derived analytically. As in the earlier chapters, at very high inference temperatures ($\beta \sim 0$), the effective Curie-Weiss internal distribution in (4.1) simplifies as follows:

$$P(\{\xi^u\})\,\mathbb{E}_{\{\xi^u\}} e^{\frac{1}{2}\sum_{u\neq v}\left(\frac{\gamma}{y}+\hat{t}\right)\xi^u\xi^v}\left[\sum_u \xi^u\left(\hat{m}+z\sqrt{\hat{q}}\right)-\sum_{u\neq v}\hat{q}\,\xi^u\xi^v\right].$$

In this regime, we assume that the order parameters $m = q \sim 0$ and that a continuous phase transition occurs as the temperature decreases. Given that the order parameters are small, the same assumption applies to their conjugates. By expanding the equation for the magnetization toward the signal, we obtain

$$m = \langle\langle\sum_u \xi^u\rangle_{\{\xi^u\}|z,s}\rangle_{z,s} = \frac{\alpha\hat{\beta}\beta m}{(1-\hat{\beta})\,(1-\beta-\beta\,t(y-1))}\,\Omega_y\left(\frac{1}{y}\sum_{u,v}\xi^u\xi^v\right)+\mathcal{O}(m)\,, \tag{4.19}$$

where we use the fact that $\Omega_y\left(\sum_u \xi^u\right)=0$ and that the average $\Omega_y\left(g(\{\xi^u\})\right)$ of a function g of the y-student patterns is computed over the Curie-Weiss model at zero external field

$$\Omega_y\left(g(\{\xi^u\})\right) = Z^{-1}_{CW_y}\sum_{\{\xi^u\}} e^{\frac{1}{2}\sum_{u\neq v}\left(\frac{\gamma}{y}+\hat{t}\right)\xi^u\xi^v} g(\{\xi^u\})\,, \tag{4.20}$$

$$Z_{CW_y} = \sum_{\{\xi^u\}} e^{\frac{1}{2}\sum_{u\neq v}\left(\frac{\gamma}{y}+\hat{t}\right)\xi^u\xi^v}\,. \tag{4.21}$$

In this regime also the Eqs.(4.17-4.14) simplifies:

$$t = \langle\frac{1}{y(y-1)}\sum_{u<v}\xi^u\xi^v\rangle_{CW_y}\,, \tag{4.22}$$

$$\hat{t} = \frac{\alpha\beta^2 t}{[(1-\beta)-(y-1)\beta\,t)]\,(1-\beta+\beta t)}\,, \tag{4.23}$$

then the equation for the P-sR transition results as

$$\beta_y^{sR} = \frac{(1-\hat{\beta})(1-\beta\,t(y-1)-\beta)}{\alpha\hat{\beta}(1+(y-1)\,t)}\,, \tag{4.24}$$

where we use $\langle\sum_{u,v}\xi^u\xi^v\rangle_{CW_y} = y + y(y-1)t$. Eq. (4.24) can be solved numerically along with Eqs.(4.22-4.23).
A Spin-Glass transition line can also be derived. In this case we impose $m = 0$ and $q \sim 0$ and follow the same steps as above. Expanding for small q and $\hat{q}$ in (4.16) we obtain

$$q = \langle\langle\sum_u \xi^u\rangle^2_{\{\xi^u\}|z,s}\rangle_{z,s} = \frac{\alpha\beta^2 q}{(1-\beta-\beta\,t(y-1))^2}\,\langle\frac{1}{y}\sum_{u,v}\xi^u\xi^v\rangle^2_{CW_y}+\mathcal{O}(q)\,, \tag{4.25}$$

leading to the equation for the P-SG transition line

$$\beta_y^{SG} = \sqrt{\frac{(1-\beta\,t(y-1)-\beta)^2}{\alpha(1+(y-1)\,t)^2}}\,. \tag{4.26}$$

It is evident that when $\gamma = 0$, the sR and SG transition lines reduce to the results for a single Dual Hopfield model (2.69, 2.71). This is expected, as setting the coupling strength to zero yields a set of independent Hopfield models that behave exactly as described in Chapter 2.
In the following, we illustrate these effects using phase diagrams. Figure 4.2 shows the impact of increasing the number of students. As evident, the ferromagnetic interactions among students extend the range of inference temperatures compared to the original model in Chapter 2 ($y = 1$). However, the critical load $\alpha_c \sim 0.06$ remains independent of the number of interacting students in the model.

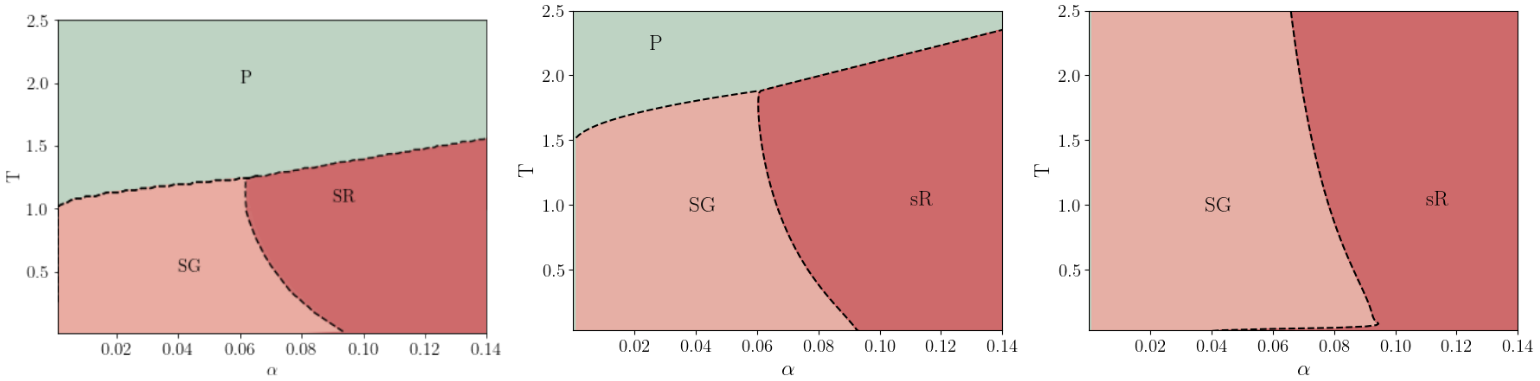


Figure 4.2: Different Phase diagram showing the sR phase for systems containing different students. The left panel represent the Dual Hopfield model of Chapter 2 for a direct compharison. The other two represent respectively $y = 2$ and $y = 10$ coupled students. The coupling strenght as well as the teacher temperature are fixed $\hat{\beta} = 0.8$, $\gamma = 1.0$. Augmenting the number of interacting students increases the retrieval temperature towards the signal. However the minimal amount of data remains the same as the one of one single Hopfield student.

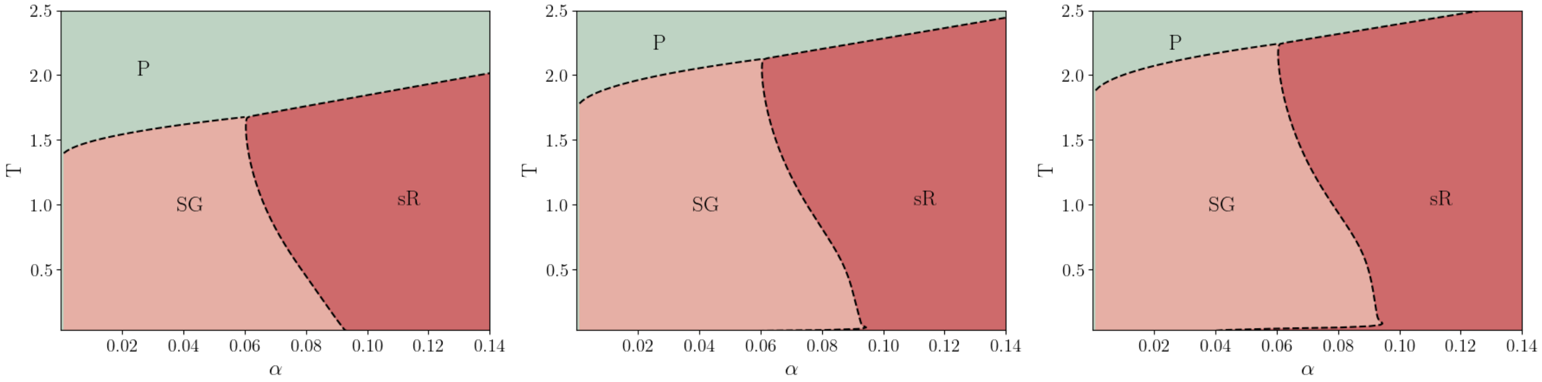


Figure 4.3: Different Phase diagram showing the sR phase for systems containing different students $y = 2, 6, 10$. The coupling strenght as well as the teacher temperature are fixed $\hat{\beta} = 0.8$, $\gamma = 1.0$, the latter is rescaled as in the following $\gamma \to \beta\gamma$. The rescaling contains the increase of the retrieval temperature towards the signal.

In Fig. 4.3, we applied a rescaling of the coupling parameter $\gamma \to \beta\gamma$. This rescaling aligns the scaling in β of the coupling strength with the one of the conjugate order parameters, appearing in the external field h in Eqs. (4.15-4.22). Additionally, this approach mitigates the influence of the ferromagnetic term, allowing us to observe more clearly how the increase in inference temperature behaves.
It becomes apparent that the system of students cannot infer at arbitrarily high temperatures. The

model inherently sets a barrier that is rapidly saturated. A similar phenomenon was observed in Chapter 3, where the self-overlap constraint determined the critical point P_c. The limiting P-sR line is further analyzed in Sec. 4.4, where the behavior in the limit $y \to \infty$ is explored.

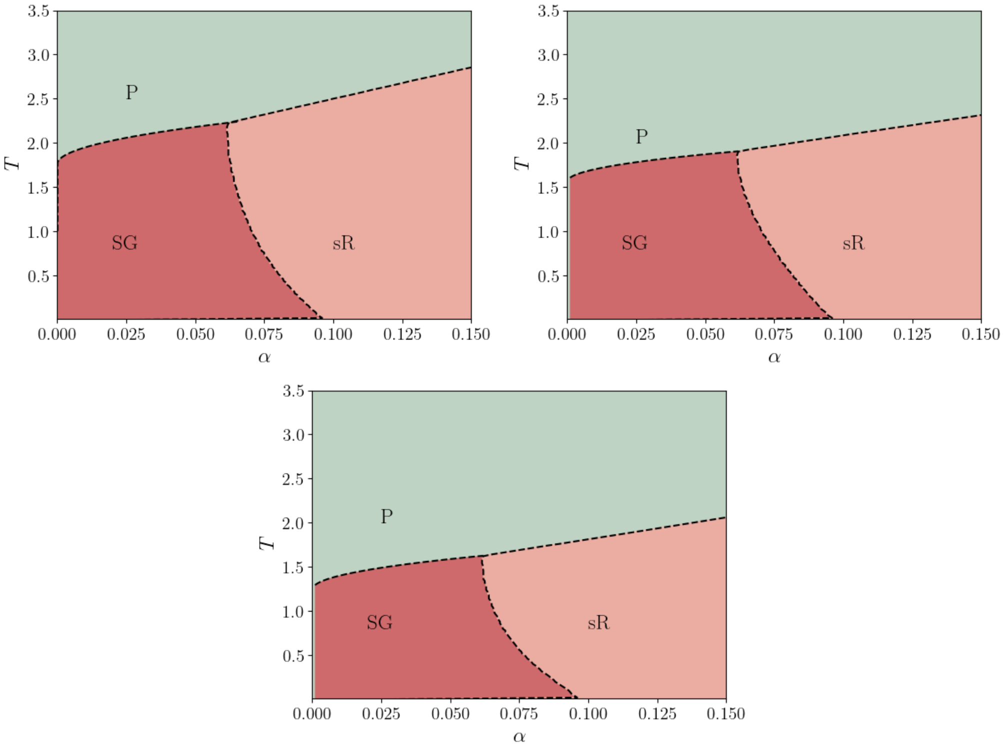


Figure 4.4: Phase diagram showing the effet of the coupling interaction $\gamma = 0.5, 2$. The bottom image is the starting point at $\gamma = 0.5$, while the other two are: Left $\gamma = 2$ without rescaling of the coupling constant, Right: same but with the rescaling $\gamma \to \beta\gamma$. Here the number of coulped students is $y = 2$

The effect of varying the coupling strength γ is illustrated in Fig.4.4. As γ increases, the learning process becomes more effective. This phenomenon can be attributed to the ferromagnetic coupling, which facilitates the sharing of information about the retrieved signal between the students. Consequently, the system transitions into the sR phase at higher temperatures, enhancing inference performance.

4.4 Infinite number of students

As discussed in the previous section and illustrated in Fig. 4.5, the interaction of many students leads to an increase in the learning temperature. However, this increment is not unbounded, as the gap between the transition lines decreases as y grows. The aim of this section is to analyze the limit $y \to \infty$ and derive the limiting transition lines, $\beta_\infty(\alpha, \gamma, \hat{\beta})^{SG}$ and $\beta_\infty(\alpha, \gamma, \hat{\beta})^{sR}$. We begin by examining the terms within the averaging distribution (4.18)

$$\sum_u \left(\hat{m} + \sqrt{\hat{q}}z\right)\xi_u + \frac{1}{2}\left(\frac{\gamma}{y} + \hat{t} - \hat{q}\right)\sum_{u \neq v} \xi^u \xi^v \,.$$

The first terms, as well as the one proportional to γ/y, are extensive in y, while the remaining part is of $\mathcal{O}(y^2)$. To preserve the linear scaling with the system size in the internal effective Curie-Weiss model, i.e., $\mathcal{O}(y)$, we assume the combination $\hat{t} - \hat{q} = \frac{\hat{\delta}}{y}$. The same assumption applies to the conjugate variables t and q.

$$t = \int Dz\, \Omega(\xi_1\xi_2) \quad q = \int Dz\, \Omega^2(\xi_1)$$

$$t - q = \int Dz\, \left(\Omega(\xi_1\xi_2) - \Omega^2(\xi_1)\right) \to_{y\to\infty} \int Dz\, \frac{\mathcal{L}(z)}{y} = \frac{\delta}{y}\,, \tag{4.27}$$

where, with $\Omega_y(O)$ we are referring to the average over the CW model:

$$\Omega(O)_y = \frac{\mathbb{E}_{\{\xi^u\}} O \exp\left\{ \sum_u \xi^u \left(\hat{m} + z\sqrt{\hat{q}}\right) + \frac{1}{2}\sum_{u\neq v}\left(\frac{\gamma}{y} + \hat{t} - \hat{q}\right)\xi^u\xi^v \right\}}{\mathbb{E}_{\{\xi^u\}} \exp\left\{ \sum_u \xi^u \left(\hat{m} + z\sqrt{\hat{q}}\right) + \frac{1}{2}\sum_{u\neq v}\left(\frac{\gamma}{y} + \hat{t} - \hat{q}\right)\xi^u\xi^v \right\}}\,.$$

