Free energies of Boltzmann Machines: self-averaging, annealed and replica symmetric approximations in the thermodynamic limit

Elena Agliari,[1] Adriano Barra,[2] and Brunello Tirozzi[3]

[1]*Dipartimento di Matematica, Sapienza Università di Roma, Italy*
GNFM-INdAM Sezione di Roma, Italy
[2]*Dipartimento di Matematica e Fisica, Università del Salento, Lecce, Italy*
GNFM-INdAM Sezione di Roma, Italy
INFN Sezione di Lecce, Italy
[3]*Dipartimento di Fisica, Sapienza Università di Roma, Italy*
Enea Research Center, Frascati, Italy

Restricted Boltzmann machines (RBMs) constitute one of the main models for machine statistical inference and they are widely employed in Artificial Intelligence as powerful tools for (deep) learning. However, in contrast with countless remarkable practical successes, their mathematical formalization has been largely elusive: from a statistical-mechanics perspective these systems display the same (random) Gibbs measure of bi-partite spin-glasses, whose rigorous treatment is notoriously difficult.
In this work, beyond providing a brief review on RBMs from both the learning and the retrieval perspectives, we aim to contribute to their analytical investigation, by considering two distinct realizations of their weights (i.e., Boolean and Gaussian) and studying the properties of their related free energies. More precisely, focusing on a RBM characterized by digital couplings, we first extend the Pastur-Shcherbina-Tirozzi method (originally developed for the Hopfield model) to prove the self-averaging property for the free energy, over its quenched expectation, in the infinite volume limit, then we explicitly calculate its simplest approximation, namely its annealed bound.
Next, focusing on a RBM characterized by analogical weights, we extend Guerra's interpolating scheme to obtain a control of the quenched free-energy under the assumption of replica symmetry (i.e., we require that the order parameters do not fluctuate in the thermodynamic limit): we get self-consistencies for the order parameters (in full agreement with the existing Literature) as well as the critical line for ergodicity breaking that turns out to be the same obtained in AGS theory. As we discuss, this analogy stems from the slow-noise universality.
Finally, glancing beyond replica symmetry, we analyze the fluctuations of the overlaps for a correct estimation of the (slow) noise affecting the retrieval of the signal, and by a stability analysis we recover the Aizenman-Contucci identities typical of glassy systems.

I. INTRODUCTION: BOLTZMANN MACHINES IN A NUTSHELL.

Boltzmann machines (BMs) play a key role in Artificial Intelligence: being able to learn internal representations (when fed by external data) and to solve difficult combinatoric problems (when suitably trained), they can be efficiently employed in machine learning and machine statistical-inference. Also, extensions of the BMs, i.e., the so-called Deep Boltzmann machines [50, 56], allow for Deep Learning, that is the novel generation of Artificial Intelligence. Their applications are broadly ranged in Science (from Particle Physics [8] to Computational Biology [44]), not to mention the applied world of technology, where their usage has become pervasive [42, 45]. Further, several models of memory formation and pattern recognition in Theoretical Immunology (regarding the adaptive branch of the immune system of mammals) [3, 4, 17, 18, 48, 49] are naturally framed in the mathematical scaffold of BMs.

In a nutshell, a BM is a network of symmetrically connected units (also called neurons or spins) divided into two or more layers. Here we shall focus on *restricted* Boltzmann machines (RBMs), made of two layers called "visible" and "hidden", respectively; the units are connected to each other across layers, but there is no intra-layer communication (this is the *restriction* in a RBM, see Fig. 1, left panel).
As anticipated, RBMs can be used to solve quite different computational problems [26, 27, 37, 38, 41, 55]. In order to achieve this, the machine has first to undergo a training process where its parameters, namely the thresholds θ for neuron firing (i.e., for spin flip in the binary case [7]) and the weights ξ associated to links (i.e., the synaptic values), are stochastically tuned according to proper algorithms (e.g., contrastive divergence [56] and simulated annealing [40]). After this process, one can initialise the visible layer in a given state (input) and make the system evolve towards its ground state[**?**]; the latter (output) provides the solution of the problem [36].
Let us explain this concept more extensively. During the training stage, a set of noisy data (assumed to be independently and identically generated by a given probability distribution) is shown to the visible layer of the RBM. This is obtained by encoding the data in terms of state configurations for the visible layer (e.g., data can be constituted by pictures, where each pixel, either black or white, is associated to the state, either $+1$ or -1, of the related binary visible unit). For each item presented to the visible layer, the system relaxes to the pertinent equilibrium distribution and, once a stable thermalization is reached, an update in the machine parameters (ξ, θ) plays as a *learning rule* in such a way that the Kullback-Leibler divergence between the

internal probabilistic representation of the data (given by the equilibrium distribution of the system) and the external distribution generating these data is minimized. In this process the hidden units, whose states are not specified by the data provided, act as latent variables and allow the RBM to store information about the learnt data. Each hidden unit turns out to be associated to a "feature" (e.g., the extent of correlations among the entries of input data) and the set of all the features is supposed to statistically characterise the data provided [43]. Actually, the nature of such features emerges during the training without control from the user. After training, if a new item (from the same distribution and possibly noisy) is shown, the machine will be able to "recognize" its statistical structure. For instance, if trained on pictures, the machine will theoretically model the distribution of pictures and can then be used for reconstruction tasks: given a partial picture as input, the machine will use the internal model to provide, as output, the complete picture. We refer to Sec. II for a more technical explanation.

Despite the success in practical applications, a rigorous control of RBMs remains a hard problem from the mathematical perspective. It is worth recalling that, in the mathematical investigations we are interested in, weights on links are assumed as quenched and extracted according to a given distribution which is meant as the result of some training process. Thus, the RBMs we are dealing with can be classified as two-party spin-glasses [10, 14, 32, 47, 51]. As a consequence, an extensive use of techniques and tools from the Statistical Mechanics of Disordered Systems is in order and, quoting Talagrand, this is still a *challenge for mathematicians* [59]. Remarkably, reaching a full, rigorous description for these models is further hindered by the fact that, due to the wide range of their applications, there exists a number of variations on theme [42] and their differences are subtle but important. These differences mainly affect the nature of links (e.g., Gaussian or Boolean) and the nature of the units (e.g., Gaussian or Boolean) of the hidden layer [9].

As a matter of fact, only a few rigorous results on this subject are available (see e.g., [23, 45]) and there is a urgent and significant gap to fill, as broadly recognized (see e.g., [23, 42]). A possible way to pave towards rigorous advances has recently been investigated [2, 5, 13, 25, 46, 64, 65] and it is based on the thermodynamic equivalence of RBMs and Hopfield neural networks [11]. The latter constitute the standard model for associative memory [35], for which several mathematical tools have been made available along the past two decades (see e.g., [15, 19, 20, 52, 53, 60, 61]). In the large volume limit, this bridge allows extending several rigorous mathematical approaches, originally developed within the framework of spin-glasses and associative neural networks, to machine learning as well. In particular, in this paper, we will focus on two variations on this theme and address their properties with two different techniques as summarized in the following:

- *RBMs with Boolean links*.
  This is a bi-partite spin-glass where the two parties (i.e., the two layers) contain variables of different nature: the visible layer is made of binary Ising spins, while the hidden layer is made of real Gaussian spins. The weights on links connecting the units belonging to different layers are binary (± 1) and extracted independently and identically with equal probability. For this model we adapt the Pastur-Shcherbina-Tirozzi scheme, originally developed for the standard Hopfield model [53, 57]. This procedure allows us to prove the self-averaging of the main observable, namely the free-energy of the model, over its quenched expectation. Next, we provide its first approximation, namely its annealed bound, which will be achieved via Jensen inequality.
  The underlying idea in this route is to construct a *matryoshka* of σ-algebras generated by the random weights associated to links and to define conditional expectations over them. Next, we show that the sequence of conditional expectations of the extensive free-energy constitutes a martingale and, exploiting its properties (mainly Doob Theorem [22]), we can bound fluctuations in the intensive free-energy and prove that it converges to its quenched expectation.

