Journal of Statistical Physics (2021) 185:10
https://doi.org/10.1007/s10955-021-02841-y

Learning and Retrieval Operational Modes for Three-Layer Restricted Boltzmann Machines

Elena Agliari[1] · **Giulia Sebastiani**[2]

Received: 26 February 2021 / Accepted: 12 October 2021 / Published online: 23 October 2021

Abstract
We consider a three-layer restricted Boltzmann machine, where the two visible layers (encoding for input and output, respectively) are made of binary neurons while the hidden layer is made of Gaussian neurons, and we show a formal equivalence with a Hopfield model. The machine architecture allows for different learning and operational modes: when all neurons are free to evolve we recover a standard Hopfield model whose size corresponds to the overall size of visible neurons; when input neurons are clamped we recover a Hopfield model, whose size corresponds to the size of the output layer, endowed with an external field as well as additional slow noise. The former stems from the signal provided by the input layer and tends to favour retrieval, the latter can be related to the statistical properties of the training set and tends to impair the retrieval performance of the network. We address this model by rigorous techniques, finding an explicit expression for its free-energy, whence a phase-diagram showing the performance of the system as parameters are tuned.

Keywords Disordered systems · Boltzmann Machine · Hopfield model

1 Introduction

In the last decade the availability of new technologies (e.g., GPUs) and huge volumes of data (the so-called "big data") have allowed the development of analytical methods and algorithms to effectively transform *information* into *knowledge*, which, in turn, can support decision-making processes. Among these, machine learning and deep networks are often used in a pseudo-empirical way, as a full rationale is not available yet. Indeed, technological and computational advances have been faster than theoretical ones and filling this gap is now an attractive and stimulating challenge for mathematicians and physicists. Of course, theoretical results may have practical outcomes: a deep understand-

Communicated by Irene Giardina.

✉ Elena Agliari
agliari@mat.uniroma1.it

[1] Dipartimento di Matematica "Guido Castelnuovo", Sapienza Università di Roma, Rome, Italy

[2] Institute of Mathematics, Goethe-Universität, Frankfurt, Germany

ing of the mechanisms underlying the operation of these machines would allow optimal use, for example by choosing a minimal number of free parameters and by identifying appropriate initializations, with consequent savings in terms of calculation time and energy.

The most widespread neural networks used for machine learning are made up of multilayer structures where each layer is made up of a certain number of neurons, which interact with each other through suitable activation functions. Generally, the first layer receives the input, the last one provides the output and the intermediate layers, called hidden, provide the degrees of freedom to be estimated during training through operations of extremization of appropriate cost functions. The purpose of the network is to learn to represent the reality that is presented to it through a series of examples; this is accomplished by suitably tuning a set of parameters that is, the interaction strengths among neurons and the external fields acting on neurons. In applications, the architecture of the network (how many hidden layers and what sizes) is often chosen through empirical methods, as there is no standard theory or accepted method [1,2]. In this paper we focus on multilayer structures with only one hidden layer (namely, the networks are *shallow*), where connections between neurons are symmetric (namely, the networks are *recurrent*), and where connections only occur between neurons belonging to different layers (namely, the networks are *restricted*). For this kind of neural networks, also referred to as restricted Boltzmann machines (RBMs), a few rigorous results are available (see e.g., [3–7]). Most of these results were obtained by leveraging a formal equivalence between two-layer RBMs and Hopfield networks (HN) [8]. More precisely, when looking at these systems as spin-like models, one can show that the Boltzmann-Gibbs equilibrium measure of the two systems are equivalent.

Here, we extend such an equivalence in order to include more general RBMs. In fact, we consider three-layer RBMs with one hidden layer and two visible layers, the input one is made of N neurons and the output one is made of K neurons. For such an architecture the accomplishment of the training stage corresponds to a parameter setting such that the expected state $\langle\cdot\rangle_-$, under the equilibrium measure where neurons are all freely evolving, is the same as the expected state $\langle\cdot\rangle_+$, obtained when (a subset of) the visible neurons are fixed in suitable configurations (namely, they are *clamped*). In particular, in the unsupervised learning, the neurons in the input layer are iteratively fixed according to the examples making up the training set and the trained network is expected to serve as a generative model or for pattern reconstruction (see e.g., [1,9]). Now, as we will show, $\langle\cdot\rangle_-$ can be recast as the expectation for an Hopfield model made of $N + K$ neurons, while $\langle\cdot\rangle_+$ can be recast as the expectation for an Hopfield model made of N neurons and subjected to an external field which depends on the input. Further, we can prove that the performance of the network, in a retrieval mode, strongly depends on the quality of the input: if this is too noisy with respect to the original pattern then the system will be no longer able to accomplish the task.

The plan of this article is as follows: in Sect. 2 we review the paradigmatic models for associative memory and machine learning, namely, the Hopfield model and the two-layer RBM, respectively, recalling a formal equivalence between them; in Sect. 3 we consider three-layer RBMs along with their training algorithm and we explore its retrieval counterpart, that is, a HN with a field which is also investigated; in Sect. 3.2 we discuss in details some possible scenario; finally, Sect. 4 is left for conclusions and discussions. Technical details are collected in Appendices A–E.

2 A Formal Equivalence Between the Hopfield Network and the Restricted Boltzmann Machine

From an equilibrium statistical-mechanics perspective, the formal "duality" between HNs and RBMs is observed through an equivalence of the relative partition functions. This result was originally obtained considering the minimal version of a RBM, where one single visible layer couples with the hidden one building a two-layer restricted machine [8,10,11]. Before proceeding in this direction, let us introduce some notation and give a brief review of the Hopfield model (see also e.g., [12] for an extensive treatment).

For each integer $N \in \mathbb{N}$ we indicate with $\mathscr{I}_N$ the usual configuration space for a spin system of size N, namely the product space $\{-1,+1\}^N$. Given a family of numbers $\{\xi_i^\mu\}_{i=1,\dots,N}^{\mu=1,\dots,P}$, with $\xi_i^\mu = \pm 1$ and $P = P(N) \in \mathbb{N}$, the Hamiltonian (or cost function) of the Hopfield model is defined for each $\sigma \in \mathscr{I}_N$ as

$$\mathcal{H}_N^{\mathrm{H}}(\sigma) = -\frac{1}{2N} \sum_{\mu=1}^{P} \sum_{i,j=1}^{N} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \tag{1}$$

$$= -\frac{N}{2} \sum_{\mu=1}^{P} \left(\frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i \right)^2 . \tag{2}$$

The vector $\sigma = (\sigma_i)_{i=1}^N$ describes the state of the system as a whole, while its i-th component, the binary (spin) variable $\sigma_i = \pm 1$, represents the state of the i-th neuron. The pair-wise interaction constants J_{ij}, consistently with the definition (1), result in the so-called *Hebb's rule*: $J_{ij} = N^{-1} \sum_\mu \xi_i^\mu \xi_j^\mu$. In the following, as standard in the Hopfield model, we take the numbers ξ_i^μ as independent Bernoulli random variables, such that $\mathbb{P}(\xi_i^\mu = \pm 1) = 1/2$. The Hamiltonian of the system defines a mean-field disordered model which is able to store *patterns* of information, represented by the P vectors $\{\xi^\mu\}_{\mu=1}^P$, $\xi^\mu = (\xi_i^\mu)_{i=1}^N$. This memory is allocated in the coupling matrix J through Hebb's rule which favors the local attractiveness of the patterns for the neuronal dynamics.

We can consider, for example, a standard Glauber's dynamics for Ising-type systems, that is a discrete-time Markov chain whose equilibrium distribution coincides with the Boltzmann-Gibbs distribution associated to the Hamiltonian (1), namely the random probability measure on the hypercube given for each $\sigma \in \mathscr{I}_N$ as

$$\mathcal{G}_N(\sigma) = \frac{e^{-\beta \mathcal{H}_N^{\mathrm{H}}(\sigma)}}{\mathcal{Z}_N^{\mathrm{H}}}, \tag{3}$$

where $\beta \in \mathbb{R}^+$ gives the inverse temperature and the normalization factor $\mathcal{Z}_N^{\mathrm{H}} = \sum_{\sigma \in \mathscr{I}_N} e^{-\beta \mathcal{H}_N^{\mathrm{H}}(\sigma)}$ is also referred to as partition function.

We define the *Mattis magnetizations* as the overlaps between a generic configuration and each stored pattern ξ^μ, that is

$$m_\mu = m_{\mu,N}(\sigma) = N^{-1} \sum_{i=1}^{N} \xi_i^\mu \sigma_i \in [-1,1], \qquad \forall \mu \in \{1,\dots,P\}. \tag{4}$$

These magnetizations play the role of order parameters for the model as they measure the resemblance of a given network configuration σ with each of the P patterns. For low temperatures, each term $Nm_\mu^2/2$ in (1) tends to align the system configuration σ with the sequence

$\{\xi_i^\mu\}_{i=1}^N$ or the sequence $\{-\xi_i^\mu\}_{i=1}^N$. If this happens, that is for some μ at equilibrium $m_\mu \sim \pm 1$, we say that the network has retrieved the pattern ξ^μ. The possibility and the quality of this retrieving behavior strongly depend both on the *noise* level, regulated by β, and on the *load* $\lambda \stackrel{\text{def}}{=} \lim_{N\to\infty} N^{-1}P(N)$, quantifying the number of stored patterns with respect to the number of neurons in the thermodynamic limit [26]. For example, in the noiseless case $\beta \to \infty$ with $\lambda > 0$ it can be shown that – in the thermodynamic limit, replica symmetric regime and uncorrelated patterns – if $\lambda < \lambda_c \sim 0.051$ the system acts with almost no errors as an associative neural network, meaning that the attractors associated to stored patterns are very stable being global minima of the quenched free energy. If, instead, $0.051 \sim \lambda_c < \lambda < \lambda_c' \sim 0.138$, the network could still work as an associative memory but spin-glass states start to dominate the landscape (the stored patterns are just local minima). When $\lambda > \lambda_c'$ the patterns are too many and a glass transition happens: the minima related to them are destroyed and solely the spin-glass panorama remains stable.

So far we presented the Hopfield model as the simplest paradigm for machine retrieval, but for an object to work as a cognitive system a learning phase must preceed the retrieval one. As anticipated, RBMs are relatively simple learning machines: in its minimal, two-layer realization, a Boltzmann machine can be introduced as a network composed of $N + P$ neurons that has been partitioned into an input-output visible layer (of N neurons) and a hidden layer (of P neurons), with $P = P(N)$. The above-mentioned equivalence with the Hopfield model can be easily highlighted considering a *hybrid* restricted machine (HBM), in which the visible layer is digital while the hidden one is analog [8]. We describe the states of the neurons in the two layers with the vectors $\sigma \in \mathscr{I}_N = \{-1, 1\}^N$ and $z \in \mathbb{R}^P$, respectively. The connectivity of the network is symmetric and couplings only regard neurons in different layers, while self-interactions are excluded. Specifically, we consider here a trained machine that has learnt some weights $\{\xi_i^\mu\}_{i=1,\dots,N}^{\mu=1,\dots,P}$, where ξ_i^μ represents the coupling between the visible unit σ_i and the hidden unit z_μ. The energy associated to each state (σ, z) is given through the Hamiltonian (or *cost*) function $\mathcal{H}_N^{\mathrm{B}} : \{-1, 1\}^N \times \mathbb{R}^P \to \mathbb{R}$ with

$$\mathcal{H}_N^{\mathrm{B}}(\sigma, z) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} \xi_i^\mu \sigma_i z_\mu. \tag{5}$$

The dynamics on the network can be given in such a way that the hidden layer z is Gaussian distributed at equilibrium, with variance regulated by the temperature.[1] The machine can be thus referred, from the point of view of statistical mechanics, as a (multipartite) mean-field spin-glass model whose partition function results in $\mathcal{Z}_N^{\mathrm{B}} = \sum_{\sigma\in\mathscr{I}_N} \int_{\mathbb{R}^P} d\boldsymbol{G}_\beta(z) e^{-\beta\mathcal{H}_N^{\mathrm{B}}(\sigma,z)}$, where $d\boldsymbol{G}_\beta(z)$ indicates the Gaussian measure on $\mathbb{R}^P$ with zero mean and variance β^{-1}, $d\boldsymbol{G}_\beta(z) = \left(\frac{\beta}{2\pi}\right)^{\frac{P}{2}} e^{-\frac{\beta}{2} z\cdot z} dz$. The integral in the definition of $\mathcal{Z}_N^{\mathrm{B}}$ can be trivially evaluated through the Gaussian integral formula giving $\mathcal{Z}_N^{\mathrm{B}} = \sum_{\sigma\in\mathscr{I}_N} e^{-\frac{\beta}{2N}\sum_\mu \left(\sum_i \xi_i^\mu \sigma_i\right)^2}$.

