Journal of Physics A:
Mathematical and
Theoretical

PURPOSE-LED
PUBLISHING™

PAPER • **OPEN ACCESS**

Replica symmetry breaking in supervised and unsupervised Hebbian networks

To cite this article: Linda Albanese *et al* 2024 *J. Phys. A: Math. Theor.* **57** 165003

View the article online for updates and enhancements.

You may also like

- Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications
Thibault Lesieur, Florent Krzakala and Lenka Zdeborová
- Spin glass field theory with replica Fourier transforms
I R Pimentel and C De Dominicis
- Statistical mechanics of complex neural systems and high dimensional data
Madhu Advani, Subhaneil Lahiri and Surya Ganguli

This content was downloaded from IP address 212.189.140.4 on 20/10/2026 at 11:42

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **57** (2024) 165003 (44pp)
https://doi.org/10.1088/1751-8121/ad38b4

Replica symmetry breaking in supervised and unsupervised Hebbian networks

Linda Albanese[1,2,3,5,*], **Andrea Alessandrelli**[4,5], **Alessia Annibale**[2] **and Adriano Barra**[6]

[1] Dipartimento di Matematica e Fisica 'Ennio De Giorgi', Università del Salento, Lecce, Italy
[2] Department of Mathematics, King's College London, The Strand, London, United Kingdom
[3] Scuola Superiore ISUFI, Università del Salento, Lecce, Italy
[4] Dipartimento di Informatica, Università di Pisa, Pisa, Italy
[5] Istituto Nazionale di Fisica Nucleare, Sezione di Lecce, Italy
[6] Dipartimento di di Scienze di Base e Applicate per l'Ingegneria, Sapienza Università di Roma, Roma, Italy

E-mail: linda.albanese@unisalento.it

Received 18 December 2023; revised 4 March 2024
Accepted for publication 26 March 2024
Published 9 April 2024



Abstract

Hebbian neural networks with multi-node interactions, often called Dense Associative Memories, have recently attracted considerable interest in the statistical mechanics community, as they have been shown to outperform their pairwise counterparts in a number of features, including resilience against adversarial attacks, pattern retrieval with extremely weak signals and supralinear storage capacities. However, their analysis has so far been carried out within a replica-symmetric theory. In this manuscript, we relax the assumption of replica symmetry and analyse these systems at one step of replica-symmetry breaking, focusing on two different prescriptions for the interactions that we will refer to as supervised and unsupervised learning. We derive the phase diagram of the model using two different approaches, namely Parisi's hierarchical ansatz for the relationship between different replicas within the replica approach, and the so-called telescope ansatz within Guerra's interpolation

* Author to whom any correspondence should be addressed.



Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2024 The Author(s). Published by IOP Publishing Ltd

method: our results show that replica-symmetry breaking does not alter the threshold for learning and slightly increases the maximal storage capacity. Further, we also derive analytically the instability line of the replica-symmetric theory, using a generalization of the De Almeida and Thouless approach.

Keywords: dense associative memory, replica symmetry breaking, Guerra's interpolation, replica trick, unsupervised learning, supervised learning

Contents

1. Introduction 2
2. 'Unsupervised' DAM 4
2.1. Model and definitions 4
2.2. 1RSB analysis via Guerra's interpolation technique 8
2.3. 1RSB analysis via Parisi's replica trick 13
2.4. Results 19
2.4.1. Limiting cases: $M \to \infty$ and $\beta \to \infty$. 21
3. 'Supervised' DAM 24
3.1. Model and definitions 24
3.2. 1RSB analysis via Guerra's interpolation technique 26
3.3. 1RSB analysis via Parisi's replica trick 27
3.4. Limiting cases: $M \to \infty$ and $\beta \to \infty$ 29
4. Conclusions and outlooks 32
Data availability statement 32
Acknowledgments 32
Appendix A. Instability of the RS solution: AT lines 33
A.1. AT line for DAM in unsupervised setting 33
A.2. AT line for DAM in supervised setting 37
Appendix B. Proofs 39
B.1. Proof of lemma 1 39
B.2. Proof of corollary 2 41
References 42

1. Introduction

Since Hopfield's seminal work on the use of biologically inspired neural networks for associative memory and pattern recognition tasks [1], there have been many important contributions applying concepts from statistical mechanics, in particular spin-glass theory, to the study of neural networks with pairwise Hebbian interactions (for an overview of the vast literature, the reader is referred to the books [2–9] and to the recent reviews [10, 11]). This connection was first pointed out by Hopfield himself in [1], where he showed that neural networks with pairwise Hebbian interactions were particular realizations of spin glasses, and it was further developed by Amit *et al* [12] who showed that the free-energy landscape of such systems is characterized by a large number of local minima, corresponding to different patterns of information, retrieved by the network from different initial conditions. In addition to neural networks

with pairwise Hebbian interactions, a large body of work has focused on extending Hebbian learning to multi-node interactions, both in the early days (see e.g. [13–15]) and more recently (see e.g. [16, 17]). These models are often referred to as *dense* associative memories [16], as they can store many more patterns than the number of neurons in the network. In particular, they can perform pattern recognition with a supra-linear storage load [15] and can work at very low signal-to-noise ratio, when compared to their pairwise counterparts [18]. In addition, they have been shown to be resilient to adversarial attacks [19] and combinations of such models can lead to exponential storage capacity [20].

In recent years, a duality between Hebbian neural networks and machine learning models has been pointed out. For example, the Hopfield model has been shown to be equivalent to a restricted Boltzmann machine [21], an archetypical model for machine learning [22], and sparse restricted Boltzmann machines have been mapped to Hopfield models with diluted patterns [23, 24]. Furthermore, restricted Boltzmann machines with generic priors have led to the definition of generalized Hopfield models [25–27] and neural networks with multi-node Hebbian interactions have recently been shown to be equivalent to higher-order Boltzmann machines [28, 29] and deep Boltzmann machines [30, 31]. As a result, multi-node Hebbian learning is receiving a second wave of interest since its foundation in the eighties [13, 14] as a paradigm to understand deep learning [17, 32].

In order to make this connection clearer, we should stress that neither the Hopfield model nor its generalized versions with multi-node Hebbian interactions, are machine learning models, in that they are not trained from data and their couplings do not evolve in time according to a learning rule. Instead, their couplings are fixed from the outset to the values that *would have been attained* by a neural network trained according to the Hebbian rule, which adjusts the strengths of the connections between the neurons based on the input patterns that the network is exposed to. In this work, we consider two different routes for exposing the network to data, as considered in [33, 34]. In particular, we analyze the scenario where the input patterns are noisy or corrupted versions of pattern archetypes. In the first route, that we will call 'unsupervised', we expose the network to all of the input patterns together, without specifying the archetype that each pattern is meant to represent [35]. In the second route, that we will call 'supervised', we split the training dataset in different classes, corresponding to the archetypes, and the network is exposed to the examples in each class, class by class, in the same way that a teacher helps a student to learn one topic at time, from different examples [36]. Beyond being more realistic than traditional Hebbian storing, this generalization allows to pose a number of new questions, such as: is the network able, in either scenario, to retrieve the archetypes by itself, i.e. find hidden and intrinsic structures in the training dataset? which is the minimal amount of examples per archetype that the network has to experience, given the noise in the examples and the number of archetypes, in the two scenarios?

These questions have been addressed in neural networks with pairwise as well as multinode Hebbian interactions, for both random uncorrelated datasets and structured datasets (including the well-known MNist or Fashion MNist) [34, 37, 38], within a replica-symmetric (RS) analysis, which is based on the assumption that different replicas (or copies) of the systems are invariant under permutations. However, since neural networks are particular realization of spin-glasses, such symmetry is expected to be broken in certain regions of the parameters space, where the system develops a multitude of degenerate states. In a recent work [39], we devised a simple and rigorous method to detect the onset of the instability of RS theories in neural networks with pairwise and multinode Hebbian interactions. Building on this work, we derive the RS instability lines for the two different protocols defined here, supervised and

unsupervised, and we re-investigate the problem defined above, previously analysed within an RS assumption, by carrying out the analysis at one step of replica-symmetry breaking (1RSB). Since Parisi's seminal work on replica symmetry breaking [2], several alternative mathematical techniques have been proposed to investigate replica-symmetry broken phases [40–42]. In this work, we will use interpolations techniques, that were pioneered by Guerra in the context of his work on spin glasses [43] and were later applied to neural networks (e.g. [44]). One advantage of this approach is that it avoids certain heuristics required by the replica method and it leads to simpler calculations. For completeness, we will also derive the same results using the replica approach, with the pedagogical aim of creating a bridge between the two methods, one being a cornerstone of the statistical physics community (Parisi's), the other constituting a golden niche in the mathematical physics community (Guerra's).

The paper is structured as follows. In section 2 we define a first model for dense associative memories (DAM) that we will shall refer to as 'unsupervised' and we analyse it at one step of replica symmetry breaking via Guerra's interpolation method and replica techniques. In section 3 we define a second model for DAM, that we will shall refer to as 'supervised' and we carry out the same analysis as in section 2. Finally, in section 4 we summarize and discuss our results. We relegate technical details to the Appendices as well as the derivation of the instability line of the RS theory, which shows the importance of analysing these systems under the assumption of replica symmetry breaking.

2. 'Unsupervised' DAM

In this section we analyse the information processing capabilities of DAM in the so-called 'unsupervised' setting, as introduced in section 2.1, via two different approaches, namely Guerra's interpolation techniques (section 2.2) and Parisi's replica approach (section 2.3). Both analyses are carried out at the first step of RSB. Results are discussed in section 2.4. The instability line of the RS theory (equivalent to the de Almeida and Thouless line in the context of spin-glasses), showing the importance of working under the assumption of broken replica symmetry, is derived in appendix A.1.

2.1. Model and definitions

We consider a system of N neurons, modelled via Ising spins $\sigma_i \in \{-1,1\}$, with $i = 1,\dots,N$, where neurons interact via P-node interactions. We assume to have K Rademacher archetypes, defined as N-dimensional vectors $\boldsymbol{\xi}^\mu$, with $\mu = 1,\dots,K$, where each entry is drawn randomly and independently from the distribution

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

Such archetypes are not provided to the network, instead the network will have to infer them by experiencing solely noisy or corrupted versions of them. In particular, we assume that for each archetype μ, M examples $\boldsymbol{\eta}^{\mu,a}$, with $a = 1,\dots,M$, are available, which are corrupted versions of the archetype, such that

$$\mathbb{P}(\eta_i^{\mu,a}|\xi_i^\mu) = \frac{1-r}{2}\delta_{\eta_i^{\mu,a},-\xi_i^\mu} + \frac{1+r}{2}\delta_{\eta_i^{\mu,a},+\xi_i^\mu}, \tag{2.2}$$

where $r \in [0,1]$ controls the *quality* of the dataset, i.e. for $r=1$ the example matches perfectly the archetype, while for $r=0$ it is purely stochastic. To quantify the information content of the dataset it is useful to introduce the variable

$$\rho = \frac{1-r^2}{Mr^2} \tag{2.3}$$

that we shall refer to as the *dataset entropy*[7]. We note that ρ vanishes either when the examples are identical to the archetypes (i.e. $r \to 1$) or when the number of examples is infinite (i.e. $M \to \infty$), or both. If we regard the traditional DAM model as storing patterns and the unsupervised DAM model as learning patterns from stored examples, then ρ can be thought of as the parameter that quantifies the difference between storing and learning (which is larger, the larger ρ).

Definition 1. The cost function (or *Hamiltonian*) of the 'unsupervised' DAM model is

$$\mathcal{H}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta}) = -\frac{1}{\mathcal{R}^{P/2} M N^{P-1}} \sum_{\mu=1}^{K} \sum_{a=1}^{M} \left(\sum_{i_1<\cdots<i_P}^{N,\cdots,N} \eta_{i_1}^{\mu,a} \cdots \eta_{i_P}^{\mu,a} \sigma_{i_1} \cdots \sigma_{i_P} \right), \tag{2.4}$$

where the constant $\mathcal{R}^{P/2}$, with $\mathcal{R} = r^2(1+\rho)$ and ρ defined in (2.3), is included in the definition for mathematical convenience, while the factor N^{P-1} ensures that the Hamiltonian is extensive in the network size N. On the left hand side (LHS), the label (P) denotes the order of the multi-node interactions and $\boldsymbol{\eta}$ is a short-hand for the collection of all the examples, for all the archetypes $\{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,a=1}^{K,M}$.

Remark 1. For $P=2$, i.e. pairwise interactions, the unsupervised DAM model reduces to the Hopfield neural network in the unsupervised setting [33]. In addition, in the absence of dataset corruption, i.e. $r=1, \mathcal{R}=1$ and the model reduces to the standard Hopfield model, as analysed in [4, 12].

Definition 2. The partition function associated to the Hamiltonian (2.4) at inverse noise level $\beta \in \mathcal{R}^+$ is defined as

$$\mathcal{Z}_N^{(P)}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\sigma}}^{2^N} \exp\left(-\beta \mathcal{H}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta})\right) =: \sum_{\boldsymbol{\sigma}}^{2^N} \mathcal{B}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta}), \tag{2.5}$$

where $\mathcal{B}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta})$ is referred to as the Boltzmann factor. At finite network size N, the free energy of the model $\mathcal{F}_N^{(P)}$ is given by

$$-\beta \mathcal{F}_N^{(P)} = \frac{1}{N} \mathbb{E} \log \mathcal{Z}_N^{(P)}(\boldsymbol{\eta}) \tag{2.6}$$

where $\mathbb{E}[.]$ denotes the average over the realizations of examples $\boldsymbol{\eta}$, regarded as quenched.

We are interested in the behaviour of the system in the limit of large system size $N \to \infty$ and finite network load α, as specified by the following

[7] It was shown in [34] that the conditional entropy which quantifies the amount of information needed to reconstruct the original pattern $\boldsymbol{\xi}^\mu$ given the set of related examples $\{\eta^{\mu,a}\}_{a=1,\ldots,M}$, is a monotonically increasing function of ρ.

Definition 3. In the thermodynamic limit $N \to \infty$, the load of the unsupervised DAM model is defined as

$$\lim_{N\to+\infty} \frac{K}{N^{P-1}} =: \alpha < \infty \tag{2.7}$$

We will mostly focus on the so-called 'saturated' regime, where $\alpha > 0$. The regime where the system is away from saturation can be inspected by taking the limit $\alpha \to 0$. For convenience, we also introduce the parameter γ, defined by

$$\alpha = \gamma \frac{2}{P!}. \tag{2.8}$$

The free energy in the thermodynamic limit is denoted as

$$\mathcal{F}^{(P)} = \lim_{N\to\infty} \mathcal{F}_N^{(P)}. \tag{2.9}$$

In the following, we focus on the ability of the network to retrieve a single archetype, say ν. Given the invariance of the Hamiltonian w.r.t. permutations of the archetypes, we will set without loss of generality $\nu = 1$. Then, following standard procedures [4], we split the sum over μ in the Hamiltonian into the contribution from $\mu = 1$ (regarded as the signal) and the contributions from the other archetypes $\mu > 1$ (regarded as slow noise). Furthermore, we add in the argument of the exponent of the Boltzmann factor a term $J\sum_i \xi_i^1 \sigma_i$, so that the free energy can serve as a moment generating functional of the so-called Mattis magnetization m_1 (*vide infra*) by taking derivatives of $\ln \mathcal{Z}_N^{(P)}$ w.r.t. J. As this term is not part of the original Hamiltonian, J will be set to zero later on.

Starting from (2.5), the resulting partition function is

$$\mathcal{Z}_N^{(P)}(\boldsymbol{\eta}) = \lim_{J\to 0} \sum_{\boldsymbol{\sigma}} \exp\left[J\sum_{i=1}^{N} \xi_i^1 \sigma_i + \frac{\beta'}{2\mathcal{R}^{P/2} M N^{P-1}} \sum_{a=1}^{M} \left(\sum_{i=1}^{N} \eta_i^{a,1} \sigma_i \right)^P \right.$$
$$\left. + \frac{\beta' P!}{2\mathcal{R}^{P/2} M N^{P-1}} \sum_{\mu>1}^{K} \sum_{a=1}^{M} \sum_{i_1<\cdots<i_P}^{N,\cdots,N} \left(\eta_{i_1}^{\mu,a} \cdots \eta_{i_P}^{\mu,a} \sigma_{i_1} \cdots \sigma_{i_P} \right) \right] \tag{2.10}$$

where we have set $\beta' = 2\beta/P!$ and for the first pattern ($\mu = 1$) we used $P!\sum_{i_1<\cdots<i_P} = \sum_{i_1\neq\cdots\neq i_P}$ and neglected terms which vanish in the thermodynamic limit. Focusing only on the last term in (2.10) we note that the product $\eta_{i_1}^{\mu,a} \ldots \eta_{i_P}^{\mu,a}$ is a random i.i.d. Binomial variable for each μ, a, and the sum over μ converges, by the central limit theorem (CLT), to a Gaussian variable with suitably defined mean and variance, i.e.

$$\frac{1}{K}\sum_{\mu>1}^{K} \left(\frac{1}{M} \sum_{a=1}^{M} \eta_{i_1}^{\mu,a} \cdots \eta_{i_P}^{\mu,a} \right) \sim \mu_1 + \lambda_{i_1,\ldots,i_P} \frac{\sqrt{\mu_2 - \mu_1^2}}{\sqrt{K}}, \qquad \lambda_{i_1\ldots i_P} \sim \mathcal{N}(0,1), \tag{2.11}$$

where

$$\mu_1 = \mathbb{E}_{\boldsymbol{\xi}} \mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})} \left[\frac{1}{M} \sum_{a=1}^{M} \eta_{i_1}^{\mu,a} \cdots \eta_{i_P}^{\mu,a} \right] = \mathbb{E}_{\boldsymbol{\xi}} \left[\left(r\xi_{i_1}^{\mu}\right) \ldots \left(r\xi_{i_P}^{\mu}\right) \right] = 0 \tag{2.12}$$

and

$$\begin{aligned}\mu_2 &= \mathbb{E}_{(\boldsymbol{\xi})}\mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})}\left[\left(\frac{1}{M}\sum_{a=1}^{M}\eta_{i_1}^{\mu,a}\cdots\eta_{i_P}^{\mu,a}\right)^2\right]\\ &= \mathbb{E}_{(\boldsymbol{\xi})}\mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})}\left[\frac{1}{M^2}\sum_{a=1}^{M}\left(\eta_{i_1}^{\mu,a}\right)^2\cdots\left(\eta_{i_P}^{\mu,a}\right)^2\right]\\ &\quad+\mathbb{E}_{(\boldsymbol{\xi})}\mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})}\left[\frac{1}{M^2}\sum_{a\neq b}^{M}\left(\eta_{i_1}^{\mu,a}\eta_{i_1}^{\mu,b}\right)\cdots\left(\eta_{i_P}^{\mu,a}\eta_{i_P}^{\mu,b}\right)\right]\\ &= \frac{1}{M}+\frac{M-1}{M}r^{2P} = r^{2P}\left(1+\frac{1-r^{2P}}{Mr^{2P}}\right).\end{aligned} \tag{2.13}$$

This enables us to write

$$\frac{1}{K}\sum_{\mu>1}^{K}\left(\frac{1}{M}\sum_{a=1}^{M}\eta_{i_1}^{\mu,a}\cdots\eta_{i_P}^{\mu,a}\right)\sim\frac{r^P}{\sqrt{K}}\sqrt{(1+\rho_P)}\lambda_{i_1\ldots i_P},\qquad\lambda_{i_1,\ldots,i_P}\sim\mathcal{N}(0,1) \tag{2.14}$$

with

$$\rho_P=\frac{1-r^{2P}}{Mr^{2P}}.$$

Thus the partition function of the unsupervised DAM model reads as

$$\begin{aligned}\mathcal{Z}_N^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J\right) = \lim_{J\to 0}\sum_{\boldsymbol{\sigma}}\exp\Bigg[&J\sum_{i=1}^{N}\xi_i^1\sigma_i+\frac{\beta'}{2\mathcal{R}^{P/2}MN^{P-1}}\sum_{a=1}^{M}\left(\sum_{i}^{N}\eta_i^{1,a}\sigma_i\right)^P\\ &+\frac{\alpha}{\sqrt{K}}\frac{\beta' P!\sqrt{(1+\rho_P)}}{2\left(1+\rho\right)^{P/2}}\left(\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\lambda_{i_1,\ldots,i_P}\,\sigma_{i_1}\cdots\sigma_{i_P}\right)\Bigg].\end{aligned} \tag{2.15}$$

To make analytical progress, it is convenient to introduce the order parameters of the model.