Then we reparametrize the free energy according to δ and $\hat{\delta}$

$$\begin{aligned} f(q, m, \delta; \hat{q}, \hat{m}, \hat{\delta}) = & \frac{\hat{q}\delta}{2\beta} + \frac{q\hat{\delta}}{2\beta} - \frac{\hat{q}q}{2\beta} + \frac{\hat{m}m}{\beta} + \frac{\hat{q}}{2\beta} + \\ & + \frac{\alpha}{2\beta}\left\{ \ln[\Delta_T\Delta] - \frac{\left(\beta q + \hat{\beta}\beta m^2/\Delta_T\right)}{\Delta_y} \right\} + \\ & + \frac{1}{\beta}\left\langle \mathbb{E}_\sigma \left(-\frac{J\bar{M}^2}{2} + \log 2 + \log\cosh\left\{\left(h(z) + J\bar{M}\right)\right\} \right) \right\rangle_z \end{aligned} \tag{4.28}$$

with

$$\begin{aligned} h(z) =& \hat{m} + z\sqrt{\hat{q}} \\ J =& \frac{\delta + \gamma}{y} \\ \bar{M}(\sigma, z, \hat{\delta}, \hat{q}, \hat{m}, \hat{p}) =& \tanh\left(h(z) + J\bar{M}\right)\,. \end{aligned}$$

From Eq.(4.27), the term inside the Gaussian average corresponds to the correlation of the CW model. As discussed in Chapter 1, these correlations vanish at infinite size, scaling as $1/y$. However, we require an estimation of the prefactor δ, which can be determined using the Linear Response theory, specifically through Eq.(1.29). This leads to an updated formulation of the saddle point equations (4.15 - 4.17)

$$\begin{aligned} m =& \mathbb{E}_\sigma \int Dz\bar{M} \\ q =& \mathbb{E}_\sigma \int Dz\bar{M}^2 \\ \delta =& \mathbb{E}_\sigma \int Dz\frac{(1-\bar{M}^2)^2 J}{1 - J(1-\bar{M}^2)}\,. \end{aligned}$$

With this new set of equations, it is possible to see that the transition lines don't grow indefinitely with the number of students. Below, we provide the expansions of the Δ terms, which were initially introduced. These expansions are derived in the $y \to \infty$ limit

$$\Delta_y = 1 - \beta + \beta q + (y-1)\beta(q-t) \simeq_{y\to\infty} 1 - \beta(1-q) - \beta\delta\,,$$
$$\Delta = 1 - \beta + \beta t \simeq_{y\to\infty} 1 - \beta(1-q)\,.$$

Our purpose is to find the transition lines and observe they are y-independent. We proceed once again by working with small values of m and q, as we did in the steps leading to Eq. (4.24). In this regime, the relevant parameters are

$$\delta = \frac{\hat{\delta} + \gamma}{1 - (\hat{\delta} + \gamma)}\,, \tag{4.29}$$

$$\hat{\delta} = \frac{\alpha\,\beta^2\delta}{(1-\beta)\Delta_y|_{q=0}}\,, \tag{4.30}$$

$$\beta = \frac{\Delta_y|_{q=0}\,\Delta_T}{(\delta+1)(\alpha\,\hat{\beta})}\,, \tag{4.31}$$

this set of equations gives an analytical expression for the critical line

$$\beta^{sR}_{\infty} = \left[\frac{2(\alpha-1)\hat{\beta}+2}{\hat{\beta}^2(\alpha+\gamma-1) - \sqrt{\left(\hat{\beta}^2(\alpha+\gamma-1) - \hat{\beta}\gamma + 1\right)^2 - 4(\hat{\beta}-1)\hat{\beta}(\gamma-1)((\alpha-1)\hat{\beta}+1)} - \hat{\beta}\gamma + 1}\right]^{-1}, \tag{4.32}$$

which can be found also after the rescaling $\gamma \to \beta\gamma$

$$\beta^{sR}_{\infty} = \left[\frac{2(\hat{\beta}(\alpha-\hat{\beta}\gamma+\gamma-1)+1)}{\hat{\beta}^2(\alpha-\gamma-1) - \sqrt{(\hat{\beta}(\hat{\beta}(\alpha-\gamma-1)+\gamma)+1)^2 - 4(\hat{\beta}-1)\hat{\beta}(-\alpha\hat{\beta}+(\hat{\beta}-1)\hat{\beta}\gamma+\hat{\beta}-1)} + \hat{\beta}\gamma + 1}\right]^{-1}. \tag{4.33}$$

On the other hand, the P-SG transition line β^{sR}_{∞}, can be found only numerically, solving together (4.29-4.30) and the following

$$\frac{\alpha\beta^2}{(\delta+1)^2\Delta_y|^2_{q=0}} = 1\,. \tag{4.34}$$

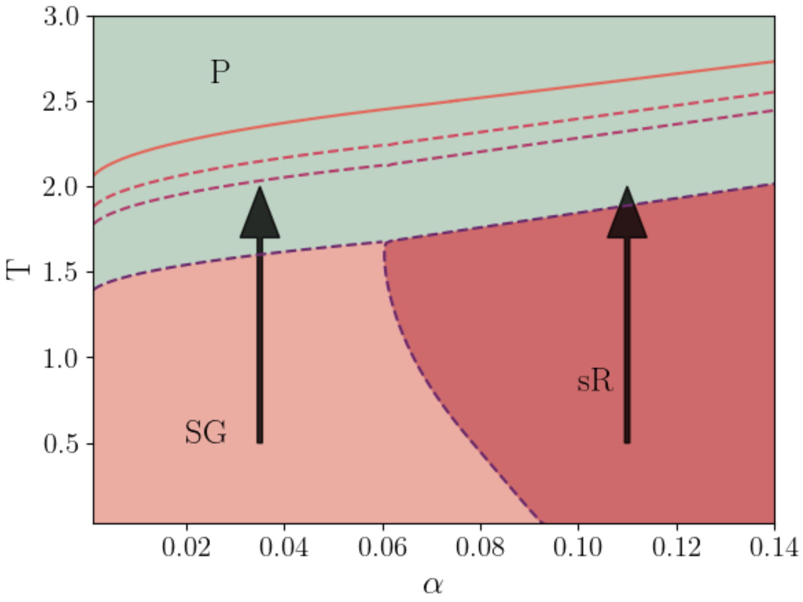


Figure 4.5: P-SG, P-sR transition lines for different number of coupled students $y = 2, 6, 10$, at fixed coupling $\gamma = 1.0$. The dashed lines represent the transition for y finite, while the solid line is the one for $y \to \infty$, obtained from Eq.(4.33)

4.5 Discussion

In the presented study we extensively use all the tools and techniques presented in Chapters 1-3 to provide the equilibrium behavior of a system of interacting Hopfield networks within the Teacher-Student inverse problem. With the use of such tools, we obtain a mean-field analysis that displays the behavior of the sR phase of such system. Even if we see the amount of data is not effected by the applied interaction, $\alpha_c \sim 0.06$, we discovered that interaction with more students enhances the regularization barrier, triggering the learning behavior. The more students are interacting, the more the collective learning enhances the learning temperature T, Fig.4.2-4.4. However, this improvement cannot be extended reaching an infinitely high inference temperature. From the ML side, this would mean that the Hopfield-RBM system could have already learned (presumably) the features from the dataset without even undergoing the training process. The limiting barrier has an analytical expression β_∞^{sR} and can be seen in Fig.4.5 when coupling an infinite amount of students. This effect is similar to the one seen in Chapter 3, but stronger as one can see by compharison between Fig.3.5 and Fig.4.2. This could suggest that other types of interaction with data are needed in order to produce fluctuations in α_c.

Appendix E

Derivation of Quenched free energy

We discuss how to compute the averaged quenched free energy (4.11) ad, consequently, (4.28). As already seen in the appendices of the previous chapters, a standard procedure in spin-glass models with quenched disorder, involves introducing a number n of replicas, with the limit $n \to 0$ taken afterwards. These replicas differ from those in the Hamiltonian (4.2) in that they are independent and there is no explicit coupling between them. In the following derivation, we explicitly distinguish between the two sets. We use indices $a, b \in \{1, \cdots, n\}$ to denote "fake" replicas and $u, v \in \{1, \cdots, y\}$ to index the students (i.e., the real replicas in the original Hamiltonian). By replicating all degrees of freedom ξ_i^u we can express the disorder average of the n^{th} power partition function

$$[Z^n]^{\boldsymbol{S}} = \sum_{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^n} \sum_{\boldsymbol{s}} e^{\frac{\hat{\beta}N}{2} m_0(\boldsymbol{s})^2} \frac{2^{N(M-1)}}{z(\hat{\beta})^M} \det\left(\Xi(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{t})\right)^{(1-M)/2} . \tag{E.1}$$

This expression can be derived by following the same steps as those used from Eq. (B.5) to (B.8), where m_0 corresponds to the magnetization of a representative teacher sample.
The only differences are related to $p_i^a = 0$, as we are only interested in the signal solutions, along with the presence of the real replicas. The major changes involve the following determinant

$$\det\left(\Xi(\boldsymbol{q}, \boldsymbol{m})\right)^{-1/2} := \int \prod_{a,u}^{n} Dz^{au} Dz \, \exp\left\{ \frac{\beta}{2} \left[\sum_{a \neq b} \sum_{u,v} z_u^a z_v^b \, q_{uv}^{ab} + \sum_{a,u} (z_u^a)^2 \right] + \frac{\hat{\beta}}{2} z^2 + \sqrt{\hat{\beta}\beta} z \sum_{a,u} z_u^a m_u^a \right\} . \tag{E.2}$$

where we have introduced the quantities

$$q_{uv}^{ab} = \frac{1}{N} \sum_i \xi_i^{au} \xi_i^{bv} , \qquad m_u^a(\boldsymbol{\xi}) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{au} , \tag{E.3}$$

which are an extended variation of the ones in (B.7) in Chapter 2. The set of order parameters represents the overlap between student u (in replica a) with the teacher signal (m_u^a) and the two-replica overlap between two student vectors u, v (q_{uv}^{ab}). The overlap matrix can be visualized in a $n \times n$ block matrix fashion

$$K = \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1n} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22} & \cdots & \mathcal{Q}_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \mathcal{Q}_{n1} & \cdots & \cdots & \mathcal{Q}_{nn} \end{pmatrix} , \tag{E.4}$$

with each block $\mathcal{Q}_{uv}$ having $y \times y$ entries, e.g.

$$\mathcal{Q}_{11} = \begin{pmatrix} q^{11}_{11} & q^{11}_{12} & \cdots & q^{11}_{1y} \\ q^{11}_{21} & q^{11}_{22} & \cdots & q^{11}_{2y} \\ \vdots & \ddots & \ddots & \vdots \\ q^{11}_{y1} & \cdots & \cdots & q^{11}_{yy} \end{pmatrix} . \tag{E.5}$$

From the definition of in (E.3) it is possible to count the number of independent overlaps as $\binom{ny}{2} = \frac{n^2y^2}{2} - \frac{ny}{2}$. We can proceed enforcing the definitions (E.3) in the partition function, using the corresponding densities of states (see Chapter 2): $\mathcal{D}(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{t})$, $\mathcal{D}(m_0)$

$$[Z^n]^{\boldsymbol{\mathcal{S}}} = \int dm_0 d\hat{m}_0 \int d\boldsymbol{q} d\boldsymbol{m} d\boldsymbol{t}\, d\hat{\boldsymbol{q}} d\hat{\boldsymbol{m}} d\hat{\boldsymbol{t}}\, e^{-\beta N \hat{f}} , \tag{E.6}$$

$$\hat{f} = \hat{f}^{(0)}(m_0, \hat{m}_0) + \hat{f}^{(1)}_n(\boldsymbol{q}, \boldsymbol{m}; \hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}) \tag{E.7}$$

$$-\beta \hat{f}^{(0)}(m_0, \hat{m}_0) = \frac{\hat{\beta}}{2} m_0^2 + (\alpha - 1) \ln 2 - \alpha f_{CW} - \hat{m}_0 m_0 \; + \ln 2, \tag{E.8}$$

$$-\beta \hat{f}^{(1)}_n(\boldsymbol{q}, \boldsymbol{m}; \hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}) = \ln \mathbb{E}_s e^{\hat{m}_0 s} \sum_{\xi^1,\dots,\xi^n} \exp\left(\sum_{u<v} \sum_a \hat{q}^{aa}_{uv}\, \xi^u_a \xi^v_a + \sum_{u,v} \sum_{a,b} \hat{q}^{ab}_{uv}\, \xi^u_a \xi^v_b + \sum_a \hat{m}^a_u\, \xi^u_a + \frac{\gamma}{y} \sum_a \sum_{u<v} \xi^u_a \xi^v_a \right) + \tag{E.9}$$

$$- \sum_a \sum_{u<v} \hat{q}^{aa}_{uv} q^{aa}_{uv} - \sum_{a<b} \sum_{u,v} \hat{q}^{ab}_{uv} q^{ab}_{uv} - \sum_{a,u} \hat{m}^a_u m^a_u - \frac{\alpha}{2} \ln\det\big(\Xi(\boldsymbol{q}, \boldsymbol{m})\big) . \tag{E.10}$$

The free energy is obtained by the following replica trick

$$-\beta f(\beta, \hat{\beta}, \alpha) = \lim_{N\to\infty} \frac{1}{N} [\ln Z]^{\boldsymbol{\mathcal{S}}} = \lim_{\substack{N\to\infty \\ n\to 0}} \frac{\ln [Z^n]^{\boldsymbol{\mathcal{S}}}}{Nn}$$

$$-\beta f \approx \lim_{\substack{N\to\infty \\ n\to 0}} \frac{1}{N y\, n} \ln \left(e^{N \,\mathrm{Extr}\, \hat{f}(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{m_0}; \hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{m}}_{\boldsymbol{0}})} \right) \approx \lim_{n\to 0} \frac{1}{y\, n} \,\mathrm{Extr}\, \hat{f}(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{m_0}; \hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{m}}_{\boldsymbol{0}}) \tag{E.11}$$

where we already factorized over the number of components $i = 1, \cdots, N$. Since $\hat{\beta} < 1$, $\hat{f}^{(0)}$ can be easily extremized, being the CW free energy at high temperature. By extremization we obtain $m_0 = 0$; which implies $\hat{f}^{(0)} = 0$, as it should be in order to prevent divergences of the exponential measure, when $n \to 0$.