A glance at the last expression is enough to realize that it equals the partition function of a Hopfield model with N neurons and patterns $\{\xi^\mu\}_{\mu=1}^P$, see (1). We can therefore state that $\mathcal{Z}_N^{\mathrm{B}} = \mathcal{Z}_N^{\mathrm{H}}$. This result connects the two Hamiltonians of the HN and the HBM and ensures that thermodynamics obtained by the first cost function (1) is the same as the one obtained by

[1] The details are given in Appendix A, where we explain how to define a dynamics that makes the hidden layer Gaussian distributed using the notation that will be introduced in Sect. 3 for the three-layer machine.

the second one (5): observable quantities stemmed from the HN are equivalent in distribution to the corresponding ones in the HBM. Simulating the dynamics of a HN, requiring the update of N neurons and the storage of $N(N-1)/2$ synapses, can be thus accomplished by a HBM, requiring the update of $N+P$ neurons but the storage of only NP synapses. In addition, the glass transition of the HN has a counterpart in the Boltzmann Machine: it corresponds to an optimum criterion for selecting the relative size of the hidden and visible layer [13,14]. We refer to [4,11,15–19] and references therein for a more extensive and general treatment, also including the case where the nature of visible and hidden neurons can span from binary to continuous, where networks are multi-layer, and where pattern entries are correlated.[2] Further, it is worth mentioning that the binary nature of the connection weights in (5) is not a strict requirement; this issue was addressed from several perspectives, both analytically and numerically, in [13,14,17].

3 The Three-Layer Boltzmann Machine and Its Training Algorithm

In this section we enrich the architecture of the Boltzmann machine by inserting an additional visible layer composed of K binary neurons, meant as the output layer (see Fig. 1). The visible layer made of N binary neurons and the hidden layer made of P Gaussian neurons are retained and the former is now meant as input, while the latter is now an intermediate layer. More precisely, we consider a three-layer HBM and, given $N, K, P \in \mathbb{N}$, $K = K(N)$, $P = P(N)$, we indicate with $\sigma = (\sigma_i)_{i=1}^{N} \in \{-1,+1\}^N$ the state of the digital N-dimensional input, $\tau = (\tau_\nu)_{\nu=1}^{K} \in \{-1,+1\}^K$ the state of the digital K-dimensional output and $z = (z_\mu)_{\mu=1}^{P} \in \mathbb{R}^P$ the state of the hidden P-dimensional analog layer. The machine is endowed with dichotomous, fixed and symmetric connections which only concern the two visible layers with respect to the hidden one, specifically:

- $\boldsymbol{\xi} = \{\xi_i^\mu\}_{i=1,\dots,N}^{\mu=1,\dots,P}$ is the $(P \times N)$-dimensional interaction matrix between the hidden and the input layers (i.e., ξ_i^μ indicates the interaction between each input unit σ_i and each hidden unit z_μ);
- $\boldsymbol{\eta} = \{\eta_\nu^\mu\}_{\nu=1,\dots,K}^{\mu=1,\dots,P}$ is the $(P \times K)$-dimensional interaction matrix between the hidden and output layers (i.e., η_ν^μ indicates the interaction between each output unit τ_ν and each hidden unit z_μ).

The Hamiltonian associated to this system is $\mathcal{H}^{\mathrm{B}} : \mathscr{I}_N \times \mathbb{R}^P \times \mathscr{I}_K \to \mathbb{R}$ with

$$\mathcal{H}_N^{\mathrm{B}}(\sigma, z, \tau) \stackrel{\mathrm{def}}{=} -\frac{1}{\sqrt{N}} \left[\sum_{\mu,i} z_\mu \xi_i^\mu \sigma_i + \sum_{\mu,\nu} z_\mu \eta_\nu^\mu \tau_\nu \right] = -\sum_{\mu} z_\mu I_\mu(\sigma, \tau), \tag{6}$$

where we posed

$$I_\mu(\sigma, \tau) \stackrel{\mathrm{def}}{=} \frac{1}{\sqrt{N}} \left(\sum_i \xi_i^\mu \sigma_i + \sum_\nu \eta_\nu^\mu \tau_\nu \right), \qquad \forall \mu \in \{1, \dots, P\}. \tag{7}$$

[2] Despite its popularity, Hebb's rule is well-known to be suboptimal in the case of correlated patterns. Revisions of this rule, for instance the pseudo-inverse rule, can yield to enhanced performances; the related HN can still be mapped into a two-layer Boltzmann machine, yet intra-layer couplings need to be allowed for [14].

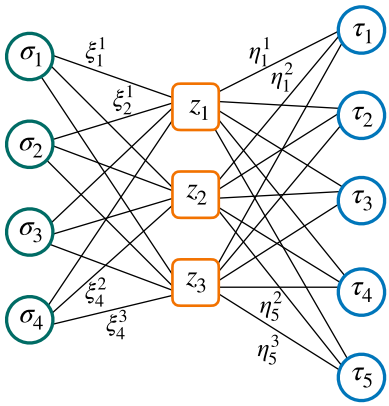

Fig. 1 A schematic picture of the three-layer RBM studied in this work. The left-most layer is the input side made of $N = 4$ binary neurons $\sigma \in \{-1, +1\}^N$, while the right-most layer it the output side made of $K = 5$ binary neurons $\tau \in \{-1, +1\}^K$; the intermediate layer is made of $P = 3$ hidden neurons $z \in \mathbb{R}^P$. The weights associated to links connecting the i-th input (output) neuron and the μ-th hidden neuron is denoted as ξ_i^μ (η_i^μ) and only a few are shown explicitly seeking for clarity

Remarkably, this is just the Hamiltonian for a bipartite[3] spin-glass and, as explained in Appendix A, the definition (6) can also be derived by studying the system from a dynamic perspective and showing that its equilibrium joint probability distribution is

$$\mathcal{G}(\sigma, z, \tau) \propto e^{-\beta\left(\frac{1}{2}\|z\|^2 - \frac{1}{\sqrt{N}} z \cdot \xi \cdot \sigma - \frac{1}{\sqrt{N}} z \cdot \eta \cdot \tau\right)}, \tag{8}$$

where the symbol "$\cdot$" is used for the standard inner product in the multidimensional Euclidean space, while "$\|\cdot\|$" will indicate the correspondent standard norm. The weight setting in (8) can be thought of as the result of a training procedure, where patterns learnt are allocated in the machine weights [13,17,19].

We recall that, for unsupervised learning, the network is trained over a sample of examples $\{(\sigma^{(k)})\}_{k=1}^M$ drawn from a certain, unknown target distribution $q(\sigma)$ which we want $\mathcal{G}(\sigma) = \sum_\tau \int dG_\beta(z)\mathcal{G}(\sigma, z, \tau)$ to mimic. Training can be carried out by iteratively tuning couplings[4] between neurons in such a way as to minimize the Kullback-Leibler divergence

$$\mathcal{D}(q||\mathcal{G}) := \sum_\sigma q(\sigma) \log_2 \left[\frac{q(\sigma)}{\mathcal{G}(\sigma)}\right]. \tag{9}$$

It can be proved that training is accomplished when the difference between the expectations $\langle\cdot\rangle_-$ and $\langle\cdot\rangle_+$ of neuron states and neuron correlations are vanishing (see e.g., [1]). As anticipated, these expectations correspond to, respectively, the following partition functions

$$\textit{Free mode}: \quad \mathcal{Z}_N^{free} \stackrel{\text{def}}{=} \sum_{\sigma\in\mathcal{I}_N} \sum_{\tau\in\mathcal{I}_K} \int_{\mathbb{R}^P} \prod_{\mu=1}^P dG_\beta(z_\mu) e^{-\beta\mathcal{H}_N^{\mathrm{B}}(\sigma,z,\tau)} \tag{10}$$

$$\textit{Clamped mode}: \quad \mathcal{Z}_N^{clamp} \stackrel{\text{def}}{=} \sum_{\tau\in\mathcal{I}_K} \int_{\mathbb{R}^P} \prod_{\mu=1}^P dG_\beta(z_\mu) e^{-\beta\mathcal{H}_N^{\mathrm{B}}(\sigma,z,\tau)} \tag{11}$$

where, again, $dG_\beta(z_\mu)$ indicates the Gaussian measure with zero mean and variance β^{-1}, namely $dG_\beta(z_\mu) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\beta z_\mu^2}{2}} dz_\mu$.

[3] The three-layer architecture allows us to distinguish an input and an output layer, but it can be recast into a two-layer struture where, possibly, only a subset of size N out of the $N + K$ neurons making up the visible layers can be clamped.

[4] In the current scenario the set of tuneable parameters include couplings between neurons, while external fields can be neglected as we are considering centered patterns.

We stress that both (10) and (11) are *random* partition functions as they depend on a specific realization of the vectors in $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$; moreover $\mathcal{Z}_N^{clamp}$ explicitly depends also on the chosen fixed input described by variables σ. Through Fubini's theorem the integral in (10) and (11) can be rewritten as

$$\prod_{\mu=1}^{P} \int_{\mathbb{R}} dG_\beta(z_\mu) e^{\beta I_\mu z_\mu}$$

and trivially evaluated through the Gaussian integral formula. We therefore get

$$\begin{aligned}\int_{\mathbb{R}^P} \prod_{\mu=1}^{P} dG_\beta(z_\mu) e^{\beta I_\mu z_\mu} &= \prod_{\mu=1}^{P} \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \int_{\mathbb{R}} e^{-\frac{\beta}{2} z_\mu^2 + \beta I_\mu z_\mu} \\ &= \prod_{\mu=1}^{P} e^{\frac{\beta}{2} I_\mu^2} = e^{\frac{\beta}{2} \sum_\mu I_\mu^2} = e^{\frac{\beta}{2} \|I\|^2}.\end{aligned}$$

Marginalizing over the analog layer each couple (σ, τ) contributes to the partition functions (10) and (11) with a quantity only depending on the Euclidean norm of the correspondent $I(\sigma, \tau)$, the vector containing the fields acting on the hidden layer

$$\mathcal{Z}_N^{free} = \sum_\sigma \sum_\tau e^{\frac{\beta}{2} \|I(\sigma,\tau)\|^2}, \qquad \mathcal{Z}_N^{clamp} = \sum_\tau e^{\frac{\beta}{2} \|I(\sigma,\tau)\|^2}. \tag{12}$$

Writing explicitly the expression for $\|I(\sigma, \tau)\|$ we obtain

$$\mathcal{Z}_N^{free} = \sum_\sigma \sum_\tau e^{-\frac{\beta}{N} \sum_\mu \left(-\frac{1}{2}(\xi^\mu \cdot \sigma)^2 - \frac{1}{2}(\eta^\mu \cdot \tau)^2 - (\xi^\mu \cdot \sigma)(\eta^\mu \cdot \tau)\right)}, \tag{13}$$

$$\mathcal{Z}_N^{clamp} = e^{\frac{\beta}{2N} \sum_\mu (\xi^\mu \cdot \sigma)^2} \sum_\tau e^{-\frac{\beta}{N} \sum_\mu \left(-\frac{1}{2}(\eta^\mu \cdot \tau)^2 - (\xi^\mu \cdot \sigma)(\eta^\mu \cdot \tau)\right)}. \tag{14}$$

Up to multiplying factors these expressions equal those for the partition functions of two HNs, as we are going to explain with more details in the next subsections. These equivalences allow us to extend techniques developed for Hopfield models to the thermodynamic study of the Boltzmann machine. Our goal now consists in investigating the thermodynamics of the Hopfield models corresponding to (13) and (14) and in inspecting possible effects arising from a poor clamping. We shall especially focus on the latter model as the peculiar external field appearing in the Hamiltonian makes it of interest *per se*. We recall that this investigation implies averaging observable quantities both on the statistical ensemble (i.e., over configurations) and noise (i.e., over the synaptic weights). In order to assess these averaged quantities we need to analyze the statistical quenched pressure (equal to the free energy up to a multiplying factor $-\beta$) which will be expressed as a function of the order parameters of the model. We thus need to define the Mattis magnetizations relative to the two visible layers for each $\mu \in \{1, \dots, P\}$

$$n_\mu(\sigma) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^N \xi_i^\mu \sigma_i}{N} \in [-1, 1], \qquad m_\mu(\tau) \stackrel{\text{def}}{=} \frac{\sum_{\nu=1}^K \eta_\nu^\mu \tau_\nu}{K} \in [-1, 1]. \tag{15}$$

Besides the Mattis magnetizations, that quantify the retrieval of the patterns, we need other parameters to measure glassy behaviors; as standard within the mean-field theory for

spin-glasses we define the following so-called 2-replica overlaps (see e.g., [20])

$$q_{\sigma\sigma'} = \frac{\sum_{i=1}^{N} \sigma_i \sigma_i'}{N} \in [-1, 1], \quad q_{\tau\tau'} = \frac{\sum_{\nu=1}^{K} \tau_\nu \tau_\nu'}{K} \in [-1, 1], \quad p_{zz'} = \frac{\sum_{\mu=2}^{P} z_\mu z_\mu'}{P-1} \in \mathbb{R}, \tag{16}$$

where (σ, z, τ) and (σ', z', τ') are two independent realizations of the machine's global state.

Clearly, since for clamped mode the input variables σ are fixed a-priori, the correspondent overlap $q_{\sigma\sigma'}$ will be needed as an order parameter only for free mode.

