Faculty of Science

PhD course in

Computational methods and mathematical models for sciences and finance

Cycle XXXVII

Explaining Machine Learning and Memorization with Statistical Mechanics

Scientific Disciplinary Sector INF/01

Candidate: Robin Thériault

Supervisor: Prof. Daniele Tantari

Coordinator: Prof. Michele Benzi

Academic year 2024/2025

Acknowledgments

To be added in the final version.

Abstract

Artificial neural networks (NNs) and machine learning (ML) algorithms are poorly understood from a theoretical perspective, which makes it difficult to fully realize their potential and overcome their weaknesses. For instance, ML algorithms train NN weights by moving them along a low-dimensional subspace of their allowed values, but this implicitly low-dimensional learning structure is not properly exploited to improve training because its nature is not well understood. Moreover, trained NNs are easily confused by pervasive adversarial attacks whose theoretical underpinnings are still unclear. This thesis aims to improve our theoretical understanding of NNs and ML, with a particular focus on adversarial attacks and implicitly low-dimensional learning. For this purpose, we use mathematical tools from statistical mechanics to study different types of NNs and ways in which they can fit the data. In particular, we study two classes of models that fit the data with various degrees of learning and memorization: dense associative memory (DAM) and restricted Boltzmann machines (RBM). In the process, we investigate connections between different versions of these models that are useful to make analytical investigations more efficient.

First, we study a type of DAM called dense Hopfield network (dense HN) in the teacher-student setting where it is trained using data generated by another dense HN. On the Nishimori line, we show that the phase where dense HNs in the teacher-student setting are able to learn data coincides with the spin-glass phase of dense HNs with random memorized patterns. Outside the Nishimori line, we investigate the noise tolerance and adversarial robustness of dense HNs. In particular, we derive an exact formula for the adversarial robustness of the student at zero temperature, and we clarify why the adversarial robustness of dense HNs changes as a function of the learning regime.

Second, we study RBMs in the teacher-student setting. When the teacher's weights are uncorrelated, we validate the conjecture that the performance of the student in learning them is independent of the number of hidden units. Moreover, we show that a student that is larger than necessary to learn the teacher's weights adopts a low-dimensional learning strategy in which only a subset of its hidden units end up correlated with those of the teacher, which we argue can be used as a toy model for studying the lottery ticket hypothesis. When the teacher's weights are correlated together rather than purely random, we show that the student crosses multiple regimes of data representation where it learns them in increasingly detailed ways as the number of samples in its training dataset increases.

Finally, we study a type of RBM that belongs to the class of DAMs and is capable of both supervised and unsupervised classification. As before, our methods are based on statistical mechanics calculations in the teacher-student setting. We propose a novel regularization scheme inspired by these calculations, which we find to make training on real data significantly more stable. Moreover, we show that the weights learned by relatively small DAMs trained on both real and synthetic data are saddle points of larger DAMs, and we implement an algorithm that uses this hierarchy to significantly accelerate training on real data.

Contents

1 Introduction **4**

1.1 Background . . . 4
1.2 The Hopfield network: a simple example of energy-based model . . . 9
1.3 Restricted Boltzmann machines and dense Hopfield networks . . . 12
1.4 Statistical mechanics of networks in the teacher-student setting . . . 15
1.5 Outline of the thesis . . . 17

2 Dense Hopfield networks in the teacher-student setting **19**

2.1 Introduction . . . 19
2.2 Overview of Gardner's results . . . 21
2.3 Teacher-student setting . . . 23
2.3.1 Matched interaction orders . . . 24
2.3.2 Mismatched interaction orders . . . 25
2.4 Results and Discussion . . . 26
2.4.1 Retrieval transition at large interaction order . . . 26
2.4.2 Transition to the ordered phases: Universality . . . 27
2.4.3 Phase diagram on the Nishimori line . . . 28
2.4.4 Inference temperature vs dataset noise . . . 31
2.4.5 Interaction order and noise tolerance . . . 33
2.4.5.1 Large noise scaling . . . 33
2.4.5.2 Finite noise scaling . . . 35
2.4.6 Robustness against adversarial attacks . . . 35
2.5 Conclusion . . . 38
2.A Gardner's Hamiltonian vs K & H's Hamiltonian . . . 38
2.B Direct model cumulant expansions . . . 41
2.C Teacher-student replicated partition function . . . 43
2.D Teacher-student free entropy . . . 46
2.E Direct model RSB ansatz . . . 54
2.F Monte Carlo simulations for various system sizes . . . 55

3 Modeling structured data learning with Restricted Boltzmann machines in the teacher–student setting **56**
3.1 Introduction . . . 56
3.2 Model . . . 58
3.3 Results and discussion . . . 60
3.3.1 Free entropy and saddle point equations . . . 60
3.3.2 Learning uncorrelated patterns . . . 62
3.3.2.1 Independence of the number of hidden units: binary patterns . . . 63
3.3.2.2 Independence of the number of hidden units: Gaussian patterns . . . 69
3.3.2.3 A simple model of the lottery ticket hypothesis . . . 69
3.3.3 Learning uniformly correlated patterns . . . 72
3.3.3.1 Permutation symmetry breaking transitions . . . 75
3.3.4 Random correlations . . . 78
3.4 Conclusion . . . 80
3.A Definitions . . . 81
3.A.1 Binary random variables with a fixed covariance matrix . . . 81
3.A.2 Projected Wishart distribution . . . 81
3.A.3 Effective Hamiltonian $\mathcal{L}_{\lambda_1\lambda_2}$ . . . 82
3.B Replicated partition function . . . 83
3.C RS free entropy . . . 85
3.D Saddle-point equations . . . 88
3.E Saddle-point equations for Gaussian ξ . . . 90
3.F Critical load . . . 91
3.G Saddle-point equations in the absence of correlations . . . 92
3.H Effect of uniform correlations . . . 95
3.I Numerical methods . . . 96
3.J Supplementary figures . . . 97

4 Saddle Hierarchy in Dense Associative Memory **101**
4.1 Introduction . . . 101
4.2 Model . . . 102
4.2.1 A dense associative memory (DAM) model . . . 103
4.2.2 Teacher-student setting . . . 106
4.3 Theoretical results . . . 106
4.3.1 Saddle-point equations . . . 106
4.3.2 Saddle-point hierarchy . . . 109
4.4 Empirical results . . . 110
4.4.1 Learning by minimizing the effective loss . . . 110
4.4.2 Dense associative memory is interpretable, even in unsupervised classification . . . 112
4.4.3 Fast training with splitting steepest descent . . . 114
4.5 Conclusion . . . 118

4.A Derivation of the model . . . 119
4.B Stationarity conditions of the loss . . . 121
4.C Integration of the von Mises-Fisher density . . . 122
4.D Replicated partition function and free entropy . . . 123
4.E Saddle-point equations . . . 127
4.F Normalization of the weights . . . 130
4.G Saddle-point hierarchy . . . 132
4.H Splitting steepest descent . . . 135
4.I Initialization and learning rate . . . 137
4.J Weights learned with unsupervised training . . . 139

Chapter 1

Introduction

1.1 Background

Artificial neural networks (NNs) and machine learning (ML) algorithms have, with stunning speed, gone from niche tools to being the state-of-the-art in solving numerous fundamental and practical problems. To name a few of their achievements, they have long surpassed humans in image recognition tasks [1], they are able to emulate written language so well that it is possible to have an open-ended conversation with them [2], and they can predict the structure of complex molecules such as proteins [3, 4], which has promising applications in drug discovery and in medicine as a whole [5]. In the process, they have also become much larger and more sophisticated, to the point where even models with hundreds of millions of parameters such as latent diffusion [6] are considered compact by today's standards. This level of complexity makes their behavior very challenging to explain theoretically. However, the general idea behind them is relatively simple.

In essence, machine learning (ML) is a set of computer algorithms for fitting high-dimensional data $\mathbf{x}$ with a function $F_{\mathbf{W}}$ by adjusting its parameters $\mathbf{W}$, which are also known as the function's *weights*. This process is commonly refer to as *training* or *learning* the weights. Here, the term high-dimensional data means data with many components $x_1, x_2, ..., x_N$, which are usually concrete *features* of the data. For example, digital grayscale images are a type of high-dimensional data whose components are the intensities of their pixels. An artificial neural network (NN) is a function $F_{\mathbf{W}}$ whose weights $\mathbf{W}$ can be interpreted as the strengths of connections in a graph or network. For example, the linear model $y = \sum_{i=1}^{N} w_i x_i$ of $\mathbf{x} = (x_1, x_2, ..., x_N)$ is an NN whose weights $w_1, w_2, ..., w_N$ can be "machine learned" from a set of high-dimensional data points $\left(\mathbf{x}^1, y^1\right), \left(\mathbf{x}^2, y^2\right), ..., \left(\mathbf{x}^M, y^M\right)$ using linear regression (see Fig. 1.1). Intuitively, each connection i is like a pipe with a width of w_i that controls how much of the input component x_i flows to the output component y (see Fig. 1.2, left panel). In general, NNs have a more sophisticated network of connections, some of which perform non-linear operations (see Fig. 1.2, right panel), which is what makes them much more complicated and powerful than the linear model [7]. The nodes at the junctions of these NN connections are typically called *neurons* or *units*.

ML algorithms are particularly useful for learning probability distribution functions that describe how data points are placed relative to each other. The goal of such models is usually to generate new data points that are close enough to the existing ones to appear realistic, such as coherent sentences in the case

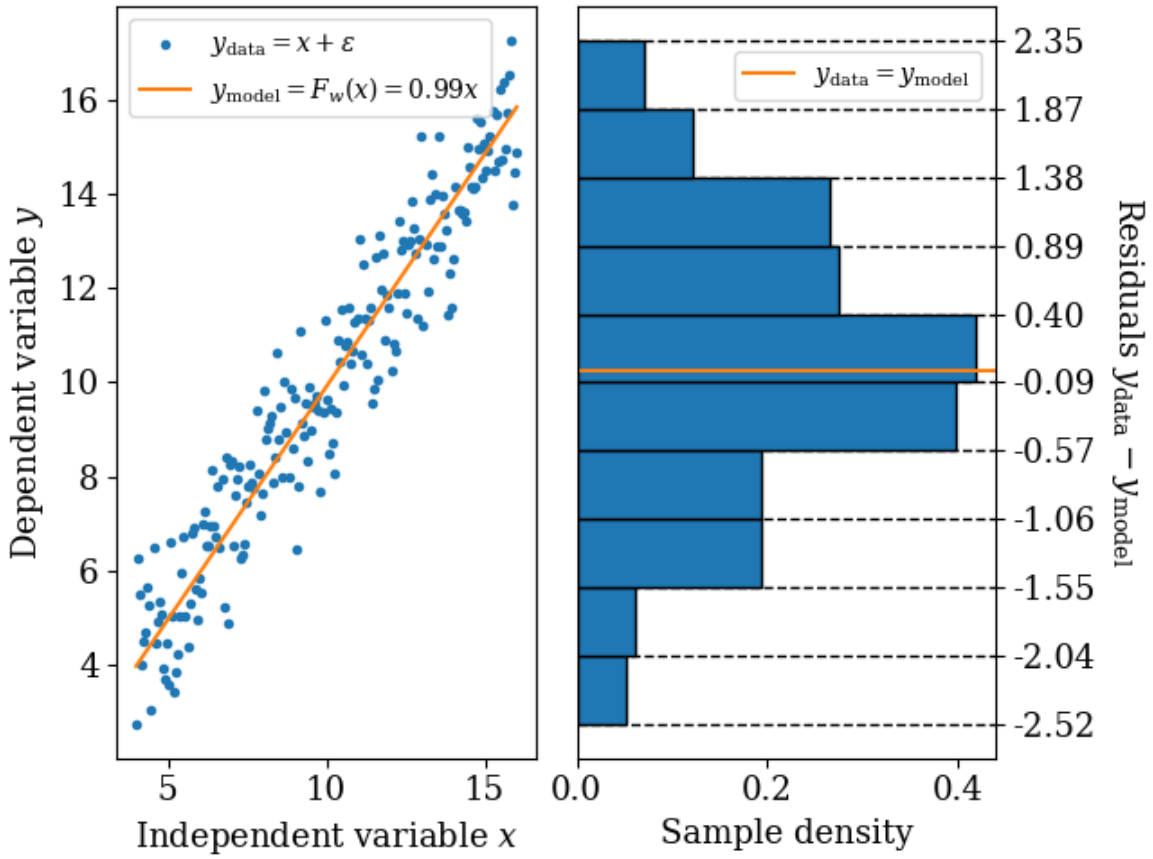

Figure 1.1: In the left panel, a linear function $y_{\text{model}} = F_w(x) = wx$ fit to data $y_{\text{data}} = x + \varepsilon$, where ε is Gaussian noise with variance $\sigma^2 = 1$ (see Fig. 1.3). The learned weight is $w = 0.99$. In the right panel, the residual distribution of the data around the fit. The orange line $y_{\text{data}} = y_{\text{model}}$ can also be interpreted as a "deterministic" distribution fit to the data.

of (large) language models [2]. Simpler NNs are also used to predict a noisy output from a deterministic input, such as for categorizing an image into one of many classes [8]. Learning a generic function and a probability distribution are two sides of the same coin. For example, linear regression is arguably equivalent to fitting data $(\mathbf{x}, y_{\text{data}})$ with the "deterministic" distribution $p(y \mid \mathbf{x}) = \delta\left(y - \sum_{i=1}^{N} w_i x_i\right)$ such that y is equal to $\sum_{i=1}^{N} w_i x_i$ with probability 1 and cannot take any other values (see Fig. 1.1, right panel). On the flip side, we can also fit the data with a proper probability distribution such as the Gaussian distribution $p(y \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left[y - \sum_{i=1}^{N} w_i x_i\right]^2\right)$ of y given $\mathbf{x}$ (illustrated in Fig. 1.3), then use the resulting w_i as the weights of the linear fit. Generating data points amounts to *sampling* the probability distribution fit to the data. In the case of the linear model, it consists of calculating $y_{\text{model}} = \sum_{i=1}^{N} w_i x_i$ for a given $\mathbf{x}$ and adding some random noise ε to the result, for example according to the Gaussian distribution mentioned above (see Figs. 1.3 and 1.1). In this context, the *variance* σ^2 represents the level of noise injected into y. When it is sufficiently small, the generated values of y are practically indistinguishable from the y_{model} obtained without adding noise.

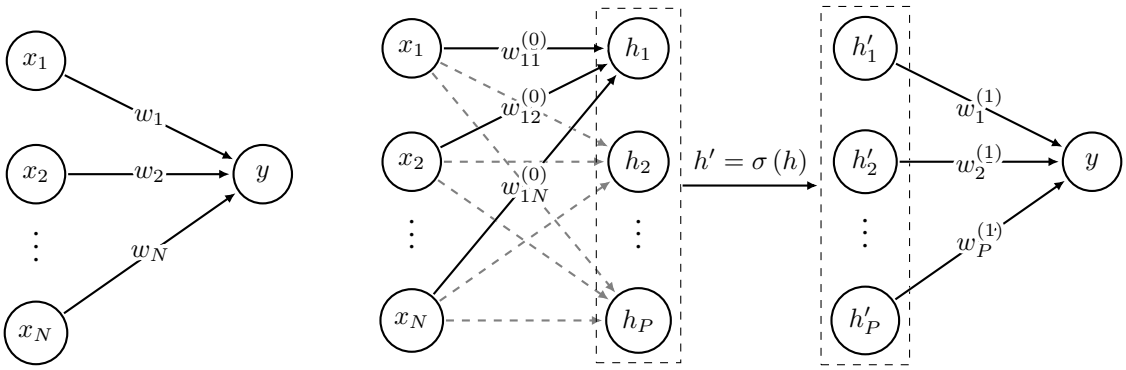

Figure 1.2: In the left panel, the network representation of the linear model $y = \sum_{i=1}^{N} w_i x_i$. In the right panel, the network representation of $y = \sum_{\mu=1}^{P} w_\mu^{(1)} \sigma\left(\sum_{i=1}^{N} w_{\mu i}^{(0)} x_i\right)$, where σ is a non-linear operation, also known as an activation function. Both models are neural networks (NNs), but the right-panel one can fit much more complicated data.

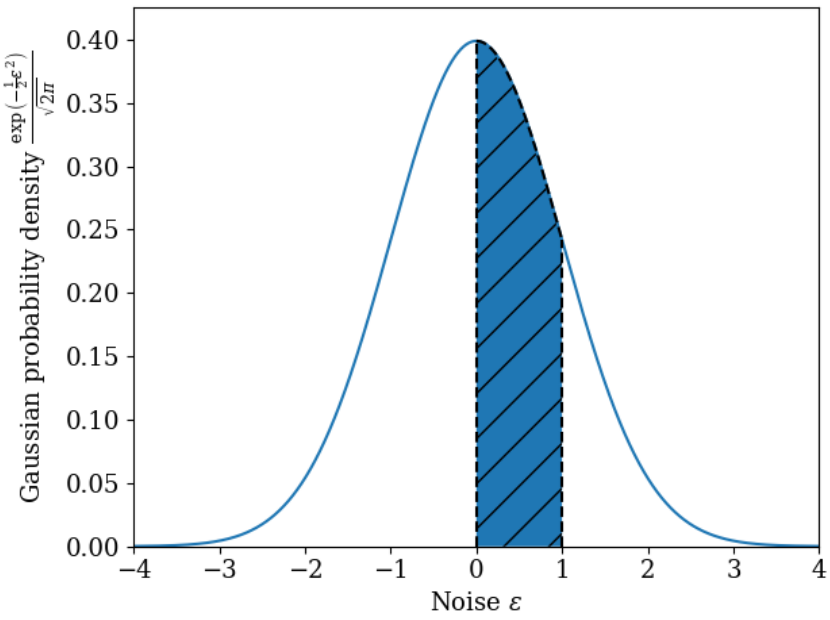

Figure 1.3: Illustration of the probability density distribution $\frac{\exp\left(-\frac{1}{2\sigma^2}\varepsilon^2\right)}{\sqrt{2\pi\sigma^2}}$ of Gaussian noise ε with $\sigma^2 = 1$. The probability that ε falls between two values is equal to the area under the curve between these two values. For example, the probability that ε ends up between 0 and 1 is given by the area of the hatched blue region. Reducing σ^2 makes the central peak of the distribution narrower and higher, and thus the noise ε more likely to be small.

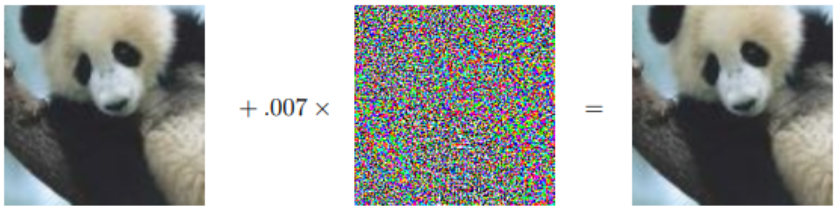

Figure 1.4: An adversarial attack making an neural network (NN) trained on the ImageNet dataset [14] recognize a panda as a gibbon. This plot comes from [15], which contains additional details about the adversarial attack and NN and in question.

Due to the difficulty in studying them theoretically and how quickly they have evolved, ML and NNs are still poorly understood, making it difficult to fully realize their potential and overcome their weaknesses. On the one hand, it was observed that ML algorithms train NN weights by moving them along a low-dimensional subspace, or *manifold*, of their allowed values [9, 10], which could potentially be used to design more efficient algorithms if it were better understood. To use an analogy, it is more efficient to locate objects on the surface of the Earth with the two-dimensional latitude-longitude coordinate system than with the three-dimensional Cartesian coordinate system. For simplicity, we call this phenomenon *implicitly low-dimensional learning* for the rest of the Introduction.

On the other hand, NNs that classify data with high accuracy can often be fooled by modifying the data with carefully crafted perturbations that are either meaningless or completely invisible to humans [11, 12] (see Fig. 1.4 for an example). This data corruption process, called an *adversarial attack*, can be dangerous in some cases. For example, it can confuse the traffic sign recognition algorithms of self-driving vehicles [13]. Adversarial attacks are not well understood, which makes them hard to circumvent.

In the same way that a bridge model made of ice cream sticks can give a rough idea of the physics of a real bridge, many aspects of ML and NNs can be understood by studying simpler high-dimensional models. For example, [16, 17] modeled high-dimensional data as lying on a low-dimensional manifold and [18, 19, 20] studied adversarial attacks in linear models. The calculations and theoretical modeling in both of these two lines of work are based on statistical mechanics, and this is not a coincidence. In fact, the field of statistical physics (also known as statistical mechanics) was originally developed to study phase transitions in materials with a large number of particles, but it can also be used to describe data points with many components, datasets with many data points, NNs with many neurons, etc. In particular, when we change their properties and those of their training data, NNs can undergo phase transitions between different regimes of data representation (uninformative, universal, specialized, etc.) [21] analogous to those between the states of matter (gaseous, liquid, solid, etc.). The scientific literature on the topic is rich and flourishing [22], recent highlights being the 2024 Nobel Prize awarded to John Hopfield and Geoffrey Hinton for their highly influential work on two models of NNs inspired by statistical physics: the Hopfield networks [23] and restricted Boltzmann machines [24], respectively.

As suggested in the previous paragraph, restricted Boltzmann machines (RBMs) and Hopfield networks (HNs) are simple enough to be studied with analytical calculations, yet still exhibit some of the most intriguing

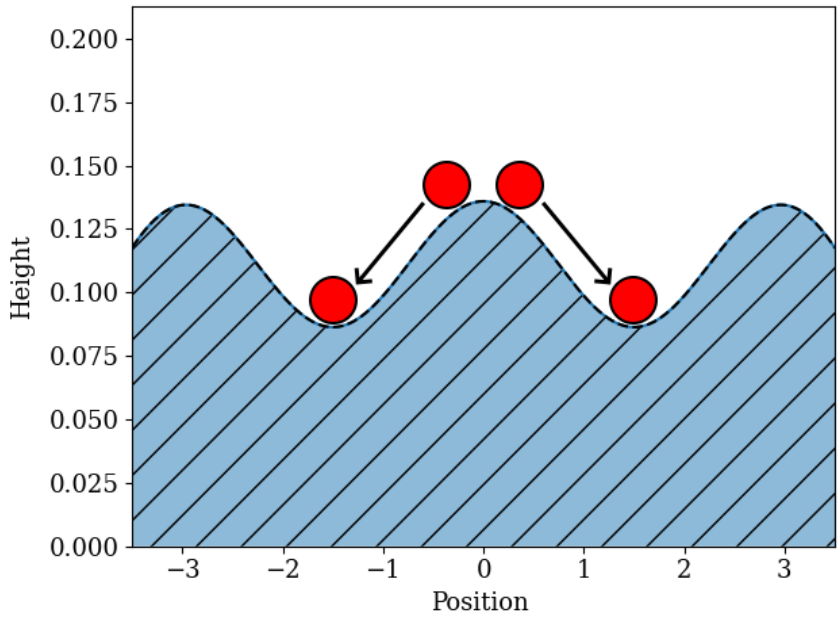

Figure 1.5: Sketch of two balls sliding down a mountain. At equilibrium, they stand motionless at the bottom of the valleys surrounding the mountain, which are local minima of the gravitational energy.

characteristics of modern ML and NNs. For example, RBMs and generalizations of HNs called *dense Hopfield networks* [25, 26, 27, 28, 29, 30] can fit data with a rich underlying structure by undergoing a multistage learning process whose first stages were found to be low dimensional [31, 32, 33, 34], a property that was then exploited to accelerate RBM training [35]. Moreover, dense HNs were observed to be either vulnerable or robust to adversarial attacks in different controlled scenarios [36], suggesting that it could be possible to improve our understanding of adversarial attacks by studying them.

RBMs and HNs fit data $\mathbf{x}$ using an energy function H, which is also known as a *Hamiltonian*. In ML terminology, they are energy-based models. The energy function has weights analogous to those of other NNs, which we will denote as $\mathbf{J}$, and it is designed so that more realistic data points typically have lower energy than less realistic ones once the weights $\mathbf{J}$ have been determined. Intuitively, a physical object leaking energy with the environment, such as a ball rolling down a mountain, eventually converges to a configuration where its energy cannot decrease anymore, which in the case of the ball corresponds to standing still at the bottom of a valley. Such a state is called an *equilibrium configuration* or a *local minimum* of the energy. The local minima $\mathbf{x}$ of an energy-based model with given weights $\mathbf{J}$ can be found analogously, and these states generally correspond to realistic data provided that the model and the strategy used to find $\mathbf{J}$ are appropriate to represent its underlying structure [23, 24, 37]. For example, the minimum energy configurations of HNs are typically preexisting data points stored in the weights by a "machine memorization" (MM) procedure (described in Section 1.2), and RBMs can learn to generate new data on which they were not trained.

This thesis studies the regimes of data representation and the phase transitions of RBMs and dense HNs [25, 26, 27, 28, 29, 30, 38] to advance our understanding of MM, ML and NNs, with a focus on implicitly low-dimensional learning and adversarial attacks. Our methods are based on statistical mechanics calculations

that are enabled by the relative simplicity of these energy-based models. On the way towards this objective, we investigate deep relationships between different versions of these models and their phase transitions [39, 40, 41, 42] that considerably simplify some of the calculations [43, 44].

The rest of this Chapter aims to introduce readers with a background in probability theory and physics to the models that we study and the tools that we use to do so. In Section 1.2, we use HNs as an example to present the general idea behind energy-based models. In Section 1.3, we build upon this foundation to describe RBMs and dense HNs. In Section 1.4, we present the teacher-student setting that we use in our analytical calculations and explain what kind of results can be derived with it. Finally, Section 1.5 uses the notions presented in Sections 1.2, 1.3 and 1.4 to put together a comprehensive outline of the other Chapters of the thesis, which present the results of our research.

Here is a high-level preview of these results. In Chapter 2, we study the adversarially vulnerable and adversarially robust regimes of dense HNs using a formal relationship between MM and ML that we investigate in detail [45]. In Chapter 3, we study the various learning strategies of RBMs as a function of the properties of the training data [46]. In particular, we show how RBMs that are larger than necessary to learn the data adopt an implicitly low-dimensional learning strategy. In Chapter 4, we develop a theory of implicitly low-dimensional learning in a kind of RBM that belongs to the class of *dense associative memory* (DAM), which is a generalization of dense HNs [47]. We then use our theory to greatly accelerate training.

1.2 The Hopfield network: a simple example of energy-based model

In this Section, we present HNs as simple examples of energy-based models. This exercise is useful for building intuition before tackling RBMs and dense HNs, which are more complicated.

HNs use their units $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{N} = (\sigma_1, \sigma_2, ..., \sigma_N)$ to represent data whose components can take only two values, such as binary black-and-white images (see Fig. 1.6, left panel). Without loss of generality, we take these two values to be $+1$ and -1. In this case, the HN Hamiltonian takes the form

$$H_{\text{HN}}\left[\boldsymbol{\sigma}; \mathbf{J}\right] = -\sum_{i=1}^{N}\sum_{j=i+1}^{N} J_{ij}\sigma_i\sigma_j = -\sum_{i<j=1}^{N} J_{ij}\sigma_i\sigma_j, \tag{1.1}$$

where the weights J_{ij} are the strengths of the connections between different components σ_i and σ_j. Contrary to those of the NNs shown in Fig. (1.2), these connections are bidirectional: they are meant to represent how σ_i and σ_j are correlated rather than how one influences the other (see Fig. 1.6, right panel). The sums in the Hamiltonian (Eq. 1.1) are restricted to $j > i$ so that the contribution of every connection to the total energy is counted exactly once.

There are many ways to fit the weights J_{ij} to data [49, 50, 51, 52], the simplest one being Hebb's rule [23, 53]

$$J_{ij} = R_{ij}\left(\boldsymbol{\xi}\right) = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \tag{1.2}$$

where $i < j$ and each pattern $\boldsymbol{\xi}^\mu = \{\xi_i^\mu\}_{i=1}^{N}$ in the set $\boldsymbol{\xi} = \{\boldsymbol{\xi}^\mu\}_{\mu=1}^{P}$ is a data point given to the model. This

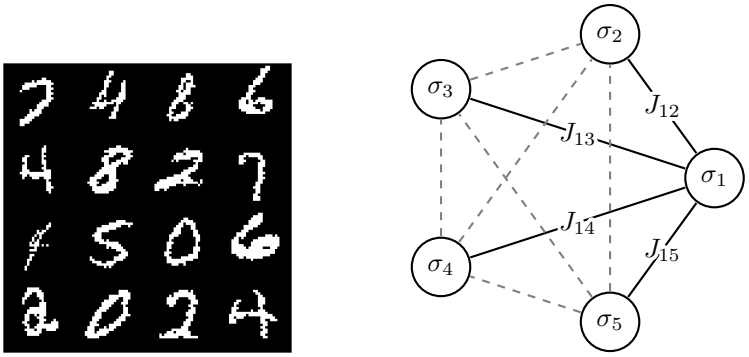

Figure 1.6: In the left panel, images from the binarized version [48] of the MNIST dataset of handwrittern digits [8]. The pixels of these images have two possible values: one for white and the other for black. In the right panel, the network representation of the Hopfield network (HN) Hamiltonian $H\left[\boldsymbol{\sigma};\mathbf{J}\right]=-\sum_{i=1}^{N}\sum_{j=i+1}^{N}J_{ij}\sigma_i\sigma_j$ with $N=5$. The connections J_{ij} are bidirectional, so they are drawn as ordinary lines rather than arrows.

rule establishes strong positive connections between neurons σ_i and σ_j that consistently take the same value ($\sigma_i=\sigma_j$), strong negative connections between neurons that consistently take opposite values ($\sigma_i=-\sigma_j$) and weak connections between neurons that do not synchronize in either of these ways. It is often summarized by the mantra "neurons that fire together, wire together. Neurons that do not sync, fail to link".

HNs can retrieve the patterns stored in their weights, thus acting as models of associative memory [23] and making Hebb's rule a form of "machine memorization" (MM). To be more precise, once its weights have been determined, an HN can be driven to local minima of the energy as described in [23]. When the weights are given by Hebb's rule with a sufficiently small number P of patterns [23, 54], the units $\boldsymbol{\sigma}$ usually converge to one of the patterns $\boldsymbol{\xi}^\mu$ or to a mixture of them [55, 56]. Conversely, when the number P of patterns exceeds a specific threshold proportional to the number N of components per data point, the HN typically no longer converges to meaningful results. In other words, it has limited storage capacity [23, 54].

As we argued in Section 1.1, learning a generic function and a probability distribution are two sides of the same coin. This observation suggests that we could learn the patterns $\boldsymbol{\xi}^\mu$ by fitting a probability distribution of the energy $H_{\text{HN}}\left[\boldsymbol{\sigma};R\left(\boldsymbol{\xi}\right)\right]=-\frac{1}{N}\sum_{i<j=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j$ (see Eqs. 1.1 and 1.2) to the data, as studied in [40, 43, 51], rather than setting the patterns directly equal to the data. For this purpose, we turn to statistical mechanics.

In statistical mechanics, a physical system exchanging energy with the environment often does not converge to a single well-defined equilibrium configuration, but rather to an *equilibrium distribution* of configurations such that more probable configurations have lower energy. The ambient temperature T controls the level of noise in the distribution in the same way as the variance σ^2 in the Gaussian distribution (see Section 1.1). In particular, as T decreases, minimum-energy configurations become more probable. Let us illustrate this point using water. At high temperature, water molecules are randomly distributed in space as vapor. With decreasing temperature, they eventually condense into ice, which is their minimum-energy configuration. Formally, the equilibrium distribution of configurations c with energy $H\left[c\right]$ is proportional to $\exp\left(-\frac{1}{T}H\left[c\right]\right)$ at temperature T. We say that they follow the *(Boltzmann)-Gibbs* distribution, a property that is assumed to hold analogously

for energy-based models. In HNs, this choice leads to the distribution

$$\mathrm{P}\left(\boldsymbol{\sigma} \mid \boldsymbol{\xi}\right) = \frac{\exp\left(-\frac{1}{T} H_{\mathrm{HN}}\left[\boldsymbol{\sigma}; R\left(\boldsymbol{\xi}\right)\right]\right)}{\sum_{\boldsymbol{\sigma}' \in \{-1,1\}^N} \exp\left(-\frac{1}{T} H_{\mathrm{HN}}\left[\boldsymbol{\sigma}'; R\left(\boldsymbol{\xi}\right)\right]\right)} \tag{1.3}$$

for the state $\boldsymbol{\sigma}$ given the patterns $\boldsymbol{\xi}$, where $\{-1,1\}^N$ is the set of all possible values of $\boldsymbol{\sigma}$, which we recall contain only $+1$'s and -1's. The denominator of Eq. (1.3) normalizes the distribution so that the total probability $\sum_{\boldsymbol{\sigma} \in \{-1,1\}^N} \mathrm{P}\left(\boldsymbol{\sigma} \mid \boldsymbol{\xi}\right)$ is equal to 1, and the temperature T is an abstract quantity representing the level of noise injected into $\boldsymbol{\sigma}$. Following statistical mechanics conventions, we define the *partition function* $Z\left(\boldsymbol{\xi}\right) = \sum_{\boldsymbol{\sigma}' \in \{-1,1\}^N} \exp\left(-\frac{1}{T} H_{\mathrm{HN}}\left[\boldsymbol{\sigma}'; R\left(\boldsymbol{\xi}\right)\right]\right)$ and the *inverse temperature* $\beta = 1/T$. We then write the Gibbs distribution as

$$\mathrm{P}_\beta\left(\boldsymbol{\sigma} \mid \boldsymbol{\xi}\right) = Z_\beta\left(\boldsymbol{\xi}\right)^{-1} \exp\left(-\beta H_{\mathrm{HN}}\left[\boldsymbol{\sigma}; R\left(\boldsymbol{\xi}\right)\right]\right), \tag{1.4}$$

where we use the subscript β to explicitly mark the dependence of $\mathrm{P}\left(\boldsymbol{\sigma} \mid \boldsymbol{\xi}\right)$ and $Z\left(\boldsymbol{\xi}\right)$ on β. Now that we have derived this probability distribution, we explain how to learn the patterns. Note that the inverse temperature β and the number of patterns P are not learned. We call such parameters *hyperparameters*.

A relatively intuitive way to fit Eq. (1.4) to a dataset $\mathcal{D} = \{\boldsymbol{\sigma}^a\}_{a=1}^M = \left(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, ..., \boldsymbol{\sigma}^M\right)$ is to find the patterns that make the dataset the most probable, a procedure known as maximum likelihood estimation. The probability that the Gibbs distribution (Eq. 1.4) produces a specific data point $\boldsymbol{\sigma}^a$ from such a dataset is $\mathrm{P}_\beta\left(\boldsymbol{\sigma}^a \mid \boldsymbol{\xi}\right)$. Similarly, the probability that it generates the whole dataset is $\mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right) = \prod_{a=1}^M \mathrm{P}_\beta\left(\boldsymbol{\sigma}^a \mid \boldsymbol{\xi}\right) = \mathrm{P}_\beta\left(\boldsymbol{\sigma}^1 \mid \boldsymbol{\xi}\right) \times \mathrm{P}_\beta\left(\boldsymbol{\sigma}^2 \mid \boldsymbol{\xi}\right) \times ... \times \mathrm{P}_\beta\left(\boldsymbol{\sigma}^M \mid \boldsymbol{\xi}\right)$. Maximum likelihood estimation amounts to finding the patterns $\boldsymbol{\xi}$ that maximize $\mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right)$. It is usually formulated as maximizing the *log likelihood*

$$L\left(\boldsymbol{\xi}\right) = \frac{1}{M} \log \mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right) = \frac{1}{M} \sum_{a=1}^M \log \mathrm{P}_\beta\left(\boldsymbol{\sigma}^a \mid \boldsymbol{\xi}\right), \tag{1.5}$$

which is mathematically equivalent and easier to handle numerically. Maximum likelihood estimation is popular for training energy-based models and NNs because there are various efficient ways to perform it numerically [24, 57, 58, 59], which we will not explain in detail here. In particular, it was used to train HNs in [40]. For historical reasons, maximum likelihood training of NNs is generally phrased as minimizing a *loss function* equal to minus the log likelihood, and the shape of the loss as a function of the learnable parameters is called the *loss landscape*.

Alternatively, we can also train the patterns $\boldsymbol{\xi}$ by sampling the *posterior* distribution $\mathrm{P}_\beta\left(\boldsymbol{\xi} \mid \mathcal{D}\right)$, which represents the probability that an HN with patterns $\boldsymbol{\xi}$ generates the data $\mathcal{D}$. Although this algorithm is generally less efficient from a numerical point of view, it can be studied analytically to characterize the fundamental limits of NN training [60], which was applied to HNs [43], RBMs [42, 44], linear models [61, 62] and far beyond [22]. The analytical calculations presented in this thesis follow this approach, which also ends up providing information about the loss landscape in Chapter 4 [47]. To understand how to calculate $\mathrm{P}_\beta\left(\boldsymbol{\xi} \mid \mathcal{D}\right)$ from $\mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right)$, we first study a similar problem in a more familiar setting. Suppose that we know the probability $\mathrm{P}\left(\text{the ground is wet} \mid \text{it rains}\right)$ that the ground is wet when it rains. We want to know the probability $\mathrm{P}\left(\text{it rains} \mid \text{the ground is wet}\right)$ that it rains when the ground is wet to predict the weather. The probability that it rains *and* that the ground is wet is equal to the probability that it rains when the ground is

wet *multiplied by* the probability that the ground is wet. Formally, this relationship is written as

$$\mathrm{P}\left(\text{it rains}, \text{the ground is wet}\right) = \mathrm{P}\left(\text{it rains} \mid \text{the ground is wet}\right) \mathrm{P}\left(\text{the ground is wet}\right).$$

Crucially, it also works in the other direction, that is

$$\mathrm{P}\left(\text{it rains}, \text{the ground is wet}\right) = \mathrm{P}\left(\text{the ground is wet} \mid \text{it rains}\right) \mathrm{P}\left(\text{it rains}\right).$$

Combining both of these equalities, we find

$$\mathrm{P}\left(\text{it rains} \mid \text{the ground is wet}\right) = \frac{\mathrm{P}\left(\text{the ground is wet} \mid \text{it rains}\right) \mathrm{P}\left(\text{it rains}\right)}{\mathrm{P}\left(\text{the ground is wet}\right)},$$

which is known as Bayes' theorem. In the case of HNs, it gives

$$\mathrm{P}_\beta\left(\boldsymbol{\xi} \mid \mathcal{D}\right) = \frac{\mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right) \mathrm{P}_\beta\left(\boldsymbol{\xi}\right)}{\mathrm{P}_\beta\left(\mathcal{D}\right)}, \tag{1.6}$$

where the *prior* $\mathrm{P}_\beta\left(\boldsymbol{\xi}\right)$ represents the knowledge that we have about the distribution of the patterns before training and $\mathrm{P}_\beta\left(\mathcal{D}\right) = \sum_{\boldsymbol{\xi} \in \{-1,1\}^{N \times P}} \mathrm{P}_\beta\left(\mathcal{D} \mid \boldsymbol{\xi}\right) \mathrm{P}_\beta\left(\boldsymbol{\xi}\right)$ normalizes the distribution in the same way as $Z_\beta\left(\boldsymbol{\xi}\right)$ in the Gibbs distribution (Eq. 1.4). By analogy, we call the normalization constant $\mathrm{P}_\beta\left(\mathcal{D}\right)$ the *posterior partition function* and write it as $\mathcal{Z}_\beta\left(\mathcal{D}\right)$.

1.3 Restricted Boltzmann machines and dense Hopfield networks

As for HNs, RBMs and dense HNs are energy-based models whose corresponding Hamiltonians and Gibbs distributions can be fit on data using maximum likelihood estimation or posterior sampling. Here, we summarize the differences between them.

The generalized HNs that are nowadays called *dense* were developed shortly after the original to improve upon its storage capacity [25, 26, 27, 28, 29]. These models have been receiving renewed attention under the name *dense associative memory* (DAM) since they were used for pattern recognition in [30]. In the process, the meaning of the term dense associative was also broadened to encompass more general models with large storage capacities [63, 64], which were in turn related to various other ML paradigms [38, 65, 66, 67, 68]. This evolution in terminology is the reason why we adopt the convention of calling the generalized HNs first considered in [25, 26, 27, 28, 29] dense HNs rather than DAMs. The dense HN Hamiltonian is

$$H_{\text{DHN}}\left[\boldsymbol{\sigma}; \mathbf{J}\right] = -\sum_{i_1 < \ldots < i_p = 1}^{N} J_{i_1 \ldots i_p} \sigma_{i_1} \ldots \sigma_{i_p}, \tag{1.7}$$

$$\text{where} \quad J_{i_1 \ldots i_p} = R_{i_1 \ldots i_p}\left(\boldsymbol{\xi}\right) = \frac{p!}{N^{p-1}} \sum_{\mu=1}^{M} \xi_{i_1}^{\mu} \ldots \xi_{i_p}^{\mu}.$$

The weights of Eq. (1.7) connect the units $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{N} = (\sigma_1, \sigma_2, ..., \sigma_N)$ within groups of size p. These

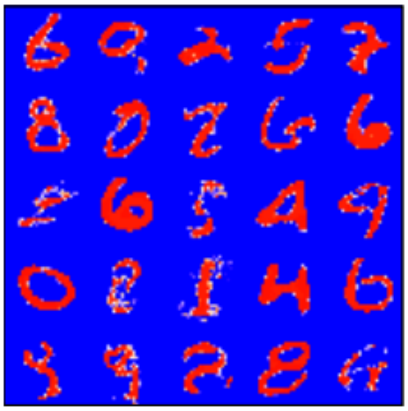

Figure 1.7: Patterns $\boldsymbol{\xi}^\mu$ learned by the dense Hopfield network (HN) studied in [30] when trained on the MNIST dataset of handwritten digits. This plot comes from [30], which contains additional details about the dense HN in question.

connections, which are called *p-body interactions*, reduce to those of the original HN when $p = 2$.[1]

Dense HNs bridge the gap between ML and MM in the sense that they learn prototypes of the data [30] when trained with an ML algorithm. For example, their patterns become prototypes of digits when trained on the MNIST dataset of handwritten digits (see Fig. 1.7). These dense HNs were observed to have a variable level of adversarial robustness as a function of p [36], offering an interesting opportunity to study adversarial attacks and how to mitigate them.

Although they have 2-body interactions like the original HN, RBMs differ from it in that they represent data on two levels of abstraction. Intuitively, these two levels are roughly analogous to how we see objects and how we describe them mentally (for example, thinking of a fire as "hot", "red" and "bright"). In practice, however, the abstract concepts that RBMs associate with data are different from our own mental representations. Formally, the RBM Hamiltonian is

$$H_{\text{RBM}}\left[\mathbf{v}, \mathbf{h}; \mathbf{J}\right] = -\sum_{i=1}^{N}\sum_{\mu=1}^{P} J_i^\mu v_i h_\mu, \tag{1.8}$$

where the visible layer $\mathbf{v} = \{v_i\}_{i=1}^N = (v_1, v_2, ..., v_N)$ is the data expressed in terms of its components, the hidden layer $\mathbf{h} = \{h_\mu\}_{\mu=1}^P$ is a high-level representation of the data in terms of abstract concepts, and the learnable weights $\mathbf{J} = \{J_i^\mu\}_{1\leq i\leq N}^{1\leq \mu\leq P}$ are connections that the RBM makes between $\mathbf{v}$ and $\mathbf{h}$. As in HNs, these connections are bidirectional, which means that they represent how $\mathbf{v}$ and $\mathbf{h}$ are correlated rather than how one influences the other (see Fig. 1.8). The units v_i and h_μ of the two layers can be bound to ± 1 as those of HNs, but various other choices are also possible. For example, they can be integers within an interval [69] or arbitrary real numbers, thus defining multiple types of RBMs [51]. Every such RBM has a Gibbs distribution of the form

$$\mathrm{P}_\beta\left(\mathbf{v}, \mathbf{h} \mid \mathbf{J}\right) = Z_\beta\left(\mathbf{J}\right)^{-1} \mathrm{P}_0\left(\mathbf{v}\right) \mathrm{P}_0\left(\mathbf{h}\right) \exp\left(-\beta H\left[\mathbf{v}, \mathbf{h}; \mathbf{J}\right]\right), \tag{1.9}$$

where the inverse temperature β regulates the strength of the interactions that $\mathbf{v}$ and $\mathbf{h}$ exchange through H,

[1] Except for a factor of two that can be removed by rescaling the inverse temperature.

the priors $\mathrm{P}_0(\mathbf{v})$ and $\mathrm{P}_0(\mathbf{h})$ are the distributions of $\mathbf{v}$ and $\mathbf{h}$ when they are decoupled, i.e. $\beta = 0$, and the partition function $Z_\beta(\mathbf{J})$ is again defined so that the total probability is equal to 1. The priors on both layers, which represent the knowledge that we have on the data before training, usually have a simple form such that their units are (a priori) independent and identically distributed (i.i.d.) with respect to each other. For example, neurons restricted to ± 1 usually have a *Rademacher* prior

$$\mathrm{P}_0(v_i) = \begin{cases} 1/2 & \text{if } v_i = \pm 1 \\ 0 & \text{otherwise,} \end{cases}$$

and units that can be arbitrary real numbers usually have a Gaussian prior $\mathrm{P}_0(v_i) = \frac{\exp\left(-\frac{1}{2\sigma^2} v_i^2\right)}{\sqrt{2\pi\sigma^2}}$. Simple priors with non-i.i.d. units are also possible. For example, RBM layers whose units are restricted to always be at a distance of $\sqrt{\sum_{i=1}^{N} v_i^2} = 1$ from the origin can be described using the *hyperspherical* prior

$$\mathrm{P}_0(\mathbf{v}) = \begin{cases} \frac{1}{\Omega_N(\beta)} & \text{if } \sqrt{\sum_{i=1}^{N} v_i^2} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\Omega_N(0)$ is the size of the region of space where $\sqrt{\sum_{i=1}^{N} v_i^2} = 1$, which is called the unit *hypersphere*. Additionally, the categorical (or *Potts* [70, 71]) prior

$$\mathrm{P}_0(\mathbf{v}) = \begin{cases} \frac{1}{N+1} & \text{if } \mathbf{v} \in \{0,1\}^N \text{ and } \sum_{i=1}^{N} v_i \in \{0,1\} \\ 0 & \text{otherwise,} \end{cases}$$

is used in RBM layers where at most one unit v_i can be activated at once (equal to 1), while all the others are turned off (equal to 0).

Given that the priors have a simple form, using a higher temperature $T = 1/\beta$ encourages the RBM to learn simpler weights J_i^μ during training, which is known as *regularization* in ML. Intuitively, regularization encourages NNs not to overthink problems with relatively simple solutions, which can help them perform better.

RBMs are trained by fitting to the data their marginal distribution $\mathrm{P}_\beta(\mathbf{v} \mid \mathbf{J}) = \int_{\mathbf{h}} d\mathbf{h}\, \mathrm{P}_\beta(\mathbf{v}, \mathbf{h} \mid \mathbf{J})$, which embodies how they represent the data with all their hidden units simultaneously. The marginal distribution often does not have a simple closed form and must be approximated, such as by repeatedly sampling the conditional distributions $\mathrm{P}_\beta(\mathbf{h} \mid \mathbf{v}, \mathbf{J})$ and $\mathrm{P}_\beta(\mathbf{v} \mid \mathbf{h}, \mathbf{J})$ following the contrastive divergence algorithm introduced in [24]. In some cases, the marginal distributions of RBMs with specific priors reduce to the Gibbs distributions of other energy models [72]. For example, the marginal distribution of an RBM with a Rademacher prior on its visible units and a Gaussian prior on its hidden units simplifies to

$$\mathrm{P}_\beta(\mathbf{v} \mid \mathbf{J}) = Z_\beta(\mathbf{J})^{-1} \exp\left(\beta \sum_{i<j=1}^{N} \sum_{\mu=1}^{P} J_i^\mu J_j^\mu v_i v_j\right),$$

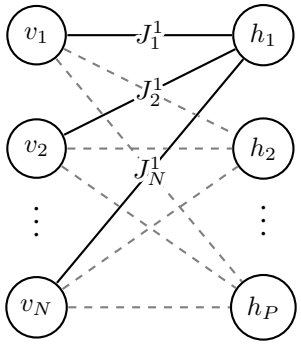

Figure 1.8: The network representation of the restricted Boltzmann machine (RBM) Hamiltonian $H\left[\mathbf{v}, \mathbf{h}; \mathbf{J}\right] = -\sum_{i=1}^{N}\sum_{\mu=1}^{P} J_i^\mu v_i h_\mu$. The connections J_i^μ are bidirectional, so they are represented using ordinary lines instead of arrows.

which is the Gibbs distribution of a Hopfield network with inverse temperature βN, memorized patterns $\mathbf{J}^\mu = \{J_i^\mu\}_{i=1}^N$ and units $\mathbf{v}$. In other words, RBMs with these priors are formally equivalent to HNs [51]. By analogy, the rows $\mathbf{J}^\mu$ of the weights are occasionally called patterns and written as $\boldsymbol{\xi}^\mu$ even for other choices of priors.

RBMs with a hyperspherical prior on their visible units and a categorical prior on their hidden units become *dense associative memory* (DAM) models that learn prototypes of the data like dense HNs. Formally, the marginal distribution of such an RBM simplifies to

$$\mathrm{P}_\beta\left(\mathbf{v} \mid \mathbf{J}\right) = \frac{1}{P+1}\sum_{\mu=1}^{P}\frac{\exp\left(\beta\sum_{i=1}^{N} J_i^\mu v_i\right)}{\Omega_N\left(\beta\right)} + \frac{1}{P+1}\frac{1}{\Omega_N\left(0\right)}, \tag{1.10}$$

where the normalization constant $\Omega_N\left(\beta\right)$ is defined in Chapter 4. In that Chapter, we discuss such DAMs and their affinity for prototype learning in more detail.

1.4 Statistical mechanics of networks in the teacher-student setting

We now present the main theoretical tool that we use to study the statistical mechanics of RBMs and dense HNs: the *teacher-student setting* [61]. This tool has been used numerous times to solve various fundamental problems in the field of ML and NNs [22].

In the teacher-student setting, a *teacher* NN with weights $\mathbf{J}^*$ generates a sample dataset $\mathcal{D}$ on which a *student* NN then trains its own weights $\mathbf{J}$. In other words, the teacher shows the dataset $\mathcal{D}$ to the student to teach it how to mimic its contents. This framework provides a controlled environment for studying NN behavior with analytical calculations. Although NNs with suitable weights can represent arbitrarily complex data [7], the teacher weights $\mathbf{J}^*$ are usually determined according to a simple criterion to simplify calculations. For example, they can be fixed samples from a simple probability distribution. Moreover, in the case of dense HNs, we assume that the weights $\mathbf{J}^*$ and $\mathbf{J}$ decompose into patterns $\boldsymbol{\xi}^{*\mu}$ and $\boldsymbol{\xi}^\nu$ following the second line of Eq. (1.7).

In our work, the student has limited knowledge of the architecture of the teacher. To be more precise, it

knows whether the teacher is a dense HN or an RBM and, in the latter case, the priors on the RBM's neurons. However, unless explicitly stated otherwise, it does not know the hyperparameters of the teacher, such as its inverse temperature β^*, its number of patterns P^*, its interaction order p^*, etc. Therefore, it adopts the same form as the teacher, but with hyperpameters that are generally different, which we write without an asterisk (*) to distinguish them from those of the teacher. Following statistical mechanics conventions, we call the special case where the student has the same hyperparameters as the teacher the *Nishimori line* [73, 74, 75, 76].

Given that the teacher and the student are two instances of the same type of model, the degree of similarity between their weights is a good indication of the performance of the student in learning the data. We measure this degree of similarity using the *overlaps* $Q\left(\boldsymbol{\xi}^{*\mu}, \boldsymbol{\xi}^{\nu}\right) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^{*\mu}\xi_i^{\nu}$ between the patterns of the student and those of the teacher, which we recall are the rows $\mathbf{J}^{\nu}$ and $\mathbf{J}^{*\mu}$ of the weights for RBMs. In the case where all components $\xi_i^{*\mu}$ and ξ_i^{ν} of the patterns $\boldsymbol{\xi}^{*\mu}$ and $\boldsymbol{\xi}^{\nu}$ are equal to ± 1, $Q\left(\boldsymbol{\xi}^{*\mu}, \boldsymbol{\xi}^{\nu}\right) = 1$ means that the student learns the teacher pattern $\boldsymbol{\xi}^{*\mu}$ perfectly. We use other types of overlaps to probe various other aspects of learning. For example, the overlaps $Q\left(\boldsymbol{\xi}^{1\mu}, \boldsymbol{\xi}^{2\nu}\right) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^{1\mu}\xi_i^{2\nu}$ between two learning attempts $\boldsymbol{\xi}^1 = \left\{\xi_i^{1\mu}\right\}_{1\leq i\leq N}^{1\leq \mu\leq P}$ and $\boldsymbol{\xi}^2 = \left\{\xi_i^{2\mu}\right\}_{1\leq i\leq N}^{1\leq \mu\leq P}$ of the student represent its tendency to stay frozen in specific pattern configurations rather than visiting all possible values of $\boldsymbol{\xi}$. $Q\left(\boldsymbol{\xi}^{1\mu}, \boldsymbol{\xi}^{2\nu}\right)$ can sometimes be relatively large even when $Q\left(\boldsymbol{\xi}^{*\mu}, \boldsymbol{\xi}^{\nu}\right)$ is close to 0, which represents the situation where the student mistakenly learns a very different pattern from that of the teacher, or in other words misunderstands what the teacher is trying to teach it. In general, the overlaps change with the hyperparameters of the teacher-student setting and the number of samples M in the training dataset $\mathcal{D}$, defining different regimes of data representation.

Our theoretical analyses are based on computing mean overlaps, which are also known as *order parameters*, using the *replica method* (explained at length in [74]). This method is well-established in the statistical mechanics community, where it is used to study a type of physical system called *spin glass* [22, 74]. Spin glasses are characterized by disordered components that are perpetually frozen, or *quenched*, as the system converges to equilibrium. In the case of NNs in the teacher-student setting, the quenched disorder is the patterns of the teacher and the dataset that it generates, which are kept fixed during learning. Conversely, the student patterns, which are allowed to evolve throughout training, are *not* quenched. Although they can sometimes end up freezing in disordered *spin-glass* states characterized by $Q\left(\boldsymbol{\xi}^{*\mu}, \boldsymbol{\xi}^{\nu}\right) \approx 0$ and $Q\left(\boldsymbol{\xi}^{1\mu}, \boldsymbol{\xi}^{2\nu}\right) > 0$, it is a consequence of the learning task rather than a hard constraint.

We use these calculations to characterize the learning phases of RBMs and dense HNs as a function of the number of samples M in the training dataset $\mathcal{D}$ and the hyperparameters of the teacher-student setting. In particular, we calculate phase diagrams representing the different phases and phase transitions of RBMs and dense HNs as a function of the hyperparameters and the size of the training dataset. For example, Fig. (1.9), which is taken from [43], is the phase diagram of HNs in the teacher-student setting when the student and the teacher have a single pattern and the same inverse temperature $\beta = \beta^*$. We see three phases and two phase transitions (black lines) as a function of the amount of data $\gamma = M/N$ and the temperature $T = 1/\beta$, which we will now explain in terms of the mean teacher-student overlap (subsequently m) and the mean overlap between the teacher pattern and the samples that it generates (subsequently r). Below the first phase transition, at the bottom of the plot, the temperature $T = 1/\beta$ is low enough for the teacher to generate highly informative training samples ($r > 0$) and for the student HN to perfectly learn the teacher pattern through them ($m = 1$). Between the two phase transitions, the training dataset is much noisier ($r = 0$), but the student

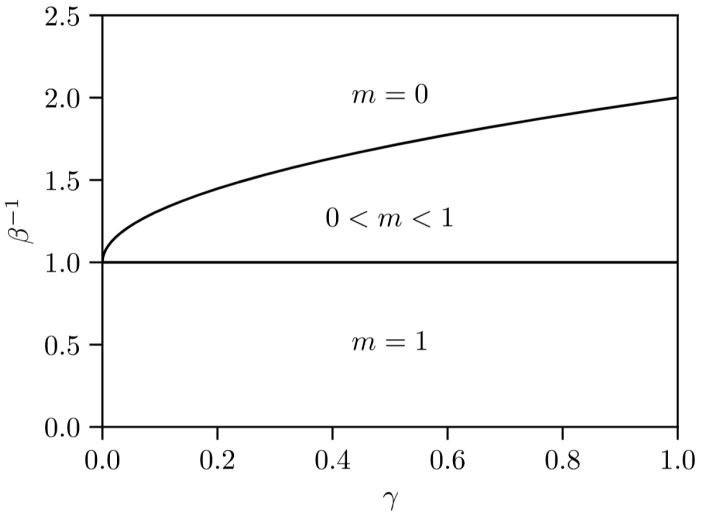

Figure 1.9: Phase diagram of Hopfield networks (HNs) in the teacher-student setting when the student and the teacher have a single pattern and the same inverse temperature $\beta = \beta^*$. The axes measure the amount of data $\gamma = M/N$ and the temperature $T = 1/\beta$. The phases, separated by black lines, are explained in the text. This plot comes from [43].

can still learn the teacher pattern imperfectly ($0 < m < 1$). Finally, above the second phase transition, the temperature $T = 1/\beta$ is too high and the dataset too small for the student HN to learn anything about the teacher pattern ($m = 0$).

1.5 Outline of the thesis

In Chapter 2 (based on a work with Daniele Tantari, published as [45]), we study dense HNs with a single pattern in the teacher-student setting. In particular, we investigate the relationship between a dense HN with a single pattern trained using M data points (the inverse model) and a dense HN with M random memorized patterns (the direct model), the latter of which is well understood thanks to previous studies [25, 26, 27, 28, 29, 30, 41]. To be more precise, we show analytically that the phase where the inverse model with $\beta = \beta^*$ and $p = p^*$ is capable of learning coincides with the phase where the direct model freezes in random states (called the spin-glass phase), refining the well-established *quiet planting* bound that the learning phase of the inverse model extends at least as far as the spin-glass phase of the direct model [60, 77, 78, 79]. Pushing our analysis of this relationship further, we study noise tolerance and adversarial attacks in dense HNs memorizing as patterns the M examples generated by the teacher. In particular, we derive a mathematical formula quantifying the adversarial robustness of dense HNs at zero temperature, and we clarify why the adversarial robustness of the dense HNs studied in [36] changes as a function of the interaction order p. Our results lay the groundwork for a broader statistical mechanics theory of adversarial attacks, contemporaneously with complementary studies of adversarial attacks in linear models [18, 19, 20].

In Chapter 3 (based on a work with Francesco Tosello and Daniele Tantari, published as [46]), we study RBMs in the teacher-student setting where the numbers of hidden units P^* and P are free to be any natural numbers much smaller than the number of samples in the training dataset. This work generalizes the results

of [42], which studied the teacher-student setting with $P^* = P = 2$ hidden units. First, we investigate the relationship between our framework and RBMs in the teacher-student setting with $P^* = P = 1$ hidden units. To be more precise, we show that an RBM with P hidden units learns data generated by a teacher with $P^* = P$ random patterns with the same overlaps as P RBMs with one hidden unit each, thus validating a conjecture formulated in [39]. Going one step further, we study the scenario where the student has more hidden units than necessary to learn the training data, i.e. $P > P^*$. We show that, in this context, only P^* of its patterns end up learning the data, while the others freeze in disordered states. Empirically, we investigate the connection between this low-dimensional learning strategy and the lottery ticket hypothesis according to which a generic randomly initialized overparameterized NN contains subnetworks that fit data with similar accuracy as the entire trained network when they are extracted from it and trained independently [9]. Finally, we show that, when the teacher patterns are correlated together rather than purely random, the student undergoes successive phase transitions where it learns the data in increasingly detailed ways as the size of the training dataset increases.

In Chapter 4 (based on a work with Daniele Tantari, available as a preprint [47]), we study the teacher-student setting of an RBM that fits into the class of *dense associative memory* (DAM) models. This DAM generalizes Eq. (1.10), which is itself a close relative of the DAM studied in [38, 67, 80], to a form capable of both supervised and unsupervised pattern classification. As in Chapter 3, we study a model with a generic number of hidden units. Based on our results, we propose a novel regularization scheme that makes training significantly more stable. Inspired by the observation that dense HNs cross multiple saddle points of the loss landscape during training [34], we study the saddle points of our DAM. We demonstrate that the weights learned by DAMs with a relatively small number of hidden units are saddle points of the loss landscape of larger DAMs (see Eq. 1.5), thus extending studies of an analogous property in related models [72, 81, 82, 83, 84, 85, 86]. We exploit this hierarchy by letting our DAM grow during learning [87, 88], which makes it significantly faster to train as a function of its final size.

Chapter 2

Dense Hopfield networks in the teacher-student setting

Based on the article [45], doi: 10.21468/SciPostPhys.17.2.040

2.1 Introduction

Hopfield networks are artificial neural networks that model associative memory [23]. In the Hopfield model, examples $\sigma \in \{-1,1\}^N$ of memories $\xi^\mu \in \{-1,1\}^N$, $\mu = 1,\dots,M$, are retrieved by sampling the Gibbs distribution of a 2-body Hamiltonian $H[\sigma \mid \xi]$ at a given temperature T [89]. Hopfield networks can be trained in a biologically plausible way using Hebb's rule [23, 53], which leads to $H[\sigma \mid \xi] = -\frac{1}{N}\sum_{\mu=1}^{M}\left(\sum_{i=1}^{N}\xi_i^\mu\sigma_i\right)^2$. However, they can only store up to $M \sim \mathcal{O}(N)$ i.i.d. random memories in the limit of large N [23, 90, 54]. One way to find this scaling is to study the phase diagram of $H[\sigma \mid \xi]$ as a function of the temperature T and load $\alpha = \frac{M}{N}$ [54], where the so-called ferromagnetic phase, which extends up to $\alpha \approx 0.14$, corresponds to accurate retrieval.

Since Hopfield's seminal work, several generalizations have been investigated in relation to their critical storage capacity and retrieval capabilities. For example, parallel retrieval has been studied in relation to pattern sparsity [91, 92, 93, 94, 95] or hierarchical interactions [96, 97, 98, 99, 100], and non-universality has been shown with respect to more general pattern entries and unit priors [39, 101, 102, 103, 104, 105, 106]. Efforts to overcome the $\mathcal{O}(N)$ limitation of the capacity led to the development of a novel class of modern Hopfield networks [38, 107, 63], which are sometimes called dense due to their faculty to store much more memories than the original Hopfield model [30]. These neural networks surpass $\mathcal{O}(N)$ storage capacity by using higher-order interactions instead of the original 2-body couplings [25, 26, 27, 28, 108, 29]. In particular, Gardner [28] calculated the replica-symmetric (RS) phase diagram of the Hamiltonian $H[\sigma \mid \xi] = -\sum_{i_1<\dots<i_p=1}^{N} J_{i_1\dots i_p}\sigma_{i_1}\dots\sigma_{i_p}$ with p-body interactions $J_{i_1\dots i_p} = \frac{p!}{N^{p-1}}\sum_{\mu=1}^{M}\xi_{i_1}^\mu\dots\xi_{i_p}^\mu$ conditioned on i.i.d.

random memories $\xi^\mu \in \{-1,1\}^N$, finding a $M = \mathcal{O}\left(N^{p-1}\right)$ storage capacity. These calculations were later extended to include the effects of one-step replica symmetry breaking (1RSB) [41].

Although they draw a rather detailed picture of the retrieval of individual i.i.d. random memories, these results are not the end of the story. First of all, 1RSB calculations allegedly struggle to find the paramagnetic to spin-glass phase transition accurately at large p because of numerical instability issues [41]. Second of all, dense Hopfield networks have been rapidly gaining a renewed attention for reasons other than their storage capacity since a recent paper [30] by Krotov and Hopfield (K & H), where they were used as a trainable machine learning architecture. For instance, they have been related to transformers [38, 109] and diffusion models [65, 66], and they were found to be significantly more explainable and adversarially robust than feedforward neural networks with ReLU activation functions [30, 36].

One such aspect of dense Hopfield networks that is still poorly understood is their performance as generative models for unsupervised learning, where they are trained over some given dataset to reproduce its probability distribution. As far as we are aware, this problem has not yet been studied theoretically for p-body models with $p \geq 3$. However, it was studied for the original 2-body Hopfield network by using the teacher-student setting [43] first described in [39, 101, 40]. In the teacher-student setting, which is also called inverse problem in opposition to the direct problem of random pattern retrieval, a student model $H\left[\xi \mid \sigma\right]$ is trained with M teacher examples $\sigma^a \sim H\left[\sigma^a \mid \xi^*\right]$ conditioned on the planted pattern ξ^*. In other words, the student tries to infer the pattern ξ^* of the teacher using a structured set of examples σ^a.

At finite load $\alpha = \frac{M}{N}$, two regimes of pattern retrieval were found: example retrieval (eR) and signal retrieval (sR). In the eR phase, the student tries to reconstruct ξ^* by directly retrieving the examples σ^a, which is a good strategy provided that they are strongly correlated with ξ^*. In the sR phase, on the other hand, retrieval is done by extracting subtle cues from weakly correlated examples. The two types of examples used in these two retrieval strategies are respectively called prototypes and features of ξ^* [30]. Interestingly, a prototype regime and a feature regime were also observed by K & H in dense Hopfield networks trained to classify real data [30], where it was found that the prototype regime is significantly more adversarially robust than the feature regime. In other words, the prototype regime is more resistant than the feature regime to small data perturbations that are specifically designed to cause incorrect classification [11, 12]. This prototype approach is arguably a big step towards designing adversarially robust neural networks, a long-standing problem that still lacks a fully satisfying solution [15, 110, 111].

In this work, we study the performance of p-body Hopfield networks in the teacher-student setting, revealing a prototype regime and a feature regime as in the 2-body model. In Section 2.2, we review Gardner's main results in studying p-body Hopfield models and summarize what the rest of the literature on spin-glass models with p-body interactions tell us about the paramagnetic to spin-glass phase transition in p-body Hopfield models. In Section 2.3, we compute the phase diagram of these p-body models in the teacher-student setting. In Section 2.4.1, we discuss the transition to the retrieval phase in the inverse problem. In Section 2.4.2, we compare this retrieval transition against the transition to the spin-glass phase in the direct problem. Despite their different nature, we show that these two transitions are equivalent on the Nishimori line where the teacher and the student have the same p and T [73, 74, 75, 76]. In Section 2.4.3, we discuss the phase diagram on the Nishimori line in more details. In Section 2.4.4 and Section 2.4.5, we discuss the phase diagram outside of the Nishimori line. First of all, we investigate the effect of using an inference temperature different from the dataset noise. Second of all, we reveal that using a larger p for the student than the teacher gives the student an

extensive tolerance against both teacher noise and pattern interference. Finally, in Section 2.4.6, we derive a closed-form expression that measures the adversarial robustness of the student at zero temperature and explain what our results reveal about the nature of adversarial attacks.

2.2 Overview of Gardner's results

Consider the p-body Hamiltonian

$$H\left[\sigma \mid \xi\right] = -\sum_{i_1<\ldots<i_p=1}^{N} J_{i_1\ldots i_p}\sigma_{i_1}\ldots\sigma_{i_p} = -\frac{p!}{N^{p-1}}\sum_{i_1<\ldots<i_p=1}^{N}\sum_{\mu=1}^{M}\xi_{i_1}^{\mu}\ldots\xi_{i_p}^{\mu}\sigma_{i_1}\ldots\sigma_{i_p}\,, \tag{2.1}$$

conditioned on a set of $M = \frac{\alpha N^{p-1}}{p!}$ quenched memories $\xi^{\mu} \in \{-1,1\}^N$, $\mu = 1,\ldots,M$, sampled i.i.d. from the Rademacher distribution $\frac{1}{2}\left[\delta\left(\xi_i^{\mu}-1\right)+\delta\left(\xi_i^{\mu}+1\right)\right]$. In the *direct model*, patterns σ are in turn sampled from the equilibrium Gibbs distribution $P(\sigma \mid \xi) = Z^{-1}e^{-\beta H[\sigma|\xi]}$, where $\beta \geq 0$ is the inverse temperature and $Z = \sum_{\sigma} e^{-\beta H[\sigma|\xi]}$ is the system's partition function. The so-called *direct problem* studied by Gardner [28] consists of quantifying the performance of this model as a method of memory retrieval. In that context, the overlap $\frac{1}{N}\sum_i \xi_i^{\mu}\sigma_i$ is a good measure of retrieval accuracy, and its expected value can be derived from the quenched free entropy $f = \frac{1}{N}\left\langle \log Z\right\rangle_{\xi}$ in the thermodynamic limit $N \to \infty$. At finite p, Gardner used the (non-rigorous) replica trick [112] to evaluate the RS approximation of f (see also Appendix 2.B) in terms of a variational principle of the form

$$f = \lim_{N\to\infty}\frac{1}{N}\left\langle \log Z\right\rangle_{\xi} = \lim_{N\to\infty,L\to 0}\left(\frac{\partial}{\partial L}\left[\frac{1}{N}\log\left\langle Z^L\right\rangle_{\xi}\right]\right) = \underset{m,k,q,k,r}{\mathrm{Extr}}\; f(m,k,q,r)\,,$$

whose solution is

$$\begin{aligned}
q &= \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\tanh^2\left(\beta\left[\sqrt{\alpha r}x + k\right]\right)\,,\\
m &= \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\tanh\left(\beta\left[\sqrt{\alpha r}x + k\right]\right)\,,\\
r &= pq^{p-1}\,,\\
k &= pm^{p-1}\,,
\end{aligned} \tag{2.2}$$

and the order parameters m and q are to be interpreted as expected overlaps. To be more precise, m can be shown to be the expected overlap of a retrieval attempt σ against one memory in the thermodynamic limit, i.e. $m = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \xi_i^{\mu}\sigma_i\right\rangle_{\xi,\sigma}$. Similarly, q is the expected overlap between two retrieval attempts σ^1 and σ^2, i.e. $q = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \sigma_i^1\sigma_i^2\right\rangle_{\xi,\sigma}$ or equivalently $q = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \left\langle\sigma_i\right\rangle_{\sigma}^2\right\rangle_{\xi}$. Intuitively, q measures the tendency of the system to stay frozen in specific configurations rather than visiting all possible values of σ.

The resulting RS phase diagram (see Fig. 2.1) are derived from the value of the order parameters as a function of three *hyperparameters*: the interaction order p, temperature $T = 1/\beta$ and load $\alpha = \frac{Mp!}{N^{p-1}}$. There are four different phases:

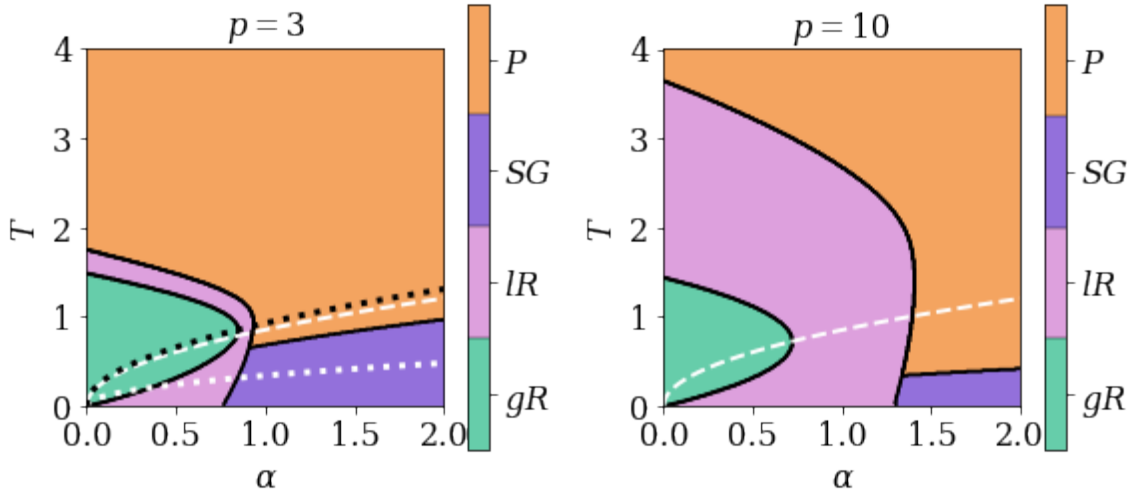

Figure 2.1: RS phase diagrams of the direct models with $p = 3$ on the left and $p = 10$ on the right. Accurate pattern retrieval is not possible in the paramagnetic phase (P) or in the spin-glass phase (SG), but it is possible in the local retrieval phase (lR) and in the global retrieval phase (gR). The ferromagnetic fixed point corresponding to accurate pattern retrieval is globally stable in the gR phase, but locally stable in the lR phase. The phase diagrams are inexact below the white dashed line where the total entropy of the paramagnetic phase becomes negative. The black dotted line overlaying the $p = 3$ diagram is the (exact) 1RSB P-SG transition temperature $T_s\,(\alpha, 3)$, which is obtained by rescaling by $\sqrt{2\alpha}$ the corresponding transition temperature of the spin-glass model with p-body Gaussian interactions. The d1RSB transition $T_d\,(\alpha, 3)$ is very close to $T_s\,(\alpha, 3)$ throughout the displayed range of α. The white dotted line in the $p = 3$ plot is the temperature $T_G\,(\alpha, 3)$ below which multiple steps of RSB are required to compute the free entropy. It is also obtained by rescaling by $\sqrt{2\alpha}$ the corresponding transition temperature of the Gaussian spin-glass model.

- In the Paramagnetic phase (P), the overlaps m and q both vanish. The network does not retrieve any specific pattern: sampled configurations are completely random.
- In the Spin-Glass phase (SG), m vanishes but $q > 0$. In other terms, the network does not retrieve individual stored memories but rather converges to spurious patterns depending on all the memories in a non-trivial way.
- In the signal Retrieval phases (lR and gR), $m \neq 0$ and $q > 0$, which means that the network is able to retrieve the stored memories. lR and gR are respectively locally stable and globally stable. In other words, local retrieval lR is only attainable from initial conditions in a limited neighborhood of a memory ξ^μ, while global retrieval gR is accessible from any initial conditions given enough time. These two phases are said to be ferromagnetic.

Gardner also calculated the exact $p \to \infty$ phase diagram without making any assumptions about replica symmetry [28]. In this limit, the resulting paramagnetic to spin-glass (P-SG) phase transition occurs at a temperature $T_E(\alpha)$ that coincides with the boundary of the region where the total entropy of the paramagnetic phase becomes negative, given by $\beta^2\alpha = 2\log 2$ (white dashed line in Fig. 2.1).

At finite p, Gardner's results only tell us that the model cannot be in the paramagnetic phase below $T_E(\alpha)$. Therefore, a spin-glass transition should occur at a temperature $T_s\,(\alpha, p) \geq T_E(\alpha)$. Since the RS spin glass

solution of Eqs. (2.2) exists only below $T_E(\alpha)$ (violet region in Fig. 2.1), the spin-glass transition must be towards a RSB spin-glass phase.

Outside of the signal retrieval phases, the free entropy of the direct model is the same as for the spin-glass model with p-body Gaussian interactions where the temperature is rescaled by a factor of $\sqrt{2\alpha}$ [113, 114]. Therefore, the spin-glass and paramagnetic solutions are the same in the direct model as in this Gaussian spin-glass model, and we expect the exact phase diagrams of both models to be identical when the direct model is not in its signal retrieval phases. According to previous work on the Gaussian model with finite p [114], a 1RSB solution with $m = k = 0$ exists and is globally stable throughout a whole phase below $T_s(\alpha, p) \geq T_E(\alpha)$ (see Fig. 2.1). This solution becomes unstable at a lower transition temperature $T_G(\alpha, p)$ (see Fig. 2.2), below which multiple steps of RSB are required. In the limit of $p \to \infty$, it holds that $T_s(\alpha, p) \to T_E(\alpha)$ and $T_G(\alpha, p) \to 0$. In other terms, the direct model becomes 1RSB, which is consistent with the fact that it is converging to a random energy model with temperature rescaled by $\sqrt{2\alpha}$ [115, 113, 28]. Finally, we mention that this type of models exhibits a random first order transition phenomenology [116, 117, 118, 119]: there is in fact a range of temperatures $T_s(\alpha, p) \leq T \leq T_d(\alpha, p)$ where the dynamics get trapped in an exponential number of metastable clusters, with an emerging RSB structure that does not affect the free energy (see Fig. 2.2). This range of temperatures thus defines a so-called dynamical 1RSB (d1RSB) phase. Below $T_s(\alpha, p)$, the number of clusters is no longer exponential, and the system undergoes the thermodynamic 1RSB phase transition that we mentioned previously. The critical temperatures $T_G(\alpha, p)$, $T_s(\alpha, p)$ and $T_d(\alpha, p)$ can all be obtained by standard RSB methods, but the resulting saddle-point equations can be prone to numerical instability at large p [41]. In Sections 2.4.2 and 2.4.3, we discuss an alternative way to obtain $T_s(\alpha, p)$ and $T_d(\alpha, p)$.

2.3 Teacher-student setting

On our end, we study a dense Hopfield network with Hamiltonian (2.1) as a generative model for unsupervised learning. In that context, the memories ξ are model parameters that have to be trained in such a way that the examples of a given dataset $\{\sigma^a\}_{a=1}^M$ result as typical network configurations.

In particular, we study a controlled teacher-student setting in which the examples are sampled from the probability distribution $P(\sigma^a \mid \xi^*)$ of a so-called *teacher* dense Hopfield network conditioned on a single *planted* pattern $\xi^* \in \{-1, 1\}^N$ whose entries are quenched Rademacher random variables. A *student* dense Hopfield network, also known as the *inverse model*, then samples its own student pattern ξ from the posterior distribution

$$P(\xi \mid \sigma) = \frac{P(\xi) \prod_{a=1}^M P(\sigma^a \mid \xi)}{P(\sigma)} = \frac{P(\xi)}{P(\sigma)} \prod_{a=1}^M Z^{-1} \exp\left(-\beta H[\sigma^a \mid \xi]\right) ,$$

where $P(\sigma^a \mid \xi)$ is the Gibbs distribution of the direct model with a single memory ξ, and $P(\xi)$ is the prior on ξ that is chosen to be uniform. Since the direct model has only a single pattern, Z does not depend on ξ (see Appendix 2.C), and the posterior simplifies to

$$P(\xi \mid \sigma) = \mathcal{Z}^{-1}(\sigma) \exp\left(-\beta H[\xi \mid \sigma]\right) .$$

In sum, the student posterior distribution is that of a dense Hopfield network where ξ plays the role of the sampled pattern and the examples σ act like the M quenched memories. Our task, called the *inverse problem*, consists of quantifying the student's capability to infer the teacher pattern, which we will also call the *signal*. Like Gardner, we calculate a free entropy of the form $f = \frac{1}{N} \langle \log \mathcal{Z} \rangle_\sigma$ in the thermodynamic limit $N \to \infty$. This time, however, the average $\langle \cdot \rangle_\sigma$ is over a structured set of examples σ. In fact, we recall that, unlike the i.i.d. memories studied by Gardner, the examples σ^a are sampled from the teacher distribution $P(\sigma^a \mid \xi^*)$.