- *RBMs with Gaussian links*.
  This is a bi-partite spin-glass where the two parties (i.e., the two layers) contain variables of different nature: the visible layer is made of binary Ising spins, while the hidden layer is made of real Gaussian spins. The weights on links connecting the units belonging to different layers are real valued and sampled identically and independently from a Gaussian distribution $\mathcal{N}[0,1]$.
  In order to get an explicit expression of the infinite volume limit of the quenched free-energy for these machines we adapt an interpolation scheme, originally developed by Guerra for the Sherrington-Kirkpatrick spin-glass [12]. Here the underlying idea is to compare the original network, where the two parties interact extensively, with an effective model where the two parties are no longer interacting, rather, they experience random external fields, whose statistical distributions mirror the real ones, namely each neuron feels an external field which mimics the internal field generated by neurons it is connected with. This makes the calculations feasible because the effective model is one-body thus, in principle, integrable.
  As standard practice in neural networks [7, 24], we assume that, in the infinite volume limit, the order parameters do not fluctuate (their distributions get δ-peaked over their thermodynamic values), namely we will assume replica symmetry.
  Since the slow noise generated by the not-retrieved patterns is *universal* (namely, in the thermodynamic limit, the replica-symmetric expression for the noises stemming from pattern entries sampled from a Bernoullian distribution and from a Gaussian one coincide), the results of this section naturally extend to the machine with digital links investigated in the previous one. This remark prompts to a stability analysis, where we investigate overlap fluctuations (hence glancing beyond replica symmetry) and we find, as standard in glassy systems, that these fluctuations obey the (suitably adapted) Aizenman-Contucci constraints.

Summary of the paper: In the next section we present our perspective on the way RBMs actually work, while we leave the next Secs. III-IV to novel and rigorous results, Sec. V for deepening these results from an applied perspective and, finally, the last section contains our conclusions. Particularly long proofs (those of Theorem 1, Proposition 2, Theorem 3 and Corollary 1) are left in the Appendices to lighten the exposition, while all the other proofs are brief enough to be included in the main text. Just for the next section we will relax the mathematical rigour in order to allow for a minimal but fluent presentation and provide a streamlined review.

II. A QUICK OVERVIEW ON RESTRICTED BOLTZMANN MACHINES

As anticipated in the previous section, a RBM is a two-party system, where one party (or *layer* to keep the original jargon [31, 56]), referred to as visible, receives input data from the outside world, while the other party (or layer), referred to as hidden, is dedicated to figure out correlations in these input data. Typically, a set of $M > 0$ data vectors $\{\sigma_1, \sigma_2, ..., \sigma_M\}$ (i.e., the so-called *training set* [31]) is presented to the machine and, under the assumption that these data have been generated by the same probability distribution $Q(\sigma)$, the ultimate goal of the machine is to make an inner representation of $Q(\sigma)$, say $P(\sigma|\xi, \theta)$, that is as close as possible to the original one, i.e., it has to reconstruct the signal. Clearly, in order to get a good representation, the more complicated $Q(\sigma)$, the larger the training set[**?**].
Each layer is composed by spins (also called neurons in this context), N for the visible layer and P for the hidden layer, and these spins can be chosen with high generality, ranging from discrete-valued (e.g., Ising spins), to real-valued (e.g., Gaussian spins). The thermodynamic limit of the ratio between the layer sizes, denoted as $\alpha = \lim_{N\to\infty} P/N$, is a control parameter and usually one splits the case $\alpha = 0$ (possibly yielding to under-fitting) and the case $\alpha \in \mathbb{R}^+$ (possibly yielding to over-fitting) [31], the latter being mathematically much more challenging.
Analogously, the entries of the weight matrix can be either real or discrete. Generally speaking, continuous weights allows for learning rules (e.g., the contrastive divergence involving weight derivatives) which are more powerful than their discrete counterparts (the typical learning rule for binary weights is the Hebbian one [7]) and are therefore more convenient during the learning stage; on the other hand, binary weights are more performing in the so-called retrieval phase, that is, once the machine has learnt and is ready to perform the task it has been trained for. This trade-off gave rise to a number of variations on theme within the world of RBMs, where the extremal cases are probably given by a machine with binary (i.e., Boolean) versus real (i.e., Gaussian) weights, equipped with a binary visible layer and a real hidden layer: in the present work we will focus on both these cases and we will try to highlight equivalences (but also crucial differences) among these extrema.

Before presenting in details these cases, it is useful to summarize the mechanisms underlying the functioning of a standard RBM and, to this aim, we now introduce its effective Hamiltonian (or "cost function" in a machine learning jargon) as

$$H_N(\sigma, z|\xi, \theta) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} \xi_i^\mu \sigma_i z_\mu - \sum_{i=1}^{N} \theta_i \sigma_i, \tag{1}$$

where σ_i ($i \in [1, ..., N]$) denotes the state of the i-th visible unit, z_μ ($\mu \in [1, ..., P]$) denotes the state of the μ-th hidden unit, ξ_i^μ denotes the weight associated to the link connecting the neurons labelled i and μ, and the factor $1/\sqrt{N}$ ensures the linear extensivity of the Hamiltonian with respect to the system volume [16]. The scalars θ_i ($i \in [1, ..., N]$) can be interpreted as external fields acting on the visible units and provide thresholds for neuron firing: given a certain internal field $\sum_\mu \xi_i^\mu z_\mu/\sqrt{N}$ over σ_i, the larger θ_i and the more likely for the i-th neuron to fire, namely to be in an active state $\sigma_i = +1$.
Now, this system is made to evolve by applying algorithms mimicking cognitive processes [1, 35]. More precisely, one splits *cognition* into two separate acts, namely distinguishing between *learning* (information) and *retrieval* (of the learnt information). The former occurs on a slower time scale and implies a synaptic dynamics which is modeled by properly rearranging the set of weights and thresholds. The latter occurs on a faster time scale and implies a neuronal dynamics which is modeled by properly rearranging the spin configuration, while keeping the weights quenched. Given the gap between the time scales characterizing these dynamical processes[**?**], one can treat them adiabatically, as done in the following subsections: the next one is dedicated to synaptic dynamics (i.e., rearrangement of the weights), while the successive one to neural dynamics (i.e., rearrangement of the spins).

A. A brief digression on slow variable's dynamics: learning

In this subsection we focus on the algorithms underlying the learning stage and which imply the dynamic of weights (we refer to [24] for a more extensive treatment). As mentioned in the beginning of Sec. II, the goal is to obtain an inner representation $P(\sigma|\xi, \theta)$ which approximates $Q(\sigma)$; this is usually achieved by the minimization of the Kullback-Leibler cross entropy $D(Q, P)$,

defined as

$$D(Q,P) \doteq \sum_{\sigma} Q(\sigma) \ln \left[\frac{Q(\sigma)}{P(\sigma)} \right], \tag{2}$$

where the sum runs over all the possible configurations of the visible layer and we have dropped the dependence on the parameters (ξ,θ) of $P(\sigma|\xi,\theta)$ to lighten the notation. To the same purpose we also introduce $\tilde{\xi}_i^{\mu} \doteq \xi_i^{\mu}/\sqrt{N}$, $\forall i,\mu$. Notice that $D(Q,P)$ is minimal (and equal to zero) if and only if $P(\sigma)$ and $Q(\sigma)$ are identical. Now, by updating the weights and the thresholds by a gradient descent rule

$$\Delta\tilde{\xi}_i^{\mu} = -\varepsilon \frac{\partial D(Q,P)}{\partial \tilde{\xi}_i^{\mu}}, \tag{3}$$

$$\Delta\theta_i = -\varepsilon \frac{\partial D(Q,P)}{\partial \theta_i}, \tag{4}$$

where ε is a small parameter (also called learning rate), we get

$$\begin{aligned} \Delta D(Q,P) &= \sum_{i,\mu} \frac{\partial D(Q,P)}{\partial \tilde{\xi}_i^{\mu}} \Delta\tilde{\xi}_i^{\mu} + \sum_i \frac{\partial D(Q,P)}{\partial \theta_i} \Delta\theta_i \\ &= -\varepsilon \left[\sum_{i,\mu} \left(\frac{\partial D(Q,P)}{\partial \tilde{\xi}_i^{\mu}} \right)^2 + \sum_i \left(\frac{\partial D(Q,P)}{\partial \theta_i} \right)^2 \right] \leq 0, \end{aligned} \tag{5}$$

that is the cross-entropy $D(Q,P)$ decreases monotonically until a stationary state is reached (which, still, does not necessarily correspond to $D(Q,P)=0$). Now, in order to make this learning rule an explicit, operational, algorithm a bit of work is still necessary. A key point is that weights in the BM are symmetric (i.e., they are undirected) and this, for (non-pathologic) stochastic dynamics, implies *detailed balance* which, in turn, ensures that the invariant measure is the Gibbs one given by

$$P(\sigma,z) = \frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\xi,\theta)}, \tag{6}$$