Definition 4. The order parameters of the unsupervised DAM model are the Mattis magnetization m_1 of the archetype $\boldsymbol{\xi}^1$, the Mattis magnetizations $n_{a,1}$ of each example $a=1\ldots M$ of the archetype $\boldsymbol{\xi}^1$ and the two-replica overlaps q_{lm}, which quantifies the correlations between the variables $\boldsymbol{\sigma}$ in two copies l and m of the system, sharing the same disorder (i.e. two *replicas*):

$$m_1\left(\boldsymbol{\sigma}\right):=\frac{1}{N}\sum_{i=1}^{N}\xi_i^1\sigma_i \tag{2.16}$$

$$n_{1,a}\left(\boldsymbol{\sigma}\right):=\frac{r}{\mathcal{R}}\frac{1}{N}\sum_{i=1}^{N}\eta_i^{1,a}\sigma_i, \tag{2.17}$$

$$q_{lm}\left(\boldsymbol{\sigma}^{(l)},\boldsymbol{\sigma}^{(m)}\right):=\frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(l)}\sigma_i^{(m)}. \tag{2.18}$$

In the next subsection we calculate the free energy of the system in the thermodynamic limit using Guerra's interpolation technique, assuming one step of replica symmetry breaking (1RSB).

2.2. 1RSB analysis via Guerra's interpolation technique

The key idea of Guerra's interpolation method is to define an auxiliary free energy, $\mathcal{F}^{(P)}(t)$, which is function of a parameter $t \in [0,1]$ that interpolates between the free energy of the original DAM model (obtained at $t=1$) and the free energy of an exactly solvable one-body model (obtained at $t=0$), whose effective fields mimic those of the original model. As the direct calculation of $\mathcal{F}^{(P)}(t=1)$ is cumbersome, this is obtained by using the fundamental theorem of calculus, namely we first evaluate $\mathcal{F}^{(P)}(t=0)$ and then we obtain $\mathcal{F}^{(P)}(t=1)$ as

$$\mathcal{F}^{(P)}(t=1) = \mathcal{F}^{(P)}(t=0) + \int_0^1 \mathrm{d}s \, \frac{\mathrm{d}\mathcal{F}^{(P)}}{\mathrm{d}t}|_{t=s}. \tag{2.19}$$

To perform the analysis within the first step of replica symmetry breaking, we make the standard 1RSB assumption [2], stated below:

Assumption 1. In 1RSB assumption, the distribution of the two-replica overlap q_{12}, in the thermodynamic limit, displays two delta-peaks at the values $\bar{q}_1$ and $\bar{q}_2$, and the concentration on these two values is ruled by the parameter $\theta \in [0,1]$, namely

$$\lim_{N\to+\infty} \mathbb{P}_N(q_{12}) = \theta\delta(q_{12}-\bar{q}_1) + (1-\theta)\,\delta(q_{12}-\bar{q}_2), \tag{2.20}$$

where

$$\mathbb{P}_N(q_{12}) = \sum_{\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}} \sum_{\boldsymbol{\eta}} \mathbb{P}(\boldsymbol{\eta})\,\mathbb{P}_N\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}|\boldsymbol{\eta}\right)\delta\left(q_{12}-q_{12}\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}\right)\right)$$

and $\mathbb{P}_N(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}|\boldsymbol{\eta}) = \mathcal{B}_N(\boldsymbol{\sigma}_1|\boldsymbol{\eta})\mathcal{B}_N(\boldsymbol{\sigma}_2|\boldsymbol{\eta})$ as replicas are conditionally independent, given the disorder $\boldsymbol{\eta}$.

The Mattis magnetizations m_1 and $n_{1,a}$ are assumed to be self-averaging at their equilibrium values $\bar{m}$ and $\bar{n}$, respectively, namely

$$\lim_{N\to+\infty} \mathbb{P}_N(m_1) = \delta(m_1-\bar{m}), \tag{2.21}$$

$$\lim_{N\to+\infty} \mathbb{P}_N(n_{1,a}) = \delta(n_{1,a}-\bar{n}). \tag{2.22}$$

where $\mathbb{P}_N(m_1) = \mathbb{E}_{\boldsymbol{\eta}} \sum_{\boldsymbol{\sigma}} \mathcal{B}_N(\boldsymbol{\sigma}|\boldsymbol{\eta})\delta(m_1 - m_1(\boldsymbol{\sigma}))$ and $\mathbb{P}_N(n_{1,a}) = \mathbb{E}_{\boldsymbol{\eta}} \sum_{\boldsymbol{\sigma}} \mathcal{B}_N(\boldsymbol{\sigma}|\boldsymbol{\eta})\delta(n_{1,a} - n_{1,a}(\boldsymbol{\sigma}))$, with $\mathbb{E}_{\boldsymbol{\eta}} f(\boldsymbol{\eta}) = \sum_{\boldsymbol{\eta}} f(\boldsymbol{\eta})\mathbb{P}(\boldsymbol{\eta})$.

Remark 2. We note that for $\theta = 0$, equation (2.20) corresponds to the standard RS ansatz with the distribution of the two-replica overlap being delta-peaked at $\bar{q} = \bar{q}_2$. As θ increases, the 1-RSB Ansatz is gradually introduced. The optimal value of Parisi's parameter θ can in principle be obtained by minimizing the free energy $\mathcal{F}_N^{(P)}$ with respect to θ. However, the explicit form of the extremization condition of the free energy with respect to θ is known to be complicated and hardly used in practice, so that θ is often left as a free parameter (see e.g. [7]). In this work, we will fix θ by following a different procedure, namely by maximizing the region where the network accomplishes pattern retrieval, as done in [45] (see section 2.4 for a more detailed discussion).

Definition 5. Given the interpolating parameter $t \in [0,1]$, the constant $A_1,\ A_2,\ \psi \in \mathbb{R}$ to be fixed later on, and the i.i.d. standard Gaussian variables $Y_i^{(b)} \sim \mathcal{N}(0,1)$ for $i = 1,\dots,N$ and $b = 1,2$ (that must be averaged over as explained below), the Guerra's 1RSB interpolating partition function for the unsupervised DAM model is given by

$$\mathcal{Z}_2^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right) := \sum_{\boldsymbol{\sigma}} \mathcal{B}_{N,2}^{(P)}\left(\boldsymbol{\sigma}|\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right) \tag{2.23}$$

$$\begin{aligned} &= \sum_{\boldsymbol{\sigma}} \exp\left[J\sum_{i=1}^{N}\xi_i^1\sigma_i + \frac{t\beta' N}{2M}(1+\rho)^{P/2}\sum_{a=1}^{M} n_{1,a}^{P}(\boldsymbol{\sigma}) + \psi(1-t)N\sum_{a=1}^{M} n_{1,a}(\boldsymbol{\sigma}) \right. \\ &\quad \left. +\sqrt{1-t}\sum_{b=1}^{2}\left(A_b\sum_{i=1}^{N}Y_i^{(b)}\sigma_i\right) + \sqrt{t}\frac{\beta' P!\alpha\sqrt{(1+\rho_P)}}{2(1+\rho)^{P/2}\sqrt{K}}\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\lambda_{i_1,\dots,i_P}\sigma_{i_1}\cdots\sigma_{i_P}\right]. \end{aligned} \tag{2.24}$$

The index 2 on the LHS stands for the number of vectors $\boldsymbol{Y}^{(b)}$ that must be averaged over. Their number is equal to $k+1$ where k is the number of steps of RSB (here $k=1$).

The average over the generalised Boltzmann factor associated to the interpolating partition function $\mathcal{Z}_2^{(P)}(\boldsymbol{\eta},\boldsymbol{\eta},\boldsymbol{Y};J,t)$, can be defined as

$$\omega_t(\cdot) := \frac{1}{\mathcal{Z}_2^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)}\sum_{\boldsymbol{\sigma}}\cdot\,\mathcal{B}_{N,2}^{(P)}\left(\boldsymbol{\sigma}|\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right). \tag{2.25}$$

Remark 3. We note that for $t=1$, (2.24) recovers the original partition function (2.15), whereas for $t=0$ it reduces to the partition function of a system of N non-interacting neurons, described by the Hamiltonian $-\beta'\mathcal{H}_N^{(1)}(\boldsymbol{\sigma}|\boldsymbol{\eta},\mathbf{Y};J) = \sum_i h_i\sigma_i$, with local fields $h_i = J\xi_i^1 + \psi r\mathcal{R}^{-1}\sum_{a=1}^{M}\eta_i^{1,a} + \sum_{b=1}^{2}A_b J_i^{(b)}$, that is readily evaluated. The parameters of the one-body model ψ, A_1 and A_2 must be chosen in such a way that the t-dependent terms in $d\mathcal{F}^{(P)}/dt$ cancel out in the thermodynamic limit, under the 1RSB assumption.

In what follows, we will average over the fields $\boldsymbol{Y}^{(b)}$ for $b=1,2$ recursively, as explained by Guerra in [43] and in the statements below. To this purpose, we define

$$\mathcal{Z}_1^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y}^{(1)};J,t\right) := \left[\mathbb{E}_{\boldsymbol{Y}^{(2)}}\mathcal{Z}_2^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)^{\theta}\right]^{1/\theta} \tag{2.26}$$

$$\mathcal{Z}_0^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J,t\right) := \exp\left\{\mathbb{E}_{\boldsymbol{Y}^{(1)}}\left[\ln\mathcal{Z}_1^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y}^{(1)};J,t\right)\right]\right\} \tag{2.27}$$

$$\mathcal{Z}_N^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J,t\right) := \mathcal{Z}_0^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J,t\right), \tag{2.28}$$

where with $\mathbb{E}_{\boldsymbol{Y}^{(b)}}[.]$ we mean the average over the vectors $\boldsymbol{Y}^{(b)}$ for $b=1,2$. The interpolating quenched free energy related to the partition function (2.24) is introduced as

$$-\beta'\mathcal{F}_N^{(P)}(J,t) := \frac{1}{N}\mathbb{E}_0\left[\ln\mathcal{Z}_N^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J,t\right)\right], \tag{2.29}$$

where $\mathbb{E}_0[.]$ denotes the average over the variables $\lambda_{i_1,\dots,i_P}^{\mu}$'s and $\eta_i^{1,a}$'s. In the thermodynamic limit, assuming the limit exists, we write

$$\mathcal{F}^{(P)}(J,t) := \lim_{N\to\infty}\mathcal{F}_N^{(P)}(J,t). \tag{2.30}$$

Remark 4. We note that in the RS case, where $\mathcal{Z}_2^{(P)}(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t)\equiv\mathcal{Z}_1^{(P)}(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y}^{(1)};J,t)$ and $\boldsymbol{Y}^{(2)}=0$, we have

$$-\beta'\mathcal{F}^{(P)}(J,t):=\lim_{N\to\infty}\frac{1}{N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\left[\ln\mathcal{Z}_1^{(P)}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y}^{(1)};J,t\right)\right], \tag{2.31}$$

hence the average $\mathbb{E}_{\boldsymbol{Y}^{(1)}}$ acts on the same level as $\mathbb{E}_0$ i.e. outside the logarithm. When moving from the RS to the 1RSB scenario, an extra average is introduced, $\mathbb{E}_{\boldsymbol{Y}^{(2)}}$, which acts on a different level than $\mathbb{E}_0$ and $\mathbb{E}_{\boldsymbol{Y}^{(1)}}$ i.e. *inside* the logarithm. This reflects the presence of hierarchical valleys in the free energy landscape [46] which leads to the well-known multi-scale thermalization in spin glasses. The internal average $\mathbb{E}_{Y^{(2)}}$ is related to thermalization within a valley at the lower level of the (two-level) hierarchy, whereas the external averages account for thermalization across valleys (i.e. at the higher level of the hierarchy). The structure of the hierarchy is captured by the parameter θ, which controls the amplitudes of the valleys at the lower level of the hierarchy and can be related to *effective temperatures* [47], as discussed in [48] for Parisi's replica approach and in [49] for Guerra's interpolation techniques. Note that the distinction between the two levels of the hierarchy is not made for ψ which accounts for the signal contribution and is kept RS (as usually done for the magnetization in spin-glass models).

Now, following Guerra's prescription [43], given two copies (or replicas) of the system, we define the following averages, corresponding to thermalization within the two different levels of the hierarchy

$$\langle\cdot\rangle_{t,1}:=\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\left[\mathbb{E}_{\boldsymbol{Y}^{(2)}}\mathcal{W}_{2,t}\omega_t(\cdot)\right]^2, \tag{2.32}$$

$$\langle\cdot\rangle_{t,2}:=\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\mathbb{E}_{\boldsymbol{Y}^{(2)}}\mathcal{W}_{2,t}\left[\omega_t(\cdot)\right]^2, \tag{2.33}$$

where

$$\mathcal{W}_{2,t}:=\frac{\mathcal{Z}_2^{\theta}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)}{\mathbb{E}_{\boldsymbol{Y}^{(2)}}\mathcal{Z}_2^{\theta}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)}. \tag{2.34}$$

At this point, we are able to state our second assumption, that is at each level $a=1,2$ of the hierarchy, the two-replica overlap $q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})$ self-averages around the value $\bar{q}_a$ of the corresponding peak in the overlap distribution $\mathbb{P}_N(q)$, as given in assumption 1:

Assumption 2. For any $t\in[0;1]$

$$\left\langle\left[q_{12}\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}\right)-\bar{q}_a\right]^k\right\rangle_{t,a}\xrightarrow{N\to\infty}0\quad\text{for } a=1,2\ \text{ and } k\geqslant 2$$

Finally, we provide the explicit expression of the quenched free energy in terms of the control parameters in the next

Proposition 1. *In the thermodynamic limit $N \to \infty$, within the 1RSB Assumption and under the assumption 2, the quenched free energy for the unsupervised DAM model, reads as*

$$
\begin{aligned}
-\beta' \mathcal{F}^{(P)}(J=0) = \frac{1}{\theta}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \Bigg\{ \cosh^{\theta}\Bigg[& \beta' \frac{P}{2}\bar{n}^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M \\
& + Y^{(1)}\beta'\sqrt{\gamma\frac{P}{2}\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\bar{q}_1^{P-1}} + Y^{(2)}\beta'\sqrt{\gamma\frac{P}{2}\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)}\Bigg]\Bigg\} \\
& - \beta'\frac{1}{2}(P-1)(1+\rho)^{P/2}\bar{n}^P + \frac{\beta'^2}{4}\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\Big[1 - P\bar{q}_2^{P-1} \\
& + (P-1)\bar{q}_2^P\Big] - \frac{\beta'^2}{4}\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}(P-1)\theta\left(\bar{q}_2^P - \bar{q}_1^P\right) + \ln 2. \qquad (2.35)
\end{aligned}
$$

where

$$
\hat{\eta}_M := \frac{1}{rM}\sum_{a=1}^{M}\eta^{1,a}, \qquad (2.36)
$$

$\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} = \prod_{a=1}^{M}\mathbb{E}_{(\eta^{1,a}|\xi^1)}$ and $\mathbb{E}_{\boldsymbol{\xi}^1}$ and $\mathbb{E}_{(\eta^{1,a}|\xi^1)}$ denote the expectations over the Bernoulli distributions (2.1) and (2.2) respectively. In the above, $\bar{n}$, $\bar{q}_1$, $\bar{q}_2$ fulfill the following self-consistency equations

$$
\begin{aligned}
\bar{n} &= \frac{1}{1+\rho}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)\tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}\hat{\eta}_M\right], \\
\bar{q}_1 &= \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)\tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}\right]^2, \\
\bar{q}_2 &= \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)\tanh^2 g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}}\cosh^{\theta} g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}\right], \qquad (2.37)
\end{aligned}
$$

with

$$
\begin{aligned}
g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) = \beta'\frac{P}{2}\bar{n}^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M + Y^{(1)}\beta'\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} \\
+ Y^{(2)}\beta'\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)} \qquad (2.38)
\end{aligned}
$$

and $\beta' = 2\beta/P!$. *Furthermore, as* $\bar{m} = -\beta' \nabla_J \mathcal{F}^{(P)}(J)|_{J=0}$, *we have*

$$\bar{m} = \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) \tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)} \xi^1 \right]. \tag{2.39}$$

In order to prove the aforementioned proposition, we need to show the following

Lemma 1. *In the thermodynamic limit, the* t *derivative of the interpolating quenched free energy (2.29) under the 1RSB assumption and under assumption 2 is given by*

$$\begin{aligned} \frac{d\mathcal{F}^{(P)}(J,t)}{dt} &= \frac{1}{2}(P-1)(1+\rho)^{P/2} \bar{n}^P - \frac{\beta'}{4}\gamma \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\left[1 - P\bar{q}_2^{P-1} + (P-1)\bar{q}_2^P\right] \\ &\quad + \frac{\beta'}{4}\gamma \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}(P-1)\theta\left(\bar{q}_2^P - \bar{q}_1^P\right). \end{aligned} \tag{2.40}$$

The proof of this lemma is shown in appendix B.1. Now let us prove proposition 1:

Proof. Exploiting the fundamental theorem of calculus, we can relate the free energy of the original model $\mathcal{F}^{(P)}(J,t=1)$ and the one stemming from the one body terms $\mathcal{F}^{(P)}(J,t=0)$ via (2.19). Since, the t-derivative (2.40) does not depend on t, all we need is to add the expression above to the free energy $\mathcal{F}^{(P)}(J,t=0)$, which only contains one body terms. Let us start with the computation of the latter at finite size N:

$$\begin{aligned} -\beta' \mathcal{F}_N^{(P)}(J,t=0) &= \frac{1}{\theta N} \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \sum_{\boldsymbol{\sigma}} \exp\left[-\beta \mathcal{H}_N^{(1)}\left(\boldsymbol{\sigma}|\boldsymbol{\eta}^1,\mathbf{Y};J\right)\right] \\ &= \ln 2 + \frac{1}{\theta} \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \cosh^\theta \\ &\quad \times \left(J\xi^1 + \frac{\psi}{(1+\rho)} \hat{\eta}_M + Y^{(1)} A_1 + Y^{(2)} A_2 \right), \end{aligned} \tag{2.41}$$

with the definition of $-\beta \mathcal{H}_N^{(1)}(\boldsymbol{\sigma}|\boldsymbol{\eta};J)$ given in remark 3.

Upon setting the constants to the values (B.8) stated in the proof of lemma 1, we have

$$-\beta' \mathcal{F}_N^{(P)}(J,t=0) = \ln 2 + \frac{1}{\theta} \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta',K,N,\boldsymbol{Y}\right), \tag{2.42}$$

where $g(\beta,K,N,\boldsymbol{Y})$ reads as

$$\begin{aligned} g\left(\beta,K,N,\boldsymbol{Y}\right) &= J\xi^1 + \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \hat{\eta}_M \\ &\quad + Y^{(1)} \beta' \sqrt{\frac{P\sqrt{(1+\rho_P)}^2}{2(1+\rho)^P} \frac{P!K}{2N^{P-1}} \bar{q}_1^{P-1}} \\ &\quad + Y^{(2)} \beta' \sqrt{\frac{P\sqrt{(1+\rho_P)}^2}{2(1+\rho)^P} \frac{P!K}{2N^{P-1}} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)}. \end{aligned} \tag{2.43}$$

Recalling that $\lim_{N\to\infty} K/N^{P-1} = 2\gamma/P!$, taking the thermodynamic limit of the above equations and inserting them in the fundamental theorem of calculus (2.19), we reach (2.35).

Finally, by maximizing (2.35) w.r.t. to the order parameters $\bar{n}, \bar{q}_1, \bar{q}_2$, we obtain the self-consistency equations (2.37), hence we reach the thesis. □

2.3. 1RSB analysis via Parisi's replica trick

In this section, we derive the expression of the quenched free energy of the unsupervised DAM model provided in proposition 1 by using the replica method [2, 50], at the first step of RSB [2, 51]. The core of this approach consists in writing the quenched free energy, as defined in (2.6), as

$$-\beta' \mathcal{F}_N^{(P)}(J) = \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}\mathcal{Z}_N^n(J) - 1}{n}, \tag{2.44}$$

with $\beta' = 2\beta/P!$ and where we have used the shorthand $\mathcal{Z}_N(J) = \mathcal{Z}_N^{(P)}(\boldsymbol{\eta}; J)$ for the partition function, defined as

$$\mathcal{Z}_N(J) = \sum_{\boldsymbol{\sigma}} \exp\left(-\beta' \mathcal{H}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta}) + J \sum_{i=1}^{N} \xi_i^1 \sigma_i \right), \tag{2.45}$$

$\mathbb{E}[.]$ denotes, as in (2.6), the average over the quenched disorder $\boldsymbol{\eta}$ and in accordance with the explanation provided in section 2.1, the inclusion of the last term in the round brackets of (2.45) is intended to make the free energy a moment-generating function for the Mattis magnetization $m_1(\boldsymbol{\sigma})$.