In general, the student does not have access to the teacher generative model. In our controlled teacher-student setting, the student knows that the correct model for $P(\sigma^a \mid \xi)$ is a dense Hopfield network. Nevertheless, it does not necessarily have access to the interaction order p^* and inverse temperature β^* used by the teacher. Therefore, we denote the student hyperparameters by p and β and emphasize that they are not necessarily equal to p^* and β^*. As previously stated, we calculate the free entropy

$$f = \frac{1}{N} \langle \log \mathcal{Z} \rangle_\sigma = 2^{-N} \sum_{\xi^*} \sum_{\sigma} [Z^*]^{-M} \exp\left(\beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a=1}^{M} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \right)$$
$$\times \log \sum_{\xi} \exp\left(\beta \frac{p!}{N^{p-1}} \sum_{a=1}^{M} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \right), \tag{2.3}$$

in the thermodynamic limit $N \to \infty$. We then draw phase diagrams of the inverse problem as a function of p^*, $T^* = 1/\beta^*$, p, $T = 1/\beta$ and α, where α is M normalized to $\mathcal{O}(1)$. Unless explicitly specified otherwise, we use $\alpha = \frac{Mp!}{N^{p-1}}$.

2.3.1 Matched interaction orders

We first consider the case where $p^* = p$ and the only possible mismatch between the teacher and student networks is in the inverse temperature, i.e. $\beta^* \neq \beta$. At low T^*, the student's task is easy. In fact, below the critical temperature T_{crit} of the direct problem with one pattern (see Fig. 2.1, $\alpha = 0$ axis), the teacher produces examples σ^a that cluster around ξ^*. Therefore, the student can infer ξ^* by aligning its pattern ξ with the examples σ^a. This retrieval strategy works even when using a very small amount of examples (see [43]). Since the size of our dataset is extensive, the retrieval accuracy is maximum in the thermodynamic limit. We call this region the (accurate) example Retrieval phase (eR).

Conversely, when T^* is above T_{crit}, the examples in the training set are very noisy and we do not observe a finite overlap between σ^a and ξ^* (see Fig. 2.1, $\alpha = 0$ axis). In this regime, we find that the RS approximation of the $p^* = p$ free entropy can be computed (see Appendix 2.D) in terms of the variational principle

$$f = \underset{m,k,q,r,q^*,r^*}{\text{Extr}} \Bigg\{ \beta^* \beta \alpha \, [q^*]^p - \frac{1}{2} \beta^2 \alpha q^p + \beta m^p - \beta^* \beta \alpha r^* q^*$$
$$+ \frac{1}{2} \beta^2 \alpha r q - \frac{1}{2} \beta^2 \alpha r - \beta m k + \frac{1}{2} \beta^2 \alpha + \log 2$$
$$+ \int dx \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\} \Big\langle \log \left[\cosh\left(\beta \left[\sqrt{\alpha r} x + \beta^* \alpha r^* + k z \right] \right) \right] \Big\rangle_z \Bigg\}, \tag{2.4}$$

whose solution is the saddle-point equations

$$
\begin{aligned}
q^* &= \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \left\langle \tanh\left(\beta\left[\sqrt{\alpha r}x + \beta^*\alpha r^* + kz\right]\right)\right\rangle_z , \\
q &= \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \left\langle \tanh^2\left(\beta\left[\sqrt{\alpha r}x + \beta^*\alpha r^* + kz\right]\right)\right\rangle_z , \\
m &= \int_{\mathbb{R}} dx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \left\langle z\tanh\left(\beta\left[\sqrt{\alpha r}x + \beta^*\alpha r^* + kz\right]\right)\right\rangle_z , \\
r^* &= p\left[q^*\right]^{p-1} , \\
r &= pq^{p-1} , \\
k &= pm^{p-1} ,
\end{aligned}
\tag{2.5}
$$

where z is a Rademacher random variable and $\alpha = \frac{Mp!}{N^{p-1}}$. As in the direct model described in Section 2.2, the order parameters m and q have a clear interpretation in terms of expected overlaps. $m = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \xi_i\sigma_i^a\right\rangle_{\xi^*,\sigma,\xi}$ is the expected overlap of a retrieval attempt with an example σ^a, and $q = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \langle\xi_i\rangle_\xi^2\right\rangle_{\xi^*,\sigma}$ is the expected overlap between two retrieval attempts. Similarly, q^* is the expected overlap between the teacher and student patterns, i.e. $q^* = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \xi_i^*\xi_i\right\rangle_{\xi^*,\sigma,\xi}$. Therefore, it is a good measure of inference performance. The free entropy (Eq. 2.4) is expected to be exact in absence of mismatch between the teacher and the student, i.e. $\beta^* = \beta$. This condition is known as the Nishimori line [73, 74, 75, 76]. Outside of the Nishimori region, RSB corrections are expected. Like the direct problem, the inverse problem with $T^* > T_{\text{crit}}$ has different phases characterized by the values of the order parameters:

- In the Paramagnetic phase (P), the overlaps m, q^* and q all vanish.
- In the signal Retrieval phases (lR and gR), $m = 0$ but $q^* \neq 0$ and $q > 0$. lR and gR are respectively locally stable and globally stable. In other words, local retrieval lR is only attainable from initial conditions in a limited neighborhood of ξ^*, while global retrieval gR is accessible from any initial conditions given enough time. These two phases are also said to be ferromagnetic.
- In the (inaccurate) example Retrieval phase (eR), $m \neq 0$ and $q > 0$ but $q^* = 0$.
- In the Spin-Glass phase (SG), $q > 0$ but q^* and m vanish.

In sum, when T^* is above T_{crit}, the student can only learn the teacher pattern in the signal retrieval phases. In all the other phases, the student pattern is uncorrelated with the signal, being either a random guess (P phase), aligned with a noisy example (inaccurate eR phase), or aligned with a spurious low energy state (SG phase). We stress that we cannot have $m \neq 0$ and $q^* \neq 0$ at the same time (accurate eR phase) when $T^* > T_{\text{crit}}$ because $\lim_{N\to\infty}\left\langle \frac{1}{N}\sum_i \xi_i^*\sigma_i^a\right\rangle_{\xi^*,\sigma} = 0$ in that regime (see Fig. 2.1, $\alpha = 0$ axis).

2.3.2 Mismatched interaction orders

We also investigate the $T^* > T_{\text{crit}}$ regime in the presence of a mismatch between the interaction orders of the teacher and student networks, i.e. $p^* \neq p$. We focus on the case of $p^* = 2$ and even $p \geq 3$ to study the consequences of fitting the teacher of [43] using a student with higher order interactions. We find two

different scaling regimes of the training set size M and inverse temperature β^* that make retrieval possible (see Appendix D):

- a large-noise scaling where $\beta^* \sim \mathcal{O}\left(N^{2/p-1}\right)$ and $M \sim \mathcal{O}\left(N^{p-1}\right)$, such that $\alpha = \frac{Mp!}{N^{p-1}}$ and $\lambda = \frac{[\beta^*]^{p/2}}{(p/2)!} N^{p/2-1}$ are finite;
- a finite-noise scaling where $\beta^* \sim \mathcal{O}(1)$ and $M \sim \mathcal{O}\left(N^{p/2}\right)$, such that $\alpha = \frac{M(p/2+1)!}{N^{p/2}}$ is finite.

In the large-noise scaling, we obtain saddle point equations similar to Eqs. (2.5) but with β^* replaced by λ (see Appendix 2.D). Conversely, the finite noise scaling leads to

$$\begin{aligned} q^* &= \Big\langle \tanh\left(\beta\left[\eta\alpha r^* + kz\right]\right) \Big\rangle_z , \\ m &= \Big\langle z \tanh\left(\beta\left[\eta\alpha r^* + kz\right]\right) \Big\rangle_z , \\ r^* &= p\left[q^*\right]^{p-1} , \\ k &= p m^{p-1} , \end{aligned} \tag{2.6}$$

where η generally depends on β^* and p in a non-trivial way, but we find that $\eta = \frac{2[\beta^*]^2}{(1-2\beta^*)^2}$ when $p = 4$ (see Appendix 2.D). These equations can also be derived by extrapolating the large-noise equations to $\alpha_{\text{large noise}} \to 0$ and $\lambda \to \infty$ with fixed $\lambda\alpha_{\text{large noise}} = \eta\alpha_{\text{finite noise}}$.

2.4 Results and Discussion

2.4.1 Retrieval transition at large interaction order

The paramagnetic solution of Eqs. (2.5) always exists and is globally stable in the part of the phase diagram where the temperature T is relatively large and $\alpha = \frac{Mp!}{N^{p-1}}$ is relatively small. On the other hand, the gR phase exists when $\beta^2\alpha p$ and $\beta^*\beta\alpha p$ are both large. In fact, in that limit, $q^* = q = 1$ is a fixed point of Eqs. (2.5). The critical line where gR becomes globally stable instead of P is not clear from this analysis alone, but we can at least find it analytically in the limit of infinite p. As for the direct model, the free entropy and the total entropy of the paramagnetic phase are respectively $\frac{1}{2}\beta^2\alpha + \log 2$ and $-\frac{1}{2}\beta^2\alpha + \log 2$ [28]. At the same time, the $p \to \infty$ free entropy takes the form

$$\begin{aligned} f = \text{Extr}\Bigg\{ &\beta^*\beta\alpha\,\theta\left(q^* - 1\right) - \frac{1}{2}\beta^2\alpha\,\theta\left(q-1\right) - \beta^*\beta\alpha r^* q^* + \frac{1}{2}\beta^2\alpha r q - \frac{1}{2}\beta^2\alpha r + \frac{1}{2}\beta^2\alpha \\ &+ \log 2 + \int dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \log\left[\cosh\left(\sqrt{\beta^2\alpha r}x + \beta^*\beta\alpha r^*\right)\right] \Bigg\}, \end{aligned}$$

where $\theta\left(q-1\right) := \lim_{p\to\infty} q^p$, $q \in [0,1]$, is the Heaviside step function jumping at $q = 1$, i.e. $\theta(1) = 1$ and $\theta(q) = 0\ \forall q \in [0,1)$. In this limit, the ferromagnetic phase is characterized by $q = q^* = 1$, and its free

entropy is then

$$\begin{aligned} f &= \beta^*\beta\alpha - \beta^*\beta\alpha p + \int dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \log\left[2\cosh\left(\sqrt{\beta^2\alpha p}x + \beta^*\beta\alpha p\right)\right] \\ &\approx \beta^*\beta\alpha - \beta^*\beta\alpha p + \int dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \left(\sqrt{\beta^2\alpha p}x + \beta^*\beta\alpha p\right) \\ &= \beta^*\beta\alpha\,. \end{aligned}$$

The corresponding total entropy is $s = f - \beta\frac{\partial f}{\partial \beta} = 0$, as expected from a ferromagnetic phase with $q^* = q = 1$. On the Nishimori line, $f = \beta^*\beta\alpha$ becomes larger than the free entropy of the paramagnetic phase, which triggers a phase transition, if and only if

$$T < \sqrt{\frac{\alpha}{2\log 2}}\,, \tag{2.7}$$

where $T_E = \sqrt{\frac{\alpha}{2\log 2}}$ is also the temperature below which the total entropy of the paramagnetic phase becomes negative. Outside of the Nishimori line, this inequality generalizes to $\beta^*\beta\alpha > \frac{1}{2}\beta^2\alpha + \log 2$, leading to

$$\beta^* - \sqrt{[\beta^*]^2 - \frac{2\log 2}{\alpha}} < \beta < \beta^* + \sqrt{[\beta^*]^2 - \frac{2\log 2}{\alpha}}\,,$$

while the temperature where the paramagnetic total entropy becomes negative stays the same.

2.4.2 Transition to the ordered phases: Universality

In the $p \to \infty$ limit, the transition towards gR of the inverse model on the Nishimori line is identical to the exact P-SG transition of the direct model [28]. We claim that these two critical lines are actually closely related for any p. In the Hopfield model with $p = 2$, they were already shown to be identical [43]. We will now argue that they overlap for any p and β such that $T > T_{\text{crit}}$ (see Figs. 2.2 and 2.1). In the case of $p = 2$, both lines can be obtained exactly from the RS approximation of either the direct model or the inverse model, so there is no obvious advantage to using this equivalence in calculations. In general, while the inverse problem on the Nishimori line is replica symmetric, the direct problem is not, and the $p \geq 3$ replica symmetric P-SG transition is not exact. Moreover, even the critical line calculated using 1RSB may be inaccurate due to numerical instability [41]. In this situation, the knowledge of the gR transition in the replica-symmetric inverse problem can be used to locate the exact P-SG transition of the direct problem, where symmetry breaking occurs.

For that purpose, we will argue that, given $T > T_{\text{crit}}$, *the direct model is in the paramagnetic phase if and only if the inverse model is in the paramagnetic phase.*

The converse implication comes from the fact that since (see Appendix 2.C)

$$P(\sigma) = \frac{1}{2^{MN}} \frac{\mathcal{Z}(\sigma)}{\langle \mathcal{Z} \rangle}\,, \tag{2.8}$$

the example distribution $P(\sigma)$ of the inverse problem is contiguous [120] to the uniform distribution, i.e. the

memory distribution of the direct problem, when

$$\lim_{N\to\infty}\left\{\frac{\log\mathcal{Z}-\log\langle\mathcal{Z}\rangle}{N}\right\}=0\,. \tag{2.9}$$

As determined in Appendix 2.C and 2.D, the annealed expression $\frac{1}{N}\log\langle\mathcal{Z}\rangle$ is equal to the free entropy of the paramagnetic phase. Therefore, when the inverse model is in the paramagnetic phase, $P(\sigma)$ is contiguous to the uniform distribution. This property is called quiet planting and is known to occur more generally in mean-field paramagnets [77, 78, 79, 60]. In our problem setting, it means that if the inverse model is in the paramagnetic phase, then it is equivalent to the direct model. In particular, if the inverse model is in the paramagnetic phase, then so is the direct model. In more intuitive terms, the gR transition temperature of the inverse model must be greater than or equal to the P-SG transition temperature of the direct model because the ensemble of examples σ^a generated by the teacher model is on average at least as structured as the set of i.i.d. random memories stored in the direct model.

For the direct implication, notice that the average replicated partition function of the direct model in the paramagnetic phase can be approximated as (see Appendix 2.E)

$$\begin{aligned}\left\langle Z^L\right\rangle \approx \frac{1}{\langle Z\rangle}\Bigg\langle \sum_{\sigma}\exp\Bigg(\beta N\sum_{\gamma}\sum_{\mu\in\Gamma_\gamma}\left[\frac{1}{N}\sum_i\xi_i^\mu\sigma_i^\gamma\right]^p \\ +\beta\sum_{\gamma}\sum_{\mu\in\bar{\Gamma}}\frac{p!}{N^{p-1}}\sum_{i_1<\ldots<i_p}\xi_{i_1}^\mu\ldots\xi_{i_p}^\mu\sigma_{i_1}^\gamma\ldots\sigma_{i_p}^\gamma\Bigg) \\ \sum_{\sigma_0}\exp\left(\beta\sum_{\mu\in\bar{\Gamma}}\frac{p!}{N^{p-1}}\sum_{i_1<\ldots<i_p}\xi_{i_1}^\mu\ldots\xi_{i_p}^\mu\sigma_{i_1}^0\ldots\sigma_{i_p}^0\right)\Bigg\rangle\,.\end{aligned}$$

This expression is identical to the replicated partition function of the inverse model with $T>T_{\text{crit}}$, which therefore must also be in the paramagnetic phase.

As a consequence, when $T>T_{\text{crit}}$, the P-SG transition line of the direct model must be identical to the gR transition line of the inverse model on the Nishimori line.

2.4.3 Phase diagram on the Nishimori line

On the Nishimori line, the student is fully informed about the teacher generative model and uses $\beta=\beta^*$ and $p=p^*$. In this scenario, thanks to the Nishimori identities [74], it is well known that ξ^* and ξ play symmetric roles and that $q^*=q$. For the same reason, the overlaps $\frac{1}{N}\sum_i\xi_i^*\xi_i$ and $\frac{1}{N}\sum_i\xi_i^1\xi_i^2$ have the same distribution. From the self-averaging of $\frac{1}{N}\sum_i\xi_i^*\xi_i$, it follows that the system is expected to be replica symmetric, and Eqs. (2.4) and (2.5) are expected to hold. Fig. (2.2) shows the phase diagrams obtained by solving the saddle-point equations numerically on the Nishimori line. Both $q^*=q$ and the replica symmetry condition are verified. In particular, numerical solutions of a few values of $p\geq 3$ show that the gR transition occurs at a higher T than the line $\beta^2\alpha=2\log 2$ where the total entropy of the paramagnetic phase becomes negative. In other terms, the phase transition towards gR prevents the total entropy from becoming negative when T decreases below $\sqrt{\frac{\alpha}{2\log 2}}$, which is consistent with the RS solution being exact on the Nishimori line.

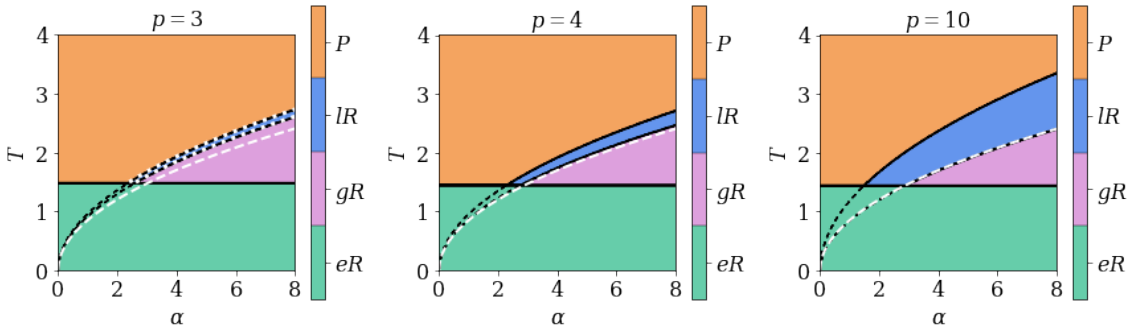

Figure 2.2: Exact RS phase diagrams of inverse models on the Nishimori line, i.e. $p^* = p$ and $\beta^* = \beta$. the left, center and right plots respectively have $p = 3$, $p = 4$ and $p = 10$. Accurate pattern retrieval is not possible in the paramagnetic phase (P), but it is possible in the local retrieval phase (lR), in the global retrieval phase (gR) and in the example retrieval phase (eR). The ferromagnetic fixed point corresponding to accurate pattern retrieval is globally stable in the gR phase, but locally stable in the lR phase. The critical temperature of the eR phase is the critical temperature T_{crit} of the direct problem with one pattern (see Fig. 2.1, $\alpha = 0$ axis). The black dashed lines mark the spurious continuation of the lR and gR phase boundaries through the eR phase. The white dashed line is the $p \to \infty$ gR critical line calculated analytically in Section 2.4.1. It matches the corresponding numerical phase boundary increasingly well as p grows larger. The white dotted lines on the $p = 3$ plot mark the 1RSB and d1RSB critical temperatures $T_s(\alpha, 3)$ and $T_d(\alpha, 3)$ of the direct model (see Section 2.2). We truncated them below T_{crit} for improved visibility. $T_s(\alpha, 3)$ and $T_d(\alpha, 3)$ are obtained by rescaling the corresponding critical temperatures found in [117] by $\sqrt{2\alpha}$.

At low T, the student can learn efficiently within the accurate eR regime. In this phase, learning is possible ($q^* \neq 0$) because the examples are correlated with the signal and the student can retrieve it by simply being aligned with them ($m \neq 0$).

At high T, learning is possible only if the amount of examples, i.e. the size of the dataset, is sufficiently large. When α is too small, Eqs. (2.5) have only a paramagnetic fixed point because the amount of information carried by the dataset is not large enough. Numerical solutions suggest that the paramagnetic fixed point always exist and it is actually locally stable in the whole high-temperature regime. When α is sufficiently large, the signal retrieval fixed point appears as a locally stable attractor (lR phase). It becomes globally stable (gR phase) as the size of the dataset is increased further or the student temperature decreases.

As per the previous Section, the critical boundary of the gR phase obtained by solving Eqs. 2.5 is identical to the 1RSB P-SG transition temperature $T_s(\alpha, p)$ of the direct model. Similarly, we observe that the metastable lR phase coincides with the d1RSB phase of the direct model (see Fig. 2.2). Our results are also consistent with the fact that $T_s(\alpha, p) \to T_E(\alpha)$ in the $p \to \infty$ limit. In fact, we find that the analytical limit boundary closely agrees with the numerical solution of the saddle-point equations with $p^* = p = 10$ and remains a good approximation even down to $p^* = p = 4$.

In the student model, σ plays a similar role as the weights of the trainable dense Hopfield network model that K & H designed for classification of data [30]. In that context, ξ is analogous to the test data whose labels are being predicted (see Fig. 2.3). In fact, the computation performed by K & H's model to recover labels is similar to the update rule used by the student to infer the teacher pattern (see Appendix 2.A). Moreover, the eR and gR phases are respectively reminiscent of the prototype and feature regimes of K & H's networks. Therefore, we believe that the student can act as a toy model of label prediction in these two regimes.

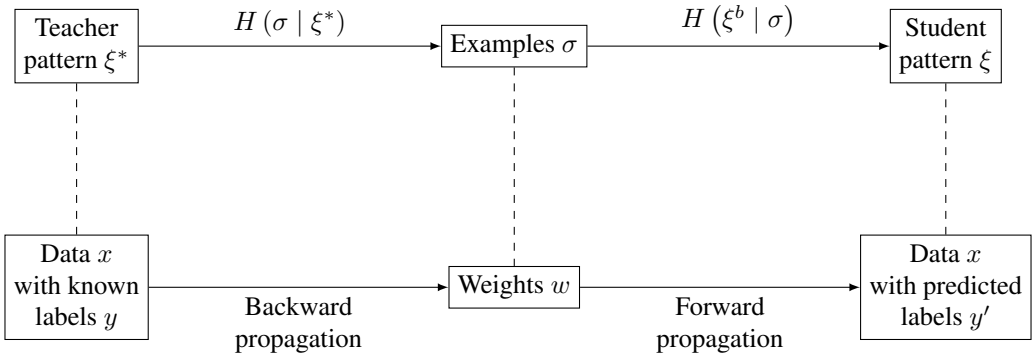

Figure 2.3: The first row of this diagram sketches how a p-body Hopfield network in the teacher-student setting can reconstruct an incomplete pattern ξ^b to match the teacher pattern ξ^* by relying on the examples σ obtained from ξ^*. The second row summarizes how a dense neural network trained by K & H can recover the labels y' of the data x given the weights w learned from x [30]. Both models tackle similar tasks using an approach where σ and ξ^b respectively play the same roles as w and (x, y'). The forward propagation algorithm used to generate y' is similar to the update rule of the student (see [30] and Appendix 2.A), but the backpropagation algorithm used to learn w is very different from the update rule of the teacher.

Comparing instead the phase diagrams of our inverse model with that of the inverse 2-body Hopfield model, we see that the eR and gR phases of the inverse p-body model with $p \geq 3$ are respectively analogous to the eR and sR (signal Retrieval) phases presented in [43]. One of the key differences between $p = 2$ and $p \geq 3$ is that the paramagnetic to signal retrieval phase transition of the p-body model is second order for $p = 2$ but first order for $p \geq 3$. On the one hand, the second order phase transition of $p = 2$ indicates that its paramagnetic fixed point is never locally stable and sets an unambiguous boundary between the sR phase where ξ^* can be recovered starting from any initial conditions and the paramagnetic phase where pattern retrieval is impossible [60]. On the other hand, the first order phase transition of $p \geq 3$ allows the retrieval and paramagnetic regimes to coexist. The lR phase is locally stable precisely because it coexists with the paramagnetic phase and has a lower free entropy. Meanwhile, the gR phase also coexists with the paramagnetic phase, but has a larger free entropy. In the presence of phase coexistence, an algorithm trying to retrieve ξ^* starting from random initial conditions can get stuck in the paramagnetic phase instead. In fact, it has been conjectured that there is no algorithm with random initial conditions that can find such a ferromagnetic fixed point in a tractable amount of time [60, 121]. That kind of metastable region was thus given the name *hard phase* [60, 122]. In summary, we expect that $p \geq 3$ models in the gR phase can only recover partially corrupted patterns whereas $p = 2$ can recover them entirely.

Fig. (2.4) shows results from Monte Carlo simulations with $p = 3$, where L replicas of the student pattern $\{\xi^b\}_{b=1}^{L}$ are initialized to the teacher pattern ξ^* corrupted by some Rademacher noise ε. In other words, the initial values of ξ_i^b are sampled from the distribution $(1-\varepsilon)\,\delta(\xi_i - \xi_i^*) + \frac{\varepsilon}{2}\left[\delta(\xi_i + 1) + \delta(\xi_i - 1)\right]$ with $\varepsilon \in [0, 1]$. The value of ε is tuned so that the simulations start relatively close to the saddle-point solutions. As explained previously, gR is a hard phase, so this initialization is necessary to make ξ^b converge to gR in a reasonable amount of time. Additionally, it is also used to make ξ^b converge to the lR phase rather than the P phase when desired. Once the simulations are over, the overlaps are averaged over all L replicas. If we fix $\varepsilon = 0$, then the simulations generally converge to the lR phase when it is a fixed point. If instead we initialize

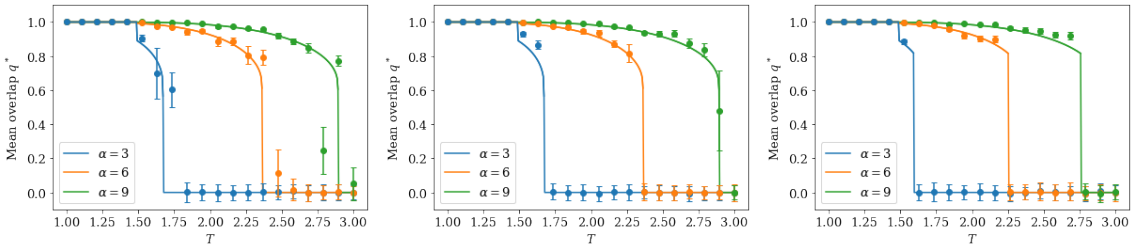

Figure 2.4: Monte Carlo simulations of the $p = 3$ inverse model compared against RS saddle-point solutions. The lR phase is included on the left and central plots, but not on the right one. The left plot has $\varepsilon = 0$, and the two other ones have a handpicked ε such that the simulations are initalized near the saddle-point solutions. The dots are simulation data at a few values of α, and the lines are slices of the saddle-point solutions at the same α. The teacher generates $M = \frac{\alpha N^{p-1}}{p!}$ examples σ^a with $N = 512$ components each, and the simulation results are then averaged over $L = 100$ student patterns. The simulation data is sometimes systematically shifted up with respect to the saddle-point solution. This difference is notably visible on the central plot, right after the fall from eR to gR when $\alpha = 3$.

them to the saddle-point solutions by handpicking ε, then they stay near the initial overlaps. In either case, the simulations converge to eR when it is globally stable. Some simulation data points might be systematically shifted up with respect to the saddle-point solutions. However, this difference decreases with the system size N, so finite size effects seem sufficient to explain it (see Fig. 2.9 in Appendix 2.F). Overall, the Monte Carlo simulations are in very good agreement with the $p = 3$ overlap landscape obtained by solving the saddle-point equations numerically.

2.4.4 Inference temperature vs dataset noise

In the two next Sections, we will discuss the phase diagram when the student is only partially informed about the teacher generative model, i.e. when the Nishimori conditions do not hold. We start with the case where $p = p^*$ but $\beta \neq \beta^*$, i.e. the inference temperature T is different from the dataset noise T^*. As we argued in Section 2.3.1, the student accurately retrieves ξ^* when $T^* < T_{\text{crit}}$. On the other hand, we must solve the saddle-points equations (see Eqs. 2.5) to study $T^* > T_{\text{crit}}$.

We show the phase diagram of this region on Fig. (2.5). At high inference temperature T, the situation is similar to Fig. (2.2): retrieval is possible if the data load α is sufficiently large, but the paramagnetic phase is always locally stable. The situation is different when the inference temperature is low. In that case, there are two phases that we did not see for $\beta = \beta^*$: the inaccurate eR phase and the SG phase. When α is relatively small, the student falls in the inaccurate eR phase. In this regime, it has finite overlap with one of the noisy examples and cannot retrieve the signal ξ^*. When α is larger, the interference among the noisy examples prevents the student to be aligned with them. In this regime, the SG phase, the student locally converge to spurious patterns that are uncorrelated with the signal.

Accurate pattern retrieval is only possible in the lR and gR phases where α is so large that the student can gather enough information from the dataset to become very close to ξ^*. The phase diagrams indicate that pattern retrieval is optimal on the Nishimori line in the sense that $\beta = \beta^*$ is the inverse temperature where the student needs the least examples to recover ξ^*. In other words, the student's performance is non-monotonic

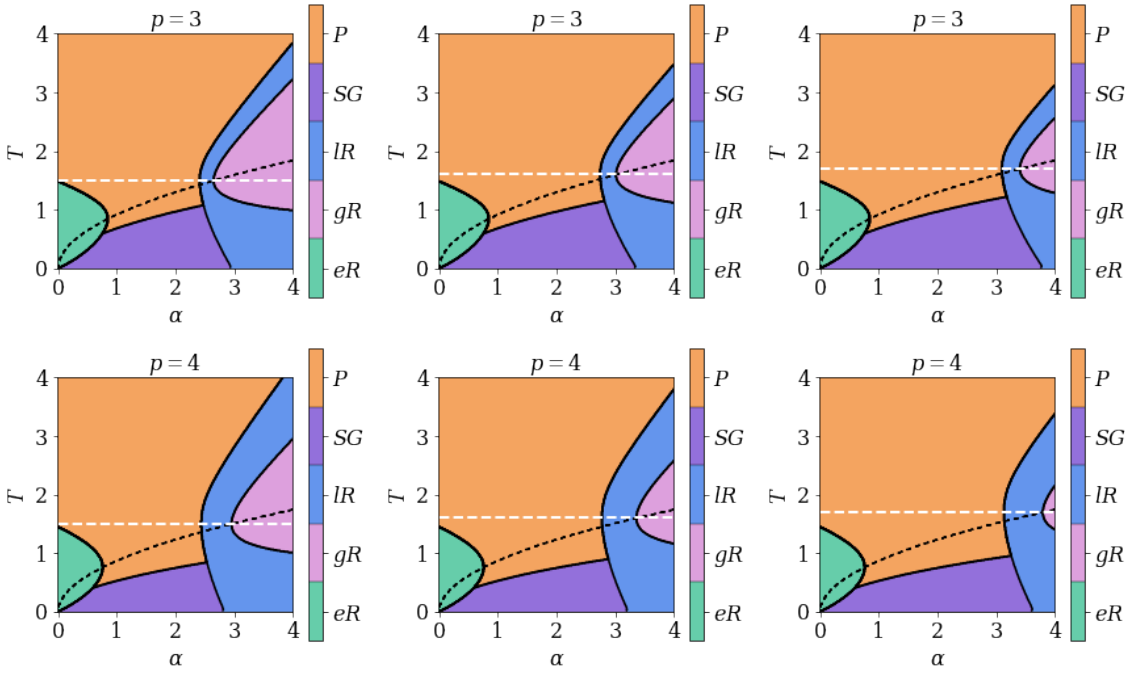

Figure 2.5: RS phase diagrams of inverse models with $p^* = p$ and fixed β^*. The top and bottom rows of plots respectively have $p^* = p = 3$ and $p^* = p = 4$. In the same way, the left, central and right columns correspond to $T^* = 1.5$, $T^* = 1.6$ and $T^* = 1.7$. Accurate pattern retrieval is not possible in the paramagnetic phase (P), in the spin-glass phase (SG) or in the example retrieval phase (eR), but it is possible in the local retrieval phase (lR) and in the global retrieval phase (gR). The ferromagnetic fixed point corresponding to accurate pattern retrieval is globally stable in the gR phase, but locally stable in the lR phase. Conversely, the SG fixed point is always locally stable and leads the student to a frozen spurious signal. The white dashed line indicates the Nishimori line $\beta^* = \beta$. The black dashed lined is the gR phase boundary on the Nishimori line. As explained in Section 2.4.3, we expect it to overlap the exact SG phase transition.

in T and peaks at $T = T^*$. These properties were also observed in the teacher-student setting of the $p = 2$ Hopfield network [43].

Contrary to what one would expect to see on the exact phase diagram [73, 74], the Nishimori line $T = T^*$ does not to cross a triple point on the RS phase diagram. The issue is that the RS phase diagram is not exact outside of the Nishimori line. In particular, the SG phase boundary is not exact. Outside of the retrieval regime, the free entropy of the inverse model is the same as the direct model. Since the transition towards gR of the inverse model on the Nishimori line overlaps the exact P-SG transition of the direct model (see Section 2.4.3), we deduce that it must also overlap the exact P-SG transition of the *inverse* model outside of the gR phase. Plotting it on the RS phase diagrams, we see that it indeed crosses the Nishimori line and the gR phase boundary at the same point, which therefore becomes a triple point, as expected.

2.4.5 Interaction order and noise tolerance

So far, we assumed that the student is informed about the interaction order used by the teacher, i.e. $p = p^*$. In this Section, we investigate the role of the student's choice of p when the task is to learn from a dataset sampled by a 2-body Hopfield network, i.e. $p^* = 2$. We study two different non trivial scalings regimes of M and β^* that make pattern inference possible (see Appendix 2.D).

2.4.5.1 Large noise scaling

We first consider a large noise scaling where $\beta^* \sim \mathcal{O}\left(N^{2/p-1}\right)$ and $M \sim \mathcal{O}\left(N^{p-1}\right)$, such that

$$\alpha = \frac{Mp!}{N^{p-1}}\,, \quad \text{and} \quad \lambda = \frac{[\beta^*]^{p/2}}{(p/2)!} N^{p/2-1}\,,$$

are finite. In this scaling, a $p \geq 3$ network requires $\mathcal{O}\left(N^{p-2}\right)$ more training examples than a $p = 2$ network with finite load $\gamma = \frac{M}{N}$, but also has a higher tolerance to teacher noise. For instance, a student with $p = 4$ interactions is able to retrieve the pattern of a teacher with $T^* \sim \mathcal{O}\left(N^{1/2}\right)$ noise when it is shown enough examples $M \sim \mathcal{O}\left(N^3\right)$ to be in the gR phase (see Fig. 2.6).

$\mathcal{O}\left(N^{1/2}\right)$ noise tolerance was also observed in the $p = 4$ direct model, where it is a consequence of the redundancy stemming from storing $\mathcal{O}\left(N\right)$ memories rather than the $\mathcal{O}\left(N^3\right)$ needed to saturate the storage capacity [123]. Our $p = 4$ inverse model exploits a different kind of redundancy by learning from $\mathcal{O}\left(N^3\right)$ examples whereas $p = 2$ only needs $\mathcal{O}\left(N\right)$. In other terms, both storing extensively less memories than the maximum allowed amount and generating extensively more examples than the minimum required amount provide enough redundancy to recover a pattern muddled in an extensive amount of noise. In both cases, there is an $\mathcal{O}\left(N^2\right)$ gap between the number of patterns used in the noise-tolerant and noise-susceptible regimes. Going beyond $p = 4$, the inverse model has $\mathcal{O}\left(N^{1-2/p}\right)$ noise tolerance as a function of p. In particular, our theory predicts that the tolerance saturates at $T^* \sim \mathcal{O}\left(N\right)$ as $p \to \infty$, but at the cost of using an intractable number of examples. This behavior is different from the $\mathcal{O}\left(N^{1/2-p/4}\right)$ tolerance of the direct p-body model in the noisy-learning regime studied in [124]. In other terms, the dataset noise that we are facing is of a different nature than the learning noise of [124]. In any case, it is interesting that both the direct and inverse models are able to tolerate an extensive amount of noise. Overall, our results suggest that it could be advantageous to use a student network with a relatively large p to learn from a large but noisy dataset when the p^* of the teacher generative model is unknown.

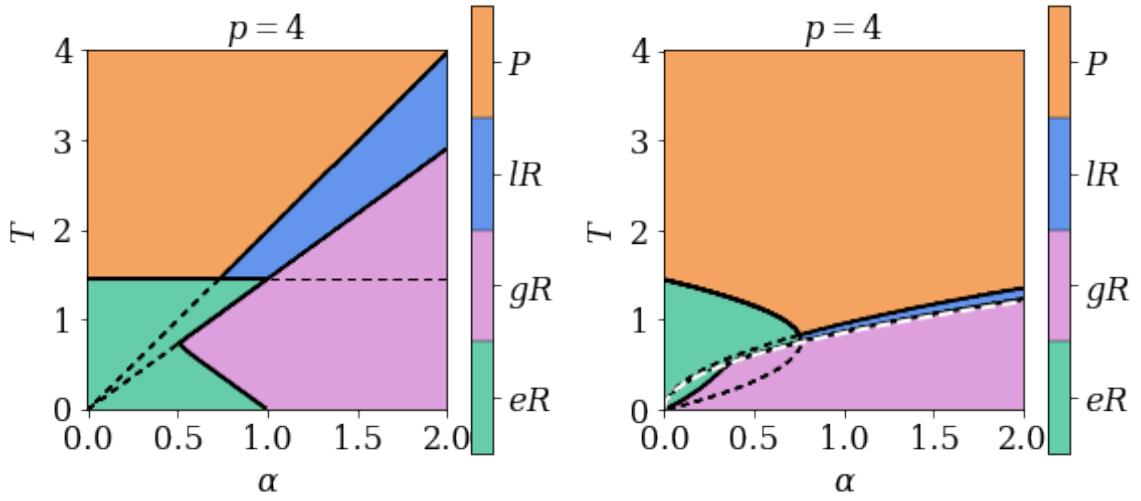

Figure 2.6: RS phase diagrams of inverse models with $p^* = 2$ and $p = 4$. The left plot is for $\alpha = \frac{M(p/2+1)!}{N^{p/2}}$, and $\beta^* = 1 - \frac{1}{\sqrt{2}}$ such that $\eta = 1$ and the right plot is for $\alpha = \frac{Mp!}{N^{p-1}}$ and $\beta^* = \sqrt{\frac{2\lambda}{N}}$ with $\lambda = \beta$. Accurate pattern retrieval is not possible in the paramagnetic phase (P) or in the example retrieval phase (eR), but it is possible in the local retrieval phase (lR) and in the global retrieval phase (gR). The ferromagnetic fixed point corresponding to accurate pattern retrieval is globally stable in the gR phase, but locally stable in the lR phase. The black dashed lines mark the metastable continuation of the eR, lR and gR phase boundaries through neighboring phases with a larger free entropy. The paramagnetic total entropy becomes negative below the white dashed line drawn on the right plot. However, the paramagnetic phase is no longer globally stable at that temperature.

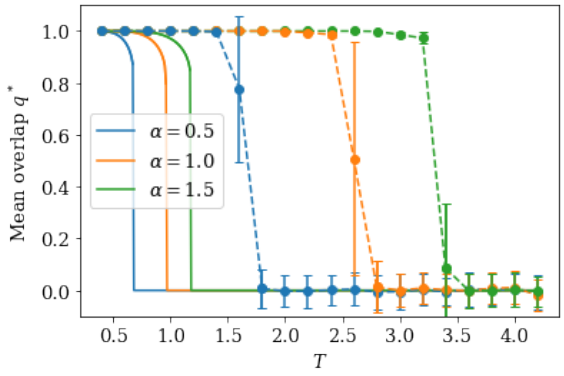

Figure 2.7: Monte Carlo simulations (dashed lines) and RS saddle-point solutions (full lines) of the inverse model in the large-noise scaling with $p^* = 2$ and $p = 4$. The teacher generates $M = \frac{\alpha N^{p-1}}{p!}$ examples σ^a with $N = 256$ components each, and the simulation results are then averaged over $L = 100$ student patterns. The student patterns are all initialized to ξ^*.

An unavoidable drawback of large teacher noise is that it always lead to uncorrelated examples, which makes accurate example retrieval impossible. Instead, it is replaced by the inaccurate example retrieval phase where the student has finite overlap m with a noisy example generated by the teacher but no overlap with the signal (see Fig. 2.6). Depending on T and α, this phase can be either globally stable or locally stable.

For the sake of clarity, we plot only the globally stable phase on our phase diagram in Fig. (2.6). The locally stable phase is arguably less important to plot because it is identical to the locally stable ferromagnetic phase previously reported in the direct model when assuming replica symmetry (see [41] and Fig. 2.1).

Given $m = 0$, the free entropy of the inverse model with $p \geq 3$, $p^* = 2$ and $\beta = \lambda$ is the same as on the Nishimori line (see Eq. 2.5 and Appendix 2.D). As a direct consequence, the total entropy is positive outside of the eR phase (see Fig. 2.6). Additionally, the $p^* = 2, p \geq 3$ phase diagrams with $\beta \neq \lambda$ are identical to the $p = p^*$ phase diagrams with $\beta \neq \beta^*$, which suggests that $\beta = \lambda$ is optimal for $p^* = 2, p \geq 3$ in the same sense as $\beta = \beta^*$ is optimal for $p = p^*$ (see Fig. 2.5). Monte Carlo simulations confirm that a student with $p \geq 3$ is able to retrieve the pattern of a teacher with $p = 2$ and $T^* \sim \mathcal{O}\left(N^{1/2}\right)$ (see Fig. 2.7). However, the lR phase transition is at a higher T in the simulations than on the $\beta = \lambda$ RS phase diagram (see Fig. 2.5), which means that RSB is necessary to describe it accurately. One could check where replica symmetry holds by evaluating the stability of the RS saddle point throughout the phase diagram.

2.4.5.2 Finite noise scaling

We also consider a different scaling regime where $\beta^* \sim \mathcal{O}(1)$ and $M \sim \mathcal{O}\left(N^{p/2}\right)$, such that

$$\alpha = \frac{M(p/2+1)!}{N^{p/2}},$$

is finite. In this finite-noise scaling, $p \geq 3$ requires $\mathcal{O}\left(N^{p/2-1}\right)$ more training examples than $p = 2$, which is a lot less than the first scaling. For instance, a student with $p = 4$ needs $\mathcal{O}\left(N^2\right)$ examples to retrieve ξ^*. As before, the phase transitions are all first order, the overlap q^* stays high throughout the gR and lR phase of $p = 4$ and gR is a hard phase. The saddle-point equations (see Eqs. 2.6) are free from the pattern interference term $\sqrt{\alpha r}x$ present in their $p^* = p$ counterparts (see Eqs. 2.5) until β^* becomes so small that is approaches $\mathcal{O}\left(N^{2/p-1}\right)$. Therefore, contrary to $p^* = p = 2$, the network is never in the SG phase. Practically, it means that $p \geq 3$ gives more freedom than $p = 2$ for tuning β and α. The only remaining restriction is that choosing α and T too small puts the network into the inaccurate eR phase resulting from the kz term (see Fig. 2.6). The saddle point equations can be derived without the RS ansatz because they do not involve q and r. Consequently, we expect them to yield an exact solution. Like on the Nishimori line, the total entropy of the paramagnetic phase is always positive, which is consistent with the solution being exact.

2.4.6 Robustness against adversarial attacks

Inverse models with $p^* = 2$ and $p \geq 3$ offer an opportunity to study adversarial attacks in a simple setting because their phase diagrams have regions where the signal retrieval phases (gR and lR) overlap with the inaccurate eR phase. Recall that, in the lR phase, a noisy student pattern ξ either converges to ξ^* or falls in the paramagnetic phase, depending on the amount of noise that ξ contains initially. The quantity of noise needed to prevent pattern retrieval becomes smaller as one approaches the lR to P phase transition and the basin of attraction of lR shrinks. Similarly, in the region of inaccurate eR where signal retrieval is metastable, patterns ξ that are corrupted by replacing some of their entries ξ_i by the components σ_i^a of an example σ^a may converge to σ^a when enough entries are replaced. The fraction ε of entries that need to be replaced becomes smaller as the basin of attraction of inaccurate eR expands and overtakes that of signal retrieval. In practice,

an adversary can use this strategy to trick the student into converging to a pattern other than ξ^*. This scenario is similar to an adversarial attack targeting the input of K & H's dense Hopfield network model because the student pattern ξ plays a similar role in the inverse model as the test data in K & H's dense Hopfield networks (see Fig. 2.3, Section 2.4.3 and Appendix 2.A). In that analogy, the examples σ are acting like the neural network weights rather than taking the role of the training data.

We will now investigate what values of the perturbation size ε are a threat by deriving a formula for the largest ε such that the student converges to the signal at zero temperature. This largest ε will be denoted ε^*, and we expect it to be a good measure of adversarial robustness. The saddle-point equations with $T = 0$ indicate that the student converges to one of the signal retrieval phases if and only if $k < \eta\alpha r^*$ (see Eqs. 2.6). Sampling the initial conditions of ξ_i from $(1-\varepsilon)\,\delta(\xi_i - \xi_i^*) + \varepsilon\,\delta(\xi_i - \sigma_i^a)$ with $\varepsilon \in [0,1]$, we get

$$r^* = p\left[\frac{1}{N}\sum_{i=1}^{(1-\varepsilon)N} \xi_i^*\xi_i^* + \frac{1}{N}\sum_{i=1}^{\varepsilon N} \xi_i^*\sigma_i^a\right]^{p-1},$$
$$k = p\left[\frac{1}{N}\sum_{i=1}^{(1-\varepsilon)N} \xi_i^*\sigma_i^a + \frac{1}{N}\sum_{i=1}^{\varepsilon N} \sigma_i^a\sigma_i^a\right]^{p-1}.$$

By the law of large numbers, $\frac{1}{\varepsilon N}\sum_{i=1}^{\varepsilon N}\xi_i^*\sigma_i^a$ and $\frac{1}{(1-\varepsilon)N}\sum_{i=1}^{(1-\varepsilon)N}\xi_i^*\sigma_i^a$ are both typically close to $m^* = \frac{1}{N}\sum_i^N \xi_i^*\sigma_i^a \approx 0$ as $N \to \infty$. If we take σ^a to be a typical example, then r^* and k reduce to

$$r^* \approx p\,(1-\varepsilon)^{p-1},$$
$$k \approx p\varepsilon^{p-1}.$$

Substituting these expressions back in $k < \eta\alpha r^*$ yields

$$\varepsilon^{p-1} < \eta\alpha\,(1-\varepsilon)^{p-1},$$
$$\varepsilon < \frac{[\eta\alpha]^{\frac{1}{p-1}}}{[\eta\alpha]^{\frac{1}{p-1}}+1}.$$

In other terms, the inverse model with $p^* = 2$ and even $p \geq 3$ is resistant to adversarial attacks of size $\varepsilon^* = \frac{[\eta\alpha]^{\frac{1}{p-1}}}{[\eta\alpha]^{\frac{1}{p-1}}+1}$ and smaller. For $p = 4$, ε^* is in good agreement with Monte Carlo simulations of the inverse model corrupted by a typical example (see Fig. 2.8). This comparison is good evidence that our solution of the finite-noise scaling is indeed exact. Additionally, ε^* is a decent approximation of empirical robustness even when the inverse model is corrupted by the example that has the largest overlap with ξ^*. A similar construction with the perturbation sampled uniformly at random gives $k \sim \mathcal{O}\left(N^{1/2-p/2}\right) \approx 0$, so adversarial attacks are much more efficient at fooling the model than random noise. Just like adversarial attacks targeting more complicated neural networks [11, 12], our example-based attack can be hard to detect at low ε because a few adversarially perturbed entries ξ_i do not look very different from a low amount of meaningless noise. Moreover, ε^* grows monotonically with α, which is consistent with the common observation that larger neural networks are also more adversarially robust [110, 125, 126, 127, 128, 129, 130]. At first glance, this effect can be counter-intuitive because adversarial vulnerability looks like a form of overfitting [15]. In our model,

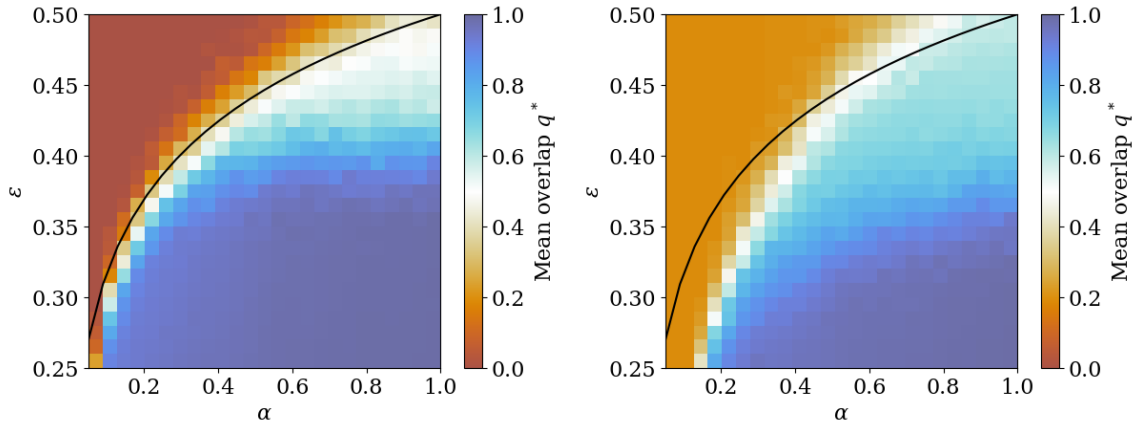

Figure 2.8: Monte Carlo simulations of the overlap q^* as a function of α and adversarial attack size ε in the inverse model with $p^* = 2$, $\beta^* = 1 - \frac{1}{\sqrt{2}}$, $p = 4$, $\beta = \infty$ and $N = 1024$. The simulation results are averaged over $L = 100$ student patterns. On the left plot, the inverse model is corrupted by an example σ^a that has a small overlap with ξ^* in absolute value. On the right plot, it is corrupted by the example that has the largest overlap with ξ^*. The black line $\varepsilon^* = \frac{\alpha^{1/3}}{\alpha^{1/3}+1}$ is our analytical formula for the largest adversarial perturbation ε such that the student retrieves ξ^* rather than the example σ^a.

however, all examples work together to stabilize the lR phase, and the best way to push the student into the eR phase is to perturb it with a single example. Therefore, it is not surprising that increasing α makes the student more robust. We recall that the examples σ are a feature-based representation of ξ^*. Interestingly, it means that the underlying mechanism of our example-based attack is conceptually similar to gradient-based attacks targeting many common types of neural networks [15]. In fact, gradient-based attacks find features stored in neural network weights and add them to the data in order to fool the network [15, 131, 132, 133]. It would be interesting to investigate, both empirically and theoretically, if only a small number of weights are involved in constructing these adversarial attacks. If it is the case, it could explain why larger neural networks are often more robust. In general, we expect this kind of one-example attack to be possible in any region of signal retrieval that overlaps with the inaccurate eR phase. Using $p \neq p^*$ may not be a necessary ingredient of adversarial vulnerability in more general models with other sources of mismatch, but in our case it ensures that the signal retrieval phases intersect the inaccurate eR phase. Conversely, the accurate eR phase is by definition robust to adversarial attacks since retrieving an example σ^a is the same as recovering ξ^*. This distinction clarifies why the dense Hopfield networks designed by K & H are adversarially robust in the prototype phase despite being adversarially vulnerable in the feature phase. In fact, K & H observed that adversarial attacks are unsuccessful in the prototype phase specifically because they retrieve stored examples that are semantically meaningful [36]. In summary, our model yields two main results concerning adversarial examples. First of all, it suggests a reason why large feature-based neural networks are more adversarially robust than smaller ones. Second of all, it clarifies why dense Hopfield networks are much more robust in the prototype phase than in the feature phase.

2.5 Conclusion

In this work, we derive the exact phase diagram of the p-dense networks in the teacher-student setting [39, 101, 28, 43]. On the Nishimori line, we find an example retrieval phase (eR) and a global retrieval phase (gR) reminiscent of the prototype and feature regimes observed empirically in dense Hopfield networks [30]. We show that the phase transition towards gR of the inverse model overlaps the paramagnetic to spin-glass (P-SG) transition of the direct model, which allows us to locate the P-SG transition much more precisely than before [28, 41]. On the other hand, we discover that inverse models outside of the Nishimori line are able to resist an extensive amount of noise. In fact, a student with $p \geq 3$ is able to learn from a teacher with $p^* = 2$ even when the teacher's inverse temperature β^* is as low as $\mathcal{O}\left(N^{2/p-1}\right)$. Moreover, such a student is immune to pattern interference until β^* reaches $\mathcal{O}\left(N^{2/p-1}\right)$. In this setting, we derive a formula measuring the adversarial robustness of the student with $p \geq 3$ and $T = 0$. We then use this formula to describe how making a neural network larger can potentially increase its robustness to adversarial attacks constructed with only a few learned weights [110, 125, 126, 127, 128, 129, 130]. Our model also clarifies why the prototype phase of dense Hopfield networks is adversarially robust [36]. We compare our key results against Monte Carlo simulations.

Dense networks with exponential interactions have been argued to be the $p \to \infty$ limit of the p-body models [134]. It would be interesting to see if they can achieve $\mathcal{O}(N)$ noise tolerance at the cost of an exponential number of training examples. More generally, studying exponential models in the teacher-student setting would be an interesting extension of this work and could be used to complement existing studies of the direct model [134, 135]. A caveat of our model is that the teacher has only one pattern. In fact, we would need to use a teacher with at least two patterns to describe more completely the kind of adversarial attack aiming to misclassify data. It should be possible to study this kind of teacher by using an approach similar to [42]. In particular, [39] and [42] argue that the performance of restricted Boltzmann machines with a finite number P of i.i.d. planted patterns is independent of P in the teacher-student setting. It would be interesting to investigate whether this characteristic also holds for p-body dense networks. On the practical side, we highlight the untapped benefits of using p-body models to either resist an extensive amount of noise in the feature phase or improve adversarial robustness in the prototype phase. Overall, we stress that further investigations of dense Hopfield networks could unlock their true potential.

Data availability

The figures can be reproduced using the code available on this public Github repository.

2.A Gardner's Hamiltonian vs K & H's Hamiltonian

Consider the generalized Hopfield Hamiltonian $H\left[\sigma \mid \xi\right] = -\sum_{i_1<\ldots<i_p=1}^{N} J_{i_1\ldots i_p}\sigma_{i_1}\ldots\sigma_{i_p}$ with p-body interactions $J_{i_1\ldots i_p} = \frac{p!}{N^{p-1}}\sum_{\mu=1}^{M}\xi_{i_1}^{\mu}\ldots\xi_{i_p}^{\mu}$ described by Gardner [28], where M indicates the number of patterns ξ^{μ} used to construct J, and N denotes the number of components of each pattern ξ^{μ} and example σ. In this Section, we will omit ξ in the argument of $H\left[\sigma \mid \xi\right]$ and write $H\left[\sigma\right]$ instead for notational simplicity. Unless indicated otherwise, we will assume a large number number of components $N \gg 1$

and patterns $M \sim \mathcal{O}\left(N^{p-1}\right)$. We will start by comparing it to the dense Hopfield network Hamiltonian $\mathcal{H}\left[\sigma\right] = -\frac{1}{N^{p-1}} \sum_{\mu} \left(\sum_{i} \xi_i^{\mu} \sigma_i\right)^p$ studied by K & H [30].

For that purpose, we rewrite H in the form $H\left[\sigma\right] = -\frac{1}{p!} \sum_{i_1 \neq ... \neq i_p} J_{i_1 ... i_p} \sigma_{i_1} ... \sigma_{i_p}$ by summing over all permutations of $\{i_1 ... i_p\}$ in place of the restricted set $i_1 < ... < i_p$ and compensating for double counting with the prefactor $\frac{1}{p!}$. This manipulation leads to

$$\begin{aligned} H\left[\sigma\right] &= -\frac{1}{p!} \sum_{i_1 \neq ... \neq i_p} J_{i_1 ... i_p} \sigma_{i_1} ... \sigma_{i_p} \\ &= -\frac{1}{N^{p-1}} \sum_{\mu} \sum_{i_1 \neq ... \neq i_p} \xi_{i_1}^{\mu} ... \xi_{i_p}^{\mu} \sigma_{i_1} ... \sigma_{i_p} \, . \end{aligned}$$

On the other hand, K & H's Hamiltonian may be rewritten

$$\begin{aligned} \mathcal{H}\left[\sigma\right] &= -\frac{1}{N^{p-1}} \sum_{\mu} \left(\sum_{i} \xi_i^{\mu} \sigma_i\right)^p \\ &= -\frac{1}{N^{p-1}} \sum_{\mu} \left(\sum_{i_1} \xi_{i_1}^{\mu} \sigma_{i_1}\right) ... \left(\sum_{i_p} \xi_{i_p}^{\mu} \sigma_{i_p}\right) \\ &= -\frac{1}{N^{p-1}} \sum_{\mu} \sum_{i_1 ... i_p} \xi_{i_1}^{\mu} ... \xi_{i_p}^{\mu} \sigma_{i_1} ... \sigma_{i_p} \, , \end{aligned}$$

where the sum over $i_1 ... i_p$ includes both the set of indices $i_1 \neq ... \neq i_p$ found in $H\left[\sigma\right]$ and other configurations where some indices are equal. For example, the configuration $i_1 \neq ... \neq i_{p-1} = i_p$ contains the fewest equal indices after $i_1 \neq ... \neq i_p$. In other words, $\mathcal{H}\left[\sigma\right]$ can be expressed as an expansion around $H\left[\sigma\right]$, and the two Hamiltonians are equivalent when the normalized residuals $\frac{\mathcal{H}[\sigma] - H[\sigma]}{N}$ vanish in the limit of large N. In this study, we encounter two cases which bring different results.

1 The Hamiltonians $\mathcal{H}\left[\sigma\right]$ and $H\left[\sigma\right]$ are dominated by a few closely packed configurations ξ^{μ} that have finite overlap $\frac{1}{N} \sum_i \xi_i^{\mu} \sigma_i \sim \mathcal{O}\left(1\right)$ with σ. We say that they are aligned with σ.

2 The Hamiltonians $\mathcal{H}\left[\sigma\right]$ and $H\left[\sigma\right]$ are dominated by many spread out configurations ξ^{μ} that have microscopic overlap $\frac{1}{N} \sum_i \xi_i^{\mu} \sigma_i \sim \mathcal{O}\left(N^{-1/2}\right)$ with σ. We say that they are misaligned with σ

We use the expansion of $\mathcal{H}\left[\sigma\right]$ to discuss both the aligned case and the misaligned case. We start by writing the $i_1 \neq ... \neq i_p$ and $i_1 \neq ... \neq i_{p-1} = i_p$ terms explicitly, which leads to the form

$$\begin{aligned} \mathcal{H}\left[\sigma\right] = &-\frac{1}{N^{p-1}} \sum_{\mu} \sum_{i_1 \neq ... \neq i_p} \xi_{i_1}^{\mu} ... \xi_{i_p}^{\mu} \sigma_{i_1} ... \sigma_{i_p} \\ &- \frac{1}{2} \frac{p(p-1)}{N^{p-1}} \sum_{\mu} \sum_{i_1 \neq ... \neq i_{p-1}} \xi_{i_1}^{\mu} ... \left(\xi_{i_{p-1}}^{\mu}\right)^2 \sigma_{i_1} ... \left(\sigma_{i_{p-1}}\right)^2 + ... \, , \end{aligned}$$

because there are $\binom{p}{2} = \frac{p(p-1)}{2}$ ways for the indices i_{p-1} and i_p to be equal. This expression can be

summarized by $\mathcal{H}[\sigma] = H[\sigma] + H'[\sigma] + ...$, where $H'[\sigma]$ simplifies to

$$\begin{aligned} H'[\sigma] &= -\frac{1}{2}\frac{p(p-1)}{N^{p-1}}\sum_{\mu}\sum_{i_1\neq...\neq i_{p-1}}\xi_{i_1}^{\mu}...\left(\xi_{i_{p-1}}^{\mu}\right)^2\sigma_{i_1}...\left(\sigma_{i_{p-1}}\right)^2 \\ &= -\frac{1}{2}\frac{p(p-1)}{N^{p-2}}\sum_{\mu}\sum_{i_1\neq...\neq i_{p-2}}\xi_{i_1}^{\mu}...\xi_{i_{p-2}}^{\mu}\sigma_{i_1}...\sigma_{i_{p-2}} \\ &= -\frac{1}{2}\frac{p!}{N^{p-2}}\sum_{\mu}\sum_{i_1<...<i_{p-2}}\xi_{i_1}^{\mu}...\xi_{i_{p-2}}^{\mu}\sigma_{i_1}...\sigma_{i_{p-2}}\,. \end{aligned}$$

In the aligned case, $H'[\sigma]$ is $\mathcal{O}(1)$ in N because the sum over $i_1 < ... < i_{p-2}$ is $\mathcal{O}\left(N^{p-2}\right)$. The terms implied by the ellipsis are even smaller because their sums are resctricted by more equality constraints. Therefore, the residuals $\frac{\mathcal{H}[\sigma]-H[\sigma]}{N}$ vanish in the limit of large N, and the two Hamiltonians are equivalent. Conversely, we find that $\mathcal{H}[\sigma]$ and $H[\sigma]$ differ from each other in the misaligned case (see Appendix 2.B for more details). Therefore, although the phases of $H[\sigma]$ that we obtain in this study are qualitatively similar to the ones observed by K & H [30, 36], the phase diagram of $H[\sigma]$ must be compared against a simulation of $H[\sigma]$ rather than $\mathcal{H}[\sigma]$ in order to test our theory quantitatively.

To understand how to sample σ in both models, consider a Monte Carlo simulation used to find the statistical equilibrium of a spin ensemble σ with Hamiltonian $G[\sigma]$. To be more specific, suppose σ is updated to a new state σ' with a randomly selected spin σ_i flipped with acceptance probability $P_i = \frac{1}{1+\exp[\beta(G[\sigma']-G[\sigma])]}$ for a large number of time-steps. This approach works well for $G[\sigma] = \mathcal{H}[\sigma]$. However, in the case of $H[\sigma]$, we find that the simulation only converges when we use the local field $h_i = \frac{p!}{N^{p-1}}\sum_{\mu}\xi_i^{\mu}\sum_{i_1<...<i_{p-1}}\xi_{i_1}^{\mu}...\xi_{i_{p-1}}^{\mu}\sigma_{i_1}...\sigma_{i_{p-1}}$ mentioned by Gardner [28] to approximate $\frac{H\left[\sigma'\right]-H[\sigma]}{2\sigma_i}$ at large N. In other words, we iteratively flip randomly chosen spins σ_i with acceptance probability $P_i = \frac{1}{1+\exp(2\beta h_i\sigma_i)}$ for a large number of time steps. For arbitrary p, it is not obvious how to compute h_i quickly as a sub-routine of the Monte Carlo simulation. However, we find that both $p = 3$ and $p = 4$ have closed-formed expressions that are easy to evaluate numerically in an efficient way. To be more precise,

- $p = 3$ leads to $h_i = 3\sum_{\mu}\xi_i^{\mu}\left[\left(\frac{1}{N}\sum_j\xi_j^{\mu}\sigma_j\right)^2 - \frac{1}{N}\right]$,
- and $p = 4$ leads to $h_i = 4\sum_{\mu}\xi_i^{\mu}\left(\frac{1}{N}\sum_j\xi_j^{\mu}\sigma_j\right)\left[\left(\frac{1}{N}\sum_j\xi_j^{\mu}\sigma_j\right)^2 - \frac{3}{N}\right]$.

For this reason and also because the number $M \sim \mathcal{O}\left(N^{p-1}\right)$ of patterns ξ^{μ} used in a Monte Carlo simulations increases exponentially with p, we choose to simulate only $p = 3$ and $p = 4$.

The output of the neural network model that K & H designed for classification of data is

$$c_j = \tanh\left[\frac{1}{2}\beta\left(\mathcal{H}[\sigma'] - \mathcal{H}[\sigma]\right)\right] \approx \tanh\left[\beta p\sum_{\mu}\xi_j^{\mu}\left(\frac{1}{N}\sum_i\xi_i^{\mu}\sigma_i\right)^{p-1}\right]$$

We omit the linear rectifier present in the original paper [30] because the overlaps $\frac{1}{N}\sum_i\xi_i^{\mu}\sigma_i$ are almost always positive (see for example the Supplement of [34]). The predicted class is then $j' = \mathrm{argmax}_j\{c_i\}$. Using $1 - P_j = \frac{1}{1+\exp[\beta(\mathcal{H}[\sigma']-\mathcal{H}[\sigma])]}$ instead of c_j does not change j' because $1 - P_j$ and c_j are related by

$1 - P_j = \frac{1}{2}[c_j + 1]$. When we evaluate P_i using H instead of $\mathcal{H}$, this relation does not always hold exactly. Rather, it should be considered an approximation.

2.B Direct model cumulant expansions

In the direct model, the average replicated partition function $\langle Z^L \rangle$ takes the form:

$$\langle Z^L \rangle = \left\langle \sum_{\sigma} \exp\left(-\beta \sum_{\gamma=1}^{L} H\left[\sigma^\gamma \mid \xi\right] \right) \right\rangle ,$$

with $\sigma = \left\{ \sigma^1 \ \ldots \ \sigma^L \right\}$. Gardner simplifies it to

$$\langle Z^L \rangle \approx \left\langle \sum_{\sigma} \exp\left(\beta N \sum_{\gamma} \sum_{\mu \in \Gamma_\gamma} \left[\frac{1}{N} \sum_i \xi_i^\mu \sigma_i^\gamma \right]^p + \beta \sum_{\gamma} \sum_{\mu \in \bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \right\rangle , \tag{2.10}$$

where the sets Γ_γ contain the patterns ξ^μ that have macroscopic overlap with σ_γ, and their complement $\bar{\Gamma} = \cap_\gamma \bar{\Gamma}_\gamma$ consists of the remaining patterns. Two approximations are used to obtain this expression:

- $\sum_{\mu \in \Gamma_\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \approx N \sum_{\mu \in \Gamma_\gamma} \left[\frac{1}{N} \sum_i \xi_i^\mu \sigma_i^\gamma \right]^p$ because this part of $H\left[\sigma^\gamma \mid \xi\right]$ is aligned with σ (see Case **1** of Appendix 2.A).
- $\sum_{\mu \in \bar{\Gamma}_\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \approx \sum_{\mu \in \bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma$ since $\bar{\Gamma}$ contains almost all of the elements in each $\bar{\Gamma}_\gamma$ when N is large.

Gardner evaluates the contribution of the $\mu \in \bar{\Gamma}$ terms via a cumulant expansion, resulting in:

$$\begin{aligned}
&\log \left\langle \exp\left(\beta \sum_{\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \right\rangle \\
&\quad \approx \beta \left\langle \sum_{\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right\rangle + \frac{1}{2} \beta^2 \left\langle \left[\sum_{\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right]^2 \right\rangle \\
&\quad \approx \frac{1}{2} \beta^2 \left\langle \left[\sum_{\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right] \left[\sum_{\delta} \frac{p!}{N^{p-1}} \sum_{j_1 < \ldots < j_p} \xi_{j_1}^\mu \ldots \xi_{j_p}^\mu \sigma_{j_1}^\delta \ldots \sigma_{j_p}^\delta \right] \right\rangle ,
\end{aligned}$$

because the product of independent spins $\xi_{i_1}^\mu \ldots \xi_{i_p}^\mu$ averages to 0. The sums are then regrouped to get

$$\begin{aligned}
&\log \left\langle \exp\left(\beta \sum_{\gamma} \frac{p!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \right\rangle \\
&\quad = \frac{1}{2} \beta^2 \left[\frac{p!}{N^{p-1}} \right]^2 \left\langle \sum_{\gamma} \sum_{\delta} \sum_{i_1 < \ldots < i_p} \sum_{j_1 < \ldots < j_p} \xi_{i_1}^\mu \xi_{j_1}^\mu \ldots \xi_{i_p}^\mu \xi_{j_p}^\mu \ \sigma_{i_1}^\gamma \sigma_{j_1}^\delta \ldots \sigma_{i_p}^\gamma \sigma_{j_p}^\delta \right\rangle .
\end{aligned}$$

Consider $\xi_i^\mu \xi_j^\mu$ for an arbitrary pair of indices i and j. There are two cases.

- If $i = j$, then $\xi_i^\mu \xi_j^\mu$ is deterministic and equal to 1.
- If $i \neq j$, then $\xi_i^\mu \xi_j^\mu$ can be either $+1$ and -1 with equal probabilities.

On the one hand, if $i_n = j_n$ for all n, then $\left\langle \xi_{i_1}^\mu \xi_{j_1}^\mu ... \xi_{i_p}^\mu \xi_{j_p}^\mu \right\rangle = 1$. On the other hand, if $i_n \neq j_n$ for some n, then $\left\langle \xi_{i_1}^\mu \xi_{j_1}^\mu ... \xi_{i_p}^\mu \xi_{j_p}^\mu \right\rangle = 0$ because $\xi_{i_1}^\mu \xi_{j_1}^\mu ... \xi_{i_p}^\mu \xi_{j_p}^\mu$ is still a product of independent random spins once the deterministic variables are removed. These two cases can be summarized by $\left\langle \xi_{i_1}^\mu \xi_{j_1}^\mu ... \xi_{i_p}^\mu \xi_{j_p}^\mu \right\rangle = \delta_{i_1 j_1} ... \delta_{i_p j_p}$, which then gives

$$
\begin{aligned}
&\log \left\langle \exp \left(\beta \sum_\gamma \frac{p!}{N^{p-1}} \sum_{i_1 < ... < i_p} \xi_{i_1}^\mu ... \xi_{i_p}^\mu \sigma_{i_1}^\gamma ... \sigma_{i_p}^\gamma \right) \right\rangle \\
&= \frac{1}{2} \beta^2 \left[\frac{p!}{N^{p-1}} \right]^2 \sum_\gamma \sum_\delta \sum_{i_1 < ... < i_p} \sum_{j_1 < ... < j_p} \delta_{i_1 j_1} ... \delta_{i_p j_p} \, \sigma_{i_1}^\gamma \sigma_{j_1}^\delta ... \sigma_{i_p}^\gamma \sigma_{j_p}^\delta \\
&= \frac{1}{2} \beta^2 \left[\frac{p!}{N^{p-1}} \right]^2 \sum_\gamma \sum_\delta \sum_{i_1 < ... < i_p} \sigma_{i_1}^\gamma \sigma_{i_1}^\delta ... \sigma_{i_p}^\gamma \sigma_{i_p}^\delta \\
&\approx \frac{1}{2} \beta^2 \frac{p!}{N^{p-1}} \frac{1}{N^{p-1}} \sum_{\gamma\delta} \left[\sum_i \sigma_i^\gamma \sigma_i^\delta \right]^p \\
&= \beta^2 \frac{p!}{N^{p-1}} N \sum_{\gamma < \delta} \left[\frac{1}{N} \sum_i \sigma_i^\gamma \sigma_i^\delta \right]^p + \frac{1}{2} \beta^2 \frac{p!}{N^{p-1}} LN \, .
\end{aligned}
$$

The order $n > 2$ terms are subdominant in N and can be neglected when $p \geq 3$ [28]. The RS free entropy is then obtained through a standard approach to the replica method. Note that Gardner's Hamiltonian is misaligned with σ when the free entropy is dominated by this cumulant expansion (see Case **2** of Appendix 2.A). In the case of K & H's Hamiltonian, we must also take into account the correction $H'[\sigma] = \frac{1}{2} \frac{p!}{N^{p-2}} \sum_\gamma \sum_{i_1 < ... < i_{p-2}} \xi_{i_1}^\mu ... \xi_{i_{p-2}}^\mu \sigma_{i_1}^\gamma ... \sigma_{i_{p-2}}^\gamma$ introduced in appendix 2.A by imposing $i_{p-1} = i_p$. In fact, a cumulant expansion of this expression gives

$$
\begin{aligned}
&\log \left\langle \exp \left(\beta p \sum_\gamma \frac{1}{2} \frac{p!}{N^{p-2}} \sum_{i_1 < ... < i_{p-2}} \xi_{i_1}^\mu ... \xi_{i_{p-2}}^\mu \sigma_{i_1}^\gamma ... \sigma_{i_{p-2}}^\gamma \right) \right\rangle \\
&\approx \frac{1}{4} \beta^2 \frac{p!}{N^{p-2}} \frac{p(p-1)}{N^{p-2}} \sum_{\gamma < \delta} \left[\sum_i \sigma_i^\gamma \sigma_i^\delta \right]^{p-2} + \frac{1}{8} \beta^2 \frac{p!}{N^{p-2}} L \\
&= \frac{1}{4} p(p-1) \beta^2 \frac{p!}{N^{p-1}} N \sum_{\gamma < \delta} \left[\frac{1}{N} \sum_i \sigma_i^\gamma \sigma_i^\delta \right]^{p-2} + \frac{1}{8} \beta^2 \frac{p!}{N^{p-1}} LN \, ,
\end{aligned}
$$

which contributes to the free energy on the same order in N as Gardner's Hamiltonian. Therefore, K & H's Hamiltonian is not equivalent to Gardner's Hamiltonian when the latter is misaligned with σ (see Case **2**). The index configurations with more equality constraints also contribute to the free entropy on the same order in N because the factors of N that are lost to equality constraints are restored when the sums get squared in the

cumulant expansion.

$p = 2$ is the only positive integer such that Gardner's Hamiltonian and Krotov's Hamiltonian are equivalent [54, 28]. In the misaligned case with a single stored pattern ξ^* (see Case **2**), the free entropy of $p = 2$ simplifies to

$$\begin{aligned}\frac{\log(Z)}{N} &= \frac{1}{N}\log\left\langle \exp\left\{\beta\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^\gamma_{i_1}\sigma^\gamma_{i_2}\right\}\right\rangle + \log 2 \\ &= \frac{1}{N}\log\left\langle \exp(-\beta)\int_{\mathbb{R}} dx\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2 + x\sqrt{\beta\frac{2}{N}}\sum_i \xi^*_i\sigma^\gamma_i\right\}\right\rangle + \log 2 \\ &= \frac{1}{N}\log\left[\int_{\mathbb{R}} dx\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}\cosh^N\left(x\sqrt{\beta\frac{2}{N}}\right)\right] - \beta\frac{1}{N} + \log 2\,,\end{aligned}$$

by using the Hubbard-Stratonovich transformation. At large N, it approximates to:

$$\begin{aligned}\frac{\log(Z)}{N} &\approx \frac{1}{N}\log\left[\int dx\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}\left(1+\beta\frac{1}{N}x^2\right)^N\right] - \beta\frac{1}{N} + \log 2 \\ &\approx \frac{1}{N}\log\left[\int_{\mathbb{R}} dx\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}\exp\left(\beta x^2\right)\right] - \beta\frac{1}{N} + \log 2 \\ &= \left(-\frac{1}{2}\log(1-2\beta) - \beta\right)\frac{1}{N} + \log 2\,,\end{aligned}$$

thanks to the well-known limit $\lim_{N\to\infty}\left(1+\frac{1}{N}z\right)^N = \exp(z)$. This free entropy is consistent with the one found in literature when $\alpha = \frac{1}{N}$ [54].