where $Z_{P,N}(\beta|\xi,\theta)$ is a normalization factor (or "partition function" in a Statistical Mechanics jargon [7, 24]) and $\beta \in \mathbb{R}^+$ encodes for the noise intrinsically present in real data sets (in Physics β plays as an inverse temperature, in proper units). Now, marginalizing $P(\sigma,z)$ over the hidden layer z, we get $P(\sigma)$. Therefore, the internal representation of the probability distribution is formally known and this allows the construction of explicit learning algorithms, among which the *contrastive divergence* that we are going to derive is probably the most applied [31]. In order to proceed with the construction of a learning algorithm we explicitly define

$$Z_{P,N}(\beta|\xi,\theta) = \int_{-\infty}^{+\infty} \prod_{\mu=1}^{P} d\mu(z_\mu) \sum_{\sigma} e^{-\beta H_N(\sigma,z|\xi,\theta)}, \tag{7}$$

$$Z_{P,N}(\beta|\sigma,\xi,\theta) = \int_{-\infty}^{+\infty} \prod_{\mu=1}^{P} d\mu(z_\mu) e^{-\beta H_N(\sigma,z|\xi,\theta)}, \tag{8}$$

$$\begin{aligned} P(\sigma) &= \int_{-\infty}^{+\infty} \prod_{\mu=1}^{P} d\mu(z_\mu) P(\sigma,z) = \int_{-\infty}^{+\infty} \prod_{\mu=1}^{P} d\mu(z_\mu) \frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\xi,\theta)} \\ &= \frac{Z_{P,N}(\beta|\sigma,\xi,\theta)}{Z_{P,N}(\beta|\xi,\theta)}, \\ P(z|\sigma) &= \frac{P(\sigma,z)}{P(\sigma)} = \frac{e^{-\beta H_N(\sigma,z|\xi,\theta)}}{Z_{P,N}(\beta|\sigma,\xi,\theta)}, \end{aligned} \tag{9}$$

where, summations are meant over all possible spin configurations and $d\mu(z_\mu)$ is the Gaussian measure ($d\mu(z_\mu) = \exp(-z_\mu^2\beta/2)\sqrt{\beta/(2\pi)}$, for $\mu = 1,...,P$). Thus, $Z_{P,N}(\beta|\xi,\theta)$ is the partition function of a system where both variable sets are free to evolve, while $Z_{P,N}(\beta|\sigma,\xi,\theta)$ is the partition function of a system where the visible layer is "clamped", namely forced to be in the configuration $\{\sigma\}$ encoded by one of the input data. Also, $P(\sigma)$ is the marginalized distribution and $P(z|\sigma)$ is the distribution for the configuration of the hidden layer being the visible layer clamped. At this point we have to evaluate each

single term inside (5):

$$\begin{aligned}
\frac{\partial D(Q,P)}{\partial \tilde{\xi}_i^\mu} &= -\sum_\sigma Q(\sigma)\frac{\partial \ln P(\sigma)}{\partial \tilde{\xi}_i^\mu} \\
&= -\sum_\sigma Q(\sigma)\frac{\partial}{\partial \tilde{\xi}_i^\mu}\left(\ln Z_{P,N}(\beta|\sigma,\xi,\theta) - \ln Z_{P,N}(\beta|\xi,\theta)\right) \\
&= \beta\sum_\sigma Q(\sigma)\left(\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z_\mu)P(z|\sigma,\xi,\theta)\frac{\partial H_N(\sigma,z|\xi,\theta)}{\partial \tilde{\xi}_i^\mu}\right. \\
&\quad \left. -\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z'_\mu)\sum_{\sigma'}P(z',\sigma'|\xi,\theta)\frac{\partial H_N(\sigma',z'|\xi,\theta)}{\partial \tilde{\xi}_i^\mu}\right) \\
&= -\beta\left(\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z_\mu)\sum_\sigma Q(\sigma)P(z|\sigma,\xi,\theta)\sigma_i z_\mu\right. \\
&\quad \left. -\sum_\sigma Q(\sigma)\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P} d\mu(z'_\mu)\sum_{\sigma'}P(z',\sigma'|\xi,\theta)\sigma'_i z'_\mu\right), \\
&= -\beta\left(\langle\sigma_i z_\mu\rangle_{clamped} - \langle\sigma_i z_\mu\rangle_{free}\right), \qquad (10)
\end{aligned}$$

where, in the first passage we used the definition (2), recalling that $Q(\sigma)$ does not depend on ξ; in the second passage we used (9); in the third passage we used (7) and (8); in the fourth passage we recalled that $\partial H_N(\sigma,z|\xi)/\partial\tilde{\xi}_i^\mu = -\sigma_i z_\mu$ and the subscript *clamped* means that the averages of the two-points correlation functions must be evaluated when the visible layer is forced to assume data values, while *free* means that the averages are the standard, statistical-mechanical ones. For the updating rule of the thresholds $\theta_i (i = 1,...,N)$, one performs analogous calculations and, recalling $\partial H_N(\sigma,z|\xi)/\partial\theta_i = -\sigma_i$, one gets

$$\frac{\partial D(Q,P)}{\partial \theta_i} = -\beta\left(\langle\sigma_i\rangle_{clamped} - \langle\sigma_i\rangle_{free}\right). \qquad (11)$$

Thus, the learning rule (5) ultimately tries to make the theoretical one-point and two-point correlation functions as close as possible to the empirical ones[**?**]. Under this rule the machine will eventually be able to reproduce the statistics of the training data correctly, and this means that the parameters (ξ,θ) have been rearranged such that, if the machine is now asked to generate vectors with $P(\sigma)$, the statistical properties of these vectors will coincide with those of the input data generated by $Q(\sigma)$. In this case we say that the machine *has learnt* a representation of the reality it has been fed with. Note that this approach allows a proper statistical reproduction of mean averages and variances, hence, when $Q(\sigma)$ violates the central limit theorem, a two-layer RBM is no longer suitable for statistical inference.

B. A brief digression on fast variable's dynamics: retrieval

After the learning stage, the machine undergoes a final check over another bulk of data, referred to as *test set*, which stems from the same distribution that has generated the training set [31]. To fix ideas, let us assume that the machine was trained for retrieval tasks[**?**]; if the trained machine is able to retrieve correctly the items in the test set, then the test is passed and the machine is ready for the usage[**?**]. In order to move from the learning mode to the retrieval mode, the hidden layer is marginalized over: as we are going to show, following this procedure we end up with a Hopfield model (that is the standard model for pattern retrieval [35]), where each feature learnt by the hidden layer corresponds to one of the learnt patterns and the optimal parameters (ξ,θ) store information about the whole set of learnt patterns [2, 9].

To see this duality between the RBM and the Hopfield model we look at the temporal evolution of the neurons which can be described by the following stochastic differential equation and map (to fix ideas we take hidden units as continuous and visible units as binary, as before)

$$\frac{dz_\mu(t)}{dt} = -z_\mu(t) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i^\mu\sigma_i + \sqrt{\frac{2}{\beta}}\eta_\mu(t), \qquad (12)$$

$$\sigma_i(t) = \text{sign}\left[\tanh\left(\frac{\beta}{\sqrt{N}}\sum_{\mu=1}^{P}\xi_i^\mu z_\mu + \beta\theta_i\right) + \tilde{\eta}_i(t)\right]. \qquad (13)$$

In the previous equation we used t to denote the time and we set the typical timescale of the variables (σ,z) as unitary; also, we denoted with η, $\tilde{\eta}$ standard Gaussian white noises with zero mean and covariance $\langle\eta_\mu(t)\eta_\nu(t')\rangle = \delta_{\mu\nu}\delta(t-t')$. Notice that, in

the temporal evolution of the visible (respectively hidden) units, the hidden (respectively visible) units are taken as fixed (see also [24]).
Let us now focus on the hidden layer dynamics: the first term in the right-hand side of eq. (12) is the standard leakage term and the second term is the input signal over the hidden layer. This dynamics overall defines an Ornstein-Uhlembeck process, whose equilibrium distribution, at fixed σ's, reads as

$$P(z_\mu|\sigma) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}\left(z_\mu - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i^\mu \sigma_i\right)^2\right]. \tag{14}$$

