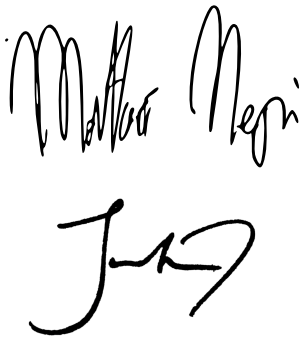
SAPIENZA
UNIVERSITÀ DI ROMA

Bidirectional Associative Memory as a model for feature extraction

Facoltà di Scienze Naturali, Fisiche e Matematiche
Corso di Laurea Magistrale in Fisica

Giovanni Colagè
ID number 1796955



Advisors
Prof. Matteo Negri
Prof. Jaron Kent-Dobias

Academic Year 2023/2024

Bidirectional Associative Memory as a model for feature extraction
Master thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: Colage.1796955@studenti.uniroma1.it

Contents

1 Introduction **1**

I State of the art **3**

2 Hopfield Model **5**
2.1 Definition of the Model 5
2.2 List of known results 7
2.3 Random Feature Hopfield Model 8
2.3.1 Definition of the Model 9
2.3.2 Learning transition 10

3 Bidirectional Associative Memory **13**
3.1 Model 13
3.2 Equilibrium analysis in the High-Load regime 15
3.2.1 Replica symmetric computation 15
3.2.2 Replica symmetric phase diagram 18
3.2.3 Replica symmetric phase diagram at $T = 0$ 20

II New results **23**

4 Random Feature BAM **25**
4.1 Definition of the Model 25
4.2 Replica symmetric computation 26
4.2.1 feature retrieval 28
4.2.2 Saddle point equation for the block matrixes of A 30
4.2.3 RS ansatz 31
4.2.4 Saddle point equations 33
4.2.5 Limit $\beta \to \infty$ 34
4.2.6 Limit $\alpha \to \infty$ (from $\beta \to \infty$) 35
4.2.7 Limit $\alpha_D \to \infty$ (from $\beta \to \infty$) 37

III Conclusions **39**

5 Discussion **41**

A Decoupling and equivalence between BAM and RBMs **43**

B Replica symmetric computation for the BAM **45**

C Replica symmetric computation for Random Feature BAM **49**
C.1 Integrating pattern magnetizations 50
C.2 Integrating over the feature magnetizations 51
C.3 Saddle point equation for $\hat{A}$ 54
C.4 Replica symmetric ansatz 55

D Algebra of RS matrixes **59**
D.1 Block matrix of RS matrixes 59
D.2 Inverse of a RS matrix 63

E Matrix calculus **65**

Bibliography **67**

Chapter 1

Introduction

The Hopfield model was introduced by Hopfield in [21] as a neural network model to emulate and understand the mechanism of associative memory that our brain performs when storing and recalling informations. Since the original work of Amit, Gutfreund, and Sompolinsky [1, 2, 3], who first developed a mean-field theory for the Hopfield model, several extensions and generalizations have been made. One of the most studied generalizations of the Hopfield model is a bipartite structure, where the network is divided into two layers of neurons and connections are allowed only between neurons of different layers. A neural network based on a bipartite topology is the bidirectional associative memories (BAMs), introduced by Kosko in 1988 [24]. Its simplest architecture is defined by two layers of neurons, with synaptic connections only between units of different layers: even without internal connections within each layer, information storage and retrieval are still possible through the reverberation of neural activities passing from one layer to another. The most popular application of a bipartite topology is the restricted Boltzmann machine (RBM) [34, 33, 10], which is widely used in computer science [34, 18, 20] and can be considered a prototype for machine learning models [20]. Even though the Hopfield model was proposed as a toy model for neurophysiology that accounts for biological learning through the Hebb rule, strong similarities between the information processing mechanisms of RBMs and of the Hopfield model have been demonstrated by statistical mechanics [33, 10, 6, 27]. In this machine, as in BAM settings, there are only interactions between units of different layers. However, while the RBMs is usually trained with gradient based methods, the BAM is trained with a simple Hebb rule. Although less powerful, it has the advantage that it is exactly solvable. At the same time, the operating principle of the BAM differs from that of the Hopfield model because in the BAM retrieval, information is passed from one layer to another according to appropriate dynamic rules: this mechanism is referred to in the literature as reverberation of information [24, 7]. The pioneering works on the Hopfield model have been the prototype for simulate how our brain stores and recalls informations. The network exhibits the ability to recall entire patterns from partial or noisy inputs, making it an archetype model in neuroscience and Artificial Intelligence (AI). But these early works do not provide theoretical insights for deep learning because they only consider unstructured input data (patterns), only shallow networks (at most two layers) and only interactions

through the Hebb rule, instead of gradient based methods.
Recently, the theory of Hopfield models has been extended to correlated data, providing a framework to understand how neural networks extract features from data. In fact, while most theoretical studies of (generalized) Hopfield Models assume uncorrelated distributions for the patterns [3, 12], in practical applications the patterns are samples from arbitrary complex distributions that correlate the variables [36]. In [31] this limitation has been addressed by proposing a generative model for the patterns where each pattern is produced by the linear combinations of a fixed vocabulary of so-called features, weighted by pattern specific coefficients, followed by an element-wise non-linearity. Two main advantages emerge from incorporating this structure for the patterns in the Hopfield Model as it was shown in [31]. The first is to provide theory of storage in Hopfield networks with more realistic input data. The second is that, even without modifying the Hebb rule, the model produces a completely new behaviour: if the correlations in the data are strong enough, the model switches from a storage phase to a learning phase, in the sense that attractors appear corresponding to the features in the data. This behaviour opens up a new paradigm for associative memories and shows that it may have some phenomenology in common with neural networks.
Motivated by this, we insert a structure for the input data in the BAM as well, possibly providing a first comprehension for the mechanism of feature extraction in RBMs. We plant two different kind of memories (patterns) in the model, so that the role of the two layers recall the hidden and the visible layer in RBMs: for the visible layer we chose memories derived from a linear combination of features weighted by coefficients, while for the hidden layer the memories are exactly the coefficients linked with the patterns of the visible layer. The structure of input data we chose for our model is related to the Hidden-Manifold Model [15, 13, 28], which has been used as an analytically solvable model of feed forward neural networks. Moreover the same data structure was also discussed by [29] in relation to the mapping between a Hopfield network and a restricted Boltzmann machine.
Using the replica method we obtain saddle point equations for the order parameters of the model. During the computation we give special importance to the spin condensation in the hidden layer, which is crucial for the mechanism of feature extraction. Finally we proceed to analyse at zero temperature different limits of the saddle point equations to investigate how the BAM performs its retrieval task in different regimes. For example we analyse both the limit in which the number of memories diverges with respect to the number of neurons in the visible layer, and the limit in which the number of features diverges with respect to the number of neurons in the visible layer. The following work is structured as follows. In Part I we recall the State of the Art for the Hopfield model, and the results obtained when Random-Feature data are incorporated. For the BAM, we provide a description of the model and the computation of the free energy through the RS computation as in [26], where however a small mistake occurred as pointed out in [7]. Then, in Part II we provide all the theoretical analysis for the new model introduced in this work, and the saddle point equations for the order parameters of the model. Finally in Part III we discuss conclusions and future perspectives.

Part I

State of the art

Chapter 2

Hopfield Model

Neural-network models are complex systems designed on the basis of the associative memory notion and on the principle that stable neural activities encode retrieved patterns of information. By associative memory we mean the ability of cortical modules in mammals' brain to remember names, objects, faces, schemes, images, ecc. starting from incomplete or corrupted data supply.
The Hopfield Model proposed by Hopfield in [21] view the human memory as a collective property of large interconnected neural networks and has been the theoretical prototype to simulate how our brain stores and recalls informations.
In Section 2.1 we provide the definition of the Hopfield Model and we explain how it performs the task of associative memory.
In Section 2.2 we summarize the main results that have been obtained in [1, 2, 3, 4], when they presented for the first time the statistical mechanics of Hopfield's Model far from saturation and near saturation.
Finally in Section 4.88 we introduce the known results for the Random Features Hopfield Model, focusing on the learning transition.

2.1 Definition of the Model

The Hopfield Model is a fully connected neural network consisting in N neurons $\mathbf{S}$. A schematic representation is exhibit in Figure 2.1. Each neuron is a node of a fully connected graph and it take a binary value, $S_i = \{-1, +1\}$, $\forall\ 1, \dots, N$. These values represent respectively the quiescent and the active state for the neuron. The neurons interact among each other through the couplings J_{ij}, which represent the synaptic efficacies between node i and node j.
The model is governed by the Hamiltonian:

$$H(\mathbf{S}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij} S_i S_j \tag{2.1}$$

Associative memory models are built to recognize a certain group of words or patterns, so the next step is to formalize how the information is encoded in neural networks. A pattern, or memory, is defined as a sequence of random variables $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_N\}$. Since we want to store several patterns, namely P, we have to

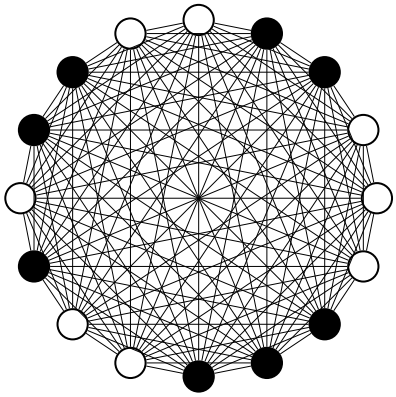


Figure 2.1. Example of a Hopfield network with $N = 16$ nodes. The binary variables $+1$ and -1 are mapped in black and white colors.

introduce another index μ that takes values $\{1, \dots, p\}$ to label the different memories $\boldsymbol{\xi}^\mu$. In doing this, we shall assume that each ξ_i^μ is independent from the others.
The choice of the synaptic coupling J_{ij}, $\forall\ i, j = 1, \dots, N$ ensuring the local attractiveness of each pattern under the neural dynamics is the one incorporating Hebb's learning rule:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{2.2}$$

This is a biologically motivated rule proposed by Hebb in [17], which states that if if two neurons are required to be in the same state ($\xi_i^\mu = \xi_j^\mu$) we increase their mutual interaction strength J_{ij}, otherwise we decrease J_{ij}.
At $T = 0$ the attractors of the dynamic, the P memories $\boldsymbol{\xi}^\mu$, satisfy the following steepest descent dynamical equations:

$$\mathbf{S}_i^{t+1} = \text{sign}\Big(\sum_j J_{ij} \mathbf{S}_j^t\Big) \ \longrightarrow\ \boldsymbol{\xi}_i^\mu = \text{sign}\Big(\sum_j J_{ij} \boldsymbol{\xi}_j^\mu\Big) \tag{2.3}$$

Where the right equation highlight how the $\boldsymbol{\xi}^\mu$ are fixed points of such a dynamics. In fact:

$$\begin{aligned} \sum_j J_{ij} \xi_j^\mu &= \frac{1}{N} \sum_j \sum_\nu \xi_j^\mu \xi_i^\nu \xi_j^\nu \\ &= \sum_\nu \xi_i^\nu \frac{1}{N} \sum_j \xi_j^\mu \xi_j^\nu \\ &= \sum_\nu \xi_i^\nu \Big(\delta_{\mu\nu} + (1 - \delta_{\mu\nu}) \frac{1}{N} \sum_j \xi_j^\mu \xi_j^\nu \Big) \quad , \end{aligned} \tag{2.4}$$

where in the $\mu \neq \nu$ case

$$\frac{1}{N} \sum_j \xi_j^\mu \xi_j^\nu = O(\frac{1}{\sqrt{N}}) \quad , \tag{2.5}$$

and can therefore be neglected.

The next step is to introduce a set spin-dependent quantities measuring the resemblance of a given network configuration with the stored patterns. These quantities will play the role of order parameters for the Hopfield model. We define p overlaps m_μ, $\mu = 1, \dots, p$ between the patterns and the neurons, also called Mattis magnetizations, as:

$$m_\mu(\mathbf{S}) = \frac{1}{N}\sum_{i=1}^{N} S_i \xi_i^\mu \tag{2.6}$$

The Hamiltonian can be nicely written in terms of these order parameters as:

$$\mathbf{H}(\mathbf{S}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij} S_i S_j = -\frac{1}{2}\frac{1}{N}\sum_{\mu=1}^{p}\left(\sum_{i=1}^{N} S_i \xi_i^\mu\right)\left(\sum_{j=1}^{N} S_j \xi_j^\mu\right) = -\frac{N}{2}\sum_{\mu=1}^{p} m_\mu^2 \tag{2.7}$$

Where it is possible to check that the patterns $\boldsymbol{\xi}^\mu$ are extremal points of the energy function.
When considering the patterns $\boldsymbol{\xi}^\mu$ as random quenched variables, the tools developed in statistical mechanics of disordered systems become essential for analysing the thermodynamic properties of the model.
Amit et al. initially computed these properties for finite p (or low-load regime) in [1, 2], and later in [3] they extended the analysis to an extensive number of patterns $p = \alpha N$, assuming ξ_i^μ to be independent and identically distributed random variables with distribution:

$$\mathrm{P}(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu + 1) + \frac{1}{2}\delta(\xi_i^\mu - 1) \tag{2.8}$$

2.2 List of known results

In this section we present the main results obtained in the theoretical works done on the Hopfield model in [1, 2, 3, 4].

- It exists a range of α where the system provides effective retrieval of memory. This is due to the fact that, despite the presence of a finite amount of noise which destabilizes the stored patterns, the modified stable states remain very near the original patterns, provided

  $$\alpha < \alpha_c = 0.138 \quad . \tag{2.9}$$

  These metastable states which are correlated with only a single memory are called retrieval states. As long as $\alpha < \alpha_c$, the overlap between them and the original pattern is greater than $m \simeq 0.97$, which is the value of the overlap at α_c. As α decreases to zero, m increases (at $T = 0$) exponentially fast towards 1,

  $$1 - m \simeq \exp\left\{-\frac{1}{2\alpha}\right\}, \quad \alpha \to \infty \quad . \tag{2.10}$$

- The maximum storage capacity of the network depends on the precision which one requires from retrieval. If we relax the constraint to have stability for all bits, and for example we accept some spins misaligned with the original patterns, then the storage capacity is given by $P < N\alpha_c$. However retrieval of patterns free of errors occurs when $P < \frac{N}{2\ln N}$

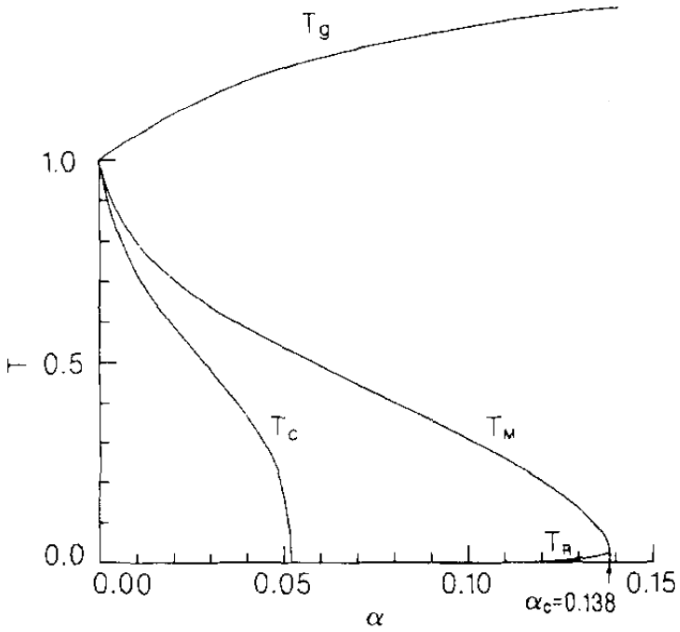


Figure 2.2. Phase diagram taken from [3] illustrating the stability landscape of the standard Hopfield Network. T_g is the transition line between the paramagnetic solution and the spin glass phase. Above the line $T_M(\alpha)$ there is the spin glass phase where the attractors of the dynamics are not correlated with the memories; between the line $T_c(\alpha)$ and $T_M(\alpha)$ the stored patterns $\boldsymbol{\xi}_\mu$ are metastable states of the system, and retrieval is possible if the initial configuration of the system has a sufficiently large overlap with a pattern; below $T_c(\alpha)$ the memories become stable global minima of the free energy.

- Retrieval states exist as thermodynamic, metastable states also at finite T, for $T < T_M(\alpha)$. The maximum temperature $T_M(\alpha)$ decreases from unity at $\alpha = 0$, to zero at $\alpha = \alpha_c$. At finite temperature we can store $P < \alpha_c(T)N$ with $\alpha_c(T) < \alpha_c$, for higher values of P the attractors are not correlated to the patterns
- At finite α there are two classes of spurious states, namely metastable states other than the single pattern retrieval states. At sufficiently small α there are mixture states, which have finite overlaps with several patterns. In addition, there is a spin-glass (SG) phase, at all finite α.

All of these results are summarized in the phase diagram reported in Figure .

2.3 Random Feature Hopfield Model

To provide theoretical insights for deep learning the memories planted in the Hopfield model must resemble more realistic input data. That is why in [31] they introduced the Random-Features Hopfield model where they lose the ansatz of uncorrelated distributions for the memories and instead they built a model where

the memorized patterns are generated from a linear combination of features that live in a hidden manifold of dimension D and that is then mapped through a non-linear function to the N-dimensional space of configurations. They showed that, for their model, it exists a thermodynamic phase where the features used to construct the patterns are the absolute minima of the free energy and they became the attractors of the dynamic. In the following sub-sections we are going to summarize the results obatined in [31].

2.3.1 Definition of the Model

The Random Feature Hopfield Model [31] is composed by N binary spins $S_i = \{-1, +1\}$, $i = 1 \dots, N$, governed by the standard Hamiltonian of the Hopfield model:

$$H(\mathbf{S}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij} S_i S_j \ . \tag{2.11}$$

The interaction matrix J_{ij}, $\forall\ i, j = 1, \dots, N$ is defined via the Hebbian rule 2.12 through a set of P patterns $\{\boldsymbol{\xi}_\nu\}_{\nu=1}^{P}$:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \ . \tag{2.12}$$

Although, as we have seen in 2.1, while in the standard statistical physics setting [3] the ξ_i^ν are independently and uniformly distributed binary spins, here instead they are structured patterns and they are given by a linear projection and a latent vector composed with a non-linearity:

$$\xi_i^\nu = \sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D} c_k^\nu f_{ki}\right) \ , \tag{2.13}$$

where $\sigma(\cdot)$ is a generic non-linear function, f_{ki} is called the matrix of features and c_k^ν is the matrix of coefficients. We consider the case of independent and identically distributed uniform binary features $f_{ki} = \pm 1$, independent identically distributed standard Gaussian coefficients c_k^ν, and $\sigma(\cdot)$ equal to a sign function, a sparse and linear version of this structure was analysed in [29].
We investigate the regime where the latent space dimension D scales linearly with N in the thermodynamic limit. Therefore, an additional external parameter compared to the standard Hopfield model is introduced:

$$\alpha_D = \frac{D}{N} \ . \tag{2.14}$$

By tuning D we can switch between weakly and strongly correlated memories. In the limit $\alpha_D \to \infty$ where the patterns become uncorrelated, the recovery of the standard Hopfield model is expected.
The dynamical update rule, at zero temperature, is given by the same formula of the classic Hopfield Model case:

$$\mathbf{S}_i^{t+1} = \text{sign}\left(\sum_j J_{ij}\mathbf{S}_j^t\right) \ . \tag{2.15}$$

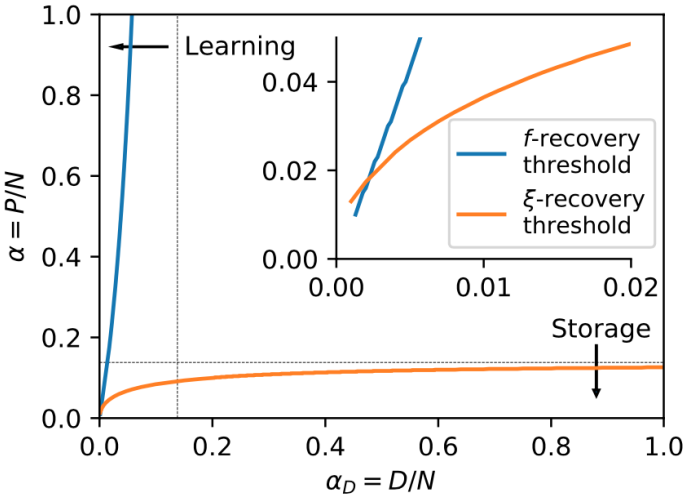


Figure 2.3. Phase diagram taken from [31] illustrating the three regimes in which the Random Feature Hopfield Model can operate: under the orange line, in the storage phase, the memories $\boldsymbol{\xi}_m u$ are attractors. Above the blue line, in the learning phase, the features $\mathbf{f}_k$ are attractors. The spin glass phase, between the two lines, where the attractors are uncorrelated with both the features and the memories. The two asymptotes are at $\alpha \simeq 0.138$ and $\alpha_D \simeq 0.138$.

The thermodynamics of the system can be examined in the replica symmetric ansatz, facilitated by a Gaussian Equivalence property verified by the model and already studied in [15, 14, 13, 28, 22, 5]. Moreover to delineate the model's phase diagram we introduced two sets of order parameters:

$$m_\mu(\mathbf{S}) = \frac{1}{\sqrt{N}} \sum_i \xi_i^\mu S_i \ , i \in [p] \ . \tag{2.16}$$

We called these pattern magnetizations, and are the same order parameters defined in 2.6, and

$$\mu_k(\mathbf{S}) = \frac{1}{\sqrt{N}} \sum_i f_{ki} S_i \ , k \in [D] \ . \tag{2.17}$$

We call these feature magnetizations.