2.C Teacher-student replicated partition function

Recall that the student samples its pattern from the posterior $P(\xi \mid \sigma) = \frac{P(\xi)\prod_a P(\sigma^a|\xi)}{P(\sigma)}$ (see Section 2.3). Given $P(\xi)$ uniform, it can be rewritten as $P(\xi \mid \sigma) = \frac{\prod_a P(\sigma^a|\xi)}{\sum_\xi \prod_a P(\sigma^a|\xi)}$, where $P(\sigma^a \mid \xi)$ is the distribution of the direct model with a single pattern ξ. To simplify $P(\xi \mid \sigma)$ further, we need to manipulate the partition function $Z = \sum_{\sigma^a}\exp(-\beta H[\sigma^a \mid \xi])$ of $P(\sigma^a \mid \xi)$ (see Appendix 2.A for the definition of $H[\sigma \mid \xi]$). Under the gauge transformation $\sigma^a_i \to \xi_i\sigma^a_i$, we may write

$$Z = \sum_{\sigma^a}\exp\left(\beta\frac{p!}{N^{p-1}}\sum_{i_1<\ldots<i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right)\,,$$

without changing the configurations of σ^a that we are summing over. Therefore, Z does not depend on ξ, and we can factor it out of the sum $\sum_\xi$, which yields

$$\begin{aligned} P\left(\xi \mid \sigma\right) &= \frac{\prod_a \frac{1}{Z} \exp\left(\beta \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}\ldots\xi_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right)}{\sum_\xi \prod_a \frac{1}{Z} \exp\left(\beta \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}\ldots\xi_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right)} \\ &= \frac{\exp\left(\beta \frac{p!}{N^{p-1}} \sum_a \sum_{i_1<\ldots<i_p} \xi_{i_1}\ldots\xi_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right)}{\sum_\xi \exp\left(\beta \frac{p!}{N^{p-1}} \sum_a \sum_{i_1<\ldots<i_p} \xi_{i_1}\ldots\xi_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right)} . \end{aligned}$$

Therefore, we define the partition function of the inverse model to be $\mathcal{Z} = \sum_\xi \exp\left(-\beta H\left[\xi \mid \sigma\right]\right)$ (again, see Appendix 2.A for the definition of $H\left[\xi \mid \sigma\right]$). The L^{th} power of $\mathcal{Z}$ and its average then take the form

$$\begin{aligned} \mathcal{Z}^L &= \sum_\xi \prod_b \exp\left(\beta \frac{p!}{N^{p-1}} \sum_a \sum_{i_1<\ldots<i_p} \xi^b_{i_1}\ldots\xi^b_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right) , \\ \left\langle \mathcal{Z}^L \right\rangle &= \sum_\sigma P\left(\sigma\right) \sum_\xi \exp\left(\beta \frac{p!}{N^{p-1}} \sum_{ab} \sum_{i_1<\ldots<i_p} \xi^b_{i_1}\ldots\xi^b_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right) , \end{aligned}$$

where $b \in \left\{1 \ \ldots \ L\right\}$ label replicas in the set of patterns $\xi = \left\{\xi^1 \ \ldots \ \xi^L\right\}$ inferred by the student. Using the definition of conditional probability, we rewrite $P\left(\sigma\right)$ as

$$\begin{aligned} P\left(\sigma\right) &= \sum_{\xi^*} P\left(\sigma \mid \xi^*\right) P\left(\xi^*\right) \\ &= \frac{1}{2^N} \sum_{\xi^*} P\left(\sigma \mid \xi^*\right) \\ &= \frac{1}{2^N} \sum_{\xi^*} \prod_a P\left(\sigma^a \mid \xi^*\right) , \end{aligned}$$

where $P\left(\sigma \mid \xi^*\right)$ has the same functional form as $P\left(\sigma \mid \xi^b\right)$, but has hyperparameters p^* and β^* in place of p and β. As we did for Z, we factor the partition function Z^* of $P\left(\sigma^a \mid \xi^*\right)$ out of the sum, which yields

$$\begin{aligned} P\left(\sigma\right) &= \frac{1}{2^N} \frac{1}{\left[Z^*\right]^M} \sum_{\xi^*} \prod_a \exp\left(\beta^* \frac{p^*!}{N^{p^*-1}} \sum_{i_1<\ldots<i_{p^*}} \xi^*_{i_1}\ldots\xi^*_{i_{p^*}}\sigma^a_{i_1}\ldots\sigma^a_{i_{p^*}}\right) \\ &= \frac{1}{2^N} \frac{\mathcal{Z}^*}{\left[Z^*\right]^M} = \frac{1}{2^{MN}} \frac{\mathcal{Z}^*}{\left[2^{N/M-N} Z^*\right]^M} , \end{aligned}$$

where $\mathcal{Z}^* = \sum_{\xi^*} \exp\left(-\beta^* H\left[\xi^* \mid \sigma\right]\right)$ is the partition function of the inverse model with interaction order p^*. Using $\sum_\sigma P\left(\sigma\right) = 1$, we immediately deduce that $\left[2^{N/M-N} Z^*\right]^M = \left\langle \mathcal{Z}^* \right\rangle$. Plugging $P\left(\sigma\right) = \frac{1}{2^{MN}} \frac{\mathcal{Z}^*}{\left\langle \mathcal{Z}^* \right\rangle}$

back in $\left\langle \mathcal{Z}^L \right\rangle$ then gives

$$\left\langle \mathcal{Z}^L \right\rangle = \frac{1}{2^{MN}} \frac{1}{\langle \mathcal{Z}^* \rangle} \sum_{\xi^*} \sum_{\sigma} \exp\left(\beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}} \right)$$
$$\sum_{\xi} \exp\left(\beta \frac{p!}{N^{p-1}} \sum_{ab} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \right) .$$

We simplify this expression to:

$$\begin{aligned}
\left\langle \mathcal{Z}^L \right\rangle = \frac{1}{2^{MN}} \frac{1}{\langle \mathcal{Z}^* \rangle} \sum_{\xi^*} \sum_{\sigma} \exp\Bigg(& \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \Gamma_*} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}} \\
& + \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \bar{\Gamma}_*} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}} \Bigg) \\
& \sum_{\xi} \exp\Bigg(\beta \frac{p!}{N^{p-1}} \sum_{b} \sum_{a \in \Gamma_b} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \\
& + \beta \frac{p!}{N^{p-1}} \sum_{b} \sum_{a \in \bar{\Gamma}_b} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \Bigg) \\
\approx \frac{1}{2^{MN}} \frac{1}{\langle \mathcal{Z}^* \rangle} \sum_{\xi^*} \sum_{\sigma} \exp\Bigg(& \beta^* N \sum_{a \in \Gamma_*} \left[\frac{1}{N} \sum_{i} \xi^*_i \sigma^a_i \right]^{p^*} \\
& + \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}} \Bigg) \\
& \sum_{\xi} \exp\Bigg(\beta N \sum_{b} \sum_{a \in \Gamma_b} \left[\frac{1}{N} \sum_{i} \xi^b_i \sigma^a_i \right]^{p} \\
& + \beta \frac{p!}{N^{p-1}} \sum_{b} \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \Bigg) ,
\end{aligned}$$

where Γ_b represents the set of inputs σ^a which have macroscopic overlap with the pattern ξ^b, and $\bar{\Gamma} = \left[\cap_b \bar{\Gamma}_b \right] \cap \bar{\Gamma}_*$ contains almost all of the elements in each $\bar{\Gamma}_b$ and $\bar{\Gamma}_*$ for $N \to \infty$. The reasoning used to build the sets Γ_*, Γ_b and $\bar{\Gamma}$ is the same as outlined at the start of appendix 2.B.

2.D Teacher-student free entropy

Assuming that the teacher is misaligned with σ (see Case **2** of Appendix 2.A), the form of $\left\langle \mathcal{Z}^L \right\rangle$ obtained in appendix 2.C simplifies to

$$\left\langle \mathcal{Z}^L \right\rangle \approx \frac{1}{2^{MN}} \frac{1}{\langle \mathcal{Z}^* \rangle} \sum_{\xi^* \xi} \sum_{\sigma} \exp\left(\beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}} \right)$$
$$\exp\left(\beta N \sum_b \sum_{a \in \Gamma_b} \left[\frac{1}{N} \sum_i \xi^b_i \sigma^a_i \right]^p + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \right).$$

In order to evaluate $\langle \mathcal{Z}^* \rangle = \left[2^{N/M - N} Z^* \right]^M$, we recall that the teacher is a special case of the direct model with a single memory (see Section 2.3). Since the teacher is in the misaligned case, its free entropy is

$$\frac{\log(Z^*)}{N} = \begin{cases} \left(-\frac{1}{2} \log(1 - 2\beta^*) - \beta^* \right) \frac{1}{N} + \log 2 \,, & p^* = 2 \,, \\ \frac{1}{2} \left[\beta^* \right]^2 \frac{p^*!}{N^{p^*-1}} + \log 2 + \mathcal{O}\left(\frac{1}{N^{3p^*/2-2)}} \right) \,, & p^* \geq 3, \end{cases}$$

as derived in Appendix 2.B. Given $\alpha^* = \frac{M p^*!}{N^{p^*-1}}$, we use it to simplify $\frac{\log \langle \mathcal{Z}^* \rangle}{N}$ to

$$\frac{\log \langle \mathcal{Z}^* \rangle}{N} = \frac{M \log \left[2^{N/M - N} Z^* \right]}{N}$$
$$= \begin{cases} \frac{1}{2} \left(-\frac{1}{2} \log(1 - 2\beta^*) - \beta^* \right) \alpha^* + \log 2 \,, & p^* = 2 \,, \\ \frac{1}{2} \left[\beta^* \right]^2 \alpha^* + \log 2 + \mathcal{O}\left(\frac{1}{N^{p^*/2-1}} \right) \,, & p^* \geq 3 \,, \end{cases}$$

which is the paramagnetic free entropy of a p^*-body Hopfield network [54, 28]. Coming back to $\left\langle \mathcal{Z}^L \right\rangle$, we fix order parameters q^{*b}, q^{bc} and m^b_a using the delta functions $\delta\left(N q^{*b} - \sum_i \xi^*_i \xi^b_i \right)$, $\delta\left(N q^{bc} - \sum_i \xi^b_i \xi^c_i \right)$ and $\delta\left(N m^b_a - \sum_i \xi^b_i \sigma^a_i \right)$, which results in

$$\left\langle \mathcal{Z}^L \right\rangle = \frac{1}{2^{MN}} \frac{1}{\langle \mathcal{Z}^* \rangle} \sum_{\xi^* \xi} \sum_{\sigma} \int_{\mathbb{R}} \prod_b dq^{*b} \prod_{b<c} dq^{bc} \prod_b \prod_{a \in \Gamma_b} dm^b_a$$
$$\delta\left(N q^{*b} - \sum_i \xi^*_i \xi^b_i \right) \delta\left(N q^{bc} - \sum_i \xi^b_i \xi^c_i \right) \delta\left(N m^b_a - \sum_i \xi^b_i \sigma^a_i \right)$$
$$\exp\left(\beta N \sum_b \sum_{a \in \Gamma_b} \left[\frac{1}{N} \sum_i \xi^b_i \sigma^a_i \right]^p \right.$$
$$+ \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_{p^*}} \xi^*_{i_1} \ldots \xi^*_{i_{p^*}} \sigma^a_{i_1} \ldots \sigma^a_{i_{p^*}}$$
$$\left. + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_p} \xi^b_{i_1} \ldots \xi^b_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p} \right).$$

In Fourier space, this expression takes the form

$$\begin{aligned}\left\langle \mathcal{Z}^L \right\rangle = \frac{1}{\left\langle \mathcal{Z}^* \right\rangle} \sum_{\xi^* \xi} \Bigg\langle \int \prod_b dq^{*b} dr^{*b} \prod_{b<c} dq^{bc} dr^{bc} \prod_b \prod_{a \in \Gamma_b} dm_a^b dk_a^b \\ \exp\left\{ \beta^* \beta \alpha \sum_b \left(\sum_i \xi_i^* \xi_i^b - N q^{*b} \right) r^{*b} + \beta^2 \alpha \sum_{b<c} \left(\sum_i \xi_i^b \xi_i^c - N q^{bc} \right) r^{bc} \right\} \\ \exp\left\{ \beta \sum_b \sum_{a \in \Gamma_b} \left(\sum_i \xi_i^b \sigma_i^a - N m_a^b \right) k_a^b + \beta N \sum_b \sum_{a \in \Gamma_b} \left[\frac{1}{N} \sum_i \xi_i^b \sigma_i^a \right]^p \right. \\ + \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_{p^*}} \xi_{i_1}^* \ldots \xi_{i_{p^*}}^* \sigma_{i_1}^a \ldots \sigma_{i_{p^*}}^a \\ \left. + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{a \in \bar{\Gamma}} \sum_{i_1 < \ldots < i_p} \xi_{i_1}^b \ldots \xi_{i_p}^b \sigma_{i_1}^a \ldots \sigma_{i_p}^a \right\} \Bigg\rangle_\sigma ,\end{aligned}$$

where the sum over σ with a pre-factor of $\frac{1}{2^{MN}}$ was replaced by the uniform average $\langle \rangle_\sigma$. Following the same reasoning as in appendix 2.B, a second order cumulant expansion of the last two terms for any $a \in \bar{\Gamma}$ yields

$$\begin{aligned}\log \left\langle \exp\left\{ \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{i_1 < \ldots < i_{p^*}} \xi_{i_1}^* \ldots \xi_{i_{p^*}}^* \sigma_{i_1}^a \ldots \sigma_{i_{p^*}}^a + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{i_1 < \ldots < i_p} \xi_{i_1}^b \ldots \xi_{i_p}^b \sigma_{i_1}^a \ldots \sigma_{i_p}^a \right\} \right\rangle \\ \approx \frac{1}{2} \beta^2 \left[\frac{p!}{N^{p-1}} \right]^2 \sum_{b \neq c} \sum_{i_1 < \ldots < i_p} \sum_{j_1 < \ldots < j_p} \xi_{i_1}^b \xi_{j_1}^c \ldots \xi_{i_p}^b \xi_{j_p}^c \left\langle \sigma_{i_1}^a \sigma_{j_1}^a \ldots \sigma_{i_p}^a \sigma_{j_p}^a \right\rangle \\ + \beta^* \beta \frac{p^*!}{N^{p^*-1}} \frac{p!}{N^{p-1}} \sum_b \sum_{i_1 < \ldots < i_{p^*}} \sum_{j_1 < \ldots < j_p} \left\langle \xi_{i_1}^* \sigma_{i_1}^a \ldots \xi_{i_{p^*}}^* \sigma_{i_{p^*}}^a \xi_{j_1}^b \sigma_{j_1}^a \ldots \xi_{j_p}^b \sigma_{j_p}^a \right\rangle \\ + \frac{1}{2} \beta^2 \frac{p!}{N^{p-1}} LN + \frac{1}{2} \left[\beta^* \right]^2 \frac{p^*!}{N^{p^*-1}} N .\end{aligned}$$

When $p^* = p$, it reduces to

$$\begin{aligned}\log \left\langle \exp\left\{ \beta^* \frac{p^*!}{N^{p^*-1}} \sum_{i_1 < \ldots < i_{p^*}} \xi_{i_1}^* \ldots \xi_{i_{p^*}}^* \sigma_{i_1}^a \ldots \sigma_{i_{p^*}}^a + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{i_1 < \ldots < i_p} \xi_{i_1}^b \ldots \xi_{i_p}^b \sigma_{i_1}^a \ldots \sigma_{i_p}^a \right\} \right\rangle \\ = \beta^2 \frac{p!}{N^{p-1}} N \sum_{b<c} \left[\frac{1}{N} \sum_i \xi_i^b \xi_i^c \right]^p + \beta^* \beta \frac{p!}{N^{p-1}} N \sum_b \left[\frac{1}{N} \sum_i \xi_i^* \xi_i^b \right]^p \\ + \frac{1}{2} \beta^2 \frac{p!}{N^{p-1}} LN + \frac{1}{2} \left[\beta^* \right]^2 \frac{p!}{N^{p-1}} N ,\end{aligned}$$

because $\left\langle \sigma_{i_n}^a \sigma_{j_n}^a \right\rangle = \delta_{i_n j_n}$ (see Appendix 2.B for more details). On the contrary, the second order expectation $\left\langle \xi_{i_1}^* \sigma_{i_1}^a \ldots \xi_{i_{p^*}}^* \sigma_{i_{p^*}}^a \xi_{j_1}^b \sigma_{j_1}^a \ldots \xi_{j_p}^b \sigma_{j_p}^a \right\rangle$ vanishes when $p^* \neq p$. In fact, spins come in pairs $\left\langle \sigma_{i_n}^a \sigma_{j_n}^a \right\rangle = \delta_{i_n j_n}$ only up to $n \leq \min\{p^*, p\}$, and the remaining single-spin averages $\left\langle \sigma_{i_n}^a \right\rangle = 0$ make the second order expectation vanish.

We need to go beyond second order to treat $p^* \neq p$. We will focus on $p^* = 2$ and $p \geq 3$ to investigate the consequences of using a p-body model to learn examples generated by the original 2-body Hopfield model.

For simplicity, we take p even so that the spins of both terms can be grouped in pairs at order $\frac{p}{2}+1$, when the teacher term $\beta^* \frac{2}{N} \sum_{i_1<i_2} \xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}$ is raised to the power of $\frac{p}{2}$ and the student term is raised to the power of 1. This restriction will simplify some of the incoming calculations. To leading order in N, the cumulant generating function reduces to

$$\begin{aligned}
&\log\left\langle \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}+\beta\frac{p!}{N^{p-1}}\sum_b\sum_{i_1<\ldots<i_p}\xi^b_{i_1}\ldots\xi^b_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right\}\right\rangle \\
&\quad\approx \log\Bigg[\left\langle \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \\
&\qquad\left\langle \exp\left\{\beta\frac{p!}{N^{p-1}}\sum_b\sum_{i_1<\ldots<i_p}\xi^b_{i_1}\ldots\xi^b_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right\}\right\rangle \\
&\qquad+\left\langle \beta\frac{p!}{N^{p-1}}\sum_b\sum_{j_1<\ldots<j_p}\xi^b_{j_1}\ldots\xi^b_{j_p}\sigma^a_{j_1}\ldots\sigma^a_{j_p}\exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle\Bigg],
\end{aligned}$$

where the last term encompasses the teacher-student coupling that allows retrieval to take place. The teacher term

$$\log\left\langle \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \approx -\frac{1}{2}\log\left(1-2\beta^*\right)-\beta^*,$$

and the student term

$$\begin{aligned}
&\log\left\langle \exp\left\{\beta\frac{p!}{N^{p-1}}\sum_b\sum_{i_1<\ldots<i_p}\xi^b_{i_1}\ldots\xi^b_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}\right\}\right\rangle \\
&\qquad\approx \beta^2\frac{p!}{N^{p-1}}N\sum_{b<c}\left[\frac{1}{N}\sum_i\xi^b_i\xi^c_i\right]^p+\frac{1}{2}\beta^2\frac{p!}{N^{p-1}}LN,
\end{aligned}$$

are both known from Appendix 2.B. Later on, we will use $\log(z^*)$ and z^* as shorthands for $-\frac{1}{2}\log\left(1-2\beta^*\right)-\beta^*$ and $\exp\left(-\frac{1}{2}\log\left(1-2\beta^*\right)-\beta^*\right)$, respectively. The coupling between the teacher and the student can be rewritten as

$$\begin{aligned}
&\left\langle \beta\frac{p!}{N^{p-1}}\sum_b\sum_{j_1<\ldots<j_p}\xi^b_{j_1}\ldots\xi^b_{j_p}\sigma^a_{j_1}\ldots\sigma^a_{j_p}\exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \\
&\quad=\left\langle \beta\frac{p!}{N^{p-1}}\sum_b\sum_{j_1<\ldots<j_p}\xi^*_{j_1}\ldots\xi^*_{j_p}\xi^b_{j_1}\ldots\xi^b_{j_p}\,\xi^*_{j_1}\ldots\xi^*_{j_p}\sigma^a_{j_1}\ldots\sigma^a_{j_p}\exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \\
&\quad=\beta\frac{p!}{N^{p-1}}\sum_b\sum_{j_1<\ldots<j_p}\xi^*_{j_1}\ldots\xi^*_{j_p}\xi^b_{j_1}\ldots\xi^b_{j_p}\left\langle \xi^*_{j_1}\ldots\xi^*_{j_p}\sigma^a_{j_1}\ldots\sigma^a_{j_p}\exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle,
\end{aligned}$$

because $\left[\xi^*_{j_n}\right]^2=1$ for every index j_n. All interacting spin tuples of the form $\xi^*_{j_1}\ldots\xi^*_{j_p}\sigma^a_{j_1}\ldots\sigma^a_{i_p}$ are statistically

equivalent as long as $j_1 < ... < j_p$, so the teacher-student coupling simplifies to

$$\begin{aligned}
&\left\langle \beta \frac{p!}{N^{p-1}} \sum_b \sum_{j_1<...<j_p} \xi_{j_1}^b ... \xi_{j_p}^b \sigma_{j_1}^a ... \sigma_{j_p}^a \exp\left\{ \beta^* \frac{2}{N} \sum_{i_1<i_2} \xi_{i_1}^* \xi_{i_2}^* \sigma_{i_1}^a \sigma_{i_2}^a \right\} \right\rangle \\
&= \beta \frac{p!}{N^{p-1}} \sum_b \sum_{i_1<...<i_p} \xi_{i_1}^* ... \xi_{i_p}^* \xi_{i_1}^b ... \xi_{i_p}^b \\
&\quad \left\langle \frac{p!}{N^p} \sum_{j_1<...<j_p} \xi_{j_1}^* ... \xi_{j_p}^* \sigma_{j_1}^a ... \sigma_{i_p}^a \exp\left\{ \beta^* \frac{2}{N} \sum_{i_1<i_2} \xi_{i_1}^* \xi_{i_2}^* \sigma_{i_1}^a \sigma_{i_2}^a \right\} \right\rangle \\
&= V(\beta^*, p)\, \beta \frac{p!}{N^{p-1}} \sum_b \sum_{i_1<...<i_p} \xi_{i_1}^* ... \xi_{i_p}^* \xi_{i_1}^b ... \xi_{i_p}^b ,
\end{aligned}$$

where $V(\beta^*, p) = \left\langle \frac{p!}{N^p} \sum_{j_1<...<j_p} \xi_{j_1}^* ... \xi_{j_p}^* \sigma_{j_1}^a ... \sigma_{i_p}^a \exp\left(\beta^* \frac{2}{N} \sum_{i_1<i_2} \xi_{i_1}^* \xi_{i_2}^* \sigma_{i_1}^a \sigma_{i_2}^a\right) \right\rangle$ does not depend on the microscopic details of the system. In fact, it can be expressed as a combination of the moments of z^*, which can all be derived from $\log(z^*)$. To leading order in N, the cumulant generating function expands to

$$\begin{aligned}
&\log \left\langle \exp\left\{ \beta^* \frac{2}{N} \sum_{i_1<i_2} \xi_{i_1}^* \xi_{i_2}^* \sigma_{i_1}^a \sigma_{i_2}^a + \beta \frac{p!}{N^{p-1}} \sum_b \sum_{i_1<...<i_p} \xi_{i_1}^b ... \xi_{i_p}^b \sigma_{i_1}^a ... \sigma_{i_p}^a \right\} \right\rangle \\
&\approx -\frac{1}{2} \log(1 - 2\beta^*) - \beta^* + \beta^2 \frac{p!}{N^{p-1}} N \sum_{b<c} \left[\frac{1}{N} \sum_i \xi_i^b \xi_i^c \right]^p + \frac{1}{2} \beta^2 \frac{p!}{N^{p-1}} LN \\
&\quad + [1 - 2\beta^*]^{1/2} \exp(\beta^*)\, V(\beta^*, p)\, \beta N \sum_b \left[\frac{1}{N} \sum_i \xi_i^* \xi_i^b \right]^p .
\end{aligned}$$

At this stage, we only need to find $V(\beta^*, p)$ in order to solve the system. We focus on two different scalings of M and β^* that make the teacher-student coupling leading order in N:

1 $M \sim \mathcal{O}\left(N^{p-1}\right)$ and $\beta^* \sim \mathcal{O}\left(N^{2/p-1}\right)$ will be called the large-noise scaling.

2 $M \sim \mathcal{O}\left(N^{p/2}\right)$ and $\beta^* \sim \mathcal{O}(1)$ will be called the finite-noise scaling.

The student term vanishes in the first scenario but is leading order in the second one. The case of the teacher-student coupling is more subtle. When β^* is small, we may keep only the first non-vanishing order of the exponential function present in the definition of $V(\beta^*, p)$. Since p is even, it leads to

$$\begin{aligned}
V(\beta^*, p) &\approx \frac{1}{(p/2)!} \left\langle \frac{p!}{N^p} \sum_{j_1<...<j_p} \xi_{j_1}^* ... \xi_{j_p}^* \sigma_{j_1}^a ... \sigma_{j_p}^a \left(\beta^* \frac{2}{N} \sum_{i_1<i_2} \xi_{i_1}^* \xi_{i_2}^* \sigma_{i_1}^a \sigma_{i_2}^a \right)^{p/2} \right\rangle \qquad (2.11) \\
&= \frac{[\beta^*]^{p/2}}{(p/2)!} \frac{2^{p/2}}{N^{p/2}} \frac{p!}{2^{p/2}} \\
&= \frac{[\beta^*]^{p/2}}{(p/2)!} \frac{p!}{N^{p/2}} ,
\end{aligned}$$

because there are $\prod_{n=1}^{p/2} \binom{2n}{2} = \frac{p!}{2^{p/2}}$ spin pairings with non-zero expectation that satisfy the inequality

constraints. In the large-noise scaling, we set

$$\lambda = \frac{[\beta^*]^{p/2}}{(p/2)!} N^{p/2-1} \sim \mathcal{O}(1) \ ,$$

to get the asymptotically exact expression $V\left(\left[(p/2)!\right]^{2/p} N^{1-2/p}, p\right) = \lambda \frac{p!}{N^{p-1}}$. In the finite-noise scaling, this expansion is only an order of magnitude approximation. However, it still indicates that $V(\beta^*, p)$ is $\mathcal{O}\left(N^{-p/2}\right)$ when β^* is $\mathcal{O}(1)$ in N. In other words, it shows that there is an $\mathcal{O}(1)$ parameter η such that $V(\beta^*(\eta, p), p) = \eta \frac{(p/2+1)!}{N^{p/2}}$. We will now use the cumulants $\frac{\partial \log(z^*)}{\partial \beta^*}$ and $\frac{\partial \log(z^*)}{\partial \beta^{*2}}$ of z^* to derive the value of η corresponding to $p = 4$. First of all, note that $\frac{4!}{N^4} \sum_{j_1<\ldots<j_4} \xi^*_{j_1} \ldots \xi^*_{j_4} \sigma^a_{j_1} \ldots \sigma^a_{j_4}$ can be expressed as:

$$\begin{aligned}
&\frac{24}{N^4} \sum_{j_1<j_2<j_3<j_4} \xi^*_{j_1}\xi^*_{j_2}\xi^*_{j_3}\xi^*_{j_4}\sigma^a_{j_1}\sigma^a_{j_2}\sigma^a_{j_3}\sigma^a_{j_4} \\
&= \frac{1}{N^4} \sum_{j_1\neq j_2\neq j_3\neq j_4} \xi^*_{j_1}\xi^*_{j_2}\xi^*_{j_3}\xi^*_{j_4}\sigma^a_{j_1}\sigma^a_{j_2}\sigma^a_{j_3}\sigma^a_{j_4} \\
&= \frac{1}{N^4} \left[\sum_{j_1\neq j_2} \xi^*_{j_1}\xi^*_{j_2}\sigma^a_{j_1}\sigma^a_{j_2}\right]\left[\sum_{j_3\neq j_4} \xi^*_{j_3}\xi^*_{j_4}\sigma^a_{j_3}\sigma^a_{j_4}\right] - \frac{4}{N^3}\left[\sum_{j_1\neq j_2} \xi^*_{j_1}\xi^*_{j_2}\sigma^a_{j_1}\sigma^a_{j_2}\right] - \frac{2}{N^2} \\
&= \frac{1}{N^2}\left[\frac{2}{N}\sum_{j_1<j_2} \xi^*_{j_1}\xi^*_{j_2}\sigma^a_{j_1}\sigma^a_{j_2}\right]^2 - \frac{4}{N^2}\left[\frac{2}{N}\sum_{j_1<j_2} \xi^*_{j_1}\xi^*_{j_2}\sigma^a_{j_1}\sigma^a_{j_2}\right] - \frac{2}{N^2} \ ,
\end{aligned}$$

by subtracting the diagonals where pairs of indices are equal. Therefore, $\frac{1}{z^*} V(\beta^*, p)$ reduces to

$$\begin{aligned}
\frac{1}{z^*}V(\beta^*,p) &= \left\langle \frac{24}{N^4} \sum_{j_1<j_2<j_3<j_4} \xi^*_{j_1}\xi^*_{j_2}\xi^*_{j_3}\xi^*_{j_4}\sigma^a_{i_1}\sigma^a_{i_2}\sigma^a_{i_3}\sigma^a_{i_4} \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \\
&= \frac{1}{z^*}\frac{1}{N^2}\left[\left\langle\left[\frac{2}{N}\sum_{j_1<j_2}\xi^*_{j_1}\xi^*_{j_2}\sigma^a_{i_1}\sigma^a_{i_2}\right]^2 \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle\right. \\
&\quad - 4\left\langle\left[\frac{2}{N}\sum_{j_1<j_2}\xi^*_{j_1}\xi^*_{j_2}\sigma^a_{i_1}\sigma^a_{i_2}\right] \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle \\
&\quad \left. - 2\left\langle \exp\left\{\beta^*\frac{2}{N}\sum_{i_1<i_2}\xi^*_{i_1}\xi^*_{i_2}\sigma^a_{i_1}\sigma^a_{i_2}\right\}\right\rangle\right] \\
&= \frac{1}{N^2}\left[\frac{\partial\log(z^*)}{\partial\beta^{*2}} + \left[\frac{\partial\log(z^*)}{\partial\beta^*}\right]^2 - 4\frac{\partial\log(z^*)}{\partial\beta^*} - 2\right] .
\end{aligned}$$

The cumulants evaluate to

$$\begin{aligned}
\frac{\partial\log(z^*)}{\partial\beta^*} &= \frac{\partial}{\partial\beta^*}\left[-\frac{1}{2}\log(1-2\beta^*) - \beta^*\right] = \frac{2\beta^*}{1-2\beta^*} \ , \\
\frac{\partial\log(z^*)}{\partial\beta^{*2}} &= \frac{\partial}{\partial\beta^{*2}}\left[-\frac{1}{2}\log(1-2\beta^*) - \beta^*\right] = \frac{2}{(1-2\beta^*)^2} \ ,
\end{aligned}$$

so we obtain

$$\begin{aligned}\frac{1}{z^*}V\left(\beta^*,p\right) &= \frac{1}{N^2}\left[\frac{2}{\left(1-2\beta^*\right)^2}+\frac{4\left[\beta^*\right]^2}{\left(1-2\beta^*\right)^2}-\frac{8\beta^*}{1-2\beta^*}-2\right] \\ &= \frac{6}{N^2}\frac{2\left[\beta^*\right]^2}{\left(1-2\beta^*\right)^2}\,.\end{aligned}$$

In other terms, we find $\eta = \frac{2[\beta^*]^2}{(1-2\beta^*)^2}$ when $p=4$. In summary, depending on the scaling, the teacher student coupling either simplifies to

1 $\beta\lambda\alpha\frac{N}{M}\sum_b\left[\frac{1}{N}\sum_i \xi_i^*\xi_i^b\right]^p$ where $\alpha = \frac{Mp!}{N^{p-1}}$ and $\lambda = \frac{[\beta^*]^{p/2}}{(p/2)!}N^{p/2-1}$ are finite,

2 or $\beta\eta\alpha\frac{N}{M}\sum_b\left[\frac{1}{N}\sum_i \xi_i^*\xi_i^b\right]^p$ where $\alpha = \frac{M(p/2+1)!}{N^{p/2}}$ and η are finite.

In either case, the result is similar to $p^*=p$ except for its pre-factor. We describe the rest of the derivation only for $p^*=p$ because the $p^*=2$ and $p\geq 3$ calculations are almost identical. Putting the result of the $p^*=p$ cumulant expansion back in $\left\langle \mathcal{Z}^L\right\rangle$, we get:

$$\begin{aligned}\left\langle \mathcal{Z}^L\right\rangle \approx \frac{1}{\left\langle \mathcal{Z}^*\right\rangle}\sum_{\xi^*\xi}\Bigg\langle \int\prod_b dq^{*b}dr^{*b}\prod_{b<c}dq^{bc}dr^{bc}\prod_b\prod_{a\in\Gamma_b}dm_a^b\,dk_a^b \\ \exp\left\{\beta^*\beta\alpha\sum_b\left(\sum_i\xi_i^*\xi_i^b-Nq^{*b}\right)r^{*b}+\beta^2\alpha\sum_{b<c}\left(\sum_i\xi_i^b\xi_i^c-Nq^{bc}\right)r^{bc}\right\} \\ \exp\left\{\beta\sum_b\sum_{a\in\Gamma_b}\left(\sum_i\xi_i^b\sigma_i^a-Nm_a^b\right)k_a^b+\beta N\sum_b\sum_{a\in\Gamma_b}\left[m_a^b\right]^p\right\} \\ \exp\left\{\beta^*\beta\alpha N\sum_b\left[q^{*b}\right]^p+\beta^2\alpha N\sum_{b<c}\left[q^{bc}\right]^p+\frac{1}{2}\beta^2\alpha LN+\frac{1}{2}\left[\beta^*\right]^2\alpha N\right\}\Bigg\rangle,\end{aligned}$$

where $\alpha = \frac{Mp!}{N^{p-1}}$. The saddle point of $\langle \mathcal{Z}^L \rangle$ then evaluates to

$$\begin{aligned}
\frac{\log \langle \mathcal{Z}^L \rangle}{N} \approx \underset{m,k,q,r,q^*,r^*}{\text{Extr}} \Bigg[& \beta^* \beta \alpha \sum_b \left[q^{*b}\right]^p + \beta^2 \alpha \sum_{b<c} \left[q^{bc}\right]^p + \beta \sum_b \sum_{a \in \Gamma_b} \left[m_a^b\right]^p \\
& - \beta^* \beta \alpha \sum_b r^{*b} q^{*b} - \beta^2 \alpha \sum_{b<c} r^{bc} q^{bc} - \beta \sum_b \sum_{a \in \Gamma_b} m_a^b k_a^b \\
& + \frac{1}{2} \beta^2 \alpha L + \frac{1}{2} \left[\beta^*\right]^2 \alpha - \frac{\log \langle \mathcal{Z} \rangle}{N} + \log 2 \\
& + \frac{1}{N} \log \Bigg\langle \sum_{\xi} \exp \Bigg\{ \beta \sum_b \sum_{a \in \Gamma_b} k_a^b \sum_i \xi_i^b \sigma_i^a \\
& + \beta^* \beta \alpha \sum_b r^{*b} \sum_i \xi_i^* \xi_i^b + \beta^2 \alpha \sum_{b<c} r^{bc} \sum_i \xi_i^b \xi_i^c \Bigg\} \Bigg\rangle_{\xi^* \sigma} \Bigg] \\
= \text{Extr} \Bigg[& \beta^* \beta \alpha \sum_b \left[q^{*b}\right]^p + \beta^2 \alpha \sum_{b<c} \left[q^{bc}\right]^p + \beta \sum_b \sum_{a \in \Gamma_b} \left[m_a^b\right]^p \\
& - \beta^* \beta \alpha \sum_b r^{*b} q^{*b} - \beta^2 \alpha \sum_{b<c} r^{bc} q^{bc} - \beta \sum_b \sum_{a \in \Gamma_b} m_a^b k_a^b \\
& + \frac{1}{2} \beta^2 \alpha L + \frac{1}{2} \left[\beta^*\right]^2 \alpha - \frac{\log \langle \mathcal{Z} \rangle}{N} + \log 2 \\
& + \frac{1}{N} \sum_i \log \Bigg\langle \sum_{\xi_i} \exp \Bigg\{ \beta \sum_b \sum_{a \in \Gamma_b} k_a^b \xi_i^b \sigma_i^a \\
& + \beta^* \beta \alpha \sum_b r^{*b} \xi_i^* \xi_i^b + \beta^2 \alpha \sum_{b<c} r^{bc} \xi_i^b \xi_i^c \Bigg\} \Bigg\rangle_{\xi_i^* \sigma_i} \Bigg],
\end{aligned}$$

where the average over ξ^* and σ is uniform. We use $\frac{\log \langle \mathcal{Z}^* \rangle}{N} = \frac{1}{2} \left[\beta^*\right]^2 \alpha + \log 2$ to simplify $\frac{\log \langle \mathcal{Z}^L \rangle}{N}$ to

$$\begin{aligned}
\frac{\log \langle \mathcal{Z}^L \rangle}{N} \approx \text{Extr} \Bigg[& \beta^* \beta \alpha \sum_b \left[q^{*b}\right]^p + \beta^2 \alpha \sum_{b<c} \left[q^{bc}\right]^p + \beta \sum_b \sum_{a \in \Gamma_b} \left[m_a^b\right]^p \\
& - \beta^* \beta \alpha \sum_b r^{*b} q^{*b} - \beta^2 \alpha \sum_{b<c} r^{bc} q^{bc} - \beta \sum_b \sum_{a \in \Gamma_b} m_a^b k_a^b \\
& + \frac{1}{2} \beta^2 \alpha L + \frac{1}{N} \sum_i \log \Bigg\langle \sum_{\xi_i} \exp \Bigg\{ \beta \sum_b \sum_{a \in \Gamma_b} k_a^b \xi_i^b \sigma_i^a \\
& + \beta^* \beta \alpha \sum_b r^{*b} \xi_i^* \xi_i^b + \beta^2 \alpha \sum_{b<c} r^{bc} \xi_i^b \xi_i^c \Bigg\} \Bigg\rangle_{\xi_i^* \sigma_i} \Bigg].
\end{aligned}$$

Assuming each ξ^b has macroscopic overlap with at most one pattern σ^a and using the replica-symmetric ansatz

$q^{*b} = q^*$, $q^{bc} = q$, $r^{*b} = r^*$, $r^{bc} = r$, $m_a^b = m$, $k_a^b = k$, the free entropy approximates to

$$\begin{aligned}
f &= \lim_{N\to\infty, L\to 0} \left(\frac{\partial}{\partial L} \left[\frac{1}{N} \log \left\langle Z^L \right\rangle \right] \right) \\
&\approx \operatorname{Extr} \Bigg[\beta^* \beta \alpha \left[q^* \right]^p - \frac{1}{2} \beta^2 \alpha q^p + \beta m^p - \beta^* \beta \alpha r^* q^* + \frac{1}{2} \beta^2 \alpha r q \\
&\quad - \beta m k + \frac{1}{2} \beta^2 \alpha + \lim_{L\to 0} \left(\frac{\partial}{\partial L} \left[\log \left\langle \sum_{\xi_i} \exp \left\{ \beta k \sum_b \xi_i^b \sigma_i^a \right. \right. \right. \right. \\
&\quad \left. \left. \left. \left. + \beta^* \beta \alpha r^* \sum_b \xi_i^* \xi_i^b + \beta^2 \alpha r \sum_{b<c} \xi_i^b \xi_i^c \right\} \right\rangle \right] \right) \Bigg] .
\end{aligned}$$

Furthermore, the Hubbard-Stratonovich transformation gives

$$\exp \left\{ \beta^2 \alpha r \sum_{b<c} \xi_i^b \xi_i^c \right\} \propto \exp \left\{ -\frac{1}{2} \beta^2 \alpha r L \right\} \int_{\mathbb{R}} dx \exp \left\{ -\frac{1}{2} x^2 + x \beta \sqrt{\alpha r} \sum_b \xi_i^b \right\} .$$

We can then use the factorization

$$\begin{aligned}
&\sum_{\xi_i} \exp \left\{ \beta k \sum_b \xi_i^b \sigma_i^a + \beta^* \beta \alpha r^* \sum_b \xi_i^* \xi_i^b + x \beta \sqrt{\alpha r} \sum_b \xi_i^b \right\} \\
&\quad = \prod_b \sum_{\xi_i^b} \exp \left\{ \beta k \xi_i^b \sigma_i^a + \beta^* \beta \alpha r^* \xi_i^* \xi_i^b + x \beta \sqrt{\alpha r} \xi_i^b \right\} \\
&\quad = \prod_b \left[2 \cosh \left(\beta k \sigma_i^a + \beta^* \beta \alpha r^* \xi_i^* + x \beta \sqrt{\alpha r} \right) \right] \\
&\quad = 2^L \cosh^L \left(\beta k \sigma_i^a + \beta^* \beta \alpha r^* \xi_i^* + x \beta \sqrt{\alpha r} \right) ,
\end{aligned}$$

in order to simplify the free energy to

$$\begin{aligned}
f &= \operatorname{Extr} \Bigg\{ \beta^* \beta \alpha \left[q^* \right]^p - \frac{1}{2} \beta^2 \alpha q^p + \beta m^p - \beta^* \beta \alpha r^* q^* + \frac{1}{2} \beta^2 \alpha r q - \beta m k \\
&\quad - \frac{1}{2} \beta^2 \alpha r + \frac{1}{2} \beta^2 \alpha + \lim_{L\to 0} \left(\frac{\partial}{\partial L} \left[\log \left\langle \sum_{\xi_i} \int_{\mathbb{R}} dx \exp \left\{ -\frac{1}{x} x^2 \right\} \right. \right. \right. \\
&\quad \left. \left. \left. \exp \left\{ \beta k \sum_b \xi_i^b \sigma_i^a + \beta^* \beta \alpha r^* \sum_b \xi_i^* \xi_i^b + x \beta \sqrt{\alpha r} \sum_b \xi_i^b \right\} \right\rangle \right] \right) \Bigg\} \\
&= \operatorname{Extr} \Bigg\{ \beta^* \beta \alpha \left[q^* \right]^p - \frac{1}{2} \beta^2 \alpha q^p + \beta m^p - \beta^* \beta \alpha r^* q^* + \frac{1}{2} \beta^2 \alpha r q - \beta m k \\
&\quad - \frac{1}{2} \beta^2 \alpha r + \frac{1}{2} \beta^2 \alpha + \log 2 + \lim_{L\to 0} \left(\frac{\partial}{\partial L} \left[\log \left\langle \int_{\mathbb{R}} dx \exp \left\{ -\frac{1}{x} x^2 \right\} \right. \right. \right. \\
&\quad \left. \left. \left. \cosh^L \left(\beta k \sigma_i^a + \beta^* \beta \alpha r^* \xi_i^* + x \beta \sqrt{\alpha r} \right) \right\rangle \right] \right) \Bigg\} .
\end{aligned}$$

After differentiating and taking the limit, we get

$$f = \underset{m,k,q,r,q^*,r^*}{\text{Extr}} \Bigg\{ \beta^*\beta\alpha\,[q^*]^p - \frac{1}{2}\beta^2\alpha q^p + \beta m^p - \beta^*\beta\alpha r^* q^* \\ + \frac{1}{2}\beta^2\alpha r q - \frac{1}{2}\beta^2\alpha r - \beta m k + \frac{1}{2}\beta^2\alpha + \log 2 \\ + \int dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \Big\langle \log\left[\cosh\left(\beta\left[\sqrt{\alpha r}x + \beta^*\alpha r^* + kz\right]\right)\right] \Big\rangle_z \Bigg\}.$$

In the case of $p^* = 2$ and $p \geq 3$ with finite $\alpha = \frac{Mp!}{N^{p-1}}$ and $\lambda = \frac{[\beta^*]^{p/2}}{(p/2)!} N^{p/2-1}$, the free energy has the same form but with β^* replaced by λ. On the other other hand, the free energy with finite $\alpha = \frac{M(p/2+1)!}{N^{p/2}}$ and η evaluates to:

$$f = \underset{m,k,q^*,r^*}{\text{Extr}} \left\{ \beta\eta\alpha\,[q^*]^p - \beta m^p - \beta\eta\alpha r^* q^* - \beta m k + \log 2 + \Big\langle \log\left[\cosh\left(\beta\left[\eta\alpha r^* + kz\right]\right)\right] \Big\rangle_z \right\}.$$

2.E Direct model RSB ansatz

Recall that the average replicated partition function of the direct model (see Eq. 2.10) takes the form

$$\left\langle Z^L \right\rangle \approx \left\langle \sum_{\sigma} \exp\left(\beta N \sum_{\gamma} \sum_{\mu\in\Gamma_\gamma} \left[\frac{1}{N}\sum_i \xi_i^\mu \sigma_i^\gamma\right]^p + \beta \sum_\gamma \sum_{\mu\in\bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \right\rangle. \tag{2.12}$$

Introducing a new replica σ^0, we rewrite it as

$$\left\langle Z^L \right\rangle = \left\langle \sum_{\sigma} \exp\left(\beta N \sum_{\gamma} \sum_{\mu\in\Gamma_\gamma} \left[\frac{1}{N}\sum_i \xi_i^\mu \sigma_i^\gamma\right]^p + \beta \sum_\gamma \sum_{\mu\in\bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \frac{\langle Z\rangle}{\langle Z\rangle} \right\rangle,$$

where $Z = \sum_{\sigma_0} \exp\left(\beta \sum_{\mu\in\bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^0 \ldots \sigma_{i_p}^0\right)$. Recall that, in the paramagnetic phase, we have (see [28] and also Appendix 2.B)

$$\langle Z\rangle = \exp\left(\frac{1}{2}\beta^2\alpha + \log 2 + \mathcal{O}\left(\frac{1}{N^{p/2-2}}\right)\right) = Z \exp\left(\mathcal{O}\left(\frac{1}{N^{p/2-2}}\right)\right),$$

so $\left\langle Z^L \right\rangle$ can be expressed as

$$\left\langle Z^L \right\rangle = \frac{1}{\langle Z\rangle} \Bigg\langle \sum_{\sigma} \exp\left(\beta N \sum_{\gamma} \sum_{\mu\in\Gamma_\gamma} \left[\frac{1}{N}\sum_i \xi_i^\mu \sigma_i^\gamma\right]^p + \beta \sum_\gamma \sum_{\mu\in\bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^\gamma \ldots \sigma_{i_p}^\gamma \right) \\ \sum_{\sigma_0} \exp\left(\beta \sum_{\mu\in\bar{\Gamma}} \frac{p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \xi_{i_1}^\mu \ldots \xi_{i_p}^\mu \sigma_{i_1}^0 \ldots \sigma_{i_p}^0 + \mathcal{O}\left(\frac{1}{N^{p/2-2}}\right)\right) \Bigg\rangle.$$

The $\mathcal{O}\left(\frac{1}{N^{p/2-2}}\right)$ corrections vanish to leading order in N when we calculate the free entropy.

2.F Monte Carlo simulations for various system sizes

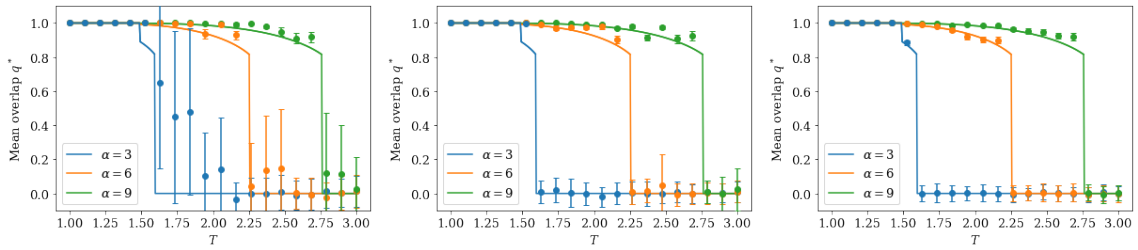

Figure 2.9: Monte Carlo simulations of the $p = 3$ inverse model compared against saddle-point solutions for different values of N. The lR phase is not included in these plots. The left plot has $N = 128$, the center plot has $N = 256$, and the right plot has $N = 512$. The dots are simulation data at a few values of α, and the lines are slices of the saddle-point solutions at the same α. There are $M = \frac{\alpha N^{p-1}}{p!}$ examples σ^a, and simulation results are averaged over $L = 100$ student patterns. The simulation data is sometimes systematically shifted up with respect to the saddle-point solution, but the size of the difference tends to decrease with N. The shift is the most visible when $\alpha = 6$ and right after the fall from eR to gR when $\alpha = 3$. As expected, the fluctuations of the paramagnetic phase also decrease with N.

Chapter 3

Modeling structured data learning with Restricted Boltzmann machines in the teacher–student setting

Based on the article [46], doi: 10.1016/j.neunet.2025.107542

3.1 Introduction

Restricted Boltzmann machines (RBM) [136, 137, 138, 24] are empirically known to fit complicated data and then sample new instances that are faithful to the underlying distribution [24, 139, 140]. For example, [69] trained RBMs to predict the interests of Netflix users by fitting a database of movie ratings, [141] used them to generate realistic textures and [142] employed them for topic modelling. There are even universal approximation theorems stating that an RBM with an arbitrary number of hidden units can approximate any distribution with binary support arbitrarily well [143, 144]. The current theoretical understanding of RBMs is largely based on statistical mechanics. For example, [31, 32] used statistical mechanics to model RBM training on real data. The statistical mechanics community also studied RBMs trained on simpler, synthetic datasets [31, 32]. In particular, many works investigated the *teacher-student setting* where a *student* RBM is trained with data produced by a *teacher* RBM [145, 146, 39, 101, 147, 42, 40, 44]. Such studies are crucial to isolate individual characteristics of structured datasets and neural network (NN) design choices in a controlled environment and explain their effects on NN training. For example, [39, 101, 44] investigated the effects of the prior chosen for the data on RBM learning. In this paper, we study the effects of data correlations and number of hidden units on RBMs in the teacher-student setting. In particular, we show that RBMs with a few hidden units in the teacher-student setting can serve as a toy model of the lottery ticket hypothesis [9, 148, 149, 150, 151].

RBMs have a visible layer $\sigma = \{\sigma_i\}_{i=1}^{N}$, a hidden layer $\tau = \{\tau_\mu\}_{\mu=1}^{P}$ and a set of internal connections $\xi = \{\xi_i^\mu\}_{1\leq i\leq N}^{1\leq \mu\leq P}$, which are commonly referred to as weights. The visible layer is a set of concrete features found directly in the data, the hidden layer is an internal representation of the data in terms of abstract concepts, and the weights represent how the input features and abstract concepts are correlated with one another. To give a caricatural example, the presence of pointy ears as an input feature could be correlated with the abstract concept of a cat in the weights of a particular RBM. Given ξ, the visible and hidden layer follow the joint distribution

$$\mathrm{P}_\beta\left(\sigma,\tau \mid \xi\right) = Z_\beta\left(\xi\right)^{-1}\mathrm{P}\left(\sigma\right)\mathrm{P}\left(\tau\right)\exp\left(-\beta H\left[\sigma,\tau;\xi\right]\right), \tag{3.1}$$

where the Hamiltonian $H\left[\sigma,\tau;\xi\right] = -\frac{1}{\sqrt{N}}\sum_{\mu=1}^{P}\tau_\mu\sum_{i=1}^{N}\xi_i^\mu\sigma_i$ weighs the cost of every RBM configuration, $\mathrm{P}\left(\sigma\right)$ and $\mathrm{P}\left(\tau\right)$ are priors on the visible and hidden layers, respectively, and $Z_\beta\left(\xi\right) = \mathbb{E}_{\sigma,\tau}\left[\exp\left(-\beta H\left[\sigma,\tau;\xi\right]\right)\right]$ is the partition function normalizing the distribution, with $\mathbb{E}_{\sigma,\tau}$ the joint expectation over the visible and hidden unit priors. Intuitively, the priors are the default distributions of σ and τ when the Hamiltonian does not contribute to $\mathrm{P}_\beta\left(\sigma,\tau \mid \xi\right)$, i.e. when β is zero. Gibbs ditributions of the form (3.1) have been deeply investigated in the mathematical physics community for their link with the Parisi theory of spin glasses [102, 152, 104, 105, 153, 154]. An RBM with a given ξ can generate data σ by sampling the marginal distribution

$$\mathrm{P}_\beta\left(\sigma \mid \xi\right) = Z_\beta\left(\xi\right)^{-1}\mathrm{P}\left(\sigma\right)\psi_\beta\left(\sigma;\xi\right) = Z_\beta\left(\xi\right)^{-1}\mathrm{P}(\sigma)\prod_{\mu=1}^{P}\phi_\beta(\xi^\mu\cdot\sigma), \tag{3.2}$$

where $\phi_\beta\left(\xi^\mu\cdot\sigma\right) = \mathbb{E}_{\tau_\mu}\left[\exp\left(\frac{\beta}{\sqrt{N}}\tau_\mu\sum_{i=1}^{N}\xi_i^\mu\sigma_i\right)\right]$, $\psi_\beta\left(\sigma;\xi\right)$ factorizes as $\psi_\beta\left(\sigma;\xi\right) = \prod_{\mu=1}^{P}\phi_\beta(\xi^\mu\cdot\sigma)$ and $Z_\beta\left(\xi\right) = \mathbb{E}_\sigma\left[\psi_\beta\left(\sigma;\xi\right)\right]$. Marginal Gibbs distributions of the form (3.2) are also known as generalized Hopfield networks [39, 101, 23, 94, 97, 96, 98]. Conversely, following Bayes' theorem, the weights of an RBM can be trained on a dataset $\boldsymbol{\sigma} = \{\sigma_i^a\}_{1\leq i\leq N}^{1\leq a\leq M}$ of M examples by sampling the posterior distribution

$$\mathrm{P}_\beta\left(\xi|\boldsymbol{\sigma}\right) = \mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)^{-1}\mathrm{P}\left(\xi\right)\prod_{a=1}^{M}\mathrm{P}_\beta\left(\sigma^a|\xi\right), \tag{3.3}$$

where $\mathrm{P}\left(\xi\right)$ is a prior on the weights and $\mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right) = \mathbb{E}_\xi\left[\prod_{a=1}^{M}\mathrm{P}_\beta\left(\sigma^a|\xi\right)\right]$ is the posterior partition function normalizing the distribution. In the teacher-student setting, the data $\boldsymbol{\sigma}$ used to train the RBM is produced by another RBM [145, 146, 39, 101, 147, 42, 40, 44]. In other words, a *student* RBM is trained using a dataset supplied by a *teacher* RBM. The student's ability to fit the teacher's data can then be evaluated in terms of the so-called overlaps $Q(\xi^{*\mu},\xi^\nu) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{*\mu}\xi_i^\nu$ between the teacher's weights $\xi^* = \left\{\xi_i^{*\mu}\right\}_{1\leq i\leq N}^{1\leq\mu\leq P^*}$ and the student's weights $\xi = \{\xi_i^\nu\}_{1\leq i\leq N}^{1\leq\mu\leq P}$, whose rows $\xi^{*\mu} = \left\{\xi_i^{*\mu}\right\}_{i=1}^{N}$ and $\xi^\mu = \{\xi_i^\nu\}_{i=1}^{N}$ are also called patterns. As such, we refer to the expected value of $Q(\xi^{*\mu},\xi^\nu)$ as the student's *performance*. This framework was used to predict the critical amount of data needed to train multiple variants of RBMs [145, 146, 39, 101, 147, 42, 44].

In [39], it was observed empirically and conjectured that the performance of a student RBM is independent of the number of hidden units when the teacher patterns are uncorrelated. It was later shown analytically that a

student RBM taught by a teacher with uncorrelated patterns achieves the same performance whether the teacher and the student have 1 or 2 hidden units each [42]. In a nutshell, this simplification occurs because an RBM with 2 uncorrelated hidden units effectively factorizes into 2 RBMs with 1 hidden unit each. The conjecture of [39] was explicitly rephrased in terms of this factorization property in [42][1]. Beyond this idealized scenario, it has long been known that machine learning models benefit from the correlations found in structured data to build an efficient internal representation [155]. This phenomenon has received considerable attention in recent theoretical studies [31, 32, 156, 16, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166], notably for RBMs with 2 hidden units in the teacher-student setting [42]. Extending the latter study to an arbitrary number of hidden units remains an intriguing open problem. In fact, although increasing the number of hidden units in the teacher model may bring us closer to the complexity of real data—complexity that RBMs are known to capture due to universal approximation theorems [143, 144]—there is still no theoretical framework that fully explains the learning performance of RBMs for an arbitrary number of hidden units.

In this work, we evaluate the student's learning performance in the teacher-student setting where both the teacher and the student have an arbitrary finite number of hidden units and the teacher patterns are allowed to be correlated with one another. In particular, we evaluate the critical data load $\alpha_{\text{crit}} = \frac{M}{N}$ above which learning becomes possible. In Section 3.2, we introduce the teacher-student setting and the replica method used in our calculations. In Section 3.3.1, we present the so-called *saddle-point* equations governing the performance and the critical load that we obtain from it. In Section 3.3.2, we discuss the case where there are no correlations between the teacher patterns. This Section is divided into three Subsections. In Subsection 3.3.2.1 and 3.3.2.2, we show that the student's performance is independent of the number of hidden units when the teacher patterns are uncorrelated. In Subsection 3.3.2.3, we argue that our teacher-student setting with uncorrelated teacher patterns can serve as a toy model of the lottery ticket hypothesis. Next, in Section 3.3.3, we discuss the effects of uniform teacher pattern correlations on the student's performance. In particular, in Subsection 3.3.3.1, we discuss the different learning phases of the teacher-student problem as a function of the correlations and the number of hidden units. Finally, in Section 3.3.4, we discuss random correlations and compare their effect on the performance to that of uniform correlations. Throughout the paper, we compare key results against Monte Carlo simulations. The code and hyperparameter values of the training algorithms used to make the figures are available at the following public Github repository [167].

3.2 Model

In the teacher-student setting, a student RBM with marginal likelihood $\mathrm{P}_{\beta}\left(\sigma|\xi\right)$ (Eq. 3.2) is trained using a dataset $\boldsymbol{\sigma} = \{\sigma_i^a\}_{1\leq i\leq N}^{1\leq a\leq M}$ of M examples $\sigma^a = \{\sigma_i^a\}_{i=1}^N$ of dimension N, generated by a teacher RBM with a prescribed marginal likelihood $\mathrm{P}_{\beta^*}\left(\sigma|\xi^*\right)$. We call $\alpha = \frac{M}{N}$ the ratio between the size of the training set and the input dimension. In this scenario, the student knows that the correct model for the data is an RBM. However, it does not necessarily know the number of hidden units P^* and the inverse temperature β^* used by the teacher. Therefore, unless explicitly stated otherwise, we will assume that the number of hidden units P and the inverse inference temperature β of the student are not necessarily the same as the teacher's. For convenience, we will frequently state our results in terms of the temperatures $T^* = 1/\beta^*$ and $T = 1/\beta$ rather

[1] See the end of Appendix C of the cited paper.

than in terms of β^* and β.

For simplicity, we assume that the visible and hidden units of both the student and the teacher take values in $\{-1, +1\}$ with a uniform prior, i.e. they are binary random variables with no prior bias towards -1 or $+1$. We impose structure in the data by taking the teacher patterns to be random variables with a fixed covariance matrix $\mathcal{Q}$. To be more precise, we assume that the columns $\xi_i^* = \left\{\xi_i^{*\mu}\right\}_{\mu=1}^{P^*}$ of ξ^* are i.i.d. random variables, with mean 0 and a well-defined $P^* \times P^*$ covariance matrix $\mathcal{Q}$. Uncorrelated teacher patterns are obtained by setting $\mathcal{Q} = \mathbf{I}$.

As previously mentioned, the student learns its patterns ξ by sampling them from the posterior distribution

$$\begin{aligned}
\mathrm{P}_\beta\left(\xi|\boldsymbol{\sigma}\right) &= \mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)^{-1}\mathrm{P}\left(\xi\right)\prod_{a=1}^{M}\mathrm{P}_\beta\left(\sigma^a|\xi\right) = \mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)^{-1}\mathrm{P}\left(\xi\right)\prod_{a=1}^{M}\left[Z_\beta\left(\xi\right)^{-1}\psi_\beta\left(\sigma^a;\xi\right)\right] \\
&= \mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)^{-1}\mathrm{P}\left(\xi\right)\prod_{a=1}^{M}\left[Z_\beta\left(\xi\right)^{-1}\mathrm{P}(\sigma^a)\prod_{\mu=1}^{P}\phi_\beta\left(\sigma^a\cdot\xi^\mu\right)\right] \\
&= \mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)^{-1}\mathrm{P}\left(\sigma\right)Z_\beta\left(\xi\right)^{-M}\prod_{\mu=1}^{P}\psi_\beta(\xi^\mu;\sigma). \qquad (3.4)
\end{aligned}$$

Compared with Eq. (3.2), the posterior distribution is composed of P generalized Hopfield distributions, one for each hidden unit. The data now plays the role of dual patterns, which interact only through the term $Z_\beta\left(\xi\right)^{-M}$. The latter is therefore responsible for encoding mutual correlation between student patterns, otherwise independent [39, 40, 43]. The student has no access to the structure of the teacher patterns, so we assume an identity covariance matrix for the student pattern prior $\mathrm{P}(\xi)$.

In general, the partition functions $Z_\beta\left(\xi\right)$ and $\mathcal{Z}_\beta\left(\boldsymbol{\sigma}\right)$ are intractable, which makes the properties of Eq. (3.4) difficult to study. However, our assumptions about the priors on the visible units, hidden units and patterns make analytical computations possible in the limit of $N, M \to \infty$.

The overlaps $Q(\xi^{*\mu}, \xi^\nu) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{*\mu}\xi_i^\nu$ are a good measure of the student's learning performance because they quantify how close the student patterns are to the teacher patterns. We exploit techniques from statistical mechanics to compute the expected value $\mathbb{E}_{\xi^*,\boldsymbol{\sigma},\xi}\left[Q(\xi^{*\mu},\xi^\nu)\right]$ of the overlaps with respect to the distribution of the teacher patterns, the generated dataset and the inferred student patterns. Specifically, in the limit of large dataset and data dimension $M, N \to \infty$, we obtain the expected value of the overlaps as a byproduct of the limiting quenched free entropy

$$f(\alpha, \beta^*, \beta, P^*, P, \mathcal{Q}) = \lim_{M,N\to\infty}\frac{1}{N}\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}\log\left[\mathcal{Z}_\beta(\boldsymbol{\sigma})\right], \qquad (3.5)$$

where $\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}$ is the expected value w.r.t. the joint distribution of the teacher patterns and generated dataset. In fact, we show in Section 3.3.1 that Eq. (3.5) can be expressed as the result of a variational principle w.r.t. a set of order parameters, whose optimum gives the expected value of the overlaps. We then use this result to investigate the effects of P and β on the student's ability to learn a dataset characterized by P^*, β^* and $\mathcal{Q}$, as well as the impact of these hyperparameters on the critical threshold α_{crit} beyond which learning becomes possible. We focus our quantitative discussion on Gaussian and binary priors $\mathrm{P}(\xi), \mathrm{P}(\xi^*)$ (see 3.A.1), but our results can be generalized to other pattern distributions that meet the other assumptions stated earlier in this Section. The Gaussian case is particularly interesting because it is closely related to RBMs used in practical

applications [141, 69, 142], which usually have continuous weights.

3.3 Results and discussion

3.3.1 Free entropy and saddle point equations

The free entropy f can be computed by exploiting a well-established statistical mechanics technique called the replica trick, which is based on the identity

$$f = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\log\left[\mathcal{Z}\right] = \lim_{N\to\infty, L\to 0}\left(\frac{1}{LN}\log\mathbb{E}\left[\mathcal{Z}^L\right]\right).$$

Calculations are shown in the Appendices. In 3.B, we calculate the average replicated partition function $\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}\left[\mathcal{Z}^L\right]$ in the limit of $N\to\infty$. In 3.C, we use $\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}\left[\mathcal{Z}^L\right]$ to evaluate the quenched free entropy f under the so-called replica-symmetric (RS) approximation. We find that the RS free entropy can be expressed in terms of the variational principle

$$\begin{aligned} f = \underset{m,\hat{m},q,\hat{q},s,\hat{s}}{\text{Extr}} \Bigg\{ & -\sum_{\mu=1}^{P^*}\sum_{\nu=1}^{P} m^{\mu\nu}\hat{m}^{\mu\nu} - \frac{1}{2}\sum_{\mu\neq\nu}^{P} s^{\mu\nu}\hat{s}^{\mu\nu} + \frac{1}{2}\sum_{\mu,\nu=1}^{P} q^{\mu\nu}\hat{q}^{\mu\nu} \\ & + \mathbb{E}_{\xi^*}\mathbb{E}_z \log\left[\mathcal{Z}\left(\mathcal{L}^C\right)\right] + \alpha\left\langle \mathbb{E}_z \log\left[\mathcal{Z}\left(\mathcal{L}^O\right)\right]\right\rangle_{\mathcal{M}_*} - \alpha\log\left[\mathcal{Z}\left(\mathcal{M}\right)\right] \Bigg\}, \end{aligned} \tag{3.6}$$

where $z = [z_{\mu\nu}]_{\mu,\nu=1}^{P}$ is a set of i.i.d. standard Gaussian random variables, and where the effective energies $\mathcal{M}_*$, $\mathcal{M}$, $\mathcal{L}^O$ and $\mathcal{L}^C$ are defined as

$$\mathcal{M}_*\left(\tau_*\right) = \frac{1}{2}\left[\beta^*\right]^2 \sum_{\mu,\nu=1}^{P^*} \mathcal{Q}_{\mu\nu}\tau_\mu^*\tau_\nu^* \tag{3.7}$$

$$\mathcal{M}\left(\tau\right) = \frac{1}{2}\beta^2 \sum_{\mu,\nu=1}^{P} s^{\mu\nu}\tau_\mu\tau_\nu \tag{3.8}$$

$$\mathcal{L}^O\left(\tau;\tau^*,z\right) = \mathcal{L}_{\beta^*,\beta}\left(\tau,\tau^*,z;m,s,q\right) \tag{3.9}$$

$$\mathcal{L}^C\left(\xi;\xi^*,z\right) = \mathcal{L}_{1,1}\left(\xi,\xi^*,z;\hat{m},\hat{s},\hat{q}\right) \tag{3.10}$$

$$\text{with}\quad \mathcal{L}_{\lambda_1,\lambda_2}\left(\xi,\xi^*,z;m,s,q\right) = \frac{1}{2}\left[\lambda_2\right]^2 \sum_{\mu,\nu=1}^{P}\left(s^{\mu\nu}-q^{\mu\nu}\right)\xi^\mu\xi^\nu \tag{3.11}$$

$$+\lambda_1\lambda_2 \sum_{\mu=1}^{P^*}\sum_{\nu=1}^{P} m^{\mu\nu}\xi^{*\mu}\xi^\nu \tag{3.12}$$

$$+\lambda_2 \sum_{\mu,\nu=1}^{P} A_{\mu\nu}\left(q\right) z_{\mu\nu}\frac{\xi^\mu+\xi^\nu}{2}$$

$$\text{and}\quad A_{\mu\nu}\left(q\right) = \sqrt{2q^{\mu\nu} - \delta_{\mu\nu}\sum_{\eta=1}^{P} q^{\mu\eta}}. \tag{3.13}$$

In particular, we use the notation $\mathcal{Z}(f) = \mathbb{E}_x\left[\exp\left\{f\left(x;\cdot\right)\right\}\right]$ for the partition function and $\langle g\rangle_f = \mathbb{E}_x\left\{g(x)\mathcal{P}\left[f\right]\left(x;\cdot\right)\right\}$ for the expectation value of an observable g w.r.t. the Gibbs distribution $\mathcal{P}\left[f\right]\left(x;\cdot\right) = \mathcal{Z}(f)^{-1}\exp\left[f\left(x;\cdot\right)\right]$ of an effective energy f, where $\mathbb{E}_x$ is the expectation over the prior of x. As explained in 3.A.3, $q^{\mu\nu}$ is assumed symmetric in Eq. (3.11). The solution of this variational principle (see 3.D for a detailed derivation) must obey the saddle-point equations

$$\begin{aligned}
m^{\mu\nu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\xi^{*\mu}\left\langle\xi^\nu\right\rangle_{\mathcal{L}^O}\right] \\
s^{\mu\nu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\left\langle\xi^\mu\xi^\nu\right\rangle_{\mathcal{L}^O}\right] \\
q^{\mu\nu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\left\langle\xi^\mu\right\rangle_{\mathcal{L}^O}\left\langle\xi^\nu\right\rangle_{\mathcal{L}^O}\right], \\
\hat{m}^{\mu\nu} &= \beta^*\beta\alpha\,\left\langle\mathbb{E}_z\left[\tau_\mu^*\left\langle\tau_\nu\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*} \\
\hat{s}^{\mu\nu} &= \beta^2\alpha\left(\left\langle\mathbb{E}_z\left[\left\langle\tau_\mu\tau_\nu\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*} - \left\langle\tau_\mu\tau_\nu\right\rangle_{\mathcal{M}}\right) \\
\hat{q}^{\mu\nu} &= \beta^2\alpha\left\langle\mathbb{E}_z\left[\left\langle\tau_\mu\right\rangle_{\mathcal{L}^O}\left\langle\tau_\nu\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*}.
\end{aligned} \tag{3.14}$$

The optimal order parameters $m \in \mathbb{R}^{P^*\times P}$, $s \in \mathbb{R}^{P\times P}$ and $q \in \mathbb{R}^{P\times P}$ solving Eqs. (3.14) are to be interpreted as expected overlaps. First of all, $m^{\mu\nu}$, which is commonly known as the Mattis magnetization, is the limiting expected overlap between the teacher pattern $\xi^{*\mu}$ and the student pattern ξ^ν, i.e.

$$m^{\mu\nu} = \lim_{N,M\to\infty}\mathbb{E}_{\xi^*,\boldsymbol{\sigma},\xi}\left[Q(\xi^{*\mu},\xi^\nu)\right] \quad \mu = 1,\dots,P^*\ \nu = 1,\dots,P, \tag{3.15}$$

where $\mathbb{E}_{\xi^*,\boldsymbol{\sigma},\xi}$ is the expectation w.r.t. the joint distribution $\mathrm{P}(\xi^*)\,\mathrm{P}_{\beta^*}(\boldsymbol{\sigma}|\xi^*)\,\mathrm{P}_\beta(\xi|\boldsymbol{\sigma})$ of teacher patterns, dataset and student patterns. Second of all, $s^{\mu\nu}$ with $\mu\neq\nu$ is the limiting expected overlap between any two student patterns ξ^μ and ξ^ν from the same sample $\xi\sim\mathrm{P}_\beta\left(\xi|\boldsymbol{\sigma}\right)$, i.e.

$$s^{\mu\nu} = \lim_{N,M\to\infty}\mathbb{E}_{\xi^*,\boldsymbol{\sigma},\xi}\left[Q(\xi^\mu,\xi^\nu)\right] \quad \mu,\nu = 1,\dots,P. \tag{3.16}$$

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

$$\begin{aligned}
q^{\mu\nu} &= \lim_{N,M\to\infty}\mathbb{E}_{\xi^*,\boldsymbol{\sigma},\xi^1\times\xi^2}\left[Q(\xi^{1\mu},\xi^{2\nu})\right] \\
&= \lim_{N,M\to\infty}\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}\left[Q(\mathbb{E}_{\xi|\boldsymbol{\sigma}}\left[\xi^\mu\right],\mathbb{E}_{\xi|\boldsymbol{\sigma}}\left[\xi^\nu\right])\right] \quad \mu,\nu = 1,\dots,P,
\end{aligned} \tag{3.17}$$

where $\mathbb{E}_{\xi|\boldsymbol{\sigma}}$ indicates the expectation w.r.t. the posterior distribution (Eq. 3.4). While s measures the effective correlation between student patterns, the so-called spin-glass order parameter q quantifies the tendency of the student to stay frozen in specific configurations rather than sampling all possible patterns uniformly. For simplicity's sake, we will usually call m the magnetization and q the spin-glass (SG) overlap. As in the Introduction, we will also occasionally refer to m as the student's performance.

The RS saddle point equations (Eqs. 3.14) can be solved by numerical iteration for any values of the hyperparameters β^*, β, α, P^* and P (see 3.I). 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. $\beta=\beta^*$, $P=P^*$ and $\mathrm{P}\left(\xi\right)=\mathrm{P}\left(\xi^*\right)$. This regime of complete information is commonly referred to as the Nishimori line [73,

76, 74, 75]. When $\beta = \beta^*$, we find two different phases (see Figs. 3.2, 3.3, 3.10 and 3.16) :

- the *paramagnetic* (P) phase, where the order parameters m, s and q all vanish;
- the *ferromagnetic* (F) phase, where they are all larger than zero.

Intuitively, the student RBM can partially learn the teacher patterns in the F phase but becomes unable to do so in the P phase. The corresponding P-F phase transition is thus also the onset of learning. When $\beta \neq \beta^*$, we also find

- the *spin-glass* (SG) phase, where $m = 0$ but $q > 0$.

In this phase, the student converges to spurious low-energy states ($q > 0$) unrelated to the teacher patterns ($m = 0$) (see Figs. 3.11 and 3.12). Looking at Figs. (3.2), (3.3), (3.10) and (3.16), the P-F phase transition appears to be second order. In general, such second-order phase transitions coincide with the onset of instability of the paramagnetic solution in the saddle-point equations (Eqs. 3.14). The instability condition (see 3.F) reads as

$$\alpha \geq \alpha_{\text{crit}} = \frac{1}{\left[\beta^*\beta\right]^2 \lambda_{\text{max}}^{\mathcal{S}}}, \tag{3.18}$$

where $\lambda_{\text{max}}^{\mathcal{S}}$ is the largest eigenvalue of the matrix $\mathcal{S} = \mathcal{Q}\mathcal{R}$, with the covariance matrix $\mathcal{R}$ defined as $\mathcal{R}_{\mu\nu} = \left\langle \tau^{*\mu}\tau^{*\nu} \right\rangle_{\mathcal{M}_*}$. As expected, the α_{crit} of Eq. (3.18) coincides with the P-F phase transition of Figs. (3.10) and (3.16). Interestingly, α_{crit} does not depend on the number of hidden units P of the student nor on its prior $\mathrm{P}(\xi)$. Outside the Nishimori line, i.e. when $(\beta, P, P(\xi)) \neq (\beta^*, P^*, P(\xi^*))$, the RS approximation is not expected to be exact. Therefore, in principle, one would need to calculate the so-called replica symmetry breaking (RSB) corrections of α_{crit}. However, we find that Eq. (3.18) is consistent with Monte Carlo simulations even when $P \neq P^*$ (see Fig. 3.5) and $\mathrm{P}(\xi) \neq \mathrm{P}(\xi^*)$ (see Fig. 3.14). As such, the RS approximation of α_{crit} seems robust to the priors and number of hidden units in practice.

3.3.2 Learning uncorrelated patterns

In this Section, including Subsections 3.3.2.1, 3.3.2.2 and 3.3.2.3, we take $\mathcal{Q} = \mathbf{I}$, i.e. there are no correlations between the teacher patterns. In that regime, we find that the teacher-student setting exhibits an effective factorization property on the student patterns, which makes the behavior of the student RBM explainable in terms of that with a single hidden unit. A first indication of this result is given by the critical load α_{crit} (see Eq. 3.18). As we explained in the previous Section, no matter $\mathcal{Q}$, Eq. (3.18) does not depend on the number P of student patterns. When $\mathcal{Q} = \mathbf{I}$, and thus $\mathcal{S} = \mathbf{I}$ (see Eqs. 3.18), we find that

$$\alpha_{\text{crit}} = \frac{1}{\left[\beta^*\beta\right]^2}, \tag{3.19}$$

which does not depend on the number P^* of teacher patterns either. Eq. (3.19) generalizes to arbitrary finite P^* and P the critical load $\alpha_{\text{crit}} = \beta^{-4}$ previously found for $P^* = P = 1$ hidden units [145, 146] and $P^* = P = 2$ uncorrelated hidden units [42] when $\beta^* = \beta$. As we show in the next Subsections, the previous universality result also holds for the order parameters, which we recall are a good measure of the student's learning performance.

3.3.2.1 Independence of the number of hidden units: binary patterns

In the case of binary patterns, Eqs. (3.14) with $P = P^* = 1$ simplify to

$$\begin{aligned} m &= \mathbb{E}_z \left[\tanh\left(\hat{m} + \sqrt{\hat{q}} z \right) \right] \\ q &= \mathbb{E}_z \left[\tanh^2\left(\hat{m} + \sqrt{\hat{q}} z \right) \right] \\ \hat{m} &= \beta^* \beta \alpha \, \mathbb{E}_z \left[\tanh\left(\beta^* \beta m + \beta \sqrt{q} z \right) \right] \\ \hat{q} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2\left(\beta^2 m + \beta \sqrt{q} z \right) \right], \end{aligned} \tag{3.20}$$

whose solution is summarized in the phase diagrams of Fig. (3.1). To be more precise, the left and right panels of Fig. (3.1) show the fixed-β^* regime and the Nishimori regime $\beta = \beta^*$, respectively. For clarity's sake, we will discuss Fig. (3.1) and the solution of Eq. (3.20) in terms of $T^* = 1/\beta^*$ and $T = 1/\beta$ for the remainder of this paragraph. We recall that the student can learn the teacher patterns in the ferromagnetic (F) phase ($m \neq 0$), but not in the paramagnetic (P) phase ($m = q = 0$) nor in the spins-glass (SG) phase ($m = 0$ and $q > 0$). For $T < T^*$, we find the SG phase between the P phase and the F phase. The corresponding P-SG transition line $\alpha_{\text{P-SG}} = T^4$ extends until the Nishimori line $T = T^*$, where it meets the F phase. Therefore, the Nishimori line crosses the triple point of the P, SG and F phases, as expected from spin glass theory [73, 74]. On the Nishimori line, it is straightforward to verify that $m = q$, which is also expected from spin glass theory [73, 74]. In particular, m and q simultaneously become non-zero above the critical load $\alpha_{\text{crit}} = [T^* T]^2 = T^4$ of the P-F phase transition on the Nishimori line, which prevents the SG phase from forming (see Fig. 3.1, right panel). As in some related inference problems in the teacher-student setting, $\alpha_{\text{P-SG}} = T^4$ is identical to the α_{crit} found on the Nishimori line [44, 43, 45]. For $T > T^*$, the SG phase does not exist either, and the P phase transitions directly to the F phase at $\alpha_{\text{crit}} = [T^* T]^2$.

Looking at the SG-F transition, we see that decreasing the inference temperature T too much can make it harder for the student to learn the teacher patterns. In particular, a student with enough data to learn the teacher patterns at a given inference temperature T_1 may fail to do so at a lower inference temperature $T_2 < T_1$. The RS approximation is probably not completely accurate in this regime, so one would need to calculate RSB corrections to study this behavior quantitatively.