Since the hidden units are independent in the RBMs under study, we can write $P(z|\sigma) = \prod_{\mu=1}^{P} P(z_\mu|\sigma)$.
As for the dynamics of the visible layer, each spin perceives an effective field (that is the sum of the overall signal and the threshold for firing) that is compared with the noise in such a way that if the signal prevails over the noise the neuron spikes. Hence, for the σ's, we can write

$$P(\sigma_i|z) = \frac{e^{\frac{\beta}{\sqrt{N}}\sigma_i \sum_\mu^P \xi_i^\mu z_\mu + \beta\theta_i\sigma_i}}{2\cosh\left(\beta\sum_\mu^P \xi_i^\mu z_\mu/\sqrt{N} + \beta\theta_i\right)}, \tag{15}$$

and, again, $P(\sigma|z) = \prod_{i=1}^{N} P(\sigma_i|z)$. In order to get the joint distribution $P(\sigma,z)$ and the marginal distributions $P(\sigma)$, we use Bayes' Theorem, i.e. $P(\sigma,z) = P(\sigma|z)P(z) = P(z|\sigma)P(\sigma)$, hence getting

$$P(\sigma,z) \propto \exp\left[-\frac{\beta}{2}\sum_{\mu=1}^{P} z_\mu^2 + \frac{\beta}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu \sigma_i z_\mu\right], \tag{16}$$

$$P(\sigma) \propto \exp\left[\frac{\beta}{2N}\sum_{i,j=1}^{N}\left(\sum_{\mu=1}^{P}\xi_i^\mu \xi_j^\mu\right)\sigma_i\sigma_j\right]. \tag{17}$$

Remarkably, one can see that the features learnt by the machine (see eq. (14)) correspond to the patterns that the machine will successively be able to retrieve (see eq. (17)), as this last equation is nothing but the Gibbs probability distribution for the original Hopfield model [7, 35].
In order to understand this from a statistical-mechanics perspective it is useful to re-write eq. (14) and eq. (17) in terms of the so-called Mattis magnetization, defined as

$$m_\mu \doteq \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu \sigma_i, \tag{18}$$

which represents the overlap between the (visible) spin configuration and the μ-th pattern ξ^μ (that is, a vector of length N). This results in the following two equations

$$P(z_\mu|m_\mu) = \sqrt{\frac{\beta}{2\pi}}\exp\left[-\frac{\beta}{2}\left(z_\mu - \sqrt{N}m_\mu\right)^2\right], \tag{19}$$

$$P(m_\mu) \propto \exp\left(\frac{N\beta}{2}\sum_{\mu=1}^{P} m_\mu^2\right). \tag{20}$$

Interestingly, equation (19) shows that the z_μ's are Gaussian variables centered at the (properly rescaled) value of the related Mattis magnetization, and can therefore be interpreted as *features detectors* because they "discover" the correct Mattis magnetization, given the available patterns. Furthermore, equation (20) can be interpreted as a Boltzmann factor for the Hamiltonian $H \propto -\sum_\mu m_\mu^2$, therefore, a minimum energy argument suffices to show that, when possible (i.e., in the low-noise limit), the most convenient σ-states are those such that $m_\mu \to \pm 1$ for some μ. In other words, once these features have been resolved from the data (and automatically stored as patterns in the network) during training, further data will be classified according to these features. An alternative, useful perspective to tackle BM learning can be understood by looking at eq. (12): one can check that the signal term in that equation is the rescaled μ-th Mattis magnetization, hence if the visible layer is fed with a pattern to be retrieved (i.e., a vector containing one of the learnt features), then the corresponding Mattis magnetization rise, hence supporting the growth of its relative hidden variable z_μ, as a staggered magnetic field. On the other hand, if the state σ does not coincide with the μ-th pattern, m_μ will be zero in the large volume limit, and no net signal will be felt by the corresponding feature detector z_μ. Also, still at this qualitative level, eq. (20) shows that the landscape where the σ's are allowed to relax is the same as the landscape of a set of P Curie-Weiss models; recalling that in the Curie-Weiss model the magnetization (in modulus) goes

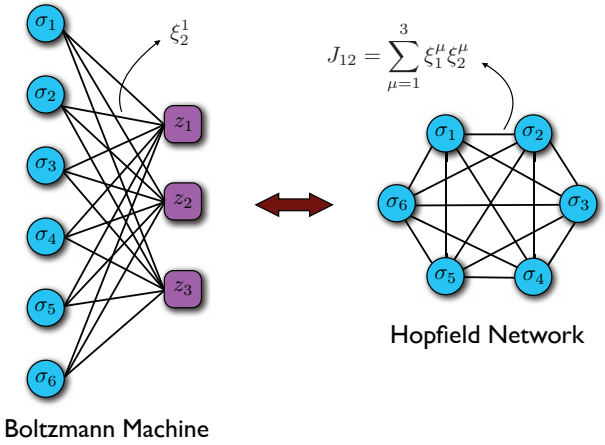

FIG. 1: **Schematic representation of the (restricted) Boltzmann machine (left panel) and its corresponding dual, the Hopfield network.** Left panel: example of a RBM equipped with six neurons in the visible layer, $\sigma_1,...,\sigma_6$ and three neurons in the hidden layer $z_1,...,z_3$. The weights among the two layers are coded by the $N \times P$ matrix ξ_i^μ. Right panel: dual example of the corresponding Hopfield model, obtained by marginalization over the hidden variables. This network uses solely the $\sigma_1,...,\sigma_6$ neurons, whose links however are now arranged according to the Hebb prescription for learning, that is $J_{ij} = \sum_{\mu=1}^{3} \xi_i^\mu \xi_j^\mu$.

to one in the small-noise limit, here we have that at least one Mattis magnetization (in modulus) will tend to one in the zero noise limit[**?**]. To fix ideas, let us say $m_1 = 1$: from definition (18) we realize that this is possible if and only if the σ's are parallel (in modulus) to the weights coded in ξ^1, and this situation corresponds to the *retrieval* of pattern ξ^1. Note that, exactly as for the ferromagnets described by the Curie-Weiss model [7], the crucial observable to detect a retrieval phase (or a ferromagnetic phase) is the model free-energy: by studying this observable (and its derivatives) with respect to the system parameters one can build the phase diagram of the model highlighting regions (and their boundaries) where the emerging behavior of the system is qualitatively different [9].

We close this section pointing out that for *discrete* weights/patterns the contrastive divergence algorithm shown in the learning section can not be applied as it requires stochastic descent over the weights that must therefore be *real* (and differentiable). For discrete pattern's entries the most widespread learning rule is the Hebbian one, which simply consists in storing patterns *adiabatically*, one after another, up to the saturation of the memory: calling J_{ij} the effective coupling between the two neurons σ_i and σ_j, given P patterns ξ^μ, namely P vectors of length N of binary entries, such a prescription results in

$$J_{ij} = \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \tag{21}$$

that, formally, does coincide with the coupling emerging from the duality between RBMs and Hopfield networks, as highlighted by eq. (20).

Before proceeding with the presentation of theorems for the RBMs, a couple of remarks are in order. First, for the sake of simplicity, in the following we will omit thresholds $\theta_i (i = 1,...,N)$ because their mathematical treatment is trivial as they do not involve correlations between spins; this does not imply any loss of generality, as thresholds can always be re-introduced at any moment. Moreover, in the next sections, we will work with random patterns, symmetrically distributed around zero: by a Shannon's compression argument this is the worst choice and therefore makes our assertions fully general with respect to the choice of patterns. In fact, if the network is able to cope with P fully random patterns, it will certainly be able to cope with at least P patterns displaying some degree of correlation (this is the case in real-world applications where correlations are always present, at least, due to the finiteness of P, and N).

III. RESTRICTED BOLTZMANN MACHINES WITH BOOLEAN PATTERNS

In this section, assuming the existence of the infinite-volume limit for the free energy of RBM with Boolean patterns, we will prove the self-average property of the free-energy around its quenched expression and we will give the explicit expression of its annealed approximation.

A. Preliminary Definitions

Definition 1 *We consider a RBM described by the following Hamiltonian:*

$$H_N(\sigma,z|\xi) = -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu z_\mu \sigma_i, \tag{22}$$

where the σ_i ($i\in[1,...,N]$) are N Ising spins forming the visible layer, the z_μ ($\mu\in[1,...,P]$) are P Gaussian $\mathcal{N}[0,\beta^{-1}]$ spins forming the hidden layer and the ξ_i^μ ($i\in[1,...,N],\mu\in[1,..,P]$) are i.i.d. random variables with discrete values ±1, which provide the weight of the link connecting the visible unit i and the hidden unit μ.