2.3.2 Learning transition

In [31] they made an ansatz on the structure of the solution for both of the order parameters reported in 2.16 and 2.17. they studied two cases: the case where the model retrieves only one of the features, feature retrieval, and the case where the model retrieves only one of the examples, pattern retrieval.
In the case of feature retrieval, it means that they searched for a solution in the form of

$$\boldsymbol{\mu} = (\mu, 0, \dots, 0) \ , \quad \mathbf{m} = (0, \dots, 0) \ , \tag{2.18}$$

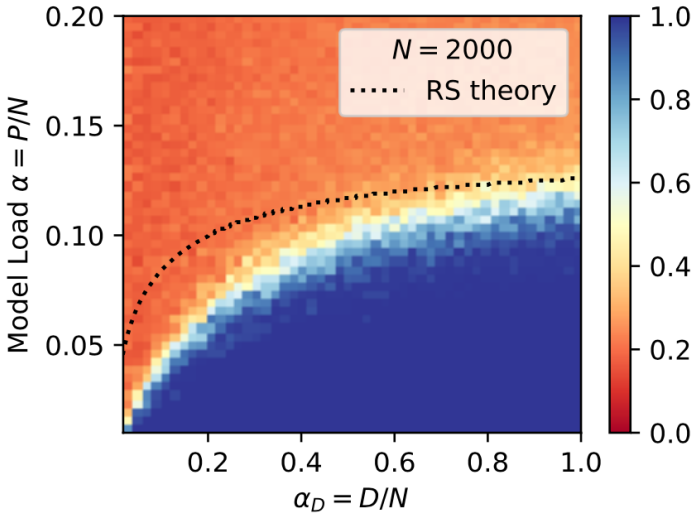


Figure 2.4. Image taken from [31], it shows the comparison with numerical results for the retrieval of one pattern. Each pixel represents the mean pattern magnetization for given values of α and α_D, averaged over 25 samples of size $N = 2000$. The simulations are performed initializing the model to a pattern $\boldsymbol{\xi}_\mu$, running the update rule 2.15, then measuring m_μ at convergence. Notably, catastrophic forgetting occurs at an α threshold lower than the predicted value, with the mismatch intensifying for lower values of α_D. This observation hints at the potential influence of strong correlations in the breakdown of the Replica Symmetry.

in the termodynamic limit. Instead in the case of pattern retrieval, they searched for a solution in the form of

$$\boldsymbol{\mu} = (0, \dots, 0) \ , \quad \mathbf{m} = (m, 0, \dots, 0) \ , \tag{2.19}$$

in the termodynamic limit.
As shown in Figure 2.3, in the limit $\beta \to \infty$ for $\alpha > \alpha_c(\alpha_D)$ the feature magnetization becomes finite with a discontinuous jump, showing that the model is actually capable of storing the features f as attractors for the model. The critical point $\alpha_c(\alpha_D)$ rapidly increases with α_D up to the point where it diverges for $\alpha_D \simeq 0.138$. It is not a coincidence that this critical value is numerically identical to the critical capacity of the Standar Hopfield Model, in fact in the $\alpha \to \infty$ limit, [31] demonstrated that:

$$\frac{1}{P} \sum_{\nu=1}^{P} \xi_i^\nu \xi_j^\nu \xrightarrow{P \to \infty} k_1^2 \frac{1}{D} \sum_{k=1}^{D} f_{ki} f_{kj} \quad . \tag{2.20}$$

Consequently, the saddle-point equations must mirror those of the standard Hopfield model, where μ plays the role of magnetization and f that of the retrieved patterns. One way to look at this behaviour in Figure 2.3 is to fix a value for α and then move horizontally in the phase diagram: with α_D low enough the model is able to retrieve the features, then, when α_D is increased and the features become too many, the equivalent of a catastrophic forgetting happens. This transition happens at the Hopfield critical capacity only if $\alpha = \infty$, where the matching between the two models is perfect.
Another interesting limit to consider is $\alpha_D \to \infty$ in this limit, as the examples become uncorrelated, the equations converge to the Standard Hopfield Model ones. By moving vertically in the phase diagram 2.3, lowering α, a thermodynamic phase with $m > 0$ emerges in the $\alpha_D \to \infty$ limit exactley at the numerical value of $\alpha = 0.138$. Decreasing α_D the example patterns become more correlated, the retrieval phase shrinks until the catastrophic forgetting happens at $\alpha = 0$ for $\alpha_D \to 0$.
Numerical simulations were performed to verify the predictions of the replica symmetric theory, the results are reported in Figure 2.4.
If we move away from the regime $\alpha_D >> 1$ where the model converges to the standard Hopfield Network we find that the replica symmetric theory predicts a value of $\alpha_c(\alpha_D)$ greater than the one obtained with the numerical simulations. Furthermore, the mismatch increases as α_D is lowered, suggesting that strong correlations might be responsible of a failure of the RS ansatz.
The behaviour that the model expresses in the learning phase resembles the mechanism of features extraction that is present in deep neural networks and shallow generative models in the unsupervised setting [25, 20, 19, 37].

Chapter 3

Bidirectional Associative Memory

The bidirectional associative memory (BAM) was introduced by Kosko in [24] as an attempt to overcome the lack of internal organization of information in the original Hopfield model and to account for structured retrieval of patterns. It is a generalization of a neural network based on a bipartite topology, where the interactions are only between units of different layers. Recentley the BAM has become topical again in the context of machine learning [32].
In Section 3.1 we provide the definition of the BAM and explain how it performs the task of associative memory for a pattern pair. Section 3.2 is divided in three subsections: In 3.2.1 we provide the free energy of the BAM through Replica Symmetry ansatz. Here we highlight only the crucial steps of the computation while a more detailed discussion is enhanced in Appendix B. In 3.2.2 the saddle point equations and the phase diagram, in the (α, T) plane, of the BAM are provided. It is moreover studied how a finite asymmetry between the two layers of the BAM affects the retrieval task of the model. In 3.2.3 it is analyzed the phase diagram at $T = 0$ considering the number of pattern pairs planted in the model and the asymmetry between the two layers. At the end it is exhibited a final discussion of the retrieval properties between the BAM and the Hopfield Model.

3.1 Model

The BAM model is defined by two layers of neurons, namely $\mathbf{S}$ and $\bar{\mathbf{S}}$. The interaction topology of the model is a bipartite graph: each unit in one layer interacts with all the units in the other layer, as it is shown in Figure 3.1.
Layer 1 has N neurons S_i, with $i = 1, \dots, N$, while layer 2 has $\bar{N}$ neurons $\bar{S}_k$, with $k = 1, \dots, \bar{N}$, and the two layers interact with each other through the coupling W_{ik}, that represents the synaptic interaction between node i in the first layer and the node k in the second layer.
Each neuron is considered as a binary variable, $S_i = \{-1, +1\}$, $\bar{S}_k = \{-1, +1\}$, $\forall\, i = 1, \dots, N$ and $k = 1, \dots, \bar{N}$, where the value $+1$ is associated with a firing state for the neuron, while -1 means the neuron is silent.

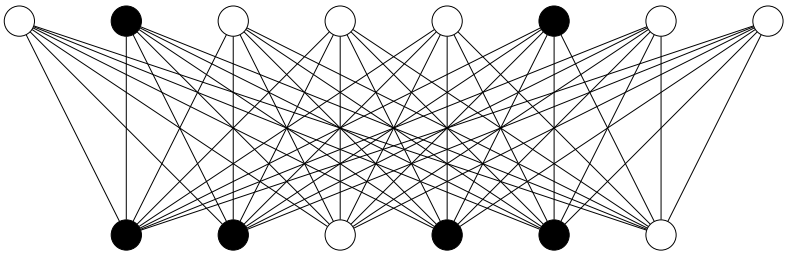


Figure 3.1. Example of a BAM network with $N = 8$ nodes in the upper layer and $\bar{N} = 6$ nodes in the downer layer. The binary variables $+1$ and -1 are mapped in black and white colors.

The BAM Hamiltonian (or cost function) is given by

$$H(\mathbf{S}, \bar{\mathbf{S}}) = -\sum_{i=1}^{N}\sum_{k=1}^{\bar{N}} W_{ik} S_i \bar{S}_k \tag{3.1}$$

The interaction matrix W is constructed in such a way that the usual Hebb's rule pattern storage is generalized to such a bipartite structure. We therefore define two sets of P memories, one for each layer, deneoted with $\boldsymbol{\xi}^\mu$ and $\bar{\boldsymbol{\xi}}^\mu$, with $\mu = 1, \dots, p$. Each memory has a dimension compatible with the corresponding layer of neurons, so that $\boldsymbol{\xi}^\mu$ and $\bar{\boldsymbol{\xi}}^\mu$ are vectors with respectively N and $\bar{N}$ components for each μ and each component is a binary variable $\xi_i^\mu = \{-1, +1\}$, $\bar{\xi}_k^\mu = \{-1, +1\}$, $\forall\ i = 1, \dots, N$ and $k = 1, \dots, \bar{N}$. The synaptic weight matrix is then constructed by using the following generalized Hebb's rule:

$$W_{ik} = \frac{1}{\sqrt{N\bar{N}}} \sum_{\mu=1}^{p} \xi_i^\mu \bar{\xi}_k^\mu \tag{3.2}$$

The scale factor $L = \sqrt{N\bar{N}}$ in the weight matrix is chosen according to reference [26, 35] and to have a non-trivial free energy density in the thermodynamic limit, where $N, \bar{N}, L \to \infty$.
To inspect the ability of the BAM to handle pairs of memories, given the bipartite structure of the network and the absence of synaptic interactions inside the same layer, it is useful to check if retrieval is feasible when both layers have a non-zero overlap with a given memory. We assume that the memories are drawn independently with independent and identically distributed components at each layer.

$$\mathrm{P}(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu + 1) + \frac{1}{2}\delta(\xi_i^\mu - 1) \tag{3.3}$$

$$\mathrm{P}(\bar{\xi}_i^\mu) = \frac{1}{2}\delta(\bar{\xi}_i^\mu + 1) + \frac{1}{2}\delta(\bar{\xi}_i^\mu - 1) \tag{3.4}$$

In the noiseless regime, the attractors of the dynamics, the p couples $(\boldsymbol{\xi}^\mu, \bar{\boldsymbol{\xi}}^\mu)$ satisfy the following steepest descent dynamical equations:

$$\mathbf{S}_i^{t+1} = \mathrm{sign}\left(\sum_k W_{ik} \bar{\mathbf{S}}_k^t\right) \longrightarrow \boldsymbol{\xi}_i^\mu = \mathrm{sign}\left(\sum_k W_{ik} \bar{\boldsymbol{\xi}}_k^\mu\right) \tag{3.5}$$

$$\bar{\mathbf{S}}_k^{t+1} = \text{sign}\left(\sum_i W_{ik}\mathbf{S}_k^t\right) \longrightarrow \bar{\boldsymbol{\xi}}_k^\mu = \text{sign}\left(\sum_i W_{ik}\boldsymbol{\xi}_i^\mu\right) \tag{3.6}$$

where the right equations highlight how the couple memories $(\boldsymbol{\xi}^\mu, \bar{\boldsymbol{\xi}}^\mu)$ are fixed points of such a dynamics.
Moreover, by introducing as order parameters of the theory the Mattis overlap $m_\mu(\mathbf{S})$, $\nu_\mu(\bar{\mathbf{S}})$, we can show that the couple memories, attractors of equations 3.5 and 3.6, are extremal point of the energy function, by writing it in the following form:

$$H(\mathbf{S}, \bar{\mathbf{S}}) = -\sum_{i=1}^{N}\sum_{k=1}^{\bar{N}} W_{ik}S_i\bar{S}_k = -\frac{1}{L}\sum_{\mu=1}^{p}\left(\sum_{i=1}^{N} S_i\xi_i^\mu\right)\left(\sum_{k=1}^{\bar{N}} \bar{S}_k\bar{\xi}_k^\mu\right) = -L\sum_{\mu=1}^{p} m_\mu(\mathbf{S})\nu_\mu(\bar{\mathbf{S}}) \tag{3.7}$$

Where the Mattis overlap $m_\mu(\mathbf{S})$, $\nu_\mu(\bar{\mathbf{S}})$ of the configurations $(\mathbf{S}, \bar{\mathbf{S}})$ with the couple memories $(\boldsymbol{\xi}^\mu, \bar{\boldsymbol{\xi}}^\mu)$ are defined as:

$$m_\mu(\mathbf{S}) = \frac{1}{N}\sum_{i=1}^{N} S_i\xi_i^\mu \tag{3.8}$$

$$\nu_\mu(\bar{\mathbf{S}}) = \frac{1}{\bar{N}}\sum_{k=1}^{\bar{N}} \bar{S}_k\bar{\xi}_k^\mu \tag{3.9}$$

3.2 Equilibrium analysis in the High-Load regime

There are two main computational regimes under which this model can be analysed: the low-load regime, in which the number of patterns is finite with respect to $L = \sqrt{N\bar{N}}$, or it grows proportionally up to $\ln L$, and the high-load regime, in which the number of patterns scales proportionally to L. The low-load scenario was analysed in [26] where it was proven that a second-order phase transition splits a paramagnetic phase (with no retrieval) and a ferromagnetic phase, where retrieval spontaneously emerges. The high-load regime was explored in [26], where results where provided only at zero temperature, and in [7], where they extended the phase diagram of the model also at $T \neq 0$. In the following subsections we are going to compute the free energy in the high-load regime using the replica method [30].

3.2.1 Replica symmetric computation

Since we are interested in the thermodynamic limit $\sqrt{N\bar{N}} = L \to \infty$, we choose a regime where N, $\bar{N}$ and p are all proportional between them. At the same time, we keep the following ratios fixed

$$\alpha = \frac{p}{N}\ , \tag{3.10}$$

$$\gamma = \sqrt{\frac{N}{\bar{N}}}\ . \tag{3.11}$$

These ratios are respectively called network load and asymmetry between layers and, combined with the thermal noise β, will be the control parameters for our model.

We need now to compute the quenched averaged free energy of the model:

$$F_Q = -\lim_{n\to 0}\frac{1}{\beta L}\langle\langle \ln Z\rangle\rangle \tag{3.12}$$

where $\langle\langle\cdot\rangle\rangle$ denotes the expectation value with respect to the distribution of the patterns of both layers, and Z is the partition function.
In order to compute the average of $\ln Z$ in equation 3.12 we use the replica method [30], that consists in writing the average of the logarithm as

$$\langle\langle \ln Z\rangle\rangle = \lim_{n\to 0}\frac{\langle\langle Z^n\rangle\rangle - 1}{n} \tag{3.13}$$

The replicated partition function averaged over the disorder reads

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \Bigg\langle\Bigg\langle \mathrm{Tr}_S\mathrm{Tr}_{\bar S}\int\prod_{\mu a} dm_\mu^a d\nu_\mu^a \exp\Big\{\beta\sum_{\mu a} m_\mu^a\nu_\mu^a\Big\} \\ \delta\Big(m_\mu^a - \frac{1}{\sqrt{N}}\sum_i \xi_i^\mu S_i^a\Big)\delta\Big(\nu_\mu^a - \frac{1}{\sqrt{\bar N}}\sum_k \bar\xi_k^\mu \bar S_k^a\Big)\Bigg\rangle_\xi\Bigg\rangle_{\bar\xi}\end{aligned} \tag{3.14}$$

where we are referring to the sum over the replicated spin configurations for both layers of spin with the notation $\mathrm{Tr}_\sigma = \sum_{\sigma^1=\{\pm1\}^N}\cdots\sum_{\sigma^n=\{\pm1\}^N}$, $\boldsymbol{\sigma}$ being both $\mathbf{S}$ and $\bar{\mathbf{S}}$. Moreover in 3.14 we also introduced the two following auxillary variables:

$$m_\mu^a = \frac{1}{\sqrt{N}}\sum_i \xi_i^\mu S_i^a, \quad a\in[n]\ , \mu\in[p], \tag{3.15}$$

$$\nu_\mu^a = \frac{1}{\sqrt{\bar N}}\sum_k \bar\xi_k^\mu \bar S_k^a, \quad a\in[n]\ , \mu\in[p]. \tag{3.16}$$

That are order parameters for this model.
To decouple the quadratic interaction in the Hamiltonian, we introduce a $\delta(\cdot)$ function and its Fourier transform. However in [26] and [7], they implemented a different decoupling transformation required by the two-layer structure of the BAM. While the final result is equivalent, this last approach can be exploited to to derive a structural analogy between the BAM's partition function and the partition function of two coupled RBMs. This topic is further discussed in Appendix A.
To describe the retrieval properties we will now distinguish between a first finite subset of $\mu = 1,\dots,l$ low components, and the remaining (extensive in the high-load regime) high components $\mu = l+1,\dots,p$, according to references [35, 26, 7]. The first set encodes for pattern pairs that can be retrieved, namely those having eventually a non-zero overlap with spin configurations, while the other extensive set will act as a quenched noise with respect to the former.
As a consequence, the partition function is split into a signal and a noise term as well and the low components, representing the signal, are rescaled according to the scheme:

$$m_\mu^a \to \sqrt{N}m_\mu^a, \quad a\in[n], \mu\in[1,\dots,l], \tag{3.17}$$

$$\nu_\mu^a \to \sqrt{\bar N}\nu_\mu^a, \quad a\in[n], \mu\in[1,\dots,l]. \tag{3.18}$$

After averaging over the high patterns and integrating in the m_μ and ν_μ fields for $\mu = l+1, \dots, p$, the partition function reads:

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \Bigg\langle\Bigg\langle \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a \exp\Bigg\{-\frac{p}{2}\ln\det\Big(\mathbb{1} - \beta^2 Q\bar{Q}\Big)\Bigg\} \\ \exp\Bigg\{-\beta\sum_a\sum_{\mu=1}^{l}\sqrt{N\bar{N}}m_\mu^a\nu_\mu^a\Bigg\} \\ \exp\Bigg\{\beta\sum_a\sum_{\mu=1}^{l}\Big(m_\mu^a\sum_i\xi_i^\mu S_i^a + \nu_\mu^a\sum_k\bar{\xi}_k^\mu\bar{S}_k^a\Big)\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2}\sum_{ab}r_{ab}q_{ab} - \frac{p\beta^2}{2N}\sum_{ab}r_{ab}\sum_i S_i^a S_i^b\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2}\sum_{ab}\bar{r}_{ab}\bar{q}_{ab} - \frac{p\beta^2}{2\bar{N}}\sum_{ab}\bar{r}_{ab}\sum_k\bar{S}_k^a\bar{S}_k^b\Bigg\}\Bigg\rangle_{\{\xi\}_1^l}\Bigg\rangle_{\{\bar{\xi}\}_1^l}\end{aligned} \tag{3.19}$$

Where we introduced the spin glasses parameters q_{ab} and $\bar{q}_{ab}$ by the definitions:

$$q_{ab} = \frac{1}{N}\sum_i S_i^a S_i^b, \quad a, b \in [n], \tag{3.20}$$

$$\bar{q}_{ab} = \frac{1}{\bar{N}}\sum_k \bar{S}_k^a \bar{S}_k^b, \quad a, b \in [n], \tag{3.21}$$

and their conjugate fields r_{ab} and $\bar{r}_{ab}$. We defined the identity matrix in the n dimensional space as:

$$\mathbb{1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}_{n\times n} \tag{3.22}$$

and with the notation Q and $\bar{Q}$ we mean the matrixes made of the spin glasses parameters:

$$Q = \begin{pmatrix} q_{11} & q_{12} & \dots & q_{1n} \\ q_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & q_{n-1n} \\ q_{n1} & \dots & q_{nn-1} & q_{nn} \end{pmatrix}_{n\times n} \tag{3.23}$$

and the same definition holds for $\bar{Q}$ and $\bar{q}_{ab}$.
We now assume replica symmetry and we make a RS ansatz for all the order

parameters:

$$q_{ab} = \delta_{ab} + q(1-\delta_{ab}) \tag{3.24}$$
$$\bar{q}_{ab} = \delta_{ab} + barq(1-\delta_{ab}) \tag{3.25}$$
$$r_{ab} = r\delta_{ab} + r(1-\delta_{ab}) \tag{3.26}$$
$$\bar{r}_{ab} = \bar{r}\delta_{ab} + \bar{r}(1-\delta_{ab}) \tag{3.27}$$
$$m_\mu^a = m_\mu \tag{3.28}$$
$$\nu_\mu^a = \nu_\mu. \tag{3.29}$$

With this ansatz we are able to compute the spin traces and the $\ln\det\Big(\mathbb{1} - \beta^2 Q\bar{Q}\Big)$, and by introducing the definitions for the control parameters 3.10 the final form for the free energy of the BAM model reads:

$$\begin{aligned} f^{RS} = &\sum_{\mu=1}^{l} m_\mu\nu_\mu + \frac{\alpha\beta}{2} r(1-q) + \frac{\alpha\beta}{2}\bar{r}(1-\bar{q}) + \\ &- \frac{\gamma}{\beta}\Big\langle \int Dz \ln\Big[2\cosh(\beta\sqrt{\bar{\gamma}\alpha r}z + \beta\bar{\gamma}\sum_\mu m_\mu\xi^\mu)\Big]\Big\rangle + \\ &- \frac{\bar{\gamma}}{\beta}\Big\langle \int D\bar{z} \ln\Big[2\cosh(\beta\sqrt{\gamma\alpha\bar{r}}\bar{z} + \beta\gamma\sum_\mu \nu_\mu\bar{\xi}^\mu)\Big]\Big\rangle + \\ &- \frac{\alpha}{2\beta}\ln\Big(1-\beta^2(1-q)(1-\bar{q})\Big) - \frac{\alpha}{2\beta}\frac{q(1-\bar{q}) + \bar{q}(1-q)}{1-\beta^2(1-q)(1-\bar{q})} \end{aligned} \tag{3.30}$$

where $\bar{\gamma} = \gamma^{-1} = \sqrt{\frac{\bar{N}}{N}}$ and z, $\bar{z}$ are two independent identically distributed standard gaussian variables. All explicit computations are reported in Appendix B.