3.1 A Formal Equivalence with a Hopfield Network with Field

In this subsection we focus on the clamped mode and present the correspondent self-consistency equations for the equilibrium states when a single pattern is candidate for retrieval. For the sake of simplicity we consider here the special case in which the number of units in the visible layers is the same, namely $N = K$; the generalization where $N/K \to$ const as $N \to \infty$ does not imply qualitative changes. Then, we define a parameter $\lambda \in (0, 1)$ expressing the relative size of the hidden layer with respect to the common size of the visible ones, that is

$$\lambda \stackrel{\text{def}}{=} \lim_{N\to\infty} \frac{P(N)}{N}. \tag{17}$$

If we look at the expression for $\mathcal{Z}_N^{clamp}$ in (14) we see that the term involved in the sum over $\tau \in \mathscr{I}_K$ equals the Boltzmann factor that would appear in an HN built with K neurons (the visible output ones) and pattern set composed by the K-dimensional vectors $\{\eta^\mu\}_{\mu=1}^{P}$, whose cardinality is the number of hidden units P. The net is then subjected to an external field reproducing the effect generated by the clamped input. Thus, if we define

$$\begin{aligned} \mathcal{H}_{N,h}^{\mathrm{H}}(\tau) &\stackrel{\text{def}}{=} -\frac{1}{2N} \sum_{\nu,\gamma,\mu} \eta_\nu^\mu \eta_\gamma^\mu \tau_\nu \tau_\gamma - \sum_\nu h_\nu \tau_\nu \\ &= -\sum_\mu \left(\frac{N}{2} m_\mu(\tau)^2 + n_\mu(\sigma) m_\mu(\tau) \right) = -\frac{N}{2} \sum_\mu m_\mu(\tau)^2 - \sum_\nu h_\nu \tau_\nu, \end{aligned} \tag{18}$$

where $h_\nu = h_\nu(\xi, \eta, \sigma)$ is the external field

$$h_\nu(\xi, \eta, \sigma) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i,\mu} \xi_i^\mu \eta_\nu^\mu \sigma_i = \sum_\mu \eta_\nu^\mu n_\mu(\sigma); \tag{19}$$

Equation (14) results translated into the following equivalence

$$\mathcal{Z}_N^{clamp} = e^{\frac{\beta N}{2} \sum_\mu n_\mu^2(\sigma)} \mathcal{Z}_{N,h}^{\mathrm{H}} \tag{20}$$

where we defined the partition function of the involved HN as

$$\mathcal{Z}_{N,h}^{\mathrm{H}} \stackrel{\text{def}}{=} \sum_\tau e^{-\beta \mathcal{H}_{N,h}^{\mathrm{H}}(\tau)}. \tag{21}$$

When considering σ fixed, the contribution provided by the interaction between σ and τ emerging from the marginalization results in a linear contribution to the Hamiltonian of the correspondent HN. Clearly, if the input is clamped in a σ "close" to a specific ξ^μ ($n_\mu(\sigma) \sim 1$) then h has a strong component parallel to the pattern η^μ: in this case the energy contribution

provided by the external field is smaller if $m_\mu(\tau) \sim 1$, namely the correspondent pattern η^μ is favored for retrieval.

As previously mentioned, in order to proceed with investigations we need the statistical quenched pressure which, for each $N \in \mathbb{N}$, is defined as

$$\mathcal{A}_N^{clamp} = \frac{1}{N}\mathbb{E} \ln \mathcal{Z}_N^{clamp}$$

where $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\eta}}$ indicates the average over the synaptic weights $\{\eta_\nu^\mu\}_{\nu=1,\dots,N}^{\mu=1,\dots,P}$.

Remark 1 Exploiting a universality hypotesis for slow noise [21–23], we can replace an extensive number of binary patterns with an extensive number of standard Gaussian patterns (those that are not retrieved, also referred to as unmarked) and a finite number of binary patterns (those candidate for retrieval, also referred to as marked). The underlying idea is that the contribution to slow noise provided by non-retrieved Gaussian patterns is the same as that of digital patterns.

Here we study the retrieval of a single pattern η^1 when the input is prescribed considering the following mixed setting:

- Each coupling η_ν^μ between output units τ_ν and hidden units z_μ is sampled independently from a standard Gaussian distribution for $\mu = 2, \dots, P$ and from a Rademacher distribution for $\mu = 1$.
- The variables σ are clamped in a specific configuration $\sigma \in \mathscr{I}_N$ and each coupling ξ_i^μ between an input unit σ_i and a hidden unit z_μ is fixed on $+1$ or -1.

The Mattis magnetizations n_1, m_1, defined as in (15), measure, respectively, the resemblance between ξ^1 and the input σ, and between η^1 and the output τ. Therefore, since the given setting is designed to study the retrieval of the first pattern, it is necessary to highlight the difference between n_1, m_1 and n_μ, m_μ for $\mu = 2, \dots, P$, the latter corresponding to unmarked patterns. In order to do so we change the notation for n_1, m_1 and set

$$n(\sigma) = \frac{\sum_{i=1}^N \xi_i^1 \sigma_i}{N}, \qquad m(\tau) = \frac{\sum_{\nu=1}^K \eta_\nu^1 \tau_\nu}{K}. \tag{22}$$

Each deterministic sequence $n(\sigma), \{n_\mu(\sigma)\}_{\mu=2}^P$ prescribes a specific input while possible interference between the patterns $\boldsymbol{\eta} \equiv \{\eta^\mu\}_{\mu=1}^P$ contributes to the slow noise. As anticipated, we indicate with $\mathbb{E}F(\boldsymbol{\eta})$ the expectation with respect to variables $\boldsymbol{\eta}$ of any quantity F depending on them:

$$\mathbb{E}F(\boldsymbol{\eta}) = \int_{\mathbb{R}^{(P-1)N}} \prod_{\mu=2}^P d\mathbf{G}_1(\eta^\mu) \sum_{\eta^1 \in \{-1,1\}^N} \frac{1}{2^N} F(\boldsymbol{\eta}). \tag{23}$$

We also need the average $\boldsymbol{\Omega}$ with respect to the Boltzmann distribution of any observable quantity $O : \mathbb{R}^P \times \{-1, 1\}^N \to \mathbb{R}$ defined as

$$\boldsymbol{\Omega}(O) = \left(\mathcal{Z}_N^{clamp}\right)^{-1} \sum_{\tau \in \{-1,1\}^N} \int \prod_{\mu=1}^P dG_\beta(z_\mu) O(z, \tau) e^{-\beta \mathcal{H}^{\mathrm{B}}(\sigma, z, \tau)}. \tag{24}$$

Remark 2 This mixed setting affects the external field (19) splitting its components as follows

$$h_\nu = \eta_\nu^1 n(\sigma) + \sum_{\mu=2}^P \eta_\nu^\mu n_\mu(\sigma). \tag{25}$$

The whole Hamiltonian $\mathcal{H}^{\mathrm{H}}_{N,h}(\tau)$ expressed in (18), when referred to our hybrid model, results in

$$\begin{aligned}\mathcal{H}^{\mathrm{H}}_{N,h}(\tau|\sigma) &= -\frac{1}{2N}\sum_{ij=1}^{N}\eta^1_i\eta^1_j\tau_i\tau_j - n(\sigma)\sum_{i=1}^{N}\eta^1_i\tau_i \\ &+ -\frac{1}{2N}\sum_{\mu=2}^{P}\sum_{ij=1}^{N}\eta^\mu_i\eta^\mu_j\tau_i\tau_j - \sum_{\mu=2}^{P}n_\mu(\sigma)\sum_{i=1}^{N}\eta^\mu_i\tau_i \\ &= -\frac{N}{2}m(\tau)^2 - Nn(\sigma)m(\tau) - \frac{N}{2}\sum_{\mu=2}^{P}m_\mu(\tau)^2 - N\sum_{\mu=2}^{P}n_\mu(\sigma)m_\mu(\tau). \quad (26)\end{aligned}$$

Performing a Mattis gauge $\eta^1_i\tau_i \to \tau_i$ on the Boolean part, the resulting Hamiltonian is written as the sum of a Curie–Weiss and an analog Hopfield term (namely an HN provided with all Gaussian patterns), both with a specific external field. Such structure is convenient since it allows us to combine the Guerra interpolation techniques available for both the models and suggests the choice of an interpolation function that combines the ones used in these cases.

Given these hypotesis and considerations we can easily establish a standard interpolation scheme [24] in order to find a sum rule for the quenched statistical pressure of the Boltzmann machine in clamped mode and under the RS approximation. The strategy is to introduce some fictitious external fields in order to"imitate" the internal, recurrently generated field, reproducing its average statistics. Then, in order to recover the second order statistics, the free energy is interpolated smoothly between the case in which all fields are external and all high-order statistics is missing, and the original case, in which fields are all internal. The fictitious fields, acting on each unit involved in the dynamics, are introduced as two classes $\{\tilde{\eta}_i\}_{i=1}^N$ and $\{\tilde{\eta}_\mu\}_{\mu=2}^P$ of i.i.d. $\mathcal{N}(0,1)$ variables, which now participate in the noise average $\mathbb{E}$. Then, with the use of an interpolating parameter $t \in [0,1]$, the following interpolating function is defined:

$$\begin{aligned}\mathcal{Z}^{clamp}_t \stackrel{\text{def}}{=} & e^{\frac{\beta N}{2}n(\sigma)^2}\sum_{\tau} e^{\frac{\beta N}{2}\left(m(\tau)^2+2m(\tau)n(\sigma)\right)t+\Phi N m(\tau)(1-t)} \cdot \\ & e^{A\sum_i \tilde{\eta}_i\tau_i\sqrt{1-t}} \\ & \cdot \int\prod_{\mu=2}^{P}dG_\beta(z_\mu)\; e^{\sum_{\mu=2}^{P}\beta\left\{\frac{d}{2}(1-t)z_\mu^2+\sqrt{N}\left[m_\mu(\tau)\sqrt{t}+n_\mu(\sigma)\right]z_\mu+B\tilde{\eta}_\mu\sqrt{1-t}\,z_\mu\right\}}. \quad (27)\end{aligned}$$

In addition to the fictitious terms $\tilde{\eta}$, we have introduced the auxiliary parameters Φ, A, B (which serve to weight the fields) and a leakage (second-order) term, parametrized by d. As explained in Appendix C, these parameters are chosen so as to separate the contribution of mean and fluctuations of the order parameters in the final expression of the free energy. It is simple to check that $\mathcal{Z}^{clamp}_1 = \mathcal{Z}^{clamp}$ and that $\mathcal{Z}_0$ is made of a series of one-body systems. Furthermore, thanks to the definition (27), we can extend the product Gibbs meausure $\mathbf{\Omega}$ (24) to its interpolating counterpart $\mathbf{\Omega}_t$. It is indeed easy to find an expression $\mathcal{H}_t(\sigma,z,\tau)$ by which $\mathcal{Z}^{clamp}_t \equiv \sum_\tau\int\prod_\mu dG_\beta(z_\mu)e^{-\beta\mathcal{H}_t(\sigma,z,\tau)}$ and

$$\mathbf{\Omega}_t(O) = \left(\mathcal{Z}^{clamp}_t\right)^{-1}\sum_{\tau\in\{-1,1\}^N}\int\prod_{\mu=1}^{P}dG_\beta(z_\mu)O(z,\tau)e^{-\beta\mathcal{H}_t(\sigma,z,\tau)}, \quad (28)$$

for any observable $O : \mathbb{R}^P \times \{-1, 1\}^N \to \mathbb{R}$. The overlaps (16), which play the role of order parameters, involve two realizations of the system, the Boltzmann averages should be thus performed over both configurations. In particular, with some abuse of notation, we use the symbols $\mathbf{\Omega}$, $\mathbf{\Omega}_t$ to also represent the Boltzmann averages over two-system configurations, namely

$$\mathbf{\Omega}_t(O) = \left(\mathcal{Z}_t^{clamp}\right)^{-2} \sum_{\tau\in\{-1,1\}^N} \sum_{\tau'\in\{-1,1\}^N} \int \prod_{\mu=1}^{P} dG_\beta(z_\mu) dG_\beta(z'_\mu) O(z, \tau, z', \tau') e^{-\beta\mathcal{H}_t(\sigma,z,\tau)} e^{-\beta\mathcal{H}_t(\sigma,z',\tau')},$$

with $\mathbf{\Omega} \equiv \mathbf{\Omega}_{t=1}$, for observables $O : \mathbb{R}^P \times \{-1, 1\}^N \times \mathbb{R}^P \times \{-1, 1\}^N \to \mathbb{R}$.

Now, assuming that $\mathcal{Z}_t^{clamp}$ is sufficiently regular, a sum rule for the quenched intensive pressure can be given through the Fundamental Theorem of Calculus:

$$\frac{1}{N} \ln \mathcal{Z}_1^{clamp} = \mathcal{A}_N^{clamp} = \frac{1}{N}\mathbb{E} \ln \mathcal{Z}_0^{clamp} + \int_0^1 \left(\frac{\partial}{\partial t}\frac{1}{N} \ln \mathcal{Z}_t^{clamp}\right)_{t=s} ds. \tag{29}$$

The first step consists in evaluating the one-body term $\frac{1}{N}\mathbb{E} \ln \mathcal{Z}_0^{clamp}$ and this is feasible with standard calculations as reported in Appendix B. In particular, we find

$$\begin{aligned}\frac{1}{N}\mathbb{E} \ln \mathcal{Z}_0^{clamp} = \frac{\beta}{2} n(\sigma)^2 + \ln 2 - \frac{\lambda}{2} \ln[1 - \beta(1 - \bar{q})] + \frac{\beta}{2[1 - \beta(1 - \bar{q})]} \sum_{\mu=2}^{P} n_\mu(\sigma)^2 \\ + \mathbb{E} \ln \cosh \left\{\beta \left[\eta^1(n + \bar{m}) + \sqrt{\lambda \bar{p}}\tilde{\eta}\right]\right\} + \frac{\lambda\beta\bar{q}}{2(1 - \beta(1 - \bar{q}))}\end{aligned} \tag{30}$$

where η^1, $\tilde{\eta}$ are respectively a Rademacher variable and a standard Gaussian one. We consider the RS approximation, under which order parameters are supposed to be self-averaging in the thermodynamic limit. Variables $\bar{m}$, $\bar{q} \in [-1, 1]$, $\bar{p} \in \mathbb{R}$ are therefore used in (30) to represent the average thermodynamic values of m, q, p.