We assume that the two layers scale linearly in their reciprocal volumes, i.e., $\lim_{N\to\infty}P/N=\alpha\in\mathbb{R}^+$. This regime corresponds to the so-called high-load [24], in contrast with the so-called low-load regime characterized by a vanishing ratio P/N as $N\to\infty$: the latter is by far less tricky as it does not require the introduction of replicas [7, 10], while the former is the mathematically challenging one [28, 39, 46, 64].
As anticipated, we are interested in the thermodynamic properties of the neurons (i.e., the *fast variables*) in both the parties, while the weights (i.e., the *slow variables*) are assumed to be quenched and drawn randomly from the distribution

$$P(\xi_i^\mu=+1)=P(\xi_i^\mu=-1)=1/2, \tag{23}$$

independently for all $i=1,...,N$ and all $\mu=1,...,P$.

Definition 2 *We introduce the (random) partition function $Z_{P,N}(\beta|\xi)$ as*

$$Z_{P,N}(\beta|\xi)=\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}dz_\mu\sqrt{\frac{\beta}{2\pi}}e^{-z_\mu^2\beta/2}\sum_\sigma e^{-\beta H_N(\sigma,z|\xi)}, \tag{24}$$

and the usual definitions of the intensive free energy $A(\alpha,\beta)$, of the quenched intensive free energy $A^Q(\alpha,\beta)$ and of the annealed intensive free energy $A^A(\alpha,\beta)$ as, respectively,

$$A(\alpha,\beta) = \lim_{N\to\infty}A_{P,N}(\beta,\xi),\quad A_{P,N}(\beta,\xi)\doteq\frac{1}{N}\log Z_{P,N}(\beta|\xi), \tag{25}$$

$$A^Q(\alpha,\beta) = \lim_{N\to\infty}A^Q_{P,N}(\beta),\quad A^Q_{P,N}(\beta)\doteq\frac{1}{N}\mathbb{E}\log Z_{P,N}(\beta|\xi), \tag{26}$$

$$A^A(\alpha,\beta) = \lim_{N\to\infty}A^A_{P,N}(\beta),\quad A^A_{P,N}(\beta)\doteq\frac{1}{N}\log\mathbb{E}Z_{P,N}(\beta|\xi), \tag{27}$$

where the subscripts P,N highlight when we are working at finite volume and the symbol $\mathbb{E}$ represents the average over the quenched variables, that is

$$\mathbb{E}\doteq\frac{1}{2^{PN}}\sum_{\{\xi_i^\mu=\pm1\}_{i=1,..,N}^{\mu=1,...,P}}. \tag{28}$$

In the following, if not otherwise specified, the thermodynamic observables are meant as intensive. Also, we recall that the free energy we use is simply a rescaling of the (probably more popular) free energy $f_{P,N}(\alpha,\beta)=-\beta^{-1}A_{P,N}(\alpha,\beta)$.

Definition 3 *Given an observable $y(\sigma,z|\xi)$ depending on the neuron configuration and on the weights, once the partition function (24) is introduced, we can also define the product Boltzmann state over s replicas of the system as*

$$\Omega(y(\sigma,z|\xi))=\omega_1(y(\sigma^1,z^1|\xi))\bigotimes...\bigotimes\omega_s(y(\sigma^s,z^s|\xi)),$$

where ω_i represents the Gibbs measure for the a-th replica ($a=1,...,s$), that is

$$\omega_a(y(\sigma^a,z^a|\xi))=\frac{\prod_{\mu=1}^P\int_{-\infty}^{+\infty}dz_\mu^a e^{-(z_\mu^a)^2\beta/2}\sqrt{\frac{\beta}{2\pi}}\sum_{\sigma^a}y(\sigma^a,z^a|\xi)e^{-\beta H_N(\sigma^a,z^a|\xi)}}{Z_{P,N}(\beta|\xi)}.$$

Finally, we introduce the symbol $\langle y(\sigma,z|\xi)\rangle$ to mean

$$\langle y(\sigma,z|\xi)\rangle=\mathbb{E}\Omega\left(y(\sigma,z|\xi)\right).$$

B. Main results

In this subsection we prove the main Theorem on the self-averaging properties of the free energy $A_{P,N}(\beta,\xi)$ (around its quenched expression, in the thermodynamic limit) of the machine defined by eq. (22), then we give the explicit expression for its annealed approximation $A^A(\alpha,\beta)$.

Theorem 1 *The free energy of the Boltzmann machine defined by the Hamiltonian (22) is a self-averaging quantity: in the thermodynamic limit its fluctuations vanish and force the latter over its quenched expectation, i.e.*

$$\lim_{N\to\infty}\mathbb{E}\left[\left(A_{P,N}(\beta|\xi)-\mathbb{E}\left(A_{P,N}(\beta|\xi)\right)\right)^2\right]=0. \tag{29}$$

If the thermodynamic limit of the free energy exists, its quenched expectation is bounded by its annealed one, i.e., $A^Q(\alpha,\beta)=\lim_{N\to\infty}N^{-1}\mathbb{E}\log Z_{P,N}(\beta|\xi)\leq\lim_{N\to\infty}N^{-1}\log\mathbb{E}Z_{P,N}(\beta|\xi)=A^A(\alpha,\beta)$. In fact, for small β (that is, $\beta<1$), $A^A(\alpha,\beta)$ coincides with the quenched free energy $A^Q(\alpha,\beta)$, and, in general, it works as its upper bound thanks to Jensen inequality [33, 34]. This motivates the next proposition.

Proposition 1 *The asymptotic value of the annealed free-energy of the RBM is*

$$A^A(\alpha,\beta)=\lim_{N\to\infty}\frac{1}{N}\ln\mathbb{E}[Z_{P,N}(\beta|\xi)]=\ln 2-\frac{\alpha}{2}\ln(1-\beta). \tag{30}$$

The proof of Theorem 1 can be found in the Appendix A, while the proof of Proposition 1 is reported hereafter[**?**].

Proof Recalling the partition function of the model

$$Z_{P,N}(\beta|\xi)=\sum_{\sigma}\prod_{\mu=1}^{P}\int_{-\infty}^{+\infty}d\mu(z_\mu)e^{\left(\frac{\beta}{\sqrt{N}}\sum_i\xi_i^\mu\sigma_i\right)z_\mu}, \tag{31}$$

we perform the Gaussian integration to get

$$Z_{P,N}(\beta|\xi)=\sum_{\sigma}\prod_{\mu=1}^{P}\sum_{k=0}^{\infty}\left(\frac{\beta}{2N}\right)^k\frac{\mathscr{K}_\mu(\sigma|\xi)^{2k}}{k!}, \tag{32}$$

where, for brevity, we posed $\mathscr{K}_\mu(\sigma|\xi)=\sum_{i=1}^N\xi_i^\mu\sigma_i$. We now average over the weights $\{\xi_i^\mu\}$ to get $\mathbb{E}[Z_{P,N}(\beta|\xi)]$ and, to this goal, we focus on $\mathbb{E}[\mathscr{K}_\mu(\sigma|\xi)^{2k}]$:

$$\begin{aligned}
\mathbb{E}[\mathscr{K}_\mu(\sigma|\xi)^0] &= 1,\\
\mathbb{E}[\mathscr{K}_\mu(\sigma|\xi)^2] &= \sum_{i,j=1}^N\sigma_i\sigma_j\mathbb{E}[\xi_i^\mu\xi_j^\mu]=\sum_{i,j=1}^N\sigma_i\sigma_j\delta_{ij}=N,\\
\mathbb{E}[\mathscr{K}_\mu(\sigma|\xi)^4] &= \sum_{i,j,k,l=1}^N\sigma_i\sigma_j\sigma_k\sigma_l\mathbb{E}[\xi_i^\mu\xi_j^\mu\xi_k^\mu\xi_l^\mu],\\
&= \sum_{i,j,k,l=1}^N[(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk})+\delta_{ijkl}-3\delta_{ijkl}]\sigma_i\sigma_j\sigma_k\sigma_l=3N^2-2N,
\end{aligned}$$

where we exploited the orthogonality of the weights and, in particular, in the last line the non-null contributions correspond to pair-wise equal indices and four-wise equal indices, properly accounting for repetitions. This calculation can be generalized to higher order, yet it gets more awkward and, ultimately, not necessary to our purposes. In fact, one can notice that, as $N\to\infty$, the leading term for $\mathbb{E}[\mathscr{K}_\mu(\sigma|\xi)^{2k}]$ is order N^k while subleading terms get vanishing under logarithm and normalization, as prescribed by the formula (27). Such leading term is given by $\prod_{i=0}^{k-1}(2k-2i-1)N^k$, which accounts for only pair-wise contributions, and

$$\left[1+\sum_{k=1}^{\infty}\left(\frac{\beta}{2N}\right)^k\frac{\prod_{i=0}^{k-1}(2k-2i-1)N^k}{k!}\right]=\frac{1}{\sqrt{1-\beta}}. \tag{33}$$