3.2.2 Replica symmetric phase diagram

The values of the order parameters at fixed control parameters can be obtained by evaluating the saddle points of f^{RS} of equation B.20. The set of self-consistent equations to be fulfilled by these values is shown below:

$$\mathbf{m} = \Big\langle \int D\bar{z}\ \bar{\boldsymbol{\xi}} \tanh\Big[\beta\sqrt{\gamma\alpha\bar{r}}\bar{z} + \beta\gamma(\bar{\boldsymbol{\xi}}\cdot\boldsymbol{\nu})\Big]\Big\rangle \tag{3.31}$$

$$\bar{\boldsymbol{\nu}} = \Big\langle \int Dz\ \boldsymbol{\xi} \tanh\Big[\beta\sqrt{\bar{\gamma}\alpha r}z + \beta\bar{\gamma}(\boldsymbol{\xi}\cdot\mathbf{m})\Big]\Big\rangle \tag{3.32}$$

$$q = \Big\langle \tanh^2\Big[\beta\sqrt{\bar{\gamma}\alpha r}z + \beta\bar{\gamma}(\boldsymbol{\xi}\cdot\mathbf{m})\Big]\Big\rangle \tag{3.33}$$

$$\bar{q} = \Big\langle \tanh^2\Big[\beta\sqrt{\gamma\alpha\bar{r}}\bar{z} + \beta\gamma(\bar{\boldsymbol{\xi}}\cdot\boldsymbol{\nu})\Big]\Big\rangle \tag{3.34}$$

$$r = \frac{\bar{q} + \beta^2 q(1-\bar{q})^2}{(1-\beta^2 q\bar{q})^2} \tag{3.35}$$

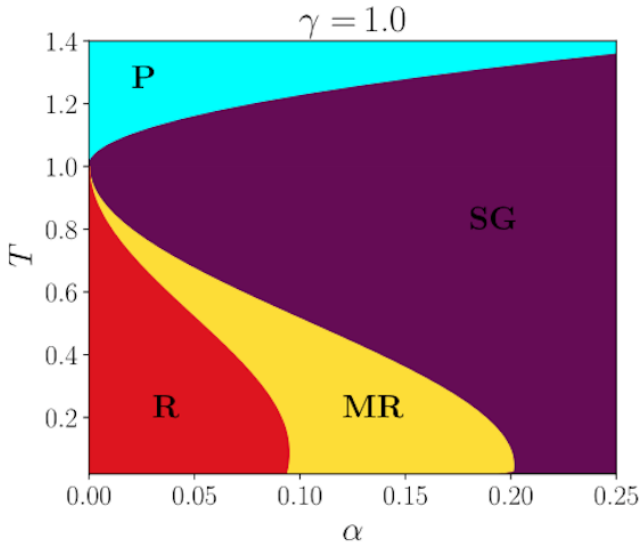


Figure 3.2. Image from [7], it shows the phase diagram of the BAM in the (α, T) plane when the two layer of the network have the same size and so present an asymmetry $\gamma = 1$. The labels $\mathbf{P}, \mathbf{SG}, \mathbf{MR}, \mathbf{R}$ stand for paramagnetic, spin glass, metastable retrieval and retrieval respectively.

$$\bar{r} = \frac{q + \beta^2 \bar{q}(1-q)^2}{(1-\beta^2 q\bar{q})^2} \tag{3.36}$$

The model phase diagram can be fully characterized by solving the above saddle point equations at any value of the three control parameters, α, β, γ and evaluating the corresponding free energy for each of the fixed points. In Figure 3.2 is reported the phase diagram that was numerically evaluated in [7] with considering the retrieval of only one pattern pair: this is equivalent to assume a solution of the form $m_\lambda = m_\mu \delta_{\lambda\mu}$ and $\nu_\lambda = \nu_\mu \delta_{\lambda\mu}$.
An important property to notice is that the above free energy B.20 (and consequently, the phase diagram) is symmetric under the exchange of the two layers, provided that all the order parameters are swapped and $\gamma \to \bar{\gamma}$.
The phase diagram of the BAM model highly resembles the phase diagram of the standard Hopfield Model: at high temperature, the equilibrium phase is paramagnetic. At $T < 1$, lowering the network load α, the model exhibits a Spin Glass phase. When $\alpha < \alpha_c(T)$ and if the temperature is sufficiently low the model exhibits a retrieval phase. A more detailed inspection shows how the retrieval phase can be divided in two sub-regions highlighted in Figure 3.2: a first (red) portion of the phase diagram where the retrieval fixed point has the lowest free energy, and another one (yellow) where the retrieval fixed point is a local stationary point.
In the asymmetric case $\gamma \neq 1$ the phase diagram looks qualitatively similar to the previous one, as showed in Figures 3.3, where are displayed respectively the result at $\gamma = 2$ and $\gamma = 5$. However, increasing the asymmetry between the layers [7] realized that retrieval loss occurs at lower values of α while the critical line

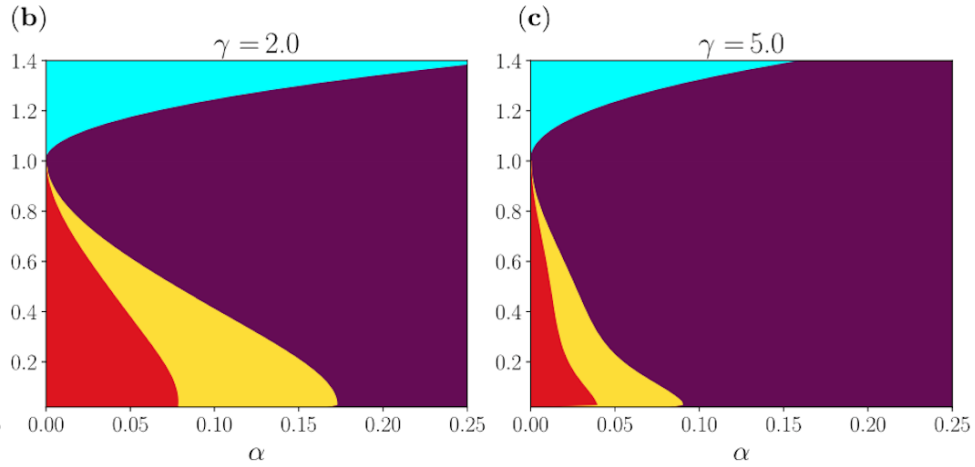


Figure 3.3. Image from [7], it shows the phase diagram of the BAM in the (α, T) plane when the two layer of the network have size one double of the other (b), and one fifth time the other (c), with asymmetries respectively of $\gamma = 2$ and $\gamma = 5$. As the asymmetry grows the images show how the spin glass phase, in violet, expands at the expense of the others.

separating the paramagnetic from the spin glass phases moves to higher temperatures. They conclude that increasing the asymmetry between the two layers extends the Spin Glass phase at the expense of the others.

3.2.3 Replica symmetric phase diagram at $T = 0$

The phase diagram in the noiseless regime can be computed by taking the $T = 0$ limit of the free energy B.20 and the corresponding saddle point equations 3.31-3.36. In this limit the order parameters scale as:

$$\delta q = \beta(1 - q), \tag{3.37}$$
$$\delta \bar{q} = \beta(1 - \bar{q}). \tag{3.38}$$

We lose the equations for r and $\bar{r}$ since the equations for the magnetizations and for the spin overlaps do not depend anymore on r and $\bar{r}$ in the $T = 0$ limit and to distinguish the different phases in the phase diagram we need just the values of $m, \bar{m}, \delta q, \delta \bar{q}$.
Therefore the saddle point equations 3.31-3.36 became:

$$\delta q = \frac{2}{\bar{\gamma}\sqrt{\pi}} \frac{y}{\mathrm{erf}(\bar{y})} e^{-y^2} \tag{3.39}$$

$$\delta \bar{q} = \frac{2}{\gamma\sqrt{\pi}} \frac{\bar{y}}{\mathrm{erf}(y)} e^{-\bar{y}^2} \tag{3.40}$$

$$\frac{1 + \delta q^2}{(1 - \delta q \delta \bar{q})^2} = \frac{\mathrm{erf}^2(y)}{2\bar{\gamma}\alpha\bar{y}^2} \tag{3.41}$$

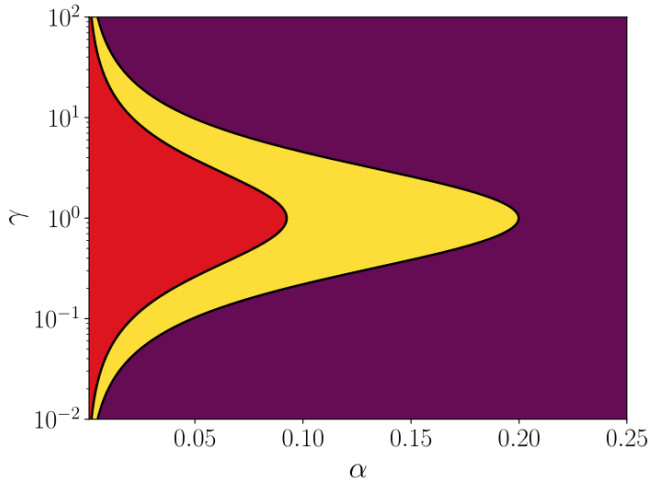


Figure 3.4. Image from [7], it shows the BAM's phase diagram at $T = 0$ in the plane (α, T): the same color code as in Figure 3.2 is used. The phase boundary separating the MR (yellow) from the SG (violet) phases defines the RS critical capacity α_c as a function of γ. As the vertical axis has a logarithmic scale, we notice again how the phase diagram is symmetric under the transformation $\gamma \to \gamma^{-1} = \bar{\gamma}$

$$\frac{1 + \delta\bar{q}^2}{(1 - \delta q \delta\bar{q})^2} = \frac{\mathrm{erf}^2(\bar{y})}{2\gamma\alpha y^2} \tag{3.42}$$

where $\mathrm{erf}(\cdot)$ denotes the error function and the new variable y and $\bar{y}$ are linked to the magnetizations:

$$\mathrm{erf}(y) = m \tag{3.43}$$
$$\mathrm{erf}(\bar{y}) = \bar{m} \tag{3.44}$$

The results of the above system 3.39 - 3.42 have been solved in [7] at different values of the control parameters α and γ and the results are displayed in Figure 3.4. The yellow retrieval phase and the violet spin glass phase are separated by the $\alpha_c(\gamma)$ line which has its maximum at $\gamma = 1$, where it has a numerical value of $\alpha_c(\gamma = 1) \simeq 0.2$, consistently with [11]. Another interesting result is that in the extremely asymmetric limit, $\gamma \to 0$, the storage capacity $\alpha_c(\gamma)$ goes to 0 linearly with γ and the rescaled capacity $\tilde{\alpha}_c = \frac{\alpha_c}{\gamma}$ tends to a finite value $\tilde{\alpha}_c \to 0.497$. By symmetry, the same result is achieved also when $\gamma \to \infty$ and $\tilde{\alpha}_c = \alpha_c\gamma$.
Finally in [7] they quantify the difference between the BAM model and the Hopfiel model in terms of their storage capabilities: first they renormalize the critical capacities of both models by the total number of neurons $N + \bar{N}$, which is kept fixed whilst varying the asymmetry between the two layers. In this notation, while the Hopfield's critical capacity is indipendent of γ and remains equal to $\alpha_c^{Hop} = \frac{P_c}{N+\bar{N}} \simeq 0.138$, on the other hand the BAM's critical capacity is given by

$$\tilde{\alpha}_c^{BAM} = \frac{P_c}{N + \bar{N}} = \frac{\alpha_c(\gamma}{\gamma + \bar{\gamma}} \tag{3.45}$$

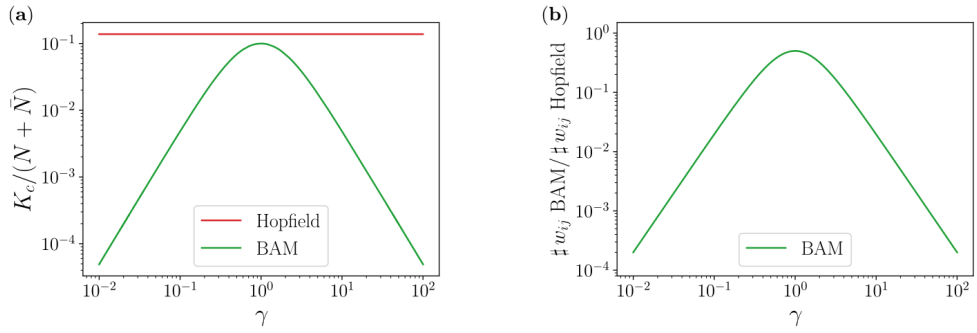


Figure 3.5. Images from [7], (a) Critical capacity of the BAM and the Hopfield model normalized by the total number of neurons in the network (therefore $N + \bar{N}$ instead of L for the BAM). The green line corresponds to equation 3.45. (b) Number of weights used to store this critical number of patterns in the BAM normalized by the analogous number of weights in the Hopfield model. In both images, each quantity is plotted as a function of the asymmetry γ.

The discrepancy between this two values with respect to γ is shown in Figure 3.5 in panel (a). As expected the BAM's critical capacity is lower than the Hopfield model even in its most stable configuration, but, as shown by [7] in Figure 3.5, panel (b), the BAM is more efficient from the point of view of the number of weights stored in the network. In fact, for a fixed number of patterns, the BAM requires at most about half of the weights compared to the Hopfield model to make the network operate in its retrieval phase.

Part II

New results

Chapter 4

Random Feature BAM

In the previous chapters we summarized the state of the art from which our original work begins.
In chapter 2 we summarized the results obtained by [31] where they showed that, by including a generative model for the patterns planted in the Hopfield model, the phase diagram of the model changes completely and it appear a transition from a storage phase to a learning phase, in the sense that attractors appear corresponding to the features in the data. This behaviour opened up a new paradigm for the model.
In chapter 3 we summarized the state of the art for the BAM model showing the results obtained in [7] where they demonstrated how, with a finite asymmetry between the two layers, the BAM can store information more efficiently than the Hopfield model by requiring less parameters to encode a fixed number of patterns.
In this chapter we combined the results and the ideas of the previous two to study, using the replica method from the statistical physics of disordered systems, what happens to the BAM when random feature data are incorporated.

4.1 Definition of the Model

Just like the BAM model described in Chapter 3 the model we introduce consists of 2 spin layers, $\mathbf{S}$ and $\bar{\mathbf{S}}$, composed of N and $\bar{N}$ components (neurons), respectively S_i with $i = 1, \dots, N$ and $\bar{S}_k$ with $k = 1, \dots, \bar{N}$. The free-energy of the system is described by the Hamiltonian:

$$H = -\sum_{i=1}^{N}\sum_{k=1}^{\bar{N}} W_{ik} S_i \bar{S}_k \ , \tag{4.1}$$

where the pair-wise interaction between the two layers is carried by the coupling W_{ik}, which represents the link between neuron i in the first layer and neuron k in the second, and there are no connections between neurons in the same layer. The coupling matrix $\mathbf{W}$ is defined through a set of p couple of patterns $(\boldsymbol{\xi}^\mu, \bar{\boldsymbol{\xi}}^\mu)$ via the Hebbian rule

$$W_{ik} = \frac{1}{\sqrt{N\bar{N}}} \sum_{\mu=1}^{p} \xi_i^\mu \bar{\xi}_k^\mu \ , \tag{4.2}$$

In Chapter 3 the results obtained hold as long as the memories $(\boldsymbol{\xi}^\mu, \bar{\boldsymbol{\xi}}^\mu)$ are independent identically distributed random variables with zero mean and unit variance. We test our approach by studying a generalization of the BAM model in which the patterns are no longer independent random variables. We shall study the case where the patterns have a correlation, created from the following structure:

$$\xi_i^\mu = \frac{1}{\sqrt{D}} \sum_{k=1}^{D} c_k^\mu f_{ki} \ . \tag{4.3}$$

The specific case we consider through this work is the one of independent identically distributed binary features $f_{ki} = \pm 1$ and independent identically distributed standard Gaussian coefficients c_k^μ. The type of disorder generated by 4.3 has been called in literature combinatorial disorder [29]. Moreover this data structure is also deeply related to the Hidden Manifold model [15, 13, 28].
The task of the model is to retrieve p pairs of memories $(\boldsymbol{\xi}^\mu, \mathbf{c}^\mu)$ for the two layers. We indeed chose this distinction between the patterns implanted in the two layers to emulate the work of visible and hidden layers in RBMs. When the visible layer $\mathbf{S}$ recovers a given memory $\boldsymbol{\xi}^\mu$ with $\mu = 1, \dots, p$, the hidden layer $\bar{\mathbf{S}}$ recovers the weights coefficients of the features that composed that pattern. This means that we built the model such as one specific feature is represented by exactly one neuron in the hidden layer: practically this means that $\bar{N} = D$, so from now on we will lose the $\bar{N}$ and we will refer as the number of neurons in the hidden layer as D.
With these new premises we can rewrite the Hamiltonian of the Random feature BAM we are studying as:

$$H = -\frac{1}{\sqrt{ND}} \sum_{i=1}^{N} \sum_{k=1}^{D} S_i \bar{S}_k \sum_{\mu=1}^{p} \xi_i^\mu c_k^\mu \ , \tag{4.4}$$

where the $\boldsymbol{\xi}^\mu$ are structured as shown in 4.3.
Since the $\mathbf{c}^\mu$ are generated according to a gaussian distribution we have to consider each neuron in the hidden layer as a continuous variable $\bar{S}_k$, for $k = 1, \dots, D$. For this reason we will consider a spherical constrain:

$$\sum_{k=1}^{D} \bar{S}_k \bar{S}_k = D \ . \tag{4.5}$$

Instead, each neuron in the visible layer $\mathbf{S}$ is a binary variable, $S_i = \{-1, +1\}$, $\forall\ i = 1, \dots, N$.

4.2 Replica symmetric computation

Since we are interested in the thermodynamic limit $\sqrt{ND} = L \to \infty$, we choose a regime where p, D and N are all proportional between them. At the same time, we keep the following ratios fixed

$$\alpha = \frac{p}{N} \ , \tag{4.6}$$

$$\alpha_D = \frac{D}{N} \ . \tag{4.7}$$

These, combined with the thermal noise β, will be the control parameters for our model. They are related via the relation $\alpha = \alpha_D \alpha_T$, where $\alpha_T = \frac{p}{D}$. We need now to compute the quenched averaged free energy of the model:

$$F_Q = -\lim_{n\to 0} \frac{1}{\beta L} \langle\langle \ln Z \rangle\rangle \tag{4.8}$$

where now, differentley from 3.12, $\langle\langle\cdot\rangle\rangle$ means that we have two sources of disorder and that we have to average on both the coefficients c and the features f, while Z is still the partition function.
In order to compute the average of $\ln Z$ in equation 4.8 we use the replica method [30], that consists in writing the average of the logarithm as

$$\langle\langle \ln Z \rangle\rangle = \lim_{n\to 0} \frac{\langle\langle Z^n \rangle\rangle - 1}{n} \tag{4.9}$$

The replicated partition function averaged over the disorder reads

$$\begin{aligned} \langle\langle\ Z^n \rangle\rangle\ = \Bigg\langle \Bigg\langle \mathrm{Tr}_S \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a d\bar{S}_k^a \exp\Big\{\beta \sum_{\mu a} m_\mu^a \nu_\mu^a\Big\} \delta\Big(D - \sum_k \bar{S}_k^a \bar{S}_k^a\Big) \\ \delta\Big(m_\mu^a - \frac{1}{\sqrt{N}} \sum_i \xi_i^\mu S_i^a\Big) \delta\Big(\nu_\mu^a - \frac{1}{\sqrt{D}} \sum_k c_k^\mu \bar{S}_k^a\Big) \Big\rangle_f \Bigg\rangle_c \end{aligned} \tag{4.10}$$

where we are referring to the sum over the replicated spin configurations for the visible layer of spin with the notation $\mathrm{Tr}_S = \sum_{\mathbf{S}^1=\{\pm 1\}^N} \cdots \sum_{\mathbf{S}^n=\{\pm 1\}^N}$. The hidden layer is integrated, instead of being summed over, due to the continuous nature of the variables $\bar{S}_k^a$. Finally in 4.10 we also introduced the two following auxillary variables:

$$m_\mu^a = \frac{1}{\sqrt{N}} \sum_i \xi_i^\mu S_i^a, \quad a \in [n]\ , \mu \in [p]\ , \tag{4.11}$$

$$\nu_\mu^a = \frac{1}{\sqrt{D}} \sum_k c_k^\mu \bar{S}_k^a, \quad a \in [n]\ , \mu \in [p]\ . \tag{4.12}$$

These are the pattern magnetizations for this model.
We now want also to introduce the feature magnetizations as it was done in [31] for the Random-feature Hopfield model. By using the Fourier transform of the $\delta(\cdot)$ function the partition function now reads:

$$\begin{aligned} \langle\langle\ Z^n \rangle\rangle\ = \Bigg\langle \Bigg\langle \mathrm{Tr}_S \int \prod_{\mu a k} d\bar{S}_k^a dm_\mu^a d\nu_\mu^a d\hat{m}_\mu^a d\hat{\nu}_\mu^a d\mu_k^a d\hat{\mu}_k^a dg_a \\ \exp\Big\{\sum_{\mu a} (\beta(m_\mu^a \nu_\mu^a) - \beta(\hat{m}_\mu^a m_\mu^a) - \beta(\hat{\nu}_\mu^a \nu_\mu^a))\Big\} \exp\Big\{-\sum_a \frac{g_a}{2} (D - \sum_k \bar{S}_k^a \bar{S}_k^a)\Big\} \\ \exp\Big\{-\sum_{ka} \hat{\mu}_k^a \mu_k^a + \frac{1}{\sqrt{N}} \sum_{ka} \hat{\mu}_k^a \sum_i f_{ki} S_i^a\Big\} \\ \exp\Big\{\frac{\beta}{\sqrt{D}} \sum_{\mu a} \hat{m}_\mu^a \sum_k c_k^\mu \mu_k^a + \frac{\beta}{\sqrt{D}} \sum_{\mu a} \hat{\nu}_\mu^a \sum_k c_k^\mu \bar{S}_k^a\Big\} \Big\rangle_f \Bigg\rangle_c\ , \end{aligned} \tag{4.13}$$

where the parameter g_a is a Lagrange multiplier introduced to keep track of the spherical constraint 4.5 of the hidden layer for each of the n replicas. In 4.13 we also have introduced the set of order parameters called feature magnetizations:

$$\mu_k^a = \frac{1}{\sqrt{N}} \sum_i f_{ki} S_i^a, \quad a \in [n] \ , k \in [D] \ . \tag{4.14}$$

We want to see if, by tuning the control parameters α and α_D we can find a solution for $\mu_k > 0$ for some k.
Similarly to [3] and [31], we make an ansatz on the structure of the solution for this set of order parameter. Since we are interested for a solution for any k, for simplicity we study the case where the model retrieves only one of the features.