When P, P^* are arbitrary finite numbers and $\mathcal{Q} = \mathbf{I}$, we search for solutions of Eqs. (3.14) by making the ansatz

$$\begin{aligned} m^{\mu\nu} &= \delta_{\mu\nu} m, & \hat{m}^{\mu\nu} &= \delta_{\mu\nu} \hat{m}, \\ s^{\mu\nu} &= \delta_{\mu\nu}, & \hat{s}^{\mu\nu} &= 0, \\ q^{\mu\nu} &= \begin{cases} \delta_{\mu\nu} q & \mu, \nu \leq P^* \\ \delta_{\mu\nu} g & \text{otherwise}, \end{cases} & \hat{q}^{\mu\nu} &= \begin{cases} \delta_{\mu\nu} \hat{q} & \mu, \nu \leq P^* \\ \delta_{\mu\nu} \hat{g} & \text{otherwise}, \end{cases} \end{aligned} \tag{3.21}$$

which describes the situation where the student learns the teacher patterns one-to-one. Following the nomenclature introduced in [42], we will call Eq. (3.21) the permutation symmetry breaking (PSB) ansatz. We have assumed without loss of generality that the $\min\{P, P^*\}$ non-zero magnetizations are on the main diagonal, as any hidden unit permutation would give an equivalent solution. According to this ansatz, the first $\min\{P, P^*\}$ student patterns converge one-to-one to the first $\min\{P, P^*\}$ teacher patterns with magnetization m and SG overlap q. When $P > P^*$, the remaining $P - P^*$ student patterns (i.e. $P \geq \nu > P^*$) are aligned in spurious

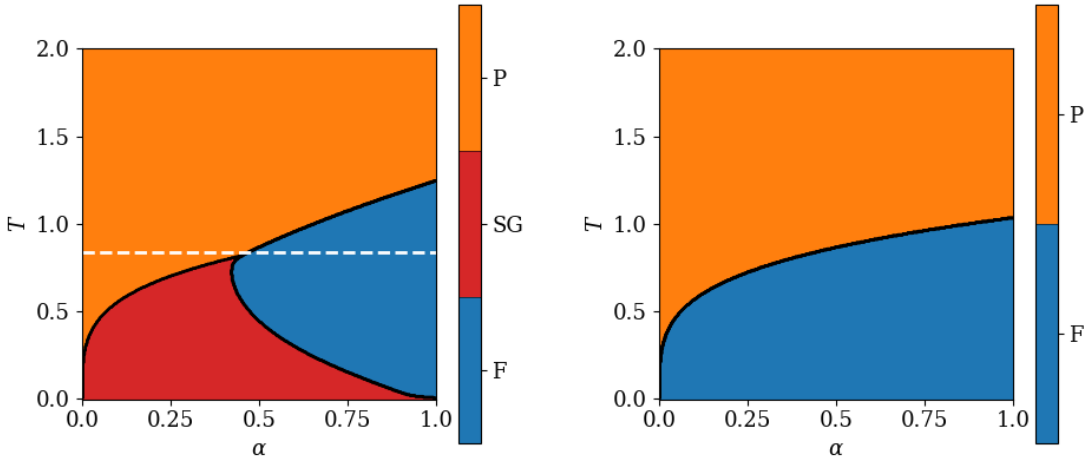

Figure 3.1: RS phase diagrams of the teacher-student setting with $P = P^* = 1$ obtained by solving Eqs. (3.20). The left diagrams has $\beta^* = 1.2$, and the right one, $\beta^* = \beta = 1/T$. The student is unable to learn the teacher patterns in the paramagnetic phase (P) and in the spin-glass phase (SG), but it is able to do so in the ferromagnetic phase (F). The white dashed line in the left plot is the Nishimori line $\beta^* = \beta$. The magnetization m and SG overlap q solving Eqs. (3.20) along this line are plotted on Fig. (3.2).

directions (i.e. $m^{\mu\nu} = 0 \; \forall \mu = 1, ..., P^*$) with a different SG overlap $g \neq q$. Conversely, when $P < P^*$, $P^* - P$ of the teacher patterns are not learned for lack of student hidden units. The latter case is described in terms of only m and q, and so is $P = P^*$. Under the PSB ansatz, Eqs. (3.14) decouple into P independent systems of equations for the P hidden units $\nu = 1, ..., P$ (see 3.G). In other words, the student factorizes into P students with one hidden unit each. The systems of equations for the first $\min\{P, P^*\}$ patterns are all identical to each other and equivalent to Eqs. (3.20). When $P > P^*$, the $P - P^*$ remaining systems of equations all take the form

$$\begin{aligned} g &= \mathbb{E}_z \left[\tanh^2 \left(\sqrt{\hat{g}} z \right) \right] \\ \hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta \sqrt{g} z \right) \right] . \end{aligned} \tag{3.22}$$

In summary, Eqs. (3.14) with binary patterns and $\mathcal{Q} = \mathbf{I}$ reduce to Eqs. (3.20, 3.22) under the PSB ansatz. In particular, the solution of Eqs. (3.20, 3.22), which we will call the PSB solution, has $m = q$ even in the region outside the Nishimori line where $\beta = \beta^*$ and $P \neq P^*$. In Fig. (3.2), we verify that the PSB solution can be found by iterating the original saddle-point equations (Eqs. 3.14) with binary student patterns (see 3.A.1), $\mathcal{Q} = \mathbf{I}$ and near-diagonal initial conditions $m^{\mu\nu} = \delta_{\mu\nu} m_0 + (1 - \delta_{\mu\nu}) \, \varepsilon$, where $\varepsilon \ll m_0$. In other words, we verify that the solution of Eqs. (3.20, 3.22) is stable for the full saddle-point iteration. We plot only the case of $P \geq P^*$ because $P < P^*$ yields similar results.

The original saddle-point equations (Eq. 3.14) with binary patterns, $\mathcal{Q} = \mathbf{I}$ and $P > P^*$ also have other solutions that cannot be expressed as cleanly by a set of simplified saddle-point equations (see Fig. 3.3). These solutions can be found by initializing $m^{\mu\nu}$ with $m_0 \gg \varepsilon$ on the diagonal and at least one off-diagonal

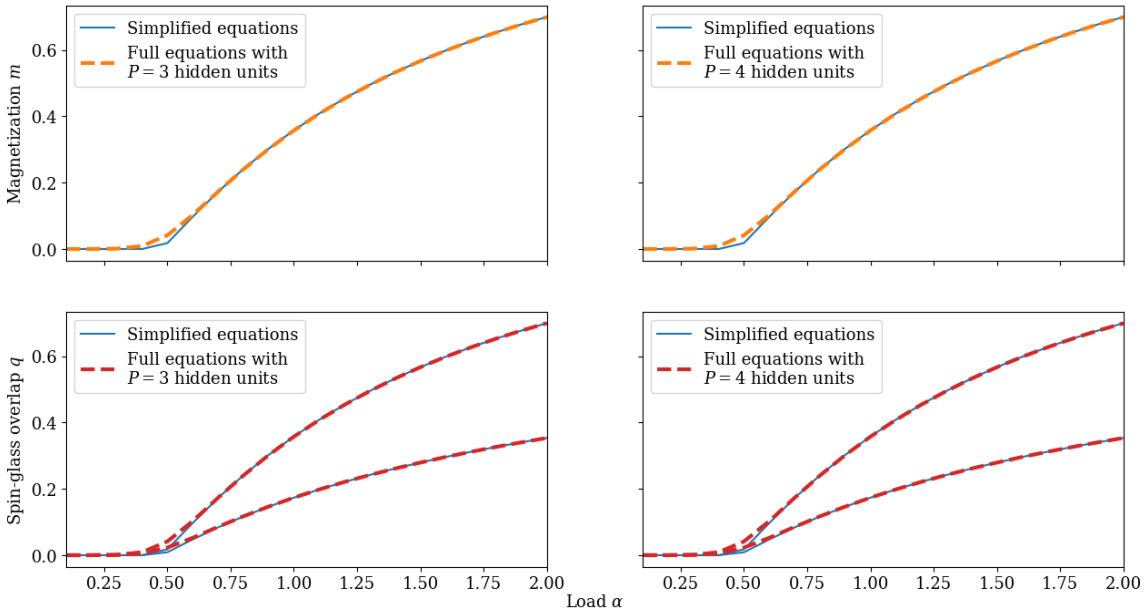

Figure 3.2: Permutation symmetry breaking (PSB) solution of Eqs. (3.14) for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q} = \mathbf{I}$, in red and orange, compared against the solution of Eqs. (3.20, 3.22), in blue. We plot the Mattis magnetization m in the top row, and the spin-glass (SG) overlap q in the bottom row. The magnetization plots and the top lines of the SG overlap plots show that the student patterns that converge to teacher patterns have the same m and q as the solution of Eqs. (3.20), and thus also satisfy $m = q$. Conversely, the bottom lines of the SG overlap plots show that the student patterns that do not converge to a teacher pattern have the same SG overlap g as the solution of Eqs. (3.22). We use $P = 3$ and $P^* = 2$ in the left column and $P = 4$ and $P^* = 3$ in the right column. All plots have $\beta^* = \beta = 1.2$.

entry close to m_0, a process we will refer to as off-diagonal initialization. They correspond to a distributed representation where some of the student patterns learn the same teacher pattern, so we dub these solutions partial PSB. Throughout this paper and its figures, we focus on the case where at most two student patterns learn the same teacher pattern, but larger numbers are also possible. Within a partial PSB solution, some student patterns may still learn teacher patterns one-to-one. These student patterns ξ^μ_{PSB} have the same m and q as Eq. (3.20), and in particular satisfy $m = q$ (see Fig. 3.3). Comparison of the RS free entropy (Eq. 3.6) of the PSB and partial PSB solutions of Eqs. (3.14) suggests that the former is always favored in the limit $N, M \to \infty$. The free entropy difference between them decreases with increasing $T = T^*$ and vanishes at the onset of the paramagnetic phase (see Fig. 3.17). These results are confirmed by numerical simulations. In low $T = T^*$ simulations, students with binary patterns always converge to the PSB solution, even when we use random initial conditions (see Figs. 3.5), which is solid evidence that PSB is favored at low $T = T^*$. When $T = T^*$ is relatively close to the paramagnetic phase, the simulations have relatively large error bars and repeatedly jump between the different solutions rather than converging to a single mode. This kind of instability can be attributed to finite-size fluctuations.

The PSB solution is independent of P^* and P, which means that, in terms of the student's performance m, learning P^* i.i.d. patterns stored in a single teacher RBM is as difficult as learning P^* i.i.d. patterns from

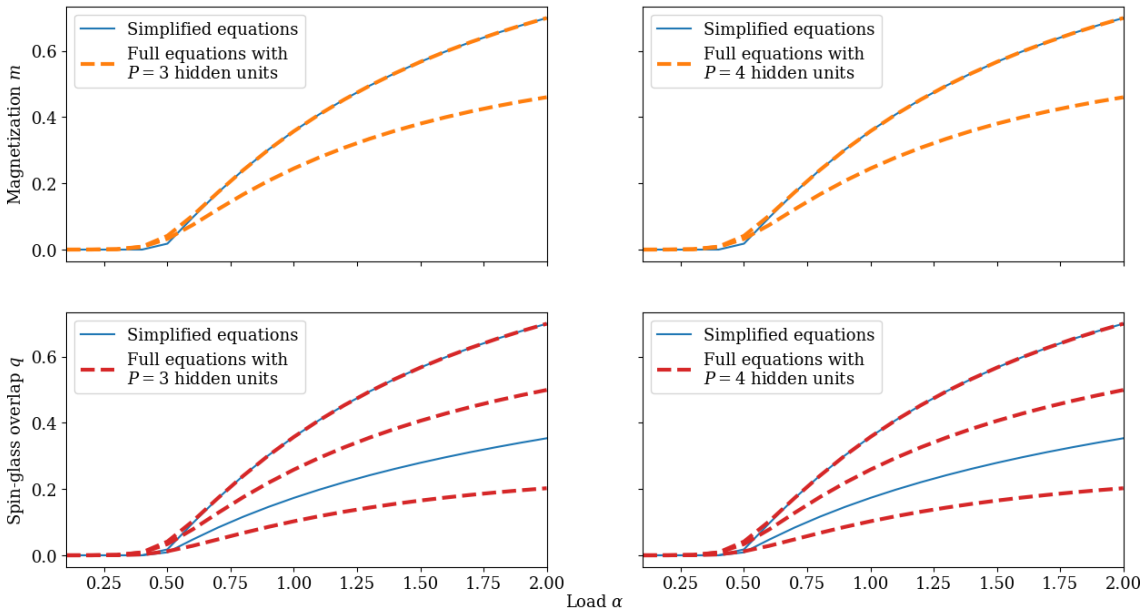

Figure 3.3: Partial permutation symmetry breaking (partial PSB) solutions of Eqs. (3.14) for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q} = \mathbf{I}$, in red and orange, compared against the solution of Eqs. (3.20, 3.22), in blue. We plot the Mattis magnetization m in the top row, and the SG overlap q in the bottom row. The top lines of the plots show that the student patterns ξ^{μ}_{PSB} that converge to teacher patterns one-to-one have the same m and q as the solution of Eqs. (3.20), and thus also satisfy $m = q$. Conversely, the other lines show that the student patterns ξ^{μ}_{PS} that converge to a common teacher pattern have a smaller m and a different q. To be more precise, the central and bottom branches of q are the spin-glass order parameters corresponding to $Q(\xi^{1\mu}_{PS}, \xi^{2\mu}_{PS})$ and $Q(\xi^{1\mu}_{PS}, \xi^{2\nu}_{PS})$ with $\mu \neq \nu$, respectively (see Section 3.2). They are both different from the g of Eq. (3.22). The Mattis magnetization and SG overlaps omitted from this Figure all vanish. We use $P = 3$ and $P^* = 2$ in the left column and $P = 4$ and $P^* = 3$ in the right column. All plots have $\beta^* = \beta = 1.2$.

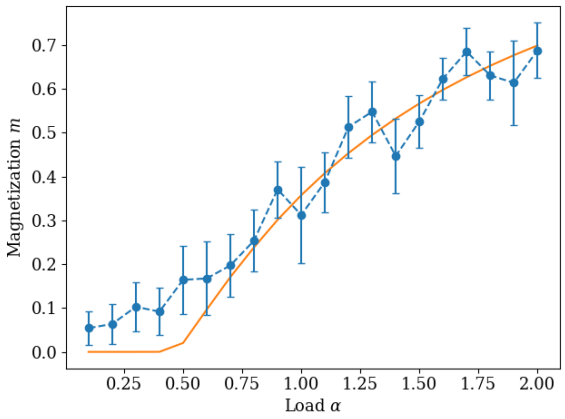

Figure 3.4: The magnetization m solving Eqs. (3.22), in orange, compared against $N = 512$ dimensional Monte Carlo simulations, in blue, of the teacher-student problem where the student has $P = 2$ binary patterns with a uniform prior and the teacher has $P^* = P = 2$ binary patterns with covariance $\mathcal{Q} = \mathbf{I}$. The blue dots and error bars represent the means and standard deviations, respectively, of the diagonal of the magnetization m during the simulations. The inverse temperature is set to $\beta^* = \beta = 1.2$, and the simulations have a small external field biasing the student towards the PSB solution.

P^* separate RBMs with one hidden unit each. Such independence of the student's performance from the number of hidden units was first conjectured in [39] based on empirical observations. The performance on the Nishimori line was then shown to be the same for $P = P^* = 2$ [42] as for $P = P^* = 1$ [145, 146]. Our PSB solution extends this result to $\beta = \beta^*$ and any finite P, P^*. On the Nishimori line, i.e. when $\beta = \beta^*$ and $P = P^*$, the teacher-student setting is replica symmetric, so we expect the PSB solution to be exact, which is confirmed by simulations (see Fig. 3.4). When $\beta = \beta^*$ and $P \neq P^*$, replica symmetry is not guaranteed, but the PSB solution is still in good agreement with simulations (see Fig. 3.5). Finally, we expect RSB corrections when $\beta \neq \beta^*$. Interestingly, we observe a weaker form of independence from the number of hidden units in the suboptimal partial PSB solutions, in the sense that the partial PSB solutions that we find at given P^* and P are also present at larger P^* and P. These results seem related to the embedding principle stating that a neural network contains all minima of narrower neural networks with the same architecture [168].

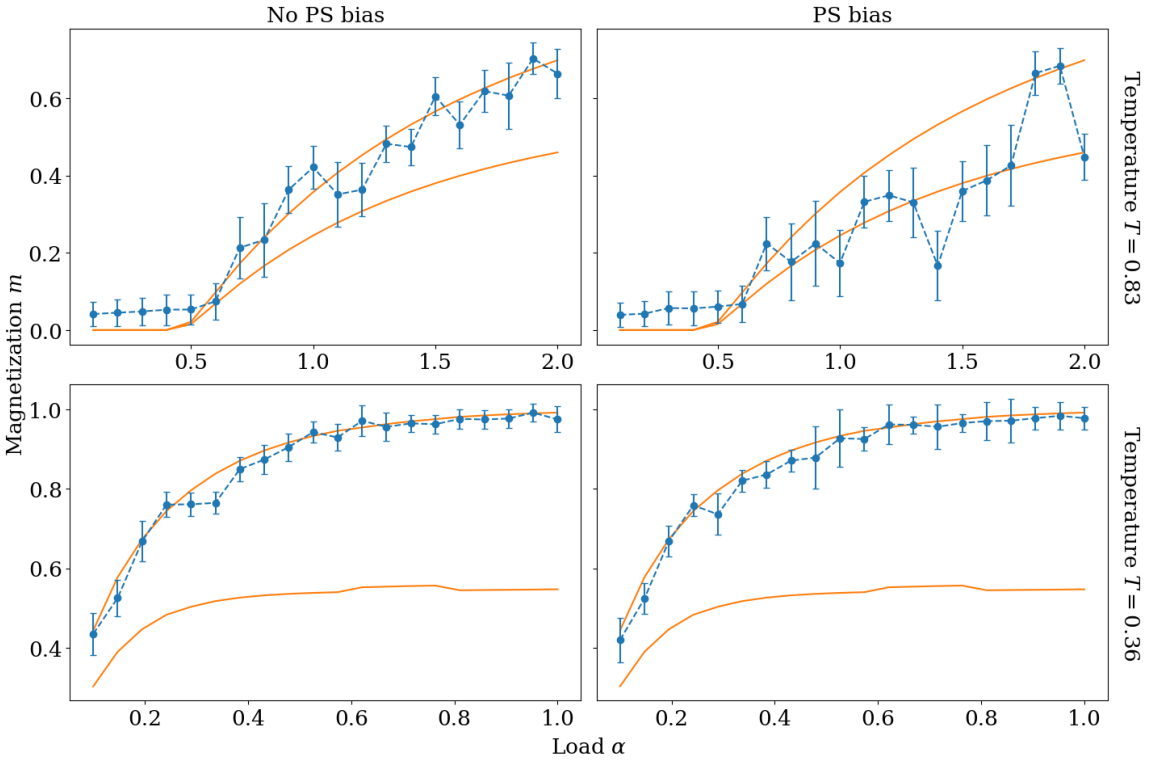

Figure 3.5: Solutions of Eqs. (3.14) shown in Fig. (3.3), in orange, compared against $N = 512$ dimensional Monte Carlo simulations, in blue, of the teacher-student problem where the student has $P = 2$ binary patterns with a uniform prior and the teacher has a single binary pattern. The blue dots and error bars represent the means and standard deviations, respectively, of a single diagonal coefficient of m during the simulations. The inverse temperature is set to $\beta^* = \beta = 1.2 \approx 1/0.83$ and $\beta^* = \beta = 2.8 \approx 1/0.36$ in the top and bottom rows, respectively. In the left column, the student with $T^* = T = 0.83$ is biased towards the PSB solution by near-diagonal initial conditions, while the student with $T^* = T = 0.36$ has random initial conditions. The students of the right column have off-diagonal initial conditions. The top left and top right students are also biased by a small external field pointing towards the PSB and partial PSB solutions, respectively. The bottom students are not biased by an external field.

3.3.2.2 Independence of the number of hidden units: Gaussian patterns

When the student patterns ξ are real-valued variables with an i.i.d. standard Gaussian prior, Eqs. (3.14) under the PSB ansatz (Eqs. 3.21) simplify to (see 3.G)

$$\begin{aligned} m &= \frac{\hat{m}}{1+\hat{q}} \\ q &= \frac{\hat{m}^2}{(1+\hat{q})^2} + \frac{\hat{q}}{(1+\hat{q})^2} \\ g &= \frac{\hat{g}}{(1+\hat{g})^2} \\ \hat{m} &= \beta^* \beta \alpha \, \mathbb{E}_z \left[\tanh\left(\beta^* \beta m + \beta \sqrt{q} z\right) \right] \\ \hat{q} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2\left(\beta^2 m + \beta \sqrt{q} z\right) \right] \\ \hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2\left(\beta \sqrt{g} z\right) \right]. \end{aligned} \tag{3.23}$$

This regime is interesting because RBMs used in practical applications usually have continous weights [141, 69, 142]. Similarly as before, we compare Eqs. (3.23) against the PSB and partial PSB solutions of Eqs. (3.14) for real-valued student patterns with an i.i.d. standard Gaussian prior and teacher pattern covariance $\mathcal{Q} = \mathbf{I}$. The resulting plots are qualitatively similar to Figs. (3.2) and (3.3), so we show them in 3.J rather than in the main text (see Figs. 3.18 and 3.19). As before, we observe that the solutions of the Gaussian equations are independent of P^* and P. One important difference between the student model with binary patterns and the student model with real-valued patterns is that they are simulated differently. In fact, binary patterns are obtained using the standard random walk Metropolis-Hastings algorithm, while real-valued patterns are learned via underdamped stochastic Langevin dynamics [169, 170] of the RBM marginal likelihood (Eq. 3.2) gradient. In a nutshell, the latter algorithm is a variant of stochastic gradient ascent where noise is added to the momentum vector to sample the RBM posterior (Eq. 3.3) rather than to find one of its modes. We use contrastive divergence, i.e. alternate Gibbs sampling of the visible and hidden units, to evaluate the marginal likelihood [24, 171]. At low temperature, simulations of the student model with real-valued patterns usually converge to the solution of Eqs. (3.23) (see Fig. 3.6). They can also sometimes stay stuck near the partial PSB solution without converging to it when the learning rate is reduced too aggressively during training. This behavior is shown in Fig. (3.6) by the simulation data points that are scattered around the partial PSB solution but do not agree with it within their error bars. In fact, although simulation error is useful for measuring the convergence of the Langevin learning algorithm, its sensitivity to the learning rate schedule makes it poorly representative of the equilibrium distribution of the magnetization m.

3.3.2.3 A simple model of the lottery ticket hypothesis

Many works studying feedforward neural networks found a training regime where some of the hidden units learn the underlying data distribution while the others take a back seat [172, 173, 174, 168]. Interestingly, this behavior is similar to our PSB solution with $P > P^*$, where only a subset of the student patterns learn the teacher patterns (see Sections 3.3.2.1 and 3.3.2.2). As proposed in [168], we investigate the relationship between this kind of training regime and the lottery ticket hypothesis [9].

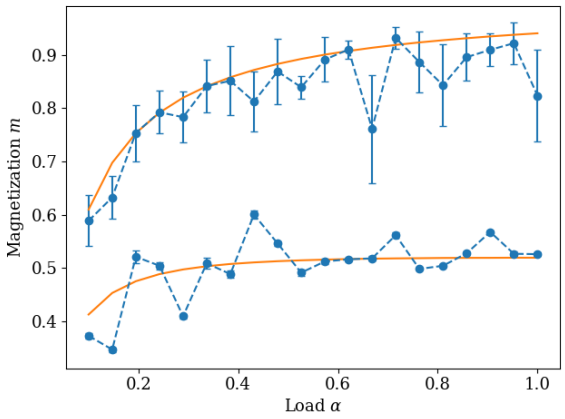

Figure 3.6: Solutions of Eqs. (3.14) for real-valued student patterns with a standard Gaussian prior and teacher pattern covariance $\mathcal{Q} = \mathbf{I}$, in orange, compared against $N = 512$ dimensional Monte Carlo simulations, in blue, of the teacher-student problem where the student has $P = 2$ real-valued patterns with a standard Gaussian prior and the teacher has a single Gaussian pattern. The top orange line is the magnetization m of student patterns that converge to teacher patterns one-to-one, so it is also the m solving Eqs. (3.23). On the other hand, the bottom orange line is the m of student patterns that converge to a common teacher pattern in the partial PSB solution. The blue dots and error bars represent the means and standard deviations, respectively, of the largest entry of the magnetization m during the simulations. The top and bottom blue lines result from simulations with i.i.d. standard Gaussian initial conditions and off-diagonal initial conditions, respectively. The learning rate was reduced more quickly for the bottom line than for the top line. The inverse temperature is set to $\beta^* = \beta = 4$.

The lottery ticket hypothesis states that a generic randomly initialized overparameterized neural network contains subnetworks that fit data with similar accuracy as the entire trained network when they are extracted from it and trained independently [9, 148, 149, 150, 151]. These special subnetworks, which are commonly referred to as winning lottery tickets, are thought to have fortuitous initial conditions that facilitate training [9]. The PSB solution studied in Sections 3.3.2.1 and 3.3.2.2 makes it clear that any student network of size $P > P^*$ contains subnetworks of size $\tilde{P} \in \{P^*, ..., P-1\}$ that learn the teacher patterns to the same extent, and thus fit the data at least as well. It is not obvious, however, whether any of these subnetworks can have lucky initial conditions such that they train more easily than with i.i.d. random initial conditions. As such, we apply to our teacher-student setting a variant of the magnitude pruning algorithm traditionally used to identify winning tickets [9, 151] and check if it finds a subnetwork that converges especially quickly. Consider one teacher and three students with real-valued patterns, hereby referred to as *the teacher*, *student 0*, *student A* and *student B*, respectively, where A is a control network and B is a candidate winning ticket obtained from 0. We perform the following numerical experiment:

- Initialize student 0 with $P = 8$ i.i.d. Gaussian patterns and the teacher with $P^* = 4$ i.i.d. Gaussian patterns. Save the initial value of the patterns of student 0 as ξ_0.
- Train student 0 on data generated by the teacher. Compute the Euclidean norms of the learned patterns.
- Initialize student A with $P = 4$ i.i.d. Gaussian patterns and student B with the $P = 4$ patterns of ξ_0 that evolved to have the largest Euclidean norm when trained.

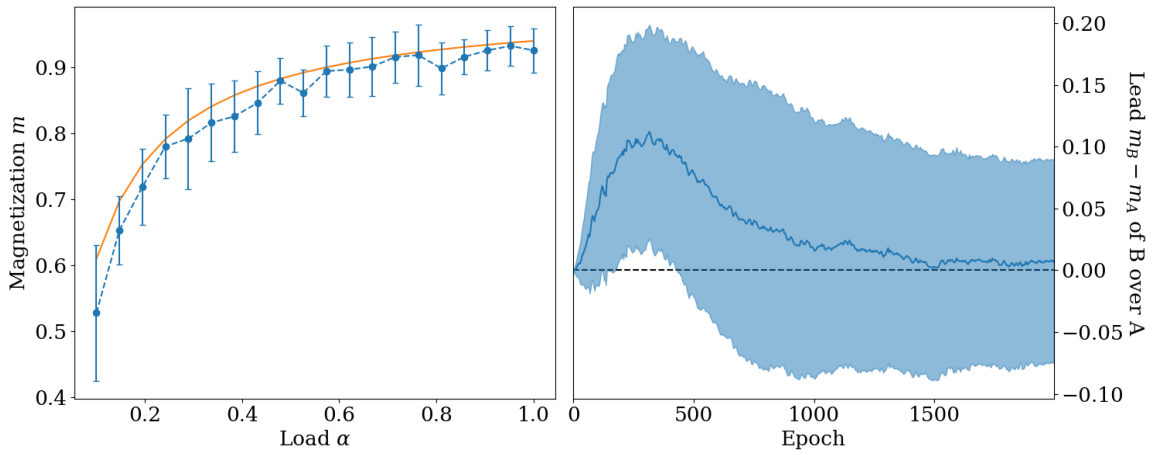

Figure 3.7: Results of the lottery ticket experiment described in Section 3.3.2.3. In the left panel, $N = 512$ dimensional Monte Carlo simulations of student B, in blue, are compared against the solution of Eqs. (3.23), in orange. The blue dots and error bars represent the means and standard deviations, respectively, of the diagonal of the magnetization m during the simulations. The right panel shows the difference $m_B - m_A$ of the magnetizations of A and B as a function of the simulation epochs. The solid blue line and the shaded region represent the median of $m_A - m_B$ over $\alpha \in [0, 1]$ and the corresponding mean absolute deviation around the median, respectively. $m_A - m_B$ goes to zero when the number of elapsed epochs is large, so student A converges to the solution of Eqs. (3.23) like student B. The inverse temperature is set to $\beta^* = \beta = 4$.

- Train student A and student B on data generated by the teacher. Record their respective magnetizations m_A and m_B as a function of the training epochs in order to compare their convergence speeds.

This procedure is slightly different from the usual magnitude pruning algorithm in that it prunes the patterns with the smallest norms rather than the entries of the weight matrix with the smallest absolute values [9, 151]. As in Section 3.3.2.2, we train students 0, A and B with underdamped stochastic Langevin dynamics [169, 170] of the RBM marginal likelihood (Eq. 3.2). Our results are shown in Fig. (3.7). In the left panel, we verify that student B converges to the PSB solution (i.e. the solution of Eqs. 3.23) in the range of α that we are studying. As in Section 3.3.2.2, the error in m is useful for measuring the convergence of the learning algorithm, but poorly representative of the equilibrium distribution. There seems to be a small systematic shift in the simulations at low α, which could be due to finite-size effects or order $\frac{P}{M}$ corrections [39]. In the right panel, we plot the median of $m_B - m_A$ over α as a function of the epochs. B initially converges more quickly than A, and the lead of B eventually shrinks to zero because of the ergodicity of the training algorithm. In other words, magnitude pruning is able to identify a winning ticket that converges more quickly than networks with i.i.d. random initial conditions. The patterns of student 0 that converge to the teacher patterns the fastest have a small initial magnetization (see Fig. 3.7, epoch 0). Therefore, the shape of the loss function at the initial conditions and the basin of attraction in which they are located may be more important than the initial conditions themselves, as suggested by [9].

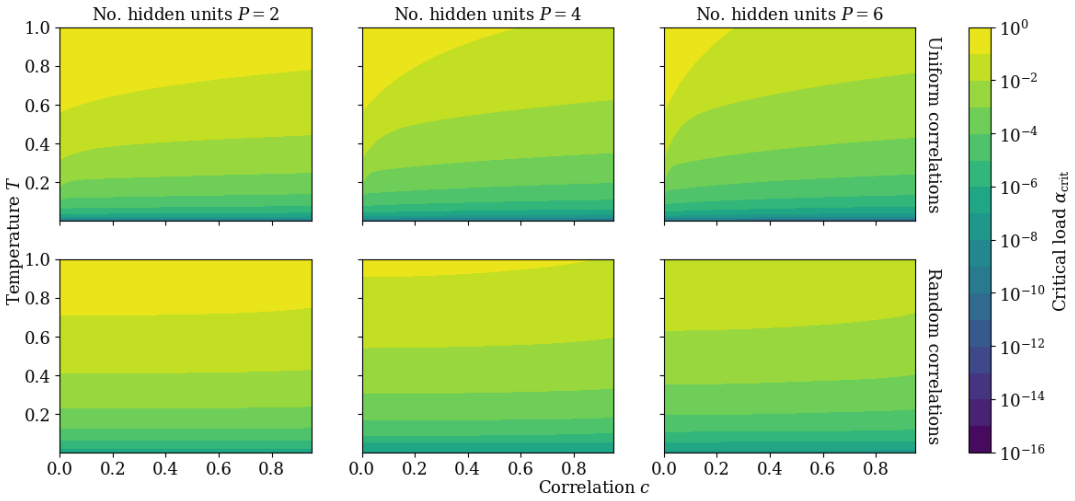

Figure 3.8: Critical load α_{crit} for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the temperature T and the correlation c. α_{crit} is obtained from Eq. (3.18). The top row has $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, so the max eigenvalue $\lambda^{\mathcal{S}}_{\text{max}}$ is that of Eq. (3.24). The bottom row is the arithmetic mean $\overline{\alpha_{\text{crit}}}$ over correlation matrices $\mathcal{Q}$ sampled from the projected Wishart distribution $\mathcal{W}(c, P)$ defined in 3.A.2.

3.3.3 Learning uniformly correlated patterns

In this Section, including Subsection 3.3.3.1, we introduce uniform correlations in the teacher patterns by fixing the covariance matrix $\mathcal{Q}$ of ξ^* to $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, where $c \in [0, 1)$ controls the correlation strength. In the presence of uniform correlation, the Hamiltonian $\mathcal{M}_*$ (Eq. 3.7) is that of the Curie-Weiss model with coupling constant $\frac{1}{2}\left[\beta^*\right]^2 c$. Therefore, its correlation matrix $\mathcal{R}$ has the same form as $\mathcal{Q}$ but with a different off-diagonal element d. In 3.H, we show that the maximum eigenvalue $\lambda^{\mathcal{S}}_{\text{max}}$ of Eq. (3.18) is

$$\lambda^{\mathcal{S}}_{\text{max}} = \left(P^* - 1\right)^2 cd + \left(P^* - 1\right)(c + d) + 1. \tag{3.24}$$

As expected, the corresponding critical load is also the onset of non-zero Mattis magnetization in the regime of $\beta = \beta^*$ where only the paramagnetic and ferromagnetic phases exist, which we show explicitly in Fig. (3.10) for binary patterns. We show in 3.J that the same holds for real-valued student patterns with a standard Gaussian prior (see Fig. 3.20). Eq. (3.24) extends to arbitrary finite P^* the critical load obtained for $P^* = 2$ in [42]. In fact, when $P^* = 2$, we find $d = \tanh\left(\left[\beta^*\right]^2 c\right)$ and Eq. (3.18) reduces to the critical load of [42]. Plotting Eq. (3.24) (see Fig. 3.9) and the corresponding α_{crit} (Fig. 3.8) as a function of T, c and P^*, we see that α_{crit} decreases with P^* and c. For $c \gg \left[T^*\right]^2$, the spins of Curie-Weiss Hamiltonian $\mathcal{M}_*$ all align, so their correlation is $d = 1$, and we obtain

$$\lambda^{\mathcal{S}}_{\text{max}} = \left(\left(P^* - 1\right) c + 1\right) P^*.$$

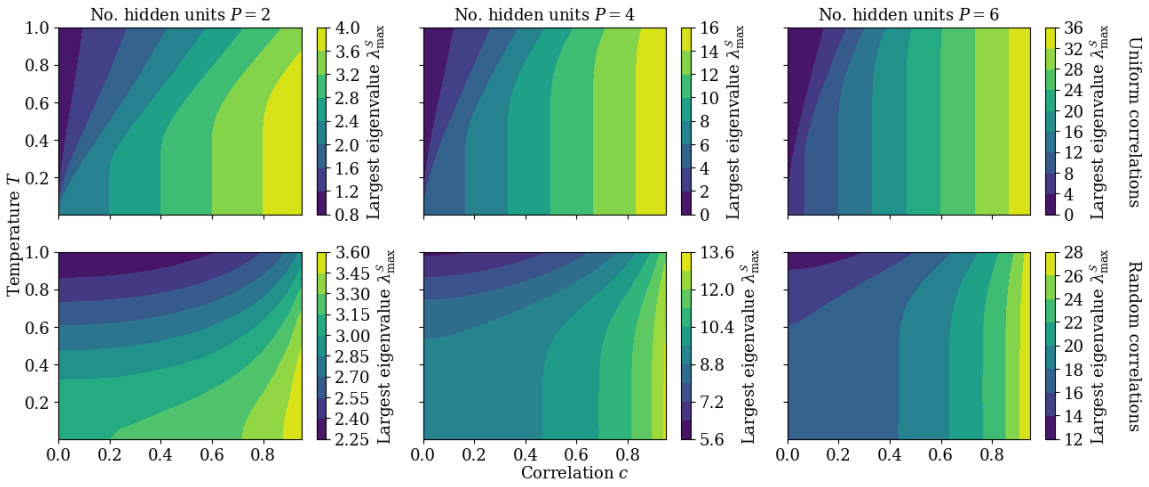

Figure 3.9: Largest eigenvalue $\lambda^{\mathcal{S}}_{\text{max}}$ of $\mathcal{S} = \mathcal{Q}\mathcal{R}$ (see 3.F) for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the temperature T and the correlation c. The top row has $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, so the max eigenvalue $\lambda^{\mathcal{S}}_{\text{max}}$ is that of Eq. (3.24). The bottom row is the harmonic mean $\left[\overline{1/\lambda^{\mathcal{S}}_{\text{max}}}\right]^{-1}$ over correlation matrices $\mathcal{Q}$ sampled from the projected Wishart distribution $\mathcal{W}(c, P)$ defined in 3.A.2.

In this limit, α_{crit} is roughly inversely proportional to both c and $[P^*]^2$. Conversely, for $c \ll [T^*]^2$, the correlation of $\mathcal{M}_*$ is $d = 0$, and we obtain

$$\lambda^{\mathcal{S}}_{\text{max}} = (P^* - 1)\, c + 1. \tag{3.25}$$

In this limit, α_{crit} decreases less quickly by a factor of P^* than for $c \gg [T^*]^2$. However, relatively small correlations can still significantly decrease α_{crit} when P^* is large. In other words, the critical load benefits from structured data with many correlated underlying abstract concepts even when the correlation between the different concepts is rather small. Overall, these results shed light upon the way α_{crit} depends on P^*, which was previously unclear given that previous work [42] focused on $P = P^* = 2$.

Neural networks are often regularized by feature decorrelation techniques that reduce or eliminate correlations in their inputs and hidden layers [175, 176, 177, 178]. Decorrelation is known to increase training speed [176, 177, 178, 179]. However, our work suggests that it can also increase the critical load as a drawback, which may hinder performance when the data is noisy (i.e. β^* finite). [180, 181] observed that batch normalization [176], which was motivated by decorrelation [177, 8], makes neural networks less robust to noise. This effect could potentially be related to our findings. Arguably, the main caveat to our analysis is that we assumed that P^* was finite when deriving Eq. (3.24), so the critical load is probably different when P^* is $\mathcal{O}(N)$ [39].

Figs. (3.11) and (3.12) display the magnetization m and the SG overlap q, respectively, found by solving Eqs. (3.14) at fixed β^*. As usual, we find the paramagnetic (P) phase ($m = q = 0$) at small α and low T. At low c and P, the spin glass (SG) phase ($m = 0$ and $q > 0$) occupies the medium α and low T region of the α, T plane. It transitions to the ferromagnetic (F) phase ($m \neq 0$) as α, c and P increase. At high c, P and α, the SG overlap q seems to be non-monotonic in T. As expected, the critical load of Eq. (3.18)

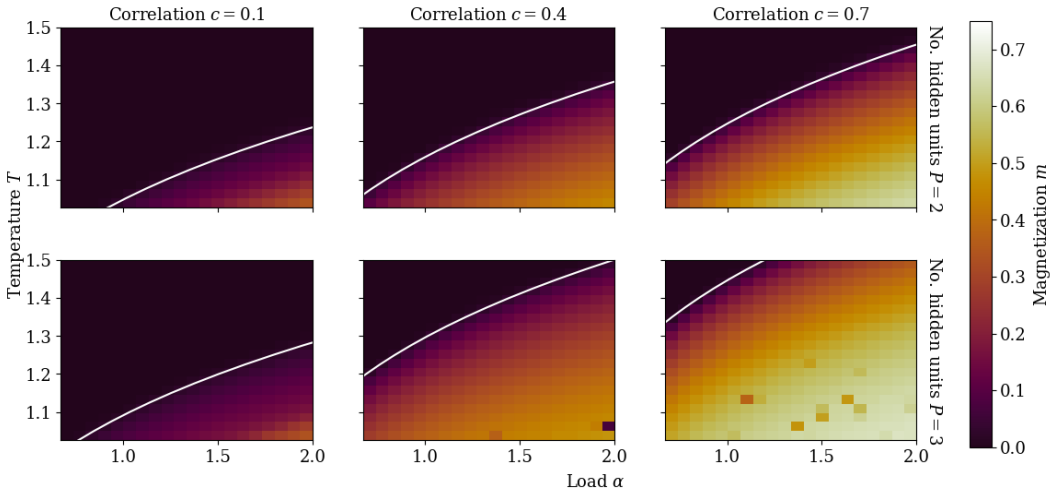

Figure 3.10: Mattis magnetization m for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. m is obtained by solving Eqs. (3.14) numerically for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, where $c \in [0, 1)$ (see 3.H). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the phase transition described by Eq. (3.18) with $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24). The speckles in the plots with $P = 3$, $c = 0.4$ and $P = 3$, $c = 0.7$ are due to the occasional numerical instability of the saddle-point equations (see 3.I).

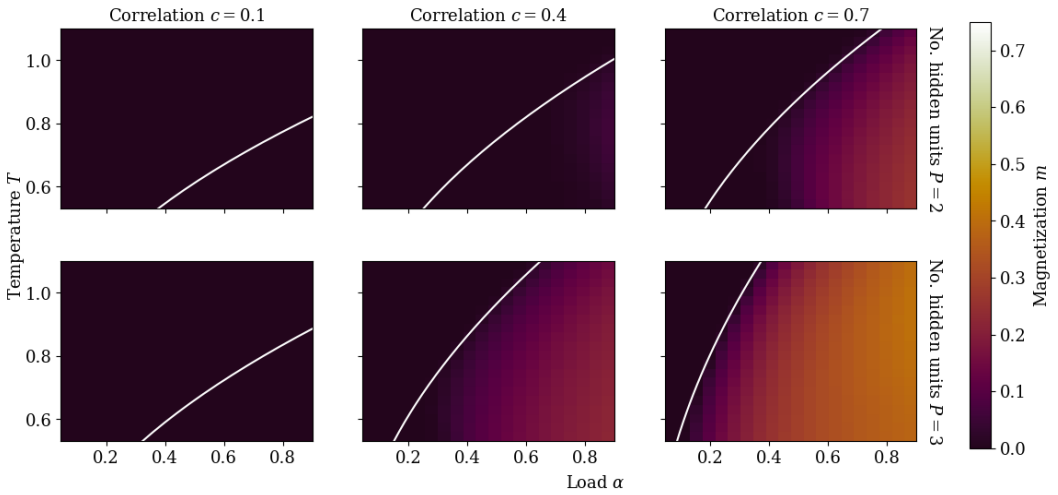

Figure 3.11: Mattis magnetization m for $\beta^* = 0.8$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. m is obtained by solving Eqs. (3.14) numerically for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, where $c \in [0, 1)$ (see 3.H). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the critical load of Eq. (3.18) with $\beta^* = 0.8$ and $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24).

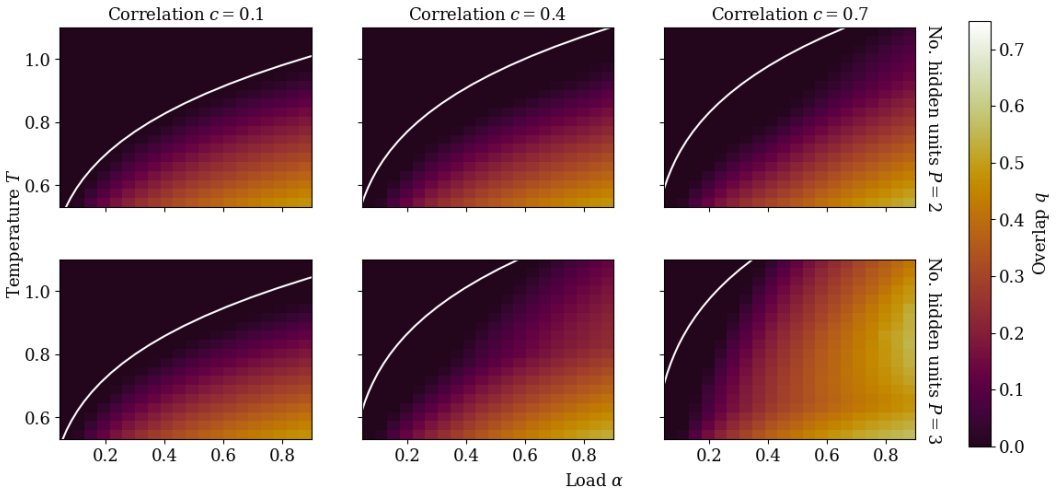

Figure 3.12: SG overlap q for $\beta^* = 0.8$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. q is obtained by solving Eqs. (3.14) numerically for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, where $c \in [0, 1)$ (see 3.H). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the critical load of Eq. (3.18) with $\beta = \beta^*$ and $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24).

with $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24) follows the onset of non-zero magnetization corresponding to the P-F phase transition, but not the SG-F phase transition. The critical load of $\beta = \beta^*$ approximately follows the P-SG phase transition when c is small, which is consistent with Section 3.3.2.1 and previous works [43, 45]. As in Section 3.3.2.1, decreasing the inference temperature T too much can make it harder for the student to learn the teacher patterns (see Fig. 3.11, top right panel).

3.3.3.1 Permutation symmetry breaking transitions

As in Section 3.3.2, the critical load marking the onset of learning (see Eq. 3.18) again does not depend on the number of hidden units P of the student RBM. Despite this, a single wide RBM does not learn teacher patterns as would multiple separate RBMs with one hidden unit each. Moreover, distinct phases emerge based on the level of correlation c and data load α, each characterized by a different learning strategy.

RBMs with $P = P^* = 2$ hidden units learning correlated patterns were previously found to have three distinct ferromagnetic phases [42]:

- in the *spontaneous symmetry breaking* (SSB) phase, both student patterns converge to a single configuration that has the same overlap with the two teacher patterns. In other words, the student learns only the features that the teacher patterns have in common;
- in the *student permutation symmetry breaking* (PSB_s) phase, the student patterns converge to two distinct configurations that have the same overlap with the teacher patterns. In other words, the student is able to learn distinct features of the teacher patterns;
- finally, in the *teacher permutation symmetry breaking* (PSB_t) phase, the student can tell the two teacher patterns apart and learns them one-to-one as in the PSB phase of Section (3.3.2.1).

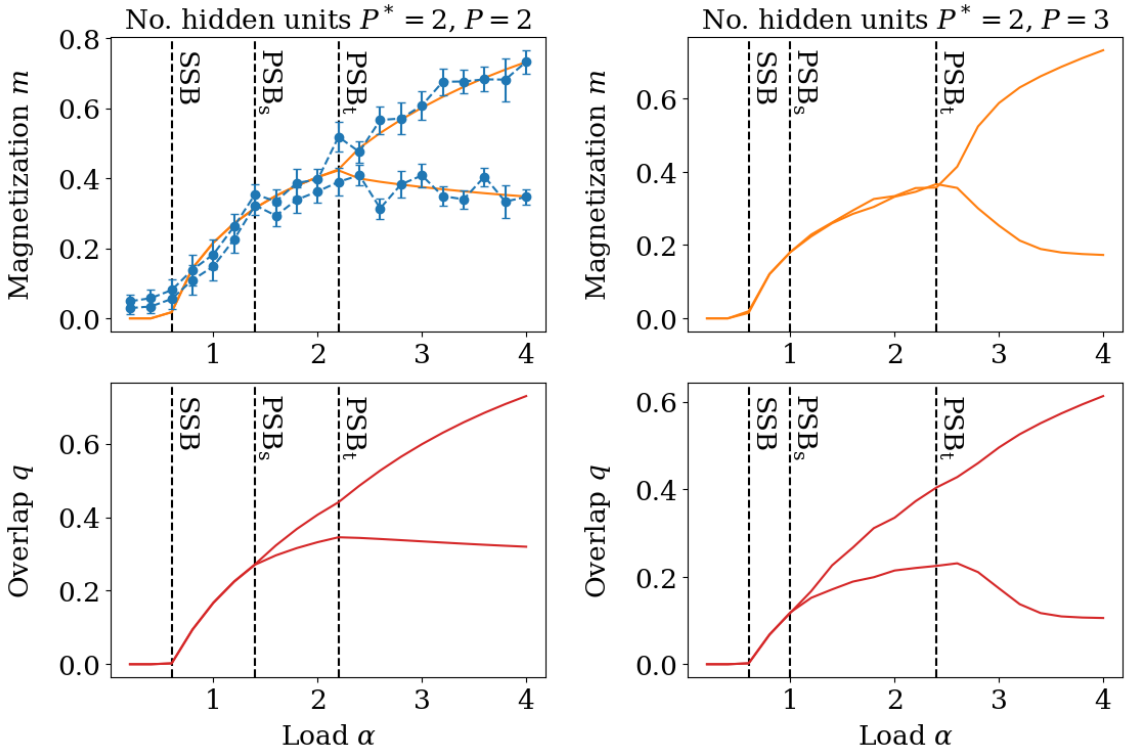

Figure 3.13: Mattis magnetization m and SG overlap q solving Eqs. (3.14), in orange and red, as a function of the load α and number of student patterns P for $\beta = \beta^* = 1$, $P^* = 2$ and $c = 0.3$. The top and bottom branches of the plots are respectively the diagonal and off-diagonal coefficients of m and q. In the top-left panel, m is compared against $N = 512$ dimensional Monte Carlo simulations, in blue. The blue dots and error bars represent the means and standard deviations, respectively, of the diagonal and off-diagonal coefficients of the magnetization m during the simulations. Simulation results do not agree with the predictions of the saddle-point equations shown in the right panel when $\alpha > \alpha_{\text{crit}}$ (see Eq. 3.18). Besides that, they are not very insightful, so we do not plot them for the sake of clarity.

These three phases can be identified by comparing the values of the order parameters $m^{\mu\nu}$ and $q^{\mu\nu}$ on and off the diagonal. As shown in the left panel of Fig. (3.13), bifurcations occur as α increases, marking second-order phase transitions at the critical loads $\alpha_{\text{crit}}^{\text{SSB}} = \alpha_{\text{crit}}$, $\alpha_{\text{crit}}^{\text{PSB}_s} > \alpha_{\text{crit}}^{\text{SSB}}$ and $\alpha_{\text{crit}}^{\text{PSB}_t} > \alpha_{\text{crit}}^{\text{PSB}_s}$. In this context, the student does not know the level of correlation c of the teacher patterns. As such, it uses a uniform prior $\mathrm{P}(\xi) \neq \mathrm{P}(\xi^*)$, and we are always outside the Nishimori regime. Despite this, the RS critical thresholds and magnetizations are in good agreement with Monte Carlo simulations (see Fig. 3.13, left panel).

As shown in Figs. (3.13) and (3.14), the same three phases also appear at larger P and P^*. $\alpha_{\text{crit}}^{\text{SSB}}$ once again coincides with the onset of the ferromagnetic phase, which occurs at the same load α_{crit} regardless of the number of hidden units P of the student. On the other hand, $\alpha_{\text{crit}}^{\text{PSB}_s}$ is smaller for $P = 3$ than for $P = 2$ (see Fig. 3.13). We think that the term $A_{\mu\mu}(q) = \sqrt{2q^{\mu\mu} - \sum_{\eta=1}^{P} q^{\mu\eta}}$ in $\mathcal{L}_{\lambda_1\lambda_2}$ (see Eqs. 3.11 and 3.13) penalizes values of q with a large off-diagonal sum, i.e. $\sum_{\eta\neq\mu}^{P} q^{\mu\eta}$, compared to the diagonal $q^{\mu\mu}$. As the number of terms in the sum grows with P, the individual off-diagonal coefficients $q^{\mu\nu}$ may be encouraged to become smaller, pushing permutation symmetry breaking of the student patterns (PSB$_s$) to occur at a lower

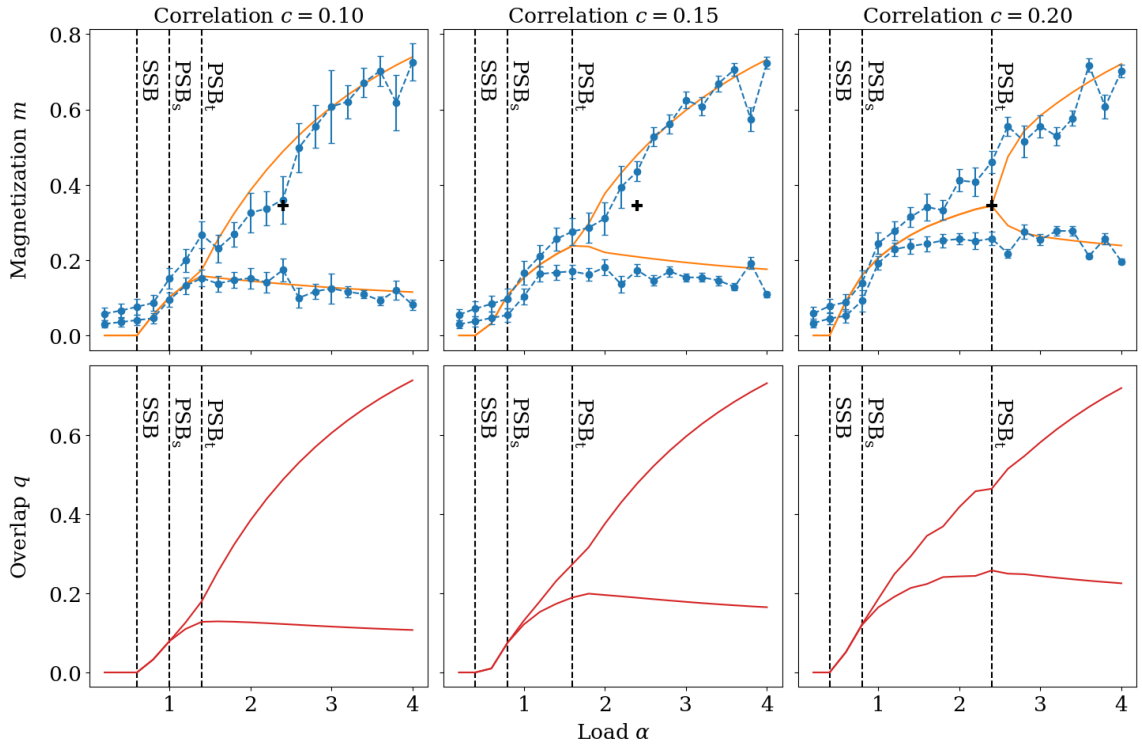

Figure 3.14: Mattis magnetization m and SG overlap q solving Eqs. (3.14), in orange and red, as a function of the load α and teacher pattern correlations c for $\beta = \beta^* = 1$ and $P = P^* = 3$. The top and bottom branches of the plots are respectively the diagonal and off-diagonal coefficients of m and q. m is compared against $N = 512$ dimensional Monte Carlo simulations, in blue. The blue dots and error bars represent the means and standard deviations, respectively, of the diagonal and off-diagonal coefficients of the magnetization m during the simulations. The dashed lines indicate the approximate locations of the SSB, PSB_s and PSB_t phase transitions of Eqs. (3.14). The black crosses in the top plots mark the α and m at which the PSB_t transition occurs in the right plots. It serves as a visual guide to show that the m predicted by Eqs. (3.14) can sometimes decrease significantly with c, but not the m of the simulations.

critical threshold $\alpha_{\text{crit}}^{\text{PSB}_\text{s}}$.

In [42], the critical load $\alpha_{\text{crit}}^{\text{PSB}_\text{t}}$ increases with the correlation strength c. As such, increasing c can trigger a phase transition from PSB_t to PSB_s and thus decrease the magnetization. In other words, large correlations in the teacher patterns can undermine the student's ability to learn them accurately. Based on these findings, Hou et al. [42] formulated the hypothesis that the best c for learning a relatively low load of data is non-zero but still relatively small. In the case of $P, P^* \geq 3$ (see Fig. 3.14), we also find that $\alpha_{\text{crit}}^{\text{PSB}_\text{t}}$ grows with c. However, unlike $P = P^* = 2$, Monte Carlo simulations of $P, P^* \geq 3$ do not completely agree with the RS approximation at large c. In particular, increasing c does not seem to decrease the magnetization significantly in the simulations, even when it does so according to the RS approximation (see Fig. 3.14, black crosses).

Repeating the lottery ticket experiment of Section 3.3.2.3 for $c = 0.05$, we find that the lead of the winning ticket B over its randomly-initialized counterpart A has a smaller maximum (see Fig. 3.15). In other words, small correlations in the data appear to make the student less sensitive to initial conditions, which also reduces the benefits of magnitude pruning (see Section 3.3.2.3). As in Sections 3.3.2.2 and 3.3.2.3, the error in m (see Fig. 3.15, left panel) is useful for measuring the convergence of the learning algorithm, but poorly representative of the equilibrium distribution.

Outside the teacher-student setting, RBMs typically undergo a sequence of second-order phase transitions at the beginning of training [31, 32, 33, 35]. After the first transition, they learn a rank-one matrix, which is reminiscent of our SSB phase. The second transition then breaks the rank-one symmetry like our PSB_s transition. The subsequent phase transitions are harder to interpret from the point of view of permutation symmetry breaking, and it is not clear whether a transition analogous to PSB_t can occur without an explicit teacher. PSB_s phase transitions were also observed in Gaussian mixture models [84, 85, 86, 72] and modern Hopfield networks [34], which suggests that they could be common in machine learning models.

3.3.4 Random correlations

In this section, we take the covariance matrix $\mathcal{Q}$ to be random. By definition, $\mathcal{Q}$ must be positive semi-definite. Moreover, when the teacher weights are binary, $\mathcal{Q}$ must have ones on the diagonal. We sample $\mathcal{Q}$ from the projected Wishart distribution $\mathcal{W}(c, D)$ defined in 3.A.2 because it satisfies these two requirements. By the law of large numbers, $\mathcal{Q} \sim \mathcal{W}(c, D)$ approaches $\mathcal{C} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$ in probability as $D \to \infty$ (see 3.A.2). Therefore, $\lambda_{\text{max}}^{\mathcal{S}}$ and α_{crit} are the same as in Section 3.3.3 when D is large. At finite $D = P$, the dependence of α_{crit} on T is still qualitatively similar to that of uniform correlations. However, the arithmetic mean $\overline{\alpha_{\text{crit}}}$ and the harmonic mean $\left[\overline{1/\lambda_{\text{max}}^{\mathcal{S}}}\right]^{-1}$ over many independent samples $\mathcal{Q} \sim \mathcal{W}(c, P = D)$ are very different from the α_{crit} and $\lambda_{\text{max}}^{\mathcal{S}}$ of uniform correlations as a function of c and P (see Figs. 3.8 and 3.9). For instance, $\overline{\alpha_{\text{crit}}}$ decreases much more slowly with c than the critical load of uniform correlations. At small c and high T, $\overline{\alpha_{\text{crit}}}$ tends to be smaller than the critical load of uniform correlations. Conversely, at large c and low T, the critical load of uniform correlations is usually smaller. At any given c, the entries of a random correlation matrix can sometimes be larger than c, which can make learning possible even when T is too high for correlations of size c or smaller to be picked up on by the student. However, this advantage disappears at larger c and lower T where correlations of size c and smaller are no longer muddled in noise. In summary, the behavior of the critical load as a function of P and c is in general very different for random correlations and for uniform correlations.

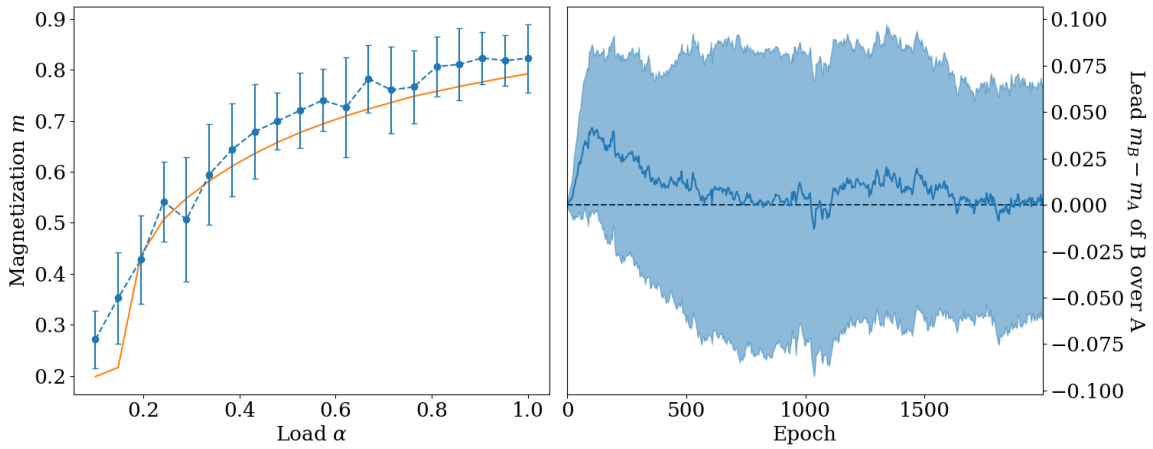

Figure 3.15: Results of the lottery ticket experiment of Section 3.3.2.3 when the teacher patterns have a uniform correlation matrix with $c = 0.05$ instead of being i.i.d. In the left panel, $N = 512$ dimensional Monte Carlo simulations of student B, in blue, are compared against the solution of Eqs. (3.22), in orange. The blue dots and error bars represent the means and standard deviations, respectively, of the diagonal of the magnetization m during the simulations. The right panel shows the difference $m_B - m_A$ of the magnetizations of A and B as a function of the simulation epochs. The solid blue line and the shaded region represent the median of $m_A - m_B$ over $\alpha \in [0, 1]$ and the corresponding mean absolute deviation around the median, respectively. $m_A - m_B$ goes to zero when the number of elapsed epochs is large, so student A converges to the solution of Eqs. (3.23) like student B. The inverse temperature is set to $\beta^* = \beta = 4$.

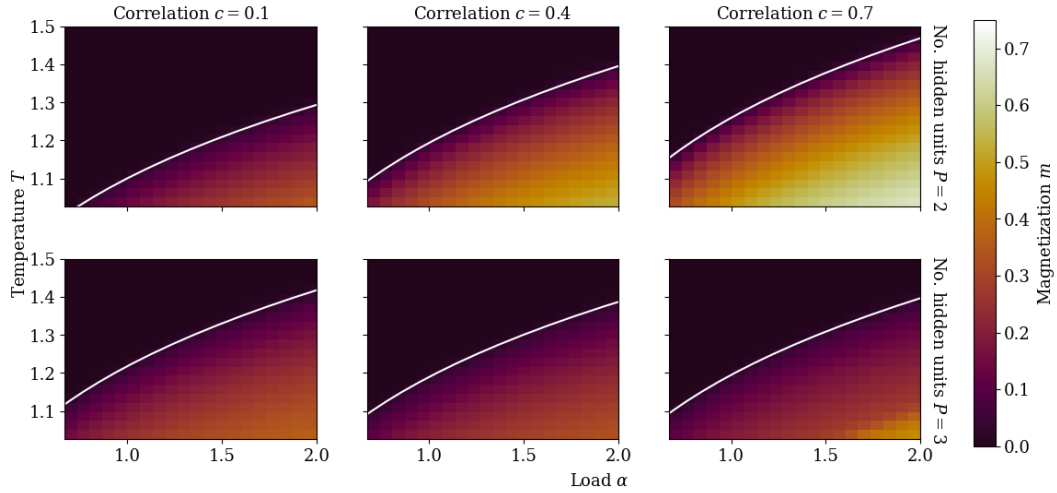

Figure 3.16: Mattis magnetization m for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. m is obtained by solving Eqs. (3.14) for binary student patterns with a uniform prior and binary teacher patterns with covariance $\mathcal{Q}_{\mu\nu} \sim \mathcal{W}(c, P)$, where $c \in [0, 1)$ (see 3.A.2). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the phase transition described by Eq. (3.18) with λ_{max}^{S} given by Eq. (3.24).

3.4 Conclusion

In this paper, we theoretically study the learning performance of restricted Boltzmann machines (RBM) [136, 137, 138, 24] with a finite number of hidden units in the teacher-student setting [145, 146, 39, 101, 147, 42, 40] using the replica method under the replica-symmetric (RS) approximation [112]. Given M data points and N visible units, we compute the critical data load $\alpha_{\text{crit}} = \frac{M}{N}$ above which learning becomes possible. Our findings validate the conjecture that the student's performance is independent of the number of hidden units when the patterns of the teacher are uncorrelated [39], generalizing the results of [42] to any number of hidden units P and teacher patterns P^* much smaller than M and N. In particular, we confirm that an RBM with P uncorrelated hidden units factorizes into P RBMs with one hidden unit each [42]. Additionally, we show that the teacher-student setting without correlations has both a permutation symmetry breaking (PSB) solution in which the student learns the teacher patterns one-to-one and a metastable partial PSB solutions in which multiple student patterns can converge to the same teacher pattern. We argue that the teacher-student setting complies with the lottery ticket hypothesis [9, 148, 149, 150, 151], and demonstrate that the student can be trained efficiently using a variant of the magnitude pruning algorithm [9, 151]. Given teacher patterns with uniform correlations $c \in [0, 1)$, we find a closed-form expression for α_{crit}, which generalizes the one found in [42] for $P = P^* = 2$, and show that it decreases with both c and P^*. Still in the case of uniform correlations, we find that the teacher-student setting undergoes a sequence of spontaneous symmetry breaking (SSB) and permutation symmetry breaking transitions that generalizes the one found in [42] to $P^*, P \geq 3$. Both with and without correlations, we find that decreasing the inference temperature T too much can prevent the student from learning the teacher's solution even when the dataset is relatively large. Throughout the paper, we compare key results against Monte Carlo simulations and observe that the RS ansatz is often a solid approximation even in the region outside the Nishimori line [73, 76, 74, 75] where the teacher and the student have the same temperature, but a different number of hidden units and a different prior on their patterns.

One could study the learning dynamics of the magnitude pruning experiment described in Section 3.3.2.3 or characterize the distribution of winning ticket initial conditions to gain additional insight into the lottery ticket hypothesis. More generally, it would be interesting to extend our study to the case where P and P^* are of order N. In fact, RBMs trained on real data typically have a large number of hidden units [24, 139, 140, 141, 182]. One could also introduce correlations between the columns $\xi_i^* = \left\{\xi_i^{*\mu}\right\}_{\mu=1}^{P}$ of the teacher's weight matrix $\xi^* = \left\{\xi_i^{*\mu}\right\}_{1\leq i\leq N}^{1\leq \mu\leq P}$ in addition to the correlations between patterns $\xi^{*\mu} = \left\{\xi_i^{*\mu}\right\}_{i=1}^{N}$ that we study in this work. Such a modification could be achieved by using a weight matrix with a low-rank SVD structure like in [31, 32] and would allow the teacher to generate data with even more structure than in this paper. One could use this framework to investigate whether the distinction between PSB_s and PSB_t persists in real data and has a measurable effect on the critical slowing down patterns observed in RBM training [31, 32, 33, 35]. Another promising research avenue would be to study other generative models with a finite number of hidden units in the teacher-student setting, such as modern (a.k.a. dense) Hopfield networks [23, 45, 25, 26, 30]. In sum, the teacher-student setting still has a lot of untapped potential for studying generative models with many hidden units.

Data availability

The code and hyperparameter values of the training algorithms used to make the figures are available at the following public repository [167].

Acknowledgements

I am grateful to Francesco Tosello, who made a significant contribution to this work. He wrote a first draft of the calculations presented in the Appendices, which Daniele Tantari and I then generalized and expanded upon to write this Chapter.

3.A Definitions

3.A.1 Binary random variables with a fixed covariance matrix

We generate binary random variables $\xi_i^* \in \{-1, +1\}^{P^*}$ with mean 0 and a fixed covariance matrix $\mathcal{Q}$ in two steps:

- sample $x_i^* \sim \mathcal{N}\left(0, \sin\left(\frac{\pi}{2}\mathcal{Q}\right)\right)$ from a multivariate Gaussian distribution with mean 0 and covariance matrix $\sin\left(\frac{\pi}{2}\mathcal{Q}\right)$,
- set $\xi_i^* = \operatorname{sign}\left(x_i^*\right)$.

This sampling method is commonly known as the arcsine law and originates from [183]. It implicitly defines a probability distribution $\mathrm{P}\left(\xi^*\right)$ for correlated binary teacher patterns ξ^*.

3.A.2 Projected Wishart distribution

We generate random positive definite matrices $\mathcal{B}$ by sampling them from the Wishart distribution [184]. That is, we sample the columns of a $P \times D$ matrix $\mathcal{A}$ from the Gaussian distribution $\mathcal{N}\left(0, \mathcal{C}\right)$, then set $\mathcal{B} = \mathcal{A}\mathcal{A}^T$. Let $\mathcal{D}$ be the diagonal matrix with the same values as $\mathcal{B}$ on the diagonal. We obtain a positive definite matrix $\mathcal{Q}$ with ones on the diagonal by normalizing $\mathcal{B}$ according to $\mathcal{Q} = \mathcal{D}^{-1/2}\mathcal{B}\mathcal{D}^{-1/2}$. In other words, $\mathcal{B}$ is a generic covariance matrix, and the entries $\mathcal{Q}_{\mu\nu}$ of $\mathcal{Q}$ are the Pearson correlation coefficients corresponding to $\mathcal{B}_{\mu\nu}$. For simplicity's sake, we take the covariance matrix $\mathcal{C}$ of the Gaussian distribution $\mathcal{N}\left(0, \mathcal{C}\right)$ to be $\mathcal{C}_{\mu\nu} = \delta_{\mu\nu} + \left(1 - \delta_{\mu\nu}\right) c$. In this paper, we say that $\mathcal{Q}$ is sampled from the projected Wishart distribution $\mathcal{W}\left(c, D\right)$.

3.A.3 Effective Hamiltonian $\mathcal{L}_{\lambda_1\lambda_2}$

The effective Hamiltonian $\mathcal{L}_{\lambda_1\lambda_2}$ plays a crucial role in deriving the free entropy and the saddle-point equations of the problem studied in this work. In this Section, we define it to be

$$\mathcal{L}_{\lambda_1,\lambda_2}\left(\xi,\xi^*,z;m,s,q\right)=\frac{1}{2}\left[\lambda_2\right]^2\sum_{\mu,\nu=1}^{P}\left(s^{\mu\nu}-\frac{q^{\mu\nu}+q^{\nu\mu}}{2}\right)\xi^\mu\xi^\nu+\lambda_1\lambda_2\sum_{\gamma=1}^{P^*}\sum_{\mu=1}^{P}m^{\gamma\mu}\xi^{*\gamma}\xi^\mu$$
$$+\lambda_2\sum_{\mu,\nu=1}^{P}\sqrt{q^{\mu\nu}+q^{\nu\mu}-\delta_{\mu\nu}\sum_{\eta=1}^{P}\frac{q^{\mu\eta}+q^{\eta\mu}}{2}}z_{\mu\nu}\frac{\xi^\mu+\xi^\nu}{2}.$$

where we set the diagonal of s to $s^{\mu\mu}=0$ to regroup $s^{\mu\nu}$ and $\frac{q^{\mu\nu}+q^{\nu\mu}}{2}$ in the same sum. In $\mathcal{L}^O$ (see Eq. 3.9), the diagonal of s is arbitrary because it does not affect the saddle-point equations (see Eqs. 3.14). In $\mathcal{L}^C$ (see Eq. 3.10), the diagonal of $\hat{s}$ always vanishes (also see Eqs. 3.14). When q and $\hat{q}$ are symmetric, we obtain the simpler form

$$\mathcal{L}_{\lambda_1,\lambda_2}\left(\xi,\xi^*,z;m,s,q\right)=\frac{1}{2}\left[\lambda_2\right]^2\sum_{\mu,\nu}\left(s^{\mu\nu}-q^{\mu\nu}\right)\xi^\mu\xi^\nu+\lambda_1\lambda_2\sum_{\gamma,\mu}m^{\gamma\mu}\xi^{*\gamma}\xi^\mu$$
$$+\lambda_2\sum_{\mu,\nu}\sqrt{2q^{\mu\nu}-\delta_{\mu\nu}\sum_{\eta}q^{\mu\eta}}z_{\mu\nu}\frac{\xi^\mu+\xi^\nu}{2},$$

which is the definition used in the main text (see Eq. 3.11). We symmetrize the initial conditions of $q^{\mu\nu}$ before solving the saddle-point equations numerically because any $q^{\mu\nu}$ solving them must be symmetric (see Eqs. 3.14 and 3.I). Defining

$$A_{\mu\nu}\left(q\right)=\sqrt{2q^{\mu\nu}-\delta_{\mu\nu}\sum_{\eta}q^{\mu\eta}},$$

these two expressions can be written more compactly as

$$\mathcal{L}_{\lambda_1,\lambda_2}\left(\xi,\xi^*,z;m,s,q\right)=\frac{1}{2}\left[\lambda_2\right]^2\sum_{\mu,\nu}\left(s^{\mu\nu}-q^{\mu\nu}\right)\xi^\mu\xi^\nu+\lambda_1\lambda_2\sum_{\gamma,\mu}m^{\gamma\mu}\xi^{*\gamma}\xi^\mu \tag{3.26}$$
$$+\lambda_2\sum_{\mu,\nu}A_{\mu\nu}\left(\frac{q+q^T}{2}\right)z_{\mu\nu}\frac{\xi^\mu+\xi^\nu}{2}\quad\text{for generic } q \text{ and}$$
$$\mathcal{L}_{\lambda_1,\lambda_2}\left(\xi,\xi^*,z;m,s,q\right)=\frac{1}{2}\left[\lambda_2\right]^2\sum_{\mu,\nu}\left(s^{\mu\nu}-q^{\mu\nu}\right)\xi^\mu\xi^\nu+\lambda_1\lambda_2\sum_{\gamma,\mu}m^{\gamma\mu}\xi^{*\gamma}\xi^\mu \tag{3.27}$$
$$+\lambda_2\sum_{\mu,\nu}A_{\mu\nu}\left(q\right)z_{\mu\nu}\frac{\xi^\mu+\xi^\nu}{2}\quad\text{for symmetric } q.$$

This way of writing $\mathcal{L}_{\lambda_1,\lambda_2}$ is useful to make the derivation of the saddle-point equations more concise (see 3.D). The third term of $\mathcal{L}_{\lambda_1,\lambda_2}$ can also be written as

$$\lambda_2 \sum_{\mu,\nu} A_{\mu\nu}(q)\, z_{\mu\nu} \frac{\xi^\mu + \xi^\nu}{2} = \lambda_2 \sum_{\mu,\nu} A_{\mu\nu}(q)\, \frac{z_{\mu\nu} + z_{\nu\mu}}{2} \xi^\mu$$

by interchanging summation indices. The two ways of writing the third term have different numerical implementations, but we did not see a significant difference in computational complexity and accuracy between them.