Consequently, stressing that the average over ξ allows getting rid of the σ variables, so that the sum over all σ configurations simply provides a factor 2^N, and the dependence on μ is dropped so the product over μ simply provides a power P, we can write

$$\lim_{N\to\infty}\frac{1}{N}\ln\mathbb{E}[Z_{P,N}(\beta|\xi)]=\lim_{N\to\infty}\frac{1}{N}\ln\left[2^N(1-\beta)^{-P/2}\right]=\ln 2-\frac{\alpha}{2}\ln(1-\beta). \tag{34}$$

IV. RESTRICTED BOLTZMANN MACHINES WITH GAUSSIAN PATTERNS.

In this section, assuming the existence of the infinite-volume limit for the free energy of the RBM with Gaussian patterns, we will provide an explicit expression for the quenched free energy, under the assumption of replica symmetry.
We remark that in this entire section we will cover solely the technical aspects from a mathematical perspective, while we leave the next section for technical aspects from an applicative perspective.

A. Preliminary Definitions

Definition 4 *We consider a RBM described by the following Hamiltonian*

$$H_N(\sigma,z|\xi) = -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu z_\mu \sigma_i, \tag{35}$$

where σ_i $(i\in[1,...,N])$ are N Ising spins forming the visible layer, z_μ $(\mu\in[1,...,P])$ are P Ising spins forming the hidden layer, while ξ_i^μ $(i\in[1,...,N],\mu\in[1,..,P])$ are quenched and, in particular, this time are chosen as i.i.d. Gaussian random variables $\mathcal{N}[0,1]$, namely the weights connecting the two layers are analogical [13, 15].

Again, we assume the two layers to scale linearly in the reciprocal volume, i.e., $\lim_{N\to\infty} P/N = \alpha \in \mathbb{R}^+$. The definitions of the partition function and the free energy (as well as its quenched and annealed expectations) coupled to the Hamiltonian (35) are analogous to those given in the previous section, apart for the average over the patterns: the symbol $\mathbb{E}$ now represents the average over all the Gaussian distributed (and no longer Boolean) pattern entries, namely

$$\mathbb{E} \doteq \int_{-\infty}^{+\infty}\prod_{i=1}^{N}\prod_{\mu=1}^{P} d\xi_i^\mu \frac{e^{-(\xi_i^\mu)^2/2}}{\sqrt{2\pi}}. \tag{36}$$

Definition 5 *The (spin-glass) order parameters of the RBM defined by eq. (35) are the two-replica overlap in the visible layer $q_{12}(\sigma)$ and the two-replica overlap in the hidden layer $p_{12}(z)$ defined as*

$$q_{12} \doteq \frac{1}{N}\sum_{i=1}^{N}\sigma_i^1\sigma_i^2, \qquad p_{12} \doteq \frac{1}{P}\sum_{\mu=1}^{P} z_\mu^1 z_\mu^2,$$

B. Preliminary results

In this subsection we present our strategy of investigation; namely we prove some theorems and propositions whose implications will be exploited in the following subsection. Taken a real scalar $n\in(0,1]$

$$\mathbb{E}\ln Z_{P,N}(\beta|\xi) = \lim_{n\to 0}\frac{1}{n}\ln\mathbb{E}\left[Z^n(\beta|\xi)\right], \tag{37}$$

and we use this relation in order to write $A^Q(\alpha,\beta) = \lim_{N\to\infty}\lim_{n\to 0}\phi_{P,N}(\beta,n)$, where

Definition 1 *The n-quenched free energy $\phi_{P,N}(\beta,n)$ has the following interpolating functional*

$$\phi_{P,N}(\beta,n) = \frac{1}{Nn}\ln\mathbb{E}\left[Z_{P,N}^n(\beta|\xi)\right]. \tag{38}$$

The scalar n is thought of as an interpolating parameter such that $\phi_{P,N}(\beta,n)$ recovers the annealed free-energy when $n=1$ and the quenched free-energy when $n\to 0$. The former statement can be checked by directly replacing $n=1$ in (38) and comparing with (26); the latter statement is given by the following

Theorem 2 *The following relations between the n-quenched free energy and the quenched free energy exist*

$$\lim_{n\to 0}\phi_{P,N}(\beta,n) = A_{P,N}^Q(\beta), \tag{39}$$

furthermore

$$\phi_{P,N}(\beta,n) \geq A_{P,N}^Q(\beta), \tag{40}$$

for any n.

Proof We can Taylor-expand the n-quenched free energy for n close to 0 to get

$$\begin{aligned}\ln \mathbb{E}\left[Z_{P,N}^n(\beta|\xi)\right] &= 0+\frac{\mathbb{E}[Z_{P,N}^n(\beta|\xi)\log Z_{P,N}^n(\beta|\xi)]}{\mathbb{E}Z_{P,N}^n(\beta|\xi)}n+o(n^2)\\ &= \mathbb{E}\left[\ln Z_{P,N}(\beta|\xi)\right]\cdot n+o(n^2)\Rightarrow\\ \lim_{n\to 0}\phi_{P,N}(\beta,n) &= \lim_{n\to 0}\frac{1}{nN}\left[\mathbb{E}\left(\ln Z_{P,N}(\beta|\xi)\cdot n+o(n^2)\right)\right]=A_{P,N}^{Q}(\beta),\end{aligned} \tag{41}$$

which proves (39), while (40) is guaranteed by the Jensen inequality. □

C. Main results

Our strategy is based on two interchained interpolation schemes: one, anticipated before, that works on replicas, the other, described hereafter, that works on spin coupling. Before proceeding, we stress that we will exploit our strategy solely within the so-called *replica symmetric approximation*, namely under the assumption that, in the asymptotic limit, the order parameters do not fluctuate (i.e., their distributions get delta-peaked over their thermodynamic averages).
Let us now introduce the interpolating structure acting on couplings. First, we give a few definitions which will lighten the calculations along the way.

Definition 2 *We introduce, as an interpolating parameter, a real scalar $t\in[0,1]$, further we need N i.i.d. $\mathcal{N}[0,1]$ random variables, referred to as J_i, $i\in(1,...,N)$, and P i.i.d. $\mathcal{N}[0,1]$ random variables, referred to as J_μ, $\mu\in(1,...,P)$, and we define the interpolating partition function Z_t as*

$$Z_t=\sum_{\sigma}\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu)e^{\sqrt{t}\left(\frac{\sqrt{\beta}}{\sqrt{N}}\sum_{i,\mu}^{N,P}\xi_i^\mu\sigma_i z_\mu\right)+\sqrt{1-t}\left(\sqrt{\alpha\beta\bar{p}}\sum_i^N J_i\sigma_i+\sqrt{\beta\bar{q}}\sum_\mu^P J_\mu z_\mu\right)+(1-t)\beta(1-\bar{q})\sum_\mu^P\frac{z_\mu^2}{2}}. \tag{42}$$

Notice that when $t=1$ we recover the partition function $Z_{P,N}(\beta|\xi)$ of the RBM under study, while when $t=0$ we get the partition function of a one-body model for the variables σ and z. We also remark that now the quenched expectation $\mathbb{E}$ applies also over (J_i,J_μ), i.e., $\mathbb{E}=\prod_{i,\mu}\mathbb{E}_{\xi_i^\mu}\prod_{i,\mu}\mathbb{E}_{J_i}\mathbb{E}_{J_\mu}$.

Definition 3 *Given a smooth function $f(\sigma,z|\xi)$ of the neurons (σ,z) and of the patterns ξ, we introduce a deformed measure $\langle f(\sigma,z|\xi)\rangle_n$ as*

$$\langle f(\sigma,z|\xi)\rangle_n\equiv\mathbb{E}\left[Z_t^n\cdot(\mathbb{E}(Z_t^n))^{-1}\cdot\Omega(f(\sigma,z|\xi))\right], \tag{43}$$

where Ω is the standard generalized 2-product Boltzmann state, namely $\Omega=\omega_1\bigotimes\omega_2$.
Note that this new measure collapses on the standard one for $n\to 0$, namely $\langle f(\sigma,z|\xi)\rangle_n\to\langle f(\sigma,z|\xi)\rangle$ as $n\to 0$.