For integer n the function $\mathcal{Z}_N^n$ is the product of a system of n identical replicas of the original system

$$\mathbb{E}\mathcal{Z}_N^n(J) = \mathbb{E} \sum_{\boldsymbol{\sigma}^1 \dots \boldsymbol{\sigma}^n} \exp\left(-\beta' \sum_{a=1}^{N} \mathcal{H}_N^{(P)}\left(\boldsymbol{\sigma}^{(a)}|\boldsymbol{\eta}\right) + J \sum_{a=1}^{n} \sum_{i=1}^{N} \xi_i^1 \sigma_i^{(a)} \right) \tag{2.46}$$

namely

$$\mathbb{E}\mathcal{Z}_N^n(J) = \mathbb{E} \sum_{\boldsymbol{\sigma}^1 \dots \boldsymbol{\sigma}^n} \exp\left[\frac{\beta' P!}{2\mathcal{R}^{P/2} M N^{P-1}} \sum_{a=1}^{n} \sum_{\mu=1}^{K} \sum_{A=1}^{M} \left(\sum_{i_1 < \dots < i_P}^{N,\dots,N} \eta_{i}^{\mu,A} \dots \eta_{i_P}^{\mu,A} \sigma_{i_1}^{(a)} \dots \sigma_{i_P}^{(a)} \right) + J \sum_{a,i}^{n,N} \xi_i^1 \sigma_i^{(a)} \right]. \tag{2.47}$$

Splitting, as before, the signal term ($\mu = 1$) from the remaining terms $\mu > 1$, which act as a quenched noise on the learning and retrieval of pattern $\boldsymbol{\xi}^1$, we can write

$$\mathbb{E}\mathcal{Z}_N^n(J) = \sum_{\boldsymbol{\sigma}^1 \dots \boldsymbol{\sigma}^n} \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \exp\left[\frac{\beta'(1+\rho)^{P/2}}{2MN} \sum_{a=1}^{n} \sum_{A=1}^{M} \left(\frac{r}{N\mathcal{R}} \sum_{i} \eta_i^{1,A} \sigma_i^{(a)} \right)^P + J \sum_{a,i}^{n,N} \xi_i^1 \sigma_i^{(a)} \right]$$
$$\cdot \mathbb{E}_{\boldsymbol{\xi}^{\mu>1}} \mathbb{E}_{(\boldsymbol{\eta}^{\mu>1}|\boldsymbol{\xi}^{\mu>1})} \exp\left[\frac{\beta' P!}{2M\mathcal{R}^{P/2} N^{P-1}} \sum_{\mu>1,a,A} \sum_{i_1 < \dots < i_P} \eta_{i_1}^{\mu,A} \dots \eta_{i_P}^{\mu,A} \sigma_{i_1}^{(a)} \dots \sigma_{i_P}^{(a)} \right]. \tag{2.48}$$

where $\mathbb{E}_{\boldsymbol{\xi}^\mu}[.] = \prod_i \mathbb{E}_{\xi_i^\mu}[.]$ denotes the average over the pattern $\boldsymbol{\xi}^\mu$, whose entries ξ_i^μ are drawn from the distribution (2.1) and $\mathbb{E}_{(\boldsymbol{\eta}^\mu|\boldsymbol{\xi}^\mu)} = \prod_{i,a} \mathbb{E}_{(\eta_i^{\mu,a}|\xi_i^\mu)}$ with $\mathbb{E}_{\eta_i^{\mu,A}}$ denoting the expectation over the conditional distribution (2.2). As before, exploiting the CLT to rewrite the last term in (2.48) via (2.14), we obtain

$$\begin{aligned}\mathbb{E}\mathcal{Z}_N^n(J) = &\sum_{\boldsymbol{\sigma}^1\dots\boldsymbol{\sigma}^n} \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \exp\left[\frac{\beta'(1+\rho)^{P/2}N}{2M}\sum_{a=1}^{n}\sum_{A=1}^{M}\left(\frac{r}{N\mathcal{R}}\sum_i \eta_i^{1,A}\sigma_i^{(a)}\right)^P + J\sum_{a,i}^{n,N}\xi_i^1\sigma_i^{(a)}\right]\\ &\cdot \mathbb{E}_{\boldsymbol{\lambda}} \exp\left[\frac{\beta' P!\alpha\sqrt{(1+\rho_P)}}{2(1+\rho)^{P/2}\sqrt{K}}\sum_a \sum_{i_1<\dots<i_P} \lambda_{i_1,\dots,i_P}\sigma_{i_1}^{(a)}\dots\sigma_{i_P}^{(a)}\right]\\ = &\sum_{\boldsymbol{\sigma}^1\dots\boldsymbol{\sigma}^n} \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathcal{Z}_{\text{signal}}^n(J)\,\mathbb{E}_{\boldsymbol{\lambda}}\mathcal{Z}_{\text{noise}}^n, \end{aligned} \tag{2.49}$$

Treating the two terms $\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathcal{Z}_{\text{signal}}^n(J)$ and $\mathbb{E}_{\boldsymbol{\lambda}}\mathcal{Z}_{\text{noise}}^n$ separately, we insert in the former the identity

$$1 = \prod_{a,A=1,1}^{n,M}\int \mathrm{d}n_{1,A}^{(a)}\,\delta\left(n_{1,A}^{(a)} - n_{1,A}^{(a)}(\boldsymbol{\sigma})\right)$$

and use the Fourier representation of the Dirac delta, obtaining

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathcal{Z}_{\text{signal}}^n(J) = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\int\prod_{a,A}\mathrm{d}n_{1,A}^{(a)}\prod_{a,A}\frac{\mathrm{d}z_{1,A}^{(a)}}{2\pi}\exp\left[\frac{\beta' N(1+\rho)^{P/2}}{2M}\sum_{A,a}\left(n_{1,A}^{(a)}\right)^P\right.\\ \left. + i\sum_{a,A}\left(n_{1,A}^{(a)} - \frac{r}{\mathcal{R}N}\sum_i \eta_i^{1,A}\sigma_i^{(a)}\right)z_{1,A}^{(a)} + J\sum_{a,i}^{n,N}\xi_i^1\sigma_i^{(a)}\right].\end{aligned} \tag{2.50}$$

For the noise term, performing first the expectation over $\boldsymbol{\lambda}$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda}}\mathcal{Z}_{\text{noise}}^n &= \int\prod_{\boldsymbol{i}}\frac{d\lambda_{i_1,\dots,i_P}}{\sqrt{2\pi}}\exp\left[-\frac{(\lambda_{i_1,\dots,i_P})^2}{2} + \frac{\beta'\alpha\sqrt{(1+\rho_P)}}{2(1+\rho)^{P/2}\sqrt{K}}P!\sum_a^n\sum_{i_1<\dots<i_P}^{N,\dots,N}\lambda_{i_1,\dots,i_P}\sigma_{i_1}^{(a)}\cdots\sigma_{i_P}^{(a)}\right]\\ &= \exp\left[\frac{1}{2}\left(\frac{\beta'\sqrt{(1+\rho_P)}}{2(1+\rho)^{P/2}N^{P-1}}\right)^2 K(P!)^2\sum_{i_1<\dots<i_P}^{N,\dots,N}\sum_{a,b}^{n,n}\sigma_{i_1}^{(a)}\sigma_{i_1}^{(b)}\cdots\sigma_{i_P}^{(a)}\sigma_{i_P}^{(b)}\right]\\ &= \exp\left[\frac{1}{2}\left(\frac{\beta'\sqrt{(1+\rho_P)}}{2(1+\rho)^{P/2}N^{P/2-1}}\right)^2 KP!\sum_{a,b}^{n,n}\left(\frac{1}{N}\sum_i^N\sigma_{i_1}^{(a)}\sigma_{i_1}^{(b)}\right)^P\right]\end{aligned} \tag{2.51}$$

inserting

$$1 = \prod_{a,b}\int \mathrm{d}q_{ab}\,\delta\left(q_{ab} - q_{ab}(\boldsymbol{\sigma})\right)$$

and using the Fourier representation of the Dirac delta, we have

$$\mathbb{E}_{\boldsymbol{\lambda}}\mathcal{Z}^n_{\text{noise}} = \int \prod_{a,b} \mathrm{d}q_{ab} \int \prod_{a,b} \frac{\mathrm{d}p_{ab}}{2\pi} \exp\left[\frac{1}{2} \left(\frac{\beta' \sqrt{(1+\rho_P)}^2}{2(1+\rho)^{P/2} N^{P/2-1}} \right)^2 KP! \sum_{a,b}^{n,n} q_{ab}^P + i \sum_{a,b} \left(q_{ab} - \frac{1}{N} \sum_i \sigma_i^{(a)} \sigma_i^{(b)} \right) p_{ab} \right] \tag{2.52}$$

Inserting equations (2.50) and (2.52) in (2.49), we obtain

$$\mathbb{E}\mathcal{Z}^n_N(J) = \int \prod_{a,A} \mathrm{d}n^{(a)}_{1,A} \prod_{a,A} \frac{\mathrm{d}z^{(a)}_{1,A}}{2\pi} \int \prod_{a,b} \mathrm{d}q_{ab} \int \prod_{a,b} \frac{\mathrm{d}p_{a,b}}{2\pi} \exp\left(-NA\left[\boldsymbol{Q},\boldsymbol{P},\boldsymbol{Z},\boldsymbol{N}\right]\right) \tag{2.53}$$

where

$$\begin{aligned} A\left[\boldsymbol{Q},\boldsymbol{P},\boldsymbol{Z},\boldsymbol{N}\right] = &-\frac{i}{N}\sum_{a,b} p_{ab}q_{ab} - \frac{1}{4}\frac{\beta'^2\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\gamma \sum_{a,b}^{n,n} q_{ab}^P - \frac{\beta'(1+\rho)^{P/2}}{2M}\sum_{A,a}\left(n^{(a)}_{1,A}\right)^P \\ &- \frac{i}{N}\sum_{a,A} n^{(a)}_{1,A} z^{(a)}_{1,A} - \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{\left(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1\right)} \log\left[\sum_{\boldsymbol{\sigma}} \exp\left(-\frac{i}{N}\sum_{a,b,i} p_{ab}\sigma^{(a)}\sigma^{(b)} \right.\right. \\ &\left.\left. - \frac{ir}{\mathcal{R}N}\sum_{a,A}\eta^{1,A} z^{(a)}_{1,A}\sigma^{(a)} + J\sum_a \xi^1\sigma^{(a)} \right)\right]. \end{aligned} \tag{2.54}$$

and we have used the shorthand notation $\boldsymbol{Q} = \{q_{ab}\}^n_{a,b=1}$ and similarly for $\boldsymbol{P}$, $\boldsymbol{N}$ and $\boldsymbol{Z}$. Next, we assume that the limit $n \to 0$ in

$$-\beta' \mathcal{F}^{(P)}(\beta) = \lim_{N\to\infty} \frac{1}{N} \lim_{n\to 0} \frac{\mathbb{E}\mathcal{Z}^n_N - 1}{n}, \tag{2.55}$$

can be taken by analytic continuation and that the two limits $N \to \infty$ and $n \to 0$ can be interchanged (provided they exist), so that the integrals in (2.53) can be performed by steepest descent. This leads to

$$-\beta' \mathcal{F}^{(P)}(\beta) = -\lim_{n\to 0} \frac{1}{n} A\left[\boldsymbol{Q}^\star,\boldsymbol{P}^\star,\boldsymbol{Z}^\star,\boldsymbol{N}^\star\right] \tag{2.56}$$

where $\boldsymbol{Q}^\star,\boldsymbol{P}^\star,\boldsymbol{Z}^\star,\boldsymbol{N}^\star$, are solution of the saddle point equations

$$\frac{\partial A}{\partial p_{ab}} = 0 \quad \Rightarrow \quad q^*_{ab} = \langle \sigma^a \sigma^b \rangle_{\text{eff}} \tag{2.57}$$

$$\frac{\partial A}{\partial z^{(a)}_{1,A}} = 0 \quad \Rightarrow \quad \left(n^{(a)}_{1,A}\right)^* = \frac{r}{\mathcal{R}} \sum_{A=1}^{M} \mathbb{E}\langle \eta^{1,A}\sigma^a \rangle_{\text{eff}} \tag{2.58}$$

$$\frac{\partial A}{\partial q_{ab}} = 0 \quad \Rightarrow \quad p^*_{ab} = \frac{i\beta'^2}{2}\frac{P}{2}\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)}N(q^*_{ab})^{P-1} \tag{2.59}$$

$$\frac{\partial A}{\partial n^{(a)}_{1,A}} = 0 \quad \Rightarrow \quad \left(z^{(a)}_{1,A}\right)^* = i\beta'\frac{P}{2}\frac{(1+\rho)^{P/2}}{M}N\left[\left(n^{(a)}_{1,A}\right)^*\right]^{P-1} \tag{2.60}$$

where the average $\langle\ldots\rangle_{\text{eff}}$ is computed over the distribution $p_{\text{eff}}(\boldsymbol{\sigma}) = e^{-\beta' H_{\text{eff}}(\boldsymbol{\sigma})}/Z_{\text{eff}}$ where $H_{\text{eff}}(\boldsymbol{\sigma})$ is equal to the terms in the round brackets in the second line of (2.54). Inserting (2.59) and (2.60) in $H_{\text{eff}}(\boldsymbol{\sigma})$, we obtain

$$\begin{aligned}-\beta' H_{\text{eff}}(\boldsymbol{\sigma}) = {} & \frac{\beta'^2\gamma}{2}\frac{P\sqrt{(1+\rho_P)}^2}{2(1+\rho)^P}\sum_{a,b}[q^*_{ab}]^{P-1}\sigma^{(a)}\sigma^{(b)} \\ & + \beta'\frac{P}{2}(1+\rho)^{P/2-1}\sum_a\frac{1}{rM}\sum_A\eta^{1,A}\left[\left(n^{(a)}_{1,A}\right)^*\right]^{P-1}\sigma^{(a)}\end{aligned} \tag{2.61}$$

and, substituting in (2.56)

$$\begin{aligned}-\beta'\mathcal{F}^{(P)}(J) = \lim_{n\to 0}\frac{1}{n}\Bigg\{ & \frac{\beta'^2\sqrt{(1+\rho_P)}^2\gamma}{4(1+\rho)^P}\sum_{a,b}(P-1)[q^*_{ab}]^P + \frac{\beta'(1+\rho)^{P/2}(P-1)}{2M}\sum_{a,A}\left[\left(n^{(a)}_{1,A}\right)^*\right]^P \\ & + \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\log\left[\sum_{\boldsymbol{\sigma}}\exp\left(-\beta' H_{\text{eff}}(\boldsymbol{\sigma}) + J\sum_a\xi^1\sigma^{(a)}\right)\right]\Bigg\}.\end{aligned} \tag{2.62}$$

where $q^\star_{ab}$ and $[n^{(a)}_{1,A}]^*$ must be determined from (2.57) and (2.58). Since $p_{\text{eff}}(\boldsymbol{\sigma})$ depends on $\boldsymbol{Q}$, $\boldsymbol{N}$ via (2.61), equations (2.57) and (2.58) denote a set of self-consistency equations. Now, in order to proceed, we need to find the form of $\boldsymbol{Q}$ and $\boldsymbol{N}$ in the limit $n \to 0$. We will make the 1RSB ansatz:

$$q_{ab} = \delta_{ab} + (1-\delta_{ab})(\theta\bar{q}_1 + (1-\theta)\bar{q}_2), \tag{2.63}$$

$$n^{(a)}_{1,A} = \bar{n}, \tag{2.64}$$

Using (2.63) to evaluate the first term in (2.62) and the first term in (2.61), we get

$$\sum_{ab} q^P_{ab} = n\left(1 + n\theta\bar{q}_1^P - \theta\bar{q}_1^P - \bar{q}_2^P + n\bar{q}_2^P + \theta\bar{q}_2^P - n\theta\bar{q}_2^P\right) \tag{2.65}$$

$$\begin{aligned}\sum_{ab} q^{P-1}_{ab}\sigma^{(a)}\sigma^{(b)} = \frac{1}{2}\Bigg[& 2\left(1-\bar{q}_2^{P-1}\right)n + \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)\sum_{k=1}^{n/\theta}\left(\sum_{a=1}^{\theta}\sigma^{((k-1)\theta+a)}\right)^2 \\ & + \bar{q}_1^{P-1}\left(\sum_a\sigma^{(a)}\right)^2\Bigg].\end{aligned} \tag{2.66}$$

Upon inserting in (2.62) we obtain

$$
\begin{aligned}
-\beta' \mathcal{F}^{(P)}(J) = & \frac{\beta'^2 \sqrt{(1+\rho_P)}^2 \gamma}{4(1+\rho)^P}\left[1-(P-1)\theta\left(\bar{q}_2^P-\bar{q}_1^P\right)+(P-1)\bar{q}_2^P\right] \\
& -\frac{\beta'(1+\rho)^{P/2}(P-1)}{2}\bar{n}^P-\frac{(\beta')^2\sqrt{(1+\rho_P)}^2\gamma P}{4(1+\rho)^P}\bar{q}_2^{P-1} \\
& +\lim_{n\to 0}\frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\log\left[\sum_{\boldsymbol{\sigma}}\exp\left(-\beta' H_{\text{eff}}(\boldsymbol{\sigma})+J\sum_a \xi^1\sigma^{(a)}\right)\right]
\end{aligned}
\tag{2.67}
$$

with

$$
\begin{aligned}
-\beta' H_{\text{eff}}(\boldsymbol{\sigma}) = & \beta'^2\gamma\frac{P\sqrt{(1+\rho_P)}^2}{4(1+\rho)^P}\left[\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)\sum_{k=1}^{n/\theta}\left(\sum_{b=1}^{\theta}\sigma^{((k-1)\theta+b)}\right)^2+\bar{q}_1^{P-1}\left(\sum_{a=1}^{n}\sigma^{(a)}\right)^2\right] \\
& +\beta'\frac{P}{2}(1+\rho)^{P/2-1}\hat{\eta}_M\left(\sum_{a=1}^{n}\sigma^{(a)}\right)\bar{n}^{P-1}
\end{aligned}
\tag{2.68}
$$

where we have used the definition (2.36).