The evaluation of the equilibrium behavior is performed under a Replica Symmetry (RS) ansatz, which also involves the student space. The simplest choice is to assume permutation symmetry over both the replica and student spaces. We can make this assumption without concern for the ”fake” replicated space, as it represents the most straightforward choice in any spin-glass model [11]. On the other hand, a symmetry ansatz also arises for the pairwise correlations between students, i.e., the overlaps q^{aa}_{uv}. This symmetry emerges from the fully connected topology with uniform interaction strengths between the students' weight vectors, as assumed in the Hamiltonian (4.2). Under this extended RS ansatz, we have only three order parameters (and their conjugates), namely:

$$m^u_a \;= m \qquad \hat{m}^u_a \;= \hat{m} \qquad \forall_{u,a} \tag{E.12}$$

$$q^{aa}_{uv} \;= t \qquad \hat{q}^{aa}_{uv} \;= \hat{t} \qquad \forall_{u\neq v,a} \tag{E.13}$$

$$q^{ab}_{uv} \;= q \qquad \hat{q}^{ab}_{uv} \;= \hat{q} \qquad \forall_{u,v,a\neq b} \tag{E.14}$$

The meaning of the symmetry can alternatively be expressed as follows: within the same replicated system, the interaction between different students is the same (since it is an exact copy, this seems a reasonable assumption). For what concern the overlap, we can interpret the separation of t and q, in the following way. The parameter t indicates that, within the same replicated system, the interaction between different students is identical. For q, we are stating that the interaction between the same students in different replicated systems is the same. Moreover, it is also equal to the interaction between different students in different replicated systems. This is expressed as:

$$
\begin{aligned}
\xi_i^{au}\xi_i^{bu} &= \xi_i^{a'u}\xi_i^{b'u}\,,\\
\xi_i^{au}\xi_i^{bv} &= \xi_i^{a'u}\xi_i^{b'v}\,.
\end{aligned}
$$

The fact that $\xi_i^{au}\xi_i^{bu}=\xi_i^{au}\xi_i^{bv}$ can be interpreted as meaning that the student u in replica a does not distinguish between being replicated in replica b or being another student in replica b.
With this choice we can update $\hat{f}_n^{(1)}$ of Eq.(E.9). Let start first with the explicit computation of the $\ln\det\big(\Xi(q,t,m)\big)$:

$$
\begin{aligned}
\det\big(\Xi(q,t,m)\big) &= \int\prod_{au}^{ny} Dz_{au}Dz\exp\left(\frac{\beta t}{2}\sum_{u\neq v}\sum_{a} z_{au}z_{av}+\frac{\beta q}{2}\sum_{a\neq b}\sum_{u,v} z_{au}z_{bv}+\frac{\beta}{2}\sum_{a} z_{au}^2+\frac{\hat{\beta}z^2}{2}+\sqrt{\hat{\beta}\beta}m\sum_{au} z_{au}z\right)\\
&= \frac{1}{\sqrt{1-\hat{\beta}}}\int\prod_{au}^{ny} Dz_{au}\exp\Big(\frac{\hat{\beta}\beta m^2}{2(1-\hat{\beta})}\sum_{a,b}\sum_{u,v} z_{au}z_{bv}+\frac{\beta t}{2}\sum_{u\neq v}\sum_{a} z_{au}z_{av}+\frac{\beta q}{2}\sum_{a\neq b}\sum_{u,v} z_{au}z_{bv}+\\
&\qquad +\frac{\beta}{2}\sum_{a}\left(z_{au}\right)^2\Big)\,.
\end{aligned}
$$

As it can be seen, the form of the matrix in $\det\big(\Xi(q,t,m)\big)$ is

$$
\Xi(q,t,m)=\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & \cdots & \boldsymbol{B}\\ \boldsymbol{B} & \boldsymbol{A} & \cdots & \boldsymbol{B}\\ \vdots & \ddots & \ddots & \vdots\\ \boldsymbol{B} & \cdots & \cdots & \boldsymbol{A}\end{pmatrix} \tag{E.15}
$$

$$
\begin{aligned}
A_{ij} &= \delta_{ij}\,a_0+(1-\delta_{ij})a_1\\
B_{ij} &= \delta_{ij}b_0+(1-\delta_{ij})b_1\,,
\end{aligned}
$$

where $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{y\times y}$ and $\Xi\in\mathbb{R}^{ny\times ny}$. In our case, the entries are respectively $a_0=1-\frac{\hat{\beta}\beta m^2}{(1-\hat{\beta})}-\beta$, $a_1=-\beta t-\frac{\hat{\beta}\beta m^2}{(1-\hat{\beta})}$ and $b_0=b_1=-\frac{\hat{\beta}\beta m^2}{(1-\hat{\beta})}-\beta q$.
Using the following definitions: $\Delta_a=a_0-a_1$, $\Delta_b=b_0-b_1$ and $\Delta_{ab}=a_1-b_1$, the log-determinant of a block matrix with these characteristics, in the $n\to 0$ limit is

$$
\begin{aligned}
\log\det\Xi &\approx n(y-1)\left\{\log[\Delta_a-\Delta_b]+\frac{\Delta_b}{\Delta_a-\Delta_b}\right\}+\\
&\quad +n\log[\Delta_a-\Delta_b+y\Delta_{ab}]+n\frac{\Delta_b+y\,b_1}{\Delta_a-\Delta_b+y\Delta_{ab}}\\
&\approx n(y-1)\log[(1-\hat{\beta})(1-\beta+\beta t)]+n\log\Big((1-\hat{\beta})\,[(1-\beta+\beta q)+(y-1)\beta(q-t)]\Big)+\\
&\quad -n\frac{y\left(\beta q+\hat{\beta}\beta m^2/(1-\hat{\beta})\right)}{1-\beta+\beta q+(y-1)\beta(q-t)}\,.
\end{aligned}
$$

Then we can update (E.9) as

$$-\beta\hat{f}^{(1)}_n(q,t,m;\hat{q},\hat{t},\hat{m}) = \ln\sum_{\{\boldsymbol{\xi}^{au}\}}\exp\left(\hat{t}\sum_a\sum_{u<v}\xi^u_a\xi^v_a + \hat{q}\sum_{a<b}\sum_{u,v}\xi^u_a\xi^v_b + \hat{m}\sum_{a,u}\xi_{au} + \frac{\gamma}{y}\sum_a\sum_{u<v}\xi^u_a\xi^v_a\right) + \tag{E.16}$$

$$- n\frac{y(y-1)}{2}\hat{t}t - y^2\frac{n(n-1)}{2}\hat{q}q - ny\hat{m}m + \frac{\beta}{2}nym^2 + \frac{\alpha}{2}n\left[\frac{y\left(\beta q + \hat{\beta}\beta m^2/(1-\hat{\beta})\right)}{1-\beta+\beta q+(y-1)\beta(q-t)}\right] + \tag{E.17}$$

$$- \frac{\alpha}{2}n\left\{(y-1)\ln[(1-\hat{\beta})(1-\beta+\beta t)] + \ln\left((1-\hat{\beta})\left[(1-\beta+\beta q)+(y-1)\beta(q-t)\right]\right)\right\} \tag{E.18}$$

Under the RS approximation $\hat{f}^{(1)}_n$ is linear in n. To see this we have to elaborate more on the first term of (E.9). Followig the same steps of Appendix B we need just to linearize the quadratic term $\hat{q}\sum_{a<b}\sum_{u,v}\xi^u_a\xi^v_b$, obtaining

$$\begin{aligned}
&\log\sum_{\{\boldsymbol{\xi}^{au}\}}\exp\left(\hat{t}\sum_a\sum_{u<v}\xi^u_a\xi^v_a + \hat{q}\sum_{a<b}\sum_{u,v}\xi^u_a\xi^v_b + \hat{m}\sum_{a,u}\xi_{au} + \frac{\gamma}{y}\sum_a\sum_{u<v}\xi^u_a\xi^v_a\right) \approx \qquad \text{(E.19)}\\
&\approx \log e^{-ny\frac{\hat{q}}{2}}\left(1 + n\int\mathcal{D}z\,\log\left[2^y\,\mathbb{E}_{\xi^u}\exp\left\{\sum_u\xi^u\left(\hat{m}+z\sqrt{\hat{q}}\right) + \frac{1}{2}\left(\hat{t}-\hat{q}+\frac{\gamma}{y}\right)\sum_{u\neq v}\xi^u\xi^v\right\}\right]\right)\\
&\approx -ny\frac{\hat{q}}{2} + n\left\langle\log\left[2^y\,\mathbb{E}_{\xi^u}\exp\left\{\sum_u\xi^u\left(\hat{m}+z\sqrt{\hat{q}}\right) + \frac{1}{2}\left(\hat{t}-\hat{q}+\frac{\gamma}{y}\right)\sum_{u\neq v}\xi^u\xi^v\right\}\right]\right\rangle_z ,
\end{aligned}$$

where in the last step we use the properties of log and expand it for small n, finally we substitute with $\langle\cdot\rangle_z$ the Normal average.

Recollecting all together we have that $\hat{f}^{(1)}_n$ is linear in n. This allow us to search for the free energy as in the following

$$-\beta f \approx \text{Extr}\,\frac{1}{y}\hat{f}^{(1)}(q,m,t;\hat{q},\hat{m},\hat{t}) \tag{E.20}$$

$$\begin{aligned}
-\beta\hat{f}^{(1)}(q,t,m;\hat{q},\hat{t},\hat{m}) &= \left\langle\log\left[\mathbb{E}_{\xi^u}\exp\left\{\sum_u\xi^u\left(\hat{m}+z\sqrt{\hat{q}}\right) + \frac{1}{2}\left(\hat{t}-\hat{q}+\frac{\gamma}{y}\right)\sum_{u\neq v}\xi^u\xi^v\right\}\right]\right\rangle_z + \qquad \text{(E.21)}\\
&+ y\log 2 - y\frac{\hat{q}}{2} - \frac{y(y-1)}{2}\hat{t}t + \frac{y^2}{2}\hat{q}q - y\hat{m}m + \frac{\alpha}{2}\left[\frac{y\left(\beta q + \hat{\beta}\beta m^2/(1-\hat{\beta})\right)}{1-\beta+\beta q+(y-1)\beta(q-t)}\right]\\
&- \frac{\alpha}{2}\left\{(y-1)\ln[(1-\hat{\beta})(1-\beta+\beta t)] + \ln\left((1-\hat{\beta})\left[(1-\beta+\beta q)+(y-1)\beta(q-t)\right]\right)\right\} ,
\end{aligned}$$

which is exactly Eq.(4.11) in the case $p = 0$. The equations to be fullfilled, obtaining the extremization are

$$\begin{aligned} \hat{m} &= \alpha \frac{\hat{\beta}\beta m}{\Delta_T\,\Delta_y}\,, \\ \hat{q} &= \alpha\beta^2 \left(\frac{q + \hat{\beta} m^2/\Delta_T}{\Delta_y^2} \right)\,, \\ \hat{t} &= \frac{\alpha\beta^2(t-q)}{\Delta\,\Delta_y} + \hat{q}\,, \end{aligned}$$

$$m \;=\; \langle \frac{1}{y} \Omega_y \Big(\sum_u \xi^u \Big) \rangle_z = \langle \Omega_y \left(\xi^1 \right) \rangle_z\,, \tag{E.22}$$

$$q \;=\; \langle \frac{1}{y^2} \Omega_y \Big(\sum_u \xi^u \Big)^2 \rangle_z = \langle \Omega_y \left(\xi^1 \right)^2 \rangle_z\,, \tag{E.23}$$

$$t \;=\; \langle \frac{1}{y(y-1)} \Omega_y \Big(\sum_{u<v} \xi^u \xi^v \Big) \rangle_z = \langle \Omega_y \left(\xi^1 \xi^2 \right) \rangle_z\,, \tag{E.24}$$

where $\langle \cdot \rangle_z$ is a Normal average, and $\Omega_y(\cdot)$ stands for the expectation over

$$P(\{\xi^u\})\, \mathbb{E}_{\{\xi^u\}} \exp \left\{ \sum_u \xi^u \Big(\hat{m} + z\sqrt{\hat{q}} \Big) + \frac{1}{2} \sum_{u \neq v} \Big(\frac{\gamma}{y} + \hat{t} - \hat{q} \Big) \xi^u \xi^v \right\}\,, \tag{E.25}$$

which is the average over a Curie-Weiss model with couplings $J = \frac{\gamma}{y} + \hat{t} - \hat{q}$ and external field $\hat{m} + z\sqrt{\hat{q}}$. $P(\{\xi^u\})$ stands for the product measure of the binary distribution for all the students' patterns.

It is possible to further simplify the above expresion. In particular we can linearize the interacting term $J \sum_{u<v} \xi^u \xi^v$ as for (E.19) and rewrite a new version of

$$\begin{aligned} -\beta \hat{f}^{(1)}(q,t,m;\hat{q},\hat{t},\hat{m}) = & \int Dz \log \int D\eta \left[\mathbb{E}_\xi \exp \left\{ \Big(h(z) + \sqrt{J}\,\eta \Big)\, \xi \right\} \right]^y - \frac{y}{2} J + \\ & + y \log 2 - y\frac{\hat{q}}{2} - \frac{y(y-1)}{2} \hat{t}t + \frac{y^2}{2} \hat{q}q - y\hat{m}m + \frac{\alpha}{2} \left[\frac{y \left(\beta q + \hat{\beta}\beta m^2/(1-\hat{\beta}) \right)}{1 - \beta + \beta q + (y-1)\beta(q-t)} \right] \\ & - \frac{\alpha}{2} \left\{ (y-1) \ln[(1-\hat{\beta})(1-\beta+\beta t)] + \ln \Big((1-\hat{\beta}) \left[(1-\beta+\beta q) + (y-1)\beta(q-t) \right] \Big) \right\}\,, \end{aligned}$$

leading to the following second version of saddle point equations

$$m = \int Dz \frac{\int D\eta y \cosh^y \left(h(z) + \sqrt{J}\eta \right) \tanh \left(h(z) + \sqrt{J}\eta \right)}{\int D\eta \cosh^y \left(h(z) + \sqrt{J}\eta \right)}\,, \tag{E.26}$$

$$q = \int Dz \left(\frac{\int D\eta y \cosh^y \left(h(z) + \sqrt{J}\eta \right) \tanh \left(h(z) + \sqrt{J}\eta \right)}{\int D\eta \cosh^y \left(h(z) + \sqrt{J}\eta \right)} \right)^2\,, \tag{E.27}$$

$$t = \int Dz \left(\frac{\int D\eta \cosh^y \tanh^2 \left(h(z) + \sqrt{J}\eta \right)}{\int D\eta \cosh^y \left(h(z) + \sqrt{J}\eta \right)} \right)\,. \tag{E.28}$$

Infinite number of students

To analyze the $y \to \infty$ limit, we already assume the order parameters to bem, q and δ, along with their conjugates, as stated in (4.27). This is done by allowing an extensive CW free energy with respect to its size y, as discussed in Chapter 1. By imposing the combination $t - q = \delta/y$ and taking the limit in Δ_y and Δ

$$\Delta_y = 1 - \beta + \beta q + (y-1)\beta(q-t) \simeq_{y \to \infty} 1 - \beta(1-q) - \beta\delta \,,$$
$$\Delta = 1 - \beta + \beta t \simeq_{y \to \infty} 1 - \beta(1-q) \,,$$

which transform the free energy density as in the following

$$\frac{1}{y}\hat{f}^{(1)}(q, m, \delta; \hat{q}, \hat{m}, \hat{\delta}) = \frac{\hat{q}\delta}{2\beta} + \frac{q\hat{\delta}}{2\beta} - \frac{\hat{q}q}{2\beta} + \frac{\hat{m}m}{\beta} + \frac{\hat{q}}{2\beta} + \tag{E.29}$$

$$+\frac{\alpha}{2\beta}\left\{\ln[\Delta_T\Delta] - \frac{\left(\beta q + \hat{\beta}\beta m^2/\Delta_T\right)}{\Delta_y}\right\} +$$

$$+\frac{1}{\beta y}\left\langle \log\left[\mathbb{E}_{\xi^u}\exp\left\{\sum_u \xi^u\left(\hat{m} + z\sqrt{\hat{q}}\right) + \frac{1}{2}\left(\frac{\hat{\delta}}{y} + \frac{\gamma}{y}\right)\sum_{u\neq v}\xi^u\xi^v\right\}\right]\right\rangle_z \,.$$