The calculations concerning the second term in the right hand side of (29) are long but straightforward, they are handled in Appendix C and yield to

$$\frac{1}{N}\frac{\partial}{\partial t} \ln \mathcal{Z}_t = -\frac{\beta}{2}\bar{m}^2 - \frac{\beta^2\lambda}{2}\bar{p}(1 - \bar{q}) + \frac{\beta}{2}\langle (m - \bar{m})^2 \rangle_t - \frac{\beta^2\lambda}{2}\langle (p_{zz'} - \bar{p})(q_{\tau\tau'} - \bar{q}) \rangle_t \tag{31}$$

where we used the symbol $\langle\cdot\rangle_t$ for $\mathbb{E}\mathbf{\Omega}_t(\cdot)$. The self-average assumption of the order parameters corresponds to ignoring fluctuations, namely we neglect the last two terms in (31). Then, using the definition (33) in (30) and plugging it together with (31) into (29), we obtain the final expression for the statistical pressure of the three-layer HBM in clamped mode. We call it $\mathcal{A}_{RS}^{clamp}$, since it does not include fluctuations of the order parameters, and this corresponds to the RS solution

$$\begin{aligned}\mathcal{A}_{RS}^{clamp} = \frac{\beta}{2} n(\sigma)^2 + \ln 2 - \frac{\lambda}{2} \ln(1 - \beta(1 - \bar{q})) + \frac{\beta}{2(1 - \beta(1 - \bar{q}))} \sum_{\mu=2}^{P} n_\mu(\sigma)^2 \\ + \mathbb{E} \ln \cosh \left[\beta \left(\eta^1(n + \bar{m}) + \sqrt{\lambda \bar{p}}\tilde{\eta}\right)\right] + \frac{\lambda\beta\bar{q}}{2(1 - \beta(1 - \bar{q}))} \\ - \frac{\beta}{2}\bar{m}^2 - \frac{\beta^2\lambda}{2}\bar{p}(1 - \bar{q}).\end{aligned} \tag{32}$$

Thanks to this explicit expression for $\mathcal{A}_{RS}^{clamp}$ we can detect a single quantity whose physical meaning concerns the overall effect on the statistical quenched pressure of the components of h parallel to unmarked patterns. That is the fourth term in (32) which involves all the Mattis magnetizations $n_{\mu,N}(\sigma)$ for $\mu = 2, \dots, P$. Considering its thermodynamic limit, we define

$$C \stackrel{\text{def}}{=} \lim_{N\to\infty} \lambda^{-1} \sum_{\mu=2}^{P(N)} n_{\mu,N}(\sigma)^2 = \lim_{N\to\infty} \frac{\sum_{\mu=2}^{P(N)} (\xi^\mu \cdot \sigma)^2}{P(N)N}, \tag{33}$$

and consequently write

$$\begin{aligned} \mathcal{A}_{RS}^{clamp} = & \frac{\beta}{2} n(\sigma)^2 + \ln 2 - \frac{\lambda}{2} \ln(1 - \beta(1-\bar{q})) + \frac{\beta\lambda C}{2(1-\beta(1-\bar{q}))} \\ & + \mathbb{E} \ln \cosh \left[\beta \left(\eta^1 (n + \bar{m}) + \sqrt{\lambda \bar{p}} \tilde{\eta} \right) \right] + \frac{\lambda\beta\bar{q}}{2(1-\beta(1-\bar{q}))} \\ & - \frac{\beta}{2} \bar{m}^2 - \frac{\beta^2 \lambda}{2} \bar{p}(1-\bar{q}). \end{aligned} \tag{34}$$

In the high-storage regime $\lambda > 0$, on which we are focusing, the sum $\sum_{\mu=2}^{P(N)} n_{\mu,N}(\sigma)^2$ becomes an infinite sum, we thus need to consider the analytical assumption $C < \infty$. An analysis of various ways in which it is possible to prescribe the input consistently with this assumption can be found in Sect. 3.2.

To find the value of the free energy of the system, according to the minimum energy principle and the maximum entropy principle, we have to look for values of the order parameters in which $\mathcal{A}_{RS}^{clamp}$ is maximized (and the free energy minimized).

To achieve this we derive Eq. (34) with respect to p, q, m (dropping the bar over $\bar{p}$, $\bar{q}$ and $\bar{m}$ to lighten notation) and consider the stationarity condition given by $\partial_p \mathcal{A}_{RS}^{clamp} = 0$, $\partial_q \mathcal{A}_{RS}^{clamp} = 0$, $\partial_m \mathcal{A}_{RS}^{clamp} = 0$. Explicit calculations are performed in Appendix D and bring to the following self-consistency equations:

$$q = \mathbb{E}_{\eta^1} \int_{\mathbb{R}} dG_1(\tilde{\eta}) \tanh^2 \left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1 (n+m)) \right] \tag{35a}$$

$$m = \mathbb{E}_{\eta^1} \eta^1 \int dG_1(\tilde{\eta}) \tanh \left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1 (n+m)) \right] \tag{35b}$$

$$p = \frac{q + C}{(1-\beta(1-q))^2}. \tag{35c}$$

The whole phase space is described by the four parameters $(\beta, \lambda, n, C) \in \mathbb{R}^+ \times \mathbb{R}^+ \times [-1, 1] \times \mathbb{R}^+$, where the additional ones n and C reflect the clamped input and, respectively, can help and disturb the retrieval of the marked pattern η^1.

Remark 3 Turning off the external field h (25) acting on the output layer and reflecting the presence of a clamped input is equival to setting $n, C = 0$. As expected this makes equations (35a) (35b) (35c) the ones for a mixed HN with one marked pattern [22].

Remark 4 It is worth comparing Eqs. (35) to the analogous ones pertaining to the standard Hopfield model with external field components h'_μ, with $\mu = 1, \dots, P$: they are equivalent as long as we identify n with h'_1 and we set $C = 0$, (see e.g., [26]). In fact, according to Eq. 33, C provides a measure of the correlation between the marked pattern and the unmarked ones and, in the standard HN accounting for Rademacher patterns, this is vanishing in the average.

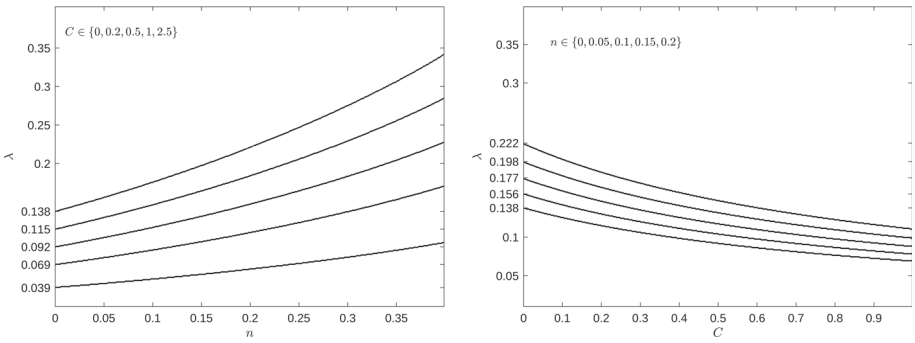

Fig. 2 Left panel: critical lines in the plane (n, λ) between the spin-glass and ferromagnetic phases when $T \to 0$ and under the RS assumption. Lines from top to bottom correspond to different increasing values of the noise parameter C. Setting $C = 0$ (namely looking at the first line from top) the standard Hopfield model with one marked pattern is recovered. Right panel: RS critical lines at zero temperature in the plane (C, λ). Lines from top to bottom correspond to different decreasing values of the signal parameter n. Setting $n = 0$ the case of a standard HN with one marked pattern and load $\lambda(1 + C)$ is recovered, namely the relative line follows the law $\lambda(C) = 0.138(1 + C)^{-1}$

Also, noticing that C appears explicitly in the numerator of the expression (35c) for p, that is known to quantify the noise due to unmarked patterns, we see that the parameter C tends to impair retrieval.

We can easily handle Eqs. (35a), (35b), (35c) when approaching the zero temperature case. Specifically, in the Appendix E we see how in the limit $T^{-1} = \beta \to \infty$ they can be re-written as a single equation in the variable $y = (2\lambda p)^{-1}(n + m)$ given by

$$y = \frac{\text{erf}(y) + n}{\frac{2}{\sqrt{\pi}} e^{-y^2} + \sqrt{2\lambda(1 + C)}}. \tag{36}$$

Equation (36) is the one for a standard HN at zero temperature with one marked pattern [25] and load $\lambda(1 + C)$ when an external field parallel to the marked pattern is considered in the Hamiltonian and has intensity n. In particular the effect caused by the components of h parallel to unmarked patterns and reflecting the information coming from the correspondent input pattern is encoded in $C > 0$ and is equivalent to an increase of the intrinsic noise as if the load was bigger and equal to $\lambda(1 + C) \geq \lambda$.

A first analysis is thus immediate when thinking to the term $\lambda(1 + C)$ as a unique noise term (see e.g., [1,25]).

When $n \neq 0$ a nonzero Mattis magnetization m with the marked pattern η^1 develops continously from zero, linearly with n. Moreover, for each $n \in [-1, 1]$ there exists a critical value for $\lambda(1 + C)$ above which only the spin-glass state exists, while for $\lambda(1 + C)$ lower than this critical value a retrieval state, with a high m, appears. This critical value increases with increasing n (as visible in the first panel of Fig. 2): this means that for $n \neq 0$ the retrieval is possible even for values $\lambda(1 + C) > \lambda_c'$, where λ_c' is the critical value for the load in a standard HN when no external field is present.

The first panel in Fig. 2 shows the critical lines increasing as functions of n for different values of C. For each C their values at $n = 0$ follow the law $\lambda(1 + C) = \lambda_c'$. As C grows, the retrieval region is reduced from the one valid for $C = 0$. For small values of n, as λ exceeds the relative critical value, the Mattis magnetization m with the marked pattern is abruptly reduced to zero. The reduction is less intense as n grows and visible in Fig. 3 where the trade

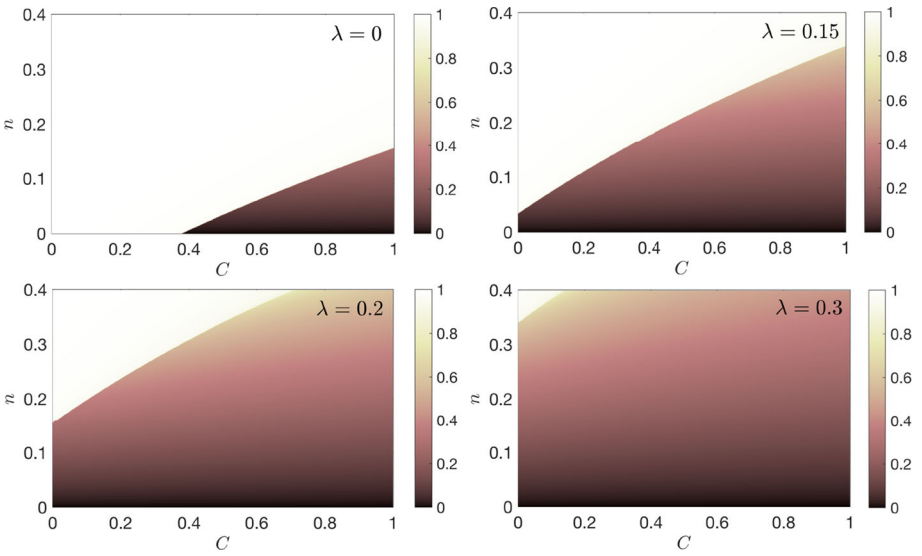

Fig. 3 Intensity plots of the magnetization m in the limit $T \to 0$ in the (C, n) plane; each panel corresponds to a different value of the load λ. Data depicted are obtained as numerical solutions of the zero temperature equation (36) setting $m = \mathrm{erf}(y)$ through a fixed point method

off between parameters (n, C) is observed looking at the transition of m for fixed values of λ.

The second panel in Fig. 2, analogous to the first, shows critical lines as functions of C for different values of n and tells that even for very small values of n the retrieval is strongly improved. For example setting $n = 0.05$ the critical load approaches the value 0.156 as $C \to 0$. When $\lambda(1 + C)$ further decreases the retrieval state becomes globally stable; eventually the spin glass state disappears and only the retrieval one persists.

3.2 Examples

In this subsection we resume the self-consistency equations (35) at temperature $T > 0$ and briefly discuss some interesting settings as for the input layer.

(1) Orthogonal input patterns

We can consider, for instance, the case in which the machine has learnt with no errors some input informations encoded by orthogonal patterns. This is the simplest setting that can be investigated giving the clamped input patterns $\boldsymbol{\xi}$ as uniformly orthogonal deterministic vectors, $\xi^\mu \cdot \xi^\nu = 0 \; \forall N \in \mathbb{N}$. If then $\sigma = \xi^1$ we immediately see from their definitions (22), (33) that in this case $n = 1$, $C = 0$: the field component related η^1 is the strongest possible, corresponding to an Hopfield network with external field of intensity 1 parallel to η^1. Clearly, clamping σ not exactly on ξ^1, but allowing for a certain percentage of errors, changes the values of (n, C) for each realization. Figure 4 exhibits intensity plots in the plane (C, n) of the Mattis magnetization relative to the output layer for different fixed values of the temperature and load. We observe that the transition seems to occur less sharply as T and λ increase but still very small values of the fixed magnetization of the input layer n strongly favors the retrieval despite the presence of positive values for C.