Next, we apply a Gaussian transformation to the last term in (2.67) to linearize the squared terms in the Hamiltonian (2.68)

$$
\begin{aligned}
& \lim_{n\to 0}\frac{1}{n}\ln\left[\sum_{\boldsymbol{\sigma}}\exp\left(-\beta' H_{\text{eff}}(\boldsymbol{\sigma})+J\sum_a\xi^1\sigma^{(a)}\right)\right] \\
& \quad =\lim_{n\to 0}\frac{1}{n}\ln\left\{\int\mathcal{D}u\int\left[\prod_{k=1}^{n/\theta}\mathcal{D}v_k\right]\sum_{\boldsymbol{\sigma}}\exp\left[\frac{\beta' P(1+\rho)^{P/2-1}}{2}\hat{\eta}_M\bar{n}^{P-1}\sum_{a=1}^{n}\sigma^{(a)}\right.\right. \\
& \qquad +J\sum_{a=1}^{n}\xi^1\sigma^{(a)}+\beta' u\sum_{a=1}^{n}\sigma^{(a)}\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} \\
& \qquad \left.\left.+\beta'\sum_{k=1}^{n/\theta}v_k\sum_{b=1}^{\theta}\sigma^{((k-1)\theta+b)}\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)}\right]\right\}
\end{aligned}
\tag{2.69}
$$

where we have set $\mathcal{D}u = \mathrm{d}u\, e^{-u^2/2}/\sqrt{2\pi}$ and $\mathcal{D}v_k = \mathrm{d}v_k\, e^{-v_k^2/2}/\sqrt{2\pi}$ for $k = 1,\dots,n/\theta$. Now, writing

$$
\sum_{a=1}^{n}\sigma^{(a)} = \sum_{k=1}^{n/\theta}\sum_{b=1}^{\theta}\sigma^{((k-1)\theta+b)},
$$

the sum over the spin configuration can be done explicitly, finding

$$
\begin{aligned}
&= \lim_{n\to 0}\frac{1}{n}\ln\Bigg\{\int \mathcal{D}u \prod_{k=1}^{n/\theta}\int \mathcal{D}v_k \prod_{b=1}^{\theta}\sum_{\sigma^{((k-1)\theta+b)}} \exp\Bigg(\frac{\beta' P(1+\rho)^{P/2-1}\hat{\eta}_M \bar{n}^{P-1}\sigma^{((k-1)\theta+b)}}{2} \\
&\quad + J\xi^1\sigma^{((k-1)\theta+b)} + \beta' u\,\sigma^{((k-1)\theta+b)}\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} \\
&\quad + \beta' v_k\sigma^{((k-1)\theta+b)}\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)}\Bigg)\Bigg\} \\
&= \lim_{n\to 0}\frac{1}{n}\ln\Bigg\{\int \mathcal{D}u \prod_{k=1}^{n/\theta}\Bigg[\int \mathcal{D}v_k 2^\theta\cosh^\theta\Bigg(\frac{\beta' P(1+\rho)^{P/2-1}}{2}\hat{\eta}_M\bar{n}^{P-1} + J\xi^1 \\
&\quad + \beta' u\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} + \beta' v_k\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)}\Bigg)\Bigg]\Bigg\}.
\end{aligned}
$$

Now applying the $n\to 0$ limit [50] and exploiting the relation $\lim_{n\to 0}\frac{1}{n}\ln\mathbb{E}_{Y^{(1)}}[f^n(Y^{(1)})] = \mathbb{E}_{Y^{(1)}}[\ln f(Y^{(1)})]$, we get

$$
\begin{aligned}
&= \lim_{n\to 0}\frac{1}{n}\ln\Bigg\{\mathbb{E}_u\Bigg[\mathbb{E}_v 2^\theta\cosh^\theta\Bigg(\beta' u\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} \\
&\quad + \beta' v\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)} + \frac{\beta' P(1+\rho)^{P/2-1}}{2}\hat{\eta}_M\bar{n}^{P-1} + J\xi^1\Bigg)\Bigg]^{n/\theta}\Bigg\} \\
&= \frac{1}{\theta}\mathbb{E}_u\ln\Bigg\{\mathbb{E}_v 2^\theta\cosh^\theta\Bigg[\beta'\Bigg(u\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}_1^{P-1}} + v\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left(\bar{q}_2^{P-1}-\bar{q}_1^{P-1}\right)} \\
&\quad + \frac{P(1+\rho)^{P/2-1}\hat{\eta}_M}{2}\bar{n}^{P-1} + J\xi^1\Bigg)\Bigg]\Bigg\}.
\end{aligned}
$$

Finally, inserting this expression in (2.67) and denoting with $\mathbb{E}_{Y^{(1)}}[.]$ and $\mathbb{E}_{Y^{(2)}}[.]$ the Gaussian averages w.r.t. $Y^{(1)}$ and $Y^{(2)}$ respectively, we reach the following expression for the free energy

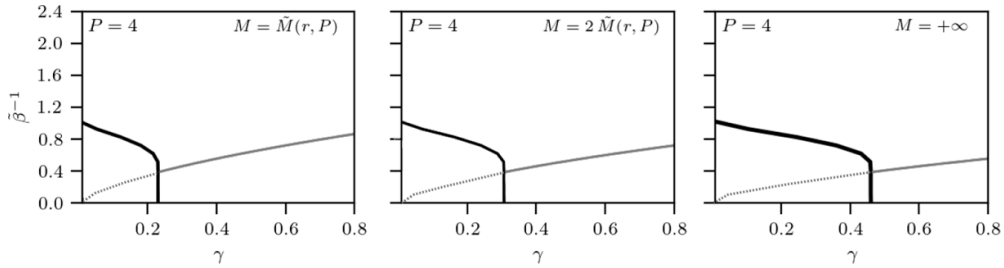


Figure 1. Phase diagram of the unsupervised DAM model with $P = 4$, in the parameters space $(\gamma, \tilde{\beta})$, where γ is the storage load defined in (2.8) and $\tilde{\beta} = \beta'/(1+\rho)^{P/2}$ is the scaled inverse noise, where ρ is the dataset entropy defined in (2.3). Different panels refer to different sizes M of the training set, as shown in the legend, where $\tilde{M}(r, P) = r^{-2P}$ and r has been set to $r = 0.2$. For high values of the noise $\tilde{\beta}^{-1}$, the system is ergodic, while lowering the noise two phases appear: a retrieval phase at small γ (below the black line), and a spin glass phase (below the grey line) at high values of γ. The spin-glass phase within the retrieval region (delimited by the dashed line) is unstable. As M increases, the spin glass region shrinks and the retrieval region expands.

$$
\begin{aligned}
-\beta' \mathcal{F}^{(P)}(J) = \ln 2 + \frac{1}{\theta}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \Bigg\{ \cosh^{\theta} \Bigg[& J\xi^1 + \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \hat{\eta}_M \\
& + Y^{(1)} \beta' \sqrt{\gamma \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \bar{q}_1^{P-1}} + Y^{(2)} \beta' \sqrt{\gamma \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} \Bigg] \Bigg\} \\
& - \frac{\beta'}{2}(P-1)(1+\rho)^{P/2} \bar{n}^P + \frac{\beta'^2}{4} \gamma \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \left[1 - P\bar{q}_2^{P-1} + (P-1)\bar{q}_2^P\right] \\
& - \frac{\beta'^2}{4} \gamma \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} (P-1)\theta \left(\bar{q}_2^P - \bar{q}_1^P\right),
\end{aligned}
$$

where $\bar{n}, \bar{q}_1$ and $\bar{q}_2$ must fulfill the same self-consistency equations (2.37) obtained by using Guerra's interpolation technique, as reported in proposition 1.

2.4. Results

Solving numerically the self consistency equations (2.37), we obtain the phase diagram shown in figure 1, in the parameters space $(\gamma, \tilde{\beta})$, where γ is the storage load defined in (2.8) and $\tilde{\beta} = \beta'/(1+\rho)^{P/2}$ is a scaled inverse temperature, where ρ is the dataset entropy defined in (2.3). Panels show results for $P = 4$ and three different values of M, as shown in the legend. The result for $M \to \infty$ follows from the analysis carried out in section 2.4.1. In each panel, the grey curve marks the transition from the ergodic phase with $\bar{m} = \bar{q}_1 = \bar{q}_2 = 0$ (above) to the spin glass phase, where either $\bar{q}_1$ or $\bar{q}_2$ becomes non-zero (below). The black curve marks the transition from the retrieval phase with $\bar{m} \neq 0$ (left) to $\bar{m} = 0$ (right). The spin glass solution within the retrieval region is always unstable and it is delimited by the dotted curve (we refer to this as the instability region). In figure 2 we show the results for $P = 6, 8, 10$, as shown in the legend. As P and M grow, the spin glass region shrinks and the retrieval region expands. For $M = +\infty$ the critical storage lines become equal to those of traditional DAM models where archetypes are encoded in the interactions, rather than their noisy examples [45]. This is as expected: when

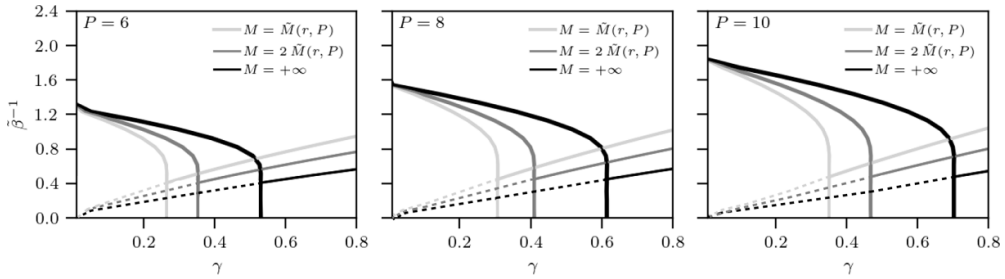


Figure 2. Phase diagrams of the unsupervised DAMs model for $P = 6, 8, 10$, and different values of M, as shown in the legends. $\tilde{M}(r, P)$ and r are as in figure 1. As P increases, the spin glass region shrinks and the retrieval region expands.

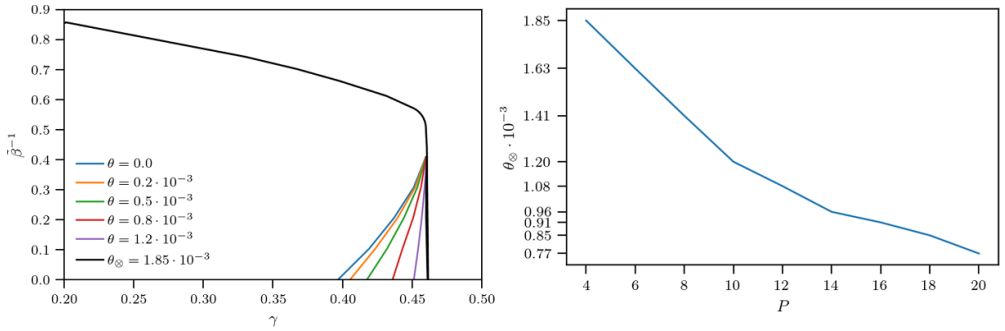


Figure 3. Left: critical line separating the retrieval (left) from the spin-glass or ergodic region (right) for different values of θ. For $\theta = 0$, where the theory is RS, the retrieval region exhibits a re-entrance.As θ is increased the retrieval region expands, reaching its maximum at $\theta = \theta_\otimes$, where the re-entrance disappears completely. Right: $\theta_\otimes$ versus P. As P increases the value of $\theta_\otimes$ decreases, indicating that the larger P the better the RS approximation is expected to be.

an infinite number of examples is provided, the system reaches the same performance as a neural network where archetypes are stored in the interactions directly. For finite M, however, the network's ability to retrieve the archetypes degrades at values of the storage load which are lower than the storage capacity in traditional DAM models. Both figures 1 and 2 have been obtained for the value of $\theta = \theta_\otimes$ which maximizes the retrieval region, as shown in figure 3(left panel). In the latter, we show the line that separates the retrieval region from the spin-glass or the ergodic region, for different values of θ. For $\theta = 0$ (which corresponds to the RS theory), the line exhibits a re-entrance, as commonly observed in spin-glasses and associative memories [45, 50, 52, 53]. As θ is increased, the retrieval region expands, reaching its maximum at $\theta = \theta_\otimes$, where the re-entrance disappears completely, showing the same qualitative behaviour as in the Hopfield model [12] and traditional DAM [45]. Increasing θ above $\theta_\otimes$ the re-entrance appears again and gets more pronounced as θ approaches 1, where the transition line becomes identical to the RS (and the $\theta = 0$) case, as expected. In figure 3(right panel) we show the dependence of $\theta_\otimes$ on P. As P increases, the value of $\theta_\otimes$ decreases. This is as expected from spin-glass theory, as for large P, P-spin models are known to converge to the REM model [54], which is RS. Interestingly, we find that $\theta_\otimes$ takes the same value as in the DAM model

considered in [45], for all values of M, hence the optimal 1RSB breaking parameter is not influenced by the use of corrupted examples rather than archetypes.

2.4.1. Limiting cases: $M \to \infty$ and $\beta \to \infty$. Finally, we consider two instructive limiting cases. One is the limit $M \to \infty$, where the number of available examples is large. Although idealised, this scenario is becoming less utopian nowadays in a number of applications, and it is instructive as an explicit relation between the archetype magnetization and the mean magnetization of the examples naturally emerges. The second scenario is the zero noise limit $\beta \to \infty$, where the information processing capabilities of the network are expected to be maximal. We now state the following

Corollary 1. *In the limit $M \to \infty$, the 1RSB self consistent equations for the order parameters of the unsupervised DAM model are*

$$\begin{aligned}
\bar{n} &= \frac{\bar{m}}{1+\rho} + \beta' \frac{P}{2} \rho (1+\rho)^{P/2-2} \left[1 + (\theta - 1)\bar{q}_2 - \theta \bar{q}_1\right] \bar{n}^{P-1}, \\
\bar{q}_1 &= \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y}) \tanh g(\beta,\gamma,\boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y})} \right]^2, \\
\bar{q}_2 &= \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y}) \tanh^2 g(\beta,\gamma,\boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y})} \right], \\
\bar{m} &= \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y}) \tanh g(\beta,\gamma,\boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta,\gamma,\boldsymbol{Y})} \right].
\end{aligned} \tag{2.70}$$

where

$$\begin{aligned}
g(\beta,\gamma,\boldsymbol{Y}) = \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} + Y^{(2)} \beta' \sqrt{\gamma \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} \\
+ \beta' Y^{(1)} \sqrt{\rho \left(\frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \right)^2 + \gamma \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \bar{q}_1^{P-1}}
\end{aligned} \tag{2.71}$$

and $\beta' = 2\beta/P!$.

Proof. In the limit $M \to \infty$, we can apply the CLT to the sum of the examples defined in (2.36), to write

$$\hat{\eta}_M \sim \xi^1 (1 + \lambda \sqrt{\rho}), \tag{2.72}$$

where λ is a standard Gaussian variable $\lambda \sim \mathcal{N}(0,1)$.

Inserting (2.72) in (2.38) and in the expression for $\bar{n}$ provided in (2.37), we get

$$\begin{aligned}
g\left(\beta,\gamma,\boldsymbol{Y},\lambda,\xi^1\right) = \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \xi^1 (1 + \sqrt{\rho}\lambda) + Y^{(1)} \beta' \sqrt{\frac{\gamma P \sqrt{(1+\rho_P)}^2}{2(1+\rho)^P} \bar{q}_1^{P-1}} \\
+ Y^{(2)} \beta' \sqrt{\frac{\gamma \sqrt{(1+\rho_P)}^2 P}{2(1+\rho)^P} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)},
\end{aligned} \tag{2.73}$$

and

$$\bar{n} = \frac{1}{(1+\rho)} \mathbb{E}_{\xi^1} \mathbb{E}_{\lambda} \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^{\theta} g\left(\beta, \gamma, \boldsymbol{Y}, \lambda, \xi^1\right) \tanh^{\theta} g\left(\beta, \gamma, \boldsymbol{Y}, \lambda, \xi^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^{\theta} g\left(\beta, \gamma, \boldsymbol{Y}, \lambda, \xi^1\right)} \cdot \xi^1 \left(1 + \lambda \sqrt{\rho}\right) \right].$$

Hence, by applying to the standard Gaussian variable λ the Stein's lemma, which states that for a standard Gaussian variable $J \sim N(0,1)$ and a generic function $f(J)$, for which the two expectations $\mathbb{E}\left(Jf(J)\right)$ and $\mathbb{E}\left(\partial_J f(J)\right)$ both exist, one has

$$\mathbb{E}\left[Jf(J)\right] = \mathbb{E}\left[\frac{\partial f(J)}{\partial J}\right], \tag{2.74}$$

and explicitly averaging over ξ we get the self-consistency equations in (2.70).

□

Finally, it has been shown in [38], within a RS analysis, that neglecting the second term in the equation for $\bar{n}$, given in (2.70), has negligible impact on the solutions of the self-consistency equations, in the relevant regime of low noise, where the network works as an associative memory. As the equation for $\bar{n}$ is the same in the RS theory and in the 1RSB theory that we are considering here, the same truncation of $\bar{n}$ can be adopted here

$$\bar{n} = \frac{\bar{m}}{1+\rho}. \tag{2.75}$$

This leads to a simplified expression for g that is

$$\begin{aligned} g\left(\beta, \gamma, \boldsymbol{Y}\right) = \tilde{\beta} \frac{P}{2} \bar{m}^{P-1} \left(1+\rho\right)^{P/2-1} + \tilde{\beta} Y^{(2)} \sqrt{\gamma \frac{P}{2} \sqrt{\left(1+\rho_P\right)}^2 \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} \\ + \tilde{\beta} Y^{(1)} \sqrt{\rho \left(\frac{P}{2} \bar{m}^{P-1}\right)^2 + \gamma \frac{P}{2} \sqrt{\left(1+\rho_P\right)}^2 \bar{q}_1^{P-1}} \end{aligned} \tag{2.76}$$

where we re-scaled the noise $\tilde{\beta} = \beta' / (1+\rho)^{P/2}$. Solving numerically the self consistency equations (2.70) where the equation for $\bar{n}$ is replaced with its truncated version (2.75), leads to the phase diagram shown by the lines $M = +\infty$ in figures 1 and 2.

Next, we turn to the analysis of the ground state, i.e. to the limit $\beta \to \infty$, and we state the following

Corollary 2. *In the limit $\beta \to \infty$ and for $M \gg 1$, the 1RSB self consistency equations for the order parameters of the unsupervised DAM model are*

$$\bar{m} = 1 - 2\mathbb{E}_{Y^{(1)}} \left[\frac{1}{1 + \exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right) \mathcal{Q}\left(A\right)} \right] \tag{2.77}$$

$$\Delta\bar{q} = \bar{q}_2 - \bar{q}_1 = 4\mathbb{E}_{Y^{(1)}} \left\{ \frac{\exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right) \mathcal{Q}\left(A\right)}{\left[1 + \exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right) \mathcal{Q}\left(A\right)\right]^2} \right\} \tag{2.78}$$

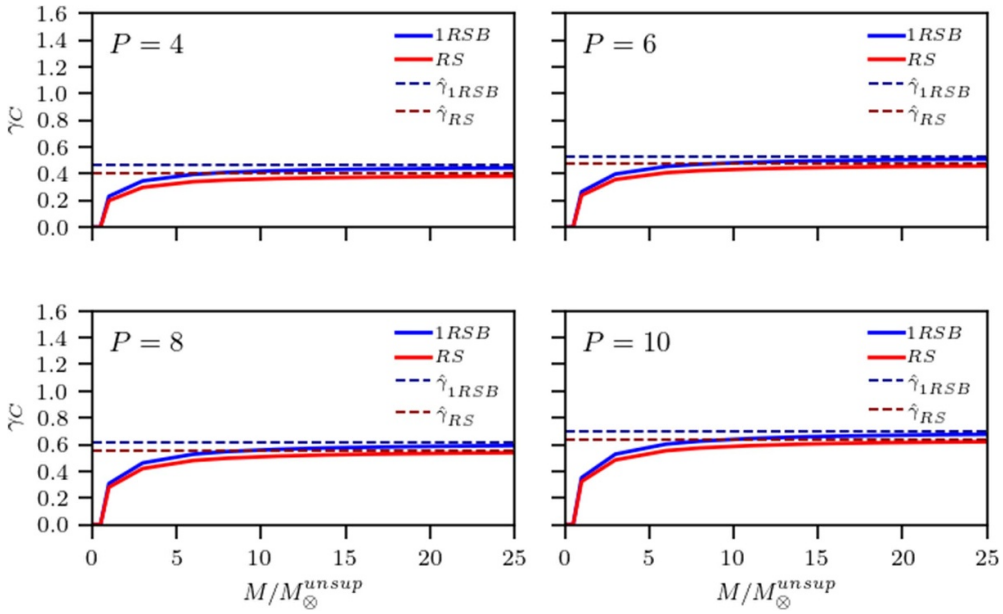


Figure 4. Critical storage capacity γ_c in the ground state for the unsupervised DAM model, for example noise $r = 0.2$, as a function of the ratio $M/M_{\otimes}^{\text{unsup}}$ both for the RS (red line) and 1RSB (blue line) assumptions. We notice that γ_c increases with M and, for $M \gg M_{\otimes}^{\text{unsup}}$, it saturates to the critical load ($\hat{\gamma}$) found numerically for the RS case in [33].

where $D = \tilde{\beta}\theta$ *and*

$$\mathcal{Q}(A) = \frac{1+\text{erf}(K^+)}{1+\text{erf}(K^-)}, \qquad K^{\pm} = \frac{DA_3^2 \pm (A_1 + A_2 Y^{(1)})}{A_3\sqrt{2}},$$

$$A_1 = \frac{P}{2}\bar{m}^{P-1}, \qquad A_2 = \sqrt{\rho A_1^2 + \gamma\sqrt{(1+\rho_P)}^2 \frac{P}{2}}, \tag{2.79}$$

$$A_3 = \sqrt{\gamma\sqrt{(1+\rho_P)}^2 \frac{P(P-1)}{2}\Delta\bar{q}}.$$

We report the proof of this corollary in appendix B.2.