Definition 4 *Once introduced a scalar parameter $t\in[0,1]$, the following interpolating functional*

$$\phi_{P,N}(\beta,n,t)=\frac{1}{Nn}\ln\mathbb{E}(Z_t^n) \tag{44}$$

bridges the n-quenched free energy of the RBM to an integrable one-body model.

With these definitions we can state

Proposition 2 *The following boundary values for the interpolating n-quenched free energy hold*

$$\phi_{P,N}(\beta,n,t=1) = \frac{1}{Nn}\ln\mathbb{E}(Z_1^n)=\phi_{P,N}(\beta,n), \tag{45}$$

$$\begin{aligned}\phi_{P,N}(\beta,n,t=0) &= \ln 2+\frac{1}{n}\ln\int_{-\infty}^{+\infty}d\mu(x)\cosh^n\left(x\sqrt{\alpha\beta\bar{p}}\right)-\frac{\alpha}{2}\ln\left[1-\beta(1-\bar{q})\right]\\ &+ \frac{\alpha\beta}{2}\frac{\bar{q}}{1-\beta(1-\bar{q})}.\end{aligned} \tag{46}$$

Theorem 3 *In the replica symmetric approximation, the thermodynamic limit of the free energy $A_{P,N}(\beta,\xi)$ converges to its quenched expectation, that reads as*

$$\begin{aligned} A^Q(\alpha,\beta) &= \ln 2 + \int_{-\infty}^{+\infty} d\mu(x) \ln\cosh\left(x\sqrt{\alpha\beta\bar{p}}\right) - \frac{\alpha}{2}\ln\left[1-\beta(1-\bar{q})\right] \\ &+ \frac{\alpha\beta}{2}\frac{\bar{q}}{1-\beta(1-\bar{q})} - \frac{\alpha\beta}{2}\bar{p}(1-\bar{q}). \end{aligned} \tag{47}$$

whose extremization with respect to $\bar{q}$, $\bar{p}$ returns

$$\frac{\partial A^Q(\alpha,\beta)}{\partial \bar{q}} = 0 \Rightarrow \bar{p} = \frac{\beta\bar{q}}{\left[1-\beta(1-\bar{q})\right]^2}, \tag{48}$$

$$\frac{\partial A^Q(\alpha,\beta)}{\partial \bar{p}} = 0 \Rightarrow \bar{q} = \int d\mu(x) \tanh^2\left(x\sqrt{\alpha\beta\bar{p}}\right). \tag{49}$$

Corollary 1 *The RBM has an ergodic regime (where the annealed approximation holds and $A^A(\alpha,\beta)$ equals the quenched replica symmetric free energy $A^Q(\alpha,\beta)$ evaluated at $\bar{q}=\bar{p}=0$) that is bounded by the following transition line, in the α,β plane of tuneable parameters*

$$\beta_c = \lim_{n\to 0} \frac{1}{1+\sqrt{\alpha/(1-n)}} = \frac{1}{1+\sqrt{\alpha}}. \tag{50}$$

The proofs of Proposition 2, Theorem 3 and Corollary 1 can be found in the Appendix B.

Remark 1 *We note that the critical line delimiting the ergodic region for the RBMs is the same critical line that traces the ergodic boundary for the Hopfield model [7, 15, 53]. However, while here α is the ratio between the hidden and visible layer sizes, in the Hopfield model α is the ratio between the stored patterns and the neurons necessary to handle them.*
Further, the annealed expression for the free energy of this RBM is the same as the one obtained for the model treated in Sec. 3 in the high noise limit (see Proposition 1). In fact, by setting $\bar{q}=\bar{p}=0$ in eq. (47) we get

$$A^A(\alpha,\beta) = \ln 2 - \frac{\alpha}{2}\ln(1-\beta). \tag{51}$$

V. THE SIGNAL AND THE NOISE

In this section we deepen the inferential capabilities of these machines and the properties of the inner (slow) noise. More precisely, we introduce a "signal", namely an external field which favors the retrieval of a given pattern and, by inspecting its corresponding Mattis magnetization, we check whether and how robustly (with respect to the parameters (α,β) this pattern can be detected over the noise. Next, we further address the noise and we provide an argument for its universality to hold, namely we will show that, in the thermodynamic limit, it does not matter if we consider Boolean or Gaussian patterns, they contribute in the same way to the (slow) noise.

Proposition 3 *We consider one Boolean pattern ξ^1 and $P-1$ Gaussian patterns ξ^μ, $\mu=2,...,P$; given a signal which favours the retrieval of pattern ξ^1, we measure the accuracy of the signal reconstruction in terms of the related Mattis magnetization $m_1(\sigma)$ which fulfils an Hopfield-like self-consistent equation.*

Proof The proof works by extending the interpolating scheme coded by eq. (42) in order to account for a signal term: in this way we get a self-consistent equation for the related Mattis magnetization $\bar{m}_1$.
Let us consider the interpolating partition function (42): it is enough to add to its exponent the terms $(1-t)\beta\bar{m}_1\sum_i \xi_i^1\sigma_i + t\beta m_1^2(\sigma)$ to extend the previous quenched free energy, still at the replica symmetric level (i.e., assuming $\lim_{N\to\infty}P(m_1) = \delta(m_1-\bar{m}_1)$ to obtain

$$A^Q(\alpha,\beta) = \ln 2 + \int_{-\infty}^{+\infty} d\mu(x)\ln\cosh\left(x\sqrt{\alpha\beta\bar{p}}+\beta\bar{m}_1\right) - \frac{\alpha}{2}\ln\left[1-\beta(1-\bar{q})\right] \tag{52}$$

$$+ \frac{\alpha\beta}{2}\frac{\bar{q}}{1-\beta(1-\bar{q})} - \frac{\alpha\beta}{2}\bar{p}(1-\bar{q}) - \frac{\beta}{2}\bar{m}_1^2. \tag{53}$$

whose extremization w.r.t. $\bar{m}_1$ returns a self consistency equation for the signal that reads

$$\frac{\partial A^Q(\alpha,\beta)}{\partial \bar{m}_1} = 0 \Rightarrow \bar{m}_1 = \int_{-\infty}^{+\infty} d\mu(x)\tanh\left(x\sqrt{\alpha\beta\bar{p}}+\beta\bar{m}_1\right). \tag{54}$$

By replacing $\bar{p}$ with the expression in terms of $\bar{q}$ given by (48), eq. (54) recovers the self-consistent equation for the Mattis magnetization in the Hopfield scenario as traced by the AGS theory [7, 52]. □

This result is naturally consistent with the intrinsic analogies between the Hopfield model and the RBM, as depicted by eqs. (16,17).
We remark further that the slow noise enters and plays a crucial role in the self-consistency for the signal, thus this should be properly estimated in machine learning applications. It is therefore important to state the next

Proposition 4 *The noise generated by the not-retrieved patterns over the signal is universal, namely, in the thermodynamic limit, it is the same if the pattern entries are either sampled from a Boolean distribution or from a Gaussian distribution.*

Proof Let us consider the n-product of the partition function, taking ξ^1 as the retrieved pattern (whose signal is detected by $m_1(\sigma)$) we perform the quenched average over the $P-1$ quenched patterns (overall acting as a slow-noise against the retrieval of ξ^1) in complete generality as

$$\mathbb{E}Z^n_{P,N}(\beta|\xi) = \mathbb{E}\prod_{a=1}^{n}\prod_{i=1}^{N}\sum_{\sigma_i^a}\int_{-\infty}^{+\infty}\prod_{a=1}^{n}\prod_{\mu=1}^{P}d\mu(z_\mu^a)e^{\left(\frac{\beta N}{2}\sum_a^n m_1^2(\sigma^a)+\sqrt{\frac{\beta}{N}}\sum_a^n\sum_i^N\sum_\mu^P\xi_i^\mu\sigma_i^a z_\mu^a\right)} \tag{55}$$

$$= \prod_{a=1}^{n}\prod_{i=1}^{N}\sum_{\sigma_i^a}\int_{-\infty}^{+\infty}\prod_{a=1}^{n}\prod_{\mu=1}^{P}d\mu(z_\mu^a)e^{\left(\frac{\beta N}{2}\sum_a^n m_1^2(\sigma^a)\right)}\mathbb{E}e^{\left(\sqrt{\frac{\beta}{N}}\sum_a^n\sum_i^N\sum_\mu^P\xi_i^\mu\sigma_i^a z_\mu^a\right)} \tag{56}$$