4.2.1 feature retrieval

In order to analyse the retrieval of one feature only we impose the ansatz where all the value for the order parameters condensate in just one of the k: we say that the variables with $k = 1$ are of $O(1)$ while the remaining variables scales as $O(\frac{1}{\sqrt{N}})$ or $O(\frac{1}{\sqrt{D}})$. This means that, in the termodynamic limit, we are looking for a solution as

$$\boldsymbol{\mu} = (\mu, 0, \dots, 0) \ . \tag{4.15}$$

With this ansatz we separate the contribution $k = 1$ from the rest with $k > 1$ and we rescale properly the order parameters:

$$\mu_k^a = \frac{1}{\sqrt{N}} \sum_i f_{ki} S_i^a, \quad a \in [n] \ , k \in [D] \ , \tag{4.16}$$

$$\mu_1^a = \frac{1}{N} \sum_i f_{1i} S_i^a \ . \tag{4.17}$$

Moreover, given that the model needs to retrieve a pattern pair, we also need to rescale the spin value in the hidden layer because with our ansatz we are saying that the main contribution from the features come only from the one with $k = 1$. So we rescale the value of the $k = 1$ hidden neuron as $\bar{S}_1^a \to \sqrt{D}\bar{S}_1^a$ so that now this is of the same size by respect to the remaining sum of $k > 1$.
Performing then the average over the coefficient matrix and separating the $k = 1$ from the $k > 1$ terms, the partition function now reads:

$$\begin{aligned}
\langle\langle \ Z^n \rangle\rangle \ = \ \Big\langle \mathrm{Tr}_S \int \prod_{\mu a k} & d\bar{S}_k^a dT_\mu^a d\mu_k^a d\hat{\mu}_k^a dg_a \\
& \exp\Big\{ -\sum_a \frac{g_a}{2} (D - \sum_{k>1} \bar{S}_k^a \bar{S}_k^a - D\bar{S}_1^a \bar{S}_1^a) \Big\} \\
& \exp\Big\{ \Big(\sum_\mu \Big(-\frac{1}{2} \sum_{ab} T_\mu^a J_{ab} T_\mu^b \Big) \Big\} \\
& \exp\Big\{ -\sum_a \sum_{k>1} \hat{\mu}_k^a \mu_k^a + \frac{1}{\sqrt{N}} \sum_a \sum_{k>1} \hat{\mu}_k^a \sum_i f_{ki} S_i^a \Big\} \\
& \exp\Big\{ -\sum_a \hat{\mu}_1^a \mu_1^a + \frac{1}{N} \sum_a \hat{\mu}_1^a \sum_i f_{1i} S_i^a \Big\} \Big\rangle_f
\end{aligned} \tag{4.18}$$

Where now the remaining mean $\langle\cdot\rangle$ is only over the features f_{ki}.
In 4.18 we introduced a new variable $T_\mu^a = (\sqrt{\beta}\hat{m}_\mu^a, \sqrt{\beta}\hat{\nu}_\mu^a)$, so that we could perform the gaussian integral over the T_μ^a.
The matrix J_{ab} is a block matrix with the following form:

$$J = -\begin{pmatrix} \frac{\beta}{D}\sum_{k>1}\mu_k^a\mu_k^b + \beta\frac{N}{D}\mu_1^a\mu_1^b & \frac{\beta}{D}\sum_{k>1}\mu_k^a\bar{S}_k^b + \beta\sqrt{\frac{N}{D}}\mu_1^a\bar{S}_1^b - \mathbb{1} \\ \\ \frac{\beta}{D}\sum_{k>1}\mu_k^b\bar{S}_k^a + \beta\sqrt{\frac{N}{D}}\mu_1^b\bar{S}_1^a - \mathbb{1} & \frac{\beta}{D}\sum_{k>1}\bar{S}_k^a\bar{S}_k^b + \beta\bar{S}_1^a\bar{S}_1^b \end{pmatrix} \tag{4.19}$$

Where with the notation $\mathbb{1}$ we define the identity matrix in the n dimensional space.

$$\mathbb{1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}_{n\times n} \tag{4.20}$$

The computation technique used to solve this kind of gaussian integrals by defining a couple variable and a block matrix has already been used, for example in [23] for the mixed p-spin spherical model. After a series of standard manipulations whose details are reported in Appendix C, we show the final result for the partition function before computing the saddle point equations:

$$\begin{aligned} \langle\langle\, Z^n \rangle\rangle \;&= \mathrm{Tr}_S \int \prod_{a,b=1}^{n} dq_{ab} d\hat{q}_{ab} d\mu_1^a d\hat{\mu}_1^a d\bar{S}_1^a dg_a \prod_{c,d=1}^{2n} dA_{cd} \\ &\exp\left\{-\frac{p}{2}\ln\det(J)\right\}\exp\left\{-\frac{D}{2}\ln\det(q)\right\}\exp\left\{\frac{D}{2}\ln\det A\right\} \\ &\exp\left\{-\sum_a \frac{g_a}{2}(D - D\bar{S}_1^a\bar{S}_1^a)\right\}\exp\left\{-\frac{N\alpha}{2}\sum_{ab}\hat{q}_{ab}q_{ab} + \frac{\alpha}{2}\sum_{ab}\hat{q}_{ab}\sum_i S_i^a S_i^b\right\} \\ &\exp\left\{nD - \frac{D}{2}Tr\left[A^T\begin{pmatrix} q^{-1} & 0 \\ 0 & -g\mathbb{1}\end{pmatrix}\right]\right\} \\ &\left\langle \exp\left\{-N\sum_a \hat{\mu}_1^a\mu_1^a + \sum_a\hat{\mu}_1^a\sum_i f_{1i}S_i^a\right\}\right\rangle_{f_{1i}} \end{aligned} \tag{4.21}$$

In this equation we have defined the following quantities

$$q_{ab} = \frac{1}{N}\sum_i S_i^a S_i^b, \quad a\ ,\ b \in [n] \tag{4.22}$$

which is the order parameter that represents the overlap between the spins in the visible layer.
For convenience we also defined:

$$A = \begin{pmatrix} \frac{\beta}{D}\sum_{k>1}\mu_k^a\mu_k^b & \frac{\beta}{D}\sum_{k>1}\mu_k^a\bar{S}_k^b \\ \\ \frac{\beta}{D}\sum_{k>1}\mu_k^b\bar{S}_k^a & \frac{\beta}{D}\sum_{k>1}\bar{S}_k^a\bar{S}_k^b \end{pmatrix} = \begin{pmatrix} P_{ab} & R_{ab} \\ \\ (R_{ab})^T & \bar{Q}_{ab} \end{pmatrix} \tag{4.23}$$

which is the block matrix of the order parameters P_{ab}, overlap between the feature magnetizations, $\bar{Q}_{ab}$ overlap between the spins in the hidden layer and R_{ab}, overlap between the spins in the hidden layer and the feature magnetizations, all for $k > 1$. While the order parameters P_{ab} appeared also in [31] and $\bar{Q}_{ab}$ appeared also in [7], the order parameter R_{ab} is new and it is typical of this model where the correleted structure of the pattern is combined with the presence of a second layer of neurons.

4.2.2 Saddle point equation for the block matrixes of A

From here we are going to compute the saddle point equation for the matrix A. We are going to proceed in the same as we did for the matrix $\hat{A}$ in Appendix C.3. Reminder that the J matrix is composed as follows:

$$J = -\beta A - \begin{pmatrix} \beta\frac{N}{D}\mu_1^a\mu_1^b & \beta\sqrt{\frac{N}{D}}\mu_1^a\bar{S}_1^b - \mathbb{1} \\ \beta\sqrt{\frac{N}{D}}\mu_1^b\bar{S}_1^a - \mathbb{1} & \beta\bar{S}_1^a\bar{S}_1^b \end{pmatrix} = -\beta A - \mathbb{A}_1 + \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix} \tag{4.24}$$

Where the matrix A was defined in 4.33 and where the matrix $\mathbb{A}_1$ is defined as.

$$\mathbb{A}_1 = \begin{pmatrix} \beta\frac{N}{D}\mathbb{P}_1 & \beta\sqrt{\frac{N}{D}}\mathbb{R}_1 \\ \beta\sqrt{\frac{N}{D}}(\mathbb{R}_1)^T & \beta\bar{\mathbb{Q}}_1 \end{pmatrix} \tag{4.25}$$

Let's also notice that for J, being a $2n \times 2n$ matrix the following equivalence holds

$$\ln\det\left(-\beta A - \mathbb{A}_1 + \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}\right) = \ln\det\left(\beta A + \mathbb{A}_1 - \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}\right) \tag{4.26}$$

By differentiating in $\partial_A(\langle\langle Z^n\rangle\rangle)$, we obtain the following equation:

$$\begin{aligned} \partial_A(\langle\langle Z^n\rangle\rangle) = &-\frac{\beta p}{2}\left(\beta A + \mathbb{A}_1 - \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}\right)^{-1} + \frac{D}{2}A^{-1} - \frac{D}{2}\begin{pmatrix} (q^T)^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} = 0 \\ &\implies \\ &-\beta pA + D\left(\beta A + \mathbb{A}_1 - \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}\right) + \\ &-DA\begin{pmatrix} (q^T)^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix}\left(\beta A + \mathbb{A}_1 - \begin{pmatrix} 0 & \mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix}\right) = 0 \end{aligned} \tag{4.27}$$

Let's now compute the matrix product. We are going to lose the transpose for all the matrixes as in the RS ansatz we are about to make all matrixes are symmetric. From all of this by computing the matrix product of equation 4.27 we get four

matrix equations for the blocks of the matrix A.

$$\begin{aligned}&-\alpha_D\beta\mathbb{P}\cdot\mathbb{P}+\alpha_D\beta\mathbb{P}\cdot\mathbb{q}-\alpha_D\alpha_T\beta\mathbb{P}\cdot\mathbb{q}-\beta\mathbb{P}_1\cdot\mathbb{P}-g\alpha_D\mathbb{R}\cdot\mathbb{q}+\\&\beta\mathbb{P}_1\cdot\mathbb{q}+g\alpha_D\beta\mathbb{R}\cdot\mathbb{q}\cdot\mathbb{R}+g\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\mathbb{R}\cdot\mathbb{q}=0,\end{aligned}\tag{4.28}$$

$$\begin{aligned}&\mathbb{P}\alpha_D-\mathbb{q}\alpha_D-\alpha_D\beta\mathbb{P}\cdot\mathbb{R}-\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\mathbb{P}+\alpha_D\beta\mathbb{R}\cdot\mathbb{q}-\alpha_D\alpha_T\beta\mathbb{R}\cdot\mathbb{q}+\\&\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\mathbb{q}+g\alpha_D\beta\mathbb{R}\cdot\mathbb{q}\cdot\bar{\mathbb{Q}}+g\alpha_D\beta\bar{\mathbb{Q}}_1\cdot\mathbb{R}\cdot\mathbb{q}=0,\end{aligned}\tag{4.29}$$

$$\begin{aligned}&-\mathbb{q}\alpha_D-g\alpha_D\bar{\mathbb{Q}}\cdot\mathbb{q}-\alpha_D\beta\mathbb{R}\cdot\mathbb{P}+\alpha_D\beta\mathbb{R}\cdot\mathbb{q}-\alpha_D\alpha_T\beta\mathbb{R}\cdot\mathbb{q}-\beta\mathbb{P}_1\cdot\mathbb{R}+\\&\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\mathbb{q}+g\alpha_D\beta\bar{\mathbb{Q}}\cdot\mathbb{q}\cdot\mathbb{R}+g\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\bar{\mathbb{Q}}\cdot\mathbb{q}=0,\end{aligned}\tag{4.30}$$

$$\begin{aligned}&\mathbb{R}\alpha_D+\alpha_D\beta\bar{\mathbb{Q}}\cdot\mathbb{q}-\alpha_D\alpha_T\beta\bar{\mathbb{Q}}\cdot\mathbb{q}-\alpha_D\beta\mathbb{R}\cdot\mathbb{R}-\sqrt{\alpha_D}\beta\mathbb{R}_1\cdot\mathbb{R}+\\&\alpha_D\beta\bar{\mathbb{Q}}_1\cdot\mathbb{q}+g\alpha_D\beta\bar{\mathbb{Q}}\cdot\mathbb{q}\cdot\bar{\mathbb{Q}}+g\alpha_D\beta\bar{\mathbb{Q}}_1\cdot\bar{\mathbb{Q}}\cdot\mathbb{q}=0.\end{aligned}\tag{4.31}$$

Where we used the $\mathbb{P}$ notation to distinguish matrixes from constants, we multiplied every equation with the $\mathbb{q}=q$ matrix to avoid the inverse and the matrix $\mathbb{g}$ is a matrix of the following form:

$$\mathbb{g}=\begin{pmatrix}g_1&0&\dots&0\\0&\ddots&\ddots&\vdots\\\vdots&\ddots&\ddots&0\\0&\dots&0&g_n\end{pmatrix}_{n\times n}\tag{4.32}$$

4.2.3 RS ansatz

We now make an RS ansatz for all the order parameters. We Remind that A is not an order parameter that will receive the RS ansatz. It is indeed a block matrix, where the single blocks are order parameters:

$$A=\begin{pmatrix}P&R\\R^T&\bar{Q}\end{pmatrix}\tag{4.33}$$

So the RS ansatz for all the order parameters reads as follows:

$$q_{ab}=\delta_{ab}+q(1-\delta_{ab})\tag{4.34}$$
$$\hat{q}_{ab}=\delta_{ab}+\hat{q}(1-\delta_{ab})\tag{4.35}$$
$$\bar{Q}_{ab}=\bar{Q}_{\mathrm{d}}\delta_{ab}+\bar{Q}(1-\delta_{ab})\tag{4.36}$$
$$P_{ab}=P_{\mathrm{d}}\delta_{ab}+P(1-\delta_{ab})\tag{4.37}$$
$$R_{ab}=R_{\mathrm{d}}\delta_{ab}+R(1-\delta_{ab})\tag{4.38}$$
$$\mu_k^a=\mu_k\tag{4.39}$$
$$\hat{\mu}_k^a=\hat{\mu}_k\tag{4.40}$$
$$\bar{S}_1^a=\bar{S}_1\tag{4.41}$$
$$g_a=g\tag{4.42}$$

The RS ansatz allows us to linearise the term $S^a S^b$ by computing the Tr_S. We can also compute explicitly the log of the determinants of equation 4.21.

$$\ln\det(J) \tag{4.43}$$
$$\ln\det(A) \tag{4.44}$$
$$\ln\det(q) \tag{4.45}$$

We remind that q is the $n \times n$ matrix defined in 4.22, while A and J are both $2n \times 2n$ block matrices defined respectively above in 4.23 and 4.24. Finally the RS ansatz allow us to compute also the $Tr\left[A^T \begin{pmatrix} q^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix}\right]$. All this computation are explicitly displayed in Appendix C.
The order parameters involved in the RS free energy are then:

$$\begin{aligned} -\beta F_Q(q,\hat{q},\bar{q}_d,\bar{q},\bar{S}_1,&P_d,P,R_d,R,\mu_1,\hat{\mu}_1,g) = \\ &\lim_{n\to 0}\lim_{L\to\infty}\frac{1}{nL}\ln\left(\langle\langle Z^n\rangle\rangle\right) \end{aligned} \tag{4.46}$$

And the final form for the RS free energy is:

$$\begin{aligned} f^{RS} =& \frac{\alpha_T\sqrt{\alpha_D}}{2\beta}\Bigg[\ln\left(\beta^2(P_d-P)(\bar{q}_d-\bar{q})-(\beta R_d-1-\beta R)^2\right)+ \\ &\frac{\beta^2(\bar{q}+\bar{S}_1^2)(P_d-P)+\frac{\beta^2}{\alpha_D}(\alpha_D P+\mu_1^2)(\bar{q}_d-\bar{q})-\frac{2\beta}{\sqrt{\alpha_D}}(\sqrt{\alpha_D}R+\mu_1\bar{S}_1)(\beta R_d-1-\beta R)}{\beta^2(P_d-P)(\bar{q}_d-\bar{q})-(\beta R_d-1-\beta R)^2}\Bigg]+ \\ &\frac{\sqrt{\alpha_D}}{2\beta}\left[\ln\left(\frac{1-q}{(P_d-P)(\bar{q}_d-\bar{q})-(R_d-R)^2}\right)\right]+ \\ &+\frac{\sqrt{\alpha_D}}{2\beta}\left[\frac{q}{1-q}+\frac{\bar{q}(P_d-P)+P(\bar{q}_d-\bar{q})-2R(R_d-R)}{(P_d-P)(\bar{q}_d-\bar{q})-(R_d-R)^2}-1\right]+ \\ &+\frac{\sqrt{\alpha_D}}{2\beta}\left(\frac{P_d-2qP_d+Pq}{(1-q)(1-q)}\right)+ \\ &+\frac{1}{\sqrt{\alpha_D}}\hat{\mu}_1\mu_1+\frac{g\sqrt{\alpha_D}}{2}(1-\bar{q}_d-\bar{S}_1\bar{S}_1)-\frac{\alpha\beta}{2\sqrt{\alpha_D}}\hat{q}(q+1)+ \\ &-\frac{1}{\beta\sqrt{\alpha_D}}\left\langle\int Dz\ln\left[2\cosh(\beta(z\sqrt{\alpha\hat{q}}+\hat{\mu}_1 f_{1i}))\right]\right\rangle_{f_{1i}} \end{aligned} \tag{4.47}$$

Where with the notation $\int Dz$ we imply the Gaussian integral over the variable z:

$$\int Dz = \int dz \exp\left\{-\frac{z^2}{2}\right\} \tag{4.48}$$

In 4.47 we recognize the control parameters: $\alpha = \frac{p}{N}$, $\alpha_D = \frac{D}{N}$ and $\alpha_T = \frac{p}{D} = \frac{\alpha}{\alpha_D}$. It is usefull now to make a discussion on our choice for the definition of the control parameter of the model by comparing them to those analysed in the BAM 3.10, 3.11 and the one analysed in the Random Feature Hopfield model 2.14. We can see in fact that we chose the network load α to represent how the p memories scale with

respect to the N neurons in the visible layer, while the information on how the memories scale with respect to the D neurons in the hidden layer is contained in α_T. The control parameter α_D is more trickier: it represents both the asymmetry control parameter 3.11 we exploited in the BAM to define the comparison between the two layers of the model, and the information on how the D features scale with respect to the N neurons in the visible layer.

4.2.4 Saddle point equations

We take the derivatives of equation 4.47 with respect to the order parameters $\bar{S}_1$, g, $\hat{\mu}$, μ, $\hat{q}$ and q getting the following saddle point equations:

$$\bar{S}_1\bar{S}_1 = 1 - \bar{q}_d \tag{4.49}$$

$$q = \int Dz \tanh^2\left(\beta(z\sqrt{\alpha\hat{q}} + \hat{\mu})\right) \tag{4.50}$$

$$\mu = \int Dz \tanh\left(\beta(z\sqrt{\alpha\hat{q}} + \hat{\mu})\right) \tag{4.51}$$

$$g = \frac{\alpha_T}{\sqrt{\alpha_D}} \frac{\mu_1 - \beta\mu_1(R_d - R) + \sqrt{\alpha_D}\beta(P_d - P)\bar{S}_1}{(\beta^2(P_d - P)(\bar{q}_d - \bar{q}) - (1 + \beta(R_d - R))^2)\bar{S}_1} \tag{4.52}$$

$$\hat{\mu}_1 = \alpha_T \frac{\beta\mu_1(\bar{q}_d - \bar{q}) + \sqrt{\alpha_D}\bar{S}_1(1 - \beta(R_d - R))}{(1 - \beta(R_d - R))^2 - \beta^2(P_d - P)(\bar{q}_d - \bar{q})} \tag{4.53}$$

$$\hat{q} = \frac{\alpha_D}{\alpha} \frac{(1 - q)(P_d + q) - (1 + q)(P_d - P)}{\beta^2(1 - q)^3} \tag{4.54}$$

Then, we add to these the equations 4.28-4.31, that we got when we differentiated the partition function with respect to A.
By doing the RS ansatz to the matrixes in equations 4.28-4.31 we get a system of 8 equations for 6 order parameters, namely the diagonal and non-diagonal elements of the matrixes P, R, $\bar{Q}$:

$$\begin{aligned} &- g(R - 2R + qR_d)\alpha_D + 2P(P - P_d)\alpha_D\beta + (P - 2Pq + P_dq)\alpha_D\beta + \\ &g(R - R_d)((-2 + 3q)R - qR_d)\alpha_D\beta - (P - 2Pq + P_dq)\alpha_D\alpha_T\beta + \\ &g(-1 + q)(R - R_d)S_1\sqrt{\alpha_D}\beta\mu_1 + (P - P_d)\beta\mu_1^2 - (-1 + q)\beta\mu_1^2 = 0, \end{aligned} \tag{4.55}$$

$$\begin{aligned} &g(qR - R_d)\alpha_D + (P^2 - P_d^2)\alpha_D\beta + (P_d - Pq)\alpha_D\beta + \\ &g(R - R_d)((-1 + 2q)R - R_d)\alpha_D\beta - (P_d - Pq)\alpha_D\alpha_T\beta + \\ &g(-1 + q)(R - R_d)S_1\sqrt{\alpha_D}\beta\mu_1 + (P - P_d)\beta\mu_1^2 - (-1 + q)\beta\mu_1^2 = 0, \end{aligned} \tag{4.56}$$

$$\begin{aligned} &P\alpha_D - q\alpha_D + (R - 2qR + qR_d)\alpha_D\beta - (P_dR + P(-2R + R_d))\alpha_D\beta + \\ &g(\bar{q}((-2 + 3q)R + R_d - 2qR_d) + \bar{q}_d(R - 2qR + qR_d))\alpha_D\beta + \\ &g(-1 + q)(R - R_d)S_1^2\alpha_D\beta - (R - 2qR + qR_d)\alpha_D\alpha_T\beta + \\ &(P - P_d)S_1\sqrt{\alpha_D}\beta\mu_1 - (-1 + q)S_1\sqrt{\alpha_D}\beta\mu_1 = 0, \end{aligned} \tag{4.57}$$

$$
\begin{aligned}
&-\alpha_D + P_d\alpha_D + (-qR + R_d)\alpha_D\beta + (PR - P_dR_d)\alpha_D\beta + \\
&g(\bar{q}_d(-qR + R_d) - \bar{q}(R - 2qR + qR_d))\alpha_D\beta + \\
&g(-1+q)(R - R_d)S_1^2\alpha_D\beta + (qR - R_d)\alpha_D\alpha_T\beta + \\
&(P - P_d)S_1\sqrt{\alpha_D}\beta\mu_1 - (-1+q)S_1\sqrt{\alpha_D}\beta\mu_1 = 0,
\end{aligned} \tag{4.58}
$$

$$
\begin{aligned}
&-q\alpha_D - g(\bar{q} - 2q\bar{q} + q\bar{q}_d)\alpha_D + (R - 2qR + qR_d)\alpha_D\beta + \\
&-(P_dR + P(-2R + R_d))\alpha_D\beta + \\
&g(\bar{q}((-2+3q)R + R_d - 2qR_d) + \bar{q}_d(R - 2qR + qR_d))\alpha_D\beta + \\
&-(R - 2qR + qR_d)\alpha_D\alpha_T\beta - (-1+q)S_1\sqrt{\alpha_D}\beta\mu_1 + \\
&g(-1+q)(\bar{q} - \bar{q}_d)S_1\sqrt{\alpha_D}\beta\mu_1 + (R - R_d)\beta\mu_1^2 = 0,
\end{aligned} \tag{4.59}
$$

$$
\begin{aligned}
&\alpha_D(-1 + PR\beta - qR\beta + R_d\beta - P_dR_d\beta \\
&+ qR\alpha_T\beta - R_d\alpha_T\beta - g(\bar{q}_d + \bar{q}R\beta + q\bar{q}_dR\beta \\
&- \bar{q}_dR_d\beta + q\bar{q}(-1 - 2R\beta + R_d\beta))) \\
&+ (-1+q)(-1 + g(\bar{q} - \bar{q}_d))S_1\sqrt{\alpha_D}\beta\mu_1 + (R - R_d)\beta\mu_1^2 = 0,
\end{aligned} \tag{4.60}
$$

$$
\begin{aligned}
&g(\bar{q} - \bar{q}_d)((-2+3q)\bar{q} - q\bar{q}_d)\alpha_D\beta + 2R(R - R_d)\alpha_D\beta + \\
&(\bar{q} - 2q\bar{q} + q\bar{q}_d)\alpha_D\beta - (-1+q)S_1^2\alpha_D\beta + \\
&(R - R_d)S_1\sqrt{\alpha_D}\beta\mu_1 + g(-1+q)(\bar{q} - \bar{q}_d)S_1^2\alpha_D\beta + \\
&R\alpha_D - (\bar{q} - 2q\bar{q} + q\bar{q}_d)\alpha_D\alpha_T\beta = 0,
\end{aligned} \tag{4.61}
$$

$$
\begin{aligned}
&R_d\alpha_D + g(\bar{q} - \bar{q}_d)((-1+2q)\bar{q} - \bar{q}_d)\alpha_D\beta + (-q\bar{q} + \bar{q}_d)\alpha_D\beta + \\
&(R^2 - R_d^2)\alpha_D\beta - (-1+q)S_1^2\alpha_D\beta + g(-1+q)(\bar{q} - \bar{q}_d)S_1^2\alpha_D\beta + \\
&(q\bar{q} - \bar{q}_d)\alpha_D\alpha_T\beta + (R - R_d)S_1\sqrt{\alpha_D}\beta\mu_1 = 0.
\end{aligned} \tag{4.62}
$$