3.B Replicated partition function

In the following appendices, we omit the subscripts β^* and β to make notation lighter. The sums over γ or ρ are from 1 to P^*, and the sums over μ or ν, from 1 to P. Given a set of M examples $\boldsymbol{\sigma} = \{\sigma^a\}_{a=1}^M$, the probability distribution of a single replica b takes the form

$$\mathrm{P}\left(\xi^b \mid \boldsymbol{\sigma}\right) = \mathcal{Z}(\boldsymbol{\sigma})^{-1}\, \mathrm{P}\left(\xi^b\right) \prod_a \mathrm{P}\left(\sigma^a \mid \xi^b\right),$$

by using Bayes' theorem (see Section 3.2). The partition function is $\mathcal{Z}(\boldsymbol{\sigma}) = \mathbb{E}_{\xi^b}\left[\prod_a \mathrm{P}\left(\sigma^a \mid \xi^b\right)\right]$, which leads to

$$\begin{aligned}
\mathbb{E}_{\xi^*,\boldsymbol{\sigma}}\left[\mathcal{Z}^L\right] &= \sum_{\sigma} \mathbb{E}_{\xi^*}\left[\prod_a \mathrm{P}\left(\sigma^a \mid \xi^*\right)\right] \mathcal{Z}(\boldsymbol{\sigma})^L \\
&= \sum_{\sigma} \mathbb{E}_{\xi^*}\left[\prod_a \mathrm{P}\left(\sigma^a \mid \xi^*\right)\right] \mathbb{E}_{\xi}\left[\prod_a \mathrm{P}\left(\sigma^a \mid \xi\right)\right] \\
&= \mathbb{E}_{\xi^*\xi}\left[\prod_a \sum_{\sigma^a} \mathrm{P}\left(\sigma^a \mid \xi^*\right) \mathrm{P}\left(\sigma^a \mid \xi\right)\right] \\
&= \mathbb{E}_{\xi^*\xi}\left[\left(\sum_{\sigma^a} \mathrm{P}\left(\sigma^a \mid \xi^*\right) \mathrm{P}\left(\sigma^a \mid \xi\right)\right)^M\right]
\end{aligned}$$

where $\xi = \left\{\xi^b\right\}_{b=1}^L$ is a set of L replicas and $\mathrm{P}\left(\sigma^a \mid \xi\right) = \prod_b \mathrm{P}\left(\sigma^a \mid \xi^b\right)$. As per Eq. (3.2), $\mathrm{P}\left(\sigma^a \mid \xi^b\right)$ with binary units takes the form

$$\mathrm{P}\left(\sigma^a \mid \xi^b\right) = Z\left(\xi^b\right)^{-1} \psi\left(\sigma^a; \xi^b\right),$$

where $\psi\left(\sigma^a; \xi^b\right) = \mathrm{P}\left(\sigma^a\right) \mathbb{E}_{\tau_b}\left[\exp\left(\frac{\beta}{\sqrt{N}} \sum_\mu \tau_{b\mu} \sum_i \xi_i^{b\mu} \sigma_i^a\right)\right]$ and $Z\left(\xi^b\right) = \sum_{\sigma^a} \psi\left(\sigma^a; \xi^b\right)$. In particular, we have

$$\sum_{\sigma^a} \mathrm{P}\left(\sigma^a \mid \xi^*\right) \mathrm{P}\left(\sigma^a \mid \xi\right) = \sum_{\sigma^a} Z\left(\xi^*\right)^{-1} \left[\prod_b Z\left(\xi^b\right)^{-1}\right] \psi\left(\sigma^a; \xi^*\right) \left[\prod_b \psi\left(\sigma^a; \xi^b\right)\right].$$

We will now take $\mathrm{P}\left(\sigma^a\right)$ uniform and eliminate σ^a from $Z\left(\xi^*\right)\prod_b Z\left(\xi^b\right)$ in order to factor it out of the sum over σ^a. We first rewrite $Z\left(\xi^b\right)$ as

$$\begin{aligned}
Z\left(\xi^b\right) &= 2^{-N}\sum_{\sigma^a}\mathbb{E}_{\tau_b}\left[\exp\left(\frac{\beta}{\sqrt{N}}\sum_{\mu}\tau_{b\mu}\sum_i \xi_i^{b\mu}\sigma_i^a\right)\right] \\
&= 2^{-N}\mathbb{E}_{\tau_b}\left[\prod_i\sum_{\sigma_i^a}\exp\left(\frac{\beta}{\sqrt{N}}\sum_{\mu}\tau_{b\mu}\xi_i^{b\mu}\sigma_i^a\right)\right] \\
&= \mathbb{E}_{\tau_b}\left[\prod_i\cosh\left(\frac{\beta}{\sqrt{N}}\sum_{\mu}\tau_{b\mu}\xi_i^{b\mu}\right)\right] \\
&= \mathbb{E}_{\tau_b}\left[\exp\left(\sum_i\log\cosh\left[\frac{\beta}{\sqrt{N}}\sum_{\mu}\tau_{b\mu}\xi_i^{b\mu}\right]\right)\right],
\end{aligned}$$

then we expand the $\log\cosh$ function in small $\frac{P}{N}$ to obtain

$$\begin{aligned}
Z\left(\xi^b\right) &\approx \mathbb{E}_{\tau_b}\left[\exp\left(\sum_i\frac{1}{2}\left[\frac{\beta}{\sqrt{N}}\sum_{\mu}\tau_{b\mu}\xi_i^{b\mu}\right]^2\right)\right] \\
&= \mathbb{E}_{\tau_b}\left[\exp\left(\frac{1}{2}\beta^2\sum_{\mu,\nu}\tau_{b\mu}\tau_{b\nu}\frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{b\nu}\right)\right].
\end{aligned}$$

The equivalent expression for $Z\left(\xi^*\right)$ is identical except for the asterisk $*$ replacing the replica index b and the sum running from 1 to P^* rather than 1 to P. Once $Z\left(\xi^*\right)\prod_b Z\left(\xi^b\right)$ is factored out of the sum, $\sum_{\sigma^a}\psi\left(\sigma^a;\xi^*\right)\prod_b\psi\left(\sigma^a;\xi^b\right)$ simplifies in a similar way to

$$\begin{aligned}
\sum_{\sigma^a}\psi\left(\sigma^a;\xi^*\right)\prod_b\psi\left(\sigma^a;\xi^b\right) &\approx \mathbb{E}_{\tau_*\tau}\left[\exp\left(\sum_i\frac{1}{2}\left[\frac{\beta^*}{\sqrt{N}}\sum_{\gamma}\tau_{*\gamma}\xi_i^{*\gamma}+\frac{\beta}{\sqrt{N}}\sum_{\mu;b}\tau_{b\mu}\xi_i^{b\mu}\right]^2\right)\right] \\
&= \mathbb{E}_{\tau_*\tau}\left[\exp\left(\frac{1}{2}\left[\beta^*\right]^2\sum_{\gamma,\rho}\tau_{*\gamma}\tau_{*\rho}\frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{*\rho}\right.\right. \\
&\quad+\beta^*\beta\sum_{\gamma,\mu;b}\tau_{*\gamma}\tau_{b\mu}\frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{b\mu} \\
&\quad\left.\left.+\ \frac{1}{2}\beta^2\sum_{\mu,\nu;b,c}\tau_{b\mu}\tau_{c\nu}\frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{c\nu}\right)\right],
\end{aligned}$$

up to an irrelevant multiplicative factor of 2^{-LN}. Combining all these expressions, we get

$$\begin{aligned}\mathbb{E}\left[\mathcal{Z}^L\right] = \mathbb{E}_{\xi^*\xi}\Bigg[\exp\Bigg\{M\Bigg(\log\Bigg[\mathbb{E}_{\tau_*\tau}\exp\Bigg([\beta^*]^2\sum_{\gamma<\rho}\tau_{*\gamma}\tau_{*\rho}\frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{*\rho} \\ + \beta^*\beta\sum_{\gamma,\mu;b}\tau_{*\gamma}\tau_{b\mu}\frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{b\mu} + \beta^2\sum_{\mu<\nu;b}\tau_{b\mu}\tau_{b\nu}\frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{b\nu} \\ + \beta^2\sum_{\mu,\nu;b<c}\tau_{b\mu}\tau_{c\nu}\frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{c\nu}\Bigg)\Bigg] \\ - \log\left[\mathbb{E}_{\tau_*}\exp\left([\beta^*]^2\sum_{\gamma<\rho}\tau_{*\gamma}\tau_{*\rho}\frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{*\rho}\right)\right] \\ - \log\left[\mathbb{E}_{\tau}\exp\left(\beta^2\sum_{\mu<\nu;b}\tau_{b\mu}\tau_{b\nu}\frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{b\nu}\right)\right]\Bigg)\Bigg\}\Bigg].\end{aligned}$$

3.C RS free entropy

In this Section, we introduce order parameters and conjugate order parameters for the various overlaps between patterns present in $\mathbb{E}\left[\mathcal{Z}^L\right]$. We define

$$\begin{aligned} & m^{b\gamma\mu} \text{ and } \hat{m}^{b\gamma\mu} && \text{for } \frac{1}{N}\sum_i\xi_i^{*\gamma}\xi_i^{b\mu}, \\ & s^{b\mu\nu} \text{ and } \hat{s}^{b\mu\nu} && \text{for } \frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{b\nu} \text{ where } \mu\neq\nu, \\ & q^{bc\mu\nu} \text{ and } \hat{q}^{bc\mu\nu} && \text{for } \frac{1}{N}\sum_i\xi_i^{b\mu}\xi_i^{c\nu} \text{ where } b\neq c, \end{aligned}$$

where $m^{b\gamma\mu}$, $s^{b\mu\nu}$ and $q^{bc\mu\nu}$ are the ordinary order parameters, $\hat{m}^{b\gamma\mu}$, $\hat{s}^{b\mu\nu}$ and $\hat{q}^{bc\mu\nu}$ are the conjugate order parameters and $s^{b\mu\nu}$ and $\hat{s}^{b\mu\nu}$ are symmetric in μ and ν by construction. We use a Fourier transform to rewrite

$\mathbb{E}\left[\mathcal{Z}^L\right]$ as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^L\right] = \mathbb{E}_{\xi^*\xi}\Bigg[& \int_{i\mathbb{R}} \prod_{\gamma,\mu;b} d\hat{m}^{b\gamma\mu} \prod_{\mu<\nu;b} d\hat{s}^{b\mu\nu} \prod_{\mu,\nu;b<c} d\hat{q}^{bc\mu\nu} \\
& \int_{\mathbb{R}} \prod_{\gamma,\mu;b} dm^{b\gamma\mu} \prod_{\mu<\nu;b} ds^{b\mu\nu} \prod_{\mu,\nu;b<c} dq^{bc\mu\nu} \exp\left\{ \sum_{\gamma,\mu;b} \hat{m}^{b\gamma\mu} \left(\sum_i \xi_i^{*\gamma} \xi_i^{b\mu} - N m^{b\gamma\mu} \right) \right\} \\
& \exp\left\{ \sum_{\mu<\nu;b} \hat{s}^{b\mu\nu} \left(\sum_i \xi_i^{b\mu} \xi_i^{b\nu} - N s^{b\mu\nu} \right) + \sum_{\mu,\nu;b<c} \hat{q}^{bc\mu\nu} \left(\sum_i \xi_i^{b\mu} \xi_i^{c\nu} - N q^{bc\mu\nu} \right) \right\} \\
& \exp\Bigg\{ M \Bigg(\log \Bigg[\mathbb{E}_{\tau_*\tau} \exp\Bigg([\beta^*]^2 \sum_{\gamma<\rho} \tau_{*\gamma}\tau_{*\rho} \frac{1}{N} \sum_i \xi_i^{*\gamma} \xi_i^{*\rho} \\
& + \beta^*\beta \sum_{\gamma,\mu;b} m^{b\gamma\mu} \tau_{*\gamma} \tau_{b\mu} + \beta^2 \sum_{\mu<\nu;b} s^{b\mu\nu} \tau_{b\mu}\tau_{b\nu} + \beta^2 \sum_{\mu,\nu;b<c} q^{bc\mu\nu} \tau_{b\mu}\tau_{c\nu} \Bigg) \Bigg] \\
& - \log \left[\mathbb{E}_{\tau_*} \exp\left([\beta^*]^2 \sum_{\gamma<\rho} \tau_{*\gamma}\tau_{*\rho} \frac{1}{N} \sum_i \xi_i^{*\gamma} \xi_i^{*\rho} \right) \right] \\
& - \log \left[\mathbb{E}_{\tau} \exp\left(\beta^2 \sum_{\mu<\nu;b} s^{b\mu\nu} \tau_{b\mu}\tau_{b\nu} \right) \right] \Bigg) \Bigg\} \Bigg].
\end{aligned}$$

By hypothesis, the columns $\xi_i^* = \left\{\xi_i^{*\gamma}\right\}_{\gamma=1}^{P^*}$ of $\xi^* \sim \mathrm{P}\left(\xi^*\right)$ are i.i.d. random variables and their distribution $\mathrm{P}\left(\xi_i^*\right)$ has a well-defined $P^* \times P^*$ covariance matrix $\mathcal{Q}$ (see Section 3.1). Therefore, by the law of large numbers, $\frac{1}{N}\sum_i \xi_i^{*\gamma}\xi_i^{*\rho}$ converges in probability to the covariance $\mathcal{Q}_{\gamma\rho}$ as $N \to \infty$. $\mathbb{E}\left[\mathcal{Z}^L\right]$ then simplifies to

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^L\right] = \int & \prod_{\gamma,\mu;b} d\hat{m}^{b\gamma\mu} dm^{b\gamma\mu} \prod_{\mu<\nu;b} d\hat{s}^{b\mu\nu} ds^{b\mu\nu} \prod_{\mu,\nu;b<c} d\hat{q}^{bc\mu\nu} dq^{bc\mu\nu} \\
& \exp\left\{ N \log\left[\mathbb{E}_{\xi_i^*\xi_i} \exp\left\{ H_S\left(\xi_i, \xi_i^*; \hat{m}, \hat{s}, \hat{q}\right) \right\} \right] - N H_Q\left(m, s, q, \hat{m}, \hat{s}, \hat{q}\right) \right\} \\
& \exp\left\{ \alpha N \log\left[\left\langle \mathbb{E}_\tau \exp\left\{ H_E\left(\tau, \tau_*; m, s, q\right) \right\} \right\rangle_{\mathcal{M}_*} \right] - \alpha N \log\left[\mathcal{Z}\left(\mathcal{M}\right)\right] \right\}
\end{aligned}$$

where the thermal average $\langle\cdot\rangle_{\mathcal{M}_*}$ and the partition function $\mathcal{Z}\left(\mathcal{M}\right)$ are defined in Section 3.3.1 using Eqs. (3.7) and (3.8), respectively, $\alpha = \frac{M}{N}$ and

$$\begin{aligned}
H_Q\left(m, s, q, \hat{m}, \hat{s}, \hat{q}\right) &= \sum_{\gamma,\mu;b} \hat{m}^{b\gamma\mu} m^{b\gamma\mu} + \sum_{\mu<\nu;b} \hat{s}^{b\mu\nu} s^{b\mu\nu} + \sum_{\mu,\nu;b<c} \hat{q}^{bc\mu\nu} q^{bc\mu\nu}, \\
H_S\left(\xi_i, \xi_i^*; \hat{m}, \hat{s}, \hat{q}\right) &= \sum_{\gamma,\mu;b} \hat{m}^{b\gamma\mu} \xi_i^{*\gamma} \xi_i^{b\mu} + \sum_{\mu<\nu;b} \hat{s}^{b\mu\nu} \xi_i^{b\mu} \xi_i^{b\nu} + \sum_{\mu,\nu;b<c} \hat{q}^{bc\mu\nu} \xi_i^{b\mu} \xi_i^{c\nu}, \\
H_E\left(\tau, \tau_*; m, s, q\right) &= \beta^*\beta \sum_{\gamma,\mu;b} m^{b\gamma\mu} \tau_{*\gamma}\tau_{b\mu} + \beta^2 \sum_{\mu<\nu;b} s^{b\mu\nu} \tau_{b\mu}\tau_{b\nu} + \beta^2 \sum_{\mu,\nu;b<c} q^{bc\mu\nu} \tau_{b\mu}\tau_{c\nu}.
\end{aligned}$$

We use the replica symmetry ansatz to simplify $\mathbb{E}\left[\mathcal{Z}^L\right]$ further. To be more precise, we assume that

$$\begin{aligned}
m^{b\gamma\mu} = m^{\gamma\mu} \text{ and } \hat{m}^{b\gamma\mu} = \hat{m}^{\gamma\mu} &\qquad \text{for all } b;\gamma,\mu,\\
s^{b\mu\nu} = s^{\mu\nu} \text{ and } \hat{s}^{b\mu\nu} = \hat{s}^{\mu\nu} &\qquad \text{for all } b;\mu\neq\nu,\\
q^{bc\mu\nu} = q^{\mu\nu} \text{ and } \hat{q}^{bc\mu\nu} = \hat{q}^{\mu\nu} &\qquad \text{for all } b\neq c;\mu,\nu.
\end{aligned}$$

Under this hypothesis, the term $\sum_{\mu,\nu;b<c} q^{\mu\nu}\tau_{b\mu}\tau_{c\nu}$ can be rewritten as

$$\begin{aligned}
\sum_{\mu,\nu;b<c} q^{\mu\nu}\tau_{b\mu}\tau_{c\nu} &= \frac{1}{2}\sum_{\mu,\nu;b,c} q^{\mu\nu}\tau_{b\mu}\tau_{c\nu} - \frac{1}{2}\sum_{\mu,\nu;b} q^{\mu\nu}\tau_{b\mu}\tau_{b\nu}\\
&= \frac{1}{4}\sum_{\mu,\nu} q^{\mu\nu}\left[\sum_b (\tau_{b\mu}+\tau_{b\nu})\right]^2 - \frac{1}{4}\sum_{\mu,\nu} q^{\mu\nu}\left[\sum_b \tau_{b\mu}\right]^2\\
&\quad - \frac{1}{4}\sum_{\mu,\nu} q^{\mu\nu}\left[\sum_b \tau_{b\nu}\right]^2 - \frac{1}{2}\sum_{\mu,\nu;b} q^{\mu\nu}\tau_{b\mu}\tau_{b\nu}\\
&= \sum_{\mu,\nu}\frac{q^{\mu\nu}+q^{\nu\mu}}{2}\left[\sum_b \frac{\tau_{b\mu}+\tau_{b\nu}}{2}\right]^2 - \frac{1}{2}\sum_{\mu,\nu}\frac{q^{\mu\nu}+q^{\nu\mu}}{2}\left[\sum_b \tau_{b\mu}\right]^2\\
&\quad - \frac{1}{2}\sum_{\mu,\nu;b} q^{\mu\nu}\tau_{b\mu}\tau_{b\nu}\\
&= \frac{1}{2}\sum_{\mu,\nu}\left[q^{\mu\nu}+q^{\nu\mu}-\delta_{\mu\nu}\sum_{\tilde{\eta}}\frac{q^{\tilde{\mu}\tilde{\eta}}+q^{\tilde{\eta}\tilde{\mu}}}{2}\right]\left[\sum_b \frac{\tau_{b\mu}+\tau_{b\nu}}{2}\right]^2\\
&\quad - \frac{1}{2}\sum_{\mu,\nu;b} q^{\mu\nu}\tau_{b\mu}\tau_{b\nu}.
\end{aligned}$$

Subsequently, the Hubbard-Stratonovich transformation gives

$$\begin{aligned}
H_E(\tau,\tau_*;m,s,q) &= \beta^*\beta\sum_{\gamma,\mu;b} m^{\gamma\mu}\tau_{*\gamma}\tau_{b\mu} + \beta^2\sum_{\mu<\nu;b} s^{\mu\nu}\tau_{b\mu}\tau_{b\nu} + \beta^2\sum_{\mu,\nu;b<c} q^{\mu\nu}\tau_{b\mu}\tau_{c\nu}\\
&= \log\left(\mathbb{E}_z\left[\prod_b \exp\left\{\mathcal{L}_{\beta^*,\beta}(\tau_b,\tau_*,z;m,s,q)\right\}\right]\right)
\end{aligned}$$

where $\mathcal{L}_{\lambda_1,\lambda_2}(\xi,\xi^*,z;m,s,q)$ is defined in 3.A.3 and $z=[z_{\mu\nu}]_{\mu,\nu=1}^P$ is a matrix of independent standard Gaussian random variables $z_{\mu\nu}$. Similarly, we get

$$\begin{aligned}
H_S(\xi,\xi^*;\hat{m},\hat{s},\hat{q}) &= \sum_{\gamma,\mu;b}\hat{m}^{\gamma\mu}\xi^{*\gamma}\xi^{b\tilde{\mu}} + \sum_{\mu<\nu;b}\hat{s}^{\mu\nu}\xi^{b\tilde{\mu}}\xi^{b\tilde{\nu}} + \sum_{\mu,\nu;b<c}\hat{q}^{\mu\nu}\xi^{b\tilde{\mu}}\xi^{c\nu}\\
&= \log\left(\mathbb{E}_z\left[\prod_b \exp\left\{\mathcal{L}_{1,1}\left(\xi^b,\xi^*,z;\hat{m},\hat{s},\hat{q}\right)\right\}\right]\right).
\end{aligned}$$

We then factor $\mathbb{E}\left[\mathcal{Z}^L\right]$ over the replicas and take the limit of $L \to 0, N \to \infty$ to obtain

$$
\begin{aligned}
f &= \underset{m,\hat{m},q,\hat{q},s,\hat{s}}{\text{Extr}} f\left(m,\hat{m},q,\hat{q},s,\hat{s}\right) \\
&= \underset{m,\hat{m},q,\hat{q},s,\hat{s}}{\text{Extr}} \Bigg\{ -\sum_{\gamma,\mu} m^{\gamma\mu}\hat{m}^{\gamma\mu} - \frac{1}{2}\sum_{\mu\neq\nu} s^{\mu\nu}\hat{s}^{\mu\nu} + \frac{1}{2}\sum_{\mu,\nu} q^{\mu\nu}\hat{q}^{\mu\nu} \\
&\quad + \mathbb{E}_{\xi^*}\mathbb{E}_z \log\left[\mathcal{Z}\left(\mathcal{L}^C\right)\right] + \alpha\left\langle \mathbb{E}_z \log\left[\mathcal{Z}\left(\mathcal{L}^O\right)\right]\right\rangle_{\mathcal{M}_*} - \alpha\log\left[\mathcal{Z}\left(\mathcal{M}\right)\right] \Bigg\},
\end{aligned}
\tag{3.28}
$$

where we used the replica trick $\lim_{L\to 0}\left(\frac{1}{L}\log\mathbb{E}_z\left[\mathcal{Z}\left(\mathcal{L}_{\lambda_1,\lambda_2}\right)^L\right]\right) = \mathbb{E}_z\log\left[\mathcal{Z}\left(\mathcal{L}_{\lambda_1,\lambda_2}\right)\right]$ to simplify the expectations over z. The partition functions $\mathcal{Z}\left(\mathcal{L}^O\right)$ and $\mathcal{Z}\left(\mathcal{L}^C\right)$ are defined in Section 3.3.1 using Eqs. (3.9) and (3.10), respectively.

3.D Saddle-point equations

Our goal is to find the values of the order parameters for which the derivatives of $f\left(m,\hat{m},q,\hat{q},s,\hat{s}\right)$ vanish (see Eq. 3.28). For that purpose, we need to evaluate

$$
\begin{aligned}
\partial_{m^{\rho\iota}}\log\left[\mathcal{Z}\left(\mathcal{L}_{\lambda_1,\lambda_2}\right)\right] &= \left\langle \partial_{m^{\rho\iota}}\mathcal{L}_{\lambda_1,\lambda_2}\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
\partial_{s^{\iota\kappa}}\log\left[\mathcal{Z}\left(\mathcal{L}_{\lambda_1,\lambda_2}\right)\right] &= \left\langle \partial_{s^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
\partial_{q^{\iota\kappa}}\log\left[\mathcal{Z}\left(\mathcal{L}_{\lambda_1,\lambda_2}\right)\right] &= \left\langle \partial_{q^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}},
\end{aligned}
$$

so we must calculate the partial derivatives inside the expectation values. We use equation (3.26) for $\mathcal{L}_{\lambda_1,\lambda_2}$ because we do not know that q is symmetric before deriving the saddle-point equations (see 3.A.3). We obtain

$$
\begin{aligned}
\partial_{m^{\rho\iota}}\mathcal{L}_{\lambda_1,\lambda_2} &= \lambda_1\lambda_2\xi^{*\rho}\xi^{\iota} \\
\partial_{s^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2} &= \frac{1}{2}\left[\lambda_2\right]^2\xi^{\iota}\xi^{\kappa} \\
\partial_{q^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2} &= \frac{1}{4}\lambda_2\left(A^{-1}_{\iota\kappa}\left[z_{\iota\kappa}+z_{\kappa\iota}\right]\left[\xi^{\iota}+\xi^{\kappa}\right] - A^{-1}_{\iota\iota}z_{\iota\iota}\xi^{\iota} - A^{-1}_{\kappa\kappa}z_{\kappa\kappa}\xi^{\kappa}\right) \\
&\quad - \frac{1}{2}\left[\lambda_2\right]^2\xi^{\iota}\xi^{\kappa},
\end{aligned}
$$

where $A_{\mu\nu}$ is defined in 3.A.3. Using the first two equalities, we immediately get

$$
\begin{aligned}
\hat{m}^{\gamma\mu} &= \beta^*\beta\alpha\,\left\langle \mathbb{E}_z\left[\tau_{*\gamma}\left\langle\tau_{\mu}\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*} \\
\hat{s}^{\mu\nu} &= \beta^2\alpha\left(\left\langle \mathbb{E}_z\left[\left\langle\tau_{\mu}\tau_{\nu}\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*} - \left\langle\tau_{\mu}\tau_{\nu}\right\rangle_{\mathcal{M}}\right) \\
m^{\gamma\mu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\xi^{*\gamma}\left\langle\xi^{\mu}\right\rangle_{\mathcal{L}^C}\right] \\
s^{\mu\nu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\left\langle\xi^{\mu}\xi^{\nu}\right\rangle_{\mathcal{L}^C}\right],
\end{aligned}
$$

where the thermal averages $\langle\cdot\rangle_{\mathcal{M}_*}$, $\langle\cdot\rangle_{\mathcal{M}}$, $\langle\cdot\rangle_{\mathcal{L}^O}$ and $\langle\cdot\rangle_{\mathcal{L}^C}$ are defined in Section 3.3.1 using Eqs. (3.7), (3.8), (3.9) and (3.10), respectively. The case of the third derivative is a bit more involved. We find that its expectation with respect to the Gaussian variables z can be expressed as

$$\begin{aligned}
\mathbb{E}_z\left[\langle\partial_{q^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right] &= \mathbb{E}_z\Bigg[\frac{1}{4}\lambda_2\Big(A^{-1}_{\iota\kappa}\left[z_{\iota\kappa}+z_{\kappa\iota}\right]\langle\xi^\iota+\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} - A^{-1}_{\iota\iota}z_{\iota\iota}\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
&\quad - A^{-1}_{\kappa\kappa}z_{\kappa\kappa}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\Big) - \frac{1}{2}\left[\lambda_2\right]^2\langle\xi^\iota\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\Bigg] \\
&= \mathbb{E}_z\Bigg[\frac{1}{4}\lambda_2\Big(A^{-1}_{\iota\kappa}\Big[\partial_{z_{\iota\kappa}}\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} + \partial_{z_{\kappa\iota}}\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
&\quad + \partial_{z_{\iota\kappa}}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} + \partial_{z_{\kappa\iota}}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\Big] - A^{-1}_{\iota\iota}\partial_{z_{\iota\iota}}\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
&\quad - A^{-1}_{\kappa\kappa}\partial_{z_{\kappa\kappa}}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\Big) - \frac{1}{2}\left[\lambda_2\right]^2\langle\xi^\iota\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\Bigg]
\end{aligned}$$

by using integration by parts. As known from linear response theory [185], any Gibbs distribution with a generic Hamiltonian $\mathcal{H}$ verifies the identity

$$\partial_x\langle\theta\rangle_{\mathcal{H}} = \langle\theta\partial_x\mathcal{H}\rangle_{\mathcal{H}} - \langle\theta\rangle_{\mathcal{H}}\langle\partial_x\mathcal{H}\rangle_{\mathcal{H}} \tag{3.29}$$

for any order parameter θ that does not depend on x. We use this identity to get

$$\begin{aligned}
\partial_{z_{\iota\kappa}}\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} &= \langle\xi^\iota\partial_{z_{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} - \langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\langle\partial_{z_{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
&= \lambda_2\left\langle A_{\iota\kappa}\xi^\iota\frac{\xi^\iota+\xi^\kappa}{2}\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} - \lambda_2\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\left\langle A_{\iota\kappa}\frac{\xi^\iota+\xi^\kappa}{2}\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} \\
&= \frac{1}{2}\lambda_2 A_{\iota\kappa}\left(\langle\xi^\iota\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} - \langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right) \\
&\quad + \frac{1}{2}\lambda_2 A_{\iota\kappa}\left(\left\langle\left[\xi^\iota\right]^2\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} - \left[\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right]^2\right),
\end{aligned}$$

Coming back to the previous expression, we obtain

$$\mathbb{E}_z\left[\langle\partial_{q^{\iota\kappa}}\mathcal{L}_{\lambda_1,\lambda_2}\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right] = -\frac{1}{2}\left[\lambda_2\right]^2\mathbb{E}_z\left[\langle\xi^\iota\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\langle\xi^\kappa\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right].$$

From this result, we find that the two remaining saddle-point equations are

$$\begin{aligned}
\hat{q}^{\mu\nu} &= \beta^2\alpha\left\langle\mathbb{E}_z\left[\langle\tau_\mu\rangle_{\mathcal{L}^O}\langle\tau_{\tilde{\nu}}\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*} \\
q^{\mu\nu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\langle\xi^\mu\rangle_{\mathcal{L}^C}\langle\xi^\nu\rangle_{\mathcal{L}^C}\right].
\end{aligned}$$

3.E Saddle-point equations for Gaussian ξ

In this Appendix, we take the prior $\mathrm{P}(\xi)$ on the student pattern to be a standard Gaussian distribution. As per Section 3.3.1,

$$\begin{aligned}\mathcal{P}\left[\mathcal{L}^C\right](\xi;\xi^*,z) &= \mathcal{Z}\left(\mathcal{L}^C\right)^{-1}\frac{1}{\sqrt{(2\pi)^P}}\exp\left(-\frac{1}{2}\sum_\mu[\xi^\mu]^2\right)\exp\left[\mathcal{L}^C(\xi;\xi^*,z)\right]\\ &= \mathcal{Z}\left(\mathcal{L}^C\right)^{-1}\frac{1}{\sqrt{(2\pi)^P}}\exp\left(\mathcal{L}^C(\xi;\xi^*,z)-\frac{1}{2}\xi^T\xi\right),\end{aligned}$$

where the second line is written in matrix notation. This expression will allow us to simplify the thermal averages $\langle\cdot\rangle_{\mathcal{L}^C}$ in the saddle-point equations (Eqs. 3.14). By completing the square, $\mathcal{L}^C(\xi;\xi^*,z)-\frac{1}{2}\xi^T\xi$ can be written as

$$\begin{aligned}\mathcal{L}^C(\xi,\xi^*,z)-\frac{1}{2}\xi^T\xi = -\frac{1}{2}&\left(\xi-[I+\hat{q}-\hat{s}]^{-1}\left[\hat{m}^T\xi^*+\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\right]\right)^T\\ &(I+\hat{q}-\hat{s})\left(\xi-[I+\hat{q}-\hat{s}]^{-1}\left[\hat{m}^T\xi^*+\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\right]\right)\\ &+\frac{1}{2}\left[\hat{m}^T\xi^*+\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\right]^T\\ &[I+\hat{q}-\hat{s}]^{-1}\left[\hat{m}^T\xi^*+\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\right],\end{aligned}$$

where the function $\mathrm{diag}(\hat{q})$ returns the diagonal of $\hat{q}$. We read out the mean and variance as

$$\begin{aligned}\langle\xi\rangle_{\mathcal{L}^C} &= [I+\hat{q}-\hat{s}]^{-1}\left[\hat{m}^T\xi^*+\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\right],\\ \left\langle\xi\xi^T\right\rangle_{\mathcal{L}^C}-\langle\xi\rangle_{\mathcal{L}^C}\langle\xi\rangle_{\mathcal{L}^C}^T &= [I+\hat{q}-\hat{s}]^{-1}.\end{aligned}$$

By evaluating the thermal averages $\langle\cdot\rangle_{\mathcal{L}^C}$ of the saddle-point equations, we then obtain

$$\begin{aligned}m &= \mathbb{E}_{\xi^*}\left[\xi^*\xi^{*T}\hat{m}\,[I+\hat{q}-\hat{s}]^{-1}\right]\\ q &= \mathbb{E}_{\xi^*}\left[[I+\hat{q}-\hat{s}]^{-1}\hat{m}^T\xi^*\xi^{*T}\hat{m}\,[I+\hat{q}-\hat{s}]^{-1}\right]\\ &\quad+\mathbb{E}_z\left[[I+\hat{q}-\hat{s}]^{-1}\,\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}^T[I+\hat{q}-\hat{s}]^{-1}\right]\\ s &= [I+\hat{q}-\hat{s}]^{-1}+\mathbb{E}_{\xi^*}\left[[I+\hat{q}-\hat{s}]^{-1}\hat{m}^T\xi^*\xi^{*T}\hat{m}\,[I+\hat{q}-\hat{s}]^{-1}\right]\\ &\quad+\mathbb{E}_z\left[[I+\hat{q}-\hat{s}]^{-1}\,\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}\mathrm{diag}\left\{A(\hat{q})\frac{z+z^T}{2}\right\}^T[I+\hat{q}-\hat{s}]^{-1}\right].\end{aligned}$$

The expected value over z simplifies to

$$\begin{aligned}
&\mathbb{E}_z\left[\operatorname{diag}\left\{A\left(\hat{q}\right)\frac{z+z^T}{2}\right\}\operatorname{diag}\left\{A\left(\hat{q}\right)\frac{z+z^T}{2}\right\}^T\right]_{\mu\nu}\\
&=\mathbb{E}_z\left[\sum_{\tilde{\kappa}}A_{\tilde{\mu}\tilde{\kappa}}\left(\hat{q}\right)\frac{z_{\tilde{\kappa}\tilde{\mu}}+z_{\tilde{\mu}\tilde{\kappa}}}{2}\sum_{\tilde{\iota}}A_{\tilde{\nu}\tilde{\iota}}\left(\hat{q}\right)\frac{z_{\tilde{\iota}\tilde{\nu}}+z_{\tilde{\nu}\tilde{\iota}}}{2}\right]\\
&=\frac{1}{4}\mathbb{E}_z\left[\sum_{\tilde{\kappa},\tilde{\iota}}A_{\tilde{\mu}\tilde{\kappa}}\left(\hat{q}\right)A_{\tilde{\nu}\tilde{\iota}}\left(\hat{q}\right)z_{\tilde{\kappa}\tilde{\mu}}z_{\tilde{\iota}\tilde{\nu}}\right]\\
&\quad+\frac{1}{4}\mathbb{E}_z\left[\sum_{\tilde{\kappa},\tilde{\iota}}A_{\tilde{\mu}\tilde{\kappa}}\left(\hat{q}\right)A_{\tilde{\nu}\tilde{\iota}}\left(\hat{q}\right)z_{\tilde{\kappa}\tilde{\mu}}z_{\tilde{\nu}\tilde{\iota}}\right]\\
&\quad+\frac{1}{4}\mathbb{E}_z\left[\sum_{\tilde{\kappa},\tilde{\iota}}A_{\tilde{\mu}\tilde{\kappa}}\left(\hat{q}\right)A_{\tilde{\nu}\tilde{\iota}}\left(\hat{q}\right)z_{\tilde{\mu}\tilde{\kappa}}z_{\tilde{\iota}\tilde{\nu}}\right]\\
&\quad+\frac{1}{4}\mathbb{E}_z\left[\sum_{\tilde{\kappa},\tilde{\iota}}A_{\tilde{\mu}\tilde{\kappa}}\left(\hat{q}\right)A_{\tilde{\nu}\tilde{\iota}}\left(\hat{q}\right)z_{\tilde{\mu}\tilde{\kappa}}z_{\tilde{\nu}\tilde{\iota}}\right]\\
&=\frac{1}{2}\delta_{\mu\nu}\sum_{\tilde{\iota}}\left[A_{\tilde{\mu}\tilde{\iota}}\left(\hat{q}\right)\right]^2+\frac{1}{2}\left[A_{\mu\nu}\left(\hat{q}\right)\right]^2\\
&=\hat{q}^{\mu\nu}.
\end{aligned}$$

In the end, we find

$$\begin{aligned}
m &= \mathcal{Q}\hat{m}\left[I+\hat{q}-\hat{s}\right]^{-1}\\
q &= \left[I+\hat{q}-\hat{s}\right]^{-1}\hat{m}^T\mathcal{Q}\hat{m}\left[I+\hat{q}-\hat{s}\right]^{-1}\\
&\quad+\left[I+\hat{q}-\hat{s}\right]^{-1}\hat{q}\left[I+\hat{q}-\hat{s}\right]^{-1}\\
s &= \left[I+\hat{q}-\hat{s}\right]^{-1}+\left[I+\hat{q}-\hat{s}\right]^{-1}\hat{m}^T\mathcal{Q}\hat{m}\left[I+\hat{q}-\hat{s}\right]^{-1}\\
&\quad+\left[I+\hat{q}-\hat{s}\right]^{-1}\hat{q}\left[I+\hat{q}-\hat{s}\right]^{-1}.
\end{aligned}\tag{3.30}$$

3.F Critical load

In the paramagnetic phase, the order parameters all vanish. The paramagnetic to ferromagnetic phase transition of the student RBM is the line where the paramagnetic solution of the saddle-point equations (see Eq. 3.14) becomes unstable to leading order in the order parameters (see Fig. 3.10). As a consequence of Eq. (3.29), any Hamiltonian of the form $\mathcal{H}\left(\xi\right)=\frac{1}{2}\sum_{\mu\neq\nu}J_{\mu\nu}\xi^\mu\xi^\nu+\sum h_\mu\xi^\mu$ has

$$\left\langle\xi^\mu\right\rangle_{\mathcal{H}}\approx h_\mu$$

to first order in the parameters $J_{\mu\nu}$ and h_μ. Therefore, given that the prior on ξ has a mean of zero, we have

$$\begin{aligned}\mathbb{E}_z\left[\xi^{*\gamma}\left\langle\xi^\mu\right\rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}\right] &\approx \lambda_1\lambda_2\mathbb{E}_z\left[\xi^{*\gamma}\sum_\rho m^{\rho\mu}\xi^{*\rho}\right]\\ &= \lambda_1\lambda_2\sum_\rho m^{\rho\mu}\xi^{*\gamma}\xi^{*\rho}\end{aligned}$$

to first order in m, s and q. The saddle-point equations for $\hat{m}$ and m then simplify to

$$\begin{aligned}\hat{m}^{\gamma\mu} &= \beta^*\beta\alpha\,\left\langle\mathbb{E}_z\left[\tau_{*\gamma}\left\langle\tau_\mu\right\rangle_{\mathcal{L}^O}\right]\right\rangle_{\mathcal{M}_*}\\ &= \left[\beta^*\beta\right]^2\alpha\sum_\rho\left\langle\tau_{*\gamma}\tau_{*\rho}\right\rangle_{\mathcal{M}_*}m^{\rho\mu}\\ &= \left[\beta^*\beta\right]^2\alpha\sum_\rho\mathcal{R}_{\mu\rho}m^{\rho\mu}\\ m^{\gamma\mu} &= \mathbb{E}_{\xi^*}\mathbb{E}_z\left[\xi^{*\gamma}\left\langle\xi^\mu\right\rangle_{\mathcal{L}^O}\right]\\ &= \sum_\rho\mathbb{E}_{\xi^*}\left[\xi^{*\gamma}\xi^{*\rho}\right]\hat{m}^{\rho\mu}\\ &= \sum_\rho\mathcal{Q}_{\mu\rho}\hat{m}^{\rho\mu},\end{aligned}$$

where $\mathcal{Q}$ and $\mathcal{R}$ are the covariance matrices of ξ^* and $\mathcal{M}_*$, respectively. We see that the behavior of $\hat{m}$ and m is much simpler near criticality than at an arbitrary location in the phase diagram. In fact, the stationary values of $\hat{m}$ and m do not depend on the other order parameters. We can even rewrite the equations for $\hat{m}$ and m in the more compact form

$$m^{\gamma\mu} = \left[\beta^*\beta\right]^2\alpha\sum_{\iota,\kappa}\mathcal{Q}_{\mu\iota}\mathcal{R}_{\iota\kappa}m^{\kappa\mu}.$$

Let the largest eigenvalue of $\mathcal{S}_{\mu\kappa} = \sum_\iota\mathcal{Q}_{\mu\iota}\mathcal{R}_{\iota\kappa}$ be $\lambda^{\mathcal{S}}_{\text{max}}$. As known from stability theory [186], the paramagnetic solution $m^{\gamma\mu} = 0$ is unstable when $\left[\beta^*\beta\right]^2\alpha\,\mathcal{S}$ has at least one eigenvalue larger than 1. In other words, the student is able to learn the teacher patterns above a critical load of

$$\alpha_{\text{crit}} = \frac{1}{\left[\beta^*\beta\right]^2\lambda^{\mathcal{S}}_{\text{max}}}.$$

3.G Saddle-point equations in the absence of correlations

In the absence of correlations, we have $\mathcal{Q} = \mathbf{I}$. The effective Hamiltonian $\mathcal{M}_*$ (see Eq. 3.7) then simplifies to

$$\mathcal{M}_*\left(\tau_*\right) = \frac{1}{2}P^*\left[\beta^*\right]^2,$$

so the expectation $\langle\cdot\rangle_{\mathcal{M}_*}$ is uniform. When $P \leq P^*$, i.e. the student has at most the same number of hidden units as the teacher, we make the ansatz

$$\begin{aligned} m^{\gamma\mu} &= \delta_{\gamma\mu} m, & \hat{m}^{\gamma\mu} &= \delta_{\gamma\mu}\hat{m}, \\ s^{\mu\nu} &= \delta_{\mu\nu}, & \hat{s}^{\mu\nu} &= 0, \\ q^{\mu\nu} &= \delta_{\mu\nu} q, & \hat{q}^{\mu\nu} &= \delta_{\mu\nu}\hat{q}, \end{aligned} \tag{3.31}$$

under which $\mathcal{M}$ (Eq. 3.8) and $\mathcal{L}_{\lambda_1,\lambda_2}$ (Eq. 3.11) respectively simplify to

$$\mathcal{M}(\tau) = \frac{1}{2} P \beta^2 \quad \text{and}$$

$$\mathcal{L}_{\lambda_1,\lambda_2}(\xi^*, \xi, z; m, s, q) = \frac{1}{2} P \left[\lambda_2\right]^2 (1 - q) + \sum_{\mu=1}^{P} \left(\lambda_1 \lambda_2 m \xi^{*\mu} + \lambda_2 \sqrt{q} z_\mu\right) \xi^\mu.$$

The spins ξ^μ do not interact with one another in $\mathcal{L}_{\lambda_1,\lambda_2}(\xi^*, \xi, z; m, s, q)$. In other words, they are independent with respect to the Gibbs distribution with Hamiltonian $\mathcal{L}_{\lambda_1,\lambda_2}$. Therefore, all of the spins with $\nu \neq \mu$ can be marginalized from the thermal average $\langle \xi^\mu \rangle_{\mathcal{L}_{\lambda_1,\lambda_2}}$. In particular, when the student patterns ξ are binary random variables, we obtain

$$\begin{aligned} \langle \xi^\mu \rangle_{\mathcal{L}_{\lambda_1,\lambda_2}} &= \sum_{\xi^\mu = \pm 1} \frac{\exp\left(\left[\lambda_1 \lambda_2 m \xi^{*\mu} + \lambda_2 \sqrt{q} z_\mu\right] \xi^\mu\right) \xi^\mu}{\exp\left(\lambda_1 \lambda_2 m \xi^{*\mu} + \lambda_2 \sqrt{q} z_\mu\right) + \exp\left(-\lambda_1 \lambda_2 m \xi^{*\mu} - \lambda_2 \sqrt{q} z_\mu\right)} \\ &= \tanh\left(\lambda_1 \lambda_2 m \xi^{*\mu} + \lambda_2 \sqrt{q} z_\mu\right). \end{aligned}$$

Similarly, the hidden unit spins τ_μ are independent with respect to the Gibbs distribution with Hamiltonian $\mathcal{M}(\tau)$. By the independence of all these spins, the saddle-point equations (Eqs. 3.14) with binary ξ reduce to

$$\begin{aligned} \hat{m}^{\gamma\mu} &= \beta^* \beta \alpha \, \delta_{\gamma\mu} \, \left\langle \mathbb{E}_z \left[\tau_{*\mu} \tanh\left(\beta^* \beta m \tau_{*\mu} + \beta \sqrt{q} z_\mu\right)\right] \right\rangle_{\mathcal{M}_*} \\ \hat{s}^{\mu\nu} &= 0 \\ \hat{q}^{\mu\nu} &= \beta^2 \alpha \, \delta_{\mu\nu} \, \left\langle \mathbb{E}_z \left[\tanh^2\left(\beta^2 m \tau_{*\mu} + \beta \sqrt{q} z_\mu\right)\right] \right\rangle_{\mathcal{M}_*} \\ m^{\gamma\mu} &= \delta_{\gamma\mu} \, \mathbb{E}_{\xi^*} \mathbb{E}_z \left[\xi^{*\mu} \tanh\left(\hat{m} \xi^{*\mu} + \sqrt{\hat{q}} z_\mu\right)\right] \\ s^{\mu\nu} &= 0 \\ q^{\mu\nu} &= \delta_{\mu\nu} \, \mathbb{E}_{\xi^*} \mathbb{E}_z \left[\tanh^2\left(\hat{m} \xi^{*\mu} + \sqrt{\hat{q}} z_\mu\right)\right]. \end{aligned}$$

Assume the teacher patterns ξ^* are also binary. Since $\tanh$ is an odd function, we factor the spins out of it according to $\tanh(\xi^{*\mu} y) = \xi^{*\mu} \tanh(y)$ and use the change of variables $z = \xi^{*\mu} z_\mu$ to simplify the

saddle-point equations to

$$\begin{aligned}
\hat{m} &= \beta^* \beta \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^* \beta m + \beta \sqrt{q} z \right) \right] \\
\hat{q} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta^2 m + \beta \sqrt{q} z \right) \right] \\
m &= \mathbb{E}_z \left[\tanh \left(\hat{m} + \sqrt{\hat{q}} z \right) \right] \\
q &= \mathbb{E}_z \left[\tanh^2 \left(\hat{m} + \sqrt{\hat{q}} z \right) \right].
\end{aligned}$$

On the Nishimori line $\beta^* = \beta$, we have

$$\begin{aligned}
\hat{m} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^2 m + \beta \sqrt{m} z \right) \right] \\
m &= \mathbb{E}_z \left[\tanh \left(\hat{m} + \sqrt{\hat{m}} z \right) \right].
\end{aligned}$$

When $P > P^*$, the saddle-point equations are slightly different. We make the ansatz

$$\begin{aligned}
&m^{\gamma\mu} = \delta_{\gamma\mu} m, && \hat{m}^{\gamma\mu} = \delta_{\gamma\mu} \hat{m}, \\
&s^{\mu\nu} = \delta_{\mu\nu}, && \hat{s}^{\mu\nu} = 0, \\
&q^{\mu\nu} = \begin{cases} \delta_{\mu\nu} q & \mu, \nu \leq P^* \\ \delta_{\mu\nu} g & \text{otherwise,} \end{cases} && \hat{q}^{\mu\nu} = \begin{cases} \delta_{\mu\nu} \hat{q} & \mu, \nu \leq P^* \\ \delta_{\mu\nu} \hat{g} & \text{otherwise,} \end{cases}
\end{aligned} \tag{3.32}$$

and obtain

$$\mathcal{L}_{\lambda_1, \lambda_2} \left(\xi^*, \xi, z; m, s, q \right) = \sum_{\mu=1}^{P^*} \left(\lambda_1 \lambda_2 m \xi^{*\mu} + \lambda_2 \sqrt{q} z_\mu \right) \xi^\mu + \sum_{\mu=P^*+1}^{P} \lambda_2 \sqrt{g} z_\mu \xi^\mu.$$

Following the same steps as for $P \leq P^*$, we get

$$\begin{aligned}
\hat{m} &= \beta^* \beta \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^* \beta m + \beta \sqrt{q} z \right) \right] \\
\hat{q} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta^2 m + \beta \sqrt{q} z \right) \right] \\
\hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta \sqrt{g} z \right) \right] \\
m &= \mathbb{E}_z \left[\tanh \left(\hat{m} + \sqrt{\hat{q}} z \right) \right] \\
q &= \mathbb{E}_z \left[\tanh^2 \left(\hat{m} + \sqrt{\hat{q}} z \right) \right] \\
g &= \mathbb{E}_z \left[\tanh^2 \left(\sqrt{\hat{g}} z \right) \right].
\end{aligned}$$

When $\beta = \beta^*$, these equations reduce to

$$\begin{aligned}
\hat{m} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^2 m + \beta \sqrt{m} z \right) \right] \\
\hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta \sqrt{g} z \right) \right] \\
m &= \mathbb{E}_z \left[\tanh \left(\hat{m} + \sqrt{\hat{m}} z \right) \right] \\
g &= \mathbb{E}_z \left[\tanh^2 \left(\sqrt{\hat{g}} z \right) \right].
\end{aligned}$$

When the student patterns ξ are Gaussian random variables, we find instead

$$\begin{aligned}
\hat{m} &= \beta^* \beta \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^* \beta m + \beta \sqrt{q} z \right) \right] \\
\hat{q} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta^2 m + \beta \sqrt{q} z \right) \right] \\
\hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta \sqrt{g} z \right) \right] \\
m &= \frac{\hat{m}}{1 + \hat{q}} \\
q &= \frac{\hat{m}^2}{\left(1 + \hat{q} \right)^2} + \frac{\hat{q}}{\left(1 + \hat{q} \right)^2} \\
g &= \frac{\hat{g}}{\left(1 + \hat{g} \right)^2}.
\end{aligned}$$

When $\beta = \beta^*$, these equations reduce to

$$\begin{aligned}
\hat{m} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh \left(\beta^2 m + \beta \sqrt{m} z \right) \right] \\
\hat{g} &= \beta^2 \alpha \, \mathbb{E}_z \left[\tanh^2 \left(\beta \sqrt{g} z \right) \right] \\
m &= \frac{\hat{m}}{1 + \hat{m}} \\
g &= \frac{\hat{g}}{\left(1 + \hat{g} \right)^2}.
\end{aligned}$$

3.H Effect of uniform correlations

We introduce uniform correlations in the teacher patterns by fixing the covariance matrix $\mathcal{Q}$ of ξ^* to $\mathcal{Q}_{\gamma\rho} = \delta_{\gamma\rho} + (1 - \delta_{\gamma\rho}) \, c$, where $c \in [0, 1)$. Given this particular $\mathcal{Q}$, the Hamiltonian $\mathcal{M}_*$ (see Eq. 3.7) simplifies to

$$\mathcal{M}_* \left(\tau_* \right) = \frac{1}{2} \left[\beta^* \right]^2 c \sum_{\gamma \neq \rho} \tau_{*\gamma} \tau_{*\rho} + \frac{1}{2} P^* \left[\beta^* \right]^2 . \tag{3.33}$$

The interaction between any two spins $\tau_{*\gamma}$ and $\tau_{*\rho}$ does not depend on the sites $\gamma \neq \rho$. Therefore, the covariance matrix $\mathcal{R}$ of τ_* has the same form as $\mathcal{Q}$, but with a different coefficient d outside the diagonal.

$\mathcal{S} = \mathcal{Q}\mathcal{R}$ then reduces to

$$\begin{aligned}\mathcal{S}_{\gamma\rho} &= \sum_{\tau} \left(c + (1-c)\,\delta_{\gamma\tau}\right)\left(d + (1-d)\,\delta_{\tau\rho}\right) \\ &= P^* cd + c\,(1-d) + (1-c)\,d + (1-c)\,(1-d)\,\delta_{\gamma\rho}\end{aligned}$$

Any $P^* \times P^*$ matrix of the form $\mathcal{A}_{\gamma\rho} = a + b\delta_{\gamma\rho}$ has eigenvalues

$$\begin{aligned}&\lambda_1^{\mathcal{A}} = P^* a + b \text{ with corresponding eigenvector } e = \frac{1}{\sqrt{P^*}}\begin{bmatrix}1 & \dots & 1\end{bmatrix}, \\ &\lambda_2^{\mathcal{A}} = b \text{ with corresponding eigenspace } \left\{ x \in \mathbb{R}^{P^*} \,\middle|\, \textstyle\sum_i x_i = 0 \right\}.\end{aligned}$$

Therefore, $\mathcal{S}$ has eigenvalues

$$\begin{aligned}\lambda_1^{\mathcal{S}} &= P^*\left(P^* cd + c\,(1-d) + (1-c)\,d\right) + (1-c)\,(1-d)\,, \\ \lambda_2^{\mathcal{S}} &= (1-c)\,(1-d)\,.\end{aligned}$$

d is positive because the interaction between any two spins $\tau_{*\gamma}$ and $\tau_{*\rho}$ is positive (see Eq. 3.33 and [187]). Therefore, the largest eigenvalue $\lambda_{\text{max}}^{\mathcal{S}}$ of $\mathcal{S}$ is $\lambda_1^{\mathcal{S}}$. In sum,

$$\lambda_{\text{max}}^{\mathcal{S}} = \lambda_1^{\mathcal{S}} = \left(P^* - 1\right)^2 cd + \left(P^* - 1\right)(c + d) + 1.$$

3.I Numerical methods

We solve the saddle-point equations (Eqs. 3.14) by numerical iteration. To be more specific, we iterate

$$\begin{aligned}\hat{m}^{\gamma\mu}(t+1) &= \hat{m}^{\gamma\mu}(t) + \Delta t\left(\beta^*\beta\alpha\,\mathbb{E}_{\mathcal{M}_*}\mathbb{E}_z\left[\tau_{*\gamma}\left\langle\tau_\mu\right\rangle_{\mathcal{L}^O(t+1)}\right] - \hat{m}^{\gamma\mu}(t)\right) \\ \hat{s}^{\mu\nu}(t+1) &= \hat{s}^{\mu\nu}(t) + \Delta t\left(\beta^2\alpha\left(\mathbb{E}_{\mathcal{M}_*}\mathbb{E}_z\left[\left\langle\tau_\mu\tau_{\tilde{\nu}}\right\rangle_{\mathcal{L}^O(t)}\right] - \left\langle\tau_\mu\tau_{\tilde{\nu}}\right\rangle_{\mathcal{M}(t)}\right) - \hat{s}^{\mu\nu}(t)\right) \\ \hat{q}^{\mu\nu}(t+1) &= \hat{q}^{\mu\nu}(t) + \Delta t\left(\beta^2\alpha\mathbb{E}_{\mathcal{M}_*}\mathbb{E}_z\left[\left\langle\tau_\mu\right\rangle_{\mathcal{L}^O(t)}\left\langle\tau_{\tilde{\nu}}\right\rangle_{\mathcal{L}^O(t)}\right] - \hat{q}^{\mu\nu}(t)\right) \\ m^{\gamma\mu}(t+1) &= m^{\gamma\mu}(t) + \Delta\tau\left(\mathbb{E}_{\xi^*}\mathbb{E}_z\left[\xi^{*\gamma}\left\langle\xi^\mu\right\rangle_{\mathcal{L}^C(t+1)}\right] - m^{\gamma\mu}(t)\right) \\ s^{\mu\nu}(t+1) &= s^{\mu\nu}(t) + \Delta\tau\left(\mathbb{E}_{\xi^*}\mathbb{E}_z\left[\left\langle\xi^\mu\xi^\nu\right\rangle_{\mathcal{L}^C(t+1)}\right] - s^{\mu\nu}(t)\right) \\ q^{\mu\nu}(t+1) &= q^{\mu\nu}(t) + \Delta\tau\left(\mathbb{E}_{\xi^*}\mathbb{E}_z\left[\left\langle\xi^\mu\right\rangle_{\mathcal{L}^C(t+1)}\left\langle\xi^\nu\right\rangle_{\mathcal{L}^C(t+1)}\right] - q^{\mu\nu}(t)\right),\end{aligned}$$

with time steps $\Delta t, \Delta\tau \in (0, 1]$. By construction, Eqs. (3.14) are a fixed point of the iteration. Empirically, the iteration is the most stable when one of the two time steps is equal to one and the other one is small. Even then, it is still occasionally unstable at large α, large teacher pattern correlations and low T, which introduces a few easily identifiable spurious discontinuities in Figs. (3.10). We symmetrize the initial conditions of q before the iteration because any q solving the saddle-point equations must be symmetric (see Eqs. 3.14).

We use Monte Carlo integration over whitened samples to estimate the Gaussian expectations $\mathbb{E}_z\left[\cdot\right]$. We enforce the sample means to be 0 by symmetrizing around the origin each set of samples $z_{\mu\nu}$ that approximates

the corresponding integral over $z_{\mu\nu}$. We then constrain the sample variances to 1 using Cholesky whitening [188].

$\mathcal{L}_{\lambda_1,\lambda_2}(\xi, \xi^*, z)$ is a complex-valued function because it involves the square root of a real number that is not necessarily positive. Therefore, the thermal averages $\langle\cdot\rangle_{\mathcal{L}^C}$ and $\langle\cdot\rangle_{\mathcal{L}^O}$, are also complex-valued. By the symmetry of the Gaussian variables in the saddle point equations (see Eqs. 3.14), the imaginary part of each thermal average is equal to a and $-a$ with the same probability. Therefore, the exact Gaussian expectations must be real-valued. However, in practice, standard Monte Carlo integration for evaluating $\mathbb{E}_z[g(z)]$ with a finite number of samples is very unlikely to randomly produce both $g(z)$ and its complex conjugate $\bar{g}(z)$. Therefore, we replace $\mathbb{E}_z[\cdot]$ by $\mathbb{E}_z[\mathrm{Re}(\cdot)]$ when solving the saddle-point equations numerically. This adjustment significantly increases the stability of the numerical iteration.

3.J Supplementary figures

This Appendix contains a graph of the free entropy difference of the PSB and partial PSB solutions of Eqs. (3.14), as well as some plots of the Mattis magnetization m and spin-glass overlap q obtained for real-valued student patterns with a standard Gaussian prior. The former supports some claims made in Section 3.3.2.1, but is not strictly necessary to understand the paper. The latter are not shown in the main text because they look similar to the m and q obtained for binary student patterns with a uniform binary prior. We simplified the saddle-point equations (Eqs. 3.14) according to 3.E in order to make them.

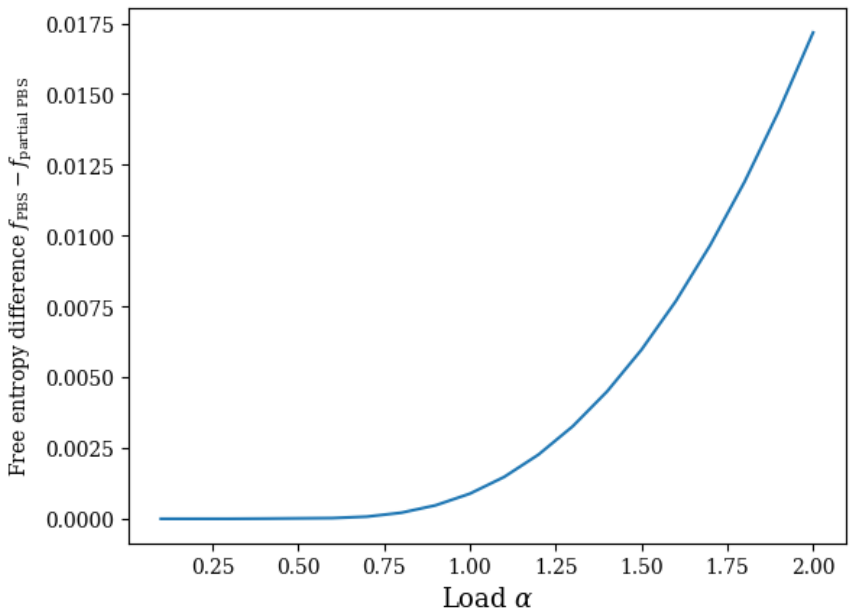

Figure 3.17: Free entropy difference of the so-called PSB and partial PSB solutions of Eqs. (3.14) shown in Figs. (3.2) and (3.3). This plot was made using $P^* = 2$ and $P = 3$, but the free entropy of $P^* = 3$ and $P = 4$ looks identical.

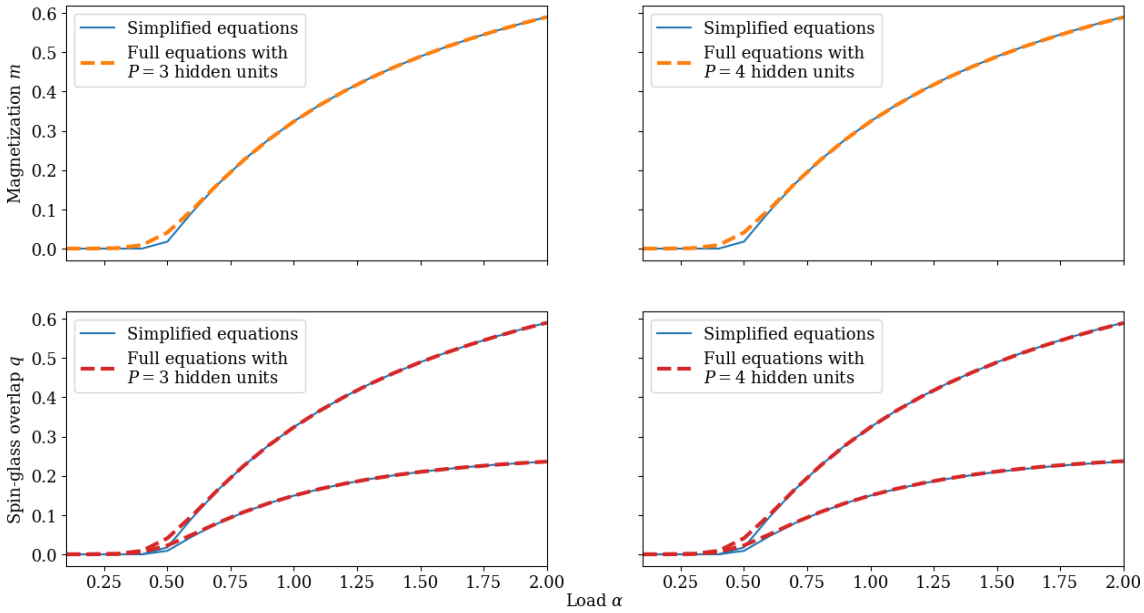

Figure 3.18: Permutation symmetry breaking (PSB) solution of Eqs. (3.14) for real-valued student patterns with a standard Gaussian prior and teacher pattern covariance $\mathcal{Q} = \mathbf{I}$, in red and orange, compared against the solution of Eqs. (3.23), in blue. We plot the Mattis magnetization m in the top row, and the SG overlap q in the bottom row. The magnetization plots and the top lines of the SG overlap plots show that the student patterns that converge to teacher patterns have the same m and q as the solution of Eqs. (3.23), and thus also satisfy $m = q$. Conversely, the bottom lines of the SG overlap plots student patterns that do not converge to a teacher pattern have a spin-glass overlap of q as in Eqs. (3.23). The top branch of q is We use $P = 3$ and $P^* = 2$ in the left column and $P = 4$ and $P^* = 3$ in the right column. All plots have $\beta^* = \beta = 1.2$.

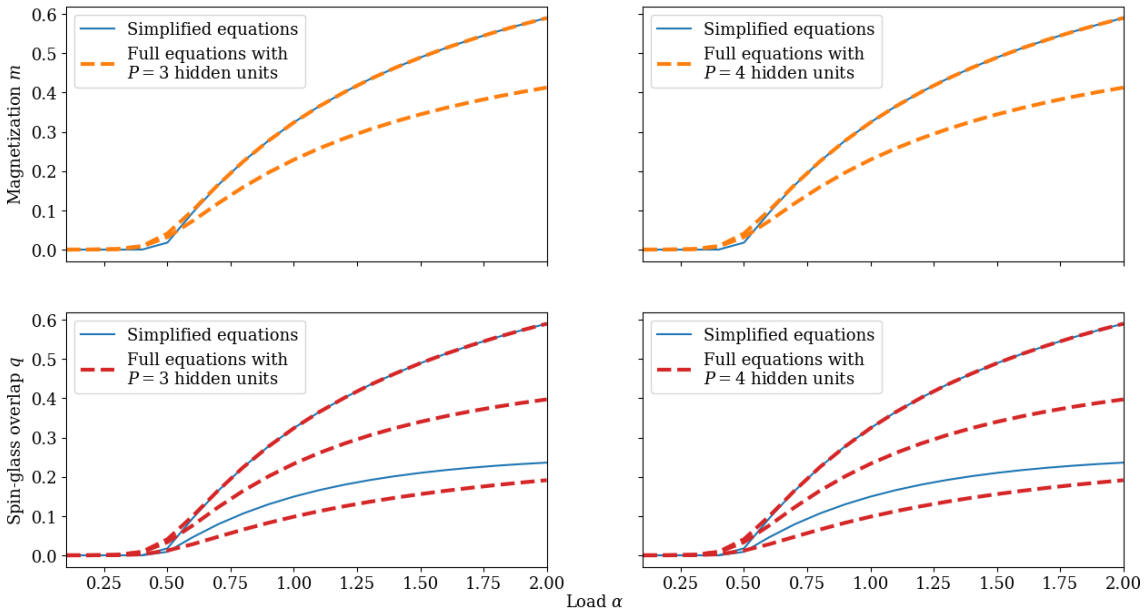

Figure 3.19: Partial permutation symmetry breaking (partial PSB) solutions of Eqs. (3.14) for real-valued student patterns with a standard Gaussian prior and teacher pattern covariance $\mathcal{Q} = \mathbf{I}$, in red and orange, compared against the solution of Eqs. (3.23), in blue. We plot the Mattis magnetization m in the top row, and the SG overlap g in the bottom row. The top lines of the plots show that the student patterns ξ^{μ}_{PSB} that converge to teacher patterns one-to-one have the same m and q as the solution of Eqs. (3.23), and thus also satisfy $m = q$. Conversely, the other lines show that the student patterns ξ^{μ}_{PS} that converge to a common teacher pattern have a smaller m and a different q. To be more precise, the central and bottom branches of q are the spin-glass order parameters corresponding to $Q(\xi^{1\mu}_{PS}, \xi^{2\mu}_{PS})$ and $Q(\xi^{1\mu}_{PS}, \xi^{2\nu}_{PS})$ with $\mu \neq \nu$, respectively (see Section 3.2). They are both different from the g of Eq. (3.23). The Mattis magnetization and SG overlaps omitted from this Figure all vanish. We use $P = 3$ and $P^* = 2$ in the left column and $P = 4$ and $P^* = 3$ in the right column. All plots have $\beta^* = \beta = 1.2$.

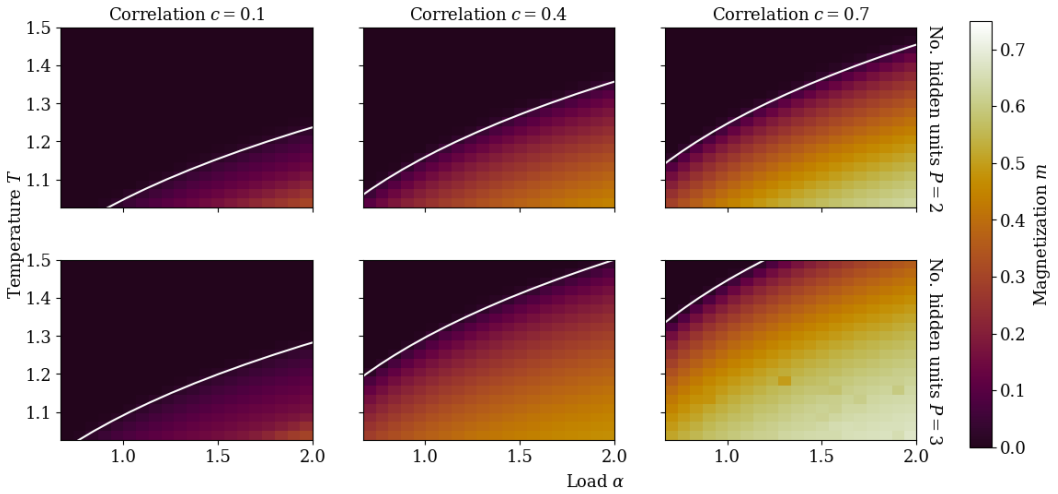

Figure 3.20: Mattis magnetization m for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. m is obtained by solving Eqs. (3.14) numerically for real-valued student patterns with a standard Gaussian prior and teacher pattern covariance $\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\, c$, where $c \in [0, 1)$ (see 3.H). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the phase transition of Eq. (3.18) with $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24). The speckles in the plots with $P = 3$, $c = 0.4$ and $P = 3$, $c = 0.7$ are due to the saddle-point iteration converging to a different solution of Eqs. (3.30) than at neighboring α and T.

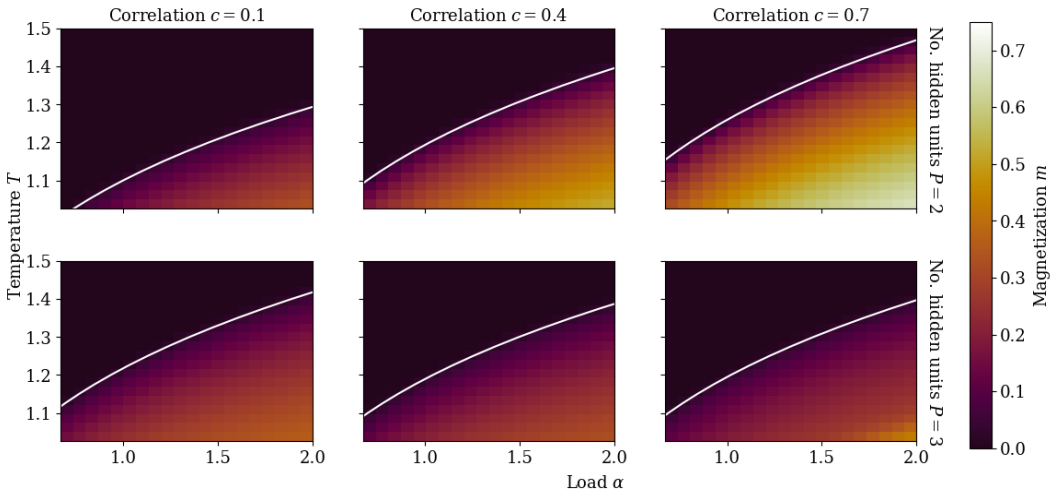

Figure 3.21: Mattis magnetization m for $\beta = \beta^*$ and $P = P^*$ as a function of the number of hidden units P, the correlation c, the temperature T and the data load α. m is obtained by solving Eqs. (3.14) for real-valued student patterns with a standard Gaussian prior and teacher pattern covariance $\mathcal{Q}_{\mu\nu} \sim \mathcal{W}(c, P)$, where $c \in [0, 1)$ (see 3.A.2). The top and bottom rows feature $P = 2$ and $P = 3$, respectively. The white lines mark the phase transition of Eq. (3.18) with $\lambda^{\mathcal{S}}_{\text{max}}$ given by Eq. (3.24).

Chapter 4

Saddle Hierarchy in Dense Associative Memory

Based on the preprint [47], doi: 10.48550/arXiv.2508.19151

4.1 Introduction

Studying the stationary points of machine learning algorithms is crucial to understand how they work. For example, [189, 190] demonstrated that local minima in the loss landscape of large artificial neural networks (NNs) are relatively close to the global minimum, explaining why they generalize well in practice. Moreover, [191, 192] showed that saddle points are much more numerous than local minima in large NNs. These breakthroughs were made by establishing deep connections between the loss landscape of NNs and the energy landscape of disordered systems studied in statistical mechanics. Beyond the broad insights provided by these studies, and despite the progress made by [193, 194, 195, 168], the classification of stationary points in machine learning algorithms remains an open problem. An interesting way forward is through dense associative memory (DAM) models [23, 30]. In fact, the interpretability of DAMs and the relative simplicity of their learning dynamics [34] suggest that their critical points could potentially be characterized using analytical calculations.

The Hopfield network was originally introduced as a paradigmatic model of biological associative memory [23]. Generalized Hopfield networks [25, 26, 27, 28, 108, 29] were then developed to improve upon the limited storage capacity of the original [23, 54]. A few years ago, these generalized networks, commonly referred to as dense associative memory (DAM) or modern Hopfield networks [30, 38], were made into trainable machine learning models capable of accurate pattern classification by Krotov and Hopfield (K & H) [30]. In a nutshell, K & H's DAM learns prototypes of patterns in a trainable weight matrix. Each prototype casts a vote for a class, and the patterns awaiting classification are assigned based on the votes of the prototypes that most closely resemble them. The resulting classification scheme is considerably more adversarially robust and interpretable

than that of feedforward NNs with ReLU activation functions [30, 36, 45]. Since their debut as trainable machine learning architectures, deep connections have been made between modern Hopfield networks and transformers [38], as well as generative diffusion models [65, 66]. In particular, modern Hopfield networks were used to implement the attention mechanism of transformers [38, 107], which has been attracting a lot of interest in fundamental [63, 196, 67, 109, 80] and applied research [107, 109, 197, 198]. Recently, it was observed that the trainable weights of K & H's DAM are channeled toward minima by a low-dimensional network of valleys in the loss landscape [34]. Moreover, the points where valleys branch out from one another were identified as saddles in the simple case where the DAM has two patterns to learn. In general, it is not straightforward to classify the stationary points of machine learning algorithms [168]. However, the results of [34] and the interpretability of DAMs suggest that their stationary points are both fundamental to their learning dynamics and easier to characterize than that of generic NNs. With this goal in mind, we revisit dense associative memory for pattern classification [30] using the framework of Boltzmann machines (BMs) [136, 137, 138, 24] and Statistical Mechanics.

In Section 4.2, we present the DAM model that we study and the tools that we use to do so. This Section is divided into two subsections. In Section 4.2.1, we derive a DAM from a BM template. In Section 4.2.2, we describe the setting in which we study DAM stationary points. In Section 4.3, we present our theoretical results. This Section is also divided into two subsections. In Section 4.3.1, we introduce saddle-point equations that characterize DAM stationary points. In Section 4.3.2, we use these equations to demonstrate the existence of a saddle-point hierarchy according to which the weights learned by DAMs of a given width are embedded in larger DAMs, where they become saddle points. In Section 4.4, we use our theoretical results to significantly improve DAM training, and we investigate the results. This Section is divided into three subsections. In Section 4.4.1, we design a theoretically motivated regularization method that facilitates supervised training. In Section 4.4.2, we show that our DAM, despite being designed for supervised learning, can learn interpretable solutions to both supervised and unsupervised classification problems. Finally, in Section 4.4.3, we relate our findings to the learning dynamics guided by valleys and saddles studied in [34], and we implement a network-growing algorithm [87, 88] that uses the saddle-point hierarchy to significantly reduce the cost of DAM training. The code and hyperparameter values used in our numerical experiments are available at this public repository [199].

4.2 Model

The Boltzmann machine (BM) is a canonical graphical model of correlations in discrete data [136]. It is customary to partition BMs into a visible layer $\mathbf{v} = \{v_i\}_{i=1}^{N}$ and a hidden layer $\mathbf{h} = \{h_\mu\}_{\mu=1}^{P}$ such that connections between the two layers are allowed, but connections within them are prohibited [137]. In this case, the visible layer represents concrete features of the data, whose mutual correlations are encoded in connections with the hidden layer. The restricted Boltzmann machine (RBM) obtained using this partition is much easier to train than a generic BM [138, 24] and still has considerable generating power [138, 143], making it more practical in machine learning applications [69, 141, 142]. The visible and hidden units of an RBM follow the

Gibbs distribution

$$\mathrm{P}_\beta\left(\mathbf{v},\mathbf{h}\mid\mathbf{J}\right)=Z_\beta\left(\mathbf{J}\right)^{-1}\mathrm{P}_0\left(\mathbf{v}\right)\mathrm{P}_0\left(\mathbf{h}\right)\exp\left(-\beta H\left[\mathbf{v},\mathbf{h};\mathbf{J}\right]\right),$$

where $\beta\geq 0$ is known as the inverse temperature, $\mathrm{P}_0\left(\mathbf{v}\right)$ and $\mathrm{P}_0\left(\mathbf{h}\right)$ are priors on $\mathbf{v}$ and $\mathbf{h}$, $H\left[\mathbf{v},\mathbf{h};\mathbf{J}\right]=-\sum_{i=1}^{N}\sum_{\mu=1}^{P}J_i^\mu v_i h_\mu$ is called the energy function or Hamiltonian, $\mathbf{J}=\{J_i^\mu\}_{1\leq i\leq N}^{1\leq\mu\leq P}$ are trainable weights, and $Z_\beta\left(\mathbf{J}\right)$ is a normalization constant called the partition function. The inverse temperature β represents the absolute strength of the RBM connections, or equivalently controls the amount of noise $T=1/\beta$ in the RBM. In this regard, $\mathrm{P}_0\left(\mathbf{v}\right)$ and $\mathrm{P}_0\left(\mathbf{h}\right)$, which restrict the form of the Gibbs distribution to help the RBM represent the data, are the marginal laws of $\mathbf{v}$ and $\mathbf{h}$ when there are no connections, i.e. $\beta=0$. Their contribution to the Gibbs distribution can be tuned with β, which can therefore be interpreted as a regularization parameter.

4.2.1 A dense associative memory (DAM) model

As mentioned in the Introduction, we will now derive a dense associative memory (DAM) model for classification from a BM template. We will explain why our model is a DAM at the end of this Section, once we have clearly defined it. We make three basic assumptions on the distribution of data to be classified:

1. the data is scale invariant, i.e. any data point $\mathbf{x}$ is equivalent to $c\mathbf{x}$;
2. the data can be partitioned in disjoint clusters;
3. the clusters can be grouped into mutually exclusive classes.

In order to exploit these three assumptions, we study a BM partitioned into three layers with different roles: the data layer $\mathbf{x}$, the hidden layer $\mathbf{h}$ and the class layer $\mathbf{q}$, which represent data, cluster membership and class membership, respectively. The corresponding energy is

$$-H\left[\mathbf{x},\mathbf{q},\mathbf{h};\mathbf{J}\right]=\sum_{i=1}^{N}\sum_{\mu=1}^{P}w_i^\mu x_i h_\mu+\sum_{y=1}^{C}\sum_{\mu=1}^{P}u_y^\mu q_y h_\mu+\sum_{\mu=1}^{P}h_\mu b^\mu, \tag{4.1}$$

where $\mathbf{J}=\{\mathbf{w},\mathbf{u},\mathbf{b}\}$ is the set of the trainable weights $\mathbf{w}=\{w_i^\mu\}_{1\leq i\leq N}^{1\leq\mu\leq P}$, $\mathbf{u}=\{u_i^\mu\}_{1\leq i\leq C}^{1\leq\mu\leq P}$ and $\mathbf{b}=\{b^\mu\}_{\mu=1}^{P}$. There are no direct interactions between the visible layer and the class layer. In other words, conditional on the cluster layer, the visible layer and the class layer are independent. Therefore, this BM is a deep Boltzmann machine (DBM) with 3 layers [200], which can also be thought of as an RBM whose visible layer $\mathbf{v}$ is further divided into $\mathbf{x}$ and $\mathbf{q}$.

Since the data is scale invariant (Assumption 1), we normalize it by its (Euclidean) norm in the data layer. In other terms, we take the data units x_i to be continuous variables with unit norm $\sqrt{\sum_{i=1}^{N}\left(x_i\right)^2}=1$. We assume no further knowledge about $\mathbf{x}$, so we take the prior $\mathrm{P}_0\left(\mathbf{x}\right)$ to be the uniform distribution on the $N-1$ dimensional unit hypersphere S^{N-1}. Data normalization is a very common practice in machine learning. For example, normalization by the Euclidean norm is frequent in text document clustering [201, 202, 203]. Various types of normalization also occur in the brain and retina [204].

Since the hidden layer and the class layer represent disjoint clusters and classes, respectively (Assumptions 2 and 3), we take their respective units to be mutually exclusive binary variables, i.e. $\mathbf{h}\in\{0,1\}^P$ with

$\sum_{\mu=1}^{P} h_\mu \in \{0,1\}$ and $\mathbf{q} \in \{0,1\}^C$ with $\sum_{y=1}^{C} q_y \in \{0,1\}$. In other words, we take each of these two layers to be the vector representation of a single categorical (or Potts [70, 71]) variable with $P+1$ and $C+1$ categories, respectively. As such, $\mathrm{P}_0(\mathbf{h})$ and $\mathrm{P}_0(\mathbf{q})$ simplify to probability mass functions $\mathrm{P}_0(\mathbf{h} = \mathbf{e}_\gamma)$ and $\mathrm{P}_0(\mathbf{q} = \mathbf{e}_y)$, where we introduce $\mathbf{e}_\gamma = \{\delta_{\gamma\mu}\}_{\mu=1}^{P}$ for $\gamma \in \{0, \ldots, P\}$ and define $\mathbf{e}_y \in \{0,1\}^C$ analogously for $y \in \{0, \ldots, C\}$. In particular $\mathbf{e}_0 = \mathbf{0}$ represents a state outside the P clusters or the C classes.

These priors on the hidden layer and the class layer can also be obtained by introducing fixed inhibitory connections within the hidden layer and the class layer, respectively [81, 82, 83]. Since at most one hidden unit h_μ can be activated at once, the hidden layer is a very sparse representation of the visible layer. In machine learning, sparsity can improve interpretability [205], generalization, computational efficiency [206], and adversarial robustness [207, 208, 209, 210, 211, 212]. The sparsity of the brain suggests that it is also beneficial for biological neural networks [213].

Given these priors $\mathrm{P}_0(\mathbf{x})$, $\mathrm{P}_0(\mathbf{h})$ and $\mathrm{P}_0(\mathbf{q})$, we derive the marginal distribution of the visible layer $(\mathbf{x}, \mathbf{q})$ (see Appendix 4.A for details). We start our derivation by showing that the conditional distribution of the data layer given the hidden layer has the form

$$\begin{aligned} \mathrm{P}_\beta(\mathbf{x}|\mu, \mathbf{J}) &:= \mathrm{P}_\beta(\mathbf{x}|\mathbf{h} = \mathbf{e}_\mu, \mathbf{J}) \\ &\propto \exp\left(\beta \sum_{i=1}^{N} w_i^\mu x_i\right) \quad \forall\, \mathbf{x} \in S^{N-1} \text{ and } \mu \in \{1, ..., P\}. \end{aligned} \tag{4.2}$$

In other words, the probability density $\mathrm{P}_\beta(\mathbf{x}|\mu, \mathbf{J})$ corresponding to each cluster $\mu > 0$ is a von Mises-Fisher (vMF) distribution centered on the direction $\mathbf{w}^\mu = \{w_i^\mu\}_{i=1}^{N}$ (see Appendix 4.C). In order to interpret the $\mathbf{w}^\mu$ as centroids for their respective clusters, we assume that they belong to S^{N-1} like the data layer. Under this assumption, we find the normalization constant of $\mathrm{P}_\beta(\mathbf{x}|\mu, \mathbf{J})$ to be $\Omega_N(\beta) = \frac{(2\pi)^{N/2} I_{N/2-1}(\beta)}{\beta^{N/2-1}}$, where $I_n(x)$ is the modified Bessel function of the first kind of order n (see Appendix 4.C).

After slightly more work, we find the marginal distribution of the visible layer to be

$$\begin{aligned} \mathrm{P}_\beta(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}) &:= \mathrm{P}_\beta(\mathbf{x}, \mathbf{q} = \mathbf{e}_y \mid \mathbf{J}) \\ &= \sum_{\mu=1}^{P} p_y^\mu \frac{\exp\left(\beta \sum_{i=1}^{N} w_i^\mu x_i\right)}{\Omega_N(\beta)} + p_y^0 \frac{1}{\Omega_N(0)} \quad \forall\, \mathbf{x} \in S^{N-1} \text{ and } y \in \{0, ..., C\}, \end{aligned} \tag{4.3}$$

where $\mathbf{p} = \left\{p_y^\gamma\right\}_{0 \le y \le C}^{0 \le \gamma \le P} = \left\{\mathrm{P}_\beta(\mathbf{q} = \mathbf{e}_y, \mathbf{h} = \mathbf{e}_\gamma|\mathbf{J})\right\}_{0 \le y \le C}^{0 \le \gamma \le P}$. $\Omega_N(0) = \frac{2\pi^{N/2}}{\Gamma(N/2)}$ is the surface area of S^{N-1}, where $\Gamma(x)$ is the Gamma function, so $\frac{1}{\Omega_N(0)}$ is the uniform distribution on S^{N-1}. The uniform distribution term of Eq. (4.3) encourages the model to ignore very noisy data during training, which may or not be desirable depending on the application.