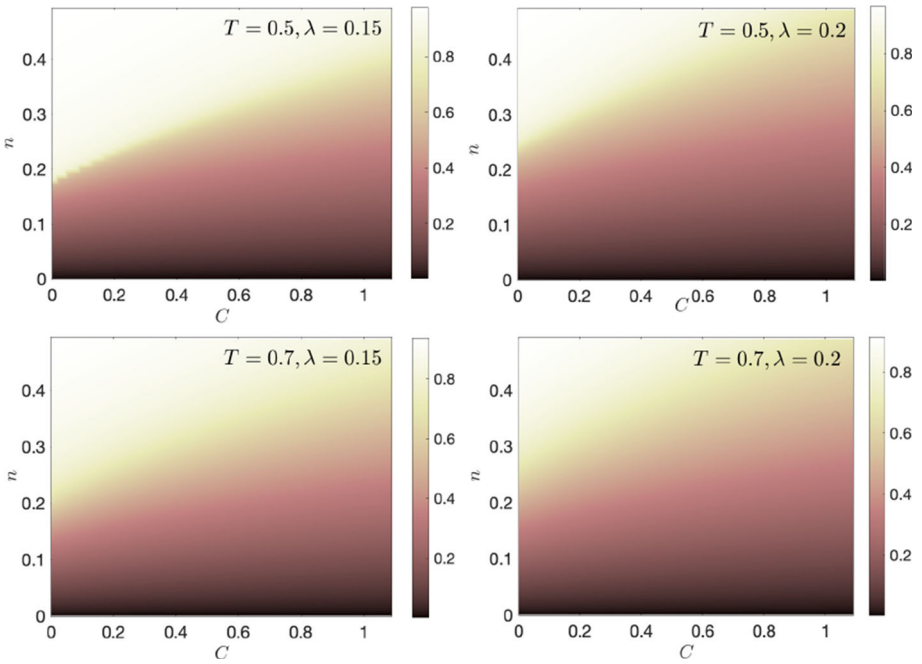

Fig. 4 Intensity plots in the (C, n) plane of the magnetization m are shown in these panels for fixed temperature and load. Different values of n and C correspond to different deterministic realizations of the clamped input layer

(2) A finite number of correlated input patterns

Another setting compatible with our analysis is the one in which the informations encoded by the patterns $\boldsymbol{\xi}$ results in non orthogonal vectors. Specifically we could assume that $\sigma = \xi$ and that there exists a finite number $P_{\text{corr}} < +\infty$ such that

$$\xi^\mu \cdot \xi = a_\mu N \quad \text{if} \quad \mu = 1, \dots, P_{\text{corr}}$$
$$\xi^\mu \cdot \xi = 0 \quad \text{if} \quad \mu > P_{\text{corr}}$$

for some values $a_\mu \in [-1, 1]$, $\mu = 1, \dots, P(N)$, uniformly in N. We would get

$$n = a_1, \quad C = \lambda^{-1} \sum_{\mu=2}^{P(N)} n_\mu^2(\sigma) = \lambda^{-1} \sum_{\mu=1}^{P_{\text{corr}}} a_\mu^2.$$

(3) Orthogonal-in-the-average inputs

Typically, the training of a machine is performed on a non-exhaustive or imperfect dataset [27,28]. Therefore, it is reasonable to face this problem from a probabilistic perspective.

To this aim we can introduce the input patterns ξ_i^μ, $\mu > 1$ as quenched variables and make them participate in the avaraging procedure considering $\mathcal{A}^B = \frac{1}{N}\mathbb{E} \ln \mathcal{Z}^{clamp}$ with $\mathbb{E} = \mathbb{E}_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{\eta}}$. For example, we can assume that the clamped layer is given as $\sigma = \xi^1$ and that each ξ_i^μ for $\mu > 1$ is independently extracted from a Rademacher distribution. Each term of the form $\xi^\mu \cdot \sigma = \xi^\mu \cdot \xi^1 = \sum_{i=1}^N \xi_i^\mu \xi_i^1$ will be thus equal to the net displacement after N steps of a random, simple and symmetric path (since in this case the products $\xi_i^\mu \xi_i$ are extracted as independent Rademacher variables). For a large number of steps we can

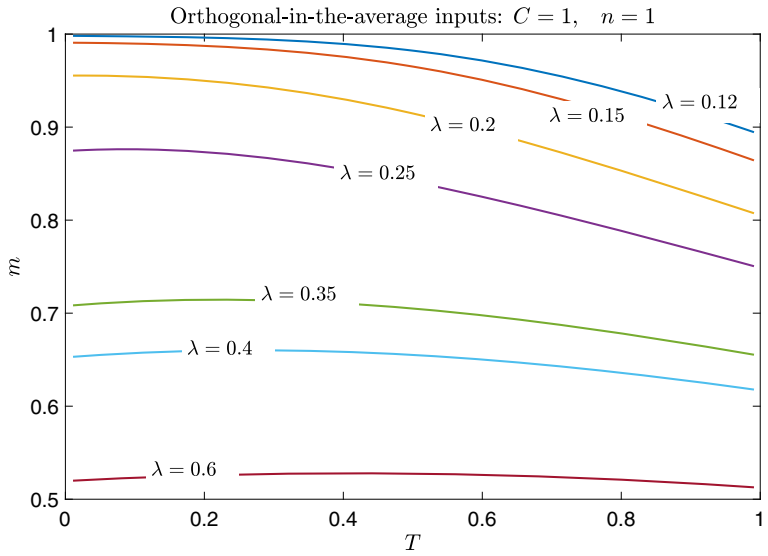

Fig. 5 The average Mattis magnetization n of the output layer obtained by numerically solving (35) is plotted as a function of the temperature for different fixed values of the load when $(C, n) = (1, 1)$. This case corresponds to the one of uncorrelated stochastic input patterns, when the input layer σ is clamped on one of them

thus approximate the distribution of $\xi^\mu \cdot \sigma$ with a Guassian centered random variable whose standard deviation is equal to the square root of the number of steps (see [26] and references therein), namely $\xi^\mu \cdot \sigma = \mathcal{O}(\sqrt{N})$, therefore

$$\sum_{\mu=2}^{P(N)} \left(\xi^\mu \cdot \sigma\right)^2 = \mathcal{O}\left(P(N)N\right).$$

Looking at the definition of the C parameter (33) we see that this case, correspondent to the one of binary uncorrelated patterns, can be inserted in our analysis taking $n = 1$ and $C = 1$. Figure 5 shows how the magnetization of the output layers remains in this case very high for high loads.

3.3 A Formal Equivalence with a Hopfield Network

We now consider what we called a free mode, correspondent to the case in which all the units $(\sigma, z, \tau) \in \mathcal{I}_N \times \mathbb{R}^P \times \mathcal{I}_K$ of the three-layer HBM are free to evolve. In this case the partition function is the one in (13) and equals

$$\mathcal{Z}_N^{free} = \sum_\sigma \sum_\tau e^{-\beta \sum_\mu \left(-\frac{1}{2N}(\xi^\mu \cdot \sigma)^2 - \frac{1}{2N}(\eta^\mu \cdot \tau)^2 - \frac{1}{N}(\xi^\mu \cdot \sigma)(\eta^\mu \cdot \tau)\right)}. \tag{37}$$

Since the case $N \neq K$ can be here easily included we give the relative size of the input layer with respect to the total number of visible units as

$$\gamma = \lim_{N \to \infty} \frac{N}{N + K(N)} > 0 \tag{38}$$

stressing that this density is fixed as N and $K = K(N)$ grows so that we can write γ instead of $N(N+K)^{-1}$. Now, through the definitions of the Mattis magnetizations (15) and using that $K(N+K)^{-1} = 1-\gamma$ we can rewrite (37) as

$$\mathcal{Z}_N^{free} = \sum_{\sigma}\sum_{\tau} e^{\frac{1}{\gamma}\frac{\beta(N+K)}{2}\sum_{\mu}\left(\gamma n_\mu(\sigma)+(1-\gamma)m_\mu(\tau)\right)^2}. \tag{39}$$

If we now define

$$\mathcal{H}_N^{\mathrm{H}}(\sigma,\tau) \stackrel{\mathrm{def}}{=} -\frac{1}{2(N+K)}\left(\sum_{i,j,\mu}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \sum_{\nu,\gamma,\mu}\eta_\nu^\mu\eta_\gamma^\mu\tau_\nu\tau_\gamma + \sum_{i,\nu,\mu}\xi_i^\mu\eta_\nu^\mu\sigma_i\tau_\nu\right) \tag{40}$$

and the correspondent partition function

$$\mathcal{Z}_N^{\mathrm{H}} = \sum_{\sigma}\sum_{\tau} e^{-\beta\mathcal{H}^{\mathrm{H}}(\sigma,\tau)}, \tag{41}$$

we see that Eq. (39) is equivalent to (41) upon a temperature rescaling $\beta \to \beta\gamma$.

In this case the notational distinction between the input and output variables σ, τ and their respective connections clearly makes the way the Hopfield Hamiltonian appears in (40) redundant, the distinction between units of type σ_i and τ_ν is actually just conceptual since we are considering the case in which input and output layers are not directly connected. The HN correspondent to the three-layer HBM evolving in free mode is built with all the $N+K$ visible units and the pattern set, whose cardinality P is the number of hidden units, is composed by the $(N+K)$-dimensional vectors $\{(\xi^\mu,\eta^\mu)\}_{\mu=1}^P$.

The replica symmetric solution, that is a sum-rule for the quenched statistical pressure $\mathcal{A}_N^{free} = (N+K)^{-1}\mathbb{E}\ln\mathcal{Z}_N^{free}$ under a self averaging hypothesis for the order parameters, can be found with tools and calculations strictly similar to those presented for the clamped case.

As pointed out before, the distinction between input and output variables is in this case redundant so that if we define for each $(\sigma,\tau),(\sigma',\tau') \in [\mathscr{I}_N \times \mathscr{I}_K]^2$ a *total* 2-replicas overlap and a *total* Mattis magnetization for $(\sigma,\tau) \in \mathscr{I}_N \times \mathscr{I}_K$ as

$$q_{\mathrm{tot}}(\sigma,\tau,\sigma',\tau') \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^N\sigma_i\sigma_i' + \sum_{\nu=1}^K\tau_\nu\tau_\nu'}{N+K} \in [-1,1],$$
$$m_{\mathrm{tot}} \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^N\xi_i^1\sigma_i + \sum_{\nu=1}^K\eta_\nu^1\tau_\nu}{N+K} \in [-1,1]$$

the RS solution when only a single pattern is candidate for retrieval, analogous to (34), would appear as

$$\begin{aligned}\mathcal{A}_{RS}^{free} = {} & \ln 2 + \mathbb{E}\ln\cosh\left(\beta m_{\mathrm{tot}}\xi + \sqrt{\beta\lambda p}\tilde{\eta}\right) \\ & - \frac{\lambda\beta}{2}\frac{q_{\mathrm{tot}}}{(1-\beta(1-q_{\mathrm{tot}}))} - \frac{\lambda}{2}\ln(1-\beta(1-q_{\mathrm{tot}})) \\ & - \frac{\beta\lambda}{2}p(1-q_{\mathrm{tot}}) - \frac{\beta}{2}m_{\mathrm{tot}}^2,\end{aligned} \tag{42}$$

where ξ is a Rademacher variable independent from $\tilde{\eta} \sim \mathcal{N}(0,1)$ and $\mathbb{E} = \mathbb{E}_\xi\mathbb{E}_{\tilde{\eta}}$. As expected, this expression is nothing but the RS free energy of a Hopfield network with a single marked pattern [22]. It is easly checked that forcing $\partial_{m_{\mathrm{tot}}}\mathcal{A}_{RS}^{free} = 0$ $\partial_{q_{\mathrm{tot}}}\mathcal{A}_{RS}^{free} = 0$

$\partial_p \mathcal{A}_{RS}^{free} = 0$ yield to the following self-consistency equations

$$m_{\text{tot}} = \mathbb{E} \tanh \left(\beta m_{\text{tot}} \xi + \beta \tilde{\eta} \frac{\sqrt{\lambda q_{\text{tot}}}}{1 - \beta(1 - q_{\text{tot}})} \right); \tag{43}$$

$$q_{\text{tot}} = \mathbb{E} \tanh^2 \left(\beta m_{\text{tot}} \xi + \beta \tilde{\eta} \frac{\sqrt{\lambda q_{\text{tot}}}}{1 - \beta(1 - q_{\text{tot}})} \right); \tag{44}$$

$$p = \frac{\beta q_{\text{tot}}}{(1 - \beta(1 - q_{\text{tot}}))^2}. \tag{45}$$

As expected, the self-consistent equations for the Hopfield model are recovered [26].

In particular, the transition line between the spin glass and the paramagnetic phase is given by equation $\frac{\beta^2 \lambda}{(1-\beta)^2} = 1$.