Under this limit Eq.(E.20) can be rewritten as

$$-\beta f \approx \lim_{y\to\infty} \operatorname{Extr} \frac{1}{y}\hat{f}^{(1)}(q, m, q; \hat{q}, \hat{m}, \hat{d}) \,.$$

The average inside $\langle\cdot\rangle_z$ is the same average of a CW model in the thermodynamic limit (here the size is $y \to \infty$). We can follow the same steps of the solution of the CW model in Chapter 1

$$\mathbb{E}_{\xi^u}\exp\left\{\sum_u \xi^u h(z) + \frac{J}{2y}\sum_{u\neq v}\xi^u\xi^v \pm \frac{J}{2y}\sum_{u=v}(\xi^u)^2\right\} \simeq e^{y\,\mathrm{Extr}_M\left[-\frac{JM^2}{2} + \log 2 + \log\cosh\{(h(z)+JM)\}\right]}$$

where we suitably redefined the coupling and field term as $h = \hat{m} + z\sqrt{\hat{q}}$ and $J = \gamma + \hat{\delta}$. This is exactly the extremization of the CW, leading to

$$\bar{M}(\sigma, z, \hat{\delta}, \hat{q}, \hat{m}, \hat{p}) = \tanh\left(h(z) + J\bar{M}\right) \,.$$

Thus one can use both the linear response theory and the fluctuation relation to interpret the internal expectation in (E.22-E.24), in particular

$$m = \langle \frac{1}{y}\Omega_y\Big(\sum_u \xi^u\Big)\rangle_z = \int Dz\bar{M}$$
$$q = \langle \frac{1}{y^2}\Omega_y\Big(\sum_u \xi^u\Big)^2\rangle_z = \int Dz\bar{M}^2$$
$$\delta = \langle \frac{1}{y(y-1)}\Omega_y\Big(\sum_{u<v}\xi^u\xi^v\Big)\rangle_z = \int Dz\frac{(1-\bar{M}^2)^2 J}{1 - J(1-\bar{M}^2)} \,.$$

Chapter 5

Architectural Optimisation in Deep Neural Networks

In this chapter, we focus on optimizing a ”learning engine” in a supervised learning setting using Deep Neural Networks (DNNs). Unlike in unsupervised or self-supervised scenarios, the machine is provided with a database containing a large dataset—such as images—where each sample is correctly labeled. The network parameters are optimized using standard backpropagation algorithms [122] [9], aligning the input-output mappings as closely as possible to the desired labels. During this process, the network is said to be in the training phase. Subsequently, a second database, known as the test set, is used to evaluate classification performance using appropriate metrics (Fig.5.1).

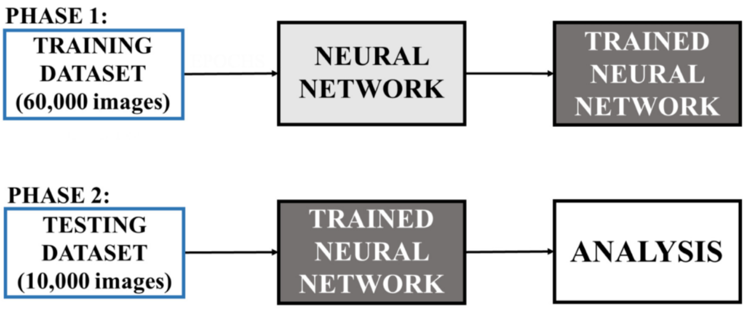


Figure 5.1: Scheme describing training and testing sessions working on a neural network.

The optimization of machine learning models has been widely addressed, particularly by tackling the complex set of hyperparameters (to distinguish them from the ones learned by the network) involved in the training process [10]. Examples of these include neuron activation functions, mini-batch size, and learning rate [123–126]. Among these, the initialization of synaptic weights is especially crucial for ensuring the machine's ability to generalize effectively. Notable in this context are *Deep Belief Networks* [18], which provide insights into the network's internal structure (hidden layers) and aid in understanding and improving overall system behavior [104] [123] [127]. Nowadays, the growing emphasis on green technologies has underscored the need to reduce the resources required to train these networks [128][129]. Early efforts in this direction include working with smaller datasets [130] or handling corrupted datasets [131] [86].

Our contribution to this challenge focuses on achieving maximum performance within a constrained set of resources. We use the number of neurons as a proxy for resource measurement and aim to determine the optimal topological architecture to maximize a classification task. This problem is particularly significant for satellites, where embedding the most efficient architecture is of paramount importance [132]. Our approach is inspired by theoretical advancements in the domain of Deep Boltzmann Machines (DBMs) [53], composed of stacked RBMs. These theoretical results suggest methods for reallocating neurons towards what we term the ”coldest area” of the network. In this study, we conducted classification experiments on a relatively small network (with three hidden layers and a total of 192 movable neurons) across three fundamental datasets: MNIST, a version of MNIST divided into even and odd digits, and Fashion MNIST [133]. Our findings demonstrate that the theoretical guidelines significantly aid in designing effective network architectures.

5.1 Results

5.1.1 Form factors and temperatures

A deep network (Fig.5.2) is built with a set of N neurons that are arranged in an arbitrary amount of $K > 2$ layers. We refer with $L^{(p)}$ at the p-th hidden layer within the set of all the ones composing the network $p \in \{1, 2, \cdots, K\}$ and the number of neurons inside the layer $L^{(p)}$ is indicated as $N^{(p)}$. Each of the *i*-th neurons of a layer $L^{(p)}$ are connected to each of the *j*-th neurons of the successive layer $L^{(p+1)}$ through a set of values called *weights* $W_{ij}^{(p)}$. A way to represent the neuron quantity in each layer is through the vector, $\boldsymbol{\alpha} = (\alpha^{(1)}, \cdots, \alpha^{(K)})$ whose entries $\alpha^{(p)}$ are called *form factors* (see Fig.(5.2)). Each form factor is defined by the relative amount of neurons with respect to the total number in the network: $\alpha^{(p)} = N^{(p)}/N$. Moreover, another useful vector describing the connections between two successive layers $\boldsymbol{\beta} = (\beta^{(1)}, \cdots, \beta^{(K)})$ is introduced. The entries of this second vector are referred as inverse temperatures and represent the scale of the fluctuations of the coupling strength of each $W_{ij}^{(p)}$ between neuron i and j, within layer $L^{(p)}$ and $L^{(p+1)}$. Their computation is done through the empirical fluctuation of the weights $W_{ij}^{(p)}$ within layers L_p and L_{p+1} , as shown in Sec. 5.1.3.

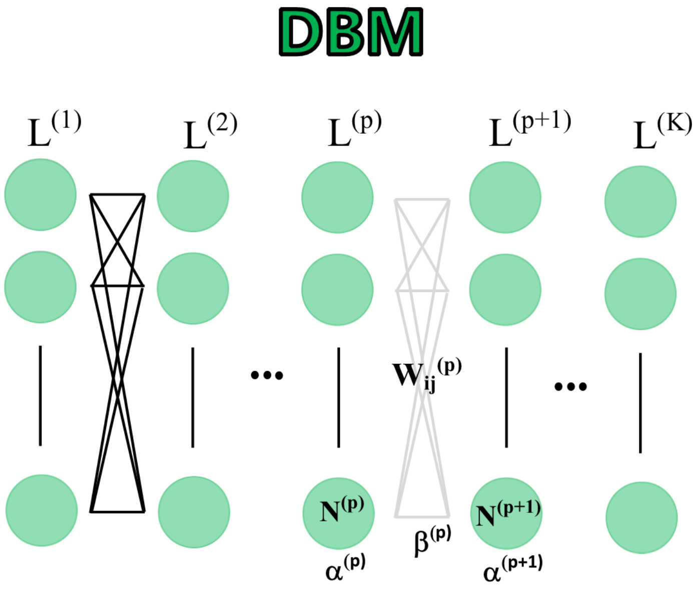


Figure 5.2: Representation of a DBM with arbitrary depth K. Each layer $L^{(p)}$ contains a certain number of neurons $N^{(p)}$, s.t. its form factors will be $\alpha^{(p)} = N^{(p)}/N$. To each couple of layers, we can associate an inverse temperature $\beta^{(p)}$.

A set of theoretical computations show [40] [20] how to estimate the spectral radius of a specific phase transition matrix which signals the exit from the highly fluctuating (paramagnetic) phase, i.e., the area of the parameters where the neural network can not reach satisfying performances. The theory suggests concentrating neurons in the coldest area of the network, namely those with the smallest weight variance. These theoretical insights form the basis of this chapter, where we empirically test the results by rearranging the neurons of a network, while keeping their total number constant.

The theoretical model proposed in [20], which inspires our work, describes the thermodynamic properties of Deep Boltzmann Machines (DBMs), such as the one shown in Fig.5.2, through the following energy function::

$$H(\sigma) = -\sqrt{\frac{2}{N}} \sum_{p}^{K-1} \beta^{(p)} \sum_{(i,j)\in L_p \times L_{p+1}} J_{ij}^{(p)} \sigma_i \sigma_j \, . \tag{5.1}$$

where each σ_i represents a spin of the layer L_p. The neurons are connected to the successive layer's neurons through centered Gaussian weights. The spectral radius mentioned above $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is computed from the following matrix

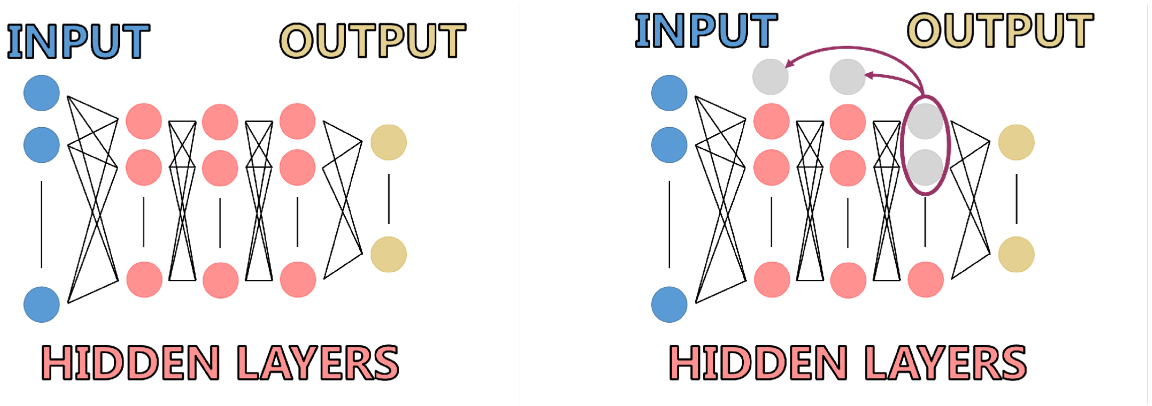
$$\begin{pmatrix} 0 & \alpha^{(2)}\left(\beta^{(1)}\right)^2 & & & \\ \alpha^{(1)}\left(\beta^{(1)}\right)^2 & 0 & \alpha^{(3)}\left(\beta^{(2)}\right)^2 & & \\ & \alpha^{(2)}\left(\beta^{(2)}\right)^2 & & & \\ & & \ddots & & \\ & & & & \alpha^{(K)}\left(\beta^{(K-1)}\right)^2 \\ & & & \alpha^{(K-1)}\left(\beta^{(K-1)}\right)^2 & 0 \end{pmatrix} . \tag{5.2}$$

This quantity acts as an order parameter describing the thermodynamic properties for phase transition of DBMs models, signaling spontaneous symmetry breaking: a transition from a convex, stable energy landscape ($\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}) < 1$) to a complex, rugged one. In the first scenario the Boltzmann machine cannot properly work, no matter of what is the database or the network architecture. In order to display some structure, we need a rich non-convex energetic landscape, where suitable minima represent the learning states [134]. The functional dependence of the spectral radius, on both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, suggests that performance can be enhanced by balancing the layer sizes with the corresponding local temperatures. The outcome is to relocate neurons in the hidden layers from the higher-temperature region to the low-temperature one, with respect to the components of $\boldsymbol{\beta}$. For example, if the layers $L^{(\hat{p})}$ and $L^{(\hat{p}+1)}$ register the highest inverse temperature $\beta^{(\hat{p})}$, the remaining neurons have to be moved towards this couple. The movement is constrained by maintaining the same number of neurons between the selected couple: $\alpha^{(\hat{p})} = \alpha^{(\hat{p}+1)}$. The rigorous mathematical treatment of the DBMs is done only for very large networks and with a simplified hypothesis for the distribution of the variables. The first assumption consists in the usage of the thermodynamic limit, which is a reasonably good hypothesis as observed in the LLM [135]. As for the second, the classical and simplified assumption is that the variance of the Gaussian distribution for the synaptic variables depends only on the layer.

These suggestions, emerging from theoretical works, must be empirically tested within simulations where the network has a finite number of neurons and the synaptic variables are real numbers obtained from specific algorithms. The network we used for this experimentation is introduced in Fig.(5.3a), it is composed by one fully-connected input layer of 784 neurons, three hidden layers of 64 neurons each, and an output layer of 10 neurons. The small neural network constructed for our experimentation is designed also to fit within the constraints of embedded hardware intended for image classification in a radiation environment. Due to the limited resources provided by embedded systems, low-dimensional neural networks are utilized [136]. The hardware is designed using the FINN firmware [137], a framework provided by AMD [138], and integrated with the Brevitas library [139] and the PYNQ framework [140]. This combination allows the implementation of such models into electronic systems.

5.1.2 Accuracy and Robustness

For the evaluation of the performance, we use two metrics: the *Accuracy* (*Acc*), that measures how accurate are the network predictions, and the *Robustness* (*Rob*), measuring the sturdiness of such predictions. Their definitions depend on the classification behaviour of the network that is explained in the following. Each of the ten classes representing the MNIST digits, have a label that we express with the index $i \in \{0, 1, ..., 9\}$. We call M_i the total amount of images in the testing set for an i-digit. For each provided j-image in M_i, to be tested, an output layer vector V_{ij} is provided.



(a)

(b)

Figure 5.3: **a**, Deep neural network for the classification of handwritten digits. The input layer has the size of the 1D-vector representing one instance of the MNIST dataset, corresponding to 784 values. The output layer is made up by 10 neurons, each one related to one of the label of the digits. The three hidden layers have the same dimension and are each composed of 64 neurons. **b**, Movement of the neurons according to the theoretical suggestion. We take two neurons from the third hidden layer of the starting network $\boldsymbol{\alpha} = (64/192, 64/192, 64/192)$ and move them to the coldest couple of layers, obtaining a new network $(65/192, 65/192, 62/192)$, maintaining the same fraction of neurons $\alpha^{(1)} = \alpha^{(2)}$.