The detailed derivation of Eq. (4.3) (in Appendix 4.A) is inspired by the derivation of Gaussian mixtures from RBMs presented in [72] based on the framework of [81, 82, 83]. In our case, the marginal distribution $\mathrm{P}_\beta(\mathbf{x}|\mathbf{w}, \mathbf{p}) = \sum_{y=0}^{C} \mathrm{P}_\beta(\mathbf{x}, y|\mathbf{w}, \mathbf{p})$ and conditional distributions $\mathrm{P}_\beta(\mathbf{x}|y; \mathbf{w}, \mathbf{p}) = \frac{\mathrm{P}_\beta(\mathbf{x}, y|\mathbf{w}, \mathbf{p})}{\sum_{\gamma=0}^{P} \mathrm{P}_\beta(y, \gamma|\mathbf{w}, \mathbf{p})}$ are convex mixtures between the uniform distribution on S^{N-1} and the vMF distributions $\mathrm{P}_\beta(\mathbf{x}|\mu, \mathbf{J})$ (see Eq. 4.2). Probabilistic models similar to $\mathrm{P}_\beta(\mathbf{x}|\mathbf{w}, \mathbf{p})$ are notably used in text document clustering [203]. Mixture distributions that have class-dependent weights like $\mathrm{P}_\beta(\mathbf{x}|y; \mathbf{w}, \mathbf{p})$ are also used in Gaussian mixture

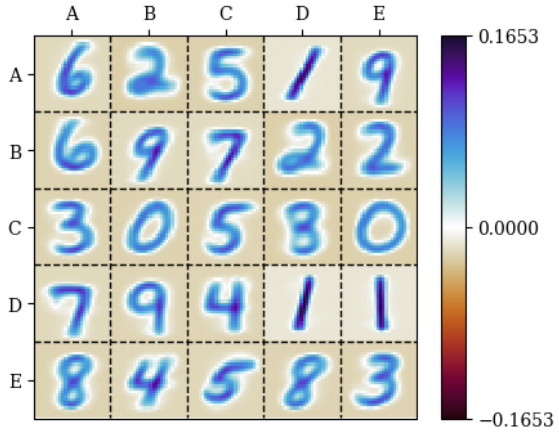

Figure 4.1: All of the $P = 25$ memories $\{\mathbf{w}^\mu\}_{\mu=1}^{25}$ learned by an instance of our model with $\beta = 16$ when it is trained on the MNIST dataset of handwritten digits [8] using constrained stochastic gradient descent (SGD) of the negative log-likelihood loss (Eq. 4.4). The hidden units are indexed using pairs of letters from A to E.

discriminant analysis [214].

The class weights $\mathbf{p}$ depend on the trainable parameters $\mathbf{u}$ and $\mathbf{b}$ of Eq. (4.1) (see Appendix 4.A). Without loss of generality, we choose to directly study (and train) $\mathbf{p}$ instead of $\mathbf{u}$ and $\mathbf{b}$. Recall that p_y^γ is a probability distribution, and in particular $p_y^\gamma \geq 0$ (see Eq. 4.3). We constrain the marginal $\sum_{y=0}^{C} p_y^\gamma$, i.e. the fraction of data in each cluster, to a fixed distribution $p_{\mathbf{h}}(\gamma)$ and the marginal $\sum_{\gamma=0}^{P} p_y^\gamma$, i.e. the proportion of data in each class y, to another fixed distribution $p_{\mathbf{q}}(y)$. In sum, we end up with the constraints $p_y^\gamma \geq 0$, $\sum_{\gamma=0}^{P} p_y^\gamma = p_{\mathbf{q}}(y)$ and $\sum_{y=0}^{C} p_y^\gamma = p_{\mathbf{h}}(\gamma)$. Since each cluster belongs to a single class (Assumption 3), we expect the p_y^γ of a trained model to be close to $\sum_{y'=0}^{C} p_{y'}^\gamma$ for a given y and close to 0 otherwise.

Given a dataset of P^* patterns $\{\mathbf{x}^{*\mu}\}_{\mu=1}^{P^*}$ with soft labels $q_y^{*\mu}$ [215], we can train the weights $\mathbf{w}$ and $\mathbf{p}$ by minimizing the negative log-likelihood loss

$$L(\mathbf{w}, \mathbf{p}) = -\frac{1}{P^*} \sum_{\mu=1}^{P^*} \sum_{y=0}^{C} q_y^{*\mu} \log \mathrm{P}_\beta(\mathbf{x}^{*\mu}, y|\mathbf{w}, \mathbf{p}), \tag{4.4}$$

which is a form of maximum likelihood estimation. We do so using constrained stochastic gradient descent with momentum (simply called SGD in this paper). We explain how the constraints on $\mathbf{w}$ and $\mathbf{p}$ are enforced in Appendix 4.F, and we briefly discuss the initial conditions and the learning rate in Appendix 4.I.

The weights $\mathbf{w}^\mu$ learned by SGD of the loss (Eq. 4.4) [8] are interpretable prototypes, or memories, of the data (see Fig. 4.1 in the case of the MNIST dataset of handwritten digits [8]), which is consistent with their role as cluster centroids. Once the model is trained, we can use its conditionals $\mathrm{P}_\beta(y|\mathbf{x}; \mathbf{w}, \mathbf{p})$ and $\mathrm{P}_\beta(\mathbf{x}|y; \mathbf{w}, \mathbf{p})$ to efficiently reconstruct y from $\mathbf{x}$ or $\mathbf{x}$ from y, respectively. In particular, we can classify unseen patterns using the Bayes classification rule $y = \mathrm{argmax}_{y'} \{\log \mathrm{P}_\beta(y'|\mathbf{x}; \mathbf{w}, \mathbf{p})\}$ and reconstruct patterns of a given class by finding the local minima of the effective energy $-\log \mathrm{P}_\beta(\mathbf{x}|y; \mathbf{w}, \mathbf{p})$ as a function of $\mathbf{x} \in S^{N-1}$.

High-dimensional probabilistic models that store or learn prototypes, such as our model, can typically reconstruct a limited number of patterns with non-trivial accuracy. In other words, they have limited storage

capacity. For instance, Hopfield's model of associative memory, the Hopfield network [23], has a capacity of up to $\mathcal{O}(N)$ patterns [23, 54]. Dense associative memory (DAM) models are a class of Hopfield network-inspired models with an asymptotically much higher capacity [30, 28]. The effective energy $-\log \mathrm{P}_\beta(\mathbf{x}|y;\mathbf{w},\mathbf{p})$ that we can minimize to reconstruct patterns with our model (see Eq. 4.3) is very similar to that of the DAM studied in [38, 67, 80], which also reconstructs patterns by energy minimization. In fact, according to [80], our model belongs to the class of DAMs with exponential capacity.

4.2.2 Teacher-student setting

Among the extensive research on the properties of artificial neural networks (NNs) from the perspective of statistical mechanics [190, 191, 192, 27, 28, 108, 29, 54, 80, 72, 216, 16, 164, 217], there have been many studies of simple RBMs trained by maximum likelihood estimation (see Eq. 4.4) [31, 32, 218, 33, 219, 220] or by averaging samples from the posterior distribution of the weights $\mathbf{J}$ given some observed data $\mathcal{D}$. The latter approach has notably been used to characterize the fundamental limits of RBM learning [145, 146, 39, 101, 147, 42, 43, 44, 46] in the *teacher-student setting* where the data used to train $\mathbf{J}$ is sampled from another RBM with planted weights $\mathbf{J}^*$ [61, 39, 101, 40]. This teacher-student setting can also be used to study our DAM. In this scenario, a *teacher* DAM with weights $\mathbf{w}^*$ and $\mathbf{p}^*$ generates a large amount $M = \alpha N$ of noisy data $\mathcal{D} = \{\mathbf{x}^c, y^c\}_{c=1}^M$ and feeds them to a *student* DAM, which then trains its weights $\mathbf{w}$ and $\mathbf{p}$ by averaging samples from the posterior distribution

$$\mathrm{P}_\beta(\mathbf{w},\mathbf{p}|\mathcal{D}) = \mathcal{Z}_\beta(\mathcal{D})^{-1}\,\mathrm{P}_0(\mathbf{w},\mathbf{p})\prod_{c=1}^{M}\mathrm{P}_\beta(\mathbf{x}^c,y^c|\mathbf{w},\mathbf{p})\,, \tag{4.5}$$

where $\mathcal{Z}_\beta(\mathcal{D}) = \mathbb{E}_{\mathbf{w},\mathbf{p}}\left[\prod_{c=1}^M \mathrm{P}_\beta(\mathbf{x}^c,y^c|\mathbf{w},\mathbf{p})\right]$ is the posterior partition function and $\mathrm{P}_0(\mathbf{w},\mathbf{p})$ is the prior on $\mathbf{w}$ and $\mathbf{p}$, which for simplicity we choose as uniform over the sets in which $\mathbf{w}$ and $\mathbf{p}$ are constrained.

We use a statistical mechanics approach to derive a single set of *saddle-point* equations that simultaneously characterize the weights that are stationary points of maximum likelihood estimation (Eq. 4.4) for generic data and the typical weight configurations obtained by averaging samples from Eq. (4.5) in the teacher-student setting. We assume that the student does not know the number of hidden units P^* and the inverse temperature β^* of the teacher, so it cannot match them with its own. In particular, we consider the case where the noise injected by the teacher in the data is relatively small, i.e. $\beta^*/N > 0$ as $N \to \infty$, while the student chooses a conservative inverse temperature $\beta \ll N$ to avoid overfitting. Moreover, we fix $\sum_{y=0}^C p_y^{*0} = p_{\mathbf{h}}^*(0) = 0$ and $\sum_{y=0}^C p_y^{*\mu} = p_{\mathbf{h}}^*(\mu) = 1/P^*$ for all $\mu > 0$ so that each $\mathbf{g}^{*\mu} := \mathrm{P}_\beta(y|\mu;\mathbf{w},\mathbf{p}) = \mathbf{p}^{*\mu}/p_{\mathbf{h}}^*(\mu) = P^*\mathbf{p}^{*\mu}$ is a soft label for the corresponding $\mathbf{w}^{*\mu}$. On the contrary, we do not give $\sum_{\gamma=0}^{P^*} p_y^{*\gamma} = p_{\mathbf{q}}^*(y)$ a restrictive form. In other words, $p_{\mathbf{q}}^*(y)$ is free to be any given probability mass function.

4.3 Theoretical results

4.3.1 Saddle-point equations

In this Section, we introduce a set of equations for the stationary points of the loss (Eq. 4.4), which we then relate to the saddle-point equations emerging from the statistical mechanics analysis of posterior sampling (see

Eq. 4.5) in the teacher-student setting.

Let us first establish a few definitions that we will use frequently throughout the Section. Given two matrices $\mathbf{w}^* \in \mathbb{R}^{P^* \times N}$ and $\mathbf{w} \in \mathbb{R}^{P \times N}$, we define the overlap matrix $\mathbf{m}(\mathbf{w}^*, \mathbf{w}) = \mathbf{w}^* \mathbf{w}^T \in \mathbb{R}^{P^* \times P}$. We write its entries as $m^{\mu_* \mu}(\mathbf{w}^*, \mathbf{w}) = \sum_{i=1}^N w_i^{*\mu_*} w_i^\mu$ and its row vectors as $m^{\mu_*}(\mathbf{w}^*, \mathbf{w})$, where $1 \leq \mu_* \leq P^*$ and $1 \leq \mu \leq P$. Moreover, for any matrix $\mathbf{m} \in \mathbb{R}^{P^* \times (P+1)}$ with entries $m^{\mu_* \gamma}$ and row vectors m^{μ_*}, we use

$$\sigma_\gamma(m^{\mu_*}) = \frac{\exp(m^{\mu_* \gamma})}{\sum_{\nu=0}^P \exp(m^{\mu_* \nu})}$$

to represent the entry number $\gamma \in \{0, 1, ..., P\}$ of the softmax function applied to the row vector m^{μ_*}. In this context, m^{μ_*} has a zeroth component $m^{\mu_* 0}$, and so does its softmax.

In Appendix 4.B, we show that the stationary points of the negative log-likelihood loss (Eq. 4.4) satisfy

$$\begin{aligned} w_i^\mu &= \frac{\bar{w}_i^\mu}{\sqrt{\sum_{j=1}^N \left[\bar{w}_j^\mu\right]^2}} \\ p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma(\bar{\mathbf{p}}; p_{\mathbf{h}})} \\ \text{with} \quad \bar{w}_i^\mu &= \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*} \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu\left(\beta m^{\mu_*}(\mathbf{x}^*, \mathbf{w}) + \log[\mathbf{p}_y]\right) \\ \bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma\left(\beta m^{\mu_*}(\mathbf{x}^*, \mathbf{w}) + \log[\mathbf{p}_y]\right) \end{aligned} \tag{4.6}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$, where $m^{\mu_* 0}(\mathbf{x}^*, \mathbf{w}) = \frac{1}{\beta} \log\left[\Omega_N(\beta) / \Omega_N(0)\right]$ and the normalization constant $\zeta_y^\gamma(\bar{\mathbf{p}}; p_{\mathbf{h}})$ is defined in Appendix 4.F.

Similar equations arise naturally in our statistical mechanics analysis, which amounts to using the replica method [74, 112] to compute the limiting free entropy

$$f(\varrho, \upsilon, \beta, P^*, P) = \lim_{\beta^*, M, N \to \infty} \frac{1}{N} \mathbb{E}_{\mathbf{w}^*, \mathbf{p}^*, \mathcal{D}} \log\left[\mathcal{Z}_\beta(\mathcal{D})\right], \tag{4.7}$$

where $\upsilon = \beta^*/N$, $\varrho = \frac{M}{P^* N}$ and $\mathbb{E}_{\mathbf{w}^*, \mathbf{p}^*, \mathcal{D}}$ is the joint expectation over the distribution of examples $\mathcal{D} = \{\mathbf{x}^c, y^c\}_{c=1}^M$ generated by the teacher and the priors on the teacher weights. To be more precise, we show that the free entropy can be computed with a variational principle of the form

$$f(\varrho, \upsilon, \beta, P^*, P) = \mathrm{Extr}_{\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}} f(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}), \tag{4.8}$$

whose extremizer $\mathbf{m} = \{m^{\mu_* \mu}\}_{1 \leq \mu \leq P}^{1 \leq \mu_* \leq P^*}$ can be interpreted as the $N \to \infty$ limit of the expected value of the teacher-student overlaps $\mathbf{m}(\mathbf{w}^*, \mathbf{w})$. We show the derivation of the variational principle in Appendix 4.D together with an explicit expression for the trial function $f(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p})$. If we assume that there are $P^* \ll N$ teacher memories and that the priors on $\mathbf{w}^{*\mu_*}$ and $\mathbf{g}^*$ are uniform like those of the student, we find that the

expected teacher-student overlaps must satisfy the saddle-point equations

$$\begin{aligned} m^{\mu_* \mu} &= \varsigma\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{\nu_*=1}^{P^*}\left[\hat{m}^{\nu_* \mu}\right]^2}\right)\frac{\hat{m}^{\mu_* \mu}}{\sqrt{\sum_{\nu_*=1}^{P^*}\left[\hat{m}^{\nu_* \mu}\right]^2}} \\ p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)} \\ \text{with} \quad \hat{m}^{\mu_* \mu} &= \sum_{y=0}^{C} p_{\mathbf{q}}^*(y)\,\sigma_\mu\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\ \bar{p}_y^\gamma &= p_{\mathbf{q}}^*(y)\sum_{\mu_*=1}^{P^*}\sigma_\gamma\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \end{aligned} \tag{4.9}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$, where $\varsigma(x) = \frac{x}{\sqrt{x^2+1}+1}$, $\beta_{\text{eff}} = \varsigma(2\upsilon)\beta$ and $m^{\mu_* 0} = \frac{1}{\beta_{\text{eff}}}\log\left[\Omega_N(\beta)/\Omega_N(0)\right]$.

If we instead clamp the teacher weights $\mathbf{w}^{*\mu_*}$ and $\mathbf{g}^{*\mu_*}$ to fixed patterns $\mathbf{x}^{*\mu_*}$ and their corresponding soft labels $\mathbf{q}^{*\mu_*}$ [215], respectively, to mimic a more general distribution for the data, then Eqs. (4.9) become (see Appendix 4.E)

$$\begin{aligned} m^{\mu_* \mu} &= \varsigma\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{i=1}^{N}\left[\bar{x}_i^\mu\right]^2}\right)\frac{\sum_{i=1}^{N} x_i^{*\mu_*}\bar{x}_i^\mu}{\sqrt{\sum_{i=1}^{N}\left[\bar{x}_i^\mu\right]^2}} \\ p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)} \\ \text{with} \quad \bar{x}_i^\mu &= \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*}\sum_{y=0}^{C} q_y^{*\mu_*}\sigma_\mu\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\ \bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*}\sigma_\gamma\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right). \end{aligned} \tag{4.10}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$, where we recall that $m^{\mu_* 0} = \frac{1}{\beta_{\text{eff}}}\log\left[\Omega_N(\beta)/\Omega_N(0)\right]$.

In the limit of $\varrho, \upsilon \to \infty$, which indicates a large number of examples and a low level of teacher noise, Eqs. (4.10) become equivalent to Eqs. (4.6) if we make the identification $\bar{x}_i^\mu = \bar{w}_i^\mu$, i.e.

$$w_i^\mu = \frac{\bar{x}_i^\mu}{\sqrt{\sum_{j=1}^{N}\left[\bar{x}_j^\mu\right]^2}}. \tag{4.11}$$

Let us explain this limit step by step. For any υ and ϱ, the M examples $\{\mathbf{x}^c\}_{c=1}^M$ provided to the student are corrupted versions of the teacher patterns $\mathbf{x}^*$ with a noise level of $1/\upsilon$. Therefore, in the limit of $\upsilon \to \infty$ with finite ϱ, each example $\mathbf{x}^c$ provided to the student is one of the original patterns $\mathbf{x}^*$, as in Eqs. (4.6). However, even in this case, the empirical distribution of the examples deviates from that of $\mathbf{x}^*$, of which it is merely a bootstrap sample. At fixed P^* and P, this mismatch disturbs the accurate learning of $\mathbf{x}^*$ when $\varrho = \frac{M}{P^* N}$ is finite, or equivalently the expected number of repetitions M/P^* of each pattern is not sufficiently large

compared to N (see Eq. 4.10). This influence progressively weakens as ϱ grows and the ς function approaches 1, reflecting the convergence of the empirical distribution of the examples to that of the teacher patterns.

In the limit of $\varrho \to \infty$ with finite υ, Eq. (4.10) is still very similar to Eq. (4.6). The only difference between them is that Eq. (4.10) has β_{eff} instead of β in the argument of σ and in the denominator of the definition of $m^{\mu_* 0}$. Using ς as a shorthand for $\varsigma\left(2\upsilon\right)$, we find that the fixed points of Eqs. (4.10) with $\varrho \to \infty$ are related, through the identification made by Eq. (4.11), to the stationary points of the effective loss

$$\mathcal{L}\left(\mathbf{w}, \mathbf{p}\right) = -\frac{1}{P^*} \sum_{\mu=1}^{P^*} \sum_{y=0}^{C} q_y^{*\mu} \log \mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}^{*\mu}, y | \mathbf{w}, \mathbf{p}\right), \tag{4.12}$$

$$\text{where} \quad \mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right) = \sum_{\mu=1}^{P} p_y^{\mu} \frac{\exp\left(\varsigma\beta \sum_{i=1}^{N} w_i^{\mu} x_i\right)}{\Omega_N\left(\beta\right)} + p_y^0 \frac{1}{\Omega_N\left(0\right)}.$$

What distinguishes this equation from the standard loss (Eq. 4.4) is that $\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right)$ has $\beta_{\text{eff}} = \varsigma\beta$ in the argument of the exponential function instead of β. As a consequence, $\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right)$ is not a probability distribution unless $\varsigma = 1$, in the limit of $\upsilon \to \infty$ (see Appendix 4.C). A value of ς less than one in the effective loss (Eq. 4.12) is reminiscent of the presence of noise in the data generation process, so we propose to use it as a regularizer for the weights. We discuss this point in more detail in Section 4.4.1.

4.3.2 Saddle-point hierarchy

As shown in [168], the loss landscape of any NN with unconstrained weights contains the stationary points of narrower NNs with the same architecture. In Appendix 4.G, we show that this result also applies to the teacher-student setting with $\varrho \to \infty$ and any non-zero υ. To be more precise, we show that, if the parameters $\bar{x}_i^{\text{fixed},\mu}$, $\bar{p}_y^{\text{fixed},\gamma}$, $m^{\text{fixed},\mu_*\gamma}$, $p_y^{\text{fixed},\gamma}$ with hidden unit prior $p_{\mathbf{h}}^{\text{given}}\left(\gamma\right)$ are a fixed point of Eqs. (4.10) with P hidden units, then the duplicated parameters

$$\bar{x}_i^{\text{dupli},\mu} = \begin{cases} \bar{x}_i^{\text{fixed},\mu} & 0 < \mu \leq P \\ \bar{x}_i^{\text{fixed},\mu-P} & P < \mu \leq P + R \end{cases}$$

$$\bar{p}_y^{\text{dupli},\gamma} = \begin{cases} \bar{p}_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2}\bar{p}_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ \bar{p}_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2}\bar{p}_y^{\text{fixed},\gamma-P} & P < \gamma \leq P + R \end{cases}$$

$$m^{\text{dupli},\mu_*\gamma} = \begin{cases} m^{\text{fixed},\mu_* 0} & \gamma = 0 \\ m^{\text{fixed},\mu_*\gamma} & 0 < \gamma \leq P \\ m^{\text{fixed},\mu_*,\gamma-P} & P < \gamma \leq P + R \end{cases} \tag{4.13}$$

$$p_y^{\text{dupli},\gamma} = \begin{cases} p_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2} p_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ p_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2} p_y^{\text{fixed},\gamma - P} & P < \gamma \leq P + R \end{cases}$$

$$\text{along with} \quad p_{\mathbf{h}}(\gamma) = \begin{cases} p_{\mathbf{h}}^{\text{given}}(0) & \gamma = 0 \\ \frac{1}{2} p_{\mathbf{h}}^{\text{given}}(\gamma) & 0 < \gamma \leq R \\ p_{\mathbf{h}}^{\text{given}}(\gamma) & R < \gamma \leq P \\ \frac{1}{2} p_{\mathbf{h}}^{\text{given}}(\gamma - P) & P < \gamma \leq P + R, \end{cases}$$

are a fixed point of the same saddle-point equations with $P + R \in \{P, ..., 2P\}$ hidden units (See Appendix 4.G for a detailed derivation). In other words, we duplicate some of the weights solving Eqs. (4.10) to construct a fixed point for a wider network. In that sense, wide DAMs contain the fixed points of narrower DAMs. In particular (see Section 4.3.1), this property also holds for the stationary points of both the standard loss (Eq. 4.4) and the effective loss (Eq. 4.12).

The saddle-point equations (Eq. 4.10) with duplicated order parameters (Eqs. 4.13) are invariant to the permutation of any hidden unit $\gamma \in \{1, ..., R\}$ and its duplicate $\gamma + P$. This kind of symmetry can be spontaneously broken if it leads to a higher free entropy (see Eqs. 4.7 and 4.22). However, symmetry-breaking transitions can be prohibitively slow when the symmetric state is stable to local perturbations [60]. Interestingly, the DAM introduced in [30] quickly undergoes many successive permutation symmetry-breaking bifurcations during training [34]. This observation and additional empirical evidence suggest that the symmetric states are unstable, or in other words that they are saddles, which was verified analytically for DAMs with only two data points to memorize [34]. In Appendix 4.G, we prove that, if β is large enough, Eqs. (4.13) is a saddle, which is a major step toward explaining why permutation symmetry-breaking transitions are relatively fast in DAMs trained with a large number of data points. We call this result *the saddle-point hierarchy principle*.

4.4 Empirical results

In this Section, we use our theoretical results to improve training, and we show empirically that our DAM learns interpretable solutions to both supervised and unsupervised classification problems. We perform all our numerical experiments on the MNIST dataset of handwritten digits [8]. The code and hyperparameter values used in our numerical experiments are available at this public repository [199].

4.4.1 Learning by minimizing the effective loss

As the number of hidden units P increases, standard maximum likelihood estimation with fixed inverse temperature β becomes progressively less apt to train our DAM to its full potential. At high P and β, many memories stay stuck in noisy states that do not contribute to classification, which is wasteful. Using a lower β helps the memories converge, but also reduces their resolution and diversity, possibly because it also lowers the

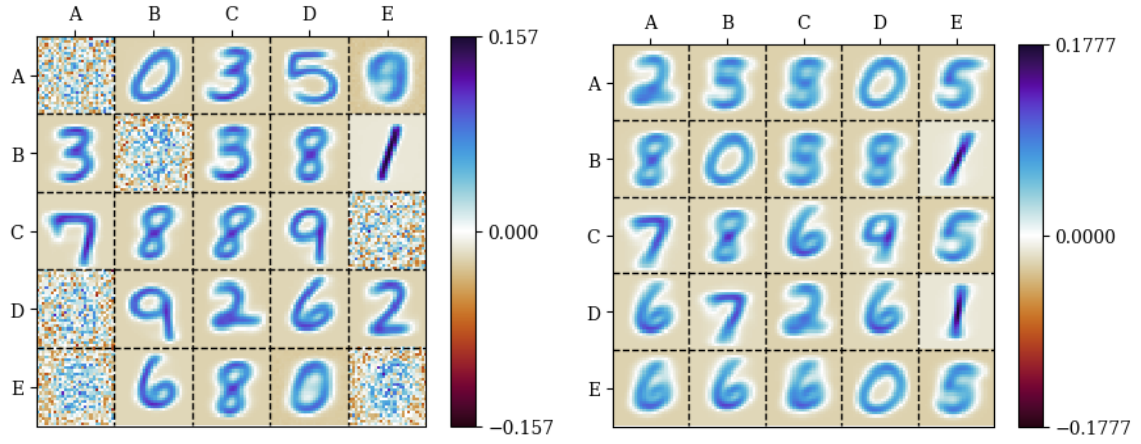

Figure 4.2: 25 of the $P = 1000$ memories $\mathbf{w}^\mu$ learned by two instances of our dense associative memory (DAM) model with different values of β. Both networks are trained on the MNIST dataset of handwritten digits [8] using constrained stochastic gradient descent (SGD) of the negative log-likelihood loss (Eq. 4.4). The left-panel model has $\beta = 18$, and the right-panel one $\beta = 6$. DAMs with $18 > \beta > 6$ learn memories that interpolate between these two pictures. The hidden units are indexed using pairs of letters from A to E.

DAM's capacity [80]. This change is far from being only cosmetic. In fact, it comes with a gradual reduction in classification accuracy (see Table 4.1).

Optimization algorithms with a parameter analogous to β often converge to better solutions when this parameter is increased from a small value during optimization. For example, annealing schedules [221, 84] can be incorporated into minimization algorithms to help them find deeper local minima than they could otherwise. This observation suggests that our DAM would learn better memories if we increased β during training. It is tempting to do so by SGD of Eq. (4.4) with respect to β, but it makes β increase so quickly that many memories still stay stuck in noisy states.

We find that optimizing the effective loss (Eq. 4.12) with a suitable ς slows down the evolution of β enough to let the DAM learn much cleaner memories (see Fig. 4.3, top panel) and achieve much better classification accuracy (see Table 4.1) than with standard training (Eq. 4.4). Looking back at Section 4.3.1, we propose to interpret ς as a regularization parameter that helps the DAM take noise from the data into account during training. There is no obvious theoretically motivated way to find the best value of ς for a generic dataset, so we choose it by hand. Despite this limitation, we believe that the simplicity and interpretability of our method still make it an interesting alternative to annealing schedules.

When we use this training method, we compute the DAM classification accuracy from the effective predictions

$$\mathcal{P}_{\beta,\varsigma}\left(y \mid \mathbf{x}; \mathbf{w}, \mathbf{p}\right) = \frac{\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right)}{\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x} \mid \mathbf{w}, \mathbf{p}\right)} \tag{4.14}$$

$$\text{where} \quad \mathcal{P}_{\beta,\varsigma}\left(\mathbf{x} \mid \mathbf{w}, \mathbf{p}\right) = \sum_{y=0}^{C} \mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right),$$

instead of the true predictions $\mathrm{P}_\beta\left(y \mid \mathbf{x}; \mathbf{w}, \mathbf{p}\right) = \mathrm{P}_\beta\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right) / \mathrm{P}_\beta\left(\mathbf{x} \mid \mathbf{w}, \mathbf{p}\right)$. This approach allows us

Inverse temperature β	Classification accuracy
6	79%
10	85%
14	89%
18	91%
Trained with $\varsigma = 0.25$	96%

Table 4.1: DAM classification accuracy (rounded down to two significant figures) for $P = 1000$ and various values of β. The last line is for β trained by SGD of the effective loss (Eq. 4.12) with $\varsigma = 0.25$.

to calculate both the accuracy and the loss through a single evaluation of $\mathcal{P}_{\beta,\varsigma}\left(y \mid \mathbf{x}; \mathbf{w}, \mathbf{p}\right)$, which is more efficient than computing $\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right)$ and $\mathrm{P}_{\beta}\left(\mathbf{x}, y \mid \mathbf{w}, \mathbf{p}\right)$ separately when monitoring the progress of training.

4.4.2 Dense associative memory is interpretable, even in unsupervised classification

Now that we understand how to train our DAM reasonably well, we investigate one of the most interesting properties of the solutions that it learns: their interpretability. As advertised in the Introduction, we will explain how to further improve training in the following Section. We already mentioned that the regularization parameter ς of the effective loss is interpretable (see Section 4.4.1 and Eq. 4.12). Here, we point out that the learned weights are as well. In fact, each learned $\mathbf{p}^{\mu}/p_{\mathbf{h}}\left(\mu\right)$ can be interpreted as a soft label for the corresponding $\mathbf{w}^{\mu}$ (see Fig. 4.3). This property was also observed by [34] in K & H's DAM for pattern classification [30].

At test time, we observe that our DAM classifies approximately 98% of the test data points into the class $y = \mathrm{argmax}_{y'} \left\{p^{\mu}_{y'}\right\}$ of the memory $\mathbf{w}^{\mu}$ to which they are the most similar. In other words, a 1-nearest neighbor classifier [222] conditioned on the memories $\mathbf{w}^{\mu}$ and their soft soft labels $\mathbf{p}^{\mu}/p_{\mathbf{h}}\left(\mu\right)$ approximates the classification of our model with 98% fidelity. This behavior is reminiscent of K & H's DAM [30], where only a few memories participate in the classification of each data point. The 1-nearest neighbor classifier approximates DAM classification more faithfully for correctly classified data points (approximately 99% fidelity) than incorrectly classified data points (approximately 70% fidelity).

We now show that the notion of interpretability described in this Section extends beyond supervised learning. Given a dataset of P^* unlabeled patterns $\{\mathbf{x}^{*\mu}\}_{\mu=1}^{P^*}$, we find that we can train our DAM for unsupervised classification by replacing the soft labels $q^{*\mu}_{y}$ in the effective loss (Eq. 4.12) with the softened DAM predictions $\left(1-\varepsilon\right)\mathcal{P}_{\beta,\varsigma}\left(y|\mathbf{x}^{*\mu}; \mathbf{w}, \mathbf{p}\right) + \varepsilon\frac{1}{C+1}$, where $\varepsilon \in [0,1]$ and $\mathcal{P}_{\beta,\varsigma}\left(y|\mathbf{x}^{*\mu}; \mathbf{w}, \mathbf{p}\right)$ is defined in Eq. (4.14). In this scenario, we are minimizing

$$\mathcal{L}_{\text{unsup}}\left(\mathbf{w}, \mathbf{p}\right) = -\frac{1}{P^*}\sum_{\mu=1}^{P^*}\sum_{y=0}^{C}\left[\left(1-\varepsilon\right)\mathcal{P}_{\beta,\varsigma}\left(y|\mathbf{x}^{*\mu}; \mathbf{w}, \mathbf{p}\right) + \varepsilon\frac{1}{C+1}\right]\log\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}^{*\mu}, y|\mathbf{w}, \mathbf{p}\right). \tag{4.15}$$

As before, we do so using constrained SGD (see Appendix 4.F) with the initialization and learning rate

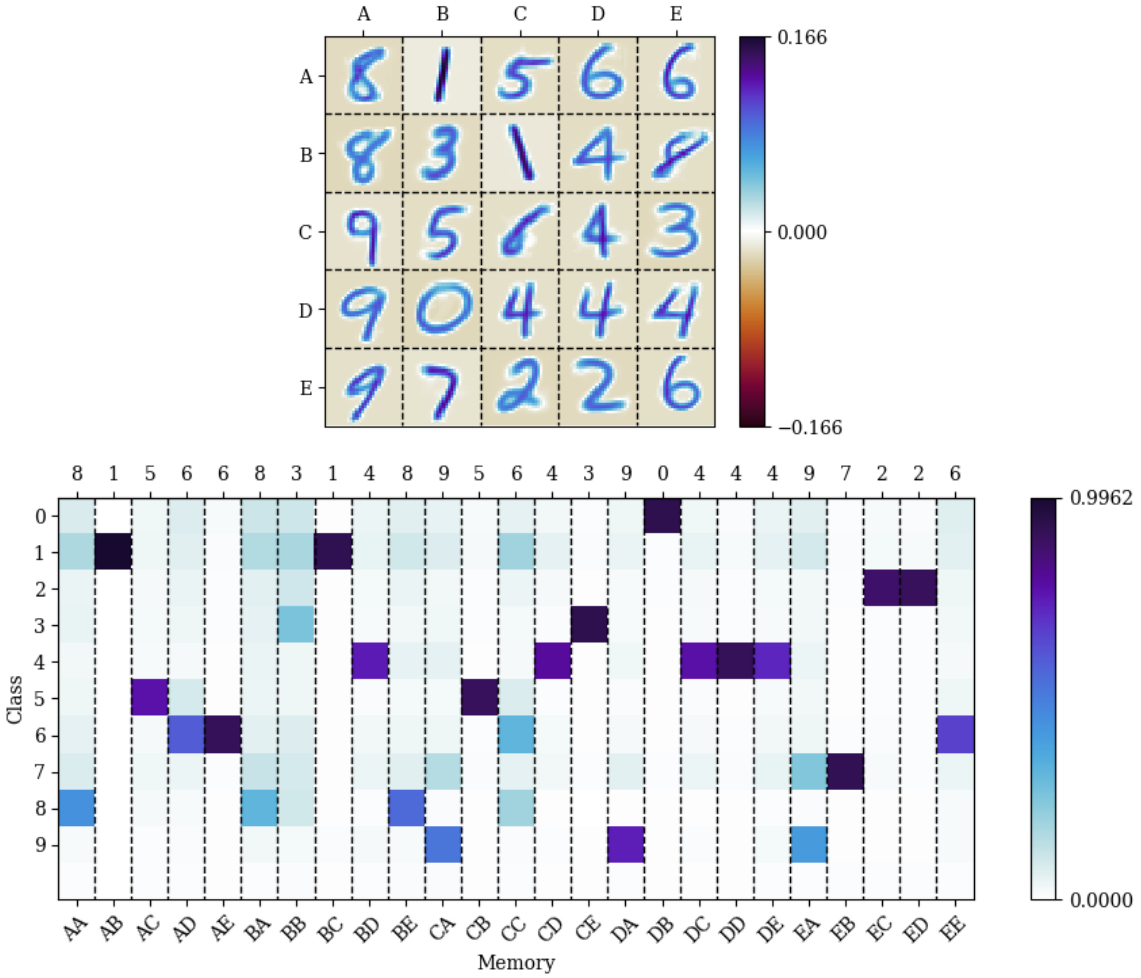

Figure 4.3: In the top panel, 25 of the $P = 1000$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained on the MNIST dataset of handwritten digits [8] using constrained stochastic gradient descent (SGD) of the effective loss (Eq. 4.4) with $\varsigma = 0.25$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices. Rescaled class weights learned with $p_{\mathbf{h}}(\gamma) \neq \frac{1}{P+1}$ are qualitatively similar to the ones shown in this figure. Approximately 98% of test digits fed to the DAM are given the class of the memory that resembles them the most. For example, a digit that looks like the memory #AA is given the class 8.

described in Appendix 4.I. Equivalently, we can also view this algorithm as minimizing the combined loss

$$\mathcal{L}_{\text{total}}(\mathbf{w}, \mathbf{p}) = \mathcal{L}_{\text{margin}}(\mathbf{w}, \mathbf{p}) + \lambda \mathcal{L}_{\text{cond}}(\mathbf{w}, \mathbf{p}), \tag{4.16}$$

$$\text{where} \quad \mathcal{L}_{\text{margin}}(\mathbf{w}, \mathbf{p}) = -\frac{1}{P^*} \sum_{\mu=1}^{P^*} \log \mathcal{P}_{\beta,\varsigma}(\mathbf{x}^{*\mu} | \mathbf{w}, \mathbf{p})$$

$$\text{and} \quad \mathcal{L}_{\text{cond}}(\mathbf{w}, \mathbf{p}) = -\frac{1}{P^*} \sum_{\mu=1}^{P^*} \sum_{y=0}^{C} \mathcal{P}_{\beta,\varsigma}(y | \mathbf{x}^{*\mu}; \mathbf{w}, \mathbf{p}) \log \mathcal{P}_{\beta,\varsigma}(y | \mathbf{x}^{*\mu}; \mathbf{w}, \mathbf{p}),$$

where $\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}^{*\mu}|\mathbf{w},\mathbf{p}\right) = \sum_{y=0}^{C}\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}^{*\mu},y|\mathbf{w},\mathbf{p}\right)$ (see Eq. 4.14). $\mathcal{L}_{\text{cond}}\left(\mathbf{w},\mathbf{p}\right)$ is called the minimum entropy regularization term in unsupervised machine learning [223]. Intuitively, it encourages the DAM to learn different class weights p_y^γ for each class y, which is not possible by minimizing only $\mathcal{L}_{\text{margin}}\left(\mathbf{w},\mathbf{p}\right)$ because $\mathcal{P}_{\beta,\varsigma}\left(\mathbf{x}^{*\mu}|\mathbf{w},\mathbf{p}\right)$ does not depend on y. Exploiting this characteristic, we train our DAM on patches of MNIST digits [8] and find that it learns reasonable latent classes $y = \text{argmax}_{y'}\left\{p_{y'}^\mu\right\}$ for the memories $\mathbf{w}^\mu$ (see Fig. 4.4 and Figs. 4.7, 4.8, 4.9 and 4.10 of Appendix 4.J). This approach could potentially be useful for feature extraction [224].

4.4.3 Fast training with splitting steepest descent

We now explain how to further improve the training method of Section 4.4.1. More specifically, we use the saddle-point hierarchy derived in 4.3.2 to accelerate training.

Machine learning models that experience the permutation symmetry-breaking bifurcations described in Section 4.3.2 during training—such as the DAM studied in [34], RBMs with binary units [42, 46] and Gaussian mixtures [72, 81, 82, 83, 84, 85, 86]—have characteristic tree-shaped learning dynamics. Intuitively, permutation symmetry-breaking transitions are points where model parameters differentiate from each other, so they correspond to the bifurcations of the tree. In the left panel of Fig. (4.5), we show that the learning dynamics of our DAM also follows a tree of permutation symmetry-breaking transitions.

The saddle-point hierarchy principle introduced in Section 4.3.2 suggests the following idea to accelerate learning: train a relatively narrow DAM for cheap, then repeatedly duplicate (or "split") some of the hidden units μ, escape the corresponding saddle point and continue training. Intuitively, we expect to save a lot of computing resources by using relatively few hidden units at the start of training, when the learning dynamics follows a few branches close to the root of the learning dynamics tree. The splitting steepest descent algorithm introduced in [87] formalizes this idea in an efficient and theoretically motivated way. We implement a variant of this algorithm that takes the constraint $\mathbf{w}^\mu \in S^{N-1}$ into account, but is otherwise very similar to the original version (see Alg. 1), with fast splitting implemented as in [88]. The constraint $\mathbf{w}^\mu \in S^{N-1}$ only matters in steps 5 and 11, which are also arguably the most conceptually difficult parts of Alg. (1), so we explain them in detail in Appendix 4.H. Fig. (4.5) shows that the learning dynamics tree of splitting steepest descent has a more sparsely populated trunk than that of SGD without splitting (see Fig. 4.5). In other words, by using a relatively small number P of memories at the beginning of training, we reduce the total number of values $P \times T$ that they take over a period of time T, which is consistent with our intuition that it can save computational resources. The situation could have been different if using fewer memories slowed down training.

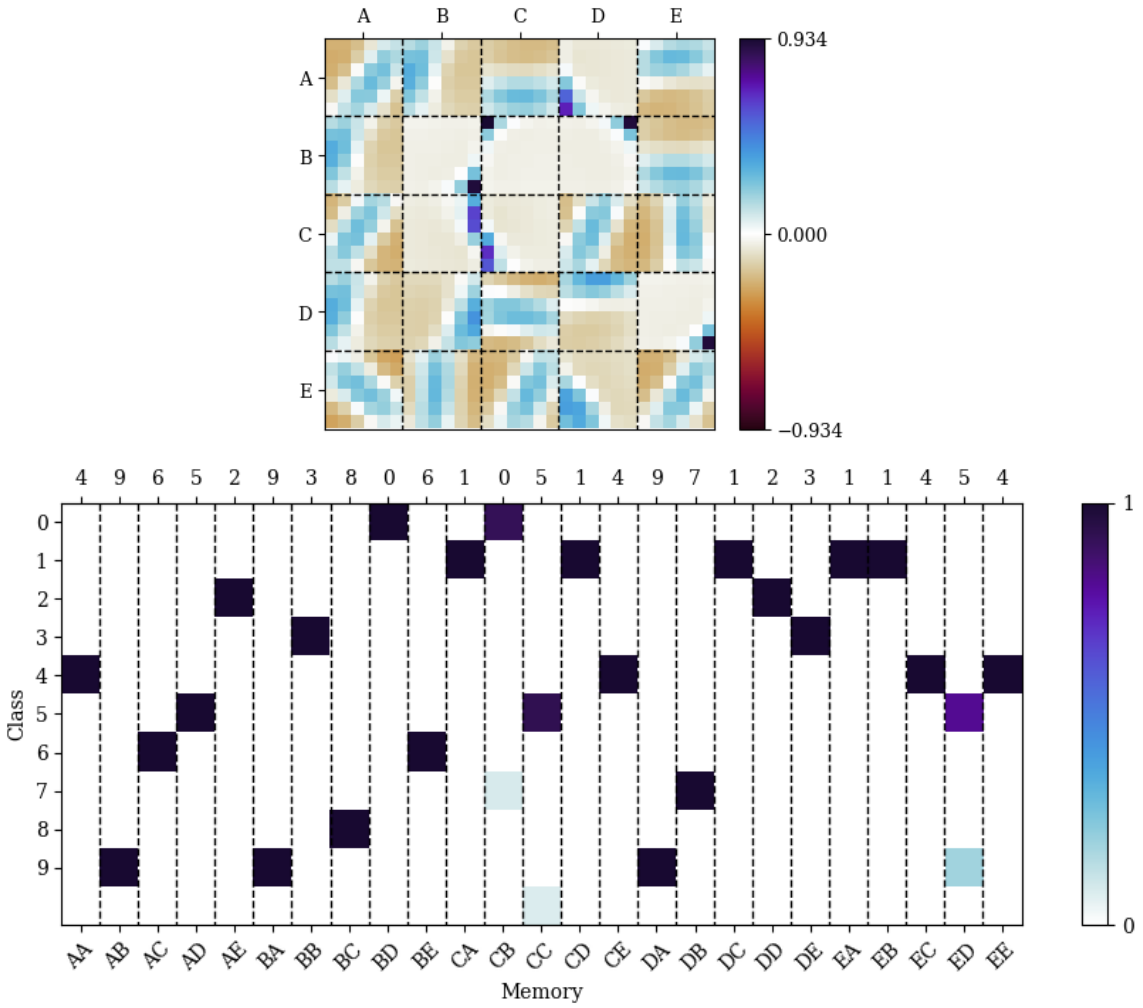

Figure 4.4: In the top panel, 25 of the $P = 100$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained in an unsupervised way (Eq. 4.15) on 6×6 patches of the MNIST dataset of handwritten digits [8] while assuming $C = 10$ latent classes and $\varsigma = 0.6$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices.

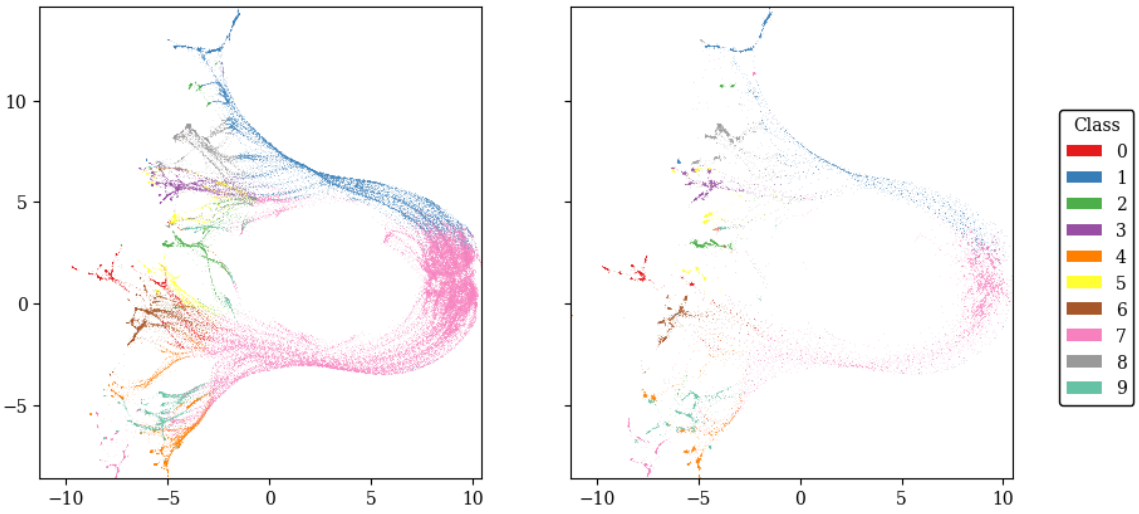

Figure 4.5: Overlaps $m^{\mu_* \mu}\left(\mathbf{x}^*, \mathbf{w}\right) = \sum_{i=1}^{N} x_i^{*\mu_*} w_i^{\mu}$ between the first 1000 digits $\mathbf{x}^{*\mu_*} = \left\{x_i^{*\mu_*}\right\}_{i=1}^{N}$ of the MNIST training set [8] and the memories $\mathbf{w}^{\mu} = \left\{w_i^{\mu}\right\}_{i=1}^{N}$ of our dense associative memory (DAM) model while it is learning them. Each point is one of the high-dimensional magnetization vectors $m_{\mu}\left(\mathbf{x}^*, \mathbf{w}\right) = \left\{m^{\mu_* \mu}\left(\mathbf{x}^*, \mathbf{w}\right)\right\}_{\mu_*=1}^{P^*}$ projected onto a two-dimensional plane using the UMAP algorithm [225]. On the left, the $m_{\mu}\left(\mathbf{x}^*, \mathbf{w}\right)$ found during training without splitting steepest descent. On the right, the $m_{\mu}\left(\mathbf{x}^*, \mathbf{w}\right)$ found with splitting steepest descent (Alg. 1). In both cases, UMAP is trained on the $m_{\mu}\left(\mathbf{x}^*, \mathbf{w}\right)$ found without splitting steepest descent. The classes $y = \text{argmax}_{y'} \left\{p_{y'}^{\mu}\right\}$ of the memories $\mathbf{w}^{\mu}$ are color-coded in the legend. 7 and 1 are the most numerous classes, so they have the top two largest entries in $\sum_{\gamma=0}^{P} p_y^{\mu} = p_{\mathbf{q}}\left(y\right)$ (see Appendix 4.I), which is why the memories are classified as either 7 or 1 at the beginning of training.

Algorithm 1 Splitting steepest descent [87, 88]

1: Preallocate space for a DAM with P_{max} hidden units and the corresponding weights $\mathbf{w}$ and $\mathbf{p}$
2: Initialize the weights $\mathbf{w}^{\mu}$ and $\mathbf{p}^{\mu}$ connected to the P_{cur} first hidden units $\mu \in \{1, ..., P_{\text{cur}}\}$, as well as $\mathbf{p}^{0}$
3: min $L(\mathbf{w}, \mathbf{p})$ with SGD
4: **while** $P_{\text{cur}} < P_{\text{max}}$ **do**
5: Identify a subset $\boldsymbol{\mu}_{\text{copy}} \subseteq \{1, ..., P_{\text{cur}}\}$ of $R \leq P_{\text{max}} - P_{\text{cur}}$ hidden units to split, **return** if empty
6: Let $\boldsymbol{\mu}_{\text{paste}} = \{P_{\text{cur}} + 1, ..., P_{\text{cur}} + R\}$
7: Build weights $\mathbf{w}^{\boldsymbol{\mu}_{\text{paste}}} = \mathbf{w}^{\boldsymbol{\mu}_{\text{copy}}}$ for $\boldsymbol{\mu}_{\text{paste}}$
8: Rescale $\mathbf{p}^{\boldsymbol{\mu}_{\text{copy}}} \leftarrow \mathbf{p}^{\boldsymbol{\mu}_{\text{copy}}}/2$ and $p_{\mathbf{h}}(\boldsymbol{\mu}_{\text{split}}) \leftarrow p_{\mathbf{h}}(\boldsymbol{\mu}_{\text{split}})/2$
9: Build weights $\mathbf{p}^{\boldsymbol{\mu}_{\text{paste}}} = \mathbf{p}^{\boldsymbol{\mu}_{\text{copy}}}$ and $p_{\mathbf{h}}(\boldsymbol{\mu}_{\text{paste}}) = p_{\mathbf{h}}(\boldsymbol{\mu}_{\text{copy}})$ for $\boldsymbol{\mu}_{\text{paste}}$
10: Update $P_{\text{cur}} \leftarrow P_{\text{cur}} + R$
11: Escape the saddle point by 2^{nd} order descent of $L(\mathbf{w}, \mathbf{p})$ w.r.t. $\mathbf{w}$
12: min $L(\mathbf{w}, \mathbf{p})$ with SGD
13: **end while**
14: **return** ▷ See Appendix 4.H and [87, 88] for details about steps 5 and 11

We now establish a methodology that we will use to compare the training times of DAMs trained using the effective loss (Eq. 4.12) on MNIST [8] with and without splitting steepest descent. It is more interesting and meaningful to compare the training times of NNs with similar performance. Therefore, we pick hyperparameters such that DAMs trained with and without splitting have similar classification accuracy. For that purpose, we find the general region of hyperparameter space where DAMs trained without splitting steepest descent have the best classification accuracy. The accuracy does not change much in this region, so we pick generic hyperparameters inside. Next, we manually tune the hyperparameters of splitting steepest descent so that the resulting DAMs have comparable accuracy. Note that we did not find hyperparameters that make the accuracy with splitting steepest descent significantly better than without it.

Once the hyperparameters are set, we collect statistics of the accuracy and training time. We run our experiment on a CPU and manually set the seed of pseudorandom number generation to make training deterministic and reproducible. Two DAMs with the same hyperparameters and the same seed are guaranteed to be trained using the same number of computer operations and to have the same classification accuracy after training. However, background processes and other external factors can change between two runs with the same seed, which adds noise to the bare training time that we want to measure. As such, we approximate the bare training time as the minimum over many runs with the same seed. We then calculate the average and standard deviation of the classification accuracy and bare training time over multiple seeds.

In Fig. (4.6), we compare DAM training time and accuracy with and without splitting steepest descent, using round markers and error bars to show their respective means and standard deviations as a function of the maximum number of hidden units P_{max}. Splitting stops at P_{max} hidden units, and DAMs without splitting have P_{max} hidden units from the start. All hidden units are split in each while-loop iteration of Alg. (1), except for the last iteration, where exactly $P_{\text{max}} - P_{\text{cur}}$ hidden units are split. As wanted, DAMs have similar accuracy with and without splitting (left panel). Moreover, the relatively small error bars of the training time indicate that the residual noise of the background processes is small and that the seed has a limited impact on the training speed (right panel). Splitting steepest descent has a significant speed advantage that scales very advantageously with P_{max} (right panel). Without splitting, the training time is proportional to P_{max}. With splitting, the training time is almost constant. The clearest sign that it increases with P_{max} is the small jump

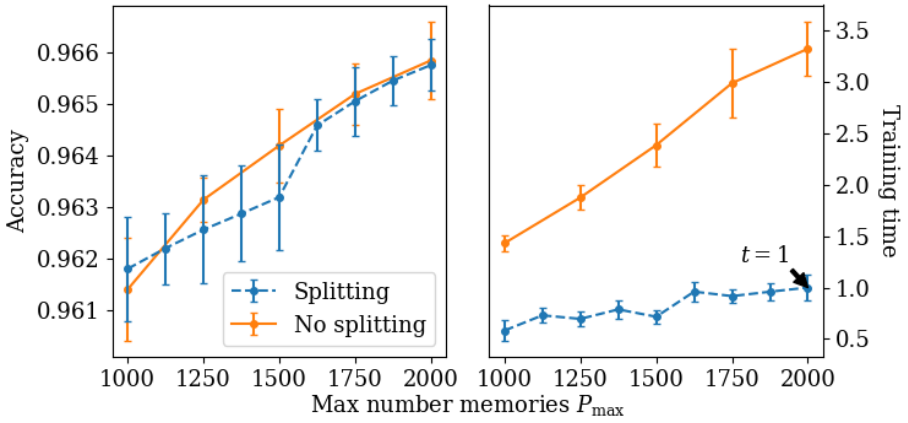

Figure 4.6: The classification accuracy and training time of dense associative memory (DAM) networks trained with and without splitting as a function of the maximum number of hidden unit $P_{\max}$. The round markers and their corresponding error bars show the means and standard deviations of the measurements made at each $P_{\max}$. Statistics are collected from 10 different random seeds. In the left panel, we verify that the DAMs with and without splitting have a similar accuracy. In the right panel, we compare their training times t. To facilitate this comparison, we normalize the t-axis such that the data point indicated by the arrow is at $t = 1$.

between $P_{\max} = 1500$ and $P_{\max} = 1625$, where the number of while-loop iterations in Alg. (1) increases by 1. In other words, the run time of Alg. (1) is better explained as proportional to $\lceil \log P_{\max} \rceil$ than $P_{\max}$ in the range of $P_{\max}$ of our numerical experiment, which is a dramatic improvement over the run time without splitting.

4.5 Conclusion

In this work, we study a dense associative memory (DAM) model capable of learning interpretable solutions to both supervised and unsupervised problems, which we construct from the framework of Boltzmann machines.

We derive two sets of equations that respectively characterize the stationary points of DAMs trained on real data using maximum likelihood estimation and the fixed points of DAMs trained on synthetic data in the teacher-student setting. Guided by their similarity, we then establish that the maximum likelihood equations are a special case of the teacher-student equations. Building on this equivalence, we introduce an effective loss function that significantly stabilizes training thanks to a regularization parameter that mimics the noise present in the data generation process.

We show that the stationary points of the effective loss of DAMs with a given number of hidden units are saddle points of larger DAMs. We considerably accelerate training using a custom implementation of the splitting steepest descent network-growing algorithm [87, 88] inspired by this *saddle-point hierarchy principle*. In our numerical experiments, the training time of splitting steepest descent is approximately logarithmic in the number of hidden units, which is a huge improvement over the linear training time that we observe without splitting steepest descent. Further research is needed to clearly determine the asymptotic functional form of the training time.

In Section 4.4.1, we briefly discuss how our DAM is interpretable. The ingredients needed to make

an interpretable neural network are not well understood, and we believe that our DAM offers a promising opportunity to explore that area. Studying the teacher-student setting with teacher inverse temperature β^* less than order N is another interesting research direction that we leave to future work. Beyond the framework of our model, [226, 227] have designed a technique that trains Gaussian mixture models by replicating mixture components. In contrast with our work, it uses a handpicked cooling schedule and breaks permutation symmetries with random noise. By design, splitting steepest descent finds the directions in the loss landscape that allow NNs to escape saddle points with permutation symmetries as quickly as possible, but finding these directions is more expensive than generating random noise (see Appendix 4.H and [87, 88]). It would be interesting to study this tradeoff and identify the regimes where each method is preferable over the other. More generally, one could investigate how permutation symmetry-breaking takes place in energy models related to DAMs, such as attentional BMs [67], and study its relationship with the permutation symmetry-breaking [42, 46] and dynamical transitions [31, 32, 228, 33, 35] found in other types of RBMs. [34, 229] have made connections between DAM learning dynamics and cellular differentiation. Similarly, it could be possible to investigate whether there are similarities between cellular division and splitting steepest descent in our model. Finally, it would be interesting to see if splitting steepest descent and our proposed regularization technique can be used to improve the training of transformers and generative diffusion.

Data availability

The code for the numerical experiments presented in this work is available at this public repository [199].

Acknowledgements

I am grateful to Carlo Lucibello and Matteo Negri for insightful discussions.

4.A Derivation of the model

Consider the RBM Hamiltonian

$$-\beta H\left[\mathbf{x}, \mathbf{q}, \mathbf{h}; \mathbf{J}\right] = \beta \sum_{\mu=1}^{P} h_\mu \sum_{i=1}^{N} w_i^\mu x_i + \beta \sum_{\mu=1}^{P} h_\mu \sum_{y=1}^{C} u_y^\mu q_y + \beta \sum_{\mu=1}^{P} h_\mu b^\mu .$$

where $\mathbf{J} = \{\mathbf{w}, \mathbf{u}, \mathbf{b}\}$ is the set of all RBM parameters, i.e. $\mathbf{w} = \{w_i^\mu\}_{1\leq i\leq N}^{1\leq \mu\leq P}$, $\mathbf{u} = \{u_i^\mu\}_{1\leq i\leq C}^{1\leq \mu\leq P}$ and $\mathbf{b} = \{b^\mu\}_{\mu=1}^{P}$. The prior densities $\mathrm{P}_0\left(\mathbf{h}\right)$ and $\mathrm{P}_0\left(\mathbf{q}\right)$ are non-zero only when $\mathbf{h} \in \{\mathbf{e}_\gamma\}_{\gamma=0}^{P}$ and $\mathbf{q} \in \{\mathbf{e}_y\}_{y=0}^{C}$, in which case we have

$$-\beta H\left[\mathbf{x}, \mathbf{e}_y, \mathbf{e}_\gamma; \mathbf{J}\right] = \begin{cases} 0 & \gamma = 0 \\ \beta \sum_{i=1}^{N} w_i^\gamma x_i + a_y^\gamma & 0 < \gamma \leq P \end{cases}$$

$$\text{where} \quad a_y^\mu = \begin{cases} \beta b^\mu & y = 0 \\ \beta u_y^\mu + \beta b^\mu & 0 < y \leq C \end{cases}$$

For convenience, we write the corresponding Gibbs distribution $\mathrm{P}_\beta\left(\mathbf{x}, \mathbf{q}=\mathbf{e}_y, \mathbf{h}=\mathbf{e}_\gamma|\mathbf{J}\right)$ as $\mathrm{P}_\beta\left(\mathbf{x}, y, \mu|\mathbf{J}\right)$, and the priors $\mathrm{P}_0\left(\mathbf{q}=\mathbf{e}_y\right)$ and $\mathrm{P}_0\left(\mathbf{h}=\mathbf{e}_\gamma\right)$ as $\pi_{\mathbf{q}}\left(y\right)$ and $\pi_{\mathbf{h}}\left(\gamma\right)$, respectively. Given a uniform prior on $\mathbf{x}$, we find

$$\mathrm{P}_\beta\left(\mathbf{x}, y, \mu|\mathbf{J}\right) = \frac{1}{Z_\beta\left(\mathbf{J}\right)}\pi_{\mathbf{q}}\left(y\right)\begin{cases}\pi_{\mathbf{h}}\left(0\right) & \gamma = 0\\ \pi_{\mathbf{h}}\left(\gamma\right)\exp\left(a_y^\gamma\right)\exp\left(\beta\sum_{i=1}^N w_i^\gamma x_i\right) & 0<\gamma\leq P,\end{cases}$$

where $Z_\beta\left(\mathbf{J}\right)$ is an unknown normalization constant. The data $\mathbf{x}$ belonging to each cluster $\mu>0$ follows a von Mises-Fisher (vMF) distribution $\mathrm{P}_\beta\left(\mathbf{x}|\mu,\mathbf{J}\right)\propto\exp\left(\beta\sum_{i=1}^N w_i^\mu x_i\right)$ centered on $\mathbf{w}^\mu=\{w_i^\mu\}_{i=1}^N$ (see Appendix 4.C). We assume that $\sqrt{\sum_{i=1}^N\left(w_i^\mu\right)^2}=1$ so that the corresponding cluster centroids $\mathbf{w}^\mu$ lie on the hypersphere S^{N-1} like the data $\mathbf{x}=\{x_i\}_{i=1}^N$. We marginalize $\mathrm{P}_\beta\left(\mathbf{x}, y, \mu|\mathbf{J}\right)$ over the hidden units μ and get

$$\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{J}\right) = \frac{1}{Z_\beta\left(\mathbf{J}\right)}\pi_{\mathbf{q}}\left(y\right)\left[\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)\exp\left(\beta\sum_{i=1}^N w_i^\mu x_i\right)+\pi_{\mathbf{h}}\left(0\right)\right].$$

We will now find the normalization constant $Z_\beta\left(\mathbf{J}\right)$. Marginalizing $\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{J}\right)$ over $\mathbf{x}$, we obtain

$$\begin{aligned}\mathrm{P}_\beta\left(y|\mathbf{J}\right) &= \int_{S^{N-1}} d\mathbf{x}\,\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{J}\right)\\ &= \frac{1}{Z_\beta\left(\mathbf{J}\right)}\pi_{\mathbf{q}}\left(y\right)\int_{S^{N-1}} d\mathbf{x}\left[\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)\exp\left(\beta\sum_{i=1}^N w_i^\mu x_i\right)+\pi_{\mathbf{h}}\left(0\right)\right]\\ &= \frac{1}{Z_\beta\left(\mathbf{J}\right)}\pi_{\mathbf{q}}\left(y\right)\Omega_N\left(\beta\right)\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)+\Omega_N\left(0\right)\pi_{\mathbf{h}}\left(0\right)\\ &= \frac{1}{Z_\beta\left(\mathbf{J}\right)}\pi_{\mathbf{q}}\left(y\right)\Omega_N\left(\beta\right)\left[\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)+\pi_{\mathbf{h}}\left(0\right)\exp\left(a^0\right)\right],\end{aligned}$$

where $a^0=\log\left[\Omega_N\left(0\right)/\Omega_N\left(\beta\right)\right]$. Normalization of $\mathrm{P}_\beta\left(y|\mathbf{J}\right)$ requires that

$$\mathrm{P}_\beta\left(y|\mathbf{J}\right) = \frac{\pi_{\mathbf{q}}\left(y\right)\left[\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)+\pi_{\mathbf{h}}\left(0\right)\exp\left(a^0\right)\right]}{\sum_{y'=0}^C \pi_{\mathbf{q}}\left(y'\right)\left[\sum_{\nu=1}^P \pi_{\mathbf{h}}\left(\nu\right)\exp\left(a_{y'}^\nu\right)+\pi_{\mathbf{h}}\left(0\right)\exp\left(a^0\right)\right]},$$

so we deduce that $Z_\beta\left(\mathbf{J}\right)=\Omega_N\left(\beta\right)\sum_{y=0}^C \pi_{\mathbf{q}}\left(y\right)\left[\sum_{\mu=1}^P \pi_{\mathbf{h}}\left(\mu\right)\exp\left(a_y^\mu\right)+\pi_{\mathbf{h}}\left(0\right)\exp\left(a^0\right)\right]$. We define

$$p_y^\gamma = \begin{cases}\frac{\pi_{\mathbf{q}}(y)\pi_{\mathbf{h}}(0)\exp\left(a^0\right)}{\sum_{y'=0}^C \pi_{\mathbf{q}}(y')\left[\sum_{\nu=1}^P \pi_{\mathbf{h}}(\nu)\exp\left(a_{y'}^\nu\right)+\pi_{\mathbf{h}}(0)\exp(a^0)\right]} & \gamma=0\\ \frac{\pi_{\mathbf{q}}(y)\pi_{\mathbf{h}}(\mu)\exp\left(a_y^\mu\right)}{\sum_{y'=0}^C \pi_{\mathbf{q}}(y')\left[\sum_{\nu=1}^P \pi_{\mathbf{h}}(\nu)\exp\left(a_{y'}^\nu\right)+\pi_{\mathbf{h}}(0)\exp(a^0)\right]} & 0<\gamma\leq P,\end{cases}$$

from which we obtain

$$\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{J}\right) = \sum_{\mu=1}^{P} p_y^\mu \frac{\exp\left(\beta \sum_{i=1}^{N} w_i^\mu x_i\right)}{\Omega_N\left(\beta\right)} + p_y^0 \frac{1}{\Omega_N\left(0\right)}.$$

The conditional distributions $\mathrm{P}_\beta\left(\mathbf{x}|y,\mathbf{J}\right)$ and their marginal $\mathrm{P}_\beta\left(\mathbf{x}|\mathbf{J}\right) = \sum_{y=0}^{C} \mathrm{P}_\beta\left(\mathbf{x}|y,\mathbf{J}\right) \mathrm{P}_\beta\left(y|\mathbf{J}\right)$ are von Mises-Fisher mixtures [203] with weights $\frac{p_y^\gamma}{\sum_{\nu=0}^{P} p_y^\nu}$ and $\sum_{y=0}^{C} p_y^\gamma$, respectively. As explained in Section 4.2, we constrain the marginals $\sum_{y=0}^{C} p_y^\gamma = \mathrm{P}_\beta\left(\gamma|\mathbf{J}\right)$ and $\sum_{\gamma=0}^{P} p_y^\gamma = \mathrm{P}_\beta\left(y|\mathbf{J}\right)$ to be equal to fixed distributions $p_{\mathbf{h}}\left(\gamma\right)$ and $p_{\mathbf{q}}\left(y\right)$, respectively. Formally, this means that $\mathbf{p} = \left\{p_y^\gamma\right\}_{0\leq y\leq C}^{0\leq\gamma\leq C}$ belongs to the transportation polytope with sum constraints $\sum_{\gamma=0}^{P} p_y^\gamma = p_{\mathbf{q}}\left(y\right)$ and $\sum_{y=0}^{C} p_y^\gamma = p_{\mathbf{h}}\left(\gamma\right)$ [230].

4.B Stationarity conditions of the loss

Since we constrain the memories $\mathbf{w}^\mu$ to have unit norm (see Section 4.2), the method of Lagrange multipliers tells us any set of memories $\mathbf{w}$ that minimizes Eq. (4.4) must solve the extremization problem

$$\underset{\mathbf{w},\varphi}{\mathrm{Extr}} \left\{ L\left(\mathbf{w},\mathbf{p}\right) + \frac{1}{2} \sum_{\mu=1}^{P} \varphi^\mu \left(\sum_{i=1}^{N} \left[w_i^\mu\right]^2 - 1 \right) \right\}.$$

The extrema are the points where the gradient vanishes, so they take the form

$$\begin{aligned} \partial_{w_i^\mu} L\left(\mathbf{w},\mathbf{p}\right) + \varphi^\mu w_i^\mu &= 0 \\ \sum_{i=1}^{N} \left[w_i^\mu\right]^2 &= 1 \end{aligned}$$

for all $1 \leq \mu \leq P$. We calculate $\partial_{w_i^\mu} L\left(\mathbf{w},\mathbf{p}\right)$ and write the solution in the more explicit form

$$\begin{aligned} \bar{w}_i^\mu &= \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*} \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu\left(\beta m^{\mu_*}\left(\mathbf{x}^*,\mathbf{w}\right) + \log\left[\mathbf{p}_y\right]\right) \\ w_i^\mu &= \frac{\bar{w}_i^\mu}{\sqrt{\sum_{j=1}^{N}\left[\bar{w}_j^\mu\right]^2}}, \end{aligned}$$

for all $1 \leq \mu \leq P$, where $\sigma_\gamma\left(x^{\gamma_*}\right) = \frac{\exp(x^{\gamma_*\gamma})}{\sum_{\nu=0}^{P}\exp(x^{\gamma_*\nu})}$, $m^{\mu_*\mu}\left(\mathbf{x}^*,\mathbf{w}\right) = \sum_{i=1}^{N} x_i^{*\mu_*} w_i^\mu$ for $1 \leq \mu \leq P$ and $m^{\mu_* 0}\left(\mathbf{x},\mathbf{w}\right) = \frac{1}{\beta}\log\left[\Omega_N\left(\beta\right)/\Omega_N\left(0\right)\right]$.

Similarly, any set of class weights $\mathbf{p}$ that minimizes Eq. (4.4) must solve the extremization problem

$$\underset{\mathbf{p},\omega,\lambda}{\mathrm{Extr}} \left\{ L\left(\mathbf{w},\mathbf{p}\right) + \sum_{y=0}^{C} \lambda_y \left(\sum_{\gamma=0}^{P} p_y^\gamma - p_{\mathbf{q}}\left(y\right) \right) + \sum_{\gamma=0}^{P} \omega^\gamma \left(\sum_{y=0}^{C} p_y^\gamma - p_{\mathbf{h}}\left(\gamma\right) \right) \right\}.$$

In Appendix 4.F, we show that its solution is

$$\begin{aligned}\bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma \left(\beta m^{\mu_*}\left(\mathbf{x}, \mathbf{w}\right) + \log\left[\mathbf{p}_y\right]\right) \\ p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)}\end{aligned}$$

for all $0 \leq \gamma \leq P$, where the normalization constant $\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$ is defined using Eqs. (4.26) of Appendix 4.F. Combining these equations with the ones for $\mathbf{w}$, we find the stationarity conditions

$$\begin{aligned}\bar{w}_i^\mu &= \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*} \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu \left(\beta m^{\mu_*}\left(\mathbf{x}^*, \mathbf{w}\right) + \log\left[\mathbf{p}_y\right]\right) \\ \bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma \left(\beta m^{\mu_*}\left(\mathbf{x}^*, \mathbf{w}\right) + \log\left[\mathbf{p}_y\right]\right) \\ w_i^\mu &= \frac{\bar{w}_i^\mu}{\sqrt{\sum_{j=1}^N \left[\bar{w}_j^\mu\right]^2}} \\ p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)}\end{aligned}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$.