4.2.5 Limit $\beta \to \infty$

We will now focus on the $\beta \to \infty$ limit, where the order parameters scale as:

$$q = 1 - \frac{\delta q}{\beta} \tag{4.63}$$

$$R = R_d - \frac{\delta R}{\beta} \tag{4.64}$$

$$P = P_d - \frac{\delta P}{\beta} \tag{4.65}$$

$$\bar{q} = \bar{q}_d - \frac{\delta \bar{q}}{\beta} \tag{4.66}$$

So the saddle point equations became

$$\delta q = \frac{2}{\sqrt{\alpha\hat{q}}} G\left(-\frac{\hat{\mu}}{\sqrt{\alpha\hat{q}}}\right) \tag{4.67}$$

$$\mu = 2H\left(-\frac{\hat{\mu}}{\sqrt{\alpha\hat{q}}}\right) - 1 \tag{4.68}$$

$$g = \frac{\alpha_T}{\sqrt{\alpha_D}} \frac{\mu - \mu\delta R + \sqrt{\alpha_D}\delta P \bar{S}_1}{(\delta P \delta\bar{q} - (1+\delta R)^2)\bar{S}_1} \tag{4.69}$$

$$\hat{\mu} = \alpha_T \frac{\mu\delta\bar{q} + \sqrt{\alpha_D}\bar{S}_1(1-\delta R)}{(1+\delta R)^2 - \delta P \delta\bar{q}} \tag{4.70}$$

$$\hat{q} = \frac{\alpha_D}{\alpha} \frac{\delta q(P_d+1) - 2\delta P}{(\delta q)^3} \tag{4.71}$$

$$\bar{S}_1\bar{S}_1 = 1 - \bar{q}_d \tag{4.72}$$

and by losing two equations, from 4.55-4.62, due to linear dependency we get 6 saddle point equations in the following variables: $P, P_d, R, R_d, \bar{q}, \bar{q}_d$:

$$\begin{aligned}&\alpha_D(-((-1+2P_d+\alpha_T)\delta P) + P_d(\delta q - \alpha_T\delta q) + g(-1+\delta R)\delta R + \\ &gR_d\delta q(-1+2\delta R)) + gS_1\sqrt{\alpha_D}\delta q\delta R\mu_1 + (-\delta P + \delta q)\mu_1^2 = 0\end{aligned} \tag{4.73}$$

$$\begin{aligned}&\alpha_D(-1+P_d+R_d(-\delta P + \delta q - \alpha_T\delta q + g\delta q\delta\bar{q}) - \alpha_T\delta R - P_d\delta R + \\ &\delta R + g\bar{q}_d\delta q\delta R + gS_1^2\delta q\delta R + g\delta\bar{q}\delta R) + S_1\sqrt{\alpha_D}(-\delta P + \delta q)\mu_1 = 0\end{aligned} \tag{4.74}$$

$$\begin{aligned}&-\alpha_D(1 + R_d(\delta P + \delta q(-1+\alpha_T - g\delta\bar{q})) - g(\bar{q}_d\delta q + \delta\bar{q})(-1+\delta R) + \\ &-\delta R + P_d\delta R + \alpha_T\delta R) + S_1\sqrt{\alpha_D}\delta q(1+g\delta\bar{q})\mu_1 - \delta R\mu_1^2 = 0\end{aligned} \tag{4.75}$$

$$\alpha_D(\delta q + \delta P\delta R + (-1+\alpha_T)\delta q\delta R + g\delta q(\delta\bar{q} - \delta\bar{q}\delta R)) = 0 \tag{4.76}$$

$$\alpha_D(\delta P^2 + (-1+\alpha_T)\delta P\delta q - g\delta q(-1+\delta R)\delta R) = 0 \tag{4.77}$$

$$\alpha_D(\delta q + \delta P(-1+\delta R) + \delta q(-1+\alpha_T - g\delta\bar{q})\delta R) = 0 \tag{4.78}$$

Where we have defined the function

$$H(x) = \frac{1}{2}\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) \tag{4.79}$$

where the complementary error function erfc reads

$$\mathrm{erfc} = 2\int_x^\infty \frac{dy}{\sqrt{\pi}} e^{-y^2} \tag{4.80}$$

4.2.6 Limit $\alpha \to \infty$ (from $\beta \to \infty$)

We are now going to perform the $\alpha \to \infty$ limit for the saddle point equations, to check how the system behaves in the regime in which the number of memories diverges with respect to the number of neurons in the visible layer, $\alpha = \frac{p}{N}$. By

inspection we obtain that in the limit $\alpha \to \infty$ the scalings for the order parameters of the model are as follows:

$$\delta q \to \frac{\alpha_D}{\alpha}\delta q \tag{4.81}$$
$$\delta P \to \frac{\alpha_D}{\alpha}\delta P \tag{4.82}$$
$$\delta R \to \frac{\alpha_D}{\alpha}\delta R \tag{4.83}$$
$$\delta \bar{q} \to \frac{\alpha_D}{\alpha}\delta \bar{q} \tag{4.84}$$
$$\hat{q} \to \frac{\alpha}{\alpha_D}\hat{q} \tag{4.85}$$
$$\hat{\mu} \to \frac{\alpha}{\alpha_D}\hat{\mu} \tag{4.86}$$
$$g \to \frac{\alpha}{\alpha_D}g \tag{4.87}$$

So from the zero temperature limit, $\beta \to \infty$, when $\alpha \to \infty$ the equations became:

$$\delta q = \frac{2}{\sqrt{\alpha_D \hat{q}}} G\left(-\frac{\hat{\mu}}{\sqrt{\alpha_D \hat{q}}}\right) \tag{4.88}$$
$$\mu = 2H\left(-\frac{\hat{\mu}}{\sqrt{\alpha_D \hat{q}}}\right) - 1 \tag{4.89}$$
$$\hat{q} = \frac{\delta q(P_d - 1) - 2\delta P}{(\delta q)^3} \tag{4.90}$$
$$\hat{\mu} = \sqrt{\alpha_D}\bar{S}_1 \tag{4.91}$$
$$g = -\frac{\mu}{\sqrt{\alpha_D}\bar{S}_1} \tag{4.92}$$
$$\bar{S}_1\bar{S}_1 = 1 - \bar{q}_d \tag{4.93}$$

$$-\alpha_D(\delta P + P_d\delta q + gR_d\delta q + g\delta R) = 0 \tag{4.94}$$

$$\alpha_D(-1 + P_d - R_d\delta q - \delta R) = 0 \tag{4.95}$$

$$-\alpha_D(1 + g\bar{q}_d\delta q + R_d\delta q + g\delta\bar{q} + \delta R) = 0 \tag{4.96}$$

$$\alpha_D(\alpha_D\delta P\delta q + g\alpha_D\delta q\delta R) = 0 \tag{4.97}$$

$$\alpha_D(-\alpha_D\delta P + \alpha_D\delta q + \alpha_D\delta q\delta R) = 0 \tag{4.98}$$

$$\alpha_D(\alpha_D\delta q + g\alpha_D\delta q\delta\bar{q} + \alpha_D\delta q\delta R) = 0 \tag{4.99}$$

Where the equations 4.94-4.99 are a system of six equation in the order parameters: $P_d, R_d, \bar{q}_d, \delta P, \delta R, \delta\bar{q}_d$. With Mathematica we found out that this system of 6

equations admits a solution:

$$P_d \to \frac{g^2}{(g+\delta q)^2} \tag{4.100}$$

$$R_d \to -\frac{g}{(g+\delta q)^2} \tag{4.101}$$

$$\bar{q}_d \to \frac{1}{(g+\delta q)^2} \tag{4.102}$$

$$\delta P \to \frac{g\delta q}{g+\delta q} \tag{4.103}$$

$$\delta R \to -\frac{\delta q}{g+\delta q} \tag{4.104}$$

$$\delta \bar{q} \to -\frac{1}{g+\delta q} \tag{4.105}$$

However we can conclude that the only possible physical solution is the trivial one of $\delta q = 0$ and $\mu = 1$ identically. In fact, being $g < 0$ always due to 4.92, the only possible value to have $P_d \leq 1$ is $\delta q = 0$. The trivial solution we obtained is indeed the one when the system is always fully magnetized: being $\delta q = 0$ and $\mu = 1$ identically in this limit means that we have always a non zero overlap for the feature magnetizations and for the spins, and we are always in the retrieval phase. Also it is important to notice that this result is dipendent free from the value of α_D, meaning that in this limit the system thas not care on the asymmetry between the layers or on how the number of features scale with respect to the visible layer.

4.2.7 Limit $\alpha_D \to \infty$ (from $\beta \to \infty$)

In the previous section we found out that in the limit $\alpha \to \infty$ the system is always fully magnetized dependet free of the control parameter α_D. We then analyzed the $\alpha_D \to \infty$ limit: we want to check how the model behaves in the regime where the asymmetry between the hidden layer and the visible layer diverges, which is the same regime where the number of features diverges with respect to the visible layer. In the following we display the saddle point equations in the $\alpha_D \to \infty$ limit for the order parameters $P_d, R_d, \bar{q}_d, \delta P, \delta R, \delta \bar{q}$:

$$\begin{aligned} &-((-1+2P_d+\alpha_T)\delta P) + P_d(\delta q - \alpha_T \delta q) + \\ &g(-R_d \delta q(\alpha_T - 2\delta R) + \delta R(-\alpha_T + \delta R)) = 0, \end{aligned} \tag{4.106}$$

$$\begin{aligned} &R_d(-\delta P + \delta q + g\delta q\delta\bar{q}) + \alpha_T(-1 + P_d - R_d\delta q - \delta R) + \\ &\delta R - P_d\delta R + g(\bar{q}_d\delta q + S_1^2\delta q + \delta\bar{q})\delta R = 0, \end{aligned} \tag{4.107}$$

$$\begin{aligned} &-R_d\delta P + R_d\delta q + gR_d\delta q\delta\bar{q} + \delta R - P_d\delta R + g\bar{q}_d\delta q\delta R + \\ &g\delta\bar{q}\delta R - \alpha_T(1 + R_d\delta q + g(\bar{q}_d\delta q + \delta\bar{q}) + \delta R) = 0, \end{aligned} \tag{4.108}$$

$$\delta P^2 + (-1+\alpha_T)\delta P\delta q + g\delta q(\alpha_T - \delta R)\delta R = 0, \tag{4.109}$$

$$\delta P - \delta q(1 + g\delta\bar{q}))\delta R + \alpha_T(-\delta P + \delta q + \delta q\delta R = 0 \tag{4.110}$$

$$\delta P - \delta q(1 + g\delta\bar{q}))\delta R + \alpha_T \delta q(1 + g\delta\bar{q} + \delta R = 0 \tag{4.111}$$

without exhibiting here the explicit solutions obtained with Mathematica, we just say that they seem much more complicated and need a more careful analysis. We have still not obtained a numerical solution for the equations 4.106-4.111, but there is the possibility to have a solution in this limit where the model allows a phase transition between a magnetized and a non magnetized phase. We expect to find this transition by tuning the control parameter α_T which represent the ratio between the number of memories and the number of features.

Part III

Conclusions

Chapter 5

Discussion

In this work we discussed the retrieval properties of the BAM when the memory pairs are structured examples and their representation, to provide a first comprehension on the mechanism of feature extraction in RBMs. More precisely, we built the model so that the two layers of the BAM would store different levels of information, emulating the role of visible and hidden layer in the RBMs: for the visible layer we chose memories derived from a linear combination of features weighted by coefficients, while for the hidden layer the memories are exactly the coefficients linked with the patterns of the visible layer. With this choice, one hidden neuron should represent one feature, and the numerical value of the i^{th} hidden neuron should represent the intensity with which the i^{th} feature contributes to the creation of a specific memory.
We computed the free energy of the model in the Replica symmetric ansatz, and we showed that the combination of the two layer architecture of the BAM with memories correlated from a latent space produces new order parameters representing the overlap between the hidden layer and the feature magnetization in the visible layer.We solved the saddle point equations in the $\alpha \to \infty$ limit starting from the $\beta \to \infty$ limit in the case where the model retrieves only one of the features. The results depicted in 4.2.6 showed that in this limit we have a numerical solution that either violates the physical and mathematical definitions of the order parameters, for example giving some of them a negative value, or it is the trivial one, with the overlap of the spins in the visible layer and the feature magnetizations being identically equal to 1. This means that the model always admits condensation, and that it is always fully magnetized. This result is also dependent free from α_D which represents the asymmetry between the two layers and the scaling between the number of features with respect to the number of neurons.
At this point we asked ourselves if a non trivial behavior for the model may happen in a different regime. From 4.2.6 we realized that scaling the number of features linearly with respect to the number of neurons in the visible layer didn't provide interesting results, and it could be that we need a larger scaling. In 4.2.7 we performed the limit $\alpha_D \to \infty$ to study what happens when the number of features scales more than linearly with respect to the dimension of the visible layer and we derived non trivial equations for which we still have to find a numerical solution. We expect to find a phase transition between a magnetized and a non magnetized phase

in this limit by tuning the control parameter α_T which represents how the number of memories scales with respect to the number of neurons in the visible layer.
Our current study leaves many future directions to be explored: the new saddle point equations in the $\alpha_D \to \infty$ limit still need to be numerically solved. Moreover there could be also other interesting limits for the control parameters that need to be analysed. A more careful inspection on the scalings for the order parameters of the model also could be performed: for example there are portions of 4.47 that diverge when the asymmetry between the two layers diverge, in the limit $\alpha_D \to \infty$, or in 4.28-4.31 there are some cancellations when $\alpha_T = 1$, so when the number of features equals the number of memories. This could be done with the help from numerical simulations, combined with our present knowledge of which scalings are non trivial. Finally, it could also be of interest to see if there is a different choice for the structure of the memory pairs, in addition to the architecture we proposed.

Appendix A

Decoupling and equivalence between BAM and RBMs

The usual technique to simplify the Boltzmann weight in fully connected spin glasses with quadratic interactions is to introduce a Hubbard–Stratonovich integral transformation. However, due to the two-layer structure of the BAM, in [26] and in [7] they used a different integral transformation involving a pair of complex conjugate variables $(z, z^\dagger)$. Such a procedure was also used to analyze other bipartite models such as the bipartite Sherrington-Kirkpatrik model [16] or RBMs [8, 9].
The procedure consists in, starting from the partition function:

$$Z = \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \exp\Big\{ -\frac{\beta}{L} \sum_{\mu=1}^{p} \sum_{a} \Big(\sum_i \xi_i^\mu S_i^a \Big) \Big(\sum_k \bar{\xi}_k^\mu \bar{S}_k^a \Big) \Big\} \tag{A.1}$$

introducing two Hubbard-Stratonovich variables u_μ^a and v_μ^a for each replica a, exploiting the identity:

$$xy = \frac{1}{4}[(x+y)^2 - (x-y)^2]. \tag{A.2}$$

We thus obtain:

$$\begin{aligned} Z =& \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} du_\mu^a dv_\mu^a \exp\Big\{ -\frac{\beta}{2} L \sum_{\mu a} (u_\mu^{a\,2} + v_\mu^{a\,2}) \Big\} \\ & \exp\Big\{ \frac{\beta}{\sqrt{2}} \sum_{\mu a} \Big[u_\mu^a \Big(\sum_i \xi_i^\mu S_i^a + \sum_k \bar{\xi}_k^\mu \bar{S}_k^a \Big) + \mathrm{i} v_\mu^a \Big(\sum_i \xi_i^\mu S_i^a + \sum_k \bar{\xi}_k^\mu \bar{S}_k^a \Big) \Big] \Big\} \end{aligned} \tag{A.3}$$

We can now introduce the two auxiliary fields z_μ^a and $z_\mu^{a\dagger}$ by defining:

$$z_\mu^a = \frac{u_\mu^a + \mathrm{i} v_\mu^a}{\sqrt{2}}, \tag{A.4}$$

$$z_\mu^{a\dagger} = \frac{u_\mu^a - \mathrm{i} v_\mu^a}{\sqrt{2}}, \tag{A.5}$$

so that

$$z_\mu^a z_\mu^{a\dagger} = \frac{1}{2}(u_\mu^{a\,2} + v_\mu^{a\,2}). \tag{A.6}$$

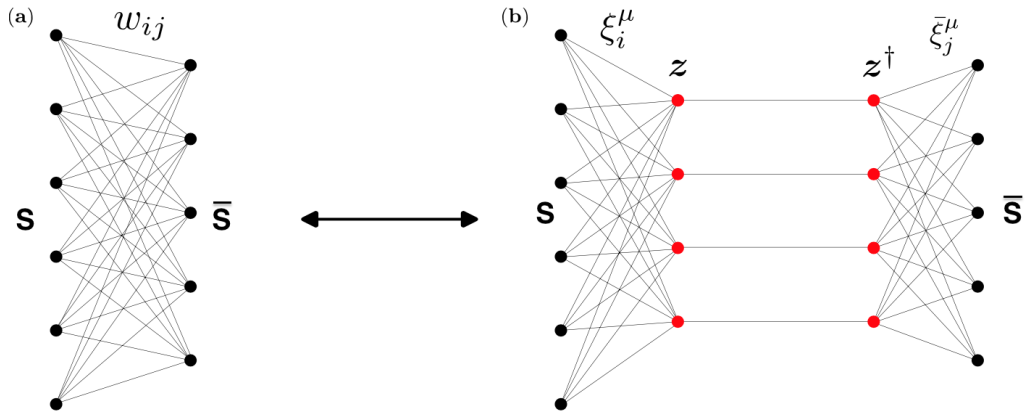


Figure A.1. Schematic representation of the equivalence between the BAM (a) and two coupled RBMs (b). Figure taken from [7]

After applying the previous transformation, the partition function reads:

$$\begin{aligned} Z =& \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dz_\mu^a dz_\mu^{a\dagger} \exp\left\{-\beta L \sum_{\mu a} z_\mu^a z_\mu^{a\dagger}\right\} \\ & \exp\left\{\beta \sum_{\mu a}\left[z_\mu^a \sum_i \xi_i^\mu S_i^a + z_\mu^{a\dagger} \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\right]\right\} \end{aligned} \tag{A.7}$$

The integral transformation that decoupled A.1 in A.7 can be exploited to derive a structural analogy between the BAM's partition function and the partition function of two coupled RBMs, by generalizing the equivalence between the Hopfield model and a Binary-Gaussian RBM [6]. In fact, for each replica, it is possible to interpret the right hand side of equation A.7 as the partition function of a 4-partite system similar to two coupled RBMs as shown in Figure A.1: The two RBMs on (b) have each one a binary visible layer (with sizes N and $\bar{N}$ respectively), the two hidden layers are encoded in the vectors $\mathbf{z}, \mathbf{z}^\dagger$: each of them has a size equal to the K number of patterns in the BAM, and for each $\mu = 1, \ldots, K$ the hidden nodes $z_\mu, z_\mu^\dagger$ are complex conjugates, interacting through a fixed potential. However, even if the analogy seems to be consistent from a structural point of view, it is not clear how to exploit it to perform learning in the resulting artificial network depicted in Figure A.1.