Next we compute, by numerically solving (2.77) and (2.78), the ground-state critical storage capacity γ_C beyond which a black-out scenario emerges, namely $\bar{m} \neq 0$ for $\gamma < \gamma_C$ and $\bar{m} = 0$ for $\gamma > \gamma_C$. In figure 4 we plot γ_C as a function of the ratio between the number M of experienced examples and the minimum number $M_{\otimes}$ of examples required by the unsupervised DAM model (with $P > 2$) to correctly learn and retrieve an archetype, within the RS theory, which has been proved to be

$$M_{\otimes}^{\text{unsup}}(r, P, \gamma) = \gamma \frac{1}{P}\frac{1}{r^{2P}} \tag{2.80}$$

in [37]. In order to ascertain whether replica symmetry breaking alters such threshold, we plot γ_C within both the 1RSB assumption (blue line) and the RS theory (red line). We see that γ_C becomes non-zero (meaning that retrieval can occur) at the same value of $M = M_{\otimes}$ for the RS

and the 1RSB theory. This shows that the minimum number of examples that a network needs to accomplish retrieval of the archetype is the same within the RS or the 1RSB assumption.

Results show that, as in the standard Hebbian storage [44, 53, 55], the phenomenon of replica symmetry breaking induces a mild improvement in terms of the maximal storage.

3. 'Supervised' DAM

In this section we analyse the information processing capabilities of the DAM model in the so-called 'supervised' setting, where the dataset given to the network is now split into different categories, one for each archetype $\boldsymbol{\xi}^\mu$, with $\mu = 1 \dots P$. Again, we analyze the model, defined in section 3.1, via both Guerra's interpolation techniques (section 3.2) and Parisi's replica approach (section 3.3), at the first step of RSB. In appendix A we derive the instability line of the RS theory.

3.1. Model and definitions

As in the unsupervised DAM model, we consider a network of N Ising neurons $\sigma_i \in \{-1,+1\}$, $i = 1,\dots,N$, interacting via P-node interactions. We assume to have K Rademacher archetypes $\boldsymbol{\xi}^\mu \in \{-1,+1\}^N$, with $\mu = 1,\dots,K$, defined as N-dimensional vectors with entries drawn randomly and independently from the distribution (2.1). In addition, we assume to have M examples $\eta^{\mu,a} \in \{-1,+1\}^N$ for each archetype, which are corrupted version of the archetypes, with entries distributed according to (2.2). We shall refer to this model as the supervised DAM model.

Definition 6. The Hamiltonian of the supervised DAM model is

$$\mathcal{H}_N^{(P)}(\boldsymbol{\sigma}|\boldsymbol{\eta}) = -\frac{1}{\mathcal{R}^{P/2}N^{P-1}M^P}\sum_{\mu=1}^{K}\sum_{a_1,\cdots,a_P=1}^{M,\cdots,M}\left(\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\eta_{i_1}^{\mu,a_1}\cdots\eta_{i_P}^{\mu,a_P}\sigma_{i_1}\cdots\sigma_{i_P}\right) \tag{3.1}$$

where the constant $\mathcal{R} := r^2 + \frac{1-r^2}{M}$ in the denominator of the r.h.s. is included for mathematical convenience and the factor N^{P-1} ensures the Hamiltonian to be $\mathcal{O}(N)$, as explained previously for the unsupervised model, see equation (2.4).

We highlight the hidden role of a teacher that, before providing the dataset to the network, has grouped examples pertaining to the same archetype together (hence the proliferation of summations in the cost function (3.1) with respect to (2.4)).

As before, we will focus (without loss of generality) on the ability of the network to store and retrieve the first archetype $\boldsymbol{\xi}^1$, hence the Mattis magnetization $m_1(\boldsymbol{\sigma})$ provided in (2.16) remains a relevant order parameter. However, the set of order parameters for the examples, previously given by $n_{1,a}(\boldsymbol{\sigma})$ with $a = 1 \dots M$ (see (2.17)) have now to be substituted with a single order parameter

$$n_1(\boldsymbol{\sigma}) = \frac{r}{\mathcal{R}}\frac{1}{NM}\sum_{i,a=1}^{N,M}\eta_i^{1,a}\sigma_i. \tag{3.2}$$

Its probability distribution, in the thermodynamic limit, is assumed self-averaging, namely

$$\lim_{N\to+\infty}\mathbb{P}_N(n_1|\boldsymbol{\sigma}) = \delta(n_1(\boldsymbol{\sigma}) - \bar{n}), \tag{3.3}$$

as for the Mattis magnetization, while the distribution for the overlap defined in (2.18) is still assumed bimodal as in assumption 1 (see (2.20)). As for the unsupervised model, we add an

extra term in the cost function, namely $J\sum_i \xi_i^1\sigma_i$, in order to generate the moments of the Mattis magnetization m_1 by taking the derivatives of the quenched free energy w.r.t. J and, as this term is not part of the original Hamiltonian, it will be set to zero at the end of the calculations. Therefore, we write the partition function as

$$\begin{aligned}\mathcal{Z}_N^{(P)}(\boldsymbol{\eta}) &= \lim_{J\to 0}\mathcal{Z}_{N,K,M,\beta}^{(P)}(\boldsymbol{\eta};J)\\ &= \lim_{J\to 0}\sum_{\boldsymbol{\sigma}}\exp\left\{J\sum_i \xi_i^1\sigma_i + \beta'\frac{N}{2}(1+\rho)^{P/2}n_1^P(\boldsymbol{\sigma}) + \frac{\beta' P!}{2N^{P-1}\mathcal{R}^{P/2}}\sum_{\mu>1}^{K}\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\right.\\ &\quad\left.\times\left[\left(\frac{1}{M}\sum_{a_1=1}^{M}\eta_{i_1}^{\mu,a_1}\right)\cdots\left(\frac{1}{M}\sum_{a_P=1}^{M}\eta_{i_P}^{\mu,a_P}\right)\right]\sigma_{i_1}\cdots\sigma_{i_P}\right\}\end{aligned} \tag{3.4}$$

where ρ is the dataset entropy defined in (2.3), $\beta' := 2\beta/P!$ and for the first pattern, $\mu = 1$, we have used the relation $P!\sum_{i_1<\cdots<i_P} = \sum_{i_1,\cdots,i_P}$ and neglected terms which vanish in the thermodynamic limit. Next, as done for the unsupervised case, we apply the CLT to the variables in the square brackets in (3.4), namely

$$\frac{1}{K}\sum_{\mu>1}^{K}\left[\left(\frac{1}{M}\sum_{a_1=1}^{M}\eta_{i_1}^{\mu,a_1}\right)\cdots\left(\frac{1}{M}\sum_{a_P=1}^{M}\eta_{i_P}^{\mu,a_P}\right)\right] \sim \mu_1 + \lambda_{i_1,\ldots,i_P}\frac{\sqrt{\mu_2-\mu_1^2}}{\sqrt{K}}, \quad \lambda_{i_1\ldots i_P}\sim\mathcal{N}(0,1), \tag{3.5}$$

where

$$\mu_1 = \mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})}\left[\left(\frac{1}{M}\sum_{a_1=1}^{M}\eta_{i_1}^{\mu,a_1}\right)\cdots\left(\frac{1}{M}\sum_{a_P=1}^{M}\eta_{i_P}^{\mu,a_P}\right)\right] = \mathbb{E}_{\boldsymbol{\xi}}\left[\left(r\xi_{i_1}^{\mu}\right)\ldots\left(r\xi_{i_P}^{\mu}\right)\right] = 0 \tag{3.6}$$

and

$$\begin{aligned}\mu_2 &= \mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{(\boldsymbol{\eta}|\boldsymbol{\xi})}\left[\left(\frac{1}{M}\sum_{a_1=1}^{M}\eta_{i_1}^{\mu,a_1}\right)^2\cdots\left(\frac{1}{M}\sum_{a_P=1}^{M}\eta_{i_P}^{\mu,a_P}\right)^2\right]\\ &= r^2\left(1+\frac{1-r^2}{Mr^2}\right)\cdots r^2\left(1+\frac{1-r^2}{Mr^2}\right) = r^{2P}\left(1+\frac{1-r^2}{Mr^2}\right)^P\end{aligned} \tag{3.7}$$

with

$$\rho = \frac{1-r^2}{Mr^2}.$$

Thus, we can write

$$\frac{1}{K}\sum_{\mu>1}^{K}\left[\left(\frac{1}{M}\sum_{a_1=1}^{M}\eta_{i_1}^{\mu,a_1}\right)\cdots\left(\frac{1}{M}\sum_{a_P=1}^{M}\eta_{i_P}^{\mu,a_P}\right)\right] = \lambda_{i_1,\ldots,i_P}\sqrt{\frac{r^{2P}(1+\rho)^P}{K}}$$

where $\lambda_{i_1,\ldots,i_P}\sim\mathcal{N}(0,1)$.

Inserting all back in (3.4) we get

$$\mathcal{Z}^{(P)}_{N,K,M,\beta}\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda};J\right)=\exp\left\{J\sum_i \xi^1_i\sigma_i+\beta'\frac{N}{2}(1+\rho)^{P/2}n_1^P(\boldsymbol{\sigma})+\frac{\beta' P!\sqrt{K}}{2N^{P-1}}\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\lambda_{i_1,\ldots,i_P}\sigma_{i_1}\cdots\sigma_{i_P}\right\}. \tag{3.8}$$

In the next subsection we calculate the free energy of the system in the thermodynamic limit using Guerra's interpolation technique, assuming one step of replica symmetry breaking (1RSB).

3.2. 1RSB analysis via Guerra's interpolation technique

As for the unsupervised case, the plan is to construct an interpolation between the original model and a simpler one-body model, whose statistical features are as close as possible to the original one, then solve the one body model and finally obtain the solution of the original model via the fundamental theorem of calculus. Thus we define the following interpolating partition function

Definition 7. Given the interpolating parameter $t\in[0,1]$, $A_1,\ A_2,\ \psi\in\mathbb{R}$ constants to be set *a posteriori*, and the i.i.d. standard Gaussian variables $Y_i^{(b)}\sim\mathcal{N}(0,1)$ for $i=1,\ldots,N$ and $b=1,2$, the Guerra's 1-RSB interpolating partition function for the DAM model, trained by a teacher, is given by

$$\begin{aligned}\mathcal{Z}^{(P)}_2\left(\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)&:=\sum_{\{\boldsymbol{\sigma}\}}\mathcal{B}^{(P)}_{N,2}\left(\boldsymbol{\sigma}|\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t\right)\\&=\sum_{\{\boldsymbol{\sigma}\}}\exp\left[J\sum_{i=1}^N\xi^1_i\sigma_i+\frac{t\beta' N}{2M}(1+\rho)^{P/2}n_1^P(\boldsymbol{\sigma})+\psi(1-t)Nn_1(\boldsymbol{\sigma})\right.\\&\quad\left.+\sqrt{t}\frac{\beta'\sqrt{K}}{N^{P-1}}\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\lambda_{i_1,\ldots,i_P}\sigma_{i_1}\cdots\sigma_{i_P}+\sqrt{1-t}\sum_{b=1}^2A_b\sum_{i=1}^N Y_i^{(b)}\sigma_i\right].\end{aligned} \tag{3.9}$$

where $\mathcal{B}^{(P)}_{N,2}(\boldsymbol{\sigma}|\boldsymbol{\eta}^1,\boldsymbol{\lambda},\boldsymbol{Y};J,t)$ is denoted as Boltzmann factor.

As before, the introduction of the interpolating partition function gives rise to a generalized measure, average and interpolating quenched free energy (that we do not repeat here). All these generalizations retrieve the standard definitions when evaluated at $t=1$. Following Guerra's method [43], we must average out the fields $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ recursively, in the interpolating free energy resulting from the interpolating partition function (3.9), as already done in the previous section for the unsupervised case, see equations (2.26)–(2.28). Proceeding in the same way and omitting obvious details due to the similarity of the proofs in the supervised and unsupervised cases, we state directly the next.

Proposition 2. *In the thermodynamic limit $N \to \infty$, within the 1RSB assumption and under the assumption 2, the quenched free energy for the supervised DAM model, reads as*

$$-\beta' \mathcal{F}^{(P)}(J) = \ln 2 + \frac{1}{\theta}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \ln \mathbb{E}_{Y^{(2)}} \left\{ \cosh^\theta \left[J\xi^1 + \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \hat{\eta}_M \right.\right.$$
$$\left.\left. + Y^{(1)} \beta' \sqrt{\gamma \frac{P}{2} \bar{q}_1^{P-1}} + Y^{(2)} \beta' \sqrt{\gamma \frac{P}{2} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1} \right)} \right] \right\} + \frac{\beta'^2}{4} \gamma \left[1 - P\bar{q}_2^{P-1} + (P-1)\bar{q}_2^P \right]$$
$$- \frac{\beta'}{2}(P-1)(1+\rho)^{P/2} \bar{n}^P - \frac{\beta'^2}{4} \gamma (P-1) \theta \left(\bar{q}_2^P - \bar{q}_1^P \right), \tag{3.10}$$

with $\bar{n}$, $\bar{q}_1$, $\bar{q}_2$ fulfilling the following self-consistency equations

$$\bar{n} = \frac{1}{1+\rho}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) \tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)} \frac{1}{r} \hat{\chi}_M \right],$$

$$\bar{q}_1 = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) \tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)} \right]^2, \tag{3.11}$$

$$\bar{q}_2 = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) \tanh^2 g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)} \right],$$

where

$$g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) = \beta' \frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \hat{\eta}_M + Y^{(1)} \beta' \sqrt{\gamma \frac{P}{2} \bar{q}_1^{P-1}} + Y^{(2)} \beta' \sqrt{\gamma \frac{P}{2} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1} \right)} \tag{3.12}$$

and $\beta' = 2\beta/P!$. Furthermore, as $\bar{m} = -\beta' \nabla_J \mathcal{F}^{(P)}(J)|_{J=0}$, we have

$$\bar{m} = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right) \tanh g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g\left(\beta,\gamma,\boldsymbol{Y},\boldsymbol{\eta}^1\right)} \xi^1 \right]. \tag{3.13}$$

The proof of the aforementioned proposition is lengthy but the steps to follow are identical to the ones provided for the unsupervised case.

3.3. 1RSB analysis via Parisi's replica trick

As stated in the previous section, the core of this approach consists in writing the logarithm of the partition function as $\lim_{n\to 0} (\mathbb{E}\mathcal{Z}(J)^n - 1)/n$ where

$$\mathbb{E}\mathcal{Z}_N^n(J) = \mathbb{E} \sum_{\boldsymbol{\sigma}^1 \dots \boldsymbol{\sigma}^n} \exp \left[\frac{\beta' P!}{2\mathcal{R}^{P/2} N^{P-1} M^P} \sum_{a=1}^{n} \sum_{\mu=1}^{K} \left(\sum_{i_1<\dots<i_P}^{N,\dots,N} \sum_{A_1,\dots,A_P}^{M,\dots,M} \eta_{i_1}^{\mu,A_1} \cdots \eta_{i_P}^{\mu,A_P} \sigma_{i_1}^{(a)} \cdots \sigma_{i_P}^{(a)} \right) \right.$$
$$\left. + J \sum_{a,i}^{n,N} \xi_i^1 \sigma_i^{(a)} \right]. \tag{3.14}$$

Proceeding as in the unsupervised case, we compute separately the signal term and the noise term. The former, accounting for the examples pertaining to the first archetype, reads as

$$\mathbb{E}\mathcal{Z}^n_{\text{signal}}(J) = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)} \sum_{\boldsymbol{\sigma}^1\dots\boldsymbol{\sigma}^n} \exp\left[\frac{\beta' N(1+\rho)^{P/2}}{2}\sum_{a=1}^{n}\left(\frac{r}{NM\mathcal{R}}\sum_{i}^{N}\sum_{A}^{M}\eta_i^{1,A}\sigma_i^{(a)}\right)^P + J\sum_{a=1}^{n}\sum_{i=1}^{N}\xi_i^1\sigma_i^{(a)}\right], \tag{3.15}$$

while the latter accounting for the examples pertaining to all the other archetypes but the first one, can be written as

$$\mathbb{E}\mathcal{Z}^n_{\text{noise}} = \mathbb{E}_{\boldsymbol{\lambda}} \sum_{\boldsymbol{\sigma}^1\dots\boldsymbol{\sigma}^n} \exp\left[\frac{\beta' P!\sqrt{K}}{2N^{P-1}}\sum_{a=1}^{n}\left(\sum_{i_1<\dots<i_P}^{N,\dots,N}\lambda_{i_1,\dots,i_P}\sigma_{i_1}^{(a)}\cdots\sigma_{i_P}^{(a)}\right)\right] \tag{3.16}$$

where now $\mathbb{E}_{\boldsymbol{\lambda}}$ is the Gaussian average w.r.t. $\boldsymbol{\lambda}$. Performing the integration over $\boldsymbol{\lambda}$ and inserting the definitions of the order parameters as done for the unsupervised setting, we can write

$$\mathbb{E}\mathcal{Z}^n_N(J) = \int\prod_a \mathrm{d}n_1^{(a)}\prod_a\frac{\mathrm{d}z_1^{(a)}}{2\pi}\int\prod_{a,b}\mathrm{d}q_{ab}\int\prod_{a,b}\frac{\mathrm{d}p_{ab}}{2\pi}\exp\left(-NA\left[\boldsymbol{Q},\boldsymbol{P},\boldsymbol{N},\boldsymbol{Z}\right]\right) \tag{3.17}$$

where

$$A\left[\boldsymbol{Q},\boldsymbol{P},\boldsymbol{N},\boldsymbol{Z}\right] = -\frac{i}{N}\sum_{a,b}p_{ab}q_{ab} - \frac{1}{4}\beta'^2\gamma\sum_{a,b}^{n,n}q_{ab}^P - \frac{\beta'(1+\rho)^{P/2}}{2}\sum_a\left(n_1^{(a)}\right)^P - \frac{i}{N}\sum_a n_1^{(a)}z_1^{(a)} - \mathbb{E}\log\left[\sum_{\boldsymbol{\sigma}}\exp\left(-\frac{i}{N}\sum_{a,b}p_{ab}\sigma^{(a)}\sigma^{(b)} - \frac{i}{N(1+\rho)}\sum_a\hat{\eta}_M z_1^{(a)}\sigma^{(a)} + J\sum_a\xi^1\sigma^{(a)}\right)\right]. \tag{3.18}$$

and $\hat{\eta}_M = (rM)^{-1}\sum_{A=1}^{M}\eta^{1,A}$. Assuming (as usual within the replica method) that the limits $N\to\infty$ and $n\to 0$ can be interchanged, the integrals can be performed by steepest descent. This leads to

$$-\beta'\mathcal{F}^{(P)}(J) = \lim_{n\to 0}\frac{1}{n}A\left[\boldsymbol{Q}^\star,\boldsymbol{P}^\star,\boldsymbol{Z}^\star,\boldsymbol{N}^\star\right] \tag{3.19}$$

where $\boldsymbol{Q}^\star,\boldsymbol{P}^\star,\boldsymbol{Z}^\star,\boldsymbol{N}^\star$, are solution of the saddle point equations

$$\frac{\partial A}{\partial p_{ab}} = 0 \quad\Rightarrow\quad q^*_{ab} = \langle\sigma^a\sigma^b\rangle_{\text{eff}} \tag{3.20}$$

$$\frac{\partial A}{\partial z_1^{(a)}} = 0 \quad\Rightarrow\quad \left(n_1^{(a)}\right)^* = \frac{r}{\mathcal{R}}\sum_{A=1}^{M}\mathbb{E}\langle\eta^{1,A}\sigma^a\rangle_{\text{eff}} \tag{3.21}$$

$$\frac{\partial A}{\partial q_{ab}} = 0 \quad\Rightarrow\quad p^*_{ab} = i\beta'^2\frac{P}{4}\gamma N(q^*_{ab})^{P-1} \tag{3.22}$$

$$\frac{\partial A}{\partial n_1^{(a)}} = 0 \quad \Rightarrow \quad \left(z_1^{(a)}\right)^* = i\beta' \frac{P}{2}(1+\rho)^{P/2} N \left[\left(n_1^{(a)}\right)^*\right]^{P-1} \tag{3.23}$$

which provides a set of self-consistency equations for $\boldsymbol{Q}$, $\boldsymbol{N}$, $\boldsymbol{P}$, and $\boldsymbol{Z}$. Using the last two equations to eliminate $\boldsymbol{P}$ and $\boldsymbol{Z}$ from the description, we finally obtain

$$\begin{aligned} -\beta' \mathcal{F}(J) = \lim_{n\to 0} \frac{1}{n} \Bigg\{ & \frac{1}{4}\beta'^2 \gamma (P-1) \sum_{a,b}^{n,n} [q_{ab}^*]^P + \frac{\beta'(1+\rho)^{P/2}}{2}(P-1) \sum_a \left[\left(n_1^{(a)}\right)^*\right]^P \\ & - \mathbb{E} \log \Bigg[\sum_{\boldsymbol{\sigma}} \exp \Bigg(\beta'^2 \frac{P}{4} \gamma \sum_{a,b} [q_{ab}^*]^{P-1} \sigma^{(a)} \sigma^{(b)} + \beta' \frac{P}{2} (1+\rho)^{P/2-1} \sum_a \hat{\eta}_M \left[\left(n_1^{(a)}\right)^*\right]^{P-1} \sigma^{(a)} \\ & + J \sum_a \xi^1 \sigma^{(a)} \Bigg) \Bigg] \Bigg\}. \end{aligned} \tag{3.24}$$

To make progress, we need to find the form of $\boldsymbol{Q}$ and $\boldsymbol{N}$ in the limit $n \to 0$. Again, we use the 1RSB ansatz for the two-replica overlap provided in (2.63), while we assume that $n_{1,a}$ is self-averaging, i.e. $n_1^{(a)} = \bar{n}$. Proceeding analogously to the unsupervised case, after taking the limit $n \to 0$ we find the same expression for the quenched free energy (see (3.10)) and order parameters (see 3.11) previously obtained by using Guerra's interpolation technique.