4 Conclusions

In this work we considered a three-layer HBM, whose constituting neurons are of different nature: the visible ones (σ, τ) are binary, while the hidden ones z are Gaussian. Such a system can be trained by means of a sample drawn from an unknown target distribution. The network architecture allows for different training modes according to whether we want the final system to be able to generate new couples $\{\hat{\sigma}^{(k)}, \hat{\tau}^{(k)}\}_k$ or to be able to reply with a certain output $\hat{\tau}^{(k)}$ to a certain input $\hat{\sigma}^{(k)}$, mimicking what we would obtain from the target distribution. These training modes imply to make the system evolve under a suitable dynamics possibly displaying constraints on the class of neurons that are free to evolve. In particular, we focused on the case where all neurons are free to evolve and on the case where only output and hidden neurons are free to evolve while input neurons are clamped. We proved that the three-layer RBM subject to such constraints is equivalent, respectively, to a HN with size given by the overall number of visible neurons and to a HN with size given by the number of output neurons and in the presence of an external field and an additional slow noise. The former recovers the well-known theory developed by Amit, Gutfreund and Sompolinsky, while for the latter we accomplished a statistical-mechanics investigation obtaining an explicit expression for the free energy in the thermodynamic limit, that is exact under the replica symmetry assumption. Remarkably, the external field and the additional noise are related to the statistical properties of the data used for training. Let us assume that data are encoded in terms of P binary vectors $\{\xi^\mu\}_{\mu=1}^P$: then, if the data set is not broad then the input σ can be set fairly aligned with a certain vector, say ξ^1, and if the data vectors display a poor correlation (i.e., $\xi^\mu \cdot \xi^\nu \approx \delta_{\mu,\nu}$), then the additional noise is vanishing and viceversa.

This result is consistent with the fact that shallow structures proves to be suitably only for structureless data as it is able to capture only first-order and second-order moments hidden in the training data.

Further, we notice that the three-layer HBM considered here, for a special choice of parameters, can be looked at as an autoencoder with one hidden layer in such a way that the above mentioned results can be read also from that perspective [13]. In particular, such a shallow autoencoder can reconstruct information correctly as long as information can be encoded in a relatively small (compared to the outer layer size) number binary vectors and as long as such information displays a vanishing structure.

Acknowledgements EA is grateful to Adriano Barra, Alberto Fachechi and Francesco Alemanno for useful discussions, and to Università Sapienza di Roma for financial support (Progetto Ateneo RM120172B8066CB0). GS is grateful to Massimiliano Viale for enlightening discussions.

Appendix A

In this appendix we give a dynamical route for the machine in such a way that the related equilibrium distribution coincide with the Gibbs measure (8). Through this approach, which reproduces the one proposed in [8], the activity in the layers follows different dynamics, including the fact that digital units change in discrete steps while analog ones change continuously in time.

For the digital visible units is imposed a standard parallel Glauber dynamics for Ising-type systems [1]. We consider a specific time-step and update each visible unit, when it is involved in the dynamics, instantaneously. At every step the probabilities of finding digital units in a specific state are dependent on the state assumed in that instant by z and determined by the total fields acting on them: $\frac{1}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu$ on σ_i and $\frac{1}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu$ on τ_ν, where the normalization factor $\sqrt{N}$ is considered in order to obtain non-trivial thermodynamic limits. These probabilities are

$$\mathcal{G}(\sigma_i|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sigma_i \sum_\mu \xi_i^\mu z_\mu}}{e^{\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu}}, \qquad \mathcal{G}(\tau_\nu|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\tau_\nu \sum_\mu \eta_\nu^\mu z_\mu}}{e^{\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu}}. \tag{A1}$$

The joint probability mass functions $\mathcal{G}(\sigma|z)$, $\mathcal{G}(\tau|z)$ are now completely determined by products of individual probabilities (see e.g. [26])

$$\mathcal{G}(\sigma|z) = \prod_i \mathcal{G}(\sigma_i|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sum_{i\mu}\sigma_i\xi_i^\mu z_\mu}}{\prod_i \left(e^{\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu}\right)}, \tag{A2}$$

$$\mathcal{G}(\tau|z) = \prod_\nu \mathcal{G}(\tau_\nu|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sum_{\nu\mu}\tau_\nu\eta_\nu^\mu z_\mu}}{\prod_\nu \left(e^{\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu}\right)}, \tag{A3}$$

$$\mathcal{G}(\sigma,\tau|z) = \mathcal{G}(\sigma|z)\mathcal{G}(\tau|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sum_\mu z_\mu\left(\sum_i \sigma_i\xi_i^\mu + \sum_\nu \tau_\nu\eta_\nu^\mu\right)}}{Z(\beta,z,\xi,\eta)} = \frac{e^{\beta z\cdot I(\sigma,\tau)}}{Z(z)} \tag{A4}$$

where

$$Z(z) \stackrel{\text{def}}{=} \prod_{i,\nu}\left(e^{\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \xi_i^\mu z_\mu}\right)\left(e^{\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu} + e^{-\frac{\beta}{\sqrt{N}}\sum_\mu \eta_\nu^\mu z_\mu}\right)$$

is the normalization factor and I is the vector defined in (7) whose P entries quantify the field felt by the hidden units $\{z_\mu\}_{\mu=1}^P$, namely $I(\sigma,\tau) = \{I_\mu(\sigma,\tau)\}_{\mu=1}^P$ with

$$I_\mu(\sigma,\tau) = \frac{1}{\sqrt{N}}\left(\sum_i \xi_i^\mu\sigma_i + \sum_\nu \eta_\nu^\mu\tau_\nu\right), \qquad \forall\mu\in\{1,\dots,P\}. \tag{A5}$$

For what concerns the activity in the hidden analog layer we consider an Ornstein-Uhlenbeck diffusion process correspondent to the following stochastic differential equation

$$\mathcal{T} dz_\mu = -z_\mu(t)dt + I_\mu(\sigma,\tau)dt + \sqrt{\frac{2\mathcal{T}}{\beta}} dW_\mu(t) \tag{A6}$$

where $\{W_\mu(t)\}_{\mu=1}^{P}$ are independent Weiner processes providing a white Gaussian noise with zero mean and covariance $\mathrm{Cov}(dW_\mu(t), dW_{\mu'}(t')) = \delta_{\mu\mu'}\delta(t-t')$. The three terms in the right hand side of (A6) are respectively a leakage term, the input signal and a noise source. While β tunes the strength of the fluctuations, the parameter $\mathcal{T} \in \mathbb{R}^+$ gives the timescale of the dynamics and is assumed to be much smaller than the time-step used to update digital units. For fixed values of σ, τ, the equilibrium distribution of z_μ is a Gaussian distribution centered in the input signal which is equal to

$$\mathcal{G}(z_\mu|\sigma,\tau) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(z_\mu - I_\mu)^2}. \tag{A7}$$

In order for this equilibrium distribution to hold, the activity of digital units σ, τ must be constant, while in fact it depends on time. However, we assumed that the timescale of diffusion is much faster than the steps at which the digital units are updated. Hence, a different equilibrium distribution for z, characterized by different values of σ, τ, holds between each subsequent update of σ, τ. Since noise is uncorrelated between different hidden units and given the mean-field nature of the model, they evolve independently and their joint equilibrium distribution is the product of individual distributions

$$\begin{aligned}\mathcal{G}(z|\sigma,\tau) = \prod_\mu \mathcal{G}(z_\mu|\sigma,\tau) &= \left(\frac{\beta}{2\pi}\right)^{\frac{P}{2}} e^{-\frac{\beta}{2}\|z - I(\sigma,\tau)\|^2} \\ &= \left(\frac{\beta}{2\pi}\right)^{\frac{P}{2}} e^{-\frac{\beta}{2}\|z\|^2 - \frac{\beta}{2}\|I(\sigma,\tau)\|^2 + \beta z\cdot I(\sigma,\tau)}.\end{aligned} \tag{A8}$$

Through the conditional distributions (A4), (A8) we can calculate the probability for the visible layers thanks to Bayes rule:

$$\mathcal{G}(\sigma,\tau) = \frac{\mathcal{G}(\sigma,\tau|z)}{\mathcal{G}(z|\sigma,\tau)}\mathcal{G}(z) = \frac{e^{\beta z\cdot I}}{Z(z)}\left(\frac{\beta}{2\pi}\right)^{-\frac{P}{2}} e^{\frac{\beta}{2}\|z\|^2 + \frac{\beta}{2}\|I\|^2 - \beta z\cdot I}\mathcal{G}(z) = c(z)e^{\frac{\beta}{2}\|I\|^2} \tag{A9}$$

where $c(z) = \left(2\pi\beta^{-1}\right)^{\frac{P}{2}} Z(z)^{-1}\mathcal{G}(z)e^{\frac{\beta}{2}\|z\|^2}$ does not depend on (σ,τ). Since the marginal distribution $\mathcal{G}(\sigma,\tau)$ can not depend on z, the factor $c(z)$ must be constant and we can write

$$\mathcal{G}(\sigma,\tau) \propto e^{\frac{\beta}{2}\|I(\sigma,\tau)\|^2}. \tag{A10}$$

We notice that considering the clamped mode consists in fixing the variables σ, in this case strictly analogous calculations would give

$$\mathcal{G}(\tau|\sigma) \propto e^{\frac{\beta}{2N}\|\eta\cdot\tau\|^2 + \frac{\beta}{N}(\xi\cdot\sigma)\cdot(\eta\cdot\tau)}, \tag{A11}$$

which reflects (20).

Using (A8) and (A10) we now get

$$\mathcal{G}(\sigma,z,\tau) = \mathcal{G}(z|\sigma,\tau)\mathcal{G}(\sigma,\tau) \propto e^{-\frac{\beta}{2}\|z\|^2 + \beta z\cdot I(\sigma,\tau)},$$

namely

$$\mathcal{G}(\sigma, z, \tau) \propto e^{-\beta\left(\frac{1}{2}\|z\|^2 - \frac{1}{\sqrt{N}} z\cdot\xi\cdot\sigma - \frac{1}{\sqrt{N}} z\cdot\eta\cdot\tau\right)}, \tag{A12}$$

which is our claim (8).

Appendix B

In this appendix we calculate the interpolating free energy for $t = 0$ when the HBM evolves in clamped mode. We have

$$\begin{aligned}
\mathcal{Z}_0^{clamp} &= e^{\frac{\beta N}{2} n(\sigma)^2} \sum_{\tau} e^{\Phi N m(\tau) + A\sum_i \tilde{\eta}_i \tau_i} \cdot \\
&\quad \cdot \int \prod_{\mu=2}^{P} dG_\beta(z_\mu) e^{\sum_{\mu=2}^{P} \beta\left[\frac{d}{2} z_\mu^2 + (B\tilde{\eta}_\mu + \sqrt{N} n_\mu(\sigma)) z_\mu\right]} \\
&= e^{\frac{\beta N}{2} n(\sigma)^2} \sum_{\tau} e^{\sum_i (A\tilde{\eta}_i + \Phi\eta_i^1)\tau_i} \cdot \frac{1}{(1-d)^{\frac{P-1}{2}}} \prod_{\mu=1}^{P-1} e^{\frac{\beta}{2(1-d)}(B\tilde{\eta}_\mu + \sqrt{N} n_\mu(\sigma))^2} \\
&= e^{\frac{\beta N}{2} n(\sigma)^2}\ 2^N \prod_{i=1}^{N} \cosh\left(\Phi\bar{\eta}_i + A\tilde{\eta}_i\right) \cdot \frac{1}{(1-d)^{\frac{P-1}{2}}} e^{\frac{\beta}{2(1-d)} \sum_{\mu=2}^{P} (B\tilde{\eta}_\mu + \sqrt{N} n_\mu(\sigma))^2}.
\end{aligned} \tag{B1}$$

Consequently, the associated intensive pressure is

$$\begin{aligned}
\frac{1}{N}\mathbb{E}\ln \mathcal{Z}_0^{clamp} &= \frac{\beta}{2} n(\sigma)^2 + \ln 2 + \mathbb{E}\ln\cosh(\Phi\eta^1 + A\tilde{\eta}) - \frac{\lambda}{2}\ln(1-d) \\
&\quad + \frac{\beta}{2N(1-d)} \sum_{\mu=2}^{P} \left(B^2 \mathbb{E}\tilde{\eta}_\mu^2 + 2B\sqrt{N} n_\mu(\sigma)\mathbb{E}\tilde{\eta}_\mu + N n_\mu(\sigma)^2\right) \\
&= \frac{\beta}{2} n(\sigma)^2 + \ln 2 - \frac{\lambda}{2}\ln(1-d) + \frac{\beta}{2(1-d)} \sum_{\mu=2}^{P} n_\mu(\sigma)^2 \\
&\quad + \frac{\lambda\beta B^2}{2(1-d)} + \mathbb{E}\ln\cosh(\Phi\eta^1 + A\tilde{\eta})
\end{aligned} \tag{B2}$$

where we used $\mathbb{E}\tilde{\eta}_\mu^2 = 1$ and $\mathbb{E}\tilde{\eta}_\mu = 0$ defining η^1 as a Rademacher random variable and $\tilde{\eta}$ as a standard Gaussian. We show in Appendix C that the following choices for the free parameters substantially simplifies the expression of the statistical pressure, making the second order fluctuations of the order parameters explicit

$$\phi = \beta(n(\sigma) + \bar{m}), \qquad d = \beta(1-\bar{q}), \qquad B = \sqrt{\bar{q}}, \qquad A = \beta\sqrt{\lambda\bar{p}}. \tag{B3}$$

Plugging these values in (B2) we find

$$\begin{aligned}
\frac{1}{N}\mathbb{E}\ln \mathcal{Z}_0^{clamp} &= \frac{\beta}{2} n(\sigma)^2 + \ln 2 - \frac{\lambda}{2}\ln(1-\beta(1-\bar{q})) + \frac{\beta}{2(1-\beta(1-\bar{q}))} \sum_{\mu=2}^{P} n_\mu(\sigma)^2 \\
&\quad + \mathbb{E}\ln\cosh\left[\beta\left(\eta^1(n+\bar{m}) + \sqrt{\lambda\bar{p}}\tilde{\eta}\right)\right] + \frac{\lambda\beta\bar{q}}{2(1-\beta(1-\bar{q}))}.
\end{aligned} \tag{B4}$$

Appendix C

In this appendix we focus on the $t-$derivative of the interpolating statistical pressure for the three-layer HBM in clamped mode $N^{-1}\ln \mathbb{E}\mathcal{Z}_t^{clamp}$. In order to see how it goes, it is useful to notice that the exponent in the interpolating expression (27) can be seen as the sum of six terms of the form

$$f(t)A(X,s) \tag{C1}$$

where $f \in C^1((0,1))$ is a function of the interpolating parameter t, s stands for a generic configuration vector, X is the $(2NP+P+N)$-dimensional vector containing all the independent random variables involved (namely the synaptic weights $\{\eta^\mu\}_{\mu=1}^P$ and the external fictitious fields $\{\tilde{\eta}_i\}_{i=1}^N$, $\{\tilde{\eta}_\mu\}_{\mu=2}^P$) and $A(X,s)$ is an observable quantity.