We can define the set of its ten components as

$$\mathbb{V}_{ij} = \{V_{ij}^{\ell},\ \ell = 0, \cdots, 9\}\ ,$$

only the highest V_{ij}^{ℓ} is associated with the prediction ℓ. The definition of the metrics are therefore:

$$Acc = \frac{1}{10}\sum_{i=0}^{9}\frac{1}{M_i}\sum_{j=1}^{M_i} f_{V_{ij}^i}(max(\mathbb{V}_{ij})) \qquad f_a(x) : \mathbb{R} \to \{0,1\} \quad f_a(x) = \begin{cases} 1, & \text{if } a = x \\ 0, & \text{otherwise} \end{cases}, \tag{5.3}$$

$$\boldsymbol{Rob} = \frac{1}{10}\sum_{i=0}^{9}\frac{1}{\boldsymbol{M_i}}\sum_{j=1}^{M_i} g_i(V_{ij}^i) - g_i(max_{\ell\neq i}(\mathbb{V}_{ij})), \tag{5.4}$$

$$\Omega_i = \bigcup_{j\in M_i} \mathbb{V}_{ij}, \qquad g_i(x) : \mathbb{R} \to \mathbb{R}_{\geq 0} \quad g_i(x) = \frac{x - min(\Omega_i)}{max(\Omega_i) - min(\Omega_i)}\ . \tag{5.5}$$

Since the number of images of each label in the testing set M_i is not constant, Acc weights the total amount of correct prediction with M_i, and a uniform average on all the classes is performed. Rob provides the difference between the prediction associated to the correct label V_{ij}^i and the highest score among the remaining ones $max_{\ell\neq i}(\mathbb{V}_{ij})$. In particular, the higher is Rob, the higher is the distance between the correct prediction and the first possible misleading digit. For this reason, we consider Rob as a measure of sturdiness of the answer given by the network. Both the metrics have been developed for having a better understanding of the behavior of the networks when noise is induced by the field radiation. If accuracy has a straightforward meaning, robustness is more intricately and directly connected to the experimental setup, particularly when the network hardware interacts with radiation [39] [37][38]. This is explained in greater detail in Appendix F.0.1.

5.1.3 Temperature estimation

In order to study the consequences of the theoretical indications over the choice of the architecture and its performance, an estimation of the temperature must be added. It turns out that the empirical fluctuations of the weight matrix elements, connecting each couple of layers, is linked to the inverse temperature [141], [142]. We therefore evaluate it as:

$$\left(\beta^{(p)}\right)^2 \sim \frac{1}{N^{(p)}N^{(p+1)}} \sum_{i=1}^{N^{(p)}} \sum_{j=1}^{N^{(p+1)}} \left(W_{ij}^{(p)} - \left\langle W^{(p)} \right\rangle\right)^2 , \tag{5.6}$$

where $\left\langle W^{(p)} \right\rangle$ is the mean value of the weights $W_{ij}^{(p)}$ of a particular couple of nearest neighbours layers. Assigned the inverse temperature $\beta^{(p)}$ and the form factors $\alpha^{(p)}$, we introduce a version of the matrix (5.2) adapted to our problem

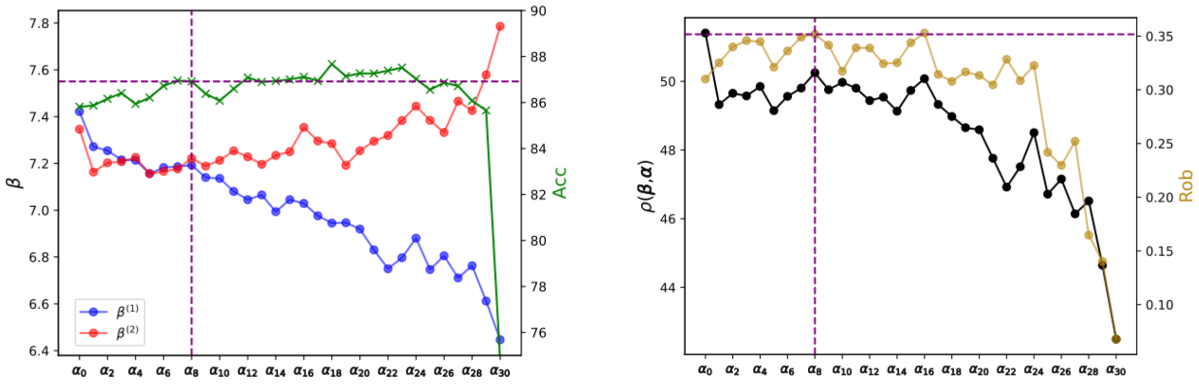
$$\begin{pmatrix} 0 & \alpha^{(2)}\left(\beta^{(1)}\right)^2 & 0 \\ \alpha^{(1)}\left(\beta^{(1)}\right)^2 & 0 & \alpha^{(3)}\left(\beta^{(2)}\right)^2 \\ 0 & \alpha^{(2)}\left(\beta^{(2)}\right)^2 & 0 \end{pmatrix} . \tag{5.7}$$

The spectral radius $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha})$ carries an important thermodynamic meaning. In order for the network to be able to have a properly classifying structure, one must have $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 1$, which is what we obtain for the network at the end of the training procedure, Fig. 5.4b.
As suggested from the theoretical results we perform the neurons reallocation, following the temperatures estimation. This strategy shows the configurations reaching thermal equilibrium $\beta^{(1)} \sim \beta^{(2)}$ are the ones giving the desired performance. These topologies are also the ones where $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is higher and coincide with the maximum of the *Robustness*. This is a remarkable fact since the *Spectral Radius* is a pure thermodynamic property of an infinite dimensional network, while the *Robustness* is an empirical quantity measured in a real finite dimensional system.

5.1.4 Architectural optimisation 1. The MNIST case study.

From this section, we refer to a *network configuration* with the vector $\boldsymbol{\alpha}_k$, containing the form factors of each hidden layer of our test network. The k-index stands for the k-th step of moving neurons, starting from the initial network configuration $\boldsymbol{\alpha}_0$. The initial condition is the equidistributed neurons network $\boldsymbol{\alpha_0} = (\alpha_0^{(1)}, \alpha_0^{(2)}, \alpha_0^{(3)})$ where $\alpha_0^{(1)} = \alpha_0^{(2)} = \alpha_0^{(3)} = 64/N$ and $N = 192$. For a network configuration, we can also associate $\boldsymbol{\beta_k} = (\beta_k^{(1)}, \beta_k^{(2)})$. Once the network is trained, we compute the associated inverse temperature values, using (5.6) and we obtain $\boldsymbol{\beta_0} = (\beta_0^{(1)} = 7.42, \beta_0^{(2)} = 7.35)$. We notice that $\beta_0^{(p)} > 1$ implies $\rho(\boldsymbol{\beta_0}, \boldsymbol{\alpha_0}) > 1$ and therefore the exit from the paramagnetic phase. The accuracy of the model that we obtain is $Acc = 85.815$, therefore, confirming the theoretical modelisation. Since $\beta_0^{(1)} > \beta_0^{(2)}$, the elementary redistribution is shifting two neurons from the last hidden layer to the first two layers, ensuring that $\alpha_1^{(1)} = \alpha_1^{(2)} = 65/192$. We continue this movement for every k, as observed in Fig.(5.3b), until $\boldsymbol{\alpha}_{30} = (\alpha_{30}^{(1)} = \alpha_{30}^{(2)} = 94/N, \alpha_{30}^{(3)} = 4/192)$. The results of this movement are shown in Fig.(5.4a) and Fig.(5.4b). In Fig. (5.4a) are represented the $Accuracy$ and $(\beta_k^{(1)}, \beta_k^{(2)})$, Fig. (5.4b) illustrates the $Robustness$ and $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha})$. After three neuronal modifications, we observe that $\beta^{(1)}$ and $\beta^{(2)}$ reach a region where they overlap. After that, we observe $\beta^{(1)}$ and $\beta^{(2)}$ diverge from each other, till reaching the bottleneck effect at $\boldsymbol{\alpha}_{28}$ where the $Accuracy$ drops drastically. Within the architectural changes, the $Accuracy$ is a slow increasing function, with its maximum reached at $\boldsymbol{\alpha}_{18}$, where it gains the 2.17% with respect to $\boldsymbol{\alpha}_0$. The total variation of this metric from the network $\boldsymbol{\alpha}_8$, inside the overlapping region, to maximum turns out to be 0.87%.

The maximum value reached for the *Robustness* at $\boldsymbol{\alpha}_8$ comes with a 13.5% improvement with respect to $\boldsymbol{\alpha}_0$. With these two metrics, we can compare better each architectural configuration and find the optimal tradeoff between the two. The configuration obtained with $\boldsymbol{\alpha}_8$ maximises the *Robustness* and makes it the most suitable configuration for resisting the radiation field. From figure (5.4b) it turns out that the maximum value of ρ is reached in proximity of configuration displaying the highest *Robustness*. This is a hint to the fact that not only the model is valuable, but it suggests a deeper link between its ideal thermodynamic properties with the practical behaviour of the neural network.



(a) (b)

Figure 5.4: **a**, Graphical representation of the $(\beta^{(1)}, \beta^{(2)})$ and Acc of a sample training for each network topology. **b**, Displays of Rob and ρ for each network configuration. The dotted line coincide with the Acc local maximum and *Rob* maximum at $\boldsymbol{\alpha}_8$.

To validate such statements, we produce a series of trained networks, each time with the neuronal exchange introduced in Fig.(5.3b), and averaging the results. In this way we can obtain an average behavior of our metrics and identify the networks belonging to the overlapping region that we name *thermal equilibrium*. Its quantitative evaluation is explained fully in Appendix F.0.2. Denoting as $\mathcal{O}$ the generic observable representing one of the metrics or the inverse temperatures, we define its average and fluctuation over repetitive trainings as:

$$\bar{\mathcal{O}} = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \mathcal{O}_t \,, \tag{5.8}$$

$$\mathcal{E}_{\mathcal{O}} = \left(\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left(\mathcal{O}_t - \bar{\mathcal{O}} \right)^2 \right)^{1/2} \,, \tag{5.9}$$

where $\mathcal{T}$ is the number of trainings, $\mathcal{O}_t$ is the observable measured at the end of the t-th training. The plots illustrating all the observables, together with their fluctuations, are displayed in Fig.(5.5). In these tests, we observe the maximum value of the *Robustness* for $\boldsymbol{\alpha}_8$, which proves to be the most stable value under repetitive trainings. This architecture is reached within the thermal equilibrium area. This suggests that reaching the equilibration of temperatures among the synapse connections is a good argument to find a suitable architecture configuration. Also the computation of the (averaged) spectral radius gives hints about the optimal solution. It suggests the higher scores of the $Robustness$ are reached for sufficiently high values of ρ. As one can see from Fig.(5.5), $\boldsymbol{\alpha}_8$

represents the correct tradeoff between performance and the network's *Robustness*, reaching an enhancement of 4.8% with respect to $\boldsymbol{\alpha}_0$.

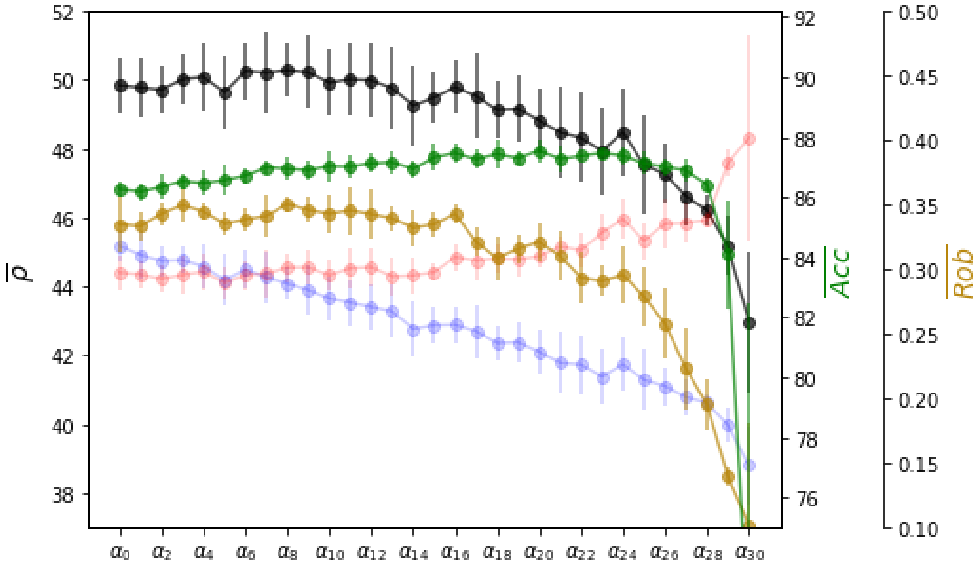
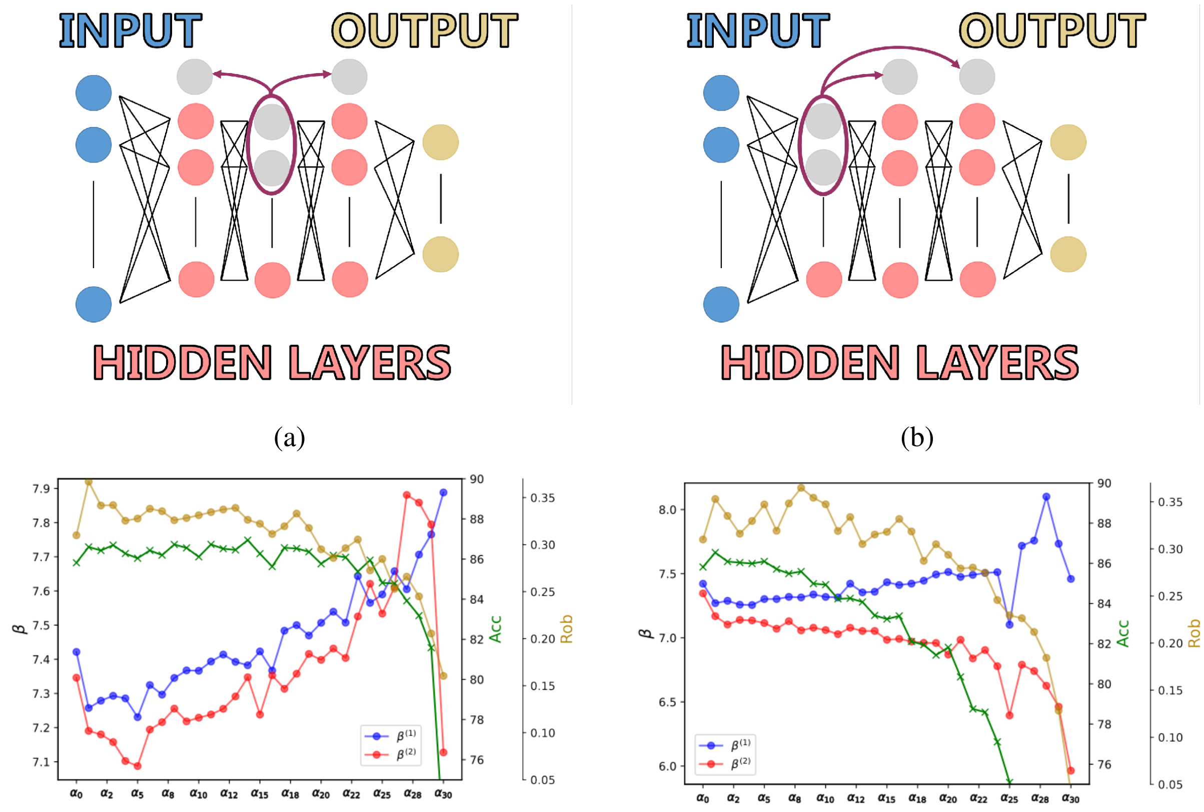


Figure 5.5: Average of the *Rob* in yellow, *Acc* in green, *SR* in black and $(\beta^{(1)}, \beta^{(2)})$ respectively in blue and red, for the MNIST dataset for each architectural configuration.