4.C Integration of the von Mises-Fisher density

The von Mises-Fisher (vMF) distribution [231] is an isotropic Gaussian distribution restricted to the $N-1$ dimensional unit sphere S^{N-1}. It takes the form

$$p\left(\mathbf{x}\right) = \Omega\left(r\right)^{-1} \exp\left(\sum_{i=1}^N r_i x_i\right),$$

where $r = \sqrt{\sum_i \left[r_i\right]^2}$ is called the concentration parameter. When $r = 0$, it reduces to the uniform distribution on the unit sphere, whose surface surface area $\frac{2\pi^{N/2}}{\Gamma(N/2)}$ is thus also the normalization constant $\Omega_N\left(0\right)$. When

$r > 0$, we define the mean direction $\hat{\mathbf{r}} = \{r_i/r\}_{i=1}^N$ and find the normalization constant to be

$$\begin{aligned}\Omega_N(r) &= \int_{S^{N-1}} d\mathbf{x} \exp\left(\sum_{i=1}^N r_i x_i\right) \\ &= \int_{S^{N-1}} d\mathbf{x} \exp\left(r \sum_{i=1}^N \hat{r}_i x_i\right) \\ &= \Omega_{N-1}(0) \int_{-1}^1 du \left(1-u^2\right)^{(N-3)/2} \exp(ru) \\ &= \Omega_N(0) \left(\frac{r}{2}\right)^{1-N/2} \Gamma\left(\frac{N}{2}\right) I_{N/2-1}(r),\end{aligned} \tag{4.17}$$

where $I_n(x)$ is the modified Bessel function of the first kind of order n. The third line of (4.17) comes from the change of variables $u = \sum_{i=1}^N \hat{r}_i x_i$, and the fourth line is a consequence of Poisson's Bessel function integral [232]. In the limit of large N with $\rho = \frac{r}{N-2} \approx \frac{r}{N}$, we find

$$\begin{aligned}\int_{S^{N-1}} d\mathbf{x} \exp\left(\sum_{i=1}^N r_i x_i\right) \approx \frac{\Omega_N(0)}{\sqrt{2\pi(N/2-1)}} \left(\frac{N}{2}-1\right)^{1-N/2} \left(1+(2\rho)^2\right)^{-1/4} \\ \Gamma\left(\frac{N}{2}\right) \exp\left[\left(\frac{N}{2}-1\right)(\eta(2\rho)+1)\right] \\ \text{where } \eta(x) = \left(1+x^2\right)^{1/2} - 1 - \log\left[1+\left(1+x^2\right)^{1/2}\right] + \log 2,\end{aligned}$$

by exploiting the large N asymptotic expansion of $I_{N/2-1}([N/2-1]\cdot 2\rho)$ found in [233, 234]. We use Stirling's approximation

$$\Gamma\left(\frac{N}{2}\right) \approx \sqrt{2\pi(N/2-1)} \left(\frac{N}{2}-1\right)^{N/2-1} \exp\left(-\frac{N}{2}+1\right)$$

to simplify it to

$$\begin{aligned}\int_{S^{N-1}} d\mathbf{x} \exp\left(\sum_{i=1}^N r_i x_i\right) \approx \Omega_N(0) \left(1+(2\rho)^2\right)^{-1/4} \exp\left[\left(\frac{N}{2}-1\right)\eta(2\rho)\right] \\ \text{where } \eta(x) = \left(1+x^2\right)^{1/2} - 1 - \log\left[1+\left(1+x^2\right)^{1/2}\right] + \log 2.\end{aligned} \tag{4.18}$$

4.D Replicated partition function and free entropy

Suppose a student dense associative memory (DAM) model is trained using a dataset $\mathcal{D} = \{\mathbf{x}^c, y^c\}_{c=1}^M$ of M i.i.d. examples $\mathbf{x}^c$ with labels y^c. By Bayes' theorem, the student weights $\mathbf{w}$ and $\mathbf{p}$ follow the distribution

$$\mathrm{P}_\beta(\mathbf{w}, \mathbf{p}|\mathcal{D}) = \mathcal{Z}(\mathcal{D})^{-1} \mathrm{P}(\mathbf{w}, \mathbf{p}) \prod_{c=1}^M \mathrm{P}_\beta(\mathbf{x}^c, y^c \mid \mathbf{w}, \mathbf{p}),$$

where $\mathcal{Z}(\mathcal{D}) = \mathbb{E}_{\mathbf{w},\mathbf{p}}\left[\prod_{c=1}^{M} \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w},\mathbf{p}\right)\right]$. Assuming the examples are sampled from a teacher DAM with weights $\mathbf{w}^*$ and $\mathbf{p}^*$, the average replicated partition takes the form

$$\begin{aligned}
\left\langle \mathcal{Z}^L \right\rangle &= \sum_{\{y^c\}_{c=1}^M} \int \left[\prod_{c=1}^{M} d\mathbf{x}^c\right] \mathbb{E}_{\mathbf{w}^*,\mathbf{p}^*}\left[\prod_{c=1}^{M} \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w}^*,\mathbf{p}^*\right)\right] \mathcal{Z}(\mathcal{D})^L \\
&= \sum_{\{y^c\}_{c=1}^M} \int \left[\prod_{c=1}^{M} d\mathbf{x}^c\right] \mathbb{E}_{\mathbf{w}^*,\mathbf{p}^*}\left[\prod_{c=1}^{M} \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w}^*,\mathbf{p}^*\right)\right] \prod_{a=1}^{L} \mathbb{E}_{\mathbf{w}^a,\mathbf{p}^a}\left[\prod_{c=1}^{M} \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w}^a,\mathbf{p}^a\right)\right] \\
&= \mathbb{E}_{\mathbf{w}^*,\mathbf{w}}\mathbb{E}_{\mathbf{p}^*,\mathbf{p}}\left[\prod_{c=1}^{M}\sum_{y^c=0}^{C}\int_{S^{N-1}} d\mathbf{x}^c\, \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w}^*,\mathbf{p}^*\right) \prod_{a=1}^{L} \mathrm{P}_\beta\left(\mathbf{x}^c, y^c|\mathbf{w}^a,\mathbf{p}^a\right)\right] \\
&= \mathbb{E}_{\mathbf{w}^*,\mathbf{w}}\mathbb{E}_{\mathbf{p}^*,\mathbf{p}}\left[\left(\sum_{y=0}^{C}\int_{S^{N-1}} d\mathbf{x}\, \mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{w}^*,\mathbf{p}^*\right) \prod_{a=1}^{L} \mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{w}^a,\mathbf{p}^a\right)\right)^M\right],
\end{aligned}$$

where we redefined $\mathbf{w} = \{\mathbf{w}^a\}_{a=1}^L$ and $\mathbf{p} = \{\mathbf{p}^a\}_{a=1}^L$. In order to simplify the argument of $(\cdot)^M$, it is convenient to write Eq. (4.3) in the form

$$\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{w}^a,\mathbf{p}^a\right) = \sum_{\gamma=0}^{P} \Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)^{-1} p_y^{a\gamma} \exp\left(\beta\left[1-\delta_{\gamma 0}\right]\sum_{i=1}^{N} w_i^{a\gamma} x_i\right), \tag{4.19}$$

where the value of $\mathbf{w}^{a0}$ is arbitrary. Until the end of this Section, all sums over the hidden units will include the index 0 unless explicitly indicated otherwise. Define $I_y(\mathbf{x}) = \mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{w}^*,\mathbf{p}^*\right)\prod_{a=1}^{L}\mathrm{P}_\beta\left(\mathbf{x}, y|\mathbf{w}^a,\mathbf{p}^a\right)$, then

$$\begin{aligned}
I_y(\mathbf{x}) = &\left[\sum_{\gamma_*=0}^{P^*} \Omega_N\left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\right)^{-1} p_y^{*\gamma_*} \exp\left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\sum_{i=1}^{N} w_i^{*\gamma_*} x_i\right)\right] \\
&\prod_{a=1}^{L}\left[\sum_{\gamma=0}^{P} \Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)^{-1} p_y^{a\gamma} \exp\left(\beta\left[1-\delta_{\gamma 0}\right]\sum_{i=1}^{N} w_i^{a\gamma} x_i\right)\right] \\
= &\sum_{\gamma_*\gamma_1\cdots\gamma_L} \Omega_N\left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\right)^{-1}\left[\prod_a \Omega_N\left(\beta\left[1-\delta_{\gamma_a 0}\right]\right)\right]^{-1} \\
&p_y^{*\gamma_*}\left[\prod_a p_y^{a\gamma_a}\right]\exp\left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\sum_i w_i^{*\gamma_*} x_i + \beta\sum_a\left[1-\delta_{\gamma_a 0}\right]\sum_i w_i^{a\gamma_a} x_i\right).
\end{aligned}$$

We will now evaluate the integral $\int_{S^{N-1}} d\mathbf{x}\, I_y(\mathbf{x})$. Using Eq. (4.18) of Appendix 4.C with $\rho = \rho_{\gamma_*\gamma} := \frac{1}{N}\sqrt{\sum_i\left[\beta^*\left[1-\delta_{\gamma_* 0}\right] w_i^{*\gamma_*} + \beta\sum_a\left[1-\delta_{\gamma_a 0}\right] w_i^{a\gamma_a}\right]^2}$, we get

$$\begin{aligned}
&\int_{S^{N-1}} d\mathbf{x} \exp\left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\sum_i w_i^{*\gamma_*} x_i + \beta\sum_a\left[1-\delta_{\gamma_a 0}\right]\sum_i w_i^{a\gamma_a} x_i\right) \\
&\approx \Omega_N(0)\left(1+\left(2\rho_{\gamma_*\gamma}\right)^2\right)^{-1/4}\exp\left[\left(\frac{N}{2}-1\right)\eta\left(2\rho_{\gamma_*\gamma}\right)\right],
\end{aligned}$$

in the limit of large N. Since $\sqrt{\sum_i \left(w_i^{*\gamma_*}\right)^2} = 1$, the square of $\rho_{\gamma_*\gamma}$ simplifies to

$$\begin{aligned}
\left(\rho_{\gamma_*\gamma}\right)^2 &= \left(\frac{1}{N}\sqrt{\sum_i \left[\beta^* \left[1-\delta_{\gamma_* 0}\right] w_i^{*\gamma_*} + \beta \sum_a \left[1-\delta_{\gamma_a 0}\right] w_i^{a\gamma_a}\right]^2}\right)^2 \\
&= \upsilon^2 \left[1-\delta_{\gamma_* 0}\right] + 2\frac{\beta}{N}\upsilon \left[1-\delta_{\gamma_* 0}\right] \sum_a \left[1-\delta_{\gamma_a 0}\right] \sum_i w_i^{*\gamma_*} w_i^{a\gamma_a} \\
&\quad + \frac{\beta^2}{N^2} \sum_{a,b} \left[1-\delta_{\gamma_a 0}\right]\left[1-\delta_{\gamma_b 0}\right] \sum_i w_i^{a\gamma_a} w_i^{b\gamma_b},
\end{aligned}$$

where $\upsilon = \frac{\beta^*}{N}$. To leading order in $\frac{\beta}{N}$ we find

$$\left(1+\left(2\rho_{\gamma_*\gamma}\right)^2\right)^{-1/4} \approx \left(1+\left(2\upsilon\left[1-\delta_{\gamma_* 0}\right]\right)^2\right)^{-1/4}$$

and

$$\begin{aligned}
\left(\frac{N}{2}-1\right)\eta\left(2\rho_{\gamma_*\gamma_a}\right) &\approx \left(\frac{N}{2}-1\right)\eta\left(2\upsilon\left[1-\delta_{\gamma_* 0}\right]\right) + \beta_{\text{eff}}\left[1-\delta_{\gamma_* 0}\right] \sum_a \left[1-\delta_{\gamma_a 0}\right] \sum_i w_i^{*\gamma_*} w_i^{a\gamma_a} \\
&\quad + \frac{\left[\beta\xi_{\gamma_*}\right]^2}{2N} \sum_{a,b} \left[1-\delta_{\gamma_a 0}\right]\left[1-\delta_{\gamma_b 0}\right] \sum_i w_i^{a\gamma_a} w_i^{b\gamma_b},
\end{aligned}$$

where $\beta_{\text{eff}} = \frac{2\upsilon}{\sqrt{[2\upsilon]^2+1}+1}\beta$ and $\xi_{\gamma_*} = \delta_{\gamma_* 0} + \sqrt{\frac{2}{\sqrt{[2\upsilon]^2+1}+1}}\left[1-\delta_{\gamma_* 0}\right]$. Assuming that $\upsilon \gg 1/N$ and $\beta \ll N$, we drop the last term. Finally, we use Eq. (4.18) backwards with $\rho = \upsilon\left[1-\delta_{\gamma_* 0}\right]$ and obtain

$$\begin{aligned}
&\Omega_N \left(\beta^*\left[1-\delta_{\gamma_* 0}\right]\right)^{-1} \int_{S^{N-1}} d\mathbf{x} \exp\left(\beta^*\left[1-\delta_{\gamma_* 0}\right] \sum_i w_i^{*\gamma_*} x_i + \beta \sum_a \left[1-\delta_{\gamma_a 0}\right] \sum_i w_i^{a\gamma_a} x_i\right) \\
&\approx \exp\left(\beta_{\text{eff}}\left[1-\delta_{\gamma_* 0}\right] \sum_a \left[1-\delta_{\gamma_a 0}\right] \sum_i w_i^{*\gamma_*} w_i^{a\gamma_a}\right),
\end{aligned}$$

from which we conclude that

$$\begin{aligned}
\sum_{y=0}^{C} \int_{S^{N-1}} d\mathbf{x}\, I_y\left(\mathbf{x}\right) \approx \sum_{\gamma_* \gamma_1 \ldots \gamma_L} \sum_y p_y^{*\gamma_*} \left[\prod_a p_y^{a\gamma_a}\right] \left[\prod_a \Omega_N\left(\beta\left[1-\delta_{\gamma_a 0}\right]\right)\right]^{-1} \\
\exp\left(\beta_{\text{eff}}\left[1-\delta_{\gamma_* 0}\right] \sum_a \left[1-\delta_{\gamma_a 0}\right] \sum_i w_i^{*\gamma_*} w_i^{a\gamma_a}\right).
\end{aligned}$$

We define $\alpha = M/N$, introduce the order parameter

$$m^{a\gamma_*\nu} \text{ for } \sum_i w_i^{*\gamma_*} w_i^{a\nu}, \tag{4.20}$$

and insert into $\left\langle \mathcal{Z}^L \right\rangle$ the identity operator

$$\begin{aligned}
1 &= \int_{\mathbb{R}} \prod_{\gamma_*,\nu;a} dm^{a\gamma_*\nu} \delta\left(m^{a\gamma_*\nu} - \sum_i w_i^{*\gamma_*} w_i^{a\nu} \right) \\
&= \int_{\mathbb{R}} \prod_{\gamma_*,\nu;a} dm^{a\gamma_*\nu} \int_{i\mathbb{R}} \prod_{\gamma_*,\nu;a} d\hat{m}^{a\gamma_*\nu} \exp\left\{ \frac{\beta_{\text{eff}}\alpha}{P^*} \sum_{\gamma_*,\nu;a} \hat{m}^{a\gamma_*\nu} \left(\sum_i w_i^{*\gamma_*} w_i^{a\nu} - m^{a\gamma_*\nu} \right) \right\}
\end{aligned}$$

so that we can rewrite it as

$$\begin{aligned}
\left\langle \mathcal{Z}^L \right\rangle = \int \prod_{\gamma_*,\nu;a} d\hat{m}^{a\gamma_*\nu} dm^{a\gamma_*\nu} \mathbb{E}_{\mathbf{w}^*,\mathbf{w}} \exp\left\{ N H_S\left(\mathbf{w}, \mathbf{w}^*; \hat{\mathbf{m}}\right) \right\} \exp\left\{ -N H_Q\left(\mathbf{m}, \hat{\mathbf{m}}\right) \right\} \\
\mathbb{E}_{\mathbf{p}^*,\mathbf{p}} \exp\left\{ \alpha N \log\left[\sum_{\gamma_*\gamma_1\cdots\gamma_L} \sum_y p_y^{*\gamma_*} \left[\prod_a p_y^{a\gamma_a} \right] \exp\left\{ H_E\left(\gamma, \gamma_*; \mathbf{m}\right) \right\} \right] \right\},
\end{aligned} \tag{4.21}$$

where
$$\begin{aligned}
H_S\left(\mathbf{w}, \mathbf{w}^*; \hat{\mathbf{m}}\right) &= \frac{\beta_{\text{eff}}\alpha}{P^*} \sum_{\gamma_*,\nu;a} \hat{m}^{a\gamma_*\nu} \sum_i w_i^{*\gamma_*} w_i^{a\nu} \\
H_Q\left(\mathbf{m}, \hat{\mathbf{m}}\right) &= \frac{\beta_{\text{eff}}\alpha}{P^*} \sum_{\gamma_*,\nu;a} \hat{m}^{a\gamma_*\nu} m^{a\gamma_*\nu}, \\
H_E\left(\gamma, \gamma_*; \mathbf{m}\right) &= -\sum_a \log\left[\Omega_N\left(\beta\left[1 - \delta_{\gamma_a 0}\right] \right) \right] + \beta_{\text{eff}}\left[1 - \delta_{\gamma_* 0}\right] \sum_a \left[1 - \delta_{\gamma_a 0}\right] m^{a\gamma_*\gamma_a}.
\end{aligned}$$

Using Eq. (4.18) of Appendix 4.C, the exponential of H_S integrates to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}} \exp\left\{ N H_S\left(\mathbf{w}, \mathbf{w}^*; \hat{\mathbf{m}}\right) \right\} \\
&= \Omega_N\left(0\right)^{-PL} \int_{S^{N-1}} d\mathbf{w} \exp\left(\frac{\beta_{\text{eff}}\alpha}{P^*} \sum_{\gamma_*,\nu;a} \hat{m}^{a\gamma_*\nu} \sum_i w_i^{*\gamma_*} w_i^{a\nu} \right) \\
&\approx \prod_{\nu;a} \exp\left[-\frac{1}{4} \log\left(1 + \left[\frac{2\beta_{\text{eff}}\alpha}{P^*} \right]^2 \sum_i \left[\sum_{\gamma_*} \hat{m}^{a\gamma_*\nu} w_i^{*\gamma_*} \right]^2 \right) \right] \\
&\quad \exp\left[\left(\frac{N}{2} - 1 \right) \eta\left(\frac{2\beta_{\text{eff}}\alpha}{P^*} \sqrt{ \sum_i \left[\sum_{\gamma_*} \hat{m}^{a\gamma_*\nu} w_i^{*\gamma_*} \right]^2 } \right) \right].
\end{aligned}$$

In order to continue the calculations, we make the replica-symmetric ansatz

$$\begin{aligned}
m^{a\gamma_*\nu} = m^{\gamma_*\nu} \text{ and } \hat{m}^{a\gamma_*\nu} = \hat{m}^{\gamma_*\nu} \text{ for all } a; \gamma_*, \nu \\
p_y^{a\nu} = p_y^{\nu} \text{ for all } a; \nu
\end{aligned}$$

so that we can use the replica trick to simplify

$$\begin{aligned}
&\frac{1}{L}\log\left[\sum_{\gamma_*\gamma_1\cdots\gamma_L}\sum_y p_y^{*\gamma_*}\left[\prod_a p_y^{a\gamma_a}\right]\exp\left\{H_E\left(\gamma,\gamma_*;\mathbf{m}\right)\right\}\right]\\
&=\frac{1}{L}\log\left[\sum_{\gamma_*}\sum_y p_y^{*\gamma_*}\sum_{\gamma_1\cdots\gamma_L}\left[\prod_a p_y^{a\gamma_a}\right]\exp\left\{H_E\left(\gamma,\gamma_*;\mathbf{m}\right)\right\}\right]\\
&\approx\sum_{\gamma_*}\sum_y p_y^{*\gamma_*}\log\left[\sum_\gamma p_y^\gamma\Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)^{-1}\exp\left(\beta_{\text{eff}}\left[1-\delta_{\gamma_* 0}\right]\left[1-\delta_{\gamma 0}\right]m^{\gamma_*\gamma}\right)\right],
\end{aligned}$$

as we take L to zero. Similarly, we get

$$\begin{aligned}
\frac{1}{L}\log\left[\mathbb{E}_{\mathbf{w}}\exp\left\{NH_S\left(\mathbf{w},\mathbf{w}^*;\hat{\mathbf{m}}\right)\right\}\right]=&-\frac{1}{4}\log\left(1+\left[\frac{2\beta_{\text{eff}}\alpha}{P^*}\right]^2\sum_i\left[\sum_{\gamma_*}\hat{m}^{\gamma_*\nu}w_i^{*\gamma_*}\right]^2\right)\\
&+\left(\frac{N}{2}-1\right)\eta\left(\frac{2\beta_{\text{eff}}\alpha}{P^*}\sqrt{\sum_i\left[\sum_{\gamma_*}\hat{m}^{\gamma_*\nu}w_i^{*\gamma_*}\right]^2}\right).
\end{aligned}$$

From now on, we assume for simplicity that $\sum_y p_y^{*0}=p_{\mathbf{h}}^*\left(0\right)=0$ and $\sum_y p_y^{*\mu_*}=p_{\mathbf{h}}^*\left(\mu_*\right)=1/P^*$ for all $\mu_*>0$. Defining $g_y^{*\mu_*}=P^*p_y^{*\mu_*}=\mathrm{P}_\beta\left(y|\mu_*,\mathbf{J}\right)$ for all $\mu_*>0$, $\mathbf{g}^*=\left\{g_y^{*\gamma_*}\right\}_{0\leq y\leq C}^{1\leq\mu_*\leq P^*}$ and $\varrho=\frac{\alpha}{P^*}=\frac{M}{P^*N}$, the free entropy f then takes the form

$$f\approx\underset{\mathbf{m},\hat{\mathbf{m}},\mathbf{p}}{\text{Extr}}\left\{f\left(\mathbf{m},\hat{\mathbf{m}},\mathbf{p}\right)\right\}\quad\text{such that}\quad\sum_{\gamma=0}^P p_y^\gamma=p_{\mathbf{q}}\left(y\right)\quad\text{and}\quad\sum_{y=0}^C p_y^\gamma=p_{\mathbf{h}}\left(\gamma\right)\tag{4.22}$$

$$\begin{aligned}
\text{with}\quad f\left(\mathbf{m},\hat{\mathbf{m}},\mathbf{p}\right)=&-\beta_{\text{eff}}\varrho\sum_{\gamma_*,\gamma=0}^{P^*,P}\hat{m}^{\gamma_*\gamma}m^{\gamma_*\gamma}+\frac{1}{2}\mathbb{E}_{\mathbf{w}^*}\left[\sum_{\gamma=0}^P\eta\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{i=1}^N\left[\sum_{\gamma_*=0}^{P^*}\hat{m}^{\gamma_*\gamma}w_i^{*\gamma_*}\right]^2}\right)\right]\\
&+\varrho\sum_{\mu_*=1}^{P^*}\sum_{y=0}^C\mathbb{E}_{\mathbf{g}^*}\left[g_y^{*\mu_*}\right]\log\left[\sum_{\gamma=0}^P p_y^\gamma\Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)^{-1}\exp\left(\beta_{\text{eff}}\left[1-\delta_{\gamma 0}\right]m^{\mu_*\gamma}\right)\right]
\end{aligned}$$

where we approximated the expectation over $\mathbf{p}$ as a (constrained) extremization problem using Laplace's method. By inspection, the order parameters $m^{0\gamma}$ and $\hat{m}^{0\gamma}$ always vanish, so we ignore them in the incoming derivation of the saddle-point equations.

4.E Saddle-point equations

In this Appendix, we adopt the convention $1\leq\mu_*\leq P^*$, $0\leq\gamma\leq P$ and $1\leq\mu\leq P$. By the Lagrange multiplier theorem, any set of class weights $\mathbf{p}$ that extremizes Eq. (4.22) must solve the extremization problem

$$\underset{\mathbf{p},\omega,\lambda}{\text{Extr}}\left\{f\left(\mathbf{m},\hat{\mathbf{m}},\mathbf{p}\right)+\sum_{y=0}^C\lambda_y\left(\sum_{\gamma=0}^P p_y^\gamma-p_{\mathbf{q}}\left(y\right)\right)+\sum_{\gamma=0}^P\omega^\gamma\left(\sum_{y=0}^C p_y^\gamma-p_{\mathbf{h}}\left(\gamma\right)\right)\right\}.$$

In Appendix 4.F, we show that its solution is

$$\bar{p}_y^{\gamma} = \sum_{\mu_*=1}^{P^*} \mathbb{E}_{\mathbf{g}^*}\left[g_y^{*\mu_*}\right] \sigma_\gamma\left(\beta_{\text{eff}}\left[1-\delta_{\gamma 0}\right] m^{\mu_*} - \log\left[\Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)\right] + \log\left[\mathbf{p}_y\right]\right)$$
$$p_y^{\gamma} = \frac{\bar{p}_y^{\gamma}}{\zeta_y^{\gamma}\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)},$$

for all $0 \leq \gamma \leq P$, where $\zeta_y^{\gamma}\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$ is defined using Eqs. (4.26) of Appendix 4.F and $\sigma_\gamma\left(x^{\mu_*}\right) = \frac{\exp(x^{\mu_*\gamma})}{\sum_{\kappa=0}^{P}\exp(x^{\mu_*\kappa})}$ is the softmax function. We extremize Eq. (4.22) with respect to the remaining parameters by solving for the gradient equal to zero. $\partial_{m^{\mu_*\mu}} f\left(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}\right) = 0$ immediately yields

$$\hat{m}^{\mu_*\gamma} = \left[1-\delta_{\gamma 0}\right] \sum_{y=0}^{C} \mathbb{E}_{\mathbf{g}^*}\left[g_y^{*\mu_*}\right] \frac{p_y^{\gamma}\Omega_N\left(\beta\left[1-\delta_{\gamma 0}\right]\right)^{-1}\exp\left(\beta_{\text{eff}}\left[1-\delta_{\gamma 0}\right] m^{\mu_*\gamma}\right)}{\sum_{\kappa=0}^{P} p_y^{\kappa}\Omega_N\left(\beta\left[1-\delta_{\kappa 0}\right]\right)^{-1}\exp\left(\beta_{\text{eff}}\left[1-\delta_{\kappa 0}\right] m^{\mu_*\kappa}\right)}. \tag{4.23}$$

On the other hand, $\partial_{\hat{m}^{\mu_*\gamma}} f\left(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}\right) = 0$ gives an equation that depends on the choice of prior for the teacher memories $\mathbf{w}^{*\mu_*}$. We first investigate the case where the teacher memories $\mathbf{w}^{*\mu_*}$ and the class weights $\mathbf{g}^*$ are distributed uniformly at random over the sets to which they are constrained (see Section 4.2.2). In this scenario, we have $\mathbb{E}_{\mathbf{g}^*}\left[g_y^{*\mu_*}\right] = p_{\mathbf{q}}^*\left(y\right)$. Moreover, assuming that $P^* \ll N$, random vectors on the unit sphere are orthonormal with high probability, which means that

$$\begin{aligned}
\sum_{i=1}^{N}\left[\sum_{\mu_*=1}^{P^*}\hat{m}^{\mu_*\gamma} w_i^{*\mu_*}\right]^2 &= \sum_{i=1}^{N}\sum_{\mu_*,\nu_*=1}^{P^*}\hat{m}^{\mu_*\gamma}\hat{m}^{\nu_*\gamma} w_i^{*\mu_*} w_i^{*\nu_*} \\
&= \sum_{\mu_*=1}^{P^*}\left[\hat{m}^{\mu_*\gamma}\right]^2 + \sum_{\mu_*\neq\nu_*}\hat{m}^{\mu_*\gamma}\hat{m}^{\nu_*\gamma}\sum_{i=1}^{N} w_i^{*\mu_*} w_i^{*\nu_*} \\
&\approx \sum_{\mu_*=1}^{P^*}\left[\hat{m}^{\mu_*\gamma}\right]^2.
\end{aligned}$$

Therefore, $\partial_{\hat{m}^{\mu_*\gamma}} f\left(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}\right) = 0$ gives

$$\begin{aligned}
0 &= \partial_{\hat{m}^{\mu_*\gamma}} f\left(\mathbf{m}, \hat{\mathbf{m}}, \mathbf{p}\right) \\
0 &= -\beta_{\text{eff}}\varrho m^{\mu_*\gamma} + \frac{1}{2}\partial_{\hat{m}^{\mu_*\gamma}}\sum_{\kappa=0}^{P}\eta\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{\nu_*=1}^{P^*}\left[\hat{m}^{\nu_*\kappa}\right]^2}\right) \\
m^{\mu_*\gamma} &= \varsigma\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{\nu_*=1}^{P^*}\left[\hat{m}^{\nu_*\gamma}\right]^2}\right)\frac{\hat{m}^{\mu_*\gamma}}{\sqrt{\sum_{\nu_*=1}^{P^*}\left[\hat{m}^{\nu_*\gamma}\right]^2}},
\end{aligned}$$

where $\varsigma\left(x\right) = \frac{x}{\sqrt{x^2+1}+1}$. $\hat{m}^{\mu_*\gamma}$ vanishes (see Eqs. 4.23) and $m^{\mu_*\gamma}$ is arbitrary (see Eqs. 4.19 and 4.20) when $\gamma = 0$. Therefore, we may update only $\hat{m}^{\mu_*\mu}$ and $m^{\mu_*\mu}$ with $1 \leq \mu \leq P$ when solving for $\nabla f = 0$ by

fixed-point iteration. Defining $m^{\mu_* 0} = \frac{1}{\beta_{\text{eff}}} \log\left[\Omega_N\left(\beta\right)/\Omega_N\left(0\right)\right]$, we then obtain the saddle-point equations

$$\begin{aligned}
\hat{m}^{\mu_* \mu} &= \sum_{y=0}^{C} p^*_{\mathbf{q}}\left(y\right) \sigma_\mu\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\
m^{\mu_* \mu} &= \varsigma\left(2\beta_{\text{eff}} \varrho \sqrt{\sum_{\nu_*=1}^{P^*} \left[\hat{m}^{\nu_* \mu}\right]^2}\right) \frac{\hat{m}^{\mu_* \mu}}{\sqrt{\sum_{\nu_*=1}^{P^*} \left[\hat{m}^{\nu_* \mu}\right]^2}},
\end{aligned}$$

for all $1 \leq \mu \leq P$, where we simplified the argument of the softmax function $\sigma_\mu\left(x^{\mu_*}\right) = \frac{\exp(x^{\mu_* \mu})}{\sum_{\kappa=0}^{P} \exp(x^{\mu_* \kappa})}$ using $m^{\mu_* 0} = \frac{1}{\beta_{\text{eff}}} \log\left[\Omega_N\left(\beta\right)/\Omega_N\left(0\right)\right]$. Putting the equations for $m^{\mu_* \mu}$ and p_y^γ together, we find

$$\begin{aligned}
\hat{m}^{\mu_* \mu} &= \sum_{y=0}^{C} p^*_{\mathbf{q}}\left(y\right) \sigma_\mu\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\
\bar{p}_y^\gamma &= p^*_{\mathbf{q}}\left(y\right) \sum_{\mu=1}^{P^*} \sigma_\gamma\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\
m^{\mu_* \mu} &= \varsigma\left(2\beta_{\text{eff}} \varrho \sqrt{\sum_{\nu_*=1}^{P^*} \left[\hat{m}^{\nu_* \mu}\right]^2}\right) \frac{\hat{m}^{\mu_* \mu}}{\sqrt{\sum_{\nu_*=1}^{P^*} \left[\hat{m}^{\nu_* \mu}\right]^2}} \\
p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)}
\end{aligned}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$. If we instead clamp $\mathbf{w}^{*\mu_*}$ and $\mathbf{g}^{*\mu_*}$ to some fixed patterns $\mathbf{x}^{*\mu_*}$ and their soft labels $\mathbf{q}^{*\mu_*}$, respectively, then the solutions of Eq. (4.22) take the form

$$\begin{aligned}
\hat{m}^{\mu_* \mu} &= \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\
\bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma\left(\beta_{\text{eff}} m^{\mu_*} + \log\left[\mathbf{p}_y\right]\right) \\
m^{\mu_* \mu} &= \varsigma\left(2\beta_{\text{eff}} \varrho \sqrt{\sum_{i=1}^{N} \left[\sum_{\nu_*=1}^{P^*} \hat{m}^{\nu_* \mu} x_i^{*\nu_*}\right]^2}\right) \frac{\sum_{i=1}^{N} x_i^{*\nu_*} \sum_{\nu_*=1}^{P^*} \hat{m}^{\nu_* \mu} x_i^{*\nu_*}}{\sqrt{\sum_{i=1}^{N} \left[\sum_{\nu_*=1}^{P^*} \hat{m}^{\nu_* \mu} x_i^{*\nu_*}\right]^2}} \\
p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)}.
\end{aligned}$$

Defining $\bar{x}_i^\mu = \sum_{\mu_*} \hat{m}^{\mu_* \mu} x_i^{*\mu_*}$, we thus find

$$\begin{aligned}
\bar{x}_i^\mu &= \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*} \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \\
\bar{p}_y^\gamma &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \\
m^{\mu_* \mu} &= \varsigma \left(2 \beta_{\text{eff}} \varrho \sqrt{\sum_{i=1}^{N} \left[\bar{x}_i^\mu \right]^2} \right) \frac{\sum_{i=1}^{N} x_i^{*\mu_*} \bar{x}_i^\mu}{\sqrt{\sum_{i=1}^{N} \left[\bar{x}_i^\mu \right]^2}} \\
p_y^\gamma &= \frac{\bar{p}_y^\gamma}{\zeta_y^\gamma \left(\bar{\mathbf{p}}; p_{\mathbf{h}} \right)},
\end{aligned}$$

for all $1 \leq \mu \leq P$ and $0 \leq \gamma \leq P$.

4.F Normalization of the weights

In order to enforce $\sqrt{\sum_{i=1}^{N} \left(w_i^\mu \right)^2} = 1$ at each SGD step, we project $\mathbf{w}^\mu$ and the gradient of the loss with respect to $\mathbf{w}^\mu$ onto the unit sphere S^{N-1} and the tangent space of $\mathbf{w}^\mu$, respectively. We divide w_i^μ by its norm to project it onto S^{N-1}, and we multiply the gradient by $\delta_{jk} - w_j^\mu w_k^\mu$ to project it onto the tangent space. Projecting the gradient onto the tangent space is a mathematically sound way to obtain the gradient of a function restricted to a manifold embedded in $\mathbb{R}^N$ [235].

The tasks of deriving the saddle-point equations for $\bar{p}_y^\gamma$ (Eqs. 4.10), finding the stationarity condition of the loss with respect to the class weights p_y^γ (Eqs. 4.6) and efficiently training p_y^γ all amount to solving the extremization problem

$$\underset{\mathbf{p}, \omega, \lambda}{\text{Extr}} \left\{ f \left(\mathbf{p} \right) + \sum_{y=0}^{C} \lambda_y \left(\sum_{\gamma=0}^{P} p_y^\gamma - p_{\mathbf{q}} \left(y \right) \right) + \sum_{\gamma=0}^{P} \omega^\gamma \left(\sum_{y=0}^{C} p_y^\gamma - p_{\mathbf{h}} \left(\gamma \right) \right) \right\}, \tag{4.24}$$

where λ_y and ω^γ are Lagrange multipliers that enforce the constraints $\sum_\gamma p_y^\gamma = p_{\mathbf{q}} \left(y \right)$ and $\sum_y p_y^\gamma = p_{\mathbf{h}} \left(\gamma \right)$. In the former task, $f \left(\mathbf{p} \right)$ is the free entropy (Eq. 4.22) at fixed $\mathbf{m}$ and $\hat{\mathbf{m}}$. In the latter two, it is the negative log-likelihood loss (Eq. 4.4) at a given $\mathbf{w}$. To find the saddle-point equations and the stationarity conditions of the loss, we derive an implicit solution of Eq. (4.24) that is useful in analytical calculations. On the other hand, to train p_y^γ, we design an algorithm to quickly compute a numerical solution of Eq. (4.24). We start by noting that, since the extrema of Eq. (4.24) are the points where the gradient vanishes, they take the form

$$\begin{aligned}
\partial_{p_y^\gamma} f \left(\mathbf{p} \right) &= \lambda_y + \omega^\gamma \\
\sum_{\gamma=0}^{P} p_y^\gamma &= p_{\mathbf{q}} \left(y \right) \\
\sum_{y=0}^{C} p_y^\gamma &= p_{\mathbf{h}} \left(\gamma \right).
\end{aligned} \tag{4.25}$$

Define $\bar{p}_y^\gamma = \partial_{p_y^\gamma} \exp\left(f\left(\mathbf{p}\right)\right) = p_y^\gamma \partial_{p_y^\gamma} f\left(\mathbf{p}\right)$, then

$$\frac{\bar{p}_y^\gamma}{p_y^\gamma} = \lambda_y + \omega^\gamma,$$

along with the previously established row and column sum constraints on $\mathbf{p}$. Rearranging terms, we get

$$p_y^\gamma = \frac{1}{\lambda_y + \omega^\gamma} \bar{p}_y^\gamma.$$

Using the row and column constraints $\sum_{\gamma=0}^{P} p_y^\gamma = p_{\mathbf{q}}\left(y\right)$ and $\sum_{y=0}^{C} p_y^\gamma = p_{\mathbf{h}}\left(\gamma\right)$, we find that the Lagrange multipliers λ_y and ω^γ solve the non-linear equations

$$\begin{aligned}
\lambda_y &= \frac{1}{p_{\mathbf{q}}\left(y\right)} \sum_{\gamma=0}^{P} \frac{\lambda_y}{\lambda_y + \omega^\gamma} \bar{p}_y^\gamma \\
\omega^\gamma &= \frac{1}{p_{\mathbf{h}}\left(\gamma\right)} \sum_{y=0}^{C} \frac{\omega^\gamma}{\lambda_y + \omega^\gamma} \bar{p}_y^\gamma,
\end{aligned} \tag{4.26}$$

For conciseness, we define $\zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right) = \lambda_y\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right) + \omega^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$, where $\lambda_y\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$ and $\omega^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$ are the λ_y and ω^γ solving Eqs. (4.26) at given P and $p_{\mathbf{h}}$. Using these definitions, we find the implicit solution $p_y^\gamma = \bar{p}_y^\gamma / \zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$. As wanted, this equation is useful in analytical calculations involving the saddle-point equations and the stationarity conditions of the loss. However, it is also quite slow to solve numerically for a given $\bar{\mathbf{p}}$, as reported in [236, 237, 238]. Therefore, although we can train p_y^γ by iterating $p_y^\gamma = \bar{p}_y^\gamma / \zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$, it is not a very efficient method.

In order to efficiently train p_y^γ, we devise a faster way to solve Eqs. (4.25) than through Eq. (4.26). Exponentiating both sides of the first line of Eqs. (4.25), we find

$$\begin{aligned}
\exp\left(\eta \partial_{p_y^\gamma} f\left(\mathbf{p}\right)\right) \exp\left(-\eta \lambda_y\right) \exp\left(-\eta \omega^\gamma\right) &= 1 \\
\exp\left(-\eta \omega^\gamma\right) p_y^\gamma \exp\left(\eta \partial_{p_y^\gamma} f\left(\mathbf{p}\right)\right) \exp\left(-\eta \lambda_y\right) &= p_y^\gamma,
\end{aligned}$$

where η is an arbitrary scalar that will play the role of a learning rate. Fix $k_y^\gamma = p_y^\gamma \exp\left(\eta \partial_{p_y^\gamma} f\left(\mathbf{p}\right)\right)$. By the Sinkhorn-Knopp theorem [239, 240, 241], there is a rescaled matrix of the form $p_y'^\gamma = D_L^\gamma\left(\mathbf{k}\right) k_y^\gamma D_R^y\left(\mathbf{k}\right)$ that satisfies the same constraints $p_y'^\gamma \geq 0$, $\sum_{\gamma=0}^{P} p_y'^\gamma = p_{\mathbf{q}}\left(y\right)$ and $\sum_{y=0}^{C} p_y'^\gamma = p_{\mathbf{h}}\left(\gamma\right)$ as p_y^γ if some technical conditions are satisfied [241]. Moreover, if $p_y'^\gamma$ exists, then it is unique [242, 243, 241], and we can quickly compute suitable scaling factors $D_L^\gamma\left(\mathbf{k}\right)$ and $D_R^\gamma\left(\mathbf{k}\right)$ for $\mathbf{k} = \left\{k_y^\gamma\right\}_{0 \leq y \leq C}^{0 \leq \gamma \leq P}$ using the Sinkhorn-Knopp algorithm [239, 240, 241, 244]. $p_y'^\gamma$ is generally not equal to p_y^γ. However, we observe that the iteration

$$\begin{aligned}
k_y^\gamma\left(t\right) &= p_y^\gamma\left(t\right) \exp\left(\eta \partial_{p_y^\gamma(t)} f\left(\mathbf{p}\left(t\right)\right)\right) \\
p_y^\gamma\left(t+1\right) &= D_L^\gamma\left(\mathbf{k}\left(t\right)\right) k_y^\gamma\left(t\right) D_R^y\left(\mathbf{k}\left(t\right)\right)
\end{aligned} \tag{4.27}$$

converges at small η (for example $\sim 0.1/P$), which means that it must converge to a solution of Eqs. (4.25). Since the Sinkhorn-Knopp algorithm is fast, it is significantly more efficient to iterate Eqs. (4.27) than

$p_y^\gamma = \bar{p}_y^\gamma / \zeta_y^\gamma\left(\bar{\mathbf{p}}; p_{\mathbf{h}}\right)$ to solve Eqs. (4.25).

In practice, we train p_y^γ using a stochastic variant of Eqs. (4.27) where the gradient is estimated over small batches of data and smoothed with a momentum hyperparameter. Furthermore, we multiply the gradient by $\delta_{yy'} - p_y^\gamma / p_{\mathbf{h}}(\gamma)$ to reduce the size of the components of $\exp\left(\eta \partial_{p_y^\gamma(t)} f\left(\mathbf{p}(t)\right)\right)$ that move p_y^γ away from the constraints $\sum_{\gamma=0}^{P} p_y^\gamma = p_{\mathbf{q}}(y)$ and $\sum_{y=0}^{C} p_y^\gamma = p_{\mathbf{h}}(\gamma)$. This step is a simple approximation of the gradient projection proposed in [245].

4.G Saddle-point hierarchy

Suppose that the set of parameters $\bar{x}_i^{\text{fixed},\mu}$, $\bar{p}_y^{\text{fixed},\gamma}$, $m^{\text{fixed},\mu_*\gamma}$, $p_y^{\text{fixed},\gamma}$ with hidden unit prior $p_{\mathbf{h}}^{\text{given}}(\gamma)$ is a fixed point of Eqs. (4.10) with P hidden units. Substitute into the same saddle-point equations with $P + R \in \{P, ..., 2P\}$ hidden units the duplicated order parameters

$$\bar{x}_i^{\text{dupli},\mu} = \begin{cases} \bar{x}_i^{\text{fixed},\mu} & 0 < \mu \leq P \\ \bar{x}_i^{\text{fixed},\mu-P} & P < \mu \leq P + R \end{cases}$$

$$\bar{p}_y^{\text{dupli},\gamma} = \begin{cases} \bar{p}_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2}\bar{p}_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ \bar{p}_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2}\bar{p}_y^{\text{fixed},\gamma-P} & P < \gamma \leq P + R \end{cases}$$

$$m^{\text{dupli},\mu_*\gamma} = \begin{cases} m^{\text{fixed},\mu_* 0} & \gamma = 0 \\ m^{\text{fixed},\mu_*\gamma} & 0 < \gamma \leq P \\ m^{\text{fixed},\mu_*,\gamma-P} & P < \gamma \leq P + R \end{cases}$$

$$p_y^{\text{dupli},\gamma} = \begin{cases} p_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2}p_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ p_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2}p_y^{\text{fixed},\gamma-P} & P < \gamma \leq P + R \end{cases}$$

$$\text{along with} \quad p_{\mathbf{h}}(\gamma) = \begin{cases} p_{\mathbf{h}}^{\text{given}}(0) & \gamma = 0 \\ \frac{1}{2}p_{\mathbf{h}}^{\text{given}}(\gamma) & 0 < \gamma \leq R \\ p_{\mathbf{h}}^{\text{given}}(\gamma) & R < \gamma \leq P \\ \frac{1}{2}p_{\mathbf{h}}^{\text{given}}(\gamma - P) & P < \gamma \leq P + R, \end{cases}$$

where the hidden units $\gamma \in \{P + 1, ..., P + R\}$ and their corresponding order parameters are duplicates, or copies, of $\gamma \in \{1, ..., R\}$. $\gamma = 0$ can also be duplicated, but the result is less interesting. By definition (see Eqs. 4.26), $\zeta_y^\gamma\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right) = \lambda_y\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right) + \omega^\gamma\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right)$, where $\lambda_y\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right)$ and $\omega^\gamma\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right)$ are

the λ_y and ω^γ solving

$$\begin{aligned}
\omega^\gamma &= \frac{1}{p_{\mathbf{h}}(\gamma)} \sum_{y=0}^{C} \frac{\omega^\gamma}{\lambda_y + \omega^\gamma} \bar{p}_y^{\text{dupli},\gamma} \\
&= \frac{1}{p_{\mathbf{h}}^{\text{given}}(\gamma)} \sum_{y=0}^{C} \frac{\omega^\gamma}{\lambda_y + \omega^\gamma} \bar{p}_y^{\text{fixed},\gamma} \\
\lambda_y &= \frac{1}{p_{\mathbf{q}}(y)} \sum_{\gamma=0}^{P+R} \frac{\lambda_y}{\lambda_y + \omega^\gamma} \bar{p}_y^{\text{dupli},\gamma} \\
&= \frac{1}{p_{\mathbf{q}}(y)} \sum_{\gamma=0}^{P} \frac{\lambda_y}{\lambda_y + \omega^\gamma} \bar{p}_y^{\text{fixed},\gamma}.
\end{aligned}$$

Therefore, $\zeta_y^\gamma\left(\bar{\mathbf{p}}^{\text{dupli}}; p_{\mathbf{h}}\right) = \zeta_y^\gamma\left(\bar{\mathbf{p}}^{\text{fixed}}; p_{\mathbf{h}}^{\text{given}}\right)$, and the saddle-point equations simplify to

$$\begin{aligned}
\bar{x}_i^{\text{fixed},\mu} &= \frac{1}{2}\left(1 + \mathbb{I}\left(\mu > R\right)\right) \sum_{\mu_*=1}^{P^*} x_i^{*\mu_*} \sum_{y=0}^{C} q_y^{*\mu_*} \sigma_\mu\left(\beta_{\text{eff}} m^{\text{fixed},\mu_*} + \log\left[\mathbf{p}_y^{\text{fixed}}\right]\right) \\
\bar{p}_y^{\text{fixed},\gamma} &= \sum_{\mu_*=1}^{P^*} q_y^{*\mu_*} \sigma_\gamma\left(\beta_{\text{eff}} m^{\text{fixed},\mu_*} + \log\left[\mathbf{p}_y^{\text{fixed}}\right]\right) \\
m^{\text{fixed},\mu_*\mu} &= \varsigma\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{i=1}^{N}\left[\bar{\mathbf{x}}_i^{*\text{fixed},\mu}\right]^2}\right) \frac{\sum_{i=1}^{N} x_i^{*\mu_*} \bar{\mathbf{x}}_i^{*\text{fixed},\mu}}{\sqrt{\sum_{i=1}^{N}\left[\bar{\mathbf{x}}_i^{*\text{fixed},\mu}\right]^2}} \\
p_y^{\text{fixed},\gamma} &= \frac{\bar{p}_y^{\text{fixed},\gamma}}{\zeta_y^\gamma\left(\bar{\mathbf{p}}^{\text{fixed}}; p_{\mathbf{h}}^{\text{given}}\right)},
\end{aligned} \tag{4.28}$$

where $\mathbb{I}\left(\mu > R\right)$ is the indicator function equal to 1 when $\mu > R$ and 0 otherwise. Assume that $\varrho \to \infty$ so that $\varsigma\left(2\beta_{\text{eff}}\varrho\sqrt{\sum_{i=1}^{N}\left[\bar{x}_i^\mu\right]^2}\right) \to \mathbb{I}\left(\sqrt{\sum_{i=1}^{N}\left[\bar{x}_i^\mu\right]^2} > 0\right)$, then the saddle-point equations are the same no matter how the prefactor of $\frac{1}{2}\left(1 + \mathbb{I}\left(\mu > R\right)\right)$ affects the norm of $\bar{x}_i^\mu$, and Eqs. (4.13) are a fixed point of the saddle-point equations with $P + R$ hidden units. In particular, Eq. (4.13) is a stationary point of the loss (Eq. 4.4) when $\beta_{\text{eff}} = \beta$.

For the rest of this Appendix, all sums are understood as having the same bounds as Eqs. (4.28) and (4.10). The stability of any fixed point of the form $\mathbf{x} = \mathbf{F}(\mathbf{x})$, such as those of Eqs. (4.10), can be evaluated using the Jacobian matrix $\mathbf{J}$ of $\mathbf{F}$. If all eigenvalues λ of the Jacobian satisfy $|\lambda| < 1$, then the fixed point is stable. Conversely, if the Jacobian has an eigenvalue λ with $|\lambda| > 1$, then the fixed point is unstable. In particular, if the quadratic form $\mathbf{v}^T\mathbf{J}\mathbf{v}$ is larger than 1 for some $\mathbf{v}$ with $\|\mathbf{v}\| = 1$, then the fixed point is unstable. For the rest of this Appendix, we evaluate the stability of Eqs. (4.10) with duplicated order parameters. Keeping $\bar{\mathbf{p}}$

and $\mathbf{p}$ fixed, the Jacobian of the saddle-point equation for $\bar{\mathbf{x}}$ is

$$\begin{aligned}
\partial_{\bar{x}_k^\nu} \bar{x}_j^\mu &= \sum_{\mu_*} x_j^{*\mu_*} \partial_{\bar{x}_k^\nu} m^{\mu_* \nu} \sum_y q_y^{*\mu_*} \partial_{m^{\mu_* \nu}} \sigma_\mu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \\
&= \beta_{\text{eff}} \sum_{\mu_*} x_j^{*\mu_*} \partial_{\bar{x}_k^\nu} m^{\mu_* \nu} \sum_y q_y^{*\mu_*} \sigma_\mu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \left(\delta_{\mu\nu} - \sigma_\nu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \right) \\
&= \beta_{\text{eff}} \sum_{\mu_*} x_j^{*\mu_*} \frac{1}{\sqrt{\sum_i \left[\bar{x}_i^\nu \right]^2}} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^\nu \right] \tilde{x}_k^\nu \right) \\
&\quad \sum_y q_y^{*\mu_*} \sigma_\mu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \left(\delta_{\mu\nu} - \sigma_\nu \left(\beta_{\text{eff}} m^{\mu_*} + \log \left[\mathbf{p}_y \right] \right) \right),
\end{aligned}$$

where $\tilde{x}_j^\nu = \frac{\bar{x}_j^\nu}{\sqrt{\sum_i \left[\bar{x}_i^\nu \right]^2}}$. Suppose that $\bar{x}_i^\mu$, $\bar{p}_y^\gamma$, $m^{\mu_* \gamma}$ and p_y^γ are the duplicated parameters of Eq. (4.13) with $0 < \mu, \nu \leq R$ or $P < \mu, \nu \leq P + R$, then

$$\begin{aligned}
\partial_{\bar{x}_k^\nu} \bar{x}_j^\mu &= \beta_{\text{eff}} \sum_{\mu_*} x_j^{*\mu_*} \frac{1}{\sqrt{\sum_i \left[\bar{x}_i^\nu \right]^2}} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^\nu \right] \tilde{x}_k^\nu \right) \\
&\quad \sum_y q_y^{*\mu_*} \frac{1}{2} \sigma_{\theta(\mu)} \left(\beta_{\text{eff}} m^{\text{fixed},\mu_*} + \log \left[\mathbf{p}_y^{\text{fixed}} \right] \right) \left(\delta_{\mu\nu} - \frac{1}{2} \sigma_{\theta(\nu)} \left(\beta_{\text{eff}} m^{\text{fixed},\mu_*} + \log \left[\mathbf{p}_y^{\text{fixed}} \right] \right) \right), \\
&\text{where} \quad \theta(\mu) = \begin{cases} \mu & 0 < \mu \leq R \\ \mu - P & 0 < \mu \leq P + R. \end{cases}
\end{aligned}$$

If β_{eff} is relatively large (for instance of order $\sqrt{N}$), then the softmax $\sigma_{\theta(\mu)}$ splits the indices μ_* into a cover $\mathcal{S}$ of sets $\mathcal{S}(\mu)$ such that $\sigma_{\theta(\mu)} \left(\beta_{\text{eff}} m^{\text{fixed},\mu_*} + \log \left[\mathbf{p}_y \right] \right) \approx \mathbb{I} \left(\mu_* \in \mathcal{S}(\mu) \right)$. Therefore, we have

$$\begin{aligned}
\partial_{\bar{x}_k^\nu} \bar{x}_j^\mu \approx \beta_{\text{eff}} \sum_{\mu_*} x_j^{*\mu_*} \frac{1}{\sqrt{\sum_i \left[\bar{x}_i^\nu \right]^2}} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^\nu \right] \tilde{x}_k^\nu \right) \\
\frac{1}{2} \mathbb{I} \left(\mu_* \in \mathcal{S}(\mu) \right) \left(\delta_{\mu\nu} - \frac{1}{2} \mathbb{I} \left(\mu_* \in \mathcal{S}(\nu) \right) \right).
\end{aligned}$$

The subcovers $\mathcal{S}(0 < \mu \leq P)$ and $\mathcal{S}(R < \mu \leq P + R)$ are both partitions of the indices μ_*. As such, $\left[\partial_{\bar{x}_k^\nu} \bar{x}_j^\mu \right]_{\mu,\nu=1}^{P+R}$ is block-diagonal with 2×2 blocks coupling the indices $0 < \mu \leq R$ and $\nu = \mu + P$. Without loss of generality, we investigate the stability of the block

$$\begin{aligned}
&\begin{bmatrix} \partial_{\bar{x}_k^1} \bar{x}_j^1 & \partial_{\bar{x}_k^{P+1}} \bar{x}_j^1 \\ \partial_{\bar{x}_k^1} \bar{x}_j^{P+1} & \partial_{\bar{x}_k^{P+1}} \bar{x}_j^{P+1} \end{bmatrix} \\
&= \frac{1}{2} \frac{\beta_{\text{eff}}}{\sqrt{\sum_i \left[x_i'^1 \right]^2}} \sum_{\mu_* \in \mathcal{S}(1)} \begin{bmatrix} x_j^{*\mu_*} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^1 \right] \tilde{x}_k^1 \right) & -x_j^{*\mu_*} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^1 \right] \tilde{x}_k^1 \right) \\ -x_j^{*\mu_*} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^1 \right] \tilde{x}_k^1 \right) & x_j^{*\mu_*} \left(x_k^{*\mu_*} - \left[\sum_i x_i^{*\mu_*} \tilde{x}_i^1 \right] \tilde{x}_k^1 \right) \end{bmatrix}.
\end{aligned}$$

where $x_i'^1 = \sum_{\mu_* \in \mathcal{S}(1)} x_i^{*\mu_*}$. Let $\mathbf{u}^1$ be an arbitrary vector orthogonal to $\tilde{\mathbf{x}}^1$ and define

$$v_j^\mu = \frac{1}{\sqrt{2}} \begin{cases} u_j^1 & \mu = 1 \\ -u_j^1 & \mu = P+1, \end{cases}$$

then we have the quadratic form

$$\sum_{\nu \in \{1, P+1\}} \sum_{k=1}^{N} v_k^\nu \, \partial_{\bar{x}_k^\nu} \bar{x}_j^\mu = \frac{\beta_{\text{eff}}}{\sqrt{\sum_i \left[x_i'^1\right]^2}} \sum_{\mu_* \in \mathcal{S}(1)} \frac{1}{\sqrt{2}} \begin{bmatrix} x_j^{*\mu_*} \sum_k x_k^{*\mu_*} u_k^1 \\ -x_j^{*\mu_*} \sum_k x_k^{*\mu_*} u_k^1 \end{bmatrix}$$

$$\sum_{\mu, \nu \in \{1, P+1\}} \sum_{j,k=1}^{N} v_j^\mu v_k^\nu \, \partial_{\bar{x}_k^\nu} \bar{x}_j^\mu = \frac{\beta_{\text{eff}}}{\sqrt{\sum_i \left[x_i'^1\right]^2}} \sum_{\mu_* \in \mathcal{S}(1)} \left[\sum_j x_j^{*\mu_*} u_j^1 \right]^2 .$$

If $\mathbf{u}^1$ is orthogonal to all the $\mathbf{x}^{*\mu_*}$ such that $\mu_* \in \mathcal{S}(1)$, and in particular if $\mathcal{S}(1)$ contains a single pattern $\mathbf{x}^{*\mu_*} = \tilde{\mathbf{x}}^1$, then $\sum_{\mu_* \in \mathcal{S}(1)} \left[\sum_j x_j^{*\mu_*} u_j^1\right]^2 = 0$, so the quadratic form vanishes. In this case, $\mathbf{u}^1$ is a stable direction of the saddle-point equations. Otherwise, the quadratic form does not vanish, so there is a β_{split} such that the direction $\mathbf{u}^1$ is unstable when $\beta_{\text{eff}} > \beta_{\text{split}}$.

4.H Splitting steepest descent

Steps 5 and 11 of splitting steepest descent (Alg. 1) involve the splitting matrices $\mathcal{S}_\mu(\mathbf{w}, \mathbf{p})$ derived in [87]. In this appendix, we explain their role in the algorithm. Consult [87] for a detailed explanation of their interpretation and theoretical underpinnings.

Define the thresholds $\tau_{\text{thres}} \in (0, 1]$ and $\lambda_{\text{thres}} \leq 0$. At step 5 of Alg. (1), we duplicate the hidden units corresponding to the $R \leq \min\{\tau_{\text{thres}} P_{\text{cur}}, P_{\max} - P_{\text{cur}}\}$ most negative minimum eigenvalues $\lambda_{\min}^\mu$ of the splitting matrices $\mathcal{S}_\mu(\mathbf{w}, \mathbf{p})$ such that $\lambda_{\min}^\mu \leq \lambda_{\text{thres}}$ [87]. To make Fig. (4.6), we pick $\tau_{\text{thres}} = 1$ and $\lambda_{\text{thres}} \approx 0$. Our splitting matrices are

$$\mathcal{S}_\mu(\mathbf{w}, \mathbf{p}) = -\mathbb{E}_{\mathbf{x}^*, y^*} \left[\frac{p_y^\mu \bar{\nabla}_{\mathbf{w}^\mu} \bar{\nabla}_{\mathbf{w}^\mu}^T \exp\left(\beta_{\text{eff}} \sum_{i=1}^N w_i^\mu x_i\right)}{\sum_{\nu=1}^P p_y^\nu \exp\left(\beta_{\text{eff}} \sum_{i=1}^N w_i^\nu x_i\right) + p_y^0 \frac{\Omega_N(\beta)}{\Omega_N(0)}} \right],$$

where $\bar{\nabla}_{\boldsymbol{\theta}}$ is the gradient constrained to the unit hypersphere S^{N-1} (see Appendix 4.F) and β_{eff} can be written explicitly in terms of β as $\beta_{\text{eff}} = \varsigma\beta$ (see Eqs. 4.12 and 4.10). At step 11 of Alg. (1), we break permutation symmetries in the memories $\mathbf{w}^\mu$ by descending along the eigenvectors $\mathbf{u}^\mu \in S^{N-1}$ corresponding to the eigenvalues $\lambda_{\min}^\mu$. To be more precise, we update the memories $\mathbf{w}^\mu$ and their duplicates $\mathbf{w}^{\text{dupli},\mu}$ according to $\mathbf{w}^\mu \leftarrow \mathbf{w}^\mu + \delta\mathbf{u}^\mu$ and $\mathbf{w}^{\text{dupli},\mu} \leftarrow \mathbf{w}^{\text{dupli},\mu} - \delta\mathbf{u}^\mu$, respectively, where δ is a relatively small learning rate [87]. In physics terminology, the eigenvectors $\mathbf{u}^\mu$ are excitation modes that break the permutation symmetries of the memories. N and P are generally large, so it is prohibitively expensive to store $\mathcal{S}_\mu(\mathbf{w}, \mathbf{p})$ explicitly for all hidden units $1 \leq \mu \leq P$. Therefore, as proposed in [88], we find the eigenvectors $\mathbf{u}^\mu$ and their eigenvalues $\lambda_{\min}^\mu$ by minimizing the Rayleigh quotients $Q_\mu[\mathbf{w}, \mathbf{p}] : S^{N-1} \ni \mathbf{u}^\mu \mapsto [\mathbf{u}^\mu]^T \mathcal{S}_\mu(\mathbf{w}, \mathbf{p}) \mathbf{u}^\mu$, which can be evaluated without constructing $\mathcal{S}_\mu(\mathbf{w}, \mathbf{p})$ explicitly. To derive $Q_\mu[\mathbf{w}, \mathbf{p}]$, we first compute

$\bar{\nabla}_{\mathbf{w}^\mu}\bar{\nabla}^T_{\mathbf{w}^\mu}\exp\left(\beta_{\text{eff}}\sum_i w_i^\mu x_i\right)$. Given $f\left(\boldsymbol{\theta}\right)=\sum_i\theta_i x_i$, we find

$$\bar{\nabla}_{\boldsymbol{\theta}}\bar{\nabla}^T_{\boldsymbol{\theta}}\exp\left(\beta_{\text{eff}}f\left(\boldsymbol{\theta}\right)\right)=\exp\left(\beta_{\text{eff}}f\left(\boldsymbol{\theta}\right)\right)\left(\beta_{\text{eff}}\bar{\nabla}_{\boldsymbol{\theta}}\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)+\beta^2_{\text{eff}}\bar{\nabla}_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)\right).$$

As mentioned at the beginning of Appendix 4.F, $\bar{\nabla}_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)$ is the unrestricted gradient of $f\left(\boldsymbol{\theta}\right)$ projected onto the tangent space of $\boldsymbol{\theta}$. It remains to calculate $\bar{\nabla}_{\boldsymbol{\theta}}\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)$. Following [235], we find it to be the iterated projected gradient

$$\begin{aligned}\left[\bar{\nabla}_{\boldsymbol{\theta}}\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)\right]_{i\ell}&=\sum_k\left(\delta_{k\ell}-\theta_k\theta_\ell\right)\partial_{\theta_i}\left(\sum_j\left(\delta_{jk}-\theta_j\theta_k\right)\partial_{\theta_j}\left[\sum_h\theta_h x_h\right]\right)\\&=\sum_k\left(\delta_{k\ell}-\theta_k\theta_\ell\right)\partial_{\theta_i}\left(x_k-\left[\sum_j\theta_j x_j\right]\theta_k\right)\\&=-\sum_k\left(\delta_{k\ell}-\theta_k\theta_\ell\right)\left(x_i\theta_k+\left[\sum_j\theta_j x_j\right]\delta_{ik}\right)\\&=-\left[\sum_j\theta_j x_j\right]\left(\delta_{i\ell}-\theta_i\theta_\ell\right),\end{aligned}$$

so we obtain

$$Q_\mu\left[\mathbf{w},\mathbf{p}\right]\left(\mathbf{u}^\mu\right)=-\mathbb{E}_{\mathbf{x}^*,y^*}\left[\frac{p_y^\mu\exp\left(\beta_{\text{eff}}\sum_{i=1}^N w_i^\mu x_i\right)}{\sum_{\nu=1}^P p_y^\nu\exp\left(\beta_{\text{eff}}\sum_{i=1}^N w_i^\nu x_i\right)+p_y^0\frac{\Omega_N(\beta)}{\Omega_N(0)}}F\left(\mathbf{u}^\mu;\mathbf{w}^\mu,\mathbf{x}\right)\right]$$

$$\begin{aligned}\text{where}\quad F\left(\boldsymbol{\varphi};\boldsymbol{\theta},\mathbf{x}\right)&=\boldsymbol{\varphi}^T\left[\beta_{\text{eff}}\bar{\nabla}_{\boldsymbol{\theta}}\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)+\beta^2_{\text{eff}}\bar{\nabla}_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)\bar{\nabla}^T_{\boldsymbol{\theta}}f\left(\boldsymbol{\theta}\right)\right]\boldsymbol{\varphi}\\&=\beta^2_{\text{eff}}\left(\sum_k\varphi_k x_k-\left[\sum_k\varphi_k\theta_k\right]\left[\sum_j\theta_j x_j\right]\right)^2\\&\quad+\beta_{\text{eff}}\left[\sum_j\theta_j x_j\right]\left[\sum_i\varphi_i\theta_i-1\right]\left[\sum_\ell\varphi_\ell\theta_\ell+1\right].\end{aligned}$$

We can directly minimize $Q\left[\mathbf{w},\mathbf{p}\right]\left(\mathbf{u}\right)=\sum_\mu Q_\mu\left[\mathbf{w},\mathbf{p}\right]\left(\mathbf{u}^\mu\right)$ to find the set of all eigenvectors $\mathbf{u}^\mu$ simultaneously. However, it is more convenient to integrate the eigenvector calculations into the DAM architecture. As such, we define the modified loss

$$\begin{aligned}\mathcal{L}_\epsilon\left(\mathbf{w},\mathbf{p},\mathbf{u}\right)&=-\log\left[\sum_{\mu=1}^P p_y^\mu\exp\left(\varsigma\beta\sum_{i=1}^N w_i^\mu x_i\right)\left[1+\epsilon F\left(\mathbf{u}^\mu;\mathbf{w}^\mu,\mathbf{x}\right)\right]+p_y^0\frac{\Omega_N\left(\beta\right)}{\Omega_N\left(0\right)}\right]\\&=-\log\left[\sum_{\mu=1}^P p_y^\mu\exp\left(\varsigma\beta\sum_{i=1}^N w_i^\mu x_i+\log\left[1+\epsilon F\left(\mathbf{u}^\mu;\mathbf{w}^\mu,\mathbf{x}\right)\right]\right)+p_y^0\frac{\Omega_N\left(\beta\right)}{\Omega_N\left(0\right)}\right],\end{aligned}$$

so that the four gradients used to trained the DAM can be calculated using the equations

$$\begin{aligned}
\nabla_{\mathbf{u}} \mathcal{Q}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}\right) &= \lim_{\epsilon \to 0} \left\{ \frac{1}{\epsilon} \nabla_{\mathbf{u}} \mathcal{L}_{\epsilon}\left(\mathbf{w}, \mathbf{p}, \mathbf{u}\right) \right\}, \\
\nabla_{\mathbf{w}} \mathcal{L}\left(\mathbf{w}, \mathbf{p}\right) &= \nabla_{\mathbf{w}} \mathcal{L}_0\left(\mathbf{w}, \mathbf{p}, \mathbf{u}\right) \\
\nabla_{\mathbf{p}} \mathcal{L}\left(\mathbf{w}, \mathbf{p}\right) &= \nabla_{\mathbf{p}} \mathcal{L}_0\left(\mathbf{w}, \mathbf{p}, \mathbf{u}\right) \\
\text{and} \quad \nabla_{\beta} \mathcal{L}\left(\mathbf{w}, \mathbf{p}\right) &= \nabla_{\beta} \mathcal{L}_0\left(\mathbf{w}, \mathbf{p}, \mathbf{u}\right).
\end{aligned}$$

This technique is based on the automatic differentiation trick proposed in [88]. We numerically implement the limit in the first equation by setting $\epsilon = 0$ during loss evaluation and $\epsilon = 1$ during gradient computation. When we optimize $\mathcal{Q}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}\right)$, the minima of the different Rayleigh quotients $Q_{\mu}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}^{\mu}\right)$ can span multiple orders of magnitude, so we normalize the gradients $\nabla_{\mathbf{u}^{\mu}} Q_{\mu}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}^{\mu}\right)$ by a running average of their magnitudes to facilitate convergence. This is inspired by the use of RMSProp in [88]. Moreover, we constrain the eigenvectors $\mathbf{u}^{\mu}$ and the gradients $\nabla_{\mathbf{u}^{\mu}} Q_{\mu}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}^{\mu}\right)$ to the unit hypersphere as we do for $\mathbf{w}^{\mu}$ (see Appendix 4.F). To make Fig. (4.6), we train the eigenvectors for 1 epoch, during which we monitor $\min_{\mu} \left\{ \frac{1}{2} \left[\mathbf{u}^{\mu}\right]^{T} \nabla_{\mathbf{u}^{\mu}} \mathcal{Q}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}^{\mu}\right) \right\} = \min_{\mu} \left\{ \mathcal{Q}\left[\mathbf{w}, \mathbf{p}\right]\left(\mathbf{u}^{\mu}\right) \right\} \sim \min_{\mu} \left\{ \lambda^{\mu}_{\text{min}} \right\}$ as a metric. The Rayleigh quotients converge very quickly, and using a higher patience does not seem to improve the quality of the learned eigenvectors.

4.I Initialization and learning rate

To make all our figures, except Fig. (4.5), we initialize the memories $\mathbf{w}^{\mu}$ uniformly at random on the unit hypersphere S^{N-1} [246]. To make Fig. 4.5, we instead use the algorithm of [247] to sample the initial memories $\mathbf{w}^{\mu}$ from a vMF distribution (see Appendix 4.C) with mean direction $\tilde{\mathbf{x}}^{*} = \bar{\mathbf{x}}^{*} / \|\bar{\mathbf{x}}^{*}\|$, where $\bar{\mathbf{x}}^{*} = \frac{1}{N} \sum_{\mu_{*}=1}^{P^{*}} \mathbf{x}^{*\mu_{*}}$ is the mean of the patterns $\mathbf{x}^{*\mu_{*}}$ and $\|\bar{\mathbf{x}}^{*}\| = \sqrt{\sum_{i=1}^{N} \left[\bar{x}_{i}^{*}\right]^{2}}$. By construction, the mean direction $\tilde{\mathbf{x}}^{*}$ is the ordered solution of Eq. (4.10) with the most permutation symmetries, or in other words the root of the learning dynamics tree shown in Fig. 4.5, so initializing the memories around it helps reveal the tree structure.

We use a learning rate of $\eta \sim 0.1$ to train the memories $\mathbf{w}^{\mu}$. Based on our experience, $\mathbf{p}$ trains well when its own learning rate is approximately $\left(1 + \frac{p_{\mathbf{h}}(0)}{1 - p_{\mathbf{h}}(0)}\right) P$ times smaller. Without this rescaling, the multiplicative update factor $\exp\left(\eta \partial_{p_{y}^{\gamma}} L\left(\mathbf{w}, \mathbf{p}\right)\right)$ described in Appendix 4.F can be relatively large even when η is relatively small, making it ill-behaved. Equivalently, we can also train $g_{y}^{\gamma} = \left(1 + \frac{p_{\mathbf{h}}(0)}{1 - p_{\mathbf{h}}(0)}\right) P \mathbf{p}$ with the same learning rate η as $\mathbf{w}$. We adopt this approach in our code available at [199].

Let $P_0 = \frac{p_{\mathbf{h}}(0)}{1 - p_{\mathbf{h}}(0)} P$ so that $g_{y}^{\gamma} = \left(P + P_0\right) p_{y}^{\gamma}$. In terms of $g_{y}^{\text{fixed},\gamma} = \left(P + P_0\right) p_{y}^{\text{fixed},\gamma}$, the duplicated parameters (Eq. 4.13) of the saddle-point hierarchy principle are

$$\bar{x}_{i}^{\text{dupli},\mu} = \begin{cases} \bar{x}_{i}^{\text{fixed},\mu} & 0 < \mu \leq P \\ \bar{x}_{i}^{\text{fixed},\mu - P} & P < \mu \leq P + R \end{cases}$$

$$\bar{g}_y^{\text{dupli},\gamma} = \frac{P+R}{P} \begin{cases} \bar{g}_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2}\bar{g}_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ \bar{g}_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2}\bar{g}_y^{\text{fixed},\gamma-P} & P < \gamma \leq P+R \end{cases}$$

$$m^{\text{dupli},\mu_*\gamma} = \begin{cases} m^{\text{fixed},\mu_* 0} & \gamma = 0 \\ m^{\text{fixed},\mu_*\gamma} & 0 < \gamma \leq P \\ m^{\text{fixed},\mu_*,\gamma-P} & P < \gamma \leq P+R \end{cases}$$

$$g_y^{\text{dupli},\gamma} = \frac{P+R}{P} \begin{cases} g_y^{\text{fixed},0} & \gamma = 0 \\ \frac{1}{2}g_y^{\text{fixed},\gamma} & 0 < \gamma \leq R \\ g_y^{\text{fixed},\gamma} & R < \gamma \leq P \\ \frac{1}{2}g_y^{\text{fixed},\gamma-P} & P < \gamma \leq P+R \end{cases}$$

$$\text{along with} \quad g_{\mathbf{h}}(\gamma) = \frac{P+R}{P} \begin{cases} g_{\mathbf{h}}^{\text{given}}(0) & \gamma = 0 \\ \frac{1}{2}g_{\mathbf{h}}^{\text{given}}(\gamma) & 0 < \gamma \leq R \\ g_{\mathbf{h}}^{\text{given}}(\gamma) & R < \gamma \leq P \\ \frac{1}{2}g_{\mathbf{h}}^{\text{given}}(\gamma - P) & P < \gamma \leq P+R, \end{cases}$$

where $g_{\mathbf{h}}^{\text{given}}(\gamma) = (P+P_0)\, p_{\mathbf{h}}^{\text{given}}(\gamma)$ and $g_{\mathbf{h}}(\gamma) = (P+P_0)\, p_{\mathbf{h}}(\gamma)$. The newly introduced scaling factor of $\frac{P+R}{P}$ must be taken into account in our implementation of splitting steepest descent (Alg. 1), which gives Alg. (2), shown below.

Algorithm 2 Rescaled splitting steepest descent [87, 88]

1: Preallocate space for a DAM with P_{max} hidden units and the corresponding weights $\mathbf{w}$ and $\mathbf{g}$
2: Initialize the weights $\mathbf{w}^\mu$ and $\mathbf{g}^\mu$ connected to the P_{cur} first hidden units $\mu \in \{1, ..., P_{\text{cur}}\}$, as well as $\mathbf{g}^0$
3: min $L(\mathbf{w}, \mathbf{g})$ with SGD
4: **while** $P_{\text{cur}} < P_{\text{max}}$ **do**
5: Identify a subset $\boldsymbol{\mu}_{\text{copy}} \subseteq \{1, ..., P_{\text{cur}}\}$ of $R \leq P_{\text{max}} - P_{\text{cur}}$ hidden units to split, **return** if empty
6: Let $\boldsymbol{\mu}_{\text{paste}} = \{P_{\text{cur}}+1, ..., P_{\text{cur}}+R\}$ and $\boldsymbol{\mu}_{\text{dupli}} = \{1, ..., P_{\text{cur}}+R\}$
7: Build weights $\mathbf{w}^{\boldsymbol{\mu}_{\text{paste}}} = \mathbf{w}^{\boldsymbol{\mu}_{\text{copy}}}$ for $\boldsymbol{\mu}_{\text{paste}}$
8: Rescale $\mathbf{g}^{\boldsymbol{\mu}_{\text{copy}}} \leftarrow \mathbf{g}^{\boldsymbol{\mu}_{\text{copy}}}/2$ and $g_{\mathbf{h}}(\boldsymbol{\mu}_{\text{split}}) \leftarrow g_{\mathbf{h}}(\boldsymbol{\mu}_{\text{split}})/2$
9: Build weights $\mathbf{g}^{\boldsymbol{\mu}_{\text{paste}}} = \mathbf{g}^{\boldsymbol{\mu}_{\text{copy}}}$ and $g(\boldsymbol{\mu}_{\text{paste}}) = g(\boldsymbol{\mu}_{\text{copy}})$ for $\boldsymbol{\mu}_{\text{paste}}$
10: Rescale $\mathbf{g}^{\boldsymbol{\mu}_{\text{dupli}}} \leftarrow \frac{P_{\text{cur}}+R}{P_{\text{cur}}}\mathbf{g}^{\boldsymbol{\mu}_{\text{dupli}}}$ and $g_{\mathbf{h}}(\boldsymbol{\mu}_{\text{dupli}}) \leftarrow \frac{P_{\text{cur}}+R}{P_{\text{cur}}} g_{\mathbf{h}}(\boldsymbol{\mu}_{\text{dupli}})$
11: Update $P_{\text{cur}} \leftarrow P_{\text{cur}} + R$
12: Escape the saddle point by 2^{nd} order descent of $L(\mathbf{w}, \mathbf{g})$ w.r.t. $\mathbf{w}$
13: min $L(\mathbf{w}, \mathbf{g})$ with SGD
14: **end while**
15: **return** ▷ See Appendix 4.H and [87, 88] for details about steps 5 and 12

We initialize the weights $\mathbf{g}^\gamma$ and the hidden unit distribution $g_{\mathbf{h}}(\gamma)$ according to

$$g_y^\gamma = \begin{cases} P_0 p_{\mathbf{q}}(y) & \gamma = 0 \\ p_{\mathbf{q}}(y) & \gamma > 0 \end{cases}$$

$$\text{and} \quad g_{\mathbf{h}}(\gamma) = \begin{cases} \frac{P_0}{P+P_0} & \gamma = 0 \\ \frac{1}{P+P_0} & \gamma > 0. \end{cases}$$

When we train our DAM without splitting steepest descent, we set $P_0 = 1$. The resulting weights are, in some sense, the most "typical" [248, 249] for the constraints $g_y^\gamma \geq 0$, $\sum_{\gamma=0}^{P} g_y^\gamma = (P+1)\, p_{\mathbf{q}}(y)$ and $\sum_{y=0}^{C} g_y^\gamma = (P+1)\, p_{\mathbf{h}}(\gamma) = 1$ (see Section 4.2). With splitting steepest descent, we set $P_0 = P_{\text{cur}}/P_{\text{final}}$ to ensure that $g_{\mathbf{h}}(0) = \sum_{y=0}^{C} g_y^0$ is approximately the same size as the other entries of $g_{\mathbf{h}}(\gamma)$ after training. For simplicity, we set $p_{\mathbf{q}}(y)$ to the proportions of classes in the data during supervised training. For unsupervised training (see Section 4.4.2), we break the permutation symmetry between the different classes by using different values for all the entries of $p_{\mathbf{q}}(y)$. Otherwise, class weights with unbroken permutation symmetries stay stuck at their initial conditions.

4.J Weights learned with unsupervised training

This Appendix contains plots of DAM weights learned in an unsupervised way (see Section 4.4.2) and sorted in increasing $y = \text{argmax}_{y'} \left\{ p_{y'}^\mu \right\}$. They are not in the main text because they take a lot of space.

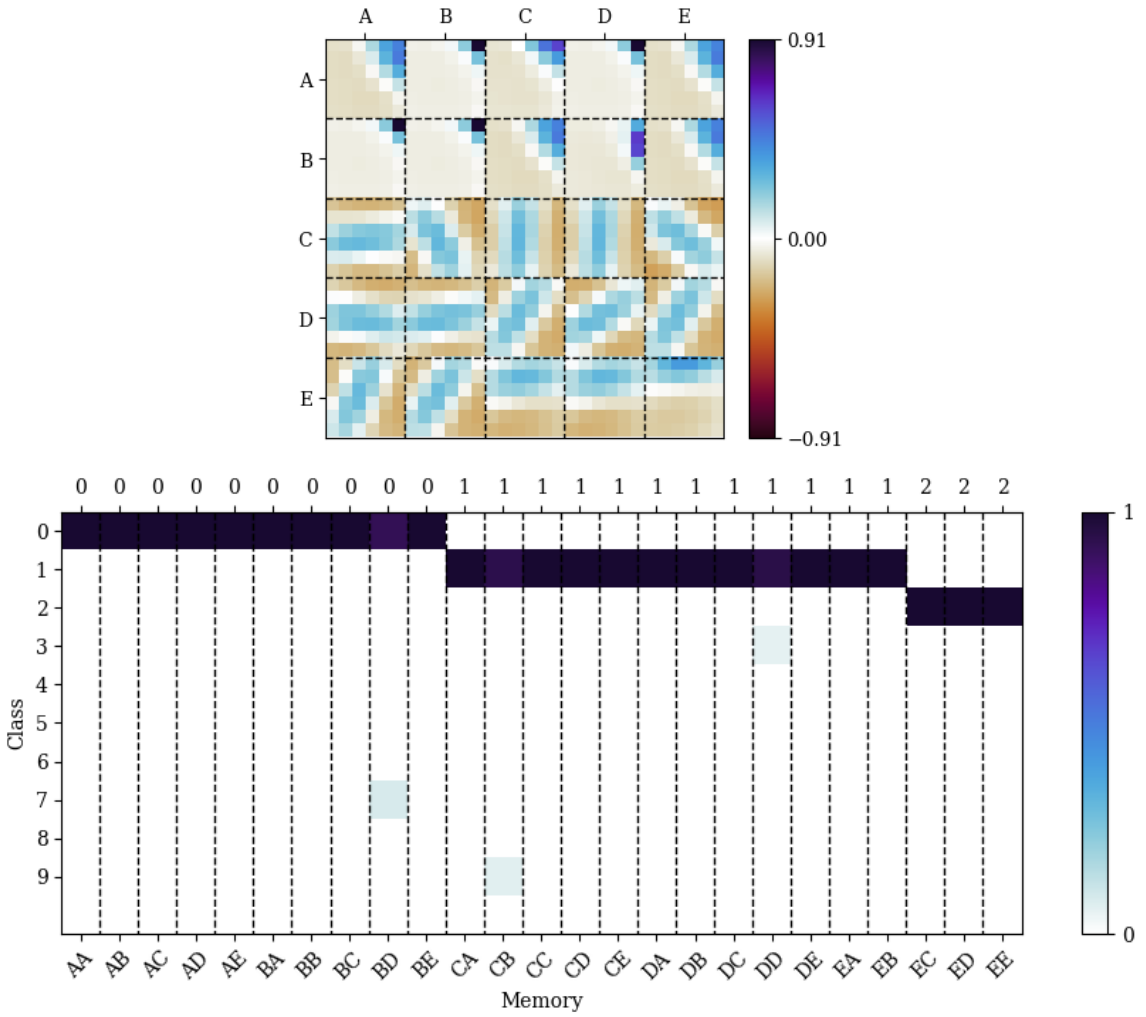

Figure 4.7: In the top panel, 25 of the $P = 100$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained in an unsupervised way (Eq. 4.15) on 6×6 patches of the MNIST dataset of handwritten digits [8] while assuming $C = 10$ latent classes and $\varsigma = 0.6$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices. $\mathbf{w}^\mu$ and $\mathbf{p}^\mu$ are sorted in increasing $y = \text{argmax}_{y'} \left\{ p^\mu_{y'} \right\}$, and this figure shows $1 \leq \mu \leq 25$.

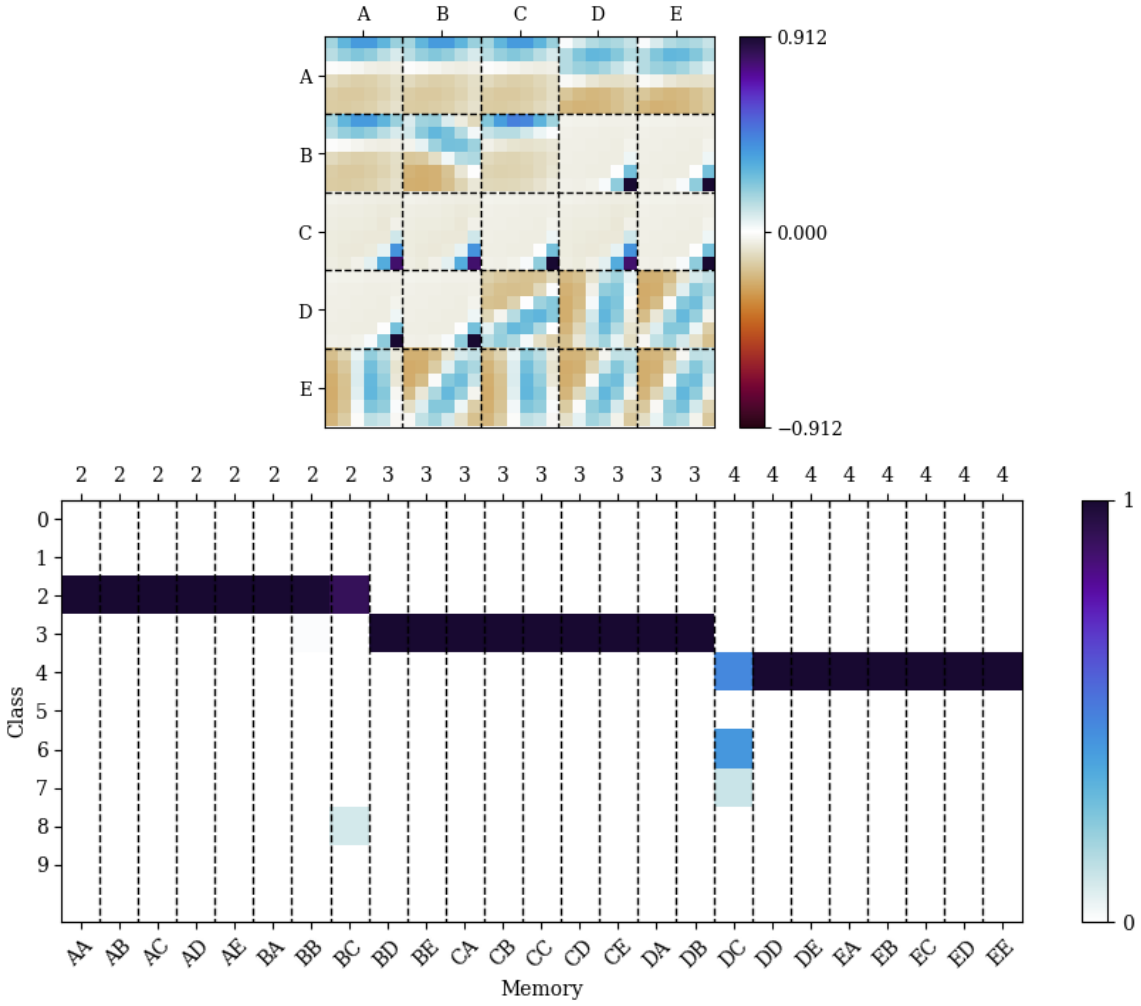

Figure 4.8: In the top panel, 25 of the $P = 100$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained in an unsupervised way (Eq. 4.15) on 6×6 patches of the MNIST dataset of handwritten digits [8] while assuming $C = 10$ latent classes and $\varsigma = 0.6$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices. $\mathbf{w}^\mu$ and $\mathbf{p}^\mu$ are sorted in increasing $y = \text{argmax}_{y'} \left\{ p^\mu_{y'} \right\}$, and this figure shows $26 \leq \mu \leq 50$.