As the distributions share unitary variance and are both symmetric, the quenched average over the patterns returns

$$\mathbb{E}\left[\exp\left(\sqrt{\frac{\beta}{N}}\sum_{a=1}^{n}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\sigma_i^a z_\mu^a\right)\right]\sim\exp\left(\frac{\beta}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{P}\sum_{a,b=1}^{n}\sigma_i^a\sigma_i^b z_\mu^a z_\mu^b\right) \tag{57}$$

□

We remark that, while differently sampled patterns behave as the same source of noise, actually they generate quite different signals [9], i.e. universality holds just for the noise.
An interesting feature of these networks, shared by several other glassy systems [6, 30], is that their overlaps (i.e., q_{12} and p_{12}) display complex, non-Gaussian fluctuations (and their complexity explains why usually researchers work at the replica symmetric level). This is well known in the spin-glass Literature, where the linear constraints fulfilled by fluctuations are usually named Aizenman-Contucci polynomials [6] (while the more general ones are known as Ghirlanda-Guerra identities [30]). Nonetheless, this feature is largely ignored among researchers in Machine Learning and hereafter we show that these constraints actually hold also for these networks. In particular, due to the theory developed up to this point, we will focus in details only on the former, namely on the Aizenman-Contucci identities (suitably adapted to the present case):

Proposition 5 *In the thermodynamic limit, β almost-everywhere, the following Aizenman-Contucci polynomials hold for the RBM's order parameters q_{12}, p_{12}:*

$$\langle q_{12}^2 p_{12}^2\rangle - 4\langle q_{12}p_{12}q_{23}p_{23}\rangle + 3\langle q_{12}p_{12}q_{34}p_{34}\rangle = 0. \tag{58}$$

Proof It is enough to explicitly write down the β streaming of their expectation, namely

$$\partial_\beta\langle q_{12}p_{12}\rangle = \frac{1}{NP}\sum_{i,\mu}\mathbb{E}\partial_\beta\Omega^2(z_\mu\sigma_i) = \frac{1}{NP}\sum_{i,\mu}\mathbb{E}2\Omega(z_\mu\sigma_i)\partial_\beta\Omega(z_\mu\sigma_i) \tag{59}$$

$$= \frac{2}{NP}\sum_{i,j,\mu,\nu}\mathbb{E}\Omega(z_\mu\sigma_i)\xi_j^\nu\left[\Omega(z_\mu\sigma_i z_\nu\sigma_j)-\Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\right], \tag{60}$$

and we use Wick's theorem on ξ to get

$$\begin{aligned}\partial_\beta\langle q_{12}p_{12}\rangle = \frac{2}{N^2P^2}\sum_{\mu,\nu,i,j}\Big\{&\Big[\Omega(z_\mu\sigma_i z_\nu\sigma_j)-\Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\Big]\Big[\Omega(z_\mu\sigma_i z_\nu\sigma_j)-\\ &+\ \Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\Big]+\Omega(z_\mu\sigma_i)\Big[\Omega(z_\mu\sigma_i\sigma_j z_\nu z_\nu\sigma_j)-\Omega(z_\mu\sigma_i z_\nu\sigma_j)\Omega(z_\nu\sigma_j)\\ &-\ \Omega(z_\mu\sigma_i)\Omega(z_\mu\sigma_i z_\nu\sigma_j)\Omega(z_\nu\sigma_j)+\Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\Omega(z_\nu\sigma_j)\Omega(z_\mu\sigma_i)\\ &-\ \Omega(z_\mu\sigma_i)\Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\Omega(z_\nu\sigma_j)+\Omega(z_\mu\sigma_i)\Omega(z_\nu\sigma_j)\Omega(z_\nu\sigma_j)\Omega(z_\mu\sigma_i)\Big]\Big\}.\end{aligned}$$

Finally, by re-introducing the overlaps and making all the cancellations that happen to appear, we can write

$$\partial_\beta \langle q_{12} p_{12} \rangle = \alpha N \Big(\langle q_{12}^2 p_{12}^2 \rangle - 4 \langle q_{12} p_{12} q_{23} p_{23} \rangle + 3 \langle q_{12} p_{12} q_{34} p_{34} \rangle \Big) \tag{61}$$

and, by a stability requirement (i.e., this streaming is bounded), again in the thermodynamic limit the thesis is proved. □

VI. CONCLUSIONS

In this work we addressed RBMs from a mathematical physics perspective.
First, we provided a basic background on these systems trying to explain, in a non-rigorous way, the mechanisms through which they are able to perform statistical learning and information retrieval; these capabilities make RBMs key tools in the rapidly expanding field of Artificial Intelligence. In particular, we showed how to obtain the contrastive divergence algorithm for learning of continuous weights and how, ultimately, this recipe displays the same mathematical representation of the Hebb rule, a well consolidated prescription for learning with discrete weights [35]. Further, by a robust Bayesian argument, we showed a formal equivalence between RBMs and Hopfield networks, where the weights learnt by the RBM are exactly the patterns successively retrieved by the (dual) Hopfield network [11]. Although learning and retrieval are typically addressed separately (thanks to adiabatic-like hypothesis), these results restore the intrinsic interplay between *learning* and *retrieval* which, in fact, are two aspects of the single and broader act of *cognition*.

Then, adopting a mathematical rigorous treatment, we proved several results concerning the two main representations of RBMs (analogical versus digital weights). From the statistical mechanical perspective, these systems are two-party spin-glasses whose external party (or *layer* to keep the original jargon) is always built up by binary Ising spins/neurons, that are fed by the external information; the inner layer, instead, detects features contained in the data presented as input by suitable learning algorithms which tune the weights connecting the two layers (e.g., contrastive divergence and simulated annealing).
Once the learning stage is over, the system relaxes to a Gibbs-measure generated by the Hamiltonian of the RBM, whose control is entirely deducible by the knowledge of its free energy: the latter thus plays as the principal observable whose investigation is in order. However, while these systems are becoming pervasive in the applied world, at the rigorous level, solely the Sherrington-Kirkpatrick is a fully understood spin-glass (the single-party [47, 51]) and any variation on theme (even possibly mild) is still a great challenge and much care should be paid in their mathematical treatment.

Here we showed that the free energy of these machines is a self-averaging quantity in the thermodynamic limit and that it self-averages over its quenched expectation. This was proven by adapting the Pastur-Shcherbina-Tirozzi argument [52, 53] (based on Doob Theorem for martingales), originally developed for standard Hopfield networks [52, 53]. Then, en route to the quenched free energy, we gave an expression of its annealed approximation that, due to Jensen inequality, plays as a natural bound for the quenched one.
Next, turning to machines with Gaussian weights, by an adaptation of Guerra's interpolation scheme [12], we provided the explicit expression of the free energy of these machines in the thermodynamic limit, under the assumption of replica symmetry. Remarkably, such a free energy turns out to share the same mathematical structure of the Hopfield one. Furthermore, we gave an argument based on the universality of the quenched noise (consistent with [29]) to generalize such a result to the previous machine with Boolean weights. The ergodicity line, that separates the region where the annealed approximation holds from the region where the quenched one prevails, turns out to be the same of the AGS theory for the Hopfield model [7, 53] thus conferring a bulk of robustness to the theory as a whole. Finally, we gave a glance at overlaps fluctuations, i.e., beyond the replica symmetry, and we found Aizenman-Contucci identities.

Acknowledgments

E.A. acknowledges financial support by Sapienza Università di Roma (project no. RG11715C7CC31E3D).
A.B. acknowledges financial support by Salento University, by INFN, by MIUR (via the basic research grant 2018) and by "Match/Pytagoras" project (FESR/FSE 2014/2020).
E.A. and A.B. acknowledge financial support by GNFM-INdAM ("Progetto Giovani 2018").
The authors are grateful to Francesco Guerra and Emanuele Mingione for usueful conversations.

Appendix A. Proof of Theorem 1

Proof First, let us explain the idea behind this proof. We consider the set of all possible 2^{NP} weight combinations $\mathscr{R}_N = \{\xi_i^\mu\}_{i=1,\dots,N}^{\mu=1,\dots,P}$ and we build a sequence of subsets

$$\mathscr{R}_k^N = \{\xi_i^\mu\}_{i\geq k}^{\mu=1,\dots,P}, \tag{62}$$

with $k = 1,...,N$. Next, we define the conditional expectations $\mathbb{E}_{<k}$ over these subsets and we show