Appendix B

Replica symmetric computation for the BAM

We will start the replica symmetric computation for the BAM from 3.14.
The next steps are to introduce the conjugate fields $\hat{m}_\mu^a$ and $\hat{\nu}_\mu^a$ by applying the Fourier transform of the delta function, and to distinguish between a first finite subset of $\mu = 1, \dots, l$ low components, and the remaining high components $\mu = l+1, \dots, p$, according to references [35, 26, 7].

$$\begin{aligned}\langle\langle\ Z^n \rangle\rangle\ = \Big\langle\Big\langle \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a d\hat{m}_\mu^a d\hat{\nu}_\mu^a \exp\Big\{\beta \sum_a \sum_{\mu=l+1}^{p} \Big(m_\mu^a \nu_\mu^a + \mathrm{i}\hat{m}_\mu^a m_\mu^a + \mathrm{i}\hat{\nu}_\mu^a \nu_\mu^a\Big)\Big\} \\ \exp\Big\{\beta \sum_a \sum_{\mu=1}^{l} \Big(m_\mu^a \nu_\mu^a + \mathrm{i}\sqrt{N}\hat{m}_\mu^a m_\mu^a + \mathrm{i}\sqrt{N}\hat{\nu}_\mu^a \nu_\mu^a\Big)\Big\} \\ \exp\Big\{-\mathrm{i}\beta \sum_a \sum_{\mu=1}^{l} \Big(\hat{m}_\mu^a \sum_i \xi_i^\mu S_i^a + \hat{\nu}_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big)\Big\} \\ \exp\Big\{-\mathrm{i}\beta \sum_a \sum_{\mu=l+1}^{p} \Big(\frac{1}{\sqrt{N}}\hat{m}_\mu^a \sum_i \xi_i^\mu S_i^a + \frac{1}{\sqrt{\bar{N}}}\hat{\nu}_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big)\Big\}\Big\rangle_\xi\Big\rangle_{\bar{\xi}}\end{aligned} \tag{B.1}$$

Where we have also scaled the low components of the partition function according to 3.17 and 3.18.
The next step is to average over the noisy memories ξ^μ and $\bar{\xi}^\mu$, with $\mu = l+1, \dots, p$. They appear only in the last exponential of the previous equation:

$$\Big\langle\Big\langle \exp\Big\{-\mathrm{i}\sum_{\mu a} \frac{\beta}{\sqrt{N}}\hat{m}_\mu^a \sum_i \xi_i^\mu S_i^a\Big\}\Big\rangle_\xi\Big\rangle_{\bar{\xi}} = \exp\Big\{-\frac{\beta^2}{2N}\sum_\mu \sum_{ab} \hat{m}_\mu^a \hat{m}_\mu^b \sum_i S_i^a S_i^b\Big\} \tag{B.2}$$

$$\Big\langle\Big\langle \exp\Big\{-\mathrm{i}\sum_{\mu a} \frac{\beta}{\sqrt{\bar{N}}}\hat{\nu}_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big\}\Big\rangle_\xi\Big\rangle_{\bar{\xi}} = \exp\Big\{-\frac{\beta^2}{2\bar{N}}\sum_\mu \sum_{ab} \hat{\nu}_\mu^a \hat{\nu}_\mu^b \sum_k \bar{S}_k^a \bar{S}_k^b\Big\} \tag{B.3}$$

Where we have used the following properties: the ξ_i^μ and the $\bar{\xi}^\mu$ are independently drawn from the distributions 3.3 and 3.4 and both the expansions $\cos(x) \sim 1 - \frac{1}{2}x^2$ and $\ln(1+x) \sim x$ hold.

Inserting the results from B.2 and B.3 in B.1 and defining the overlap q_{ab} and $\hat{q}_{ab}$, according to 3.20 and 3.21, the partition function now reads:

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \Bigg\langle\Bigg\langle \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a d\hat{m}_\mu^a d\hat{\nu}_\mu^a \exp\Bigg\{\beta \sum_a \sum_{\mu=l+1}^{p} \Big(m_\mu^a \nu_\mu^a + \mathrm{i}\hat{m}_\mu^a m_\mu^a + \mathrm{i}\hat{\nu}_\mu^a \nu_\mu^a\Big)\Bigg\} \\ \exp\Bigg\{\beta \sum_a \sum_{\mu=1}^{l} \Big(m_\mu^a \nu_\mu^a + \mathrm{i}\sqrt{N}\hat{m}_\mu^a m_\mu^a + \mathrm{i}\sqrt{\bar{N}}\hat{\nu}_\mu^a \nu_\mu^a\Big)\Bigg\} \\ \exp\Bigg\{-\mathrm{i}\beta \sum_a \sum_{\mu=1}^{l} \Big(\hat{m}_\mu^a \sum_i \xi_i^\mu S_i^a + \hat{\nu}_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big)\Bigg\} \\ \exp\Bigg\{-\frac{\beta^2}{2} \sum_\mu \sum_{ab} \hat{m}_\mu^a \hat{m}_\mu^b q_{ab}\Bigg\} \\ \exp\Bigg\{-\frac{\beta^2}{2} \sum_\mu \sum_{ab} \hat{\nu}_\mu^a \hat{\nu}_\mu^b \bar{q}_{ab}\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2} \sum_{ab} r_{ab} q_{ab} - \frac{p\beta^2}{2N} \sum_{ab} r_{ab} \sum_i S_i^a S_i^b\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2} \sum_{ab} \bar{r}_{ab} \bar{q}_{ab} - \frac{p\beta^2}{2\bar{N}} \sum_{ab} \bar{r}_{ab} \sum_k \bar{S}_k^a \bar{S}_k^b\Bigg\}\Bigg\rangle_{\{\xi\}_1^l}\Bigg\rangle_{\{\bar{\xi}\}_1^l}\end{aligned} \tag{B.4}$$

Where r_{ab} and $\bar{r}_{ab}$ are the conjugate variables for q_{ab} and $\hat{q}_{ab}$. There are now two Gaussian integral that need to be performed in the high variables $\hat{m}_\mu$ $\hat{\nu}_\mu$, $\mu = l+1, \ldots, p$:

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \Bigg\langle\Bigg\langle \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a d\hat{m}_\mu^a d\hat{\nu}_\mu^a \exp\Bigg\{\beta \sum_a \sum_{\mu=l+1}^{p} m_\mu^a \nu_\mu^a\Bigg\} \\ \exp\Bigg\{\beta \sum_a \sum_{\mu=1}^{l} m_\mu^a \nu_\mu^a + \mathrm{i}\sqrt{N}\hat{m}_\mu^a m_\mu^a + \mathrm{i}\sqrt{\bar{N}}\hat{\nu}_\mu^a \nu_\mu^a\Bigg\} \\ \exp\Bigg\{-\mathrm{i}\beta \sum_a \sum_{\mu=1}^{l} \Big(\hat{m}_\mu^a \sum_i \xi_i^\mu S_i^a + \hat{\nu}_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big)\Bigg\} \\ \exp\Bigg\{-\frac{p}{2} \ln\det(Q)\Bigg\} \exp\Bigg\{-\frac{p}{2} \ln\det(\bar{Q})\Bigg\} \\ \exp\Bigg\{-\frac{1}{2} \sum_\mu \sum_{ab} m_\mu^a m_\mu^b q_{ab}^{-1}\Bigg\} \exp\Bigg\{-\frac{1}{2} \sum_\mu \sum_{ab} \nu_\mu^a \nu_\mu^b \bar{q}_{ab}^{-1}\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2} \sum_{ab} r_{ab} q_{ab} - \frac{p\beta^2}{2N} \sum_{ab} r_{ab} \sum_i S_i^a S_i^b\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2} \sum_{ab} \bar{r}_{ab} \bar{q}_{ab} - \frac{p\beta^2}{2\bar{N}} \sum_{ab} \bar{r}_{ab} \sum_k \bar{S}_k^a \bar{S}_k^b\Bigg\}\Bigg\rangle_{\{\xi\}_1^l}\Bigg\rangle_{\{\bar{\xi}\}_1^l}\end{aligned} \tag{B.5}$$

And then we can compute the integrals over the m_μ and ν_μ fields:

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \Bigg\langle\Bigg\langle \mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \int \prod_{\mu a} dm_\mu^a d\nu_\mu^a \exp\Bigg\{-\beta \sum_a \sum_{\mu=1}^{l} \sqrt{N\bar{N}} m_\mu^a \nu_\mu^a\Bigg\} \\ \exp\Bigg\{-\frac{p}{2}\ln\det\Big(\mathbb{1} + \beta^2 Q\bar{Q}\Big)\Bigg\} \\ \exp\Bigg\{\beta \sum_a \sum_{\mu=1}^{l}\Big(m_\mu^a \sum_i \xi_i^\mu S_i^a + \nu_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a\Big)\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2}\sum_{ab} r_{ab}q_{ab} - \frac{p\beta^2}{2N}\sum_{ab} r_{ab}\sum_i S_i^a S_i^b\Bigg\} \\ \exp\Bigg\{-\frac{p\beta^2}{2}\sum_{ab} \bar{r}_{ab}\bar{q}_{ab} - \frac{p\beta^2}{2\bar{N}}\sum_{ab} \bar{r}_{ab}\sum_k \bar{S}_k^a \bar{S}_k^b\Bigg\}\Bigg\rangle_{\{\xi\}_1^l}\Bigg\rangle_{\{\bar{\xi}\}_1^l}\end{aligned} \tag{B.6}$$

Where Q and $\bar{Q}$ are the overlap matrixes defined in 3.23. Moreover we also introduced a change of variable $\mathrm{i}\hat{m}_\mu \to m_\mu$ and $\mathrm{i}\hat{\nu}_\mu \to \nu_\mu$ to lose the annoying dependency on the hat and on the imaginary unit. We also have used the property that the $\det(\mathbb{A})\det(\mathbb{B}) = \det(\mathbb{A}\mathbb{B})$.
Now it is time to introduce the Replica Simmetric ansatz for all the order parameters:

$$q_{ab} = \delta_{ab} + q(1-\delta_{ab}) \tag{B.7}$$
$$\bar{q}_{ab} = \delta_{ab} + barq(1-\delta_{ab}) \tag{B.8}$$
$$r_{ab} = r\delta_{ab} + r(1-\delta_{ab}) \tag{B.9}$$
$$\bar{r}_{ab} = \bar{r}\delta_{ab} + \bar{r}(1-\delta_{ab}) \tag{B.10}$$
$$m_\mu^a = m_\mu \tag{B.11}$$
$$\nu_\mu^a = \nu_\mu. \tag{B.12}$$

And with this ansatz we can compute all the remaining parts in the partition function:

$$-\beta \sum_a \sum_\mu \sqrt{N\bar{N}} m_\mu^a \nu_\mu^a = -\beta \sum_\mu n\sqrt{N\bar{N}} m_\mu \nu_\mu \tag{B.13}$$

$$-\frac{p\beta^2}{2}\sum_{ab} \bar{r}_{ab}\bar{q}_{ab} \simeq nr(1-q) \tag{B.14}$$

$$-\frac{p\beta^2}{2}\sum_{ab} \bar{r}_{ab}\bar{q}_{ab} \simeq n\bar{r}(1-\bar{q}) \tag{B.15}$$

$$\begin{aligned}&\mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \exp\Bigg\{\beta \sum_a \sum_\mu m_\mu^a \sum_i \xi_i^\mu S_i^a - \frac{p\beta^2}{2N}\sum_{ab} r_{ab}\sum_i S_i^a S_i^b\Bigg\} = \\ &\exp\Bigg\{\frac{\gamma}{\beta}\int Dz \ln\Big[2\cosh(\beta\sqrt{\bar{\gamma}\alpha r}z + \beta\bar{\gamma}\sum_\mu m_\mu \xi^\mu)\Big]\Bigg\}\end{aligned} \tag{B.16}$$

$$\begin{aligned}&\mathrm{Tr}_S \mathrm{Tr}_{\bar{S}} \exp\left\{\beta \sum_a \sum_\mu \nu_\mu^a \sum_k \bar{\xi}_k^\mu \bar{S}_k^a - \frac{p\beta^2}{2\bar{N}} \sum_{ab} \bar{r}_{ab} \sum_k \bar{S}_k^a \bar{S}_k^b\right\} = \\ &\exp\left\{\frac{\bar{\gamma}}{\beta} \int D\bar{z} \ln\left[2\cosh(\beta\sqrt{\gamma\alpha\bar{r}}\bar{z} + \beta\gamma \sum_\mu \nu_\mu \bar{\xi}^\mu)\right]\right\}\end{aligned} \tag{B.17}$$

$$-\frac{p}{2} \ln\det\left(\mathbb{1} + \beta^2 Q\bar{Q}\right) = n\left[\ln\left(1-\beta^2(1-q)(1-\bar{q})\right) + \beta^2 \frac{q(1-\bar{q}) + \bar{q}(1-q)}{1-\beta^2(1-q)(1-\bar{q}}\right] \tag{B.18}$$

Where in B.18 we applied the following formula

$$\ln\det \mathbb{X} = n\ln(x_d - x) + n\frac{x}{x_d - x} + O(n^2) \tag{B.19}$$

Where $\mathbb{X}$ is a $n \times n$ RS matrix with diagonal term x_d and off diagonal term x. Moreover in the previous equations we have defined some control parameters for the model: $\gamma = \frac{N}{\bar{N}}$ and $\bar{\gamma} = \gamma^{-1}$, $\alpha = \frac{p}{\sqrt{N\bar{N}}}$. Finally z and $\bar{z}$ are two independent identically distributed standard gaussian variables.
Now combining all the terms obtained with the RS ansatz we display the replicated free energy for the BAM model:

$$\begin{aligned}f^{RS} = &\sum_{\mu=1}^{l} m_\mu \nu_\mu + \frac{\alpha\beta}{2} r(1-q) + \frac{\alpha\beta}{2}\bar{r}(1-\bar{q}) + \\ &- \frac{\gamma}{\beta}\left\langle \int Dz \ln\left[2\cosh(\beta\sqrt{\bar{\gamma}\alpha r}z + \beta\bar{\gamma}\sum_\mu m_\mu \xi^\mu)\right]\right\rangle + \\ &- \frac{\bar{\gamma}}{\beta}\left\langle \int D\bar{z} \ln\left[2\cosh(\beta\sqrt{\gamma\alpha\bar{r}}\bar{z} + \beta\gamma\sum_\mu \nu_\mu \bar{\xi}^\mu)\right]\right\rangle + \\ &- \frac{\alpha}{2\beta}\ln\left(1-\beta^2(1-q)(1-\bar{q})\right) - \frac{\alpha}{2\beta}\frac{q(1-\bar{q}) + \bar{q}(1-q)}{1-\beta^2(1-q)(1-\bar{q})}\end{aligned} \tag{B.20}$$

Appendix C

Replica symmetric computation for Random Feature BAM

We will start the replica symmetric computation for the Random Feature BAM from 4.13.
The next step in the computation is to average the partition over the coefficients c_k^μ which are distributed according to a Gaussian distribution. In 4.13 the exponential that depends on the coefficients c_k^μ is the last one, so to perform the computation we can isolate it:

$$
\begin{aligned}
&< \exp\left\{ \frac{\beta}{\sqrt{D}} \sum_{\mu a} \hat{m}_\mu^a \sum_k c_k^\mu \mu_k^a + \frac{\beta}{\sqrt{D}} \sum_{\mu a} \hat{\nu}_\mu^a \sum_k c_k^\mu \bar{S}_k^a \right\} >_c = \\
&\int dc_{\mu k} \exp\left\{ -\frac{1}{2} c_{\mu k}^2 + c_{\mu k} \left(\frac{\beta}{\sqrt{D}} \sum_a \hat{m}_\mu^a \mu_k^a + \frac{\beta}{\sqrt{D}} \sum_a \hat{\nu}_\mu^a \bar{S}_k^a \right) \right\} = \\
&\exp\left\{ \frac{\beta^2}{2D} \left(\sum_{ab} \hat{m}_\mu^a \mu_k^a \hat{m}_\mu^b \mu_k^b + \sum_{ab} \hat{\nu}_\mu^a \bar{S}_k^a \hat{\nu}_\mu^b \bar{S}_k^b + \right. \right. \\
&\left. \left. \sum_{ab} \hat{m}_\mu^a \mu_k^a \hat{\nu}_\mu^b \bar{S}_k^b + \sum_{ab} \hat{m}_\mu^b \mu_k^b \hat{\nu}_\mu^a \bar{S}_k^a \right) \right\}
\end{aligned}
\tag{C.1}
$$

Now we will insert this new result in 4.13 and exhibit the ansatz for the retrieval of one feature only according to 4.16 and 4.17.
Let's also notice that the integral in dm_μ^a can be seen as a $\delta(\nu_\mu^a + \hat{m}_\mu^a)$, and then, integrating in $d\nu_\mu^a$ the first exponential in 4.13 remains only as

$$
\begin{aligned}
&\int \prod_{\mu a} dm_\mu^a d\nu_\mu^a \exp\left\{ \sum_{\mu a} (\beta(m_\mu^a \nu_\mu^a) - \beta(\hat{m}_\mu^a m_\mu^a) - \beta(\hat{\nu}_\mu^a \nu_\mu^a)) \right\} = \\
&\int \prod_{\mu a} d\nu_\mu^a \exp\left\{ -\beta \sum_{\mu a} \hat{\nu}_\mu^a \nu_\mu^a \right\} \delta(\nu_\mu^a - \hat{m}_\mu^a) = \\
&\exp\left\{ -\beta \sum_{\mu a} \hat{\nu}_\mu^a \hat{m}_\mu^a \right\} \quad .
\end{aligned}
\tag{C.2}
$$

The partition function at this point then reads:

$$
\begin{aligned}
<< \ Z^n >> = \ < Tr_S \int \prod_{\mu a k} d\bar{S}^a_k d\hat{m}^a_\mu d\hat{\nu}^a_\mu d\mu^a_k d\hat{\mu}^a_k dg_a \\
&\exp\Big\{-\beta \sum_{\mu a} \hat{\nu}^a_\mu \hat{m}^a_\mu\Big\} \exp\Big\{-\sum_a \frac{g_a}{2}(D - \sum_{k>1} \bar{S}^a_k \bar{S}^a_k - D\bar{S}^a_1 \bar{S}^a_1)\Big\} \\
&\exp\Big\{-\sum_{ka} \hat{\mu}^a_k \mu^a_k + \frac{1}{\sqrt{N}} \sum_{ka} \hat{\mu}^a_k \sum_i f_{ki} S^a_i\Big\} \\
&\exp\Big\{-\sum_a \hat{\mu}^a_1 \mu^a_1 + \frac{1}{N} \sum_a \hat{\mu}^a_1 \sum_i f_{1i} S^a_i\Big\} \\
&\exp\Big\{\frac{1}{2}\Big(\frac{\beta^2}{D} \sum_{\mu k} \sum_{ab} \hat{m}^a_\mu \mu^a_k \hat{m}^b_\mu \mu^b_k + \frac{\beta^2}{D} \sum_{\mu k} \sum_{ab} \hat{\nu}^a_\mu \bar{S}^a_k \hat{\nu}^b_\mu \bar{S}^b_k\Big)\Big\} \\
&\exp\Big\{\frac{1}{2}\Big(\frac{\beta^2}{D} \sum_{\mu k} \sum_{ab} \hat{m}^a_\mu \mu^a_k \hat{\nu}^b_\mu \bar{S}^b_k + \frac{\beta^2}{D} \sum_{\mu k} \sum_{ab} \hat{m}^b_\mu \mu^b_k \hat{\nu}^a_\mu \bar{S}^a_k\Big)\Big\} \\
&\exp\Big\{\frac{1}{2}\Big(\beta^2 \frac{N}{D} \sum_{ab} \hat{m}^a_\mu \mu^a_1 \hat{m}^b_\mu \mu^b_1 + \beta^2 \sum_{ab} \hat{\nu}^a_\mu \bar{S}^a_1 \hat{\nu}^b_\mu \bar{S}^b_1\Big)\Big\} \\
&\exp\Big\{\frac{1}{2}\Big(\beta^2 \sqrt{\frac{N}{D}} \sum_{ab} \hat{m}^a_\mu \mu^a_1 \hat{\nu}^b_\mu \bar{S}^b_1 + \beta^2 \sqrt{\frac{N}{D}} \sum_{ab} \hat{m}^b_\mu \mu^b_1 \hat{\nu}^a_\mu \bar{S}^a_1\Big)\Big\} >
\end{aligned}
\tag{C.3}
$$

Where now the remaining mean $< \cdot >$ is only over the features f_{ki}.