To fully investigate the effect of topological changes, we also tested alternative movement schemes, independently of the theoretical suggestions. Fig.(5.6) shows two cases, the former moving neurons from the middle layer towards the outer layer, as shown in Fig.(5.6a), the latter in the opposite direction of the temperature criterion, as shown in Fig.(5.6b). From the computation of the metrics, we observe that these architectural choices yield lower performance results, as shown in Fig. (5.6c) and (5.6d). Moving the neurons towards the coldest couple of layers appears to be the best choice among the tested ones.



(a) (b)

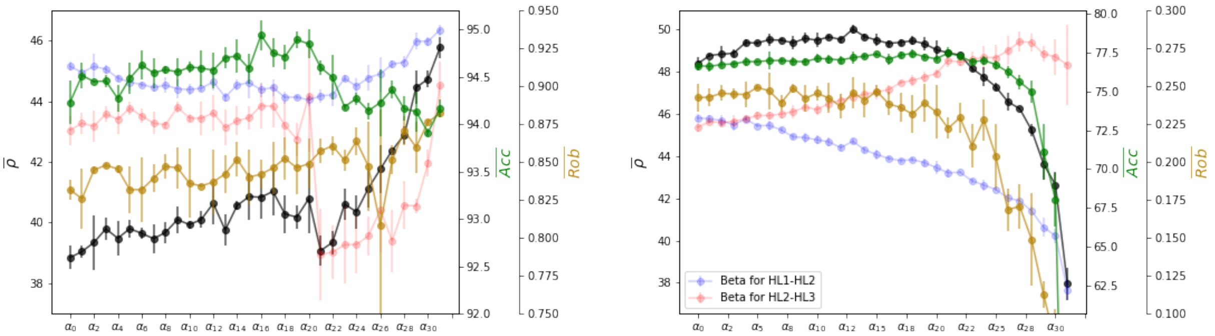
(c) (d)

Figure 5.6: **a**, Movement of neurons from the middle layer to the two outer ones. **b**, movement of neurons from the cold area toward the hotter one. **c**, performances of the network when moving neurons from the middle layers. **d**, performances of the network when moving neurons from the cold area.

5.1.5 Architectural optimisation 2. The other case studies.

The previous work has been performed utilising the same dataset, number of classes and training procedure. Therefore, we aim to visualise how the network would respond to perturbations of these attributes.

The first modification we implement is reducing the number of classes of the MNIST dataset while maintaining the same code implementation. Instead of assigning each digit to its own class, we sort them into two classes: odd and even numbers. The output layer of the network is then changed from 10 to 2 neurons. We perform the same test as before by transferring neurons from the hottest area towards the coldest one. The results are visualised in Fig.(5.7a), which displays the mean of different trainings for all the architectures. Compared to the results obtained with the MNIST dataset in Sec.5.1.4, both the $Accuracy$ and the $Robustness$ show enhanced results. This is due to the fact that the task is simplified by performing binary classification. Here, the *Spectral Radius* has a plateau in the thermal equilibrium region, where the maximum value of the $Accuracy$, $\boldsymbol{\alpha}_{16}$, is registered. At this architecture Acc increases by 0.7% and Rob increases by 1.7% with respect to $\boldsymbol{\alpha}_0$ performance. We also observe that $\rho(\boldsymbol{\alpha}, \boldsymbol{\beta})$ tends to have a similar functional form of the $Robustness$ as we continue the architectural changes. In particular, the *Robustness* peaks at the final architectural model, which no longer suffers from the bottleneck effect, with *Rob* showing a 6% improvement compared to $\boldsymbol{\alpha}_0$. However, a slight decline in accuracy is observed compared to the previous configurations. This can be explained by considering the role of ρ as the order parameter signaling the network phase transitions. As we stated, once the critical value $\rho_c = 1$ is reached, the network transitions from a stable energy landscape to a more complex one. However, continuously increasing the spectral radius does not necessarily lead to further improvements. Very large values may be associated with highly irregular energy landscapes, characterized by narrow minima that are difficult to reach during training, ultimately resulting in a degraded performance.



(a)

(b)

Figure 5.7: **a**, Graphical representation of the averaged $(\beta^{(1)}, \beta^{(2)})$, Acc, Rob and SR for the MNIST dataset with odd-even classification for each architectural configuration. **b**, Graphical representation of the $(\beta^{(1)}, \beta^{(2)})$, Acc, Rob and SR for the fashion MNIST dataset for each architectural configuration.

In the second test, we change the dataset from MNIST to fashion MNIST [133]. It consists of ten classes of clothes items, with the same input dimensions as the standard MNIST images, as described in the Appendix. The results of the study, using this dataset, are displayed in Fig.(5.7b), showing the mean results of five different studies. From the scale of the two metrics, both $Robustness$ and $Accuracy$ are significantly smaller compared to the standard MNIST results. However, we once again observe the same interesting area around thermal equilibrium, corresponding to $\boldsymbol{\alpha}_5$, where the *Robustness* peaks and the three metrics exhibit the smallest uncertainties. Here, Acc and Rob respectively reach $< 1\%$ and 2.8% with respect to the initial condition $\boldsymbol{\alpha}_0$. In Fig. 5.5, the higher scores of Rob and ρ correspond to a narrow region of architectures, whereas, in the fashion MNIST case, the higher values of ρ alone correspond to a broader region, forming a plateau in Fig. 5.7b.

5.2 Discussion

In this chapter, we have used a theoretical inspired method to fine-tune the performance of deep networks with a fixed number of neurons. We observe that the movement of the neurons from one layer to another leads to improvements in *Accuracy* and, most importantly, *Robustness*. The suggestion of moving neurons towards the coldest area of the network [20], which suitably increases the spectral radius, proves to be effective across the two tested datasets. In our cases we move the neurons across the hidden part of the network, since both input and output layers are fixed. We find that continuing the moving procedure up to the region of thermal equilibrium leads to robust architectural configurations. Additionally, the spectral radius of the phase transition matrix (5.7) serves as a reliable indicator for exploring nearby configurations. Our method suggests that the most stable and robust architecture lies inside or at the boundary of the identified thermal equilibrium region, as illustrated in Figs.(5.5, 5.7a and 5.7b). By following this approach, we achieve average robustness improvements of 4.8% and 6% for the MNIST and its modified version, and the 2.8% for the Fashion MNIST.This work primarily focused on finding a method to select the optimal architecture with the best tradeoff between *Accuracy* and *Robustness*, while adhering to a fixed number of neurons due to energy limitation. This approach is intended to produce a small network suitable for use and training in satellite hardware, which often face power limitations. Additionally, we aimed to identify architectures with the highest possible *Robustness*, being a metric affected by space ambient interactions.

Appendix F

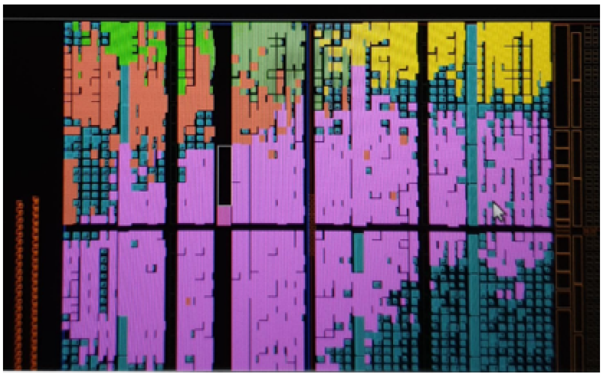
Description of network training and testing

The network is constructed using the Brevitas libraries [139], which enable the allocation of bits for deployment on embedded systems. The various features of the network are as follows:

- Dropout, representing the probability of randomly setting elements of the input tensor to zero, is set to 0.2. The input dimension is determined by the number of pixels composing each image,
- The number of layers and neurons within each layer is initially set to three hidden layers, each containing sixty-four neurons,
- The dimension of the output layer depends on the number of classes present in the dataset,
- The number of bits allocated to the weights and activations is set to 1 for both,
- The number of epochs is set to 500 to ensure good performance, mitigate the risk of overfitting, and reduce computation time,
- The learning rate is set to 0.02 and is adjusted using the ADAM optimizer [125],
- Weights are initialised with a uniform distribution between -1 and 1,
- Biases are initialised to 0,
- The loss function used is the Square Hinge loss.

F.0.1 Metrics motivations

In this section, we briefly summarize the main steps that connect this chapter to the hardware testing. The details of the hardware analysis are beyond the scope of our architectural study; for an in-depth overview, see [39]. As mentioned in the main text, one of the parallel goals of this work was to study the resilience and behavior of the tested networks under irradiation. The successful implementation of the DNN on the hardware board (Zybo Z7-10) was achieved using the Brevitas libraries, which convert all characteristics of the trained network into an electronic format, as shown in Fig. (F.1a).



(a)

Digits	CL$_i$ for 0% noise	CL$_i$ for 10% noise
0	0.4759615	0.4443430
1	0.5897323	0.4044362
2	0.3962992	0.2990106
3	0.4146421	0.3143130
4	0.4204332	0.2110623
5	0.3600811	0.2192392
6	0.4936567	0.3581108
7	0.4069328	0.3039969
8	0.3999368	0.2685075
9	0.4103263	0.2568289
Total	0.4392423	0.3099597

(b)

Figure F.1: **a**, Display of the electronics composing the hardware board. The network elements allocated to electronic are: purple for the input layer; yellow, orange and khaki for the hidden layers; bright green for the output layer. The size matches the form factors specified for the tested networks. **b**, List of the *Rob*$_i$ for the network at respectively 0% and 10%noise in the testing and training sessions. The ”Total” line corresponds to the *Rob*. This metric is the one affected by particles interaction .

The first metric introduced is the *Accuracy* (*Acc*) and is computed by averaging the accuracy of each digit, as shown in Eq.(5.3). This approach ensures homogeneity regarding the weight of each digit in the final result, given that the number of images in the testing set, of all the used databases, is not equally distributed.

The *Robustness* (*Rob*) is the second metric, and is computed by exploiting the difference between the value expected by the network, and the highest value of the non-expected predictions in the output layer. This method has been computed taking into account the consequences of the radiations interacting with hardware, as described in [37] and [38]. This metric quantifies the steadiness of the network against variations induced by field radiations, as one can observe in the table presented in Fig.(F.1b). *Rob* can also have negative values indicating the network is performing incorrect predictions.

A practical example of the computation of one element of the sum in Eq.(5.4) is explained in the following. Assume the network is testing a specific image $\hat{j} \in M_2$, therefore corresponding to the label $i = 2$ as in Fig.(F.2). We further assume that the minimum and maximum output values of all the outputs $\{V^{\ell}_{2j}\}$, $\ell = 0, \cdots, 9$, $j \in M_2$ of the tested images are respectively $max(\Omega_2) = 3$, $min(\Omega_2) = -2$.

Then we just shift all the output values of $\{V^{\ell}_{2\hat{j}}\}$ of the corresponding minimal value and normalize them, as in Eq.(F.1)

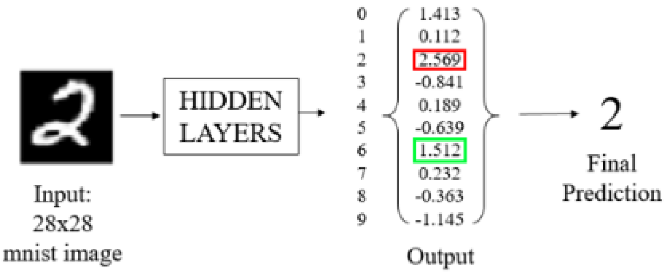


Figure F.2: This scheme describes how a network predicts the class corresponding to the input. A picture of a 2-digit is provided and then processed through the trained network and an output array is obtained. In this case, the highest value in red (2.569) at index 2, corresponding to a correct prediction of the class ”2”.

$$V^{\ell}_{2\hat{j}} - \mathbf{1}^{\ell} min(\Omega_2) = \begin{pmatrix} 2.413 \\ 2.112 \\ \boxed{4.569} \\ \vdots \\ \boxed{3.512} \\ \vdots \end{pmatrix} \quad ; \qquad V^{\ell}_{2\hat{j}} - \mathbf{1}^{\ell} \frac{min(\Omega_2)}{max(\Omega_2) - min(\Omega_2)} = \begin{pmatrix} 0.48 \\ 0.42 \\ \boxed{0.91} \\ \vdots \\ \boxed{0.70} \\ \vdots \end{pmatrix} , \tag{F.1}$$

where $\mathbf{1}^{\ell}$ is the $\ell-$dimensional vector with 10 components. After normalization we obtain $Rob_2 = g_2\left(V^2_{2\hat{j}}\right) - g_2(max_{\ell \neq 2}(\mathbb{V}_{2\hat{j}})) = 0.21$.

F.0.2 Thermal equilibrium

In a numerical setting, defining when a system has reached thermal equilibrium requires an operational criterion. Our choice is based on a dimensionless comparison between two quantities: The relative spread of the local inverse temperatures, which we denote as Δ_{L1}; the intrinsic statistical fluctuations of the local temperatures, measured via the normalized variance (variance divided by the square of the mean), denoted Δ_{L2}. Remembering (Eqs.(5.8-5.9)) that we denote as $\bar{\mathcal{O}}$ the average of an observable under repetitive training, the definitions of Δ_{L1} and Δ_{L2} are as follow:

$$\Delta_{L1} = \frac{|\bar{\beta}_1 - \bar{\beta}_2|}{\frac{\bar{\beta}_1 + \bar{\beta}_2}{2}} , \tag{F.2}$$

$$\Delta_{L2} = \sqrt{\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \frac{\left(\beta_t - \bar{\beta}\right)^2}{\bar{\beta}^2}} . \tag{F.3}$$

We say that the system is at equilibrium when $\Delta_{L1} \leq \Delta_{L2}$. The rationale is that if the spread of temperatures across the system falls within the scale of their intrinsic statistical fluctuations, then any remaining inhomogeneities can be attributed to noise rather than to a genuine out-of-equilibrium state. This criterion provides a practical and physically motivated argument to distinguish temperature equilibration. With this choice, the points of thermal equilibrium discussed in Sec.5.1.4 correspond to those in Fig.(F.3). Although the inverse temperatures signal the presence of a cold and hot area, we note the temperatures of each area of the network are $T_{1,2} = 1/\beta_{1,2} < 1$. This is directly liked with the value of $\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 1$, signaling the network is catching data structure.