3.4. Limiting cases: $M \to \infty$ and $\beta \to \infty$

Now we state the next corollaries concerning the large datasets $M \to \infty$ limit and the ground state $\beta \to \infty$ limit. We omit their proofs since these are trivial variations of those provided for the unsupervised case (cfr corollaries 1 and 2).

Corollary 3. *In the large dataset limit $M \to \infty$, the 1RSB self-consistency equations for the order parameters of the supervised DAM model can be expressed as*

$$\begin{aligned} \bar{n} &= \frac{\bar{m}}{1+\rho} + \beta' \frac{P}{2} \rho (1+\rho)^{P/2-2} [1 + (\theta - 1)\bar{q}_2 - \theta \bar{q}_1] \bar{n}^{P-1}, \\ \bar{q}_1 &= \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta, \gamma, \boldsymbol{Y}) \tanh g(\beta, \gamma, \boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta, \gamma, \boldsymbol{Y})} \right]^2, \\ \bar{q}_2 &= \mathbb{E}_{Y^{(1)}} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta, \gamma, \boldsymbol{Y}) \tanh^2 g(\beta, \gamma, \boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta, \gamma, \boldsymbol{Y})} \right], \\ \bar{m} &= \mathbb{E}_{1} \left[\frac{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta, \gamma, \boldsymbol{Y}) \tanh g(\beta, \gamma, \boldsymbol{Y})}{\mathbb{E}_{Y^{(2)}} \cosh^\theta g(\beta', \gamma, \boldsymbol{Y})} \right]. \end{aligned} \tag{3.25}$$

where the expression of g reads as

$$\begin{aligned} g(\beta, \gamma, \boldsymbol{Y}) = & \beta' \frac{\boldsymbol{P}}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} + Y^{(2)} \beta' \sqrt{\gamma \frac{P}{2} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} \\ & + Y^{(1)} \beta' \sqrt{\rho \left(\frac{P}{2} \bar{n}^{P-1} (1+\rho)^{P/2-1} \right)^2 + \gamma \frac{P}{2} \bar{q}_1^{P-1}}. \end{aligned} \tag{3.26}$$

Moreover, if we use the truncated expression for $\bar{n}$, namely

$$\bar{n} = \frac{\bar{m}}{1+\rho} \tag{3.27}$$

we get the simplified expression of (3.26) that is

$$g(\beta,\gamma,\boldsymbol{Y}) = \tilde{\beta}\frac{P}{2}\bar{m}^{P-1} + Y^{(2)}\tilde{\beta}\sqrt{\gamma\frac{P}{2}(1+\rho)^{P}\left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} + Y^{(1)}\tilde{\beta}\sqrt{\rho\left(\frac{P}{2}\bar{m}^{P-1}\right)^2 + \gamma(1+\rho)^{P}\frac{P}{2}\bar{q}_1^{P-1}}\,, \tag{3.28}$$

where $\tilde{\beta} = \beta'/(1+\rho)^{P/2}$.

Corollary 4. *The ground-state (i.e. $\beta \to \infty$) self-consistency equations for the order parameters of the theory in the large dataset limit (i.e. $M \gg 1$) and under the 1RSB assumption are*

$$\bar{n} = \frac{\bar{m}}{(1+\rho)} \tag{3.29}$$

$$\bar{m} = 1 - 2\mathbb{E}_{Y^{(1)}}\left[\frac{1}{1+\exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right)\mathcal{Q}(A)}\right] \tag{3.30}$$

$$\Delta\bar{q} = \bar{q}_2 - \bar{q}_1 = 4\mathbb{E}_{Y^{(1)}}\left\{\frac{\exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right)\mathcal{Q}(A)}{\left[1+\exp\left(2D\left(A_1 + A_2 Y^{(1)}\right)\right)\mathcal{Q}(A)\right]^2}\right\} \tag{3.31}$$

where $D = \tilde{\beta}\theta$ and

$$\mathcal{Q}(A) = \frac{1+\mathrm{erf}(K^+)}{1+\mathrm{erf}(K^-)}, \qquad K^{\pm} = \frac{DA_3^2 \pm \left(A_1 + A_2 Y^{(1)}\right)}{A_3\sqrt{2}},$$
$$A_1 = \frac{P}{2}\bar{m}^{P-1}, \qquad A_2 = \sqrt{\rho\left(\frac{P\bar{m}^{P-1}}{2}\right)^2 + \gamma(1+\rho)^{P}\frac{P}{2}}, \tag{3.32}$$
$$A_3 = \sqrt{\gamma(1+\rho)^{P}\frac{P(P-1)}{2}\Delta\bar{q}}.$$

By numerically solving the self-consistency equations (3.25), we obtain the phase diagram shown in figure 5, for different values of P and different values of the dataset entropy ρ, as shown in the legend.

In [38] it has been proven that the minimum number of examples $M_{\otimes}(r,\gamma)$ required by the supervised DAM model to correctly learn and retrieve an archetype, is given by

$$M_{\otimes}(r,\gamma) = (1+2\gamma)\frac{1}{r^2} \tag{3.33}$$

where $\gamma \neq 0$. In figure 6 we plot the critical storage γ_C in the ground state versus the ratio $M/M_{\otimes}$. As previously observed for the unsupervised setting, the phenomenon of replica symmetry breaking slightly increases the critical storage and it does not alter the minimum number of examples required for learning.

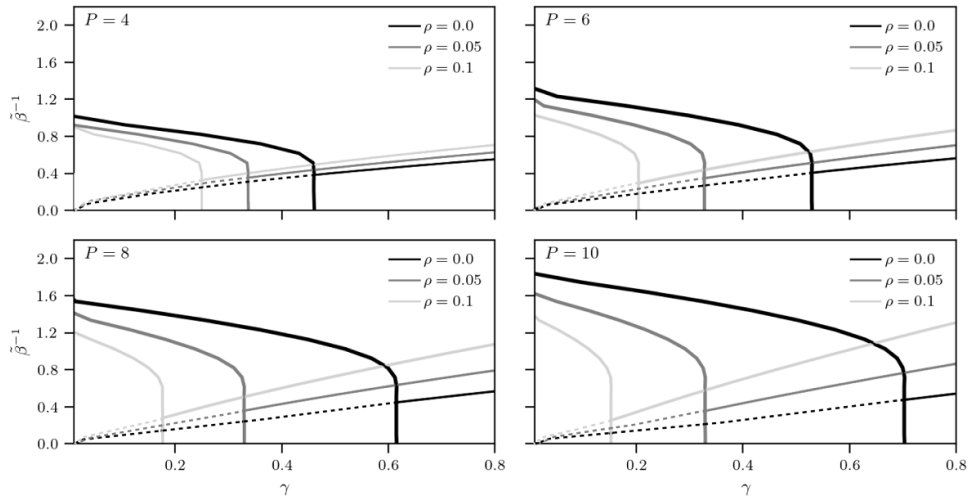


Figure 5. Phase Diagram of the supervised DAM model for different values of the dataset entropy ρ, ranging from $\rho = 0.1$ (light grey), 0.05 (grey) up to 0 (black). The interpretation of the different regions is the same as in the unsupervised model. As ρ decrease (i.e. as more information is provided to the network), the instability region shrinks and the retrieval region expands, similarly to what was observed earlier upon increasing M.

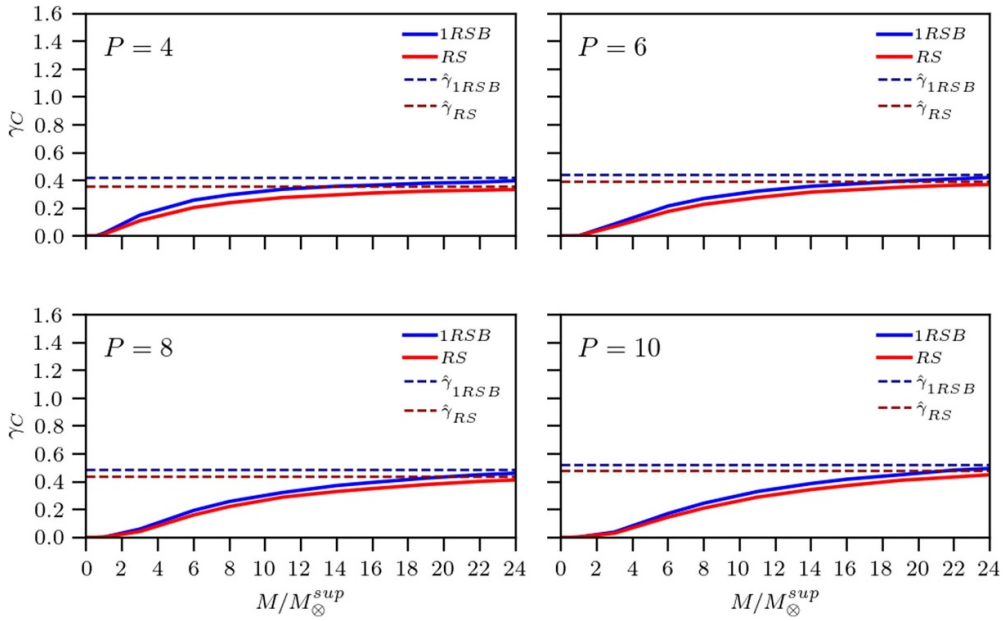


Figure 6. Critical storage capacity γ_c in the ground state for the supervised DAM model for example noise $r = 0.2$, as a function of the ratio $M/M_{\otimes}^{sup}$, both for the RS (red line) and 1RSB (blue line) assumptions. Different panels show results for different values of $P = 4,\ 6,\ 8,\ 10$.

4. Conclusions and outlooks

This manuscript analyses the equilibrium behaviour of DAM trained with or without the supervision of a teacher, within a 1RSB assumption, thus extending previous analysis carried out at a RS level [37, 38]. The unsupervised and supervised settings differ in the choice of the couplings (which involve P nodes) and are given by

$$J^{\text{unsup}}_{i_1 i_2 \dots i_P} \sim \frac{1}{MN^{P-1}} \sum_{\mu=1}^{K} \sum_{a=1}^{M} \eta^{\mu,a}_{i_1} \dots \eta^{\mu,a}_{i_P}, \tag{4.1}$$

$$J^{\text{sup}}_{i_1 i_2 \dots i_P} \sim \frac{1}{M^P N^{P-1}} \sum_{\mu=1}^{K} \sum_{a_1,\dots,a_P=1}^{M,\dots,M} \eta^{\mu,a_1}_{i_1} \dots \eta^{\mu,a_P}_{i_P} \tag{4.2}$$

respectively, where $\{\boldsymbol{\eta}^{\mu}_{a}\}^{\mu=1,\dots,K}_{a=1,\dots,M}$ are perturbed versions of the unknown archetypes $\{\boldsymbol{\xi}^{\mu}\}^{\mu=1,\dots,K}$. The network does not experience the archetypes directly, instead it has to infer them from the supplied examples. For both the settings, we obtained explicit expressions for the quenched free energy and derived full phase diagrams. In doing so, we proved a full equivalence, at 1RSB level, between two different approaches, namely Guerra's telescopic interpolation [43] and Parisi's RSB theory [2]. In addition, we derived (in appendix A) the De Almeida-Thouless line, which marks the onset of the instability of the replica symmetric description, and below which the 1RSB description should be preferred.

The main differences brought about by the RSB description, with respect to the RS one, consist in the disappearance of the instability region within the retrieval zone of the phase diagram close to saturation (as standard in glassy statistical mechanics [4]) and in a slight improvement of the value of the critical storage. Importantly, the threshold for learning, both in the supervised and unsupervised settings, is not influenced by replica symmetry breaking, i.e. the minimum number of examples required to infer the archetypes is the same in the RS and 1RSB description. Interestingly, the optimal value of the Parisi's parameter θ, that controls the distribution of the overlaps in the 1RSB scenario, is not influenced by the dataset entropy and is equal to that of the classical Hopfield model. From the mathematical viewpoint, possible future developments would be relaxing the constraints of a self-averaging Mattis magnetization and inspecting how the learning and retrieval properties of these networks change with different kind of noise: in this work we have focused on multiplicative noise, however the use of additive noise has lately gained large popularity in generative models for machine learning (see for instance [56]). Furthermore, within the framework of multiplicative noise, an interesting outlook would be considering corruptions of archetypes which also consist of blank (in addition to inverted) entries, as done recently in [57] for networks away from saturation. Their operation in the saturated regime and the effects of replica symmetry breaking have not yet been investigated: we plan to report soon on these topics.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

All the authors acknowledge the stimulating research environment provided by the Alan Turing Institute's Theory and Methods Challenge Fortnights event *Physics-informed*

Machine Learning. Albanese acknowledges Ermenegildo Zegna Founder's Scholarship, UMI (Unione Matematica Italiana), INdAM—GNFM Project (CUP E53C22001930001) and PRIN grant *Stochastic Methods for Complex Systems* N. 2017JFFHS for financial support and King's College London for kind hospitality. Alessandrelli acknowledges INdAM (Istituto Nazionale d'Alta Matematica) and Unisalento for support via PhD-AI. Barra's research is supported by MAECI via the BULBUL grant (Italy-Israel collaboration), Project N. F85F21006230001 *Brain-inspired Ultra-fast and Ultra-sharp machines for assisted healthcare* and by MUR via the PRIN-2022 grant *Statistical Mechanics of Learning Machines: from algorithmic and information-theoretical limits to new biologically inspired paradigms*, Project N. 20229T9EAT that are gratefully acknowledged.

Appendix A. Instability of the RS solution: AT lines

In the main text we have analysed the equilibrium behaviour of supervised and unsupervised DAM models under the assumption of replica symmetry breaking. While such phenomenon is expected in these models, a formal proof that the replica symmetric theory becomes unstable in certain ranges of the control parameters has not been provided in the literature. In this section we provide such a proof and we derive the critical line of the RS instability in the phase diagram, for both the unsupervised and the supervised DAM models, separately. We will use the method recently introduced in [39], which provides a simple alternative to the method originally introduced by de Almeida and Thouless in [58], as it does not require to compute the so-called replicon (i.e. the smallest eigenvalue of the spectrum) of the Hessian of the quadratic fluctuations of the free-energy around its RS value and it does not rely on the availability of an 'ansatz-free' expression for the free-energy.

A.1. AT line for DAM in unsupervised setting

Following the method introduced in [39], we aim to determine the region in the phase diagram where the quenched free energy evaluated within the 1RSB approximation, that from now on we denote for convenience as $\mathcal{F}^{(P)}_{1RSB}(\bar{n},\bar{q}_2,\bar{q}_1|\theta)$, is smaller than the free energy evaluated within the RS assumption, $\mathcal{F}^{(P)}_{RS}(\bar{n}',\bar{q})$, in the limit $\theta\to 1$, where the transition from RS to RSB is expected to occur.

We start by recalling the expression for $\mathcal{F}^{(P)}_{1RSB}(\bar{n},\bar{q}_2,\bar{q}_1|\theta)$, given in (2.35), with $\bar{n}$, $\bar{q}_1$ and $\bar{q}_2$ determined from the self-consistency equations (2.37), and by providing the expression for the quenched free energy within the RS assumption, as derived in [37]

$$\begin{aligned}-\beta'\mathcal{F}^{(P)}_{RS}(\bar{n}',\bar{q}) = &-\frac{\beta'^2(1+\rho)^{P/2}}{2}(P-1)\bar{n}'^P+\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2}{4(1+\rho)^P}\left(1-\bar{q}^P\right)\\ &-\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2}{4(1+\rho)^P}\bar{q}^{P-1}(1-\bar{q})+\ln 2\\ &+\mathbb{E}\ln\cosh\left(\beta'(1+\rho)^{P/2-1}\frac{P}{2}\hat{\eta}_M\bar{n}'^{P-1}+\beta' z\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}^{P-1}}\right)\end{aligned} \tag{A.1}$$

where $\mathbb{E} = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_z$, $\hat{\eta}_M := \frac{1}{rM}\sum_{a=1}^{M}\eta^{1,a}$, and $\bar{n}'$ and $\bar{q}$ fulfill the following self-consistency equations

$$\bar{n}' = \frac{1}{1+\rho}\mathbb{E}\left\{\tanh\left[\beta'\frac{P}{2}\bar{n}'^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M + z\beta'\sqrt{\gamma\frac{P}{2}\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\bar{q}^{P-1}}\right]\hat{\eta}_M\right\} \tag{A.2}$$

$$\bar{q} = \mathbb{E}\left\{\tanh^2\left[\beta'\frac{P}{2}\bar{n}'^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M + z\beta'\sqrt{\gamma\frac{P}{2}\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\bar{q}^{P-1}}\right]\right\}. \tag{A.3}$$

We note that for $\theta = 1$, $\bar{q}_1 = \bar{q}$, $\bar{n} = \bar{n}'$ and the 1RSB expression for the quenched free-energy reduces to the RS one. Now, we expand the 1RSB expression for the quenched free energy $\mathcal{F}^{(P)}_{\text{1RSB}}(\bar{n},\bar{q}_2,\bar{q}_1|\theta)$, as given in (2.35), around $\theta = 1$, using $\mathcal{F}^{(P)}_{\text{1RSB}}(\bar{n},\bar{q}_2,\bar{q}_1|\theta)|_{\theta=1} = \mathcal{F}^{(P)}_{\text{RS}}(\bar{n}',\bar{q})$

$$\mathcal{F}_{\text{1RSB}}(\bar{n},\bar{q}_2,\bar{q}_1|\theta) = \mathcal{F}_{\text{RS}}(\bar{n}',\bar{q}) + (\theta-1)\,\partial_\theta\mathcal{F}_{\text{1RSB}}(\bar{n},\bar{q}_2,\bar{q}_1|\theta)|_{\theta=1}. \tag{A.4}$$