C.1 Integrating pattern magnetizations

From C.3 we can se how the fields $\hat{m}^a_\mu$ and $\hat{\nu}^a_\mu$ have taken the role of pattern magnetizations. We now procede in integrating over them. To do so we introdue a couple variable $T^a_\mu = (\sqrt{\beta}\hat{m}^a_\mu, \sqrt{\beta}\hat{\nu}^a_\mu)$. With this new notation our integral is now written as:

$$
\begin{aligned}
<< \ Z^n >> = \ < Tr_S Tr_{\bar{S}} \int \prod_{\mu a k} dT^a_\mu d\mu^a_k d\hat{\mu}^a_k dg_a \\
&\exp\Big\{-\sum_a \frac{g_a}{2}(D - \sum_{k>1} \bar{S}^a_k \bar{S}^a_k - D\bar{S}^a_1 \bar{S}^a_1)\Big\} \\
&\exp\Big\{\Big(\sum_\mu \Big(-\frac{1}{2} \sum_{ab} T^a_\mu J_{ab} T^b_\mu\Big)\Big\} \\
&\exp\Big\{-\sum_a \sum_{k>1} \hat{\mu}^a_k \mu^a_k + \frac{1}{\sqrt{N}} \sum_a \sum_{k>1} \hat{\mu}^a_k \sum_i f_{ki} S^a_i\Big\} \\
&\exp\Big\{-\sum_a \hat{\mu}^a_1 \mu^a_1 + \frac{1}{N} \sum_a \hat{\mu}^a_1 \sum_i f_{1i} S^a_i\Big\} >
\end{aligned}
\tag{C.4}
$$

where the bloc matrix J_{ab} was defined in 4.19. Computing the Gaussian integral over the dT^a_μ variables we obtain the following result:

$$
\begin{aligned}
<<\ Z^n\ >> = < Tr_S Tr_{\bar{S}} \int \prod_{ak} d\mu_k^a d\hat{\mu}_k^a dg_a \\
&\exp\Big\{ -\sum_a \frac{g_a}{2}(D - \sum_{k>1} \bar{S}_k^a \bar{S}_k^a - D\bar{S}_1^a \bar{S}_1^a) \Big\} \\
&\exp\Big\{ -\frac{p}{2} \ln\det(J) \Big\} \\
&\exp\Big\{ -\sum_a \sum_{k>1} \hat{\mu}_k^a \mu_k^a + \frac{1}{\sqrt{N}} \sum_a \sum_{k>1} \hat{\mu}_k^a \sum_i f_{ki} S_i^a \Big\} \\
&\exp\Big\{ -\sum_a \hat{\mu}_1^a \mu_1^a + \frac{1}{N} \sum_a \hat{\mu}_1^a \sum_i f_{1i} S_i^a \Big\} >
\end{aligned}
\tag{C.5}
$$

C.2 Integrating over the feature magnetizations

Now we average over the f_{ki} for the first $k > 1$ terms. Those are spin variables that can have values of $+1$ or -1 with probability of $\frac{1}{2}$. Indeed looking only at the term with the variables f_{ki}, with $k > 1$, this is what happens:

$$
\begin{aligned}
&< \exp\Big\{ +\frac{1}{\sqrt{N}} \sum_{ka} \hat{\mu}_k^a \sum_i f_{ki} S_i^a \Big\} >_f = \\
&\frac{1}{2} \exp\Big\{ \frac{1}{\sqrt{N}} \sum_{ka} \hat{\mu}_k^a \sum_i S_i^a \Big\} + \frac{1}{2} \exp\Big\{ -\frac{1}{\sqrt{N}} \sum_{ka} \hat{\mu}_k^a \sum_i S_i^a \Big\} = \\
&\prod_{ki} \frac{1}{2} \Big(\exp\Big\{ \frac{1}{\sqrt{N}} \sum_a \hat{\mu}_k^a S_i^a \Big\} + \exp\Big\{ -\frac{1}{\sqrt{N}} \sum_a \hat{\mu}_k^a S_i^a \Big\} \Big) = \\
&\prod_{ki} \cosh \frac{1}{\sqrt{N}} \sum_a \hat{\mu}_k^a S_i^a = \prod_{ki} \exp\Big\{ \ln\cosh \frac{1}{\sqrt{N}} \sum_a \hat{\mu}_k^a S_i^a \Big\} = \\
&\exp\Big\{ \sum_{ki} \ln\cosh \frac{1}{\sqrt{N}} \sum_a \hat{\mu}_k^a S_i^a \Big\} = \exp\Big\{ \frac{1}{2N} \sum_{ki} \sum_{ab} \hat{\mu}_k^a S_i^a S_i^b \hat{\mu}_k^b \Big\}
\end{aligned}
\tag{C.6}
$$

Where with ch we mean the hyperbolic cosin, and in the last equality we have used the fact that $ch(x) \sim 1 + \frac{1}{2}x^2$ and that $\ln(1+x) \sim x$.
So now we have the following situation:

$$
\begin{aligned}
<<\ Z^n\ >> = Tr_S Tr_{\bar{S}} \int \prod_{ak} d\mu_k^a d\hat{\mu}_k^a dg_a \\
&\exp\Big\{ -\sum_a \frac{g_a}{2}(D - \sum_{k>1} \bar{S}_k^a \bar{S}_k^a - D\bar{S}_1^a \bar{S}_1^a) \Big\} \\
&\exp\Big\{ -\frac{p}{2} \ln\det(J) \Big\} \\
&\exp\Big\{ -\sum_a \sum_{k>1} \hat{\mu}_k^a \mu_k^a + \frac{1}{2N} \sum_{abi} \sum_{k>1} \hat{\mu}_k^a S_i^a \hat{\mu}_k^b S_i^b \Big\} \\
&\Big\langle \exp\Big\{ -\sum_a \hat{\mu}_1^a \mu_1^a + \frac{1}{N} \sum_a \hat{\mu}_1^a \sum_i f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}}
\end{aligned}
\tag{C.7}
$$

Here we will define the overlap order parameter according to 4.22. Moreover we are going to compute the integral in $d\hat{\mu}_k^a$ as it is a Gaussian integral with a linear term:

$$\begin{aligned}\int \prod_{a,k>1} d\hat{\mu}_k^a \exp\Big\{\sum_{k>1}\Big(+\frac{1}{2}\sum_{ab}\hat{\mu}_k^a(q_{ab})\hat{\mu}_k^b - \sum_a \hat{\mu}_k^a\mu_k^a\Big)\Big\} = \\ \exp\Big\{-\frac{D}{2}\ln\det(q)\Big\}\exp\Big\{-\sum_{k>1}\frac{1}{2}\Big(\sum_{ab}\mu_k^a(q_{ab})^{-1}\mu_k^b\Big)\Big\}\end{aligned} \tag{C.8}$$

From here we can compute the integral in $d\mu_k^a$ as well. We can introduce a couple variable $L_k^a = (\mu_k^a, \bar{S}_k^a)$ to rewrite the integral in $d\mu_k^a$ and $d\bar{S}_k^a$ and other components of the partition function. Using the introduction of the variable $L_k^a = (\mu_k^a, \bar{S}_k^a)$ we can also redefine the matrix J in the following version:

$$\begin{aligned}J = -\begin{pmatrix} \frac{\beta}{D}\sum_{k>1}\mu_k^a\mu_k^b + \beta\frac{N}{D}\mu_1^a\mu_1^b & \frac{\beta}{D}\sum_{k>1}\mu_k^a\bar{S}_k^b + \beta\sqrt{\frac{N}{D}}\mu_1^a\bar{S}_1^b - \mathbb{1} \\ \frac{\beta}{D}\sum_{k>1}\mu_k^b\bar{S}_k^a + \beta\sqrt{\frac{N}{D}}\mu_1^b\bar{S}_1^a - \mathbb{1} & \frac{\beta}{D}\sum_{k>1}\bar{S}_k^a\bar{S}_k^b + \beta\bar{S}_1^a\bar{S}_1^b \end{pmatrix} = \\ = -\beta A - \begin{pmatrix} \beta\frac{N}{D}\mu_1^a\mu_1^b & \beta\sqrt{\frac{N}{D}}\mu_1^a\bar{S}_1^b - \mathbb{1} \\ \beta\sqrt{\frac{N}{D}}\mu_1^b\bar{S}_1^a - \mathbb{1} & \beta\bar{S}_1^a\bar{S}_1^b \end{pmatrix}\end{aligned} \tag{C.9}$$

Where we have introduced the block matrix A as a $2n \times 2n$ matrix for which the following expression holds:

$$A_{ab} = \begin{pmatrix} \frac{1}{D}\sum_{k>1}\mu_k^a\mu_k^b & \frac{1}{D}\sum_{k>1}\mu_k^a\bar{S}_k^b \\ \frac{1}{D}\sum_{k>1}\mu_k^b\bar{S}_k^a & \frac{1}{D}\sum_{k>1}\bar{S}_k^a\bar{S}_k^b \end{pmatrix} = \frac{1}{D}\sum_{k>1}L_k^aL_k^b, \quad a,b \in [n]. \tag{C.10}$$

Note that A is the block matrix of the order parameters $\frac{1}{D}\sum_{k>1}\mu_k^a\mu_k^b$, overlap between the feature magnetizations, $\frac{1}{D}\sum_{k>1}\bar{S}_k^a\bar{S}_k^b$ overlap between the spins in the hidden layer and $\frac{1}{D}\sum_{k>1}\mu_k^a\bar{S}_k^b$, overlap between the spins in the hidden layer and the feature magnetizations, all for $k > 1$.
Introducing the definition for the block matrix A and for the couple variable

$L_k^a = (\mu_k^a, \bar{S}_k^a)$, the partition function of the model becomes

$$\begin{aligned}
<<\ Z^n\ >>\ &= Tr_S \int \prod_{k>1} \prod_{c,d=1}^{2n} dL_k^c dA_{cd} d\hat{A}_{cd} \prod_{a,b=1}^{n} dq_{ab} d\hat{q}_{ab} d\mu_1^a d\hat{\mu}_1^a d\bar{S}_1^a dg_a \\
&\exp\Big\{ -\sum_a \frac{g_a}{2}(D - D\bar{S}_1^a \bar{S}_1^a) \Big\} \\
&\exp\Big\{ -\frac{p}{2} \ln\det(J) \Big\} \exp\Big\{ -\frac{D}{2} \ln\det(q) \Big\} \\
&\exp\Big\{ -\sum_{k>1} \frac{1}{2} \Big(\sum_{c,d=1}^{2n} L_k^c \begin{pmatrix} (q_{ab})^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} L_k^d \Big) \Big\} \\
&\exp\Big\{ -\frac{N\alpha}{2} \sum_{ab} \hat{q}_{ab} q_{ab} + \frac{\alpha}{2} \sum_{ab} \hat{q}_{ab} \sum_i S_i^a S_i^b \Big\} \\
&\exp\Big\{ -\frac{D}{2} \sum_{cd} \hat{A}_{cd} A_{cd} + \frac{1}{2} \sum_{cd} \hat{A}_{cd} \sum_{k>1} L_k^c L_k^d) \Big\} \\
&\Big\langle \exp\Big\{ -\sum_a \hat{\mu}_1^a \mu_1^a + \frac{1}{N} \sum_a \hat{\mu}_1^a \sum_i f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}}
\end{aligned} \tag{C.11}$$

and we can finally compute the integral in dL_k^a as it is a Gaussian integral with a linear term:

$$\begin{aligned}
&\int \prod_{ak>1} dL_k^a \exp\Big\{ -\sum_{k>1} \frac{1}{2} \Big(\sum_{cd} L_k^c \begin{pmatrix} (q_{ab})^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} L_k^d \Big) + \frac{1}{2} \sum_{k>1} \sum_{cd} L_k^c \hat{A}_{cd} L_k^d) \Big\} = \\
&\int \prod_{ak>1} dL_k^a \exp\Big\{ \sum_{k>1} \Big[-\frac{1}{2} \Big(\sum_{cd} L_k^c \Big(-\hat{A}_{cd} + \begin{pmatrix} (q_{ab})^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} \Big) L_k^b \Big] \Big\} = \\
&\exp\Big\{ -\frac{D}{2} \ln\det\Big(-\hat{A}_{cd} + \begin{pmatrix} (q_{ab})^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} \Big) \Big\}
\end{aligned} \tag{C.12}$$

Inserting the result of this computation our equation now reads:

$$\begin{aligned}
\langle\langle\ Z^n \rangle\rangle\ &= \mathrm{Tr}_S \int \prod_{a,b=1}^{n} dq_{ab} d\hat{q}_{ab} d\mu_1^a d\hat{\mu}_1^a d\bar{S}_1^a dg_a \prod_{c,d=1}^{2n} dA_{cd} d\hat{A}_{cd} \\
&\exp\Big\{ -\sum_a \frac{g_a}{2}(D - D\bar{S}_1^a \bar{S}_1^a) \Big\} \exp\Big\{ -\frac{D}{2} \sum_{cd} \hat{A}_{cd} A_{cd} \Big\} \\
&\exp\Big\{ -\frac{p}{2} \ln\det(J) \Big\} \exp\Big\{ -\frac{D}{2} \ln\det(q) \Big\} \\
&\exp\Big\{ -\frac{D}{2} \ln\det\Big(-\hat{A} + \begin{pmatrix} q^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix} \Big) \Big\} \\
&\exp\Big\{ -\frac{N\alpha}{2} \sum_{ab} \hat{q}_{ab} q_{ab} + \frac{\alpha}{2} \sum_{ab} \hat{q}_{ab} \sum_i S_i^a S_i^b \Big\} \\
&\Big\langle \exp\Big\{ -\sum_a \hat{\mu}_1^a \mu_1^a + \frac{1}{N} \sum_a \hat{\mu}_1^a \sum_i f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}}
\end{aligned} \tag{C.13}$$

C.3 Saddle point equation for $\hat{A}$

By looking at equation C.13 we can see how there are only two terms which depend on the conjugate field $\hat{A}$. We can then try to perform a saddle point equation over those term by derivating in $\partial_{\hat{A}}(\langle\langle Z^n\rangle\rangle)$. What we obtain is the following equation:

$$\begin{aligned}\partial_{\hat{A}}(\langle\langle Z^n\rangle\rangle) = &-\frac{D}{2}A^T + \frac{D}{2}\left(-\hat{A} + \begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\right)^{-1} = 0\\ &\implies\\ &\hat{A} = -(A^T)^{-1} + \begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\end{aligned} \tag{C.14}$$

This is the first saddle point equation we derive. We use it by substituting the value of the conjugate field in equation C.13. In this way we removed in the partition function the dependency on the conjugate field. In fact by inserting the result of equation C.14 for $\hat{A}$ in our integral we obtain the following equation:

$$\begin{aligned}\langle\langle\ Z^n\rangle\rangle\ = \mathrm{Tr}_S \int \prod_{a,b=1}^{n} dq_{ab}d\hat{q}_{ab}d\mu_1^a d\hat{\mu}_1^a d\bar{S}_1^a dg_a \prod_{c,d=1}^{2n} dA_{cd}\\ \exp\left\{-\frac{p}{2}\ln\det(J)\right\}\exp\left\{-\frac{D}{2}\ln\det(q)\right\}\exp\left\{\frac{D}{2}\ln\det A\right\}\\ \exp\left\{-\sum_a \frac{g_a}{2}(D - D\bar{S}_1^a\bar{S}_1^a)\right\}\exp\left\{-\frac{N\alpha}{2}\sum_{ab}\hat{q}_{ab}q_{ab} + \frac{\alpha}{2}\sum_{ab}\hat{q}_{ab}\sum_i S_i^a S_i^b\right\}\\ \exp\left\{nD - \frac{D}{2}Tr\left[A^T\begin{pmatrix}q^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\right]\right\}\\ \left\langle\exp\left\{-N\sum_a\hat{\mu}_1^a\mu_1^a + \sum_a\hat{\mu}_1^a\sum_i f_{1i}S_i^a\right\}\right\rangle_{f_{1i}}\end{aligned} \tag{C.15}$$

Where with $Tr[\cdot]$ we mean the trace of the matrix in bracket. In the equation above we also have done the scaling $\hat{\mu}_1^a \to N\hat{\mu}_1^a$. Finally we remind the form of the J matrix:

$$J = -\beta A - \begin{pmatrix}\beta\frac{N}{D}\mu_1^a\mu_1^b & \beta\sqrt{\frac{N}{D}}\mu_1^a\bar{S}_1^b - \mathbb{1}\\ \beta\sqrt{\frac{N}{D}}\mu_1^b\bar{S}_1^a - \mathbb{1} & \beta\bar{S}_1^a\bar{S}_1^b\end{pmatrix} \tag{C.16}$$

Note that to perform all the matrices computation above and in the next paragraph we used the properties listed in Appendix E. Finally the parameter $\hat{A}_{ab}$ is the conjugate parameter of A.
Inserting equation C.14 in equation C.13 and looking only at the terms with $\hat{A}$ we

obtain:

$$\begin{aligned}
&\exp\Big\{-\frac{D}{2}\ln\det\Big(-\hat{A}_{cd}+\begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\Big)-\frac{D}{2}\sum_{cd}\hat{A}_{cd}A_{cd}\Big\}=\\
&\exp\Big\{-\frac{D}{2}\ln\det\Big(-(A^T)^{-1}\Big)-\frac{D}{2}Tr\Big[A^T\Big(-(A^T)^{-1}+\begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\Big)\Big]\Big\}=\\
&\exp\Big\{-\frac{D}{2}\ln\det\Big(-(A^T)^{-1}\Big)+\frac{D}{2}Tr\Big[\mathbb{1}\Big]-\frac{D}{2}Tr\Big[A^T\begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\Big]\Big\}=\\
&\exp\Big\{\frac{D}{2}\ln\det A^T+nD-\frac{D}{2}Tr\Big[A^T\begin{pmatrix}(q_{ab})^{-1} & 0\\ 0 & -g\mathbb{1}\end{pmatrix}\Big]\Big\}
\end{aligned}\tag{C.17}$$

where we used the following properties for the det

$$\det(\mathbb{A}^{-1})=\frac{1}{\det(\mathbb{A})}\tag{C.18}$$

$$\det(c\mathbb{A})=c^n\det(\mathbb{A})\tag{C.19}$$

where (1.66) stands for $\mathbb{A}$ as a $n\times n$ matrix. Let's also notice that

$$\det(\mathbb{A}^T)=\det(\mathbb{A})\tag{C.20}$$

C.4 Replica symmetric ansatz

We now make an RS ansatz for all the order parameters. Reminder that A is a block matrix, where the single blocks are going to be order parameters:

$$A=\begin{pmatrix}P & R\\ R^T & \bar{Q}\end{pmatrix}\tag{C.21}$$

So the RS ansatz for all the order parameters reads as follows:

$$q_{ab}=\delta_{ab}+q(1-\delta_{ab})\tag{C.22}$$

$$\hat{q}_{ab}=\delta_{ab}+\hat{q}(1-\delta_{ab})\tag{C.23}$$

$$\bar{Q}_{ab}=\bar{Q}_{\mathrm{d}}\delta_{ab}+\bar{Q}(1-\delta_{ab})\tag{C.24}$$

$$P_{ab}=P_{\mathrm{d}}\delta_{ab}+P(1-\delta_{ab})\tag{C.25}$$

$$R_{ab}=R_{\mathrm{d}}\delta_{ab}+R(1-\delta_{ab})\tag{C.26}$$

$$\mu_k^a=\mu_k\tag{C.27}$$

$$\hat{\mu}_k^a=\hat{\mu}_k\tag{C.28}$$

$$\bar{S}_1^a=\bar{S}_1\tag{C.29}$$

$$g_a=g\tag{C.30}$$

The RS ansatz allows us to linearise the term S^aS^b. In fact by isolating the terms in the integral which depend only by the spins we can perform the Tr_S by

remembering that $\mathrm{Tr}_S = \sum_{S^1} \cdots \sum_{S^n}$.

$$
\begin{aligned}
&Tr_S \Big\langle \exp\Big\{ \frac{\alpha}{2}\hat{q} \sum_{ab}\sum_{i} S_i^a S_i^b + \sum_a \hat{\mu}_1 \sum_i f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}} = \\
&\Big\langle \prod_i \sum_{S^1} \cdots \sum_{S^n} \exp\Big\{ \frac{\alpha}{2}\hat{q} \sum_{ab} S_i^a S_i^b + \sum_a \hat{\mu}_1 f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}} = \\
&\Big\langle \prod_i \exp\Big\{ -\frac{\alpha}{2}\hat{q}n \Big\} \sum_{S^1} \cdots \sum_{S^n} \exp\Big\{ \frac{\alpha}{2}\hat{q} \Big(\sum_a S_i^a \Big)^2 + \sum_a \hat{\mu}_1 f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}} = \\
&\exp\Big\{ \Big\langle \sum_i \ln\Big(\exp\Big\{ -\frac{\alpha}{2}\hat{q}n \Big\} \\
&\sum_{S^1} \cdots \sum_{S^n} \exp\Big\{ \frac{\alpha}{2}\hat{q} \Big(\sum_a S_i^a \Big)^2 + \sum_a \hat{\mu}_1 f_{1i} S_i^a \Big\} \Big\rangle_{f_{1i}} \Big) \Big\} = \\
&\exp\Big\{ \Big\langle \sum_i \Big(-\frac{\alpha}{2}\hat{q}n + \\
&\ln\Big(\sum_{S^1} \cdots \sum_{S^n} \int Dz \exp\Big\{ (z\sqrt{\alpha\hat{q}} + \hat{\mu}_1 f_{1i}) \sum_a S_i^a \Big\} \Big\rangle_{f_{1i}} \Big) \Big) \Big\} = \\
&\exp\Big\{ \sum_i \Big(-\frac{\alpha}{2}\hat{q}n \ + n \Big\langle \int Dz \ln\Big[2\cosh(z\sqrt{\alpha\hat{q}} + \hat{\mu}_1 f_{1i}) \Big] \Big\rangle_{f_{1i}} \Big) \Big\} = \\
&\exp\Big\{ N \Big(-\frac{\alpha}{2}\hat{q}n \ + n \Big\langle \int Dz \ln\Big[2\cosh(z\sqrt{\alpha\hat{q}} + \hat{\mu}_1 f_{1i}) \Big] \Big\rangle_{f_{1i}} \Big) \Big\}
\end{aligned}
\tag{C.31}
$$

Where with the notation $\int Dz$ we imply the Gaussian integral over the variable z:

$$
\int Dz = \int dz \exp\Big\{ -\frac{z^2}{2} \Big\} \tag{C.32}
$$

Thanks to the RS ansatz we can also compute explicitly other terms from equation C.15: the log of the determinants

$$
\ln\det(J) \tag{C.33}
$$

$$
\ln\det(A) \tag{C.34}
$$

$$
\ln\det(q) \tag{C.35}
$$

with the notions from Appendix D.1 and the formula:

$$
\ln\det \mathbb{X} = n\ln(x_d - x) + n\frac{x}{x_d - x} + O(n^2) \tag{C.36}
$$

Finally the RS ansatz allow us to compute also the $Tr\left[A^T \begin{pmatrix} q^{-1} & 0 \\ 0 & -g\mathbb{1} \end{pmatrix}\right]$ with the notions from Appendix D.2. After all those manipulations our equation reads

$$
\begin{aligned}
<< \ Z^n >> = \int \prod_{a,b=1}^{n} dq_{ab} d\hat{q}_{ab} d\mu_1^a d\hat{\mu}_1^a d\bar{S}_1^a dg_a \prod_{c,d=1}^{2n} dA_{cd} \\
\exp\Bigg\{ -n\frac{p}{2}\Bigg[\ln\Big(\beta^2(P_d-P)(\bar{q}_d-\bar{q}) - (\beta R_d - 1 - \beta R)^2\Big) + \\
\frac{\beta^2(\bar{q}+\bar{S}_1^2)(P_d-P) + \frac{\beta^2}{\alpha_D}(\alpha_D P + \mu_1^2)(\bar{q}_d-\bar{q}) - \frac{2\beta}{\sqrt{\alpha_D}}(\sqrt{\alpha_D}R + \mu_1\bar{S}_1)(\beta R_d - 1 - \beta R)}{\beta^2(P_d-P)(\bar{q}_d-\bar{q}) - (\beta R_d - 1 - \beta R)^2}\Bigg]\Bigg\} \\
\exp\Bigg\{ -n\frac{D}{2}\Bigg[\ln\Big((1-q)\Big) + \frac{q}{1-q}\Bigg] - nN\hat{\mu}_1\mu_1 - \frac{gDn}{2}(1-\bar{S}_1\bar{S}_1) - n(n-1)\frac{N\alpha}{2}\hat{q}q\Bigg\} \\
\exp\Bigg\{n\frac{D}{2}\Bigg[\ln\Big((P_d-P)(\bar{q}_d-\bar{q}) - (R_d-R)^2\Big) + \frac{\bar{q}(P_d-P) + P(\bar{q}_d-\bar{q}) - 2R(R_d-R)^2)}{(P_d-P)(\bar{q}_d-\bar{q}) - (R_d-R)^2}\Bigg]\Bigg\} \\
\exp\Bigg\{n\frac{D}{2} - \frac{D}{2}n\Big(\frac{P_d(1+(n-2)q) - (n-1)Pq}{(1-q)((n-1)q+1)} - g\bar{q}_d\Big)\Bigg\} \\
\exp\Bigg\{Nn\Big(-\frac{\alpha}{2}\hat{q} \ + \Big\langle \int Dz \ln\Big[2\cosh(z\sqrt{\alpha\hat{q}} + \hat{\mu}_1 f_{1i})\Big]\Big\rangle_{f_{1i}}\Big)\Bigg\}
\end{aligned}
\tag{C.37}
$$