It is easily checked that, performing the derivative of the logarithm function, the contribution provided by each term in the form (C1) to the $t-$derivative $\partial t \left(N^{-1}\mathbb{E}\ln \mathcal{Z}_t^{clamp}\right)$ appears as

$$\frac{1}{N} f'(t)\mathbb{E}\mathbf{\Omega}_t\, A(X,s). \tag{C2}$$

We proceed to the evaluation of these contributions indicating them with the symbols **I**, **II**, **III**, **IV**, **V**, **VI**, where the enumeration follows the order in which they appear as exponents in (27). Some of them (those for which $f(t)$ is a linear function of t) come from the Boolean part of the Hamiltonian, or from terms which do not involve random variables, and their calculation is completely trivial. Others (those for which $f(t)$ is a square root) involve Gaussian random variables and their calculation requires a Gaussian integration by parts. Specifically, we have:

(I) $f_1(t) \equiv t$, $A_1 \equiv \frac{\beta N}{2} m(\tau)^2 + \beta N n(\sigma) m(\tau)$.
This first term corresponds to a Curie–Weiss model with a particular external field, the relative contribution to the $t-$derivative of the interpolating pressure in the form (C2) is

$$\mathbf{I} = \frac{\beta}{2}\mathbb{E}\mathbf{\Omega}_t(m^2) + \beta n(\sigma)\mathbb{E}\mathbf{\Omega}_t(m).$$

(II) $f_2(t) \equiv \Phi N(1-t)$, $A_2 \equiv m(\tau)$.
This second term, "imitating" the first, contributes with

$$\mathbf{II} = -\Phi\mathbb{E}\mathbf{\Omega}_t(m).$$

(IV) $f_4(t) \equiv (1-t)\frac{\beta d}{2}$, $\quad A_4 \equiv \sum_{\mu=2}^P z_\mu^2$.
We thus get

$$\mathbf{IV} = -\frac{\beta d}{2N}\sum_{\mu=2}^P \mathbb{E}\mathbf{\Omega}_t(z_\mu^2).$$

Terms **III**, **V** and **VI** are the ones involving Gaussian random variables, they appear as

(III) $f_3(t) \equiv A\sqrt{1-t}$, $\quad A_3 \equiv \sum_i \tilde{\eta}_i \tau_i$;

(V) $f_5(t) \equiv \frac{\beta}{\sqrt{N}}\sqrt{t}$, $\quad A_5 \equiv N\sum_{\mu=2}^P m_\mu(\tau) z_\mu = \sum_{\mu=2}^P z_\mu \sum_i \eta_i^\mu \tau_i$;

(VI) $f_6(t) \equiv \beta B\sqrt{1-t}, \quad A_6 \equiv \sum_{\mu=2}^{P} \tilde{\eta}_\mu z_\mu;$

where the same symbol $\tilde{\eta}$ in A_3 and A_6 refers to two different independent families of i.i.d standard Gaussians. Focusing on **V**

$$\mathbf{V} = \frac{\beta}{2\sqrt{t}N\sqrt{N}} \mathbb{E}\boldsymbol{\Omega}_t \left(\sum_{\mu=2}^{P} z_\mu \sum_i \eta_i^\mu \tau_i \right),$$

and defining $F(X, \sigma, z, \tau) \stackrel{\text{def}}{=} \frac{B_t(X,\sigma,z,\tau)}{\mathcal{Z}_t(X,\sigma)}$ with B_t being the interpolating Boltzmann factor, we can write

$$\mathbf{V} = \frac{\beta}{2\sqrt{t}N\sqrt{N}} \mathbb{E}_{\eta^1} \sum_\tau \int \prod_{\mu=2}^{P} dz_\mu \sum_{\mu=2}^{P} \sum_i \mathbb{E}_{\tilde{\eta},\eta}(\eta_i^\mu F(X)) z_\mu \tau_i. \tag{C3}$$

The integration by parts of the average $\mathbb{E}_{\tilde{\eta},\eta}(\eta_i^\mu F(X))$ is now straightforward and gives

$$\mathbb{E}_{\tilde{\eta},\eta}(\xi_i^\mu F(X)) = \sum_l \mathbb{E}_{\tilde{\eta},\eta}(\eta_i^\mu X_l) \mathbb{E}_{\tilde{\eta},\eta} \left(\frac{\partial}{\partial X_l} F(X) \right) \tag{C4}$$

where the sum runs over all the indices defining the entries of X correspondent to Gaussian variables. Since all the noise variables X_l are completely independent, the only non-zero term in the right hand side of (C4) is the one provided by $X_l \equiv \eta_i^\mu$, for which $\mathbb{E}(\eta_i^\mu)^2 = 1$:

$$\mathbb{E}_{\tilde{\eta},\eta} \eta_i^\mu F(X) = \mathbb{E}_{\tilde{\eta},\eta} \frac{\partial}{\partial \eta_i^\mu} F.$$

The computation of the derivative of F with respect to the synaptic weight η_i^μ is trivial and gives

$$\frac{\partial}{\partial \eta_i^\mu} F = \frac{\beta\sqrt{t}}{\sqrt{N}} \left[\frac{\tau_i z_\mu B_t}{\mathcal{Z}_t} - \frac{B_t \sum_{\tau'} \int \prod_{\nu=2}^{P} dz'_\nu \tau'_i z'_\mu B'_t}{\mathcal{Z}_t^2} \right];$$

where B'_t is the Boltzmann factor correspondent to the "second replica configuration" (τ', z'). Plugging the latter equation into (C3), using $\tau_i^2 = 1$ and the definition of the overlaps (16) we get

$$\mathbf{V} = \frac{\beta^2}{2N} \mathbb{E} \sum_{\mu=2}^{P} \boldsymbol{\Omega}_t(z_\mu^2) - \frac{\beta^2\lambda}{2} \mathbb{E}\boldsymbol{\Omega}_t(p_{zz'} q_{\tau\tau'}).$$

Analogously, it results

$$\mathbf{III} = -\frac{A}{2N\sqrt{1-t}} \mathbb{E}\boldsymbol{\Omega}_t \left(\sum_i \tilde{\eta}_i \tau_i \right) = -\frac{A}{2N\sqrt{1-t}} \sum_i \sum_\tau \int \prod_\mu dz_\mu \mathbb{E} \left(\frac{\partial}{\partial \tilde{\eta}_i} F(X) \right) \tau_i$$
$$= -\frac{A^2}{2} + \frac{A^2}{2} \mathbb{E}\boldsymbol{\Omega}_t(q_{\tau\tau'});$$

$$\mathbf{VI} = -\frac{\beta B}{2N\sqrt{1-t}}\mathbb{E}\boldsymbol{\Omega}_t\left(\sum_\mu \tilde{\eta}_\mu z_\mu\right) = -\frac{\beta B}{2N\sqrt{1-t}}\sum_\mu\sum_\tau\int\prod_\mu dz_\mu \mathbb{E}\left(\frac{\partial}{\partial\tilde{\eta}_\mu}F(X)\right)z_\mu$$
$$= -\frac{\beta^2B^2}{2N}\sum_{\mu=2}^{P}\mathbb{E}\boldsymbol{\Omega}_t(z_\mu^2) + \frac{\beta^2B^2\lambda}{2}\mathbb{E}\boldsymbol{\Omega}_t(p_{zz'}).$$

Finally, defining the thermodynamic values of the order parameters as

$$\langle m\rangle_t = \bar{m}, \qquad \langle q_{\tau\tau'}\rangle_t = \bar{q}, \qquad \langle p_{zz'}\rangle_t = \bar{p},$$

where $\langle\cdot\rangle_t \equiv \mathbb{E}\boldsymbol{\Omega}_t(\cdot)$, we can fix the free parameters ϕ, d, B, A as

$$\phi = \beta(n(\sigma)+\bar{m}), \qquad d = \beta(1-\bar{q}), \qquad B = \sqrt{\bar{q}}, \qquad A = \beta\sqrt{\lambda\bar{p}}, \tag{C5}$$

and find expressions were second order fluctuations are explicit. Specifically, the choice for Φ makes

$$\mathbf{I}+\mathbf{II} = \frac{\beta}{2}\langle m^2 + 2mn(\sigma) - \frac{2}{\beta}\Phi m\rangle_t = \frac{\beta}{2}\langle(m-\bar{m})^2\rangle_t - \frac{\beta}{2}\bar{m}^2.$$

Similarly, since $d = \beta - \beta B^2$, the terms in $\mathbf{IV}+\mathbf{V}+\mathbf{VI}$ involving z_μ^2 vanish and we get

$$\mathbf{III}+\mathbf{IV}+\mathbf{V}+\mathbf{VI} = -\frac{\beta^2\lambda}{2}\langle p_{zz'}q_{\tau\tau'} + \bar{p}q_{\tau\tau'} + \bar{q}p_{zz'}\rangle_t - \frac{\beta^2\lambda}{2}\bar{p}$$
$$= -\frac{\beta^2\lambda}{2}\langle(p_{zz'}-\bar{p})(q_{\tau\tau'}-\bar{q})\rangle_t - \frac{\beta^2\lambda}{2}\bar{p}(1-\bar{q})$$

Since $\frac{\partial}{\partial t}\left(N^{-1}\ln\mathbb{E}\mathcal{Z}_t^{clamp}\right) = \mathbf{I}+\mathbf{II}+\mathbf{III}+\mathbf{IV}+\mathbf{V}+\mathbf{VI}$, it results

$$\frac{\partial}{\partial t}\frac{1}{N}\ln\mathbb{E}\mathcal{Z}_t^{clamp} = -\frac{\beta}{2}\bar{m}^2 - \frac{\beta^2\lambda}{2}\bar{p}(1-\bar{q}) + \frac{\beta}{2}\langle\,(m-\bar{m})^2\,\rangle_t - \frac{\beta^2\lambda}{2}$$
$$\times\langle\,(p_{zz'}-\bar{p})(q_{\tau\tau'}-\bar{q})\,\rangle_t. \tag{C6}$$

Appendix D

In this appendix we focus on the final expression for the statistical pressure of the three-layer HBM in clamped mode (34) and perform the derivatives $\partial_q\mathcal{A}_{RS}^{clamp}$, $\partial_p\mathcal{A}_{RS}^{clamp}$ and $\partial_m\mathcal{A}_{RS}^{clamp}$. The correspondent stationarity condition will let us write self-consistency equations for the order parameters m, p, q, dependent on the model's parameters $(\beta, \lambda, n, C) \in \mathbb{R}^+\times\mathbb{R}^+\times[-1,1]\times\mathbb{R}^+$. We get

$$\frac{\partial\mathcal{A}^B}{\partial m} = \mathbb{E}_{\eta^1}\frac{\partial}{\partial m}\int dG_1(\tilde{\eta})\ln\cosh\left[\beta(\eta^1(n+m)+\sqrt{\lambda p}\tilde{\eta})\right] - \beta m$$
$$= \beta\mathbb{E}_{\eta^1}\eta^1\int dG_1(\tilde{\eta})\tanh\left[\beta(\eta^1(n+m)+\sqrt{\lambda p}\tilde{\eta})\right] - \beta m. \tag{D1}$$

The stationarity condition is thus translated into

$$m = \mathbb{E}_{\eta^1}\eta^1 \int dG_1(\tilde{\eta}) \tanh\left[\beta(\eta^1(n+m) + \sqrt{\lambda p}\tilde{\eta})\right]$$
$$= \frac{1}{2}\left\{\int dG_1(\tilde{\eta}) \tanh\left[\beta(\sqrt{\lambda p}\tilde{\eta} + n + m)\right] - \int dG_1(\tilde{\eta}) \tanh\left[\beta(\sqrt{\lambda p}\tilde{\eta} - n - m)\right]\right\} \tag{D2}$$

$$\frac{\partial \mathcal{A}^B}{\partial p} = \mathbb{E}_{\eta^1}\frac{\partial}{\partial p}\mathbb{E}_{\tilde{\eta}} \ln\cosh\left[\beta\left(\eta^1(n+m) + \sqrt{\lambda p}\tilde{\eta}\right)\right] - \frac{\beta^2\lambda}{2}(1-q). \tag{D3}$$