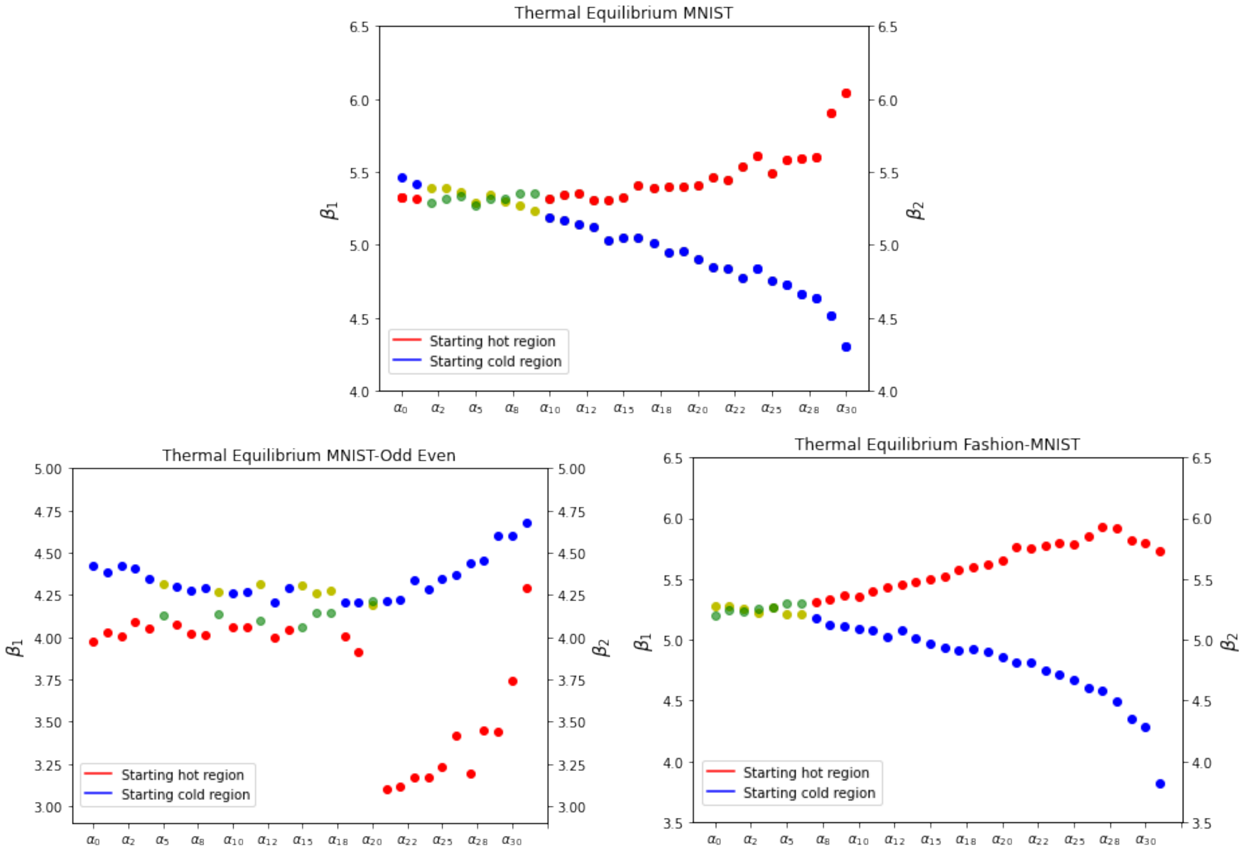
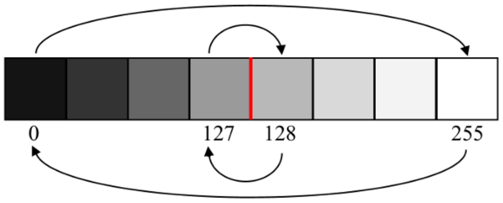


Figure F.3: Representation of the architectures with hidden layers in thermal equilibrium. In yellow and green, we have the inverse temperatures points that are indistinguishable. The yellow ones refers to the first couple of hidden layers of the network. All the temperatures points are obtained after an average over many trained networks.

F.0.3 Datasets

To mimic the interfierence of the radiation environment that affects also the dataset images [39] we convert the grayscale pixels composing the MNIST and Fashion MNIST performing a random symmetric swithc as in Fig.(F.4) . This is done both for the training and testing parts.



(a)

(b)

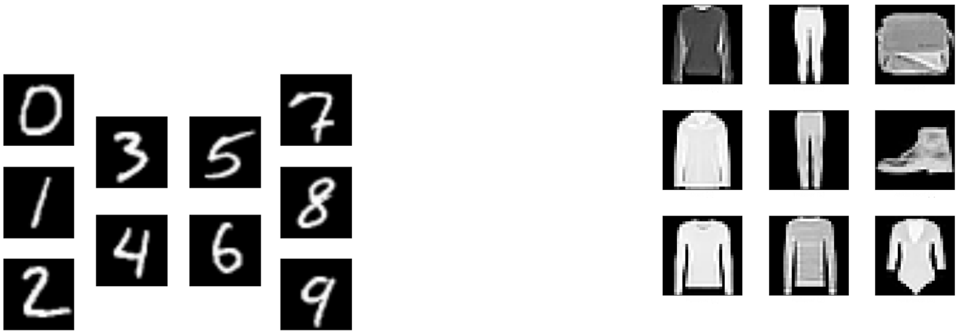
Figure F.4: Example of the implemented symmetric flip on a grayscale image of the MNIST dataset. The original class sample is on the left, while on the right an implementation of the flipping with a rate of 10% is shown.

MNIST

The MNIST dataset [143] comprises 70,000 images, with 60,000 images in the training set, and 10,000 images in the testing set. The images depict handwritten digits ranging from 0 to 9. Each image consists of 28x28 greyscale pixels. The distribution of images is heterogeneous across the various classes. This dataset is well-suited for tasks involving image recognition, and its reputation provides valuable insights into the performance of artificial intelligence systems on this dataset.

Fashion MNIST

The fashion MNIST dataset [133] comprises 70,000 images, with 60,000 images in the training set, and 10,000 images in the testing set. The images depict different types of clothes divided in 10 classes. Each image consists of 28x28 greyscale pixels. The distribution of images is homogeneous across the various classes. This dataset being very close to the MNIST dataset, it allows studying variations while sticking to the same area of image recognition.



(a)

(b)

Figure F.5: Samples of MNIST (left) and Fashion MNIST (right) datasets

F.0.4 Data and Code availability

All the datasets supporting the findings of this study as well as the code used for the analysis and training of the neural network can be found on GitHub at https://github.com/ArchitecturalOpt/Architecture-Optimisation

Chapter 6

Concluding remarks

In this thesis, we study the inference properties of different Boltzmann-Gibbs engines, which are well-known for their remarkable ability to learn from data. Our focus has been on understanding their behavior in optimization contexts, gaining insights that could enhance their learning efficiency through better selection of key properties. The first chapter provides a concise overview of the direct problem for Restricted Boltzmann Machines, with a particular focus on the Hopfield model. This serves as an essential foundation, highlighting that the celebrated generative properties of these models rely on their ability to learn the underlying structure of the data they process. Consequently, inference emerges as a fundamental step in their operation.

The second chapter delves into the Dual version of the Hopfield model, introducing the teacher-student framework. In this setup, one Hopfield network (the teacher) generates a dataset that is subsequently analyzed by another Hopfield network (the student). The challenge of recovering the teacher's planted pattern is examined in two regimes: the Bayes-optimal case, where the likelihood is known and the mismatched case, where the beyond Bayes-optimality is operated on the parameters of the student. This exploration provides insight into the equilibrium inference properties in a more realistic scenario, since the assumption that the student's machine is exactly equal to the teacher one is not realistic.

In the initial scenario of small datasets, we find that, with a suitable transformation, the examples are those derived from a Curie-Weiss (CW) model. When the CW model operates at low temperatures, these examples exhibit a macroscopic overlap with the teacher's planted pattern, enabling the student to learn by directly memorizing the data. As the dataset grows, alignment between the student and the teacher's pattern becomes easier, making learning trivial when the dataset becomes extensive.

However, in cases where the examples are noisy—such as when the teacher generates data at higher temperatures—memorization fails. The Hopfield student then transitions to a signal retrieval (sR) phase, where it extracts small fragments of useful information from the examples. This behavior mirrors the modern learning mechanisms observed in contemporary AI models. We then realize the mismathcing between the two models by adopting different temperatures of inference and generation ($\beta \neq \hat{\beta}$). In such cases, the emergence of low-energy configurations, uncorrelated with the signal, can hinder learning. These student pattern configurations may either correlate with the examples (eR phase) or remain entirely uncorrelated with both the signal and the examples (SG phase). Despite this, increasing the dataset size enables the student to effectively learn by generalization.

These findings, derived through the mean-field theory of spin glasses, are not limited to the Hopfield model but extend to the broader class of Restricted Boltzmann Machines (RBMs). Chapter 3 builds on this foundation by examining the effects of prior mismatches in RBMs within the teacher-student framework. Specifically, we explore challenges arising from differing teacher processes or student architectures. Through this approach, we derive general equations applicable across various prior configurations, with a focus on an interpolating case where the priors transition between continuous (Gaussian) and binary variables. Notably, Gaussian hidden units facilitate the student machine's entry into the sR region more efficiently, though the critical dataset size depends on the dataset's properties. This can be seen by the computation of the critical point $P_{triple} = (\alpha_c, T_c)$ in Sec.3.4, which components move according to the different choices of teacher (both α_c and T_c move) and student (only T_c moves) architecture. Unlike the direct problem approach, the posterior distribution method inherently incorporates self-regularization, eliminating the need for ad hoc constraints and avoiding issues like parameter divergence.

In Chapter 4, we address the limitations imposed by the teacher's influence on the critical point component α_c. Trying to overcome these boundaries, we introduced ferromagnetic coupling among y replicated machines, focusing our efforts on the initial case of the Hopfield inverse problem. Unlike previous approaches, this method does not achieve Bayes-optimality because the inclusion of interacting replicas distorts the likelihood's form. Interestingly, we observed that the minimum dataset size required for entering the sR regime remains unchanged, regardless of the coupling strength (γ) or the number of replicas (y). However, the sR region in the (α, T) phase space expands significantly compared to earlier architectural efficiency mechanisms (see Fig.3.5). Similar to the findings in Chapter 3, the sR area extends in the direction of higher inference temperatures (weights regularization). This expansion occurs whether the system is enhanced by increasing γ or y, suggesting that ferromagnetic coupling promotes alignment among replicas. This alignment amplifies the signal retrieval effect as soon as the data threshold (α_c) is crossed.
Nevertheless, the maximum temperature for transitioning into the sR regime remains bounded, as seen in the self-spherical constraint for the RBM teacher-student framework. This limitation becomes evident when analyzing the transition lines P-SG: β_y^{SG} and P-sR: β_y^{sR}. By plotting these lines, and with a suitable rescaling of the coupling strength, we observe that the gap between transition lines diminishes as y increases (Fig.4.3). In the limit $y \to \infty$, the coupled replicas yield transition lines (β_∞^{SG} and β_∞^{sR}) that no longer depend on y (Fig. 4.5). Notably, we derived an analytical form for β_∞^{sR}, as shown in Eqs. (4.32-4.33). This is reasonable from a machine learning perspective: if a model can enter a learning regime at arbitrarily high temperatures, it implies that learning could occur with minimal training epochs or even immediately after random weight initialization, where no structure is apparent.

These results deepen our understanding of the learning properties of RBM models under different configurations. In contrast, Chapter 5 shifts focus to the optimization of deep neural networks (DNNs), aiming to reduce the energy demands of their training processes—a pressing concern in the context of green technological initiatives. Our investigation centers on an image classification problem using a DNN with a finite number of neurons. This problem has practical implications for satellite applications, where efficient architectures must be embedded in hardware operating in radiation-rich environments. Since the visible and output layers are constrained by the specific task, we optimized the placement of neurons within the hidden layers to enhance resilience against radiation. This optimization was achieved through two approaches: introducing noise to both training and test datasets, to simulate the effects of particle interactions on hardware, and defining a robustness metric (Rob) to evaluate network sturdiness. Guided by theoretical insights from Deep Boltzmann Machines, we found that a small set of thermodynamic parameters—specifically, form factors ($\boldsymbol{\alpha}$) and the inverse temperatures of layer pairs ($\boldsymbol{\beta}$)—suffices to define an equilibrium topology with maximum Rob. This is achieved by moving the neurons from the hot part of the network towards the colder one, reaching a thermal equilibrium area where we find the desired topology. We validated these findings through repetitive training, yielding averaged performance metrics for different architectures. The results, tested on three datasets (MNIST, Fashion MNIST, and Odd-Even MNIST—a binary classification variant of the MNIST), demonstrated consistent improvements in robustness. Our procedure enhanced network resilience by 4.8%, 2.8% and 6% for the respective datasets.

Bibliography

[1] Pierluigi Contucci, Godwin Osabutey, and Filippo Zimmaro. New perspectives on growth and sustainability. glimpses into the ai industrial revolution, 2024.

[2] The Nobel Prize. The nobel prize in chemistry 2024 press release: [include brief description of the award here], 2024. Accessed: 2024-10-11.

[3] Jyri Kivinen and Christopher Williams. Multiple texture boltzmann machines. In *Artificial Intelligence and Statistics*, pages 638–646. PMLR, 2012.

[4] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

[5] Nitish Srivastava, Ruslan R Salakhutdinov, and Geoffrey E Hinton. Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865*, 2013.

[6] The Nobel Prize. The nobel prize in physics 2024 press release: They trained artificial neural networks using physics, 2024. Accessed: 2024-10-11.

[7] Pierluigi Contucci. *Rivoluzione intelligenza artificiale: Sfide, rischi, opportunità*. Edizioni Dedalo, 2024.

[8] Yann LeCun. The epistemology of deep learning, institute for advanced study. https://www.youtube.com/watch?v=gG5NCkMerHU.

[9] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[10] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade: Second edition*, pages 437–478. Springer, 2012.

[11] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[12] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[13] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

[14] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press, 2001.

[15] Matthias Löwe. On the storage capacity of hopfield models with correlated patterns. *The Annals of Applied Probability*, 8(4):1216–1250, 1998.

[16] Hanoch Gutfreund. Neural networks with hierarchically correlated patterns. *Physical Review A*, 37(2):570, 1988.

[17] Giordano De Marzo and Giulio Iannelli. Effect of spatial correlations on hopfield neural network and dense associative memories. *Physica A: Statistical Mechanics and its Applications*, 612:128487, 2023.

[18] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[19] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[20] Diego Alberici, Pierluigi Contucci, and Emanuele Mingione. Deep boltzmann machines: rigorous results at arbitrary depth. In *Annales Henri Poincaré*, volume 22, pages 2619–2642. Springer, 2021.

[21] Jorge Fernandez-de Cossio-Diaz, Clément Roussel, Simona Cocco, and Remi Monasson. Accelerated sampling with stacked restricted boltzmann machines. In *The Twelfth International Conference on Learning Representations*.

[22] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[23] Giorgio Parisi. A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

[24] A Dubey, M Sachan, and J Wieczorek. summary and discussion of:“why does unsupervised pre-training help deep learning?”. *Statistics Journal Club*, 2014.

[25] Diego Alberici, Francesco Camilli, Pierluigi Contucci, and Emanuele Mingione. The solution of the deep boltzmann machine on the nishimori line. *Communications in Mathematical Physics*, 387:1191–1214, 2021.

[26] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

[27] Aurélien Decelle, Cyril Furtlehner, Alfonso de Jesús Navas Gómez, and Beatriz Seoane. Inferring effective couplings with restricted boltzmann machines. *SciPost Physics*, 16(4):095, 2024.

[28] Haiping Huang. Statistical mechanics of unsupervised feature learning in a restricted boltzmann machine with binary synapses. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053302, 2017.

[29] Haiping Huang and Taro Toyoizumi. Unsupervised feature learning from finite data by message passing: discontinuous versus continuous phase transition. *Physical Review E*, 94(6):062310, 2016.

[30] Aurelien Decelle, Sungmin Hwang, Jacopo Rocchi, and Daniele Tantari. Inverse problems for structured datasets using parallel tap equations and restricted boltzmann machines. *Scientific Reports*, 11(1):19990, 2021.