The order parameters involved in the RS free energy are

$$
-\beta F_Q(q,\hat{q},\bar{q}_d,\bar{q},\bar{S}_1,P_d,P,R_d,R,\mu_1,\hat{\mu}_1,g) = \lim_{n\to 0}\lim_{L\to\infty}\frac{1}{nL}\ln\Big(<< Z^n >>\Big) \tag{C.38}
$$

Moreover we neglect all the $O(n^2)$ terms and we impose the following β scalings:

$$
\hat{q} \to \beta^2\hat{q} \tag{C.39}
$$

$$
\hat{\mu}_1 \to \beta\hat{\mu}_1 \tag{C.40}
$$

$$
g \to \beta g \tag{C.41}
$$

By plugging in these changes and removing a factor $-\beta$ from both sides we get:

$$
\begin{aligned}
F_Q = &\frac{\alpha_T\sqrt{\alpha_D}}{2\beta}\Bigg[\ln\Big(\beta^2(P_d-P)(\bar{q}_d-\bar{q}) - (\beta R_d - 1 - \beta R)^2\Big) + \\
&\frac{\beta^2(\bar{q}+\bar{S}_1^2)(P_d-P) + \frac{\beta^2}{\alpha_D}(\alpha_D P + \mu_1^2)(\bar{q}_d-\bar{q}) - \frac{2\beta}{\sqrt{\alpha_D}}(\sqrt{\alpha_D}R + \mu_1\bar{S}_1)(\beta R_d - 1 - \beta R)}{\beta^2(P_d-P)(\bar{q}_d-\bar{q}) - (\beta R_d - 1 - \beta R)^2}\Bigg] + \\
&\frac{\sqrt{\alpha_D}}{2\beta}\Bigg[\ln\Big(\frac{1-q}{(P_d-P)(\bar{q}_d-\bar{q}) - (R_d-R)^2}\Big)\Bigg] + \\
&+\frac{\sqrt{\alpha_D}}{2\beta}\Bigg[\frac{q}{1-q} + \frac{\bar{q}(P_d-P) + P(\bar{q}_d-\bar{q}) - 2R(R_d-R)}{(P_d-P)(\bar{q}_d-\bar{q}) - (R_d-R)^2} - 1\Bigg] + \\
&+\frac{\sqrt{\alpha_D}}{2\beta}\Big(\frac{P_d - 2qP_d + Pq}{(1-q)(1-q)}\Big) + \frac{1}{\sqrt{\alpha_D}}\hat{\mu}_1\mu_1 + \frac{g\sqrt{\alpha_D}}{2}(1-\bar{q}_d-\bar{S}_1\bar{S}_1) - \frac{\alpha\beta}{2\sqrt{\alpha_D}}\hat{q}(q+1) + \\
&-\frac{1}{\beta\sqrt{\alpha_D}}\Big\langle \int Dz \ln\Big[2\cosh(\beta(z\sqrt{\alpha\hat{q}} + \hat{\mu}_1 f_{1i}))\Big]\Big\rangle_{f_{1i}}
\end{aligned}
\tag{C.42}
$$

Appendix D

Algebra of RS matrixes

D.1 Block matrix of RS matrixes

Here we try to compute the $\ln\det G$ for a generic block matrix G:

$$G = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{D.1}$$

Where A, B, C, D are all RS matrixes $n \times n$.
Reminder that all RS matrixes have a defined structure:

$$A_{ab} = A_d\delta_{ab} + A(1 - \delta_{ab}) \tag{D.2}$$

in matrix form:

$$A = \begin{pmatrix} A_d & A & \dots & A \\ A & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A \\ A & \dots & A & A_d \end{pmatrix}_{n\times n} \tag{D.3}$$

Moreover they all have 2 different eigenvalues: $\lambda_1 = A_d - A$ with multiplicity $n-1$ and $\lambda_2 = (n-1)A + A_d$ with multiplicity 1. Using the program Mathematica we were able to find 4 different eigenvalues fot the generic block matrix G:

$$G_1 = \frac{1}{2}\Big((n-1)A + A_d + (n-1)D + D_d - \sqrt{((n-1)A + A_d - (n-1)D - D_d)^2 + 4((n-1)C + C_d)((n-1)B + B_d)}\Big) \tag{D.4}$$

$$G_2 = \frac{1}{2}\Big((n-1)A + A_d + (n-1)D + D_d + \sqrt{((n-1)A + A_d - (n-1)D - D_d)^2 + 4((n-1)C + C_d)((n-1)B + B_d)}\Big) \tag{D.5}$$

both with multiplicity 1, and

$$G_3 = \frac{1}{2}\Big(A_d - A + D_d - D\ - \sqrt{(A_d - A + D - D_d)^2 + 4(C_d - C)(B_d - B)}\Big) \tag{D.6}$$

$$G_4 = \frac{1}{2}\Big(A_d - A + D_d - D \ + \sqrt{(A_d - A + D - D_d)^2 + 4(C_d - C)(B_d - B)}\Big) \quad \text{(D.7)}$$

both with multiplicity $n-1$.
To simplify the following steps let's define in a compact way the eigenvalues of the RS matrixes, for example for the A matrix:

$$\begin{aligned} A_1 &\equiv A_d - A \\ A_2 &\equiv (n-1)A + A_d = nA + A_1 \end{aligned} \quad \text{(D.8)}$$

The eigenvalues of the other matrixes will follow the same compact definition. We can then rewrite the eigenvalues of G as

$$\begin{aligned} G_1 &= \frac{1}{2}\Big(A_2 + D_2 - \sqrt{(A_2 - D_2)^2 + 4B_2C_2}\Big) \\ G_2 &= \frac{1}{2}\Big(A_2 + D_2 + \sqrt{(A_2 - D_2)^2 + 4B_2C_2}\Big) \\ G_3 &= \frac{1}{2}\Big(A_1 + D_1 - \sqrt{(A_1 - D_1)^2 + 4B_1C_1}\Big) \\ G_4 &= \frac{1}{2}\Big(A_1 + D_1 + \sqrt{(A_1 - D_1)^2 + 4B_1C_1}\Big) \end{aligned} \quad \text{(D.9)}$$

To further simplify the computation we are going to introduce also the following variables:

$$X = A + D \quad \text{(D.10)}$$
$$Y = A - D \quad \text{(D.11)}$$
$$W = B + C \quad \text{(D.12)}$$
$$Z = B - C \quad \text{(D.13)}$$

Even these are RS matrixes, in fact RS matrixes are a closed group concerning the operation of sum, so the previous definition of the eigenvalues for the RS matrixes are effective for them too. for example:

$$\begin{aligned} X_1 &\equiv A_1 + D_1 \\ X_2 &\equiv A_2 + D_2 = nX + X_1 \end{aligned} \quad \text{(D.14)}$$

With these new variables we can write the eigenvalues of G as

$$\begin{aligned} G_1 &= \frac{1}{2}\Big(X_2 - \sqrt{Y_2^2 + W_2^2 - Z_2^2}\Big) \\ G_2 &= \frac{1}{2}\Big(X_2 + \sqrt{Y_2^2 + W_2^2 - Z_2^2}\Big) \\ G_3 &= \frac{1}{2}\Big(X_1 - \sqrt{Y_1^2 + W_1^2 - Z_1^2}\Big) \\ G_4 &= \frac{1}{2}\Big(X_1 + \sqrt{Y_1^2 + W_1^2 - Z_1^2}\Big) \end{aligned} \quad \text{(D.15)}$$

We are now ready to compute the $\ln\det G$

$$\ln\det G = \ln(G_1) + \ln(G_2) + (n-1)\ln(G_3) + (n-1)\ln(G_4) \tag{D.16}$$

We want to end with an equation of $O(n)$, and we recall that $n \to 0$. In this limit the following Taylor expansions hold:

$$\begin{aligned} \sqrt{1+n} &= 1 + \frac{n}{2} \\ \ln(1+n) &= n \end{aligned} \tag{D.17}$$

let's add another definition to smooth the notation:

$$Y_1^2 + W_1^2 - Z_1^2 = K_1^2 \tag{D.18}$$

with these assumption let's compute the first and third term of equation D.16 (the other two are the same as the one computed except for a sign). We have then for the G_1

$$\begin{aligned} &\ln(G_1) = \ln\left(\frac{1}{2}\left(X_2 - \sqrt{Y_2^2 + W_2^2 - Z_2^2}\right)\right) = \\ &\ln\left(\frac{1}{2}\left(nX + X_1 - \sqrt{(nY+Y_1)^2 + (nW+W_1)^2 - (nZ+Z_1)^2}\right)\right) = \\ &\ln\left(\frac{1}{2}\left(nX + X_1 - \sqrt{K_1^2 + 2n(YY_1 + WW_1 - ZZ_1) + O(n^2)}\right)\right) = \\ &\ln\left(\frac{1}{2}\left(nX + X_1 - K_1\left(1 + n\frac{YY_1 + WW_1 - ZZ_1}{K_1^2}\right)\right)\right) = \\ &\ln\left(\frac{1}{2K_1}\left(nXK_1 + X_1K_1 - K_1^2 - n(YY_1 + WW_1 - ZZ_1)\right)\right) = \\ &\ln\left(\frac{X_1 - K_1}{2}\left(1 + n\frac{XK_1 - YY_1 - WW_1 + ZZ_1}{X_1K_1 - K_1^2}\right)\right) = \\ &\ln\left(\frac{X_1 - K_1}{2}\right) + n\frac{XK_1 - YY_1 - WW_1 + ZZ_1}{X_1K_1 - K_1^2} \end{aligned} \tag{D.19}$$

and for the G_3 term

$$\begin{aligned} &(n-1)\ln(G_3) = (n-1)\ln\left(\frac{1}{2}\left(X_1 - \sqrt{Y_1^2 + W_1^2 - Z_1^2}\right)\right) = \\ &(n-1)\ln\left(\frac{1}{2}\left(X_1 - K_1\right)\right) = n\ln\left(\frac{X_1 - K_1}{2}\right) - \ln\left(\frac{X_1 - K_1}{2}\right) \end{aligned} \tag{D.20}$$

and equivalently for the G_2 term and the G_3 term we have

$$\ln(G_2) = \ln\left(\frac{X_1 + K_1}{2}\right) + n\frac{XK_1 + YY_1 + WW_1 - ZZ_1}{X_1K_1 + K_1^2} \tag{D.21}$$

$$(n-1)\ln(G_4) = n\ln\left(\frac{X_1 + K_1}{2}\right) - \ln\left(\frac{X_1 + K_1}{2}\right) \tag{D.22}$$

So it is clear that the $O(1)$ terms cancel each other when performing the sum. Finally we can conclude that $\ln\det G$ at the $O(n)$ in the limit $n \to 0$ reads

$$\begin{aligned}\ln\det G =& n\ln\left(\frac{X_1+K_1}{2}\right) + n\ln\left(\frac{X_1-K_1}{2}\right)+ \\ & n\frac{XK_1 - YY_1 - WW_1 + ZZ_1}{X_1K_1 - K_1^2}+ \\ & n\frac{XK_1 + YY_1 + WW_1 - ZZ_1}{X_1K_1 + K_1^2}\end{aligned} \tag{D.23}$$

where

$$K_1^2 = Y_1^2 + W_1^2 - Z_1^2 \tag{D.24}$$
$$X = A + D \tag{D.25}$$
$$X_1 = A_d + A + D_d + D \tag{D.26}$$
$$Y = A - D \tag{D.27}$$
$$Y_1 = A_d + A - D_d - D \tag{D.28}$$
$$W = B + C \tag{D.29}$$
$$W_1 = B_d + B + C_d + C \tag{D.30}$$
$$Z = B - C \tag{D.31}$$
$$Z_1 = B_d + B - C_d - C \tag{D.32}$$

Moreover by replacing those terms in equation D.23 and performing some simple calculations we obtain:

$$\begin{aligned}\ln\det G =& n\ln\Big((A_d - A)(D_d - D) - (B_d - B)(C_d - C)\Big)+ \\ & n\frac{D(A_d - A) + A(D_d - D) - B(C_d - C) - C(B_d - B)}{(A_d - A)(D_d - D) - (B_d - B)(C_d - C)}\end{aligned} \tag{D.33}$$

Which is a generic formula to compute the $\ln\det G$ when G is a $2n \times 2n$ block matrix composed of 4 $n \times n$ RS matrixes.
Note that the same result can be obtained by noticing that

$$\det G = \det(AD - BC) \tag{D.34}$$

and that the RS matrixes are a closed group concerning product and sum, so the matrix $AD - BC$ has two different eigenvalues, founded with Mathematica, one with multiplicity $n - 1$ and the other one with multiplicity 1, respectively

$$(AD - BC)_1 = (A_d - A)(D_d - D) - (B_d - B)(C_d - C) \tag{D.35}$$
$$(AD - BC)_2 = ((n-1)A + A_d)((n-1)D + D_d)+ \tag{D.36}$$
$$-((n-1)C + C_d))((n-1)B + B_d)) \tag{D.37}$$

And then by applying the following formula

$$\ln\det X_{ab} = n\ln(x_d - x) + n\frac{x}{x_d - x} + O(n^2) \tag{D.38}$$

Where X_{ab} is a $n \times n$ RS matrix, in our case $AD - BC$, and as usual x_d and x are it's diagonal and off diagonal terms.

D.2 Inverse of a RS matrix

Here we want to prove that the inverse of a RS matrix is still a RS matrix.
Let's take the $n \times n$ RS matrix V. As all other RS matrixes it shares the following structure:

$$V_{ab} = V_d \delta_{ab} + V(1 - \delta_{ab}) \tag{D.39}$$

wich in matrix form presents itself as:

$$V = \begin{pmatrix} V_d & V & \dots & V \\ V & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & V \\ V & \dots & V & V_d \end{pmatrix}_{n \times n} \tag{D.40}$$

Moreover it also shares with other RS matrixes the following eigenvalues structure, recalling the previous paragraph notation,
$V_1 = V_d - V$ with multiplicity $n - 1$ and $V_2 = (n-1)V + V_d$ with multiplicity 1.
By performing the inverse of the matrix V with mathematica we obtain the following matrix

$$V^{-1} = \frac{1}{(V_d - V)((n-1)V + V_d)} \begin{pmatrix} (n-2)V + V_d & -V & \dots & -V \\ -V & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -V \\ -V & \dots & -V & (n-2)V + V_d \end{pmatrix}_{n \times n} \tag{D.41}$$

with eigenvalues:

$$(V^{-1})_1 = \frac{1}{V_d - V} \tag{D.42}$$

$$(V^{-1})_2 = \frac{1}{(n-1)V + V_d} \tag{D.43}$$

respectively with multiplicity $n - 1$ and 1.
This is indeed still a RS matrix, in fact we can say $V^{-1} = \Lambda$ and we can show that the matrix Λ has the same construct as the others RS matrixes

$$\Lambda = \begin{pmatrix} \Lambda_d & \Lambda & \dots & \Lambda \\ \Lambda & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Lambda \\ \Lambda & \dots & \Lambda & \Lambda_d \end{pmatrix}_{n \times n} \tag{D.44}$$

where

$$\Lambda_d = \frac{(n-2)V + V_d}{(V_d - V)((n-1)V + V_d)} \tag{D.45}$$

$$\Lambda = \frac{-V}{(V_d - V)((n-1)V + V_d)} \tag{D.46}$$

And the matrix Λ has the following eigenvalues

$$\Lambda_1 = \Lambda_d - \Lambda = \frac{(n-2)V + V_d + V}{(V_d - V)((n-1)V + V_d)} = \frac{1}{V_d - V} \tag{D.47}$$

$$\Lambda_2 = (n-1)\Lambda + \Lambda_d = \frac{(n-2)V + V_d - (n-1)V}{(V_d - V)((n-1)V + V_d)} = \frac{1}{(n-1)V + V_d} \tag{D.48}$$

respectively with multiplicity $n-1$ and 1.

Appendix E

Matrix calculus

To perform the matrix computation in the subsection where we evaluate the saddle point equation for $\hat{A}$ and A we needed the following definitions, properties and formulas of the matrix calculus:

- Definition of adjugate Matrix: the adjugate of a square matrix $\mathbb{A}$ is the transpose of its cofactor matrix, $\mathbb{C}$, and is denoted by $adj(\mathbb{A})$.In more detail, suppose $\mathbb{R}$ is a unital commutative ring and $\mathbb{A}$ is an $n \times n$ matrix with entries from $\mathbb{R}$. The (i, j) minor of $\mathbb{A}$, denoted M_{ij}, is the determinant of the $(n-1) \times (n-1)$ matrix that results from deleting row i and column j of $\mathbb{A}$. The cofactor matrix of $\mathbb{A}$ is the $n \times n$ matrix $\mathbb{C}$ whose (i, j) entry is the (i, j) cofactor of $\mathbb{A}$, which is the (i, j) minor times a sign factor:

$$\mathbb{C} = ((-1)^{i+j} M_{ij})_{1 \leq i,j \leq n} \tag{E.1}$$

  The adjugate of $\mathbb{A}$ is the transpose of $\mathbb{C}$, that is, the $n \times n$ matrix whose (i, j) entry is the (j, i) cofactor of $\mathbb{A}$

$$adj(\mathbb{A}) = \mathbb{C}^T = ((-1)^{i+j} M_{ji})_{1 \leq i,j \leq n} \tag{E.2}$$

- There is a property of the $adj(\mathbb{A})$ that links it with the $det(\mathbb{A})$, namely:

$$adj(\mathbb{A}) = det(\mathbb{A})\mathbb{A}^{-1} \tag{E.3}$$

- For two $n \times n$ matrices, the following equality hold:

$$\sum_{ab} \mathbb{A}_{ab} \mathbb{X}_{ab} = Tr\left[\mathbb{A}_{ab}^T \mathbb{X}_{ab}\right] = Tr\left[\mathbb{A}_{ab} \mathbb{X}_{ab}^T\right] \tag{E.4}$$

- Jacobi's Formula:

$$d(\det(\mathbb{A})) = Tr(adj(\mathbb{A}) dA) \tag{E.5}$$

  Moreover, by using the following conversion from differential to derivative form:

$$dy = Tr(\mathbb{A} d\mathbb{X}) \longrightarrow \frac{dy}{d\mathbb{X}} = \mathbb{A} \tag{E.6}$$

  where y is a scalar, we can write the Jacobi's Formula as:

$$\frac{d(\det(\mathbb{A}))}{d\mathbb{A}} = adj(\mathbb{A}) \tag{E.7}$$

- For a function g of a scalar u that depends on a matrix $\mathbb{X}$ the following equation holds:

$$\partial_{\mathbb{X}} g(u) = \partial_u(g)\partial_{\mathbb{X}}(u) \tag{E.8}$$

- For two matrices $\mathbb{A}$ and $\mathbb{X}$ the following formula holds:

$$\partial_{\mathbb{X}} Tr(\mathbb{A}^T\mathbb{X}) = \mathbb{A}^T \tag{E.9}$$

Bibliography

[1] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[2] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

[3] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

[4] DJ Amit and Modelling Brain Function. The world of attractor neural networks. *Modeling Brain Function*, 1989.

[5] Carlo Baldassi, Clarissa Lauditi, Enrico M Malatesta, Rosalba Pacelli, Gabriele Perugini, and Riccardo Zecchina. Learning through atypical phase transitions in overparameterized neural networks. *Physical Review E*, 106(1):014116, 2022.

[6] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1–9, 2012.

[7] Adriano Barra, Giovanni Catania, Aurélien Decelle, and Beatriz Seoane. Thermodynamics of bidirectional associative memories. *Journal of Physics A: Mathematical and Theoretical*, 56(20):205005, 2023.

[8] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Spectral dynamics of learning in restricted boltzmann machines. *Europhysics Letters*, 119(6):60001, 2017.

[9] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Thermodynamics of restricted boltzmann machines and related learning dynamics. *Journal of Statistical Physics*, 172:1576–1608, 2018.

[10] Aurélien Decelle and Cyril Furtlehner. Exact training of restricted boltzmann machines on intrinsically low dimensional data. *Physical Review Letters*, 127(15):158303, 2021.

[11] H Englisch, V Mastropietro, and Benedetto Tirozzi. The bam storage capacity. *Journal de Physique I*, 5(1):85–96, 1995.

[12] E Gardner. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, 20(11):3453, 1987.

[13] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

[14] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[15] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks. *stat*, 1050:25, 2019.

[16] Gavin S Hartnett, Edward Parker, and Edward Geist. Replica symmetry breaking in bipartite spin glasses and neural networks. *Physical Review E*, 98(2):022116, 2018.

[17] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

[18] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[19] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[20] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[21] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[22] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[23] Jaron Kent-Dobias and Jorge Kurchan. How to count in hierarchical landscapes: A full solution to mean-field complexity. *Physical Review E*, 107(6):064111, 2023.

[24] B. Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):49–60, 1988.

[25] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

[26] J Kurchan, L Peliti, and M Saber. A statistical investigation of bidirectional associative memories (bam). *Journal de Physique I*, 4(11):1627–1639, 1994.

[27] Francesca Elisa Leonelli, Elena Agliari, Linda Albanese, and Adriano Barra. On the effective initialisation for restricted boltzmann machines via duality with hopfield model. *Neural Networks*, 143:314–326, 2021.

[28] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[29] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

[30] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[31] M. Negri, C. Lauditi, G. Perugini, C. Lucibello, and E. Malatesta. Storage and learning phase transitions in the random-features Hopfield model. *Physical Review Letters*, 131(25):257301, December 2023.

[32] Alejandro Pozas-Kerstjens, Gorka Muñoz-Gil, Eloy Piñol, Miguel Ángel García-March, Antonio Acín, Maciej Lewenstein, and Przemysław R Grzybowski. Efficient training of energy-based models via spin-glass control. *Machine Learning: Science and Technology*, 2(2):025026, 2021.

[33] Kai Shimagaki and Martin Weigt. Selection of sequence motifs and generative hopfield-potts models for protein families. *Physical Review E*, 100(3):032128, 2019.

[34] Paul Smolensky et al. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[35] N Sourlas. Multilayer neural networks for hierarchical patterns. *Europhysics Letters*, 7(8):749, 1988.

[36] Julia Steinberg and Haim Sompolinsky. Associative memory of structured knowledge. *Scientific Reports*, 12(1):21808, 2022.

[37] Jérôme Tubiana and Rémi Monasson. Emergence of compositional representations in restricted boltzmann machines. *Physical review letters*, 118(13):138301, 2017.