Since $\bar{q}_1$ and $\bar{q}_2$ are determined through the self-consistency equations (2.37), they depend on θ as well, so we have to expand (2.37) around $\theta = 1$ too. Following [39] closely, we obtain

$$\bar{q}_1 = \bar{q} + (\theta-1)A(\bar{n}',\bar{q},\tilde{q}_2(\bar{n}',\bar{q})) \tag{A.5}$$

$$\bar{q}_2 = \tilde{q}_2(\bar{n}',\bar{q}) + (\theta-1)B(\bar{n}',\bar{q},\tilde{q}_2(\bar{n}',\bar{q})) \tag{A.6}$$

where $\bar{q}$ is the solution of (A.3), $\tilde{q}_2(\bar{n}',\bar{q})$ is the solution of the following self-consistency equation

$$\bar{q}_2 = \mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\bar{q},\tilde{q}_2(\bar{n}',\bar{q}))\tanh^2 g(\bar{n}',\bar{q},\tilde{q}_2(\bar{n}',\bar{q}))}{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\bar{q},\tilde{q}_2(\bar{n}',\bar{q}))}\right] \tag{A.7}$$

where

$$\begin{aligned} g(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q}) = {} & \beta'\frac{P}{2}\bar{n}'^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M + Y^{(1)}\beta'\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\bar{q}^{P-1}} \\ & + Y^{(2)}\beta'\sqrt{\gamma\frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P}\frac{P}{2}\left((\tilde{q}_2(\bar{n}',\bar{q}))^{P-1} - \bar{q}^{P-1}\right)}. \end{aligned} \tag{A.8}$$

and the functions $A(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})$ and $B(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})$ are given by

$$
\begin{aligned}
&A\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) \\
&\quad = 2\mathbb{E}_{Y^{(1)}}\left\{\frac{\mathbb{E}_{Y^{(2)}} \ln\cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) \sinh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) \tanh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right)}{\mathbb{E}_{Y^{(2)}} \cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right)}\right\} \\
&\qquad - 2\mathbb{E}_{Y^{(1)}}\left\{\frac{\mathbb{E}_{Y^{(2)}} \ln\cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) \cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right)}{\left(\mathbb{E}_{Y^{(2)}} \cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right)\right)^2}\right. \\
&\qquad \left.\cdot \frac{\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \tanh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^2}\right\},
\end{aligned}
\tag{A.9}
$$

$$
\begin{aligned}
B(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) = 2\mathbb{E}_{Y^{(1)}}&\left\{\frac{\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \tanh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^2}\right. \\
&\left.\cdot \frac{\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \log\cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^2}\right\} \\
&- 2\mathbb{E}_{Y^{(1)}}\left\{\frac{\left(\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \tanh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^2}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^3}\right. \\
&\left.\cdot \frac{\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \log\cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^3}\right\},
\end{aligned}
\tag{A.10}
$$

respectively. Similarly, we can expand $\bar{n}$ as

$$
\bar{n} = \bar{n}' + (\theta - 1)\, C\left(\bar{n}', \bar{q}, \tilde{q}_2\left(\bar{n}', \bar{q}\right)\right) \tag{A.11}
$$

where

$$
\begin{aligned}
&C(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \\
&\quad = \mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \log\cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}\right] \\
&\qquad - \mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}} \sinh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q}) \log\cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}{\left(\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})\right)^2}\right].
\end{aligned}
\tag{A.12}
$$

In the limit $\theta \to 1$, equation (A.4) implies that $\mathcal{F}^{(P)}_{\mathrm{1RSB}}(\bar{n}', \bar{q}_2(\bar{q}), \bar{q}|\theta) < \mathcal{F}^{(P)}_{\mathrm{RS}}(\bar{n}', \bar{q})$ when we have that $\partial_\theta \mathcal{F}_{\mathrm{1RSB}}(\bar{n}', \bar{q}_2(\bar{q}), \bar{q}|\theta)|_{\theta=1} > 0$. Next, we evaluate

$$
\begin{aligned}
K\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) &:= \partial_\theta\left(-\beta' \mathcal{F}^{(P)}_{\mathrm{1RSB}}\left(\bar{n}', \bar{q}_2\left(\bar{q}\right), \bar{q}|\theta\right)\right)|_{\theta=1} \\
&= -\frac{\beta'^2 \gamma \sqrt{(1+\rho_P)}^2}{4(1+\rho)^P}(P-1)\left[\left(\bar{q}_2\left(\bar{q}\right)\right)^P - \bar{q}^P\right] - \frac{\beta'^2 \gamma \sqrt{(1+\rho_P)}^2}{4(1+\rho)^P} P\left[\left(\bar{q}_2\left(\bar{q}\right)\right)^{P-1} - \bar{q}^{P-1}\right] \\
&\quad - \mathbb{E}\ln\cosh\left(\beta' \frac{P}{2} \bar{n}'^{P-1}(1+\rho)^{P/2-1} \hat{\eta}_M + Y\beta' \sqrt{\gamma \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \bar{q}^{P-1}}\right) \\
&\quad + \mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}} \cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right) \log\cosh g\left(\bar{n}', \tilde{q}_2\left(\bar{n}', \bar{q}\right), \bar{q}\right)}{\mathbb{E}_{Y^{(2)}} \cosh g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})}\right].
\end{aligned}
\tag{A.13}
$$

In order to determine the sign of the expression above, it is useful to note that $K(\bar{n}',\bar{q},\bar{q})=0$, as the last term of (A.8) vanishes so, the last two addends in (A.13) elide each other. This is as expected, as $\bar{q}$ is an extremum of the RS free-energy, which is retrieved for $\theta=1$. Next, we study $K(\bar{n}',x,\bar{q})$ as a function of $x\in[0,\bar{q}]$ and locate its extrema. These are found from

$$\partial_x K(\bar{n}',x,\bar{q}) = -\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2(P-1)Px^{P-2}}{4(1+\rho)^P} \cdot\left[x-\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\sinh g(\bar{n}',\bar{q},x)\tanh g(\bar{n}',\bar{q},x)}{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\bar{q},x)}\right]\right]=0 \tag{A.14}$$

as

$$x=\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\sinh g(\bar{n}',\bar{q},x)\tanh g(\bar{n}',\bar{q},x)}{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\bar{q},x)}\right]\equiv\tilde{q}_2(\bar{n}',\bar{q})\,. \tag{A.15}$$

where the last equality follows from algebraic manipulations of trigonometric functions.

Given that $K(\bar{n}',x,\bar{q})$ vanishes for $x=\bar{q}$, if the extremum $x=\tilde{q}_2(\bar{n}',\bar{q})$ is global in the domain considered, we must have that $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})>0$ if $x=\tilde{q}_2(\bar{n}',\bar{q})$ is a maximum and $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})<0$ if $x=\tilde{q}_2(\bar{n}',\bar{q})$ is a minimum. Evaluating

$$\begin{aligned}\partial_x^2 K(\bar{n}',x,\bar{q})\,|_{x=\tilde{q}_2(\bar{n}',\bar{q})} = &-\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2(P-1)P}{4(1+\rho)^P}\left(\tilde{q}_2(\bar{n}',\bar{q})\right)^{P-2}\\ &\cdot\left[1-\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2}{2(1+\rho)^P}(P-1)P\left(\tilde{q}_2(\bar{n}',\bar{q})\right)^{P-2}\right.\\ &\left.\times\,\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\mathrm{sech}^3 g(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})}{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})}\right]\right].\end{aligned} \tag{A.16}$$

we have that if the expression in the square brackets is positive, namely if the parameter $\gamma\beta'^2$ satisfies

$$\frac{\beta'^2\gamma\sqrt{(1+\rho_P)}^2}{2(1+\rho)^P}(P-1)P\left(\tilde{q}_2(\bar{n}',\bar{q})\right)^{P-2}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\mathrm{sech}^3 g(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})}{\mathbb{E}_{Y^{(2)}}\cosh g(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})}\right]<1, \tag{A.17}$$

$K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})<0$ and $\mathcal{F}^{(P)}_{\mathrm{1RSB}}(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q}|\theta)-\mathcal{F}^{(P)}_{\mathrm{RS}}(\bar{n}',\bar{q})=-(\theta-1)K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})/\beta'<0$, hence the RS theory is unstable. We plot the line (A.17) in figure 7, for different values of M, in the parameter space $(\gamma,\tilde{\beta})$, where $\tilde{\beta}=\beta'/(1+\rho)^{P/2}$, together with the critical lines delimiting the retrieval region (top plots) and the spin-glass region (bottom plots), within the RS and the RSB theory, respectively. We note that the above recovers the expression for the RS instability line found in DAM models [39] in the limits $r\to 0$ or $M\to\infty$, where the values of ρ_P and ρ vanish.

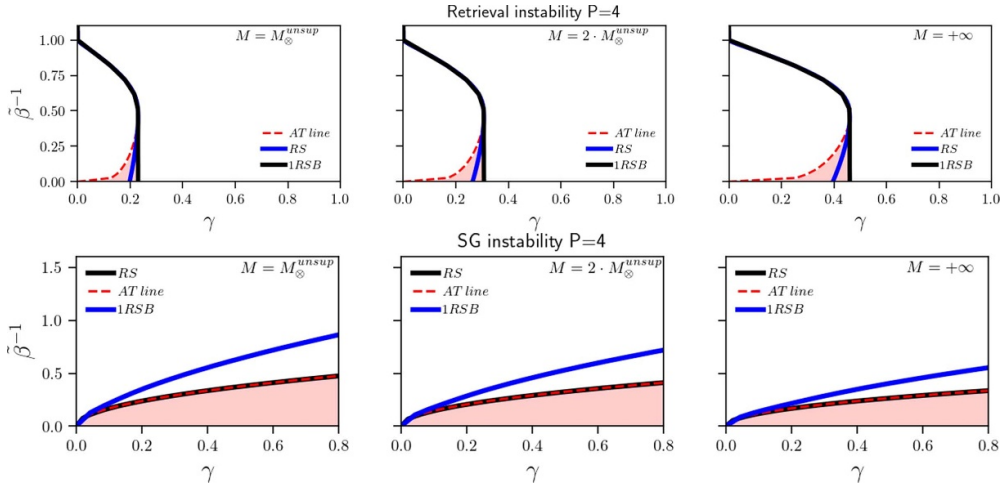


Figure 7. The AT line (marking the onset of instability of the RS theory) is shown as a dashed line in the parameter space of storage load γ and scaled noise $\tilde{\beta} = \beta'/(1+\rho)^{P/2}$ for the unsupervised DAM model, with $P = 4$. For comparison, we also show the critical lines delimiting the retrieval region (top plots) and the spin-glass region (bottom plots), within the RS and the RSB theory, respectively, as obtained in the phase diagram in figure 1. Different columns correspond to different numbers of examples, M, presented to the network, namely $M = M_{\otimes}^{\text{unsup}}$ (left), $M = 2M_{\otimes}^{\text{unsup}}$ (middle) and $M = +\infty$ (right). We note that the AT line coincides with the line delimiting the SG region withing the RS theory.

A.2. AT line for DAM in supervised setting

In this section we derive the RS instability line for the supervised DAM model. We start by providing the expression for the quenched free-energy in RS assumption as derived in [14] for the standard Dense Hopfield Model.

$$-\beta' \mathcal{F}_{RS}^{(P)}(\bar{n}', \bar{q}) = \ln 2 - \frac{\beta'^2}{2}(P-1)(1+\rho)^{P/2}\bar{n}'^P + \frac{\beta'^2\gamma}{4}\left(1-\bar{q}^P\right) - \frac{\beta'^2\gamma P}{4}\bar{q}^{P-1}(1-\bar{q}) + \mathbb{E}_\xi \mathbb{E}_{(\eta|\xi)} \mathbb{E}_z \ln\cosh\left(\beta' \frac{P}{2}(1+\rho)^{P/2-1}\hat{\eta}_M \bar{n}^{P-1} + \beta' z \sqrt{\gamma \frac{P}{2}\bar{q}^{P-1}}\right) \tag{A.18}$$

with $\beta' := 2\beta/P!$ and where $\bar{q}$ and $\bar{n}$ satisfy the self-consistency equations

$$\bar{n}' = \frac{1}{1+\rho}\mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_z\left\{\tanh\left[\beta' \frac{P}{2}(1+\rho)^{P/2-1}\hat{\eta}_M \bar{n}'^{P-1} + \beta' z\sqrt{\gamma\frac{P}{2}\bar{q}^{P-1}}\right]\hat{\eta}_M\right\},$$
$$\bar{q} = \mathbb{E}_{\boldsymbol{\xi}^1}\mathbb{E}_{(\boldsymbol{\eta}^1|\boldsymbol{\xi}^1)}\mathbb{E}_z\left\{\tanh^2\left[\beta' \frac{P}{2}(1+\rho)^{P/2-1}\hat{\eta}_M \bar{n}'^{P-1} + \beta' z\sqrt{\gamma\frac{P}{2}\bar{q}^{P-1}}\right]\right\}. \tag{A.19}$$

We also recall that the quenched free-energy within the 1RSB approximation, $\mathcal{F}_{1\text{RSB}}^{(P)}(\bar{n}, \bar{q}_1, \bar{q}_2|\theta)$, is as given in (3.10), where the order parameters $\bar{q}_2$, $\bar{q}_1$ and $\bar{n}$ satisfy the set of self-consistency equations provided in (3.11). We note that for $\theta = 1$, $\bar{q}_1 = \bar{q}$, $\bar{n} = \bar{n}'$ and the

1RSB expression for the quenched free-energy reduces to the RS one. Now, we expand, to the leading order in $\theta-1$, the 1RSB quenched free-energy around its RS expression, as shown in (A.4). Since the self-consistency equations also depend on θ, we need to expand them too. We can write $\bar{q}_1$ as in (A.5), with $A(\bar{n}',\bar{q}_2(\bar{n}',\bar{q}),\bar{q})$ given in (A.9), and $\bar{q}_2$ as given in (A.6), where $\tilde{q}_2(\bar{n}',\bar{q})$ is the solution of (A.7) and $B(\bar{n}',\bar{q}_2(\bar{n}',\bar{q}),\bar{q})$ is as given in (A.10) (we recall that now the expression of $g(\bar{n}',\bar{q}_2(\bar{n}',\bar{q}),\bar{q})$ is as given in (3.12)). With these expressions in hand, we can now compute the derivative of $\mathcal{F}^{(P)}_{\text{1RSB}}$ w.r.t. θ when $\theta=1$, as needed in (A.4)

$$\begin{aligned}K\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right) &:= \partial_\theta\left(-\beta'\mathcal{F}^{(P)}_{\text{1RSB}}\left(\bar{n}',\bar{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}|\theta\right)\right)|_{\theta=1}\\ &= -\frac{\beta'^2\gamma}{4}(P-1)\left[\left(\tilde{q}_2\left(\bar{n}',\bar{q}\right)\right)^P-\bar{q}^P\right]-\frac{\beta'^2\gamma}{4}P\left[\left(\tilde{q}_2\left(\bar{n}',\bar{q}\right)\right)^{P-1}-\bar{q}^{P-1}\right]\\ &\quad-\mathbb{E}_\xi\,\mathbb{E}_{(\eta|\xi)}\mathbb{E}_Y\ln\cosh\left(\beta'\frac{P}{2}\bar{n}'^{P-1}(1+\rho)^{P/2-1}\hat{\eta}_M+Y\beta'\sqrt{\gamma\frac{P}{2}\bar{q}^{P-1}}\right)\\ &\quad+\mathbb{E}_\xi\,\mathbb{E}_{(\eta|\xi)}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)\log\cosh g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}\right].\end{aligned} \tag{A.20}$$

Again, we have that $K(\bar{n}',\bar{q},\bar{q})=0$ (as $\bar{q}$ is the extremum of the RS free-energy). Next, we inspect the sign of $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})$. To this purpose, we study $K(\bar{n}',x,\bar{q})$ for $x\in[0,\bar{q}]$ and locate its extrema, which are found from

$$\partial_x K\left(\bar{n}',x,\bar{q}\right)=-\frac{\beta'^2\gamma(P-1)Px^{P-2}}{4(1+\rho)^P}\left[x-\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\sinh g\left(\bar{n}',\bar{q},x\right)\tanh g\left(\bar{n}',\bar{q},x\right)}{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',\bar{q},x\right)}\right]\right]=0 \tag{A.21}$$

as

$$x=\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',x,\bar{q}\right)\tanh^2 g\left(\bar{n}',x,\bar{q}\right)}{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',x,\bar{q}\right)}\right]\equiv\tilde{q}_2\left(\bar{n}',\bar{q}\right) \tag{A.22}$$

where the last equality follows from (A.7). Under the assumption that the extremum $x=\tilde{q}_2(\bar{n}',\bar{q})$ is global in the domain considered, we have that $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})>0$ if $x=\tilde{q}_2(\bar{n}',\bar{q})$ is a maximum and $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})<0$ if it is a minimum. In particular, if

$$\begin{aligned}\partial_x^2 K\left(\bar{n}',x,\bar{q}\right)|_{x=\tilde{q}_2(\bar{n}',\bar{q})} &= -\frac{\beta'^2\gamma(P-1)P\left(\tilde{q}_2\left(\bar{n}',\bar{q}\right)\right)^{P-2}}{4}\\ &\quad\cdot\left[1-\frac{\beta'^2\gamma}{2}(P-1)P\left(\tilde{q}_2\left(\bar{n}',\bar{q}\right)\right)^{P-2}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\operatorname{sech}^3 g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}\right]\right]\end{aligned} \tag{A.23}$$

is positive, $K(\bar{n}',\tilde{q}_2(\bar{n}',\bar{q}),\bar{q})<0$ and $\mathcal{F}^{(P)}_{\text{1RSB}}(\bar{n}',\bar{q}_2(\bar{n}',\bar{q}),\bar{q}|\theta)<\mathcal{F}^{(P)}_{\text{RS}}(\bar{n}',\bar{q})$. This happens when the expression in the curly brackets of the equation above is negative, i.e. when the parameter $\gamma\beta'^2$ satisfies the inequality

$$\frac{\beta'^2\gamma}{2}(P-1)P\left(\tilde{q}_2\left(\bar{n}',\bar{q}\right)\right)^{P-2}\mathbb{E}_{Y^{(1)}}\left[\frac{\mathbb{E}_{Y^{(2)}}\operatorname{sech}^3 g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}{\mathbb{E}_{Y^{(2)}}\cosh g\left(\bar{n}',\tilde{q}_2\left(\bar{n}',\bar{q}\right),\bar{q}\right)}\right]>1. \tag{A.24}$$

We note that (A.24) is functionally identical to the expression found in [39] for the RS instability line of the standard DAM model, the only difference being encoded in the term

$g(\bar{n}', \tilde{q}_2(\bar{n}', \bar{q}), \bar{q})$, which is here defined differently, as it reflects the supervised protocol. In particular, in the limit of $r \to 0$ or $M \to \infty$, where ρ and ρ_P vanish, (A.24) retrieves the RS instability line of standard DAM models, as obtained in [39].

Appendix B. Proofs

B.1. Proof of lemma 1

Defining the shorthand $\mathcal{Z}_2(J,t) = \mathcal{Z}_2^{(P)}(\boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{Y}; J, t)$, we start by computing the derivative of equation (2.29) w.r.t. t

$$\begin{aligned} -\beta' \partial_t \mathcal{F}_N^{(P)}(J,t) &= \frac{1}{N} \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \partial_t \left[\ln \left[\mathbb{E}_{\boldsymbol{Y}^{(2)}} \mathcal{Z}_2^{\theta}(J,t) \right]^{1/\theta} \right] \\ &= \frac{1}{N} \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \left[\frac{\mathbb{E}_{\boldsymbol{Y}^{(2)}} \partial_t \mathcal{Z}_2(J,t)^{\theta-1} \partial_t \mathcal{Z}_2(J,t)}{\mathbb{E}_{\boldsymbol{Y}^{(2)}} \mathcal{Z}_2(J,t)^{\theta}} \right] \\ &= \frac{1}{N} \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \mathbb{E}_{\boldsymbol{Y}^{(2)}} \left[\mathcal{W}_{2,t} \frac{\partial_t \mathcal{Z}_2(J,t)}{\mathcal{Z}_2(J,t)} \right] \end{aligned} \tag{B.1}$$

where $\mathbb{E}_0 = \mathbb{E}_{\boldsymbol{\xi}^1} \mathbb{E}_{(\boldsymbol{\eta}^1 | \boldsymbol{\xi}^1)} \mathbb{E}_{\boldsymbol{\lambda}}$ and $\mathcal{W}_{2,t}$ is defined in (2.34). Recalling the definition of $\mathcal{Z}_2(J,t)$ in (2.23) we have:

$$\begin{aligned} \partial_t \mathcal{Z}_2(J,t) = \sum_{\boldsymbol{\sigma}} \mathcal{B}_{N,2}^{(P)}(\boldsymbol{\sigma}; J, t) \Bigg[& \left(\frac{\beta' N}{2M} \left(\frac{\mathcal{R}}{r^2} \right)^{P/2} \sum_{a=1}^{M} n_{1,a}^{P}(\boldsymbol{\sigma}) \right) - \psi N \sum_{a=1}^{M} n_{1,a}(\boldsymbol{\sigma}) \\ & + \frac{1}{2} \frac{\beta' P! \sqrt{(1+\rho_P)} \sqrt{K}}{2\sqrt{t}(1+\rho)^{P/2} N^{P-1}} \sum_{i_1 < \cdots < i_P}^{N, \cdots, N} \lambda_{i_1, \ldots, i_P} \sigma_{i_1} \cdots \sigma_{i_P} - \frac{1}{2\sqrt{1-t}} \sum_{b=1}^{2} \left(A_b \sum_{i=1}^{N} Y_i^{(b)} \sigma_i \right) \Bigg]. \end{aligned} \tag{B.2}$$

where, in order to lighten the notation we have set $\mathcal{B}_2^{(P)}(\boldsymbol{\sigma} | \boldsymbol{\eta}^1, \boldsymbol{\lambda}, \boldsymbol{Y}; J, t) = \mathcal{B}_2(\boldsymbol{\sigma}; J, t)$. Inserting this expression in (B.1) and using the definition (2.25) for the average of a single replica of the system over the generalised Boltzmann factor $\mathcal{B}_2(\boldsymbol{\sigma}; J, t)$, we obtain

$$\begin{aligned} & -\beta' \partial_t \mathcal{F}_N^{(P)}(J,t) \\ & = \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \mathbb{E}_{\boldsymbol{Y}^{(2)}} \left\{ \frac{\beta'}{2M} \left(\frac{\mathcal{R}}{r^2} \right)^{P/2} \mathcal{W}_{2,t} \omega_t \left(\sum_{a=1}^{M} n_{1,a}^{P}(\boldsymbol{\sigma}) \right) - \psi \mathcal{W}_{2,t} \omega_t \left(\sum_{a=1}^{M} n_{1,a}(\boldsymbol{\sigma}) \right) \right\} \\ & \quad + \frac{1}{2} \frac{\beta' P! \sqrt{(1+\rho_P)} \sqrt{K}}{2\sqrt{t}(1+\rho)^{P/2} N^P} \sum_{i_1 < \cdots < i_P}^{N, \cdots, N} \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \mathbb{E}_{\boldsymbol{Y}^{(2)}} \left\{ \mathcal{W}_{2,t} \frac{1}{\mathcal{Z}_2(J,t)} \sum_{\boldsymbol{\sigma}} \left(\lambda_{i_1, \ldots, i_P} \sigma_{i_1} \cdots \sigma_{i_P} \right) \mathcal{B}_{N,2}^{(P)}(\boldsymbol{\sigma}; J, t) \right\} \\ & \quad - \frac{1}{2N\sqrt{1-t}} \sum_{b=1}^{2} \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \mathbb{E}_{\boldsymbol{Y}^{(2)}} \left\{ \mathcal{W}_{2,t} \frac{1}{\mathcal{Z}_2(J,t)} \sum_{\boldsymbol{\sigma}} \left(A_b \sum_{i=1}^{N} Y_i^{(b)} \sigma_i \right) \mathcal{B}_{N,2}^{(P)}(\boldsymbol{\sigma}; J, t) \right\}. \end{aligned} \tag{B.3}$$