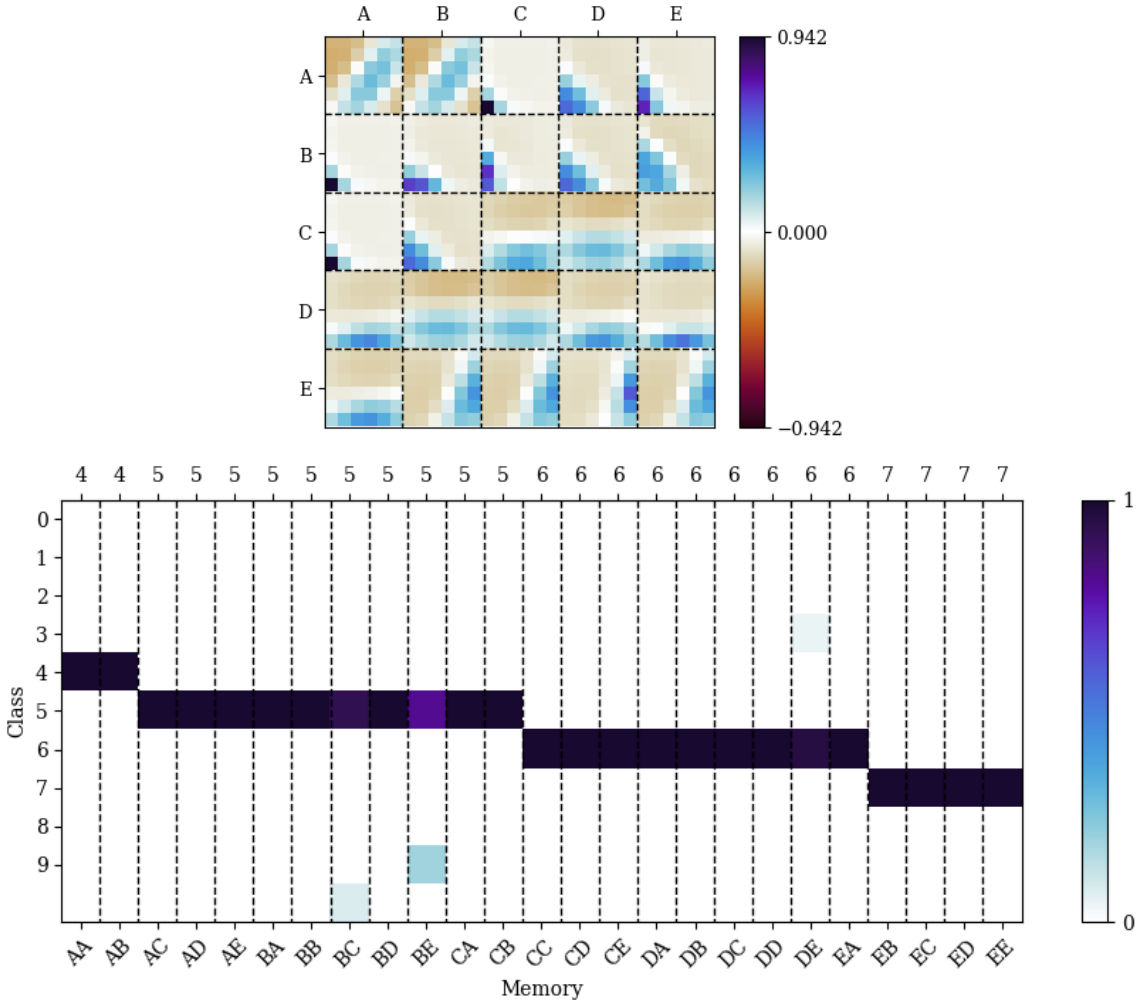

Figure 4.9: In the top panel, 25 of the $P = 100$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained in an unsupervised way (Eq. 4.15) on 6×6 patches of the MNIST dataset of handwritten digits [8] while assuming $C = 10$ latent classes and $\varsigma = 0.6$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices. $\mathbf{w}^\mu$ and $\mathbf{p}^\mu$ are sorted in increasing $y = \text{argmax}_{y'} \left\{ p^\mu_{y'} \right\}$, and this figure shows $51 \leq \mu \leq 75$.

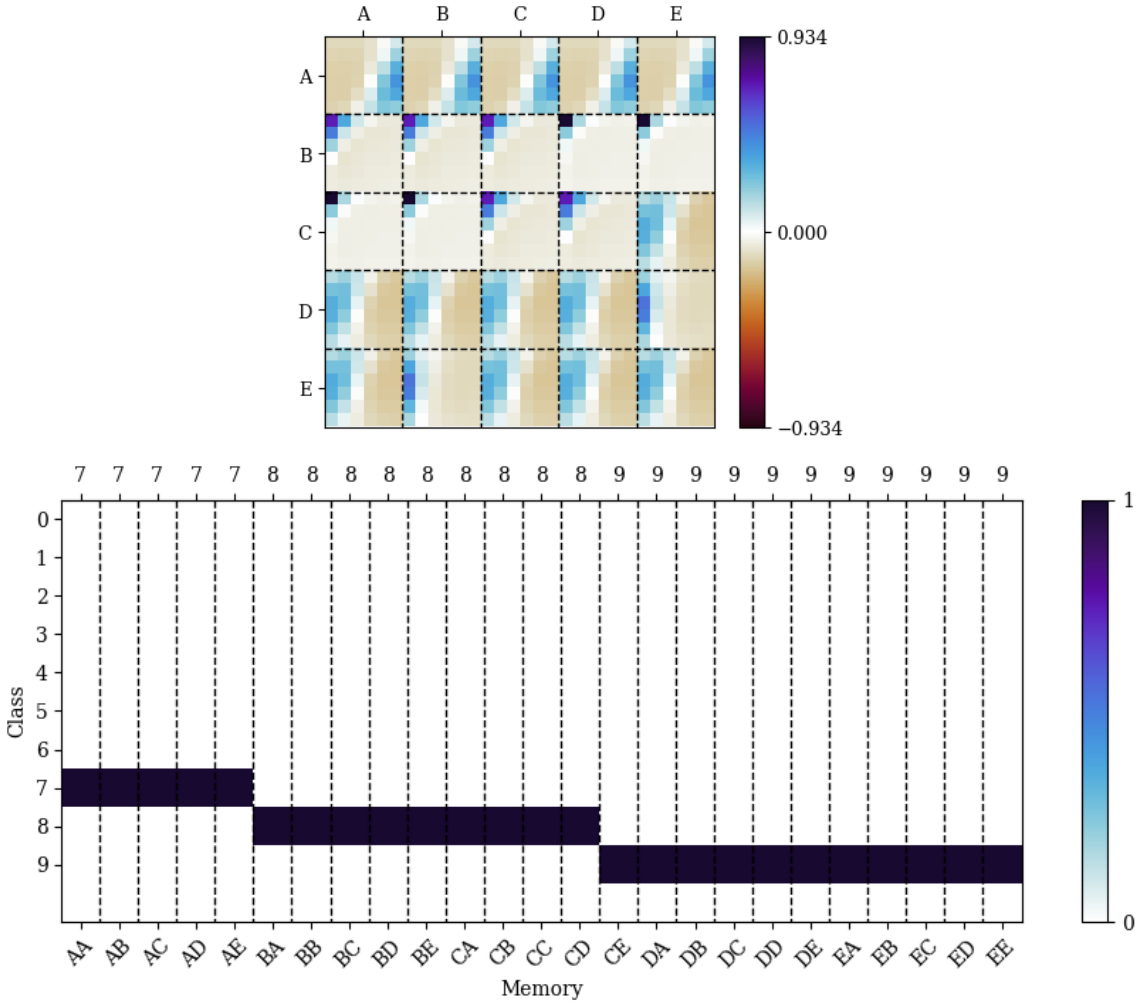

Figure 4.10: In the top panel, 25 of the $P = 100$ memories $\mathbf{w}^\mu$ learned by an instance of our dense associative memory (DAM) model trained in an unsupervised way (Eq. 4.15) on 6×6 patches of the MNIST dataset of handwritten digits [8] while assuming $C = 10$ latent classes and $\varsigma = 0.6$. In the bottom panel, the corresponding rescaled class weights $\mathbf{p}^\mu / p_{\mathbf{h}}(\mu)$, where $p_{\mathbf{h}}(\gamma) = \frac{1}{P+1}$ for all $0 \leq \gamma \leq P$. The hidden units are indexed using pairs of letters from A to E, and the column-wise maxima of the class weights are the classes of the memories with the corresponding letter indices. $\mathbf{w}^\mu$ and $\mathbf{p}^\mu$ are sorted in increasing $y = \text{argmax}_{y'} \left\{ p^\mu_{y'} \right\}$, and this figure shows $76 \leq \mu \leq 100$.

Conclusion

In this thesis, we study various kinds of dense associative memory (DAM) models and restricted Boltzmann machines (RBMs) in the teacher-student setting.

In Chapter 2, we study a DAM called dense Hopfield network (dense HN). On the Nishimori line, we show that the phase transition where student dense HNs with a single pattern become capable of learning a dataset with M samples generated by a teacher dense HN with a single pattern coincides with the spin-glass transition of dense HNs with M random patterns. Outside the Nishimori line, we study the resistance of dense HNs to noise and adversarial attacks. In particular, we derive a formula quantifying the adversarial robustness of dense HNs at zero temperature, and we clarify why the adversarial robustness of dense HNs depends on the learning regime, as observed in [36].

In Chapter 3, we study RBMs with any finite number of hidden units. When the patterns incident to the P^* hidden units of the teacher RBM are uncorrelated, we show that a student RBM with P hidden units learns the data in the same way as P separate RBMs with one hidden unit each, thus validating a conjecture formulated in [39]. Moreover, we show that student RBMs with more hidden units than their teacher learn a representation of the data where exactly P^* of their patterns align themselves with those of the teacher while the $P - P^*$ remaining ones freeze in spin-glass states. We then argue that such RBMs can be used as toy models to study the lottery ticket hypothesis [9]. When the teacher patterns are correlated, we show that the student can adopt different learning strategies depending on the hyperparameters of the teacher-student setting and the number of samples in the training dataset.

In Chapter 4, we study a kind of DAM that fits into the framework of RBMs and is capable of both supervised and unsupervised classification. Based on insights from our theoretical analysis of the teacher-student setting, we propose a novel regularized loss function that makes training significantly more stable. Moreover, we show that the fixed points of relatively small DAMs are saddle points of larger DAMs, and we leverage this saddle-point hierarchy to considerably accelerate training using the splitting steepest descent algorithm introduced in [87].

A natural continuation of this work would be to study adversarial attacks in DAMs with many hidden units, such as the model studied in Chapter 4 [47]. In particular, it would be interesting to see if the methods that we use to study adversarial attacks in Chapter 2 can be generalized and used for this purpose in combination with those used to study adversarial attacks in linear models [18, 19, 20].

One could also extend this work by studying DAM and RBM training dynamics with analytical calculations, the statistical mechanics approach of [31, 32] being a possible starting point. First, it would be interesting to build a theoretical model of the training dynamics of the lottery ticket experiment presented in Chapter 3 [46],

as it could provide crucial insight into the lottery ticket hypothesis. In particular, it could help us to understand if the initial conditions that make certain subnetworks of large NNs converge especially quickly have common properties that can be exploited to design novel initialization schemes leading to faster training. Second, it would also be interesting to study analytically the learning dynamics of the DAM studied in Chapter 4 [47], particularly near saddle points, for a rigorous comparison of the standard training time with that of splitting steepest descent.

To further our study of implicitly low-dimensional learning, it would be interesting to investigate models with low-rank weights, such as in [31, 32]. For example, one could study the teacher-student setting in which a student RBM (or DAM) with weights constrained to have a low rank learns data generated by a teacher whose weights are structured in a similar way for representing data lying on a low-dimensional manifold [16, 17]. This kind of study could be used to investigate the benefits of low-rank constraints in exploiting the structure of data to improve the speed and quality of learning.

Finally, the interpretable learning dynamics [31, 32, 34] of RBMs and DAMs make them promising surrogate models of biological systems. It was notably shown that DAM learning dynamics is similar to cellular differentiation [34], and it would be interesting to see if the splitting steepest descent algorithm used to train DAMs in Chapter 4 [47] can be related to cellular division in an analogous way.

Funding information

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. The work was also supported by the project PRIN22TANTARI ”Statistical Mechanics of Learning Machines: from algorithmic and information-theoretical limits to new biologically inspired paradigms” 20229T9EAT – CUP J53D23003640001.

Bibliography

[1] Dan Cireşan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. "A committee of neural networks for traffic sign classification". In: *The 2011 International Joint Conference on Neural Networks*. 2011, pp. 1918–1921. DOI: 10.1109/IJCNN.2011.6033458.

[2] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[3] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://doi.org/10.1038/s41586-021-03819-2.

[4] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016 (June 2024), pp. 493–500. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w. URL: https://doi.org/10.1038/s41586-024-07487-w.

[5] Zhenyu Yang, Xiaoxi Zeng, Yi Zhao, and Runsheng Chen. "AlphaFold2 and its applications in the fields of biology and medicine". In: *Signal Transduction and Targeted Therapy* 8.1 (Mar. 2023), p. 115. ISSN: 2059-3635. DOI: 10.1038/s41392-023-01381-z. URL: https://doi.org/10.1038/s41392-023-01381-z.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-Resolution Image Synthesis With Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10684–10695. DOI: 10.48550/arXiv.2112.10752.

[7] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[9] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. 2019. DOI: https://doi.org/10.48550/arXiv.1803.03635. URL: https://openreview.net/forum?id=rJl-b3RcF7.

[10] Jialin Mao et al. “The training process of many deep networks explores the same low-dimensional manifold”. In: *Proceedings of the National Academy of Sciences* 121.12 (2024), e2310002121. DOI: 10.1073/pnas.2310002121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2310002121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2310002121.

[11] Battista Biggio et al. “Evasion Attacks against Machine Learning at Test Time”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402. ISBN: 978-3-642-40994-3. DOI: https://doi.org/10.1007/978-3-642-40994-3_25.

[12] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv e-prints*, arXiv:1312.6199 (Dec. 2013), arXiv:1312.6199. DOI: 10.48550/arXiv.1312.6199. arXiv: 1312.6199 [cs.CV].

[13] Svetlana Pavlitska, Nico Lambing, and J. Marius Zöllner. “Adversarial Attacks on Traffic Sign Recognition: A Survey”. In: *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. 2023, pp. 1–6. DOI: 10.1109/ICECCME57830.2023.10252727.

[14] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES”. In: *stat* 1050, arXiv:1412.6572 (2015), p. 20. DOI: 10.48550/arXiv.1412.6572. arXiv: 1412.6572 [stat.ML]. URL: https://doi.org/10.48550/arXiv.1412.6572.

[16] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. “Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model”. In: *Phys. Rev. X* 10 (4 Dec. 2020), p. 041044. DOI: 10.1103/PhysRevX.10.041044. URL: https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

[17] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. “Generalisation error in learning with random features and the hidden manifold model*”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (Dec. 2021), p. 124013. DOI: 10.1088/1742-5468/ac3ae6. URL: https://dx.doi.org/10.1088/1742-5468/ac3ae6.

[18] Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. *A High Dimensional Statistical Model for Adversarial Training: Geometry and Trade-Offs*. 2024. DOI: 10.48550/arXiv.2402.05674. arXiv: 2402.05674 [stat.ML]. URL: https://arxiv.org/abs/2402.05674.

[19] Matteo Vilucchio, Nikolaos Tsilivis, Bruno Loureiro, and Julia Kempe. *On the Geometry of Regularization in Adversarial Training: High-Dimensional Asymptotics and Generalization Bounds*. 2024. DOI: 10.48550/arXiv.2410.16073. arXiv: 2410.16073 [stat.ML]. URL: https://arxiv.org/abs/2410.16073.

[20] Matteo Vilucchio, Lenka Zdeborová, and Bruno Loureiro. *On the existence of consistent adversarial attacks in high-dimensional linear classification*. 2025. DOI: 10.48550/arXiv.2506.12454. arXiv: 2506.12454 [stat.ML]. URL: https://arxiv.org/abs/2506.12454.

[21] Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. *Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation*. 2025. arXiv: 2501.18530 [stat.ML]. URL: https://arxiv.org/abs/2501.18530.

[22] Patrick Charbonneau et al. *Spin Glass Theory and Far Beyond*. WORLD SCIENTIFIC, 2023. DOI: 10.1142/13341. eprint: https://www.worldscientific.com/doi/pdf/10.1142/13341. URL: https://www.worldscientific.com/doi/abs/10.1142/13341.

[23] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554. URL: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

[24] Geoffrey E. Hinton. "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Computation* 14.8 (2002), pp. 1771–1800. DOI: 10.1162/089976602760128018.

[25] H. H. Chen et al. "High order correlation model for associative memory". In: *AIP Conference Proceedings* 151.1 (Aug. 1986), pp. 86–99. ISSN: 0094-243X. DOI: 10.1063/1.36224. eprint: https://pubs.aip.org/aip/acp/article-pdf/151/1/86/12091820/86_1_online.pdf. URL: https://doi.org/10.1063/1.36224.

[26] Demetri Psaltis and Cheol Hoon Park. "Nonlinear discriminant functions and associative memories". In: *AIP Conference Proceedings* 151.1 (Aug. 1986), pp. 370–375. ISSN: 0094-243X. DOI: 10.1063/1.36241. eprint: https://pubs.aip.org/aip/acp/article-pdf/151/1/370/12091772/370_1_online.pdf. URL: https://doi.org/10.1063/1.36241.

[27] Pierre Baldi and Santosh S. Venkatesh. "Number of stable points for spin-glasses and neural networks of higher orders". In: *Phys. Rev. Lett.* 58 (9 Mar. 1987), pp. 913–916. DOI: 10.1103/PhysRevLett.58.913. URL: https://link.aps.org/doi/10.1103/PhysRevLett.58.913.

[28] E Gardner. "Multiconnected neural network models". In: *Journal of Physics A: Mathematical and General* 20.11 (Aug. 1987), p. 3453. DOI: 10.1088/0305-4470/20/11/046. URL: https://dx.doi.org/10.1088/0305-4470/20/11/046.

[29] Horn, D. and Usher, M. "Capacities of multiconnected memory models". In: *J. Phys. France* 49.3 (1988), pp. 389–395. DOI: 10.1051/jphys:01988004903038900. URL: https://doi.org/10.1051/jphys:01988004903038900.

[30] Dmitry Krotov and John J. Hopfield. "Dense Associative Memory for Pattern Recognition". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. NIPS'16. Barcelona, Spain: Curran Associates, Inc., 2016, pp. 1180–1188. ISBN: 9781510838819. DOI: 10.48550/arXiv.1606.01164. arXiv: 1606.01164 [cs.NE]. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

[31] A. Decelle, G. Fissore, and C. Furtlehner. "Spectral dynamics of learning in restricted Boltzmann machines". In: *Europhysics Letters* 119.6 (Nov. 2017), p. 60001. DOI: 10.1209/0295-5075/119/60001. URL: https://dx.doi.org/10.1209/0295-5075/119/60001.

[32] A. Decelle, G. Fissore, and C. Furtlehner. "Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics". In: *Journal of Statistical Physics* 172.6 (Sept. 2018), pp. 1576–1608. ISSN: 1572-9613. DOI: 10.1007/s10955-018-2105-y. URL: https://doi.org/10.1007/s10955-018-2105-y.

[33] Dimitrios Bachtis, Giulio Biroli, Aurélien Decelle, and Beatriz Seoane. "Cascade of phase transitions in the training of energy-based models". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 55591–55619. DOI: 10.48550/arXiv.2405.14689. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/648a5a590ca6f2bb5de53f938e230160-Paper-Conference.pdf.

[34] Nacer Eddine Boukacem et al. "Waddington landscape for prototype learning in generalized Hopfield networks". In: *Phys. Rev. Res.* 6 (3 July 2024), p. 033098. DOI: 10.1103/PhysRevResearch.6.033098. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.6.033098.

[35] Nicolas Béreux, Aurélien Decelle, Cyril Furtlehner, Lorenzo Rosset, and Beatriz Seoane. *Fast training and sampling of Restricted Boltzmann Machines*. 2024. arXiv: 2405.15376 [cs.LG]. URL: https://arxiv.org/abs/2405.15376.

[36] Dmitry Krotov and John Hopfield. "Dense Associative Memory Is Robust to Adversarial Inputs". In: *Neural Computation* 30.12 (Dec. 2018), pp. 3151–3167. ISSN: 0899-7667. DOI: 10.1162/neco_a_01143. arXiv: 1701.00939 [cs.LG]. URL: https://doi.org/10.1162/neco%5C_a%5C_01143.

[37] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG]. URL: https://arxiv.org/abs/2011.13456.

[38] Hubert Ramsauer et al. "Hopfield Networks is All You Need". In: *International Conference on Learning Representations*. 2021. DOI: 10.48550/arXiv.2008.02217. arXiv: 2008.02217 [cs.NE]. URL: https://openreview.net/forum?id=tL89RnzIiCd.

[39] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. "Phase transitions in restricted Boltzmann machines with generic priors". In: *Phys. Rev. E* 96 (4 Oct. 2017), p. 042156. DOI: 10.1103/PhysRevE.96.042156. arXiv: 1612.03132 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevE.96.042156.

[40] Aurelien Decelle, Sungmin Hwang, Jacopo Rocchi, and Daniele Tantari. "Inverse problems for structured datasets using parallel TAP equations and restricted Boltzmann machines". In: *Scientific Reports* 11, 19990 (Oct. 2021), p. 19990. DOI: 10.1038/s41598-021-99353-2. arXiv: 1906.11988 [cond-mat.dis-nn].

[41] Linda Albanese, Francesco Alemanno, Andrea Alessandrelli, and Adriano Barra. "Replica Symmetry Breaking in Dense Hebbian Neural Networks". In: *Journal of Statistical Physics* 189.2, 24 (Nov. 2022), p. 24. ISSN: 1572-9613. DOI: 10.1007/s10955-022-02966-8. arXiv: 2111.12997 [cond-mat.dis-nn]. URL: https://doi.org/10.1007/s10955-022-02966-8.

[42] Tianqi Hou, K Y Michael Wong, and Haiping Huang. "Minimal model of permutation symmetry in unsupervised learning". In: *Journal of Physics A: Mathematical and Theoretical* 52.41 (Sept. 2019), p. 414001. DOI: 10.1088/1751-8121/ab3f3f. arXiv: 1904.13052 [cond-mat.dis-nn]. URL: https://dx.doi.org/10.1088/1751-8121/ab3f3f.

[43] Francesco Alemanno, Luca Camanzi, Gianluca Manzan, and Daniele Tantari. "Hopfield model with planted patterns: A teacher-student self-supervised learning model". In: *Applied Mathematics and Computation* 458 (2023), p. 128253. ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2023.128253. arXiv: 2304.13710 [cond-mat.dis-nn]. URL: https://www.sciencedirect.com/science/article/pii/S0096300323004228.

[44] Gianluca Manzan and Daniele Tantari. "The effect of priors on Learning with Restricted Boltzmann Machines". In: *Physica A: Statistical Mechanics and its Applications* 674 (2025), p. 130766. ISSN: 0378-4371. DOI: 10.1016/j.physa.2025.130766. URL: https://www.sciencedirect.com/science/article/pii/S0378437125004182.

[45] Robin Thériault and Daniele Tantari. "Dense Hopfield networks in the teacher-student setting". In: *SciPost Phys.* 17 (2024), p. 040. DOI: 10.21468/SciPostPhys.17.2.040. URL: https://scipost.org/10.21468/SciPostPhys.17.2.040.

[46] Robin Thériault, Francesco Tosello, and Daniele Tantari. "Modeling structured data learning with Restricted Boltzmann machines in the teacher–student setting". In: *Neural Networks* 189 (2025), p. 107542. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2025.107542. URL: https://www.sciencedirect.com/science/article/pii/S0893608025004216.

[47] Robin Thériault and Daniele Tantari. *Saddle Hierarchy in Dense Associative Memory*. 2025. arXiv: 2508.19151 [cs.LG]. URL: https://arxiv.org/abs/2508.19151.

[48] Ruslan Salakhutdinov and Iain Murray. "On the quantitative analysis of deep belief networks". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 872–879. ISBN: 9781605582054. DOI: 10.1145/1390156.1390266. URL: https://doi.org/10.1145/1390156.1390266.

[49] I. Kanter and H. Sompolinsky. "Associative recall of memory without errors". In: *Phys. Rev. A* 35 (1 Jan. 1987), pp. 380–392. DOI: 10.1103/PhysRevA.35.380. URL: https://link.aps.org/doi/10.1103/PhysRevA.35.380.

[50] Amos Storkey. "Increasing the capacity of a hopfield network without sacrificing functionality". In: *Artificial Neural Networks — ICANN'97*. Ed. by Wulfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 451–456.

[51] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. “On the equivalence of Hopfield networks and Boltzmann Machines”. In: *Neural Networks* 34 (2012), pp. 1–9. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2012.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0893608012001608.

[52] Ludovica Serricchio et al. “Daydreaming Hopfield Networks and their surprising effectiveness on correlated data”. In: *Neural Networks* 186 (2025), p. 107216. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2025.107216. URL: https://www.sciencedirect.com/science/article/pii/S0893608025000954.

[53] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005. ISBN: 9781410612403. DOI: https://doi.org/10.4324/9781410612403.

[54] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. “Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks”. In: *Phys. Rev. Lett.* 55 (14 Sept. 1985), pp. 1530–1533. DOI: 10.1103/PhysRevLett.55.1530. URL: https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

[55] Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. “Statistical mechanics of neural networks near saturation”. In: *Annals of Physics* 173.1 (1987), pp. 30–67. ISSN: 0003-4916. DOI: https://doi.org/10.1016/0003-4916(87)90092-3. URL: https://www.sciencedirect.com/science/article/pii/0003491687900923.

[56] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. “Information storage in neural networks with low levels of activity”. In: *Phys. Rev. A* 35 (5 Mar. 1987), pp. 2293–2303. DOI: 10.1103/PhysRevA.35.2293. URL: https://link.aps.org/doi/10.1103/PhysRevA.35.2293.

[57] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519. URL: https://doi.org/10.1037/h0042519.

[58] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[59] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. 2017. arXiv: 1609.04747 [cs.LG]. URL: https://arxiv.org/abs/1609.04747.

[60] Lenka Zdeborová and Florent Krzakala. “Statistical physics of inference: thresholds and algorithms”. In: *Advances in Physics* 65.5 (2016), pp. 453–552. DOI: 10.1080/00018732.2016.1211393. arXiv: 1511.02476 [cond-mat.stat-mech]. URL: https://doi.org/10.1080/00018732.2016.1211393.

[61] E Gardner and B Derrida. “Three unfinished works on the optimal storage capacity of networks”. In: *Journal of Physics A: Mathematical and General* 22.12 (June 1989), p. 1983. DOI: 10.1088/0305-4470/22/12/004. URL: https://dx.doi.org/10.1088/0305-4470/22/12/004.

[62] Géza Györgyi. "First-order transition to perfect generalization in a neural network with binary synapses". In: *Phys. Rev. A* 41 (12 June 1990), pp. 7097–7100. DOI: 10.1103/PhysRevA.41.7097. URL: https://link.aps.org/doi/10.1103/PhysRevA.41.7097.

[63] Dmitry Krotov and John J. Hopfield. "Large Associative Memory Problem in Neurobiology and Machine Learning". In: *International Conference on Learning Representations*. 2021. DOI: 10.48550/arXiv.2008.06996. arXiv: 2008.06996 [q-bio.NC]. URL: https://openreview.net/forum?id=X4y_10OX-hX.

[64] Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. *Modern Methods in Associative Memory*. 2025. arXiv: 2507.06211 [cs.LG]. URL: https://arxiv.org/abs/2507.06211.

[65] Benjamin Hoover et al. "Memory in Plain Sight: A Survey of the Uncanny Resemblances between Diffusion Models and Associative Memories". In: *arXiv e-prints*, arXiv:2309.16750 (Sept. 2023), arXiv:2309.16750. DOI: 10.48550/arXiv.2309.16750. arXiv: 2309.16750 [cs.LG].

[66] Luca Ambrogioni. "In search of dispersed memories: Generative diffusion models are associative memory networks". In: *arXiv e-prints*, arXiv:2309.17290 (Sept. 2023), arXiv:2309.17290. DOI: 10.48550/arXiv.2309.17290. arXiv: 2309.17290 [stat.ML].

[67] Toshihiro Ota and Ryo Karakida. "Attention in a Family of Boltzmann Machines Emerging From Modern Hopfield Networks". In: *Neural Computation* 35.8 (July 2023), pp. 1463–1480. ISSN: 0899-7667. DOI: 10.1162/neco_a_01597. eprint: https://direct.mit.edu/neco/article-pdf/35/8/1463/2143211/neco_a_01597.pdf. URL: https://doi.org/10.1162/neco%5C_a%5C_01597.

[68] Ryo Karakida, Toshihiro Ota, and Masato Taki. *Hierarchical Associative Memory, Parallelized MLP-Mixer, and Symmetry Breaking*. 2024. arXiv: 2406.12220 [cs.LG]. URL: https://arxiv.org/abs/2406.12220.

[69] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: Association for Computing Machinery, 2007, pp. 791–798. ISBN: 9781595937933. DOI: 10.1145/1273496.1273596. URL: https://doi.org/10.1145/1273496.1273596.

[70] Renfrey Burnard Potts. "Some generalized order-disorder transformations". In: *Mathematical proceedings of the cambridge philosophical society*. Vol. 48. 1. Cambridge University Press. 1952, pp. 106–109. DOI: 10.1017/S0305004100027419.

[71] Fa-Yueh Wu. "The potts model". In: *Reviews of modern physics* 54.1 (1982), p. 235. DOI: 10.1103/RevModPhys.54.235.

[72] Aurélien Decelle and Cyril Furtlehner. "Restricted Boltzmann machine: Recent advances and mean-field theory*". In: *Chinese Physics B* 30.4 (Apr. 2021), p. 040202. DOI: 10.1088/1674-1056/abd160. URL: https://dx.doi.org/10.1088/1674-1056/abd160.

[73] H Nishimori. “Exact results and critical properties of the Ising model with competing interactions”. In: *Journal of Physics C: Solid State Physics* 13.21 (July 1980), p. 4071. DOI: 10.1088/0022-3719/13/21/012. URL: https://dx.doi.org/10.1088/0022-3719/13/21/012.

[74] Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, July 2001. ISBN: 9780198509417. DOI: 10.1093/acprof:oso/9780198509417.001.0001. eprint: https://academic.oup.com/book/5185/book-pdf/54038185/acprof-9780198509400.pdf. URL: https://doi.org/10.1093/acprof:oso/9780198509417.001.0001.

[75] Pierluigi Contucci, Cristian Giardinà, and Hidetoshi Nishimori. “Spin Glass Identities and the Nishimori Line”. In: *Spin Glasses: Statics and Dynamics*. Ed. by Anne Boutet de Monvel and Anton Bovier. Basel: Birkhäuser Basel, 2009, pp. 103–121. DOI: https://doi.org/10.1007/978-3-7643-9891-0_4. arXiv: 0805.0754 [cond-mat.dis-nn].

[76] Yukito Iba. “The Nishimori line and Bayesian statistics”. In: *Journal of Physics A Mathematical General* 32.21 (May 1999), pp. 3875–3888. DOI: 10.1088/0305-4470/32/21/302. arXiv: cond-mat/9809190 [cond-mat.dis-nn].

[77] Dimitris Achlioptas and Amin Coja-Oghlan. “Algorithmic Barriers from Phase Transitions”. In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. 2008, pp. 793–802. DOI: 10.1109/FOCS.2008.11. arXiv: 0803.2122 [math.CO].

[78] Florent Krzakala and Lenka Zdeborová. “Hiding Quiet Solutions in Random Constraint Satisfaction Problems”. In: *Phys. Rev. Lett.* 102 (23 June 2009), p. 238701. DOI: 10.1103/PhysRevLett.102.238701. arXiv: 0901.2130 [cond-mat.stat-mech]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.102.238701.

[79] Lenka Zdeborová and Florent Krzakala. “Quiet Planting in the Locked Constraint Satisfaction Problems”. In: *SIAM Journal on Discrete Mathematics* 25.2 (2011), pp. 750–770. DOI: 10.1137/090750755. arXiv: 0902.4185 [cond-mat.stat-mech]. URL: https://doi.org/10.1137/090750755.

[80] Carlo Lucibello and Marc Mézard. “Exponential Capacity of Dense Associative Memories”. In: *Phys. Rev. Lett.* 132 (7 Feb. 2024), p. 077301. DOI: 10.1103/PhysRevLett.132.077301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.132.077301.

[81] Bert Kappen. “Using Boltzmann Machines for probability estimation”. In: *ICANN ’93*. Ed. by Stan Gielen and Bert Kappen. London: Springer London, 1993, pp. 521–526. ISBN: 978-1-4471-2063-6.

[82] Hilbert J. Kappen. “Deterministic learning rules for boltzmann machines”. In: *Neural Networks* 8.4 (1995), pp. 537–548. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(94)00112-Y. URL: https://www.sciencedirect.com/science/article/pii/089360809400112Y.

[83] Marcel J. Nijman and Hilbert J. Kappen. “Symmetry Breaking and Training from Incomplete Data with Radial Basis Boltzmann Machines”. In: *International Journal of Neural Systems* 08.03 (1997), pp. 301–315. DOI: 10.1142/S0129065797000318. eprint: https://doi.org/10.1142/S0129065797000318. URL: https://doi.org/10.1142/S0129065797000318.

[84] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. "Statistical mechanics and phase transitions in clustering". In: *Phys. Rev. Lett.* 65 (8 Aug. 1990), pp. 945–948. DOI: 10.1103/PhysRevLett.65.945. URL: https://link.aps.org/doi/10.1103/PhysRevLett.65.945.

[85] Martin Kloppenburg and Paul Tavan. "Deterministic annealing for density estimation by multivariate normal mixtures". In: *Phys. Rev. E* 55 (3 Mar. 1997), R2089–R2092. DOI: 10.1103/PhysRevE.55.R2089. URL: https://link.aps.org/doi/10.1103/PhysRevE.55.R2089.

[86] Shotaro Akaho and Hilbert J. Kappen. "Nonmonotonic Generalization Bias of Gaussian Mixture Models". In: *Neural Computation* 12.6 (June 2000), pp. 1411–1427. ISSN: 0899-7667. DOI: 10.1162/089976600300015439. eprint: https://direct.mit.edu/neco/article-pdf/12/6/1411/814516/089976600300015439.pdf. URL: https://doi.org/10.1162/089976600300015439.

[87] Lemeng Wu, Dilin Wang, and Qiang Liu. "Splitting Steepest Descent for Growing Neural Architectures". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. DOI: https://doi.org/10.48550/arXiv.1910.02366. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/3a01fc0853ebeba94fde4d1cc6fb842a-Paper.pdf.

[88] Dilin Wang, Meng Li, Lemeng Wu, Vikas Chandra, and Qiang Liu. "Energy-Aware Neural Architecture Optimization with Fast Splitting Steepest Descent". In: *arXiv e-prints*, arXiv:1910.03103 (Oct. 2019), arXiv:1910.03103. DOI: 10.48550/arXiv.1910.03103. arXiv: 1910.03103 [cs.LG].

[89] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. "Spin-glass models of neural networks". In: *Phys. Rev. A* 32 (2 Aug. 1985), pp. 1007–1018. DOI: 10.1103/PhysRevA.32.1007. URL: https://link.aps.org/doi/10.1103/PhysRevA.32.1007.

[90] Thomas M. Cover. "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition". In: *IEEE Transactions on Electronic Computers* EC-14.3 (1965), pp. 326–334. DOI: 10.1109/PGEC.1965.264137.

[91] Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, and Francesco Moauro. "Multitasking Associative Networks". In: *Phys. Rev. Lett.* 109 (26 Dec. 2012), p. 268101. DOI: 10.1103/PhysRevLett.109.268101. arXiv: 1111.5191 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.109.268101.

[92] E Agliari, A Annibale, A Barra, A C C Coolen, and D Tantari. "Immune networks: multi-tasking capabilities at medium load". In: *Journal of Physics A: Mathematical and Theoretical* 46.33 (July 2013), p. 335101. DOI: 10.1088/1751-8113/46/33/335101. arXiv: 1302.7259 [cond-mat.dis-nn]. URL: https://dx.doi.org/10.1088/1751-8113/46/33/335101.

[93] E. Agliari, A. Annibale, A. Barra, A. C. C. Coolen, and D. Tantari. "Immune networks: multitasking capabilities near saturation". In: *Journal of Physics A Mathematical General* 46.41, 415003 (Oct. 2013), p. 415003. DOI: 10.1088/1751-8113/46/41/415003. arXiv: 1305.5936 [cond-mat.dis-nn].

[94] Peter Sollich, Daniele Tantari, Alessia Annibale, and Adriano Barra. "Extensive Parallel Processing on Scale-Free Networks". In: *Phys. Rev. Lett.* 113 (23 Dec. 2014), p. 238106. DOI: 10.1103/PhysRevLett.113.238106. arXiv: 1404.3654 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.113.238106.

[95] E. Agliari, A. Annibale, A. Barra, A. C. C. Coolen, and D. Tantari. "Retrieving infinite numbers of patterns in a spin-glass model of immune networks". In: *Europhysics Letters* 117.2 (Mar. 2017), p. 28003. DOI: 10.1209/0295-5075/117/28003. arXiv: 1305.2076 [cond-mat.dis-nn]. URL: https://dx.doi.org/10.1209/0295-5075/117/28003.

[96] Elena Agliari et al. "Retrieval Capabilities of Hierarchical Networks: From Dyson to Hopfield". In: *Phys. Rev. Lett.* 114 (2 Jan. 2015), p. 028103. DOI: 10.1103/PhysRevLett.114.028103. arXiv: 1407.5019 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.114.028103.

[97] Elena Agliari, Danila Migliozzi, and Daniele Tantari. "Non-convex Multi-species Hopfield Models". In: *Journal of Statistical Physics* 172.5 (Sept. 2018), pp. 1247–1269. DOI: 10.1007/s10955-018-2098-6. arXiv: 1807.03609 [cond-mat.dis-nn].

[98] Elena Agliari et al. "Hierarchical neural networks perform both serial and parallel processing". In: *Neural Networks* 66 (2015), pp. 22–35. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2015.02.010. arXiv: 1409.0227 [cond-mat.dis-nn]. URL: https://www.sciencedirect.com/science/article/pii/S0893608015000441.

[99] Elena Agliari et al. "Metastable states in the hierarchical Dyson model drive parallel processing in the hierarchical Hopfield network". In: *Journal of Physics A Mathematical General* 48.1, 015001 (Jan. 2015), p. 015001. DOI: 10.1088/1751-8113/48/1/015001. arXiv: 1407.5176 [cond-mat.dis-nn].

[100] Elena Agliari et al. "Topological properties of hierarchical networks". In: *Phys. Rev. E* 91 (6 June 2015), p. 062807. DOI: 10.1103/PhysRevE.91.062807. arXiv: 1412.5918 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevE.91.062807.

[101] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. "Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors". In: *Phys. Rev. E* 97 (2 Feb. 2018), p. 022310. DOI: 10.1103/PhysRevE.97.022310. arXiv: 1702.05882 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevE.97.022310.

[102] Adriano Barra, Pierluigi Contucci, Emanuele Mingione, and Daniele Tantari. "Multi-Species Mean Field Spin Glasses. Rigorous Results". In: *Annales Henri Poincaré* 16.3 (Mar. 2015), pp. 691–708. ISSN: 1424-0661. DOI: 10.1007/s00023-014-0341-5. arXiv: 1307.5154 [math-ph]. URL: https://doi.org/10.1007/s00023-014-0341-5.

[103] Elena Agliari, Adriano Barra, Chiara Longo, and Daniele Tantari. "Neural Networks Retrieving Boolean Patterns in a Sea of Gaussian Ones". In: *Journal of Statistical Physics* 168.5 (Sept. 2017), pp. 1085–1104. DOI: 10.1007/s10955-017-1840-9. arXiv: 1703.05210 [math-ph].

[104] Adriano Barra, Giuseppe Genovese, Francesco Guerra, and Daniele Tantari. "How glassy are neural networks?" In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.7 (July 2012), p. 07009. DOI: 10.1088/1742-5468/2012/07/P07009. arXiv: 1205.3900 [cond-mat.dis-nn]. URL: https://dx.doi.org/10.1088/1742-5468/2012/07/P07009.

[105] Giuseppe Genovese and Daniele Tantari. "Legendre equivalences of spherical Boltzmann machines". In: *Journal of Physics A Mathematical General* 53.9, 094001 (Mar. 2020), p. 094001. DOI: 10.1088/1751-8121/ab6b92. arXiv: 1910.14559 [cond-mat.dis-nn]. URL: https://dx.doi.org/10.1088/1751-8121/ab6b92.

[106] Jacopo Rocchi, David Saad, and Daniele Tantari. "High storage capacity in the Hopfield model with auto-interactions—stability analysis". In: *Journal of Physics A Mathematical General* 50.46, 465001 (Nov. 2017), p. 465001. DOI: 10.1088/1751-8121/aa8fd7. arXiv: 1704.07741 [cond-mat.dis-nn].

[107] Michael Widrich et al. "Modern Hopfield Networks and Attention for Immune Repertoire Classification". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 18832–18845. DOI: 10.48550/arXiv.2007.13505. arXiv: 2007.13505 [cs.LG]. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/da4902cb0bc38210839714ebdcf0efc3-Paper.pdf.

[108] L. F. Abbott and Yair Arian. "Storage capacity of generalized networks". In: *Phys. Rev. A* 36 (10 Nov. 1987), pp. 5091–5094. DOI: 10.1103/PhysRevA.36.5091. URL: https://link.aps.org/doi/10.1103/PhysRevA.36.5091.

[109] Benjamin Hoover et al. "Energy Transformer". In: *arXiv e-prints*, arXiv:2302.07253 (Feb. 2023), arXiv:2302.07253. DOI: 10.48550/arXiv.2302.07253. arXiv: 2302.07253 [cs.LG].

[110] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations*. 2018. DOI: 10.48550/arXiv.1706.06083. arXiv: 1706.06083 [stat.ML]. URL: https://openreview.net/forum?id=rJzIBfZAb.

[111] Awais Muhammad and Sung-Ho Bae. "A Survey on Efficient Methods for Adversarial Robustness". In: *IEEE Access* 10 (Jan. 2022), pp. 118815–118830. DOI: 10.1109/ACCESS.2022.3216291.

[112] Patrick Charbonneau. "From the replica trick to the replica symmetry breaking technique". In: *arXiv e-prints*, arXiv:2211.01802 (Nov. 2022), arXiv:2211.01802. DOI: 10.48550/arXiv.2211.01802. arXiv: 2211.01802 [physics.hist-ph].

[113] D.J. Gross and M. Mezard. "The simplest spin glass". In: *Nuclear Physics B* 240.4 (1984), pp. 431–452. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(84)90237-2. URL: https://www.sciencedirect.com/science/article/pii/0550321384902372.

[114] E. Gardner. "Spin glasses with p-spin interactions". In: *Nuclear Physics B* 257 (1985), pp. 747–765. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(85)90374-8. URL: https://www.sciencedirect.com/science/article/pii/0550321385903748.

[115] Bernard Derrida. “Random-energy model: An exactly solvable model of disordered systems”. In: *Phys. Rev. B* 24 (5 Sept. 1981), pp. 2613–2626. DOI: 10.1103/PhysRevB.24.2613. URL: https://link.aps.org/doi/10.1103/PhysRevB.24.2613.

[116] Rémi Monasson. “Structural Glass Transition and the Entropy of the Metastable States”. In: *Phys. Rev. Lett.* 75 (15 Oct. 1995), pp. 2847–2850. DOI: 10.1103/PhysRevLett.75.2847. URL: https://link.aps.org/doi/10.1103/PhysRevLett.75.2847.

[117] A. Montanari and F. Ricci-Tersenghi. “On the nature of the low-temperature phase in discontinuous mean-field spin glasses”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 33.3 (June 2003), pp. 339–346. ISSN: 1434-6036. DOI: 10.1140/epjb/e2003-00174-7. URL: https://doi.org/10.1140/epjb/e2003-00174-7.

[118] A. Crisanti, L. Leuzzi, and T. Rizzo. “Complexity in mean-field spin-glass models: Ising p-spin”. In: *Phys. Rev. B* 71 (9 Mar. 2005), p. 094202. DOI: 10.1103/PhysRevB.71.094202. URL: https://link.aps.org/doi/10.1103/PhysRevB.71.094202.

[119] Silvio Franz, Giorgio Parisi, Maksim Sevelev, Pierfrancesco Urbani, and Francesco Zamponi. “Universality of the SAT-UNSAT (jamming) threshold in non-convex continuous constraint satisfaction problems”. In: *SciPost Phys.* 2 (2017), p. 019. DOI: 10.21468/SciPostPhys.2.3.019. URL: https://scipost.org/10.21468/SciPostPhys.2.3.019.

[120] George G. Roussas. *Contiguity of Probability Measures: Some Applications in Statistics*. Cambridge Tracts in Mathematics. Cambridge University Press, 1972. DOI: 10.1017/CBO9780511804373. URL: https://cir.nii.ac.jp/crid/1361137043923091072.

[121] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. “Glassy Nature of the Hard Phase in Inference Problems”. In: *Phys. Rev. X* 9 (1 Jan. 2019), p. 011020. DOI: 10.1103/PhysRevX.9.011020. arXiv: 1805.05857 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevX.9.011020.

[122] Lenka Zdeborová and Florent Krzakala. “Phase transitions in the coloring of random graphs”. In: *Phys. Rev. E* 76 (3 Sept. 2007), p. 031131. DOI: 10.1103/PhysRevE.76.031131. arXiv: 0704.1269 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevE.76.031131.

[123] Elena Agliari, Francesco Alemanno, Adriano Barra, Martino Centonze, and Alberto Fachechi. “Neural Networks with a Redundant Representation: Detecting the Undetectable”. In: *Phys. Rev. Lett.* 124 (2 Jan. 2020), p. 028301. DOI: 10.1103/PhysRevLett.124.028301. arXiv: 1911.12689 [cond-mat.dis-nn]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.124.028301.

[124] Elena Agliari and Giordano De Marzo. “Tolerance versus synaptic noise in dense associative memories”. In: *European Physical Journal Plus* 135.11, 883 (Nov. 2020), p. 883. DOI: 10.1140/epjp/s13360-020-00894-8. arXiv: 2007.02849 [cond-mat.dis-nn].

[125] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. "Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples". In: *arXiv e-prints*, arXiv:2010.03593 (Oct. 2020), arXiv:2010.03593. DOI: 10.48550/arXiv.2010.03593. arXiv: 2010.03593 [stat.ML].

[126] Hanxun Huang et al. "Exploring Architectural Ingredients of Adversarially Robust Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 5545–5559. DOI: 10.48550/arXiv.2110.03825. arXiv: 2110.03825 [cs.LG]. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/2bd7f907b7f5b6bbd91822c0c7b835f6-Paper.pdf.

[127] Sebastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. "A Law of Robustness for Two-Layers Neural Networks". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 804–820. DOI: 10.48550/arXiv.2009.14444. arXiv: 2009.14444 [cs.LG]. URL: https://proceedings.mlr.press/v134/bubeck21a.html.

[128] Sebastien Bubeck and Mark Sellke. "A Universal Law of Robustness via Isoperimetry". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 28811–28822. DOI: 10.48550/arXiv.2105.12806. arXiv: 2105.12806 [cs.LG]. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/f197002b9a0853eca5e046d9ca4663d5-Paper.pdf.

[129] Joan Puigcerver, Rodolphe Jenatton, Carlos Riquelme, Pranjal Awasthi, and Srinadh Bhojanapalli. "On the Adversarial Robustness of Mixture of Experts". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 9660–9671. DOI: 10.48550/arXiv.2210.10253. arXiv: 2210.10253 [cs.LG]. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/3effb91593c4fb42b1da1528328eff49-Paper-Conference.pdf.

[130] Antônio H. Ribeiro and Thomas B. Schön. "Overparameterized Linear Regression Under Adversarial Attacks". In: *IEEE Transactions on Signal Processing* 71 (Jan. 2023), pp. 601–614. DOI: 10.1109/TSP.2023.3246228. arXiv: 2204.06274 [stat.ML].

[131] Saumya Jetley, Nicholas Lord, and Philip Torr. "With Friends Like These, Who Needs Adversaries?" In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. DOI: 10.48550/arXiv.1807.04200. arXiv: 1807.04200 [cs.CV]. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/803a82dee7e3fbb3438a149508484250-Paper.pdf.

[132] Andrew Ilyas et al. "Adversarial Examples Are Not Bugs, They Are Features". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. DOI: https://doi.org/10.48550/arXiv.1905.02175. arXiv: 1905.02175 [stat.ML].

URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

[133] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy". In: *International Conference on Learning Representations*. 2019. DOI: 10.48550/arXiv.1805.12152. arXiv: 1805.12152 [stat.ML]. URL: https://openreview.net/forum?id=SyxAb30cY7.

[134] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. "On a Model of Associative Memory with Huge Storage Capacity". In: *Journal of Statistical Physics* 168.2 (July 2017), pp. 288–299. DOI: 10.1007/s10955-017-1806-y. arXiv: 1702.01929 [math.PR].

[135] Carlo Lucibello and Marc Mézard. "The Exponential Capacity of Dense Associative Memories". In: *arXiv e-prints*, arXiv:2304.14964 (Apr. 2023), arXiv:2304.14964. DOI: 10.48550/arXiv.2304.14964. arXiv: 2304.14964 [cond-mat.dis-nn].

[136] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. "A learning algorithm for boltzmann machines". In: *Cognitive Science* 9.1 (1985), pp. 147–169. ISSN: 0364-0213. DOI: https://doi.org/10.1016/S0364-0213(85)80012-4. URL: https://www.sciencedirect.com/science/article/pii/S0364021385800124.

[137] P. Smolensky. "Information Processing in Dynamical Systems: Foundations of Harmony Theory". In: *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, July 1986, pp. 194–281. ISBN: 9780262291408. DOI: 10.7551/mitpress/5236.003.0009. eprint: https://direct.mit.edu/book/chapter-pdf/2163042/9780262291408_caf.pdf. URL: https://doi.org/10.7551/mitpress/5236.003.0009.

[138] Yoav Freund and David Haussler. "Unsupervised learning of distributions on binary vectors using two layer networks". In: *Advances in Neural Information Processing Systems*. Ed. by J. Moody, S. Hanson, and R.P. Lippmann. Vol. 4. NIPS'91. Denver, Colorado: Morgan-Kaufmann, 1991, pp. 912–919. ISBN: 1558602224. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf.

[139] Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. "Learning a Generative Model of Images by Factoring Appearance and Shape". In: *Neural Computation* 23.3 (2011), pp. 593–650. DOI: 10.1162/NECO_a_00086.

[140] J. Tubiana and R. Monasson. "Emergence of Compositional Representations in Restricted Boltzmann Machines". In: *Phys. Rev. Lett.* 118 (13 Mar. 2017), p. 138301. DOI: 10.1103/PhysRevLett.118.138301. arXiv: 1611.06759 [physics.data-an]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.138301.

[141] Jyri Kivinen and Christopher Williams. "Multiple Texture Boltzmann Machines". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 2012, pp. 638–646. URL: https://proceedings.mlr.press/v22/kivinen12.html.

[142] Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. “Modeling documents with a Deep Boltzmann Machine”. In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. UAI’13. Bellevue, WA: AUAI Press, 2013, pp. 616–624. DOI: arXiv:1309.6865.

[143] Nicolas Le Roux and Yoshua Bengio. “Representational Power of Restricted Boltzmann Machines and Deep Belief Networks”. In: *Neural Computation* 20.6 (2008), pp. 1631–1649. DOI: 10.1162/neco.2008.04-07-510.

[144] Guido Montufar and Nihat Ay. “Refinements of Universal Approximation Results for Deep Belief Networks and Restricted Boltzmann Machines”. In: *Neural Computation* 23.5 (2011), pp. 1306–1319. DOI: 10.1162/NECO_a_00113.

[145] Haiping Huang and Taro Toyoizumi. “Unsupervised feature learning from finite data by message passing: Discontinuous versus continuous phase transition”. In: *Phys. Rev. E* 94 (6 Dec. 2016), p. 062310. DOI: 10.1103/PhysRevE.94.062310. URL: https://link.aps.org/doi/10.1103/PhysRevE.94.062310.

[146] Haiping Huang. “Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2017.5 (May 2017), p. 053302. DOI: 10.1088/1742-5468/aa6ddc. URL: https://dx.doi.org/10.1088/1742-5468/aa6ddc.

[147] Haiping Huang. “Role of zero synapses in unsupervised feature learning”. In: *Journal of Physics A: Mathematical and Theoretical* 51.8 (Jan. 2018), 08LT01. DOI: 10.1088/1751-8121/aaa631. URL: https://dx.doi.org/10.1088/1751-8121/aaa631.

[148] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. “Deconstructing lottery tickets: zeros, signs, and the supermask”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019. DOI: https://doi.org/10.48550/arXiv.1905.01067.

[149] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. “What’s Hidden in a Randomly Weighted Neural Network?” In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11890–11899. DOI: 10.1109/CVPR42600.2020.01191.

[150] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. “Proving the Lottery Ticket Hypothesis: Pruning is All You Need”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6682–6691. DOI: https://doi.org/10.48550/arXiv.2002.00585. URL: https://proceedings.mlr.press/v119/malach20a.html.

[151] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. “Linear Mode Connectivity and the Lottery Ticket Hypothesis”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 3259–3269. DOI: https://doi.org/10.48550/arXiv.1912.05671. URL: https://proceedings.mlr.press/v119/frankle20a.html.

[152] Dmitry Panchenko. "The free energy in a multi-species Sherrington–Kirkpatrick model". In: *The Annals of Probability* 43.6 (2015), pp. 3494–3513. DOI: 10.1214/14-AOP967. URL: https://doi.org/10.1214/14-AOP967.

[153] Giuseppe Genovese and Daniele Tantari. "Overlap synchronisation in multipartite random energy models". In: *Journal of Statistical Physics* 169 (2017), pp. 1162–1170. DOI: 10.1007/s10955-017-1897-5.

[154] Giuseppe Genovese and Daniele Tantari. "Non-convex multipartite ferromagnets". In: *Journal of Statistical Physics* 163 (2016), pp. 492–513. DOI: 10.1007/s10955-016-1482-3.

[155] E Gardner. "The space of interactions in neural network models". In: *Journal of Physics A: Mathematical and General* 21.1 (Jan. 1988), p. 257. DOI: 10.1088/0305-4470/21/1/030. URL: https://dx.doi.org/10.1088/0305-4470/21/1/030.

[156] Edgar Dobriban and Stefan Wager. "HIGH-DIMENSIONAL ASYMPTOTICS OF PREDICTION: RIDGE REGRESSION AND CLASSIFICATION". In: *The Annals of Statistics* 46.1 (2018), pp. 247–279. ISSN: 00905364, 21688966. DOI: 10.48550/arXiv.1507.03003. URL: https://www.jstor.org/stable/26542784 (visited on 02/20/2024).

[157] Denny Wu and Ji Xu. "On the Optimal Weighted \ell_2 Regularization in Overparameterized Linear Regression". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 10112–10123. DOI: https://doi.org/10.48550/arXiv.2006.05800. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.

[158] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "When do neural networks outperform kernel methods?*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (Dec. 2021), p. 124009. DOI: 10.1088/1742-5468/ac3a81. URL: https://dx.doi.org/10.1088/1742-5468/ac3a81.

[159] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. "A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (Dec. 2021), p. 124006. DOI: 10.1088/1742-5468/ac3a77. URL: https://dx.doi.org/10.1088/1742-5468/ac3a77.

[160] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. "Multiple Descent: Design Your Own Generalization Curve". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8898–8912. DOI: https://doi.org/10.48550/arXiv.2008.01036. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/4ae67a7dd7e491f8fb6f9ea0cf25dfdb-Paper.pdf.

[161] Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. "Optimal Regularization can Mitigate Double Descent". In: *International Conference on Learning Representations*. 2021. DOI: https://doi.org/10.48550/arXiv.2003.01897. URL: https://openreview.net/forum?id=7R7fAoUygoa.

[162] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. "Asymptotics of Ridge(less) Regression under General Source Condition". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, Apr. 2021, pp. 3889–3897. DOI: https://doi.org/10.48550/arXiv.2006.06386. URL: https://proceedings.mlr.press/v130/richards21b.html.

[163] Stéphane d'Ascoli, Marylou Gabrié, Levent Sagun, and Giulio Biroli. "On the interplay between data structure and loss function in classification problems". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8506–8517. DOI: https://doi.org/10.48550/arXiv.2103.05524. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/47a5feca4ce02883a5643e295c7ce6cd-Paper.pdf.

[164] Bruno Loureiro et al. "Learning Gaussian Mixtures with Generalised Linear Models: Precise Asymptotics in High-dimensions". English. In: *Advances in Neural Information Processing Systems 34 - 35th Conference on Neural Information Processing Systems, NeurIPS 2021*. Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jenn Wortman Vaughan. Advances in Neural Information Processing Systems. Neural information processing systems foundation, 2021, pp. 10144–10157. DOI: arXiv:2106.03791.

[165] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborova. "Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 8936–8947. DOI: https://doi.org/10.48550/arXiv.2102.11742. URL: https://proceedings.mlr.press/v139/refinetti21b.html.

[166] Yuma Ichikawa and Koji Hukushima. "Statistical-mechanical Study of Deep Boltzmann Machine Given Weight Parameters after Training by Singular Value Decomposition". In: *Journal of the Physical Society of Japan* 91.11 (2022), p. 114001. DOI: 10.7566/JPSJ.91.114001. eprint: https://doi.org/10.7566/JPSJ.91.114001. URL: https://doi.org/10.7566/JPSJ.91.114001.

[167] Robin Thériault. *Modelling Structured Data Learning with Restricted Boltzmann Machines in the Teacher-Student Setting*. Version v1.0. 2025. DOI: 10.5281/zenodo.14892573. URL: https://github.com/RobinTher/Restricted_Boltzmann_Machine.

[168] Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhiqin J Xu. "Embedding Principle of Loss Landscape of Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc.,

2021, pp. 14848–14859. DOI: https://doi.org/10.48550/arXiv.2105.14573. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/7cc532d783a7461f227a5da8ea80bfe1-Paper.pdf.

[169] Max Welling and Yee Whye Teh. "Bayesian learning via stochastic gradient langevin dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688. ISBN: 9781450306195. DOI: https://dl.acm.org/doi/10.5555/3104482.3104568.

[170] Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A. Erdogdu. "Improved Discretization Analysis for Underdamped Langevin Monte Carlo". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 36–71. DOI: https://doi.org/10.48550/arXiv.2302.08049. URL: https://proceedings.mlr.press/v195/zhang23a.html.

[171] Geoffrey E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 599–619. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_32. URL: https://doi.org/10.1007/978-3-642-35289-8_32.

[172] Lénaïc Chizat and Francis Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 3040–3050. DOI: 10.48550/arXiv.1805.09545.

[173] Chao Ma, Lei Wu, and Weinan E. "The Quenching-Activation Behavior of the Gradient Descent Dynamics for Two-layer Neural Network Models". In: *arXiv e-prints*, arXiv:2006.14450 (June 2020), arXiv:2006.14450. DOI: 10.48550/arXiv.2006.14450. arXiv: 2006.14450 [cs.LG].

[174] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. "Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit". In: *Journal of Machine Learning Research* 22.71 (2021), pp. 1–47. DOI: arXiv:2007.07497v2. URL: http://jmlr.org/papers/v22/20-1123.html.

[175] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. "Efficient BackProp". In: *Neural Networks: Tricks of the Trade: Second Edition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. Chap. 2, pp. 9–48. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_3. URL: https://doi.org/10.1007/978-3-642-35289-8_3.

[176] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 448–456. DOI: arXiv:1502.03167. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[177] L. Huangi, L. Huangi, D. Yang, B. Lang, and J. Deng. "Decorrelated Batch Normalization". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2018, pp. 791–800. DOI: 10.1109/CVPR.2018.00089. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00089.

[178] S. Zhang et al. "Stochastic Whitening Batch Normalization". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2021, pp. 10973–10982. DOI: 10.1109/CVPR46437.2021.01083. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01083.

[179] Simon Wiesler and Hermann Ney. "A convergence analysis of log-linear training". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Granada, Spain: Curran Associates Inc., 2011, pp. 657–665. ISBN: 9781618395993.

[180] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W. Taylor. "Batch Normalization is a Cause of Adversarial Vulnerability". In: *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*. 2019. DOI: 10.48550/arXiv.1905.02161. URL: https://openreview.net/forum?id=BkxOwVShhE.

[181] Omobayode Fagbohungbe and Lijun Qian. "The Effect of Batch Normalization on Noise Resistant Property of Deep Learning Models". In: *IEEE Access* 10 (2022), pp. 127728–127741. DOI: 10.1109/ACCESS.2022.3206958.

[182] Asja Fischer and Christian Igel. "Training restricted Boltzmann machines: An introduction". In: *Pattern Recognition* 47.1 (2014), pp. 25–39. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2013.05.025. URL: https://www.sciencedirect.com/science/article/pii/S0031320313002495.

[183] J.H. Van Vleck and D. Middleton. "The spectrum of clipped noise". In: *Proceedings of the IEEE* 54.1 (1966), pp. 2–19. DOI: 10.1109/PROC.1966.4567.

[184] John Wishart. "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population". In: *Biometrika* 20A.1/2 (1928), pp. 32–52. ISSN: 00063444. DOI: https://doi.org/10.2307/2331939. URL: http://www.jstor.org/stable/2331939 (visited on 02/21/2024).

[185] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987, pp. 252–254. ISBN: 9780195042764. DOI: http://dx.doi.org/10.1063/1.2811680.

[186] Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. CRC Press, May 2018. DOI: 10.1201/9780429492563. URL: https://cir.nii.ac.jp/crid/1363670318557081344.

[187] Robert B. Griffiths. "Correlations in Ising Ferromagnets. I". In: *Journal of Mathematical Physics* 8.3 (Mar. 1967), pp. 478–483. ISSN: 0022-2488. DOI: 10.1063/1.1705219. eprint: https://pubs.aip.org/aip/jmp/article-pdf/8/3/478/19327843/478_1_online.pdf. URL: https://doi.org/10.1063/1.1705219.

[188] Alex Lewin Agnan Kessy and Korbinian Strimmer. “Optimal Whitening and Decorrelation”. In: *The American Statistician* 72.4 (2018), pp. 309–314. DOI: 10.1080/00031305.2016.1277159. eprint: https://doi.org/10.1080/00031305.2016.1277159. URL: https://doi.org/10.1080/00031305.2016.1277159.

[189] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. “Random Matrices and Complexity of Spin Glasses”. In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 165–201. DOI: https://doi.org/10.1002/cpa.21422. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.21422. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21422.

[190] Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. “The Loss Surfaces of Multilayer Networks”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, May 2015, pp. 192–204. DOI: https://doi.org/10.48550/arXiv.1412.0233. URL: https://proceedings.mlr.press/v38/choromanska15.html.

[191] Razvan Pascanu, Yann N. Dauphin, Surya Ganguli, and Yoshua Bengio. “On the saddle point problem for non-convex optimization”. In: *arXiv e-prints*, arXiv:1405.4604 (May 2014), arXiv:1405.4604. DOI: 10.48550/arXiv.1405.4604. arXiv: 1405.4604 [cs.LG].

[192] Yann N Dauphin et al. “Identifying and attacking the saddle point problem in high-dimensional non-convex optimization”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. DOI: https://doi.org/10.48550/arXiv.1406.2572. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/17e23e50bedc63b4095e3d8204ce063b-Paper.pdf.

[193] K. Fukumizu and S. Amari. “Local minima and plateaus in hierarchical structures of multilayer perceptrons”. In: *Neural Networks* 13.3 (2000), pp. 317–327. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(00)00009-5. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000095.

[194] Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-ichi Mototake, and Mirai Tanaka. “Semi-flat minima and saddle points by embedding neural networks to overparameterization”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. DOI: arXiv:1906.04868v2. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/a4ee59dd868ba016ed2de90d330acb6a-Paper.pdf.

[195] Berfin Simsek et al. “Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 9722–9732. DOI: arXiv:2105.12221v2. URL: https://proceedings.mlr.press/v139/simsek21a.html.

[196] Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. "Universal Hopfield Networks: A General Framework for Single-Shot Associative Memory Models". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 15561–15583. DOI: https://doi.org/10.48550/arXiv.2202.04557. URL: https://proceedings.mlr.press/v162/millidge22a.html.

[197] Andreas Fürst et al. "CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 20450–20468. DOI: https://doi.org/10.48550/arXiv.2110.11316. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8078e76f913e31b8467e85b4c0f0d22b-Paper-Conference.pdf.

[198] Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. "Conformal Prediction for Time Series with Modern Hopfield Networks". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 56027–56074. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/aef75887979ae1287b5deb54a1e3cbda-Paper-Conference.pdf.

[199] Robin Thériault. 2025. URL: https://github.com/RobinTher/Dense_Associative_Network_vMF.

[200] Ruslan Salakhutdinov and Geoffrey Hinton. "Deep Boltzmann Machines". In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Ed. by David van Dyk and Max Welling. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, Apr. 2009, pp. 448–455. URL: https://proceedings.mlr.press/v5/salakhutdinov09a.html.

[201] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986. ISBN: 0070544840.

[202] Inderjit S. Dhillon and Dharmendra S. Modha. "Concept Decompositions for Large Sparse Text Data Using Clustering". In: *Machine Learning* 42.1 (Jan. 2001), pp. 143–175. ISSN: 1573-0565. DOI: 10.1023/A:1007612920971. URL: https://doi.org/10.1023/A:1007612920971.

[203] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions". In: *Journal of Machine Learning Research* 6.46 (2005), pp. 1345–1382. URL: http://jmlr.org/papers/v6/banerjee05a.html.

[204] Matteo Carandini and David J. Heeger. "Normalization as a canonical neural computation". In: *Nature Reviews Neuroscience* 13.1 (Jan. 2012), pp. 51–62. ISSN: 1471-0048. DOI: 10.1038/nrn3136. URL: https://doi.org/10.1038/nrn3136.

[205] Michael C Mozer and Paul Smolensky. "Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 1. Morgan-Kaufmann, 1988. URL: https://proceedings.neurips.cc/paper_files/paper/1988/file/07e1cd7dca89a1678042477183b7ac3f-Paper.pdf.

[206] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. "Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks". In: *J. Mach. Learn. Res.* 22.1 (Jan. 2021). ISSN: 1532-4435. DOI: https://doi.org/10.48550/arXiv.2102.00554.

[207] Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. "Sparse DNNs with Improved Adversarial Robustness". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. DOI: https://doi.org/10.48550/arXiv.1810.09619. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/4c5bde74a8f110656874902f07378009-Paper.pdf.

[208] Soorya Gopalakrishnan, Zhinus Marzi, Upamanyu Madhow, and Ramtin Pedarsani. "Combating Adversarial Attacks Using Sparse Representations". In: *arXiv e-prints*, arXiv:1803.03880 (Mar. 2018), arXiv:1803.03880. DOI: 10.48550/arXiv.1803.03880. arXiv: 1803.03880 [stat.ML].

[209] Justin Cosentino, Federico Zaiter, Dan Pei, and Jun Zhu. "The Search for Sparse, Robust Neural Networks". In: *arXiv e-prints*, arXiv:1912.02386 (Dec. 2019), arXiv:1912.02386. DOI: 10.48550/arXiv.1912.02386. arXiv: 1912.02386 [cs.LG].

[210] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. "Adversarial Neural Pruning with Latent Vulnerability Suppression". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6575–6585. DOI: 10.48550/arXiv.1908.04355. URL: https://proceedings.mlr.press/v119/madaan20a.html.

[211] Adrian Siraj Rakin et al. "Robust sparse regularization: Defending adversarial attacks via regularized sparse network". English (US). In: *GLSVLSI 2020 - Proceedings of the 2020 Great Lakes Symposium on VLSI*. Proceedings of the ACM Great Lakes Symposium on VLSI, GLSVLSI. Funding Information: This work is supported in part by the National Science Foundation under Grant No.1931871. Publisher Copyright: © 2020 Association for Computing Machinery.; 30th Great Lakes Symposium on VLSI, GLSVLSI 2020 ; Conference date: 07-09-2020 Through 09-09-2020. Association for Computing Machinery, Sept. 2020, pp. 125–130. DOI: 10.1145/3386263.3407651.

[212] Tong Jian, Zifeng Wang, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. "Pruning Adversarially Robust Neural Networks without Adversarial Examples". In: *2022 IEEE International Conference on Data Mining (ICDM)*. 2022, pp. 993–998. DOI: 10.1109/ICDM54844.2022.00120.

[213] Karl Friston. "Hierarchical Models in the Brain". In: *PLoS Computational Biology* 4.11 (Nov. 2008), e1000211. DOI: 10.1371/journal.pcbi.1000211.

[214] Trevor Hastie and Robert Tibshirani. "Discriminant Analysis by Gaussian Mixtures". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 155–176. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02073.x. eprint: https://academic.oup.com/jrsssb/article-pdf/58/1/155/49098770/jrsssb_58_1_155.pdf. URL: https://doi.org/10.1111/j.2517-6161.1996.tb02073.x.

[215] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, Shlens Jon, and Wojna Zbigniew. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016 (June 2016), pp. 2818–2826. DOI: 10.1109/cvpr.2016.308. URL: https://cir.nii.ac.jp/crid/1361418519657311872.

[216] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. “Optimal errors and phase transitions in high-dimensional generalized linear models”. In: *Proceedings of the National Academy of Sciences* 116.12 (2019), pp. 5451–5460. DOI: 10.1073/pnas.1802705116. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1802705116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1802705116.

[217] Koki Okajima, Xiangming Meng, Takashi Takahashi, and Yoshiyuki Kabashima. “Average case analysis of Lasso under ultra sparse conditions”. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, Apr. 2023, pp. 11317–11330. DOI: 10.48550/arXiv.2302.13093. URL: https://proceedings.mlr.press/v206/okajima23a.html.

[218] Aurélien Decelle and Cyril Furtlehner. “Gaussian-spherical restricted Boltzmann machines”. In: *Journal of Physics A: Mathematical and Theoretical* 53.18 (Apr. 2020), p. 184002. DOI: 10.1088/1751-8121/ab79f3. URL: https://dx.doi.org/10.1088/1751-8121/ab79f3.

[219] Moshir Harsh, Jérôme Tubiana, Simona Cocco, and Rémi Monasson. “‘Place-cell’ emergence and learning of invariant data with restricted Boltzmann machines: breaking and dynamical restoration of continuous symmetries in the weight space”. In: *Journal of Physics A: Mathematical and Theoretical* 53.17 (Apr. 2020), p. 174002. DOI: 10.1088/1751-8121/ab7d00. URL: https://dx.doi.org/10.1088/1751-8121/ab7d00.

[220] Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. *Learning with Restricted Boltzmann Machines: Asymptotics of AMP and GD in High Dimensions*. 2025. arXiv: 2505.18046 [cs.LG]. URL: https://arxiv.org/abs/2505.18046.

[221] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by Simulated Annealing”. In: *Science* 220.4598 (1983), pp. 671–680. DOI: 10.1126/science.220.4598.671. eprint: https://www.science.org/doi/pdf/10.1126/science.220.4598.671. URL: https://www.science.org/doi/abs/10.1126/science.220.4598.671.

[222] Evelyn Fix and J. L. Hodges. “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. In: *International Statistical Review / Revue Internationale de Statistique* 57.3 (1989), pp. 238–247. ISSN: 03067734, 17515823. DOI: 10.2307/1403797. URL: http://www.jstor.org/stable/1403797 (visited on 01/14/2025).

[223] Yves Grandvalet and Yoshua Bengio. “Semi-supervised learning by entropy minimization”. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS’04. Vancouver, British Columbia, Canada: MIT Press, 2004, pp. 529–536.

[224] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. “A survey of feature selection and feature extraction techniques in machine learning”. In: *2014 Science and Information Conference*. 2014, pp. 372–378. DOI: 10.1109/SAI.2014.6918213.

[225] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. “UMAP: Uniform Manifold Approximation and Projection”. In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861.

[226] K ROSE, E GUREWITZ, and G. C FOX. “Constrained clustering as an optimization method”. English. In: *IEEE transactions on pattern analysis and machine intelligence* 15.8 (1993), pp. 785–794. ISSN: 0162-8828. DOI: 10.1109/34.236251.

[227] K ROSE. “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”. English. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239. ISSN: 0018-9219. DOI: 10.1109/5.726788.

[228] A. Decelle and C. Furtlehner. “Exact Training of Restricted Boltzmann Machines on Intrinsically Low Dimensional Data”. In: *Phys. Rev. Lett.* 127 (15 Oct. 2021), p. 158303. DOI: 10.1103/PhysRevLett.127.158303. URL: https://link.aps.org/doi/10.1103/PhysRevLett.127.158303.

[229] Anton Grishechkin, Abhirup Mukherjee, and Omer Karin. *Hierarchical cell identities emerge from animal gene regulatory mechanisms*. 2025. arXiv: 2412.11336 [q-bio.CB]. URL: https://arxiv.org/abs/2412.11336.

[230] Jesús A. De Loera and Edward D. Kim. *Combinatorics and Geometry of Transportation Polytopes: An Update*. 2013. arXiv: 1307.0124 [math.CO]. URL: https://arxiv.org/abs/1307.0124.

[231] KV Mardia and Peter Edmund Jupp. *Directional Statistics*. English. United States: John Wiley & Sons, Ltd, 2000. ISBN: 0-471-95333-4.

[232] *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.2 of 2024-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. URL: https://dlmf.nist.gov/10.9.E4.

[233] Jürg Fröhlich and Thomas Spencer. “The Kosterlitz-Thouless transition in two-dimensional Abelian spin systems and the Coulomb gas”. In: *Communications in Mathematical Physics* 81.4 (Dec. 1981). Appendix B, pp. 527–602. ISSN: 1432-0916. DOI: 10.1007/BF01208273. URL: https://doi.org/10.1007/BF01208273.

[234] *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.0 of 2024-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. URL: https://dlmf.nist.gov/10.41.E4.

[235] Davide Barilari. *Lecture notes on Differential Geometry*. Mar. 2023.

[236] Pedro Uribe, C. G. de Leeuw, and H. Theil. "The Information Approach to the Prediction of Interregional Trade Flows1". In: *The Review of Economic Studies* 33.3 (July 1966), pp. 209–220. ISSN: 0034-6527. DOI: 10.2307/2974414. eprint: https://academic.oup.com/restud/article-pdf/33/3/209/4460034/33-3-209.pdf. URL: https://doi.org/10.2307/2974414.

[237] G J D Hewings and B N Janson. "Exchanging Regional Input—Output Coefficients: A Reply and Further Comments". In: *Environment and Planning A: Economy and Space* 12.7 (1980), pp. 843–854. DOI: 10.1068/a120843. eprint: https://doi.org/10.1068/a120843. URL: https://doi.org/10.1068/a120843.

[238] Sue McNeil and Chris Hendrickson. "A note on alternative matrix entry estimation techniques". In: *Transportation Research Part B: Methodological* 19.6 (1985), pp. 509–519. ISSN: 0191-2615. DOI: https://doi.org/10.1016/0191-2615(85)90045-1. URL: https://www.sciencedirect.com/science/article/pii/0191261585900451.

[239] Richard Sinkhorn. "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices". In: *The Annals of Mathematical Statistics* 35.2 (1964), pp. 876–879. ISSN: 00034851. DOI: 10.1214/aoms/1177703591. URL: http://www.jstor.org/stable/2238545 (visited on 05/05/2025).

[240] Paul Knopp and Richard Sinkhorn. "Concerning nonnegative matrices and doubly stochastic matrices." In: *Pacific Journal of Mathematics* 21.2 (1967), pp. 343–348. DOI: 10.2140/pjm.1967.21.343.

[241] Martin Idel. "A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps". In: *arXiv e-prints*, arXiv:1609.06349 (Sept. 2016), arXiv:1609.06349. DOI: 10.48550/arXiv.1609.06349. arXiv: 1609.06349 [math.RA].

[242] M.V. Menon and Hans Schneider. "The spectrum of a nonlinear operator associated with a matrix". In: *Linear Algebra and its Applications* 2.3 (1969), pp. 321–334. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(69)90034-2. URL: https://www.sciencedirect.com/science/article/pii/0024379569900342.

[243] Daniel Hershkowitz, Uriel G. Rothblum, and Hans Schneider. "Classifications of Nonnegative Matrices Using Diagonal Equivalence". In: *SIAM Journal on Matrix Analysis and Applications* 9.4 (1988), pp. 455–460. DOI: 10.1137/0609038. eprint: https://doi.org/10.1137/0609038. URL: https://doi.org/10.1137/0609038.

[244] W. Edwards Deming and Frederick F. Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". In: *The Annals of Mathematical Statistics* 11.4 (1940), pp. 427–444. ISSN: 00034851. DOI: 10.1214/aoms/1177731829. URL: http://www.jstor.org/stable/2235722 (visited on 05/05/2025).

[245] Ahmed Douik and Babak Hassibi. "Manifold Optimization Over the Set of Doubly Stochastic Matrices: A Second-Order Geometry". In: *IEEE Transactions on Signal Processing* 67.22 (2019), pp. 5761–5774. DOI: 10.1109/TSP.2019.2946024.

[246] Mervin E. Muller. “A note on a method for generating points uniformly on n-dimensional spheres”. In: *Commun. ACM* 2.4 (Apr. 1959), pp. 19–20. ISSN: 0001-0782. DOI: 10.1145/377939.377946. URL: https://doi.org/10.1145/377939.377946.

[247] Carlos Pinzón and Kangsoo Jung. “Fast Python sampler for the von Mises Fisher distribution”. working paper or preprint. Aug. 2023. URL: https://hal.science/hal-04004568 (visited on 03/19/2025).

[248] ALEXANDER BARVINOK. “What Does a Random Contingency Table Look Like?” In: *Combinatorics, Probability and Computing* 19.4 (2010), pp. 517–539. DOI: 10.1017/S0963548310000039.

[249] I. J. Good. “Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables”. In: *The Annals of Mathematical Statistics* 34.3 (1963), pp. 911–934. ISSN: 00034851. URL: http://www.jstor.org/stable/2238473 (visited on 03/19/2025).