We can now evaluate the Gaussian expectation through a standard integration by parts as follows

$$\frac{\partial}{\partial p}\mathbb{E}_{\tilde{\eta}} \ln\cosh\left[\beta\left(\eta^1(n+m) + \sqrt{\lambda p}\tilde{\eta}\right)\right] = \int_{\mathbb{R}} dG_1(\tilde{\eta})\frac{\partial}{\partial p}\ln\cosh$$
$$\times\left[\beta\eta^1(n+m) + \beta\sqrt{\lambda p}\tilde{\eta}\right]$$
$$= \frac{\beta\sqrt{\lambda}}{2\sqrt{p}}\int_{\mathbb{R}} dG_1(\tilde{\eta})\tilde{\eta}\tanh\left[\beta\eta^1(n+m) + \beta\sqrt{\lambda p}\tilde{\eta}\right]$$
$$= \frac{\beta^2\lambda}{2} - \frac{\beta^2\lambda}{2}\int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta\eta^1(n+m) + \beta\sqrt{\lambda p}\tilde{\eta}\right] \quad .$$

Plugging the latter into (D3) and forcing $\partial_p \mathcal{A}^B = 0$ we get

$$\frac{\partial \mathcal{A}^B}{\partial p} = \frac{\beta^2\lambda}{2} - \frac{\beta^2\lambda}{2}\mathbb{E}_{\eta^1}\int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta\eta^1(n+m) + \beta\sqrt{\lambda p}\tilde{\eta}\right] - \frac{\beta^2\lambda}{2}(1-q) = 0$$

which corresponds to

$$q = \mathbb{E}_{\eta^1}\int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta\eta^1(n+m) + \beta\sqrt{\lambda p}\tilde{\eta}\right]$$
$$= \frac{1}{2}\left\{\int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta(\sqrt{\lambda p}\tilde{\eta} + n + m)\right] + \int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta(\sqrt{\lambda p}\tilde{\eta} - n - m)\right]\right\}. \tag{D4}$$

The stationarity condition given by the statistical pressure's derivative with respect to q brings, introducing the C parameter (33), to the following

$$\frac{\beta^2\lambda}{2}p = \frac{\lambda\beta(1-\beta(1-q)) + \beta^2\lambda C - \lambda\beta(1-\beta)}{2(1-\beta(1-q))^2}$$
$$p = \frac{q+C}{(1-\beta(1-q))^2}. \tag{D5}$$

Finally, we obtain the replica symmetric self-consistency equations for the HBM in clamped mode as

$$q = \mathbb{E}_{\eta^1}\int_{\mathbb{R}} dG_1(\tilde{\eta})\tanh^2\left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m))\right], \tag{D6a}$$

$$m = \mathbb{E}_{\eta^1}\eta^1\int dG_1(\tilde{\eta})\tanh\left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m))\right], \tag{D6b}$$

$$p = \frac{q+C}{(1-\beta(1-q))^2}. \tag{D6c}$$

Appendix E

In this appendix we find an expression for the zero temperature limit $\beta \to \infty$ of the self-consistency equations (35a), (35b), (35c), correspondent to the equilibrium conditions for the three-layer HBM in clamped mode. From the properties of the hyperbolic tangent (which is bounded by 1 and converges a.e. to the sign function as its argument grows to infinity) the limit $\beta \to \infty$ of the right hand side in equation (35b) can be easily performed through dominated convergence and yields to

$$m = \text{erf}\left(\frac{n+m}{\sqrt{2p\lambda}}\right), \tag{E1}$$

where the error function $\text{erf}(x)$ is $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, and where we supposed $p \neq 0$. This is achieved through the following

$$\begin{aligned} m &= \mathbb{E}_{\eta^1} \eta^1 \int dG_1(\tilde{\eta})\ \text{sign}\left[\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m)\right] \\ &= \mathbb{E}_{\eta^1} \eta^1 \sqrt{\frac{2}{\pi}} \int_0^{\frac{\eta^1(n+m)}{\sqrt{\lambda p}}} e^{\frac{-\tilde{\eta}^2}{2}} d\tilde{\eta} = \mathbb{E}_{\eta^1} \eta^1\ \text{erf}\left(\frac{\eta^1(n+m)}{\sqrt{2p\lambda}}\right) = \text{erf}\left(\frac{n+m}{\sqrt{2p\lambda}}\right). \end{aligned}$$

Obviously $\text{sign}^2(x) = 1\ \forall x \in \mathbb{R}$ so, since the hyperbolic tangent there appears squared, equation (35a) suggests that q tends to 1 for $\beta \to \infty$. We thus need to evaluate the zero temperature limit of the term $\beta(1-q)$, which appears in the denominator of (35c) . Through the equation for q we get

$$\begin{aligned} \beta(1-q) &= \beta(1 - \mathbb{E}_{\eta^1} \int_{\mathbb{R}} dG_1(\tilde{\eta}) \tanh^2\left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m))\right] \\ &= \frac{1}{\sqrt{\lambda p}} \int dG_1(\tilde{\eta}) \frac{\partial}{\partial \tilde{\eta}} \tanh\left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m))\right]. \end{aligned} \tag{E2}$$

Integrating by parts we get

$$= \frac{1}{\sqrt{\lambda p}} \int dG_1(\tilde{\eta})\ \tilde{\eta}\ \tanh\left[\beta(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m))\right],$$

which in the limit $\beta \to \infty$ gives

$$\begin{aligned} &\frac{1}{\sqrt{\lambda p}} \int dG_1(\tilde{\eta})\ \tilde{\eta}\ \text{sign}(\sqrt{\lambda p}\tilde{\eta} + \eta^1(n+m)) \\ &= \frac{1}{\sqrt{\lambda p}} \left(\int_{-\frac{\eta^1(n+m)}{\sqrt{\lambda p}}}^{+\infty} \tilde{\eta} dG_1(\tilde{\eta}) - \int_{-\infty}^{-\frac{\eta^1(n+m)}{\sqrt{\lambda p}}} \tilde{\eta} dG_1(\tilde{\eta}) \right) \\ &= \frac{2}{\sqrt{\lambda p}} \int_{\frac{\eta^1(n+m)}{\sqrt{\lambda p}}}^{+\infty} \tilde{\eta} dG_1(\tilde{\eta}) = \sqrt{\frac{2}{\lambda p \pi}} e^{-\frac{(\eta^1(n+m))^2}{2\lambda p}} = \sqrt{\frac{2}{\lambda p \pi}} e^{-\frac{(n+m)^2}{2\lambda p}}, \end{aligned}$$

where we used $(\eta^1)^2 = 1$ and the fact that $f(x) = x e^{-\frac{x^2}{2}}$ is odd. In the limit $\beta \to \infty$ Eq. (35c) thus corresponds to

$$p = \frac{(1+C)}{\left(1-\sqrt{\frac{2}{\lambda p\pi}}e^{-\frac{(n+m)^2}{2\lambda p}}\right)^2} = \frac{\lambda p\pi}{2}\frac{(1+C)}{\left(\sqrt{\frac{\lambda p\pi}{2}}-e^{-\frac{(n+m)^2}{2\lambda p}}\right)^2};$$

$$p\left(\sqrt{\frac{\lambda p\pi}{2}}-e^{-\frac{(n+m)^2}{2\lambda p}}\right)^2 = \frac{\lambda p\pi}{2}(1+C).$$

Excluding the zero solution for p we can multiply for $\frac{4}{\pi p}$ and obtain that

$$\left(\sqrt{2\lambda p}-\frac{2}{\sqrt{\pi}}e^{-\frac{(n+m)^2}{2\lambda p}}\right)^2 = 2\lambda(1+C).$$

Taking the square root we obtain an expression for $\sqrt{2\lambda p}$ as

$$\sqrt{2\lambda p} = \frac{2}{\sqrt{\pi}}e^{-\frac{(n+m)^2}{2\lambda p}} + \sqrt{2\lambda(1+C)}. \tag{E3}$$

Equations (E1), (E3) can be reduced to one equation in the variable $y = \frac{n+m}{\sqrt{2\lambda p}}$, they indeed appear as

$$\begin{cases} y = \frac{\mathrm{erf}(y)+n}{\sqrt{2\lambda p}} \\ \sqrt{2\lambda p} = \frac{2}{\sqrt{\pi}}e^{-y^2} + \sqrt{2\lambda(1+C)} \end{cases} \tag{E4}$$

which directly corresponds to

$$y = \frac{\mathrm{erf}(y)+n}{\frac{2}{\sqrt{\pi}}e^{-y^2} + \sqrt{2\lambda(1+C)}}. \tag{E5}$$

References

1. Coolen, A.C.C., Kuehn, R., Sollich, P.: Theory of Neural Information Processing Systems. Oxford Press, Oxford (2005)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Nature, London (2006)
3. Agliari, E., Barra, A., Tirozzi, B.: Free energies of Boltzmann Machines: self-averaging properties, annealed and replica symmetric approximations in the thermodynamic limit. J. Stat. Mech. **2019**, 033301 (2019)
4. Barra, A., Genovese, G., Sollich, P., Tantari, D.: Phase diagram of restricted Boltzmann machines & generalized Hopfield models. Phys. Rev. E **97**, 022310 (2018)
5. Tubiana, J., Monasson, R.: Emergence of compositional representations in restricted Boltzmann. Phys. Rev. Lett. **118**, 138301 (2017)
6. Alberici, D., Contucci, P., Mingione, E.: Deep Boltzmann machines: rigorous results at arbitrary depth. Ann. Henri Poincaré **22**, 2619–2642 (2021)
7. Decelle, A., Furtlehner, C.: Restricted Boltzmann machine: recent advances and mean-field theory. Chin. Phys. B **30**(4), 040202 (2021)
8. Barra, A., Bernacchia, A., Santucci, E., Contucci, P.: On the equivalence of Hopfield networks and Boltzmann machines. Neur. Netw. **34**, 1–9 (2012)
9. Fischer, A., Igel, C.: Training restricted Boltzmann machines: an introduction. Pattern Recogn. **47**, 25–39 (2014)
10. Agliari, E., Barra, A., Galluzzi, A., Moauro, F., Guerra, F.: Multitasking associative networks. Phys. Rev. Lett. **109**, 268101 (2012)
11. Marullo, C., Agliari, E.: Boltzmann machines as generalized Hopfield networks: a review on recent results and outlooks. Entropy **23**(1), 34 (2020)
12. Bovier, A.: Principles of statistical mechanics. In Statistical Mechanics of Disordered Systems: A Mathematical Perspective Cambridge Series in Statistical and Probabilistic Mathematics, pp. 9–32 (2006)

13. Leonelli, F.E., Agliari, E., Albanese, L., Barra, A.: On the effective initialisation for restricted Boltzmann machines via duality with Hopfield model. Neur. Netw. **143**, 314 (2021)
14. Fachechi, A., Agliari, E., Alemanno, F., Barra, A.: Dreaming Boltzmann machines outperform standard ones, submitted (2021)
15. Agliari, E., Migliozzi, D., Tantari, D.: Non-convex multi-species Hopeld models. J. Stat. Phys. **172**, 1247–1269 (2018)
16. Agliari, E., Alemanno, F., Barra, A., Fachechi, A.: Dreaming neural networks: rigorous results. J. Stat. **2019**, 083503 (2019)
17. Smart, M., Zilman, A.: On the mapping between Hopfiled networks and restricted Boltzmann machines. In: ICLR Conference Paper (2021)
18. Agliari, E., Albanese, L., Alemanno, F., Fachechi, A.: A transport equation approach for deep neural networks, submitted (2021)
19. Agliari, E., Leonelli, F.E., Marullo, C.: Retrieval capabilities of neural networks with biased patterns. Appl. Math. Comput (2021)
20. Agliari, E., Alemanno, F., Barra, A., Fachechi, A.: Generalized Guerras interpolation schemes for dense associative neural networks. Neur. Netw. **128**, 254–267 (2020)
21. Genovese, G.: Universality in bipartite mean field spin glasses. J. Math. Phys. **53**(12), 123304 (2012)
22. Agliari, E., Barra, A., Longo, C., Tantari, D.: Neural networks retrieving Boolean patterns in a sea of Gaussian ones. J. Stat. Phys. **68**, 1085–1104 (2017)
23. Carmona, P., Hu, Y.: Universality in Sherrington–Kirkpatricks Spin glass model. Ann. Inst. Henri Poincaré (B) **42**, 215–225 (2006)
24. Guerra, F.: Sum rules for the free energy in the mean field spin glass model. Fields Inst. Comm. **30**, 161 (2001)
25. Amit, D.J., Gutfreund, H., Sompolinsky, H.: Storing infinite numbers of patterns in a spin glass model of neural networks Phys. Rev. Lett. **55**, 1530–1533 (1985)
26. Amit, D.J.: Modeling Brain Functions. Cambridge University Press, Cambridge (1989)
27. Agliari, E., De Marzo, G.: Tolerance versus synaptic noise in dense associative memories. Eur. Phys. J. Plus **135**, 883 (2020)
28. Agliari, E., Alemanno, F., De Marzo, G., Barra, A.: The emergence of a concept in shallow neural networks. arXiv:2109.00454