Next, denoting the combined average over the quenched disorder and the Boltzmann distribution as

$$\langle \cdot \rangle = \mathbb{E}_0 \mathbb{E}_{\boldsymbol{Y}^{(1)}} \mathbb{E}_{\boldsymbol{Y}^{(2)}} \left[\mathcal{W}_{2,t} \omega_t(\cdot) \right] \tag{B.4}$$

and applying the Stein's lemma (2.74) to the standard Gaussian variables $Y_i^{(b)}$ and $\lambda^{\mu}_{i_1,\dots,i_P}$, we get

$$
\begin{aligned}
-\beta'\partial_t\mathcal{F}_N^{(P)} = & \frac{\beta'}{2M}\left(\frac{\mathcal{R}}{r^2}\right)^{P/2}\sum_{a=1}^{M}\langle n_{1,a}^P(\boldsymbol{\sigma})\rangle_t - \psi\sum_{a=1}^{M}\langle n_{1,a}(\boldsymbol{\sigma})\rangle_t \\
& + \frac{\beta'^2 K\sqrt{(1+\rho_P)}^2}{8(1+\rho)^P N^{2P-1}}\sum_{i_1,\cdots,i_P}^{N,\cdots,N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\mathbb{E}_{\boldsymbol{Y}^{(2)}}\left(\mathcal{W}_{2,t}\omega_t(1)+(\theta-1)\mathcal{W}_{2,t}\omega_t^2\left(\sigma_{i_1}\cdots\sigma_{i_P}\right)\right) \\
& - \frac{\beta'^2 P!K\sqrt{(1+\rho_P)}^2}{8(1+\rho)^P N^{2P-1}}\sum_{i_1<\cdots<i_P}^{N,\cdots,N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\theta\left[\mathbb{E}_{\boldsymbol{Y}^{(2)}}\left(\mathcal{W}_{2,t}\omega_t\left(\sigma_{i_1}\cdots\sigma_{i_P}\right)\right)\right]^2 \\
& - \frac{(A_1)^2}{2}\sum_{i=1}^{N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\mathbb{E}_{\boldsymbol{Y}^{(2)}}\left(1+(\theta-1)\mathcal{W}_{2,t}\omega_t^2(\sigma_i)\right) - \frac{(A_1)^2}{2}\theta\sum_{i=1}^{N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\left[\mathcal{W}_{2,t}\omega_t(\sigma_i)\right]^2 \\
& - \frac{(A_2)^2}{2}\sum_{i=1}^{N}\mathbb{E}_0\mathbb{E}_{\boldsymbol{Y}^{(1)}}\mathbb{E}_{\boldsymbol{Y}^{(2)}}\left(1+(\theta-1)\mathcal{W}_{2,t}\omega_t^2(\sigma_i)\right). \qquad \text{(B.5)}
\end{aligned}
$$

Finally, we use the definitions (2.32) and (2.33) and we arrive at the compact expression:

$$
\begin{aligned}
-\beta'\partial_t\mathcal{F}_N^{(P)} = & \frac{\beta'}{2M}\left(\frac{\mathcal{R}}{r^2}\right)^{P/2}\sum_{a=1}^{M}\langle n_{1,a}^P(\boldsymbol{\sigma})\rangle_t - \psi\sum_{a=1}^{M}\langle n_{1,a}(\boldsymbol{\sigma})\rangle_t \\
& + \frac{\beta'^2\sqrt{(1+\rho_P)}^2}{4(1+\rho)^P}\frac{P!K}{2N^{P-1}}\left[1+(\theta-1)\langle q_{12}^P(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,2} - \theta\langle q_{12}^P(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,1}\right] \\
& - \frac{A_1^2}{2}\left[1+(\theta-1)\langle q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,2} - \theta\langle q_{12}\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}\right)\rangle_{t,1}\right] \\
& - \frac{A_2^2}{2}\left[1+(\theta-1)\langle q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,2}\right]. \qquad \text{(B.6)}
\end{aligned}
$$

Next, we take the thermodynamic limit $N\to\infty$, using assumption 2. By manipulating the expression for the moments $\langle q_{12}^P(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,a}$, with $a=1,2$, using Newton's binomial theorem

$$
\begin{aligned}
&\langle q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})^P\rangle_{t,a} \\
&= \left\langle\left[(q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})-\bar{q}_a+\bar{q}_a\right]^P\right\rangle_{t,a} = \sum_{k=0}^{P}\binom{P}{k}\langle(q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})-\bar{q}_a)^k\rangle_{t,a}\bar{q}_a^{P-k} \\
&= P\bar{q}_a^{P-1}\langle q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})\rangle_{t,a} + \sum_{k=2}^{P}\binom{P}{k}\langle(q_{12}(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)})-\bar{q}_a)^k\rangle_{t,a}\bar{q}_a^{P-k} + (1-P)\bar{q}_a^P,
\end{aligned}
$$

we obtain, in the thermodynamic limit, under assumption 2,

$$
\langle q_{12}\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}\right)^P\rangle_{t,a} = P\bar{q}_a^{P-1}\langle q_{12}\left(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}\right)\rangle_{t,a} - (P-1)\bar{q}_a^P \qquad \text{(B.7)}
$$

for $a = 1,2$. Finally, we insert the above in (B.6) and we choose the constants A_1, A_2, ψ in such a way that the terms dependent on $\langle q_{12}(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)})\rangle_{t,a}$ cancel out,

$$\begin{aligned}
\psi &= \beta' \frac{P}{2M} (1+\rho)^{P/2} \bar{n}^{P-1}, \\
A_1^2 &= \beta' \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \frac{KP!}{2N^{P-1}} \bar{q}_1^{P-1}, \\
A_2^2 &= \beta' \frac{P}{2} \frac{\sqrt{(1+\rho_P)}^2}{(1+\rho)^P} \frac{KP!}{2N^{P-1}} \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right).
\end{aligned} \tag{B.8}$$

This makes $\mathrm{d}\mathcal{F}^{(P)}/\mathrm{d}t$ t-independent and leads to the thesis (2.40).

B.2. Proof of corollary 2

Let us start from the self consistency equations in the $M \to \infty$ limit introduced in corollary 1, we recognize that as $\tilde{\beta} \to \infty$, we have $\bar{q}_2 \to 1$, therefore in order to perform the limit we will introduce the reparametrization

$$\bar{q}_2 = 1 - \frac{\delta \bar{q}_2}{\tilde{\beta}} \quad \text{as} \quad \tilde{\beta} \to \infty. \tag{B.9}$$

It is now useful to insert an additional term $\tilde{\beta} y$ in the expression of $g(\beta, \gamma, \boldsymbol{Y})$ in (2.76), which now reads as

$$\begin{aligned}
g(\beta,\gamma,\boldsymbol{Y}) = \tilde{\beta}\frac{P}{2}\bar{m}^{P-1} &+ \tilde{\beta} Y^{(2)} \sqrt{\gamma \frac{P}{2} \sqrt{(1+\rho_P)}^2 \left(\bar{q}_2^{P-1} - \bar{q}_1^{P-1}\right)} \\
&+ \tilde{\beta} Y^{(1)} \sqrt{\rho \left(\frac{P}{2}\bar{m}^{P-1}\right)^2 + \gamma \frac{P}{2} \sqrt{(1+\rho_P)}^2 \bar{q}_1^{P-1}} + \tilde{\beta} y.
\end{aligned} \tag{B.10}$$

Using this new parameter y, we can recast the equation for $\bar{q}_2$ as a derivative of the magnetization

$$\frac{\partial \bar{m}}{\partial y} = \delta \bar{q}_2 - D\Delta\bar{q} \Longrightarrow \delta \bar{q}_2 = \frac{\partial \bar{m}}{\partial y} + D\Delta\bar{q} \tag{B.11}$$

where we have used $\Delta\bar{q} = \bar{q}_2 - \bar{q}_1$ and, as $\tilde{\beta} \to \infty$, $\tilde{\beta}\theta \to D \in \mathbb{R}$. Thus, in the zero temperature limit the last three equations in equation (2.70) become

$$\begin{aligned}
\bar{m} &\to \mathbb{E}_{Y^{(1)}} \left\{ \frac{\mathbb{E}_{Y^{(2)}} \left[\mathrm{sign}[g(\boldsymbol{Y}, \bar{m}, \Delta\bar{q})]\, e^{D|g(\boldsymbol{Y},\bar{m},\Delta\bar{q})|}\right]}{\mathbb{E}_{Y^{(2)}} \left[e^{D|g(\boldsymbol{Y},\bar{m},\Delta\bar{q})|}\right]} \right\} \\
\Delta\bar{q} &\to 1 - \mathbb{E}_{Y^{(1)}} \left\{ \frac{\mathbb{E}_{Y^{(2)}} \left[\mathrm{sign}[g(\boldsymbol{Y}, \bar{m}, \Delta\bar{q})]\, e^{D|g(\boldsymbol{Y},\bar{m},\Delta\bar{q})|}\right]}{\mathbb{E}_{Y^{(2)}} \left[e^{D|g(\boldsymbol{Y},\bar{m},\Delta\bar{q})|}\right]} \right\}^2 \\
\bar{q}_2 &\to 1.
\end{aligned} \tag{B.12}$$

Now, if we suppose $\Delta\bar{q} \ll 1$ the (B.10) reduces to

$$g(\boldsymbol{Y}, \bar{m}, \Delta\bar{q}) = C_0 + C_1 Y^{(1)} + C_2 Y^{(2)} + \mathcal{O}(\Delta\bar{q}) \tag{B.13}$$

where

$$C_0 = \frac{P}{2}\bar{m}^{P-1}, C_1 = \sqrt{\rho\left(\frac{P}{2}\bar{m}^{P-1}\right)^2 + \gamma\sqrt{(1+\rho_P)^2}\frac{P}{2}},$$
$$C_2 = \sqrt{\gamma\sqrt{(1+\rho_P)^2}\frac{P(P-1)}{2}\Delta\bar{q}}. \tag{B.14}$$

Performing the integral over $Y^{(2)}$ we reach equations (2.77) and (2.78).

ORCID iDs

Linda Albanese https://orcid.org/0000-0003-1783-0775
Alessia Annibale https://orcid.org/0000-0003-4010-6742

References

[1] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci. USA* **79** 2554–8
[2] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* vol 9 (World Scientific Publishing Company)
[3] Talagrand M 2003 *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models* vol 46 (Springer Science & Business Media)
[4] Amit D J J 1989 *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press)
[5] Dotsenko V 1994 *An Introduction to the Theory of Spin Glasses and Neural Networks* (World Scientific)
[6] Engel A 2001 *Statistical Mechanics of Learning* (Cabridge Press)
[7] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Clarendon)
[8] Coolen A C C, Kuhn R and Sollich P 2005 *Theory of Neural Information Processing Systems* (Oxford University Press, Inc.)
[9] Huang H 2021 *Statistical Mechanics of Neural Networks* (Springer)
[10] Agliari E, Barra A, Sollich P and Zdeborová L 2020 Machine learning and statistical physics *J. Phys. A: Math. Theor.* **53** 500401
[11] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 Machine learning and the physical sciences *Rev. Mod. Phys.* **91** 045002
[12] Amit D J, Gutfreund H and Sompolinsky H 1985 Storing infinite numbers of patterns in a spin-glass model of neural networks *Phys. Rev. Lett.* **55** 1530
[13] Kanter I 1988 Asymmetric neural networks with multispin interactions *Phys. Rev.* A **38** 5972
[14] Gardner E 1987 Multiconnected neural network models *J. Phys. A: Math. Gen.* **20** 3453–64
[15] Baldi P and Venkatesh S S 1987 Number of stable points for spin-glasses and neural networks of higher orders *Phys. Rev. Lett.* **58** 913
[16] Krotov D and Hopfield J J 2016 Dense associative memory for pattern recognition *Advances in Neural Information Processing Systems* pp 1180–8
[17] Ramsauer H *et al* 2020 Hopfield networks is all you need (arXiv:2008.02217)
[18] Agliari E, Alemanno F, Barra A, Centonze M and Fachechi A 2020 Neural networks with a redundant representation: detecting the undetectable *Phys. Rev. Lett.* **124** 28301
[19] Krotov D and Hopfield J J 2018 Dense associative memory is robust to adversarial inputs *Neural Comput.* **30** 3151–67
[20] Lucibello C and Mezard M 2023 The exponential capacity of dense associative memories (arXiv:2304.14964)
[21] Barra A, Bernacchia A, Santucci E and Contucci P 2012 On the equivalence of Hopfield networks and Boltzmann machines *Neural Netw.* **34** 1–9

[22] Hinton G E, Sejnowski T J and Ackley D H 1984 *Boltzmann Machines: Constraint Satisfaction Networks That Learn* (Carnegie-Mellon University, Department of Computer Science)
[23] Agliari E, Barra A, Galluzzi A, Guerra F and Moauro F 2012 Multitasking associative networks *Phys. Rev. Lett.* **109** 268101
[24] Sollich P, Tantari D, Annibale A and Barra A 2014 Extensive parallel processing on scale-free networks *Phys. Rev. Lett.* **113** 238106
[25] Tubiana J and Monasson R 2017 Emergence of compositional representations in restricted Boltzmann machines *Phys. Rev. Lett.* **118** 138301
[26] Barra A, Genovese G, Sollich P and Tantari D 2017 Phase transitions in restricted Boltzmann machines with generic priors *Phys. Rev.* E **2017** 042156
[27] Barra A, Genovese G, Sollich P and Tantari D 2018 Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors *Phys. Rev.* E **2018** 022310
[28] Agliari E, Albanese L, Alemanno F and Fachechi A 2021 A transport equation approach for deep neural networks with quenched random weights *J. Phys. A: Math. Theor.* **54** 505004
[29] Agliari E and Sebastiani G 2021 Learning and retrieval operational modes for three-layer restricted Boltzmann machines *J. Stat. Phys.* **185** 10
[30] Alberici D, Barra A, Contucci P and Mingione E 2020 Annealing and replica-symmetry in deep Boltzmann machines *J. Stat. Phys.* **180** 665–77
[31] Alberici D, Contucci P and Mingione E 2021 Deep Boltzmann machines: rigorous results at arbitrary depth *Ann. Henri Poincaré* **22** 2619–42
[32] Mehta P and Schwab D 2014 An exact mapping between the variational renormalization group and deep learning (arXiv:1410.3831)
[33] Agliari E, Alemanno F, Barra A and De Marzo G 2022 The emergence of a concept in shallow neural networks *Neural Netw.* **148** 232–53
[34] Alemanno F, Aquaro M, Kanter I, Barra A and Agliari E 2023 Supervised hebbian learning *Europhys. Lett.* **141** 11001
[35] Barlow H B 1989 Unsupervised learning *Neural Comput.* **1** 295–311
[36] Cunningham P, Cord M and Delany S J 2008 Supervised learning *Machine Learning Techniques for Multimedia* (Springer) pp 21–49
[37] Agliari E, Albanese L, Alemanno F, Alessandrelli A, Barra A, Giannotti F, Lotito D and Pedreschi D 2023 Dense Hebbian neural networks: a replica symmetric picture of unsupervised learning *Physica* A **627** 129143
[38] Agliari E, Albanese L, Alemanno F, Alessandrelli A, Barra A, Giannotti F, Lotito D and Pedreschi D 2023 Dense Hebbian neural networks: a replica symmetric picture of supervised learning *Physica* A **626** 129076
[39] Albanese A, Alessandrelli L, Annibale A and Barra A 2024 About the de Almeida-Thouless line in neural networks *Physica* A **633** 129372
[40] Bovier A and Picco P 2012 *Mathematical Aspects of Spin Glasses and Neural Networks* vol 41 (Springer Science & Business Media)
[41] Gradenigo G, Angelini M C, Leuzzi L and Ricci-Tersenghi F 2020 Solving the spherical p-spin model with the cavity method: equivalence with the replica results *J. Stat. Mech.* 113302
[42] Barra A, Di Biasio A and Guerra F 2010 Replica symmetry breaking in mean-field spin glasses through the Hamilton–Jacobi technique *J. Stat. Mech.* P09006
[43] Guerra F 2003 Broken replica symmetry bounds in the mean field spin glass model *Commun. Math. Phys.* **233** 1–12
[44] Agliari E, Albanese L, Barra A and Ottaviani G 2020 Replica symmetry breaking in neural networks: a few steps toward rigorous results *J. Phys. A: Math. Theor.* **53** 415005
[45] Albanese L, Alemanno F, Alessandrelli A and Barra A 2022 Replica symmetry breaking in dense Hebbian neural networks *J. Stat. Phys.* **189** 1–41
[46] Rammal R, Toulouse G and Virasoro M A 1986 Ultrametricity for physicists *Rev. Mod. Phys.* **58** 765
[47] Cugliandolo L F 2011 The effective temperature *J. Phys. A: Math. Theor.* **44** 483001
[48] Van Mourik J and Coolen A C C 2001 Cluster derivation of Parisi's RSB solution for disordered systems *J. Phys. A: Math. Gen.* **34** L111
[49] Barra A, Guerra F and Mingione E 2012 Interpolating the Sherrington–Kirkpatrick replica trick *Phil. Mag.* **92** 78–97
[50] Sherrington D and Kirkpatrick S 1975 Solvable model of a spin-glass *Phys. Rev. Lett.* **35** 1792–6
[51] Parisi G 1979 Infinite number of order parameters for spin-glasses *Phys. Rev. Lett.* **43** 1754–8

[52] Amit D, Gutfreund H and Sompolinsky H 1987 Statistical mechanics of neural networks near saturation *Ann. Phys., NY* **173** 30–67
[53] Crisanti A, Amit D J and Gutfreund H 1986 Saturation level of the Hopfield model for neural network *Europhys. Lett.* **2** 337–41
[54] Derrida B 1981 Random-energy model: an exactly solvable model of disordered systems *Phys. Rev.* B **24** 2613
[55] Steffan H and Kühn R 1994 Replica symmetry breaking in attractor neural network models *Z. Phys.* B **95** 249–60
[56] Biroli G and Mezard M 2023 Generative diffusion in very large dimensions (arXiv:2306.03518)
[57] Agliari E, Alessandrelli A, Barra A and Ricci-Tersenghi F 2023 Parallel learning by multitasking neural networks *J. Stat. Mech.* 113401
[58] de Almeida J R L and Thouless D J 1978 Stability of the Sherrington-Kirkpatrick solution of a spin glass model *J. Phys. A: Math. Gen.* **11** 983