[31] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[32] Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071, 1980.

[33] Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, 1999.

[34] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

[35] Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. Copycat perceptron: Smashing barriers through collective learning. *Physical Review E*, 109(6):065313, 2024.

[36] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. Glassy nature of the hard phase in inference problems. *Physical Review X*, 9(1):011020, 2019.

[37] Lucas Matana Luza, Annachiara Ruospo, Daniel Söderström, Carlo Cazzaniga, Maria Kastriotou, Ernesto Sanchez, Alberto Bosio, and Luigi Dilillo. Emulating the effects of radiation-induced soft-errors for the reliability assessment of neural networks. *IEEE Transactions on Emerging Topics in Computing*, 10(4):1867–1882, 2022.

[38] Luiz Henrique Laurini, João Baptista dos Santos Martins, and Rodrigo Possamai Bastos. Investigation of edge computing hardware architectures processing tiny machine learning under neutron-induced radiation effects. *Microelectronics Reliability*, 150:115179, 2023. Special issue of 34th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis, ESREF 2023.

[39] Sacha Cormenier. *Characterisation of electronic devices and finite-dimensions deep neural network topology optimisation*. Phd thesis, Università di Roma Tre, May 2024. Available at http://www.matfis.uniroma3.it/Allegati/Dottorato/TESI/scormenier/PhD_Thesis_Cormenier.pdf other than Google Chrome.

[40] Diego Alberici, Adriano Barra, Pierluigi Contucci, and Emanuele Mingione. Annealing and replica-symmetry in deep boltzmann machines. *Journal of Statistical Physics*, 180(1):665–677, 2020.

[41] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

[42] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[43] Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. Equilibrium and non-equilibrium regimes in the learning of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 34:5345–5359, 2021.

[44] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. *Advances in neural information processing systems*, 28, 2015.

[45] Jean Barbier, Francesco Camilli, Marco Mondelli, and Manuel Saenz. Bayes-optimal limits in structured pca, and how to reach them. *arXiv preprint arXiv:2210.01237*, 2022.

[46] Kerson Huang. *Introduction to statistical physics*. Chapman and Hall/CRC, 2009.

[47] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[48] Sacha Friedli and Yvan Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017.

[49] Anthony CC Coolen, Reimer Kühn, and Peter Sollich. *Theory of neural information processing systems*. OUP Oxford, 2005.

[50] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

[51] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[52] Elena Agliari, Adriano Barra, and Brunello Tirozzi. Free energies of boltzmann machines: self-averaging, annealed and replica symmetric approximations in the thermodynamic limit. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(3):033301, 2019.

[53] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[54] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1–9, 2012.

[55] Elena Agliari, Danila Migliozzi, and Daniele Tantari. Non-convex multi-species hopfield models. *Journal of Statistical Physics*, 172(5):1247–1269, 2018.

[56] Dmitry Panchenko. The free energy in a multi-species sherrington–kirkpatrick model. 2015.

[57] Diego Alberici, Francesco Camilli, Pierluigi Contucci, and Emanuele Mingione. The multi-species mean-field spin-glass on the nishimori line. *Journal of Statistical Physics*, 182:1–20, 2021.

[58] Giuseppe Genovese. A remark on the spherical bipartite spin glass. *Mathematical Physics, Analysis and Geometry*, 25(2):14, 2022.

[59] Giuseppe Genovese and Daniele Tantari. Non-convex multipartite ferromagnets. *Journal of Statistical Physics*, 163:492–513, 2016.

[60] Giuseppe Genovese and Daniele Tantari. Legendre duality of spherical and gaussian spin glasses. *Mathematical Physics, Analysis and Geometry*, 18:1–19, 2015.

[61] Adriano Barra, Andrea Galluzzi, Francesco Guerra, Andrea Pizzoferrato, and Daniele Tantari. Mean field bipartite spin models treated with mechanical techniques. *The European Physical Journal B*, 87:1–13, 2014.

[62] Giuseppe Genovese and Daniele Tantari. Overlap synchronisation in multipartite random energy models. *Journal of Statistical Physics*, 169:1162–1170, 2017.

[63] Adriano Barra, Giuseppe Genovese, Francesco Guerra, and Daniele Tantari. About a solvable mean field model of a gaussian spin glass. *Journal of Physics A: Mathematical and Theoretical*, 47(15):155002, 2014.

[64] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. Phase diagram of restricted boltzmann machines and generalized hopfield networks with arbitrary priors. *Physical Review E*, 97(2):022310, 2018.

[65] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. Phase transitions in restricted boltzmann machines with generic priors. *Physical Review E*, 96(4):042156, 2017.

[66] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[67] E Gardner. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, 20(11):3453, 1987.

[68] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, and Francesco Moauro. Multitasking associative networks. *Physical review letters*, 109(26):268101, 2012.

[69] Elena Agliari, Alessia Annibale, Adriano Barra, ACC Coolen, and Daniele Tantari. Immune networks: multitasking capabilities near saturation. *Journal of Physics A: Mathematical and Theoretical*, 46(41):415003, 2013.

[70] Elena Agliari, Alessia Annibale, Adriano Barra, ACC Coolen, and Daniele Tantari. Immune networks: multi-tasking capabilities at medium load. *Journal of Physics A: Mathematical and Theoretical*, 46(33):335101, 2013.

[71] Elena Agliari, Alessia Annibale, Adriano Barra, Anthony CC Coolen, and Daniele Tantari. Retrieving infinite numbers of patterns in a spin-glass model of immune networks. *Europhysics Letters*, 117(2):28003, 2017.

[72] Peter Sollich, Daniele Tantari, Alessia Annibale, and Adriano Barra. Extensive parallel processing on scale-free networks. *Physical review letters*, 113(23):238106, 2014.

[73] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, Daniele Tantari, and Flavia Tavani. Retrieval capabilities of hierarchical networks: From dyson to hopfield. *Physical review letters*, 114(2):028103, 2015.

[74] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, Daniele Tantari, and Flavia Tavani. Hierarchical neural networks perform both serial and parallel processing. *Neural Networks*, 66:22–35, 2015.

[75] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, Daniele Tantari, and Flavia Tavani. Metastable states in the hierarchical dyson model drive parallel processing in the hierarchical hopfield network. *Journal of Physics A: Mathematical and Theoretical*, 48(1):015001, 2014.

[76] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, Daniele Tantari, and Flavia Tavani. Topological properties of hierarchical networks. *Physical Review E*, 91(6):062807, 2015.

[77] Adriano Barra, Giuseppe Genovese, Francesco Guerra, and Daniele Tantari. How glassy are neural networks? *Journal of Statistical Mechanics: Theory and Experiment*, 2012(07):P07009, 2012.

[78] Elena Agliari, Adriano Barra, Chiara Longo, and Daniele Tantari. Neural networks retrieving boolean patterns in a sea of gaussian ones. *Journal of Statistical Physics*, 168:1085–1104, 2017.

[79] Giuseppe Genovese and Daniele Tantari. Legendre equivalences of spherical boltzmann machines. *Journal of Physics A: Mathematical and Theoretical*, 53(9):094001, 2020.

[80] Elena Agliari, Adriano Barra, Gino Del Ferraro, Francesco Guerra, and Daniele Tantari. Anergy in self-directed b lymphocytes: a statistical mechanics perspective. *Journal of theoretical biology*, 375:21–31, 2015.

[81] Jacopo Rocchi, David Saad, and Daniele Tantari. High storage capacity in the hopfield model with auto-interactions—stability analysis. *Journal of Physics A: Mathematical and Theoretical*, 50(46):465001, 2017.

[82] José Fernando Fontanari and WK Theumann. On the storage of correlated patterns in hopfield's model. *Journal de Physique*, 51(5):375–386, 1990.

[83] R Der, VS Dotsenko, and B Tirozzi. Modified pseudo-inverse neural networks storing correlated patterns. *Journal of Physics A: Mathematical and General*, 25(10):2843, 1992.

[84] JL Van Hemmen. Hebbian learning, its correlation catastrophe, and unlearning. *Network: Computation in Neural Systems*, 8(3):V1, 1997.

[85] Elena Agliari, Francesca Elisa Leonelli, and Chiara Marullo. Storing, learning and retrieving biased patterns. *Applied Mathematics and Computation*, 415:126716, 2022.

[86] Elena Agliari, Francesco Alemanno, Adriano Barra, and Giordano De Marzo. The emergence of a concept in shallow neural networks. *Neural Networks*, 148:232–253, 2022.

[87] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

[88] Yuma Ichikawa and Koji Hukushima. Statistical-mechanical study of deep boltzmann machine given weight parameters after training by singular value decomposition. *Journal of the Physical Society of Japan*, 91(11):114001, 2022.

[89] Matteo Negri, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, and Enrico Malatesta. The hidden-manifold hopfield model and a learning phase transition. *arXiv preprint arXiv:2303.16880*, 2023.

[90] Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability theory and related fields*, 174:1133–1185, 2019.

[91] Maria Chiara Angelini and Federico Ricci-Tersenghi. Limits and performances of algorithms based on simulated annealing in solving sparse hard inference problems. *Physical Review X*, 13(2):021011, 2023.

[92] Maria Chiara Angelini, Paolo Fachin, and Simone de Feo. Mismatching as a tool to enhance algorithmic performances of monte carlo methods for the planted clique model. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(11):113406, 2021.

[93] Simona Cocco, Remi Monasson, and Vitor Sessak. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Physical Review E*, 83(5):051123, 2011.

[94] Tianqi Hou, KY Michael Wong, and Haiping Huang. Minimal model of permutation symmetry in unsupervised learning. *Journal of Physics A: Mathematical and Theoretical*, 52(41):414001, 2019.

[95] Alfredo Braunstein, Abolfazl Ramezanpour, Riccardo Zecchina, and Pan Zhang. Inference and learning in sparse systems with multiple states. *Physical Review E*, 83(5):056114, 2011.

[96] Aurélien Decelle and Federico Ricci-Tersenghi. Solving the inverse ising problem by mean-field methods in a clustered phase space with many states. *Physical Review E*, 94(1):012112, 2016.

[97] Richard S Ellis. *Entropy, large deviations, and statistical mechanics*, volume 1431. Taylor & Francis, 2006.

[98] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre GR Day, Clint Richardson, Charles K Fisher, and David J Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics reports*, 810:1–124, 2019.

[99] Haiping Huang. *Statistical Mechanics of Neural Networks*. Springer Nature, 2022.

[100] Aurélien Decelle and Cyril Furtlehner. Restricted boltzmann machine: Recent advances and mean-field theory. *Chinese Physics B*, 30(4):040202, 2021.

[101] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Spectral dynamics of learning in restricted boltzmann machines. *Europhysics Letters*, 119(6):60001, 2017.

[102] Alessandra Carbone, Aurélien Decelle, Lorenzo Rosset, and Beatriz Seoane. Fast and functional structured data generators rooted in out-of-equilibrium physics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[103] Nicolas Béreux, Aurélien Decelle, Cyril Furtlehner, Lorenzo Rosset, and Beatriz Seoane. Fast, accurate training and sampling of restricted boltzmann machines. *arXiv preprint arXiv:2405.15376*, 2024.

[104] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[105] Jorge Fernandez-de Cossio-Diaz, Simona Cocco, and Rémi Monasson. Disentangling representations in restricted boltzmann machines without adversaries. *Physical Review X*, 13(2):021003, 2023.

[106] Adriano Barra, Pierluigi Contucci, Emanuele Mingione, and Daniele Tantari. Multi-species mean field spin glasses. rigorous results. In *Annales Henri Poincaré*, volume 16, pages 691–708. Springer, 2015.

[107] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Thermodynamics of restricted boltzmann machines and related learning dynamics. *Journal of Statistical Physics*, 172:1576–1608, 2018.

[108] Adriano Barra, Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. Thermodynamics of bidirectional associative memories. *Journal of Physics A: Mathematical and Theoretical*, 56(20):205005, 2023.

[109] Fabrizio Antenucci, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Approximate survey propagation for statistical inference. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(2):023401, 2019.

[110] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[111] Jérôme Tubiana. *Restricted Boltzmann machines: from compositional representations to protein sequence analysis*. PhD thesis, Université Paris sciences et lettres, 2018.

[112] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

[113] Robin Thériault, Francesco Tosello, and Daniele Tantari. Modelling structured data learning with restricted boltzmann machines in the teacher-student setting. *arXiv preprint arXiv:2410.16150*, 2024.

[114] Robin Thériault and Daniele Tantari. Dense Hopfield networks in the teacher-student setting. *SciPost Phys.*, 17:040, 2024.

[115] Pierluigi Contucci, Cristian Giardina, and Hidetoshi Nishimori. Spin glass identities and the nishimori line. In *Spin Glasses: Statics and Dynamics: Summer School, Paris 2007*, pages 103–121. Springer, 2009.

[116] Carlo Baldassi and Riccardo Zecchina. Efficiency of quantum vs. classical annealing in nonconvex learning problems. *Proceedings of the National Academy of Sciences*, 115(7):1457–1462, 2018.

[117] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 599–619. Springer, 2012.

[118] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Local entropy as a measure for sampling solutions in constraint satisfaction problems. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(2):023301, 2016.

[119] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical review letters*, 115(12):128101, 2015.

[120] Carlo Baldassi, Christian Borgs, Jennifer T Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016.

[121] Haiping Huang, KY Michael Wong, and Yoshiyuki Kabashima. Entropy landscape of solutions in the binary perceptron problem. *Journal of Physics A: Mathematical and Theoretical*, 46(37):375002, 2013.

[122] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, may 2019.

[123] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[124] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

[125] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[126] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[127] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[128] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[129] Karen Hao. Training a single ai model can emit as much carbon as five cars in their lifetimes. *MIT technology Review*, 75:103, 2019.

[130] Elena Agliari, Francesco Alemanno, Miriam Aquaro, Adriano Barra, Fabrizio Durante, and Ido Kanter. Hebbian dreaming for small datasets. *Neural Networks*, page 106174, 2024.

[131] Francesco Alemanno, Miriam Aquaro, Ido Kanter, Adriano Barra, and Elena Agliari. Supervised hebbian learning. *Europhysics Letters*, 141(1):11001, 2023.

[132] Partha Pratim Ray. A review on tinyml: State-of-the-art and prospects. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1595–1623, 2022.

[133] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[134] Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2020.

[135] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[136] Nimit S. Sohoni, Christopher R. Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré. Low-memory neural network training: A technical report, 2022.

[137] Yaman Umuroglu, Nicholas J. Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. Finn: A framework for fast, scalable binarized neural network inference. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, FPGA '17. ACM, February 2017.

[138] Advanced Micro Devices. https://www.amd.com/en.html.

[139] Alessandro Pappalardo. Xilinx/brevitas, 2023.

[140] Advanced Micro Devices. Pynq: Python productivity for adaptive computing platforms. https://pynq.readthedocs.io/en/latest/#.

[141] J Tubiana and R Monasson. Emergence of compositional representations in restricted boltzmann machines. *Physical Review Letters*, 118(13):138301–138301, 2017.

[142] Giancarlo Fissore. *Generative modeling: statistical physics of Restricted Boltzmann Machines, learning with missing information and scalable training of Linear Flows*. PhD thesis, université Paris-Saclay, 2022.

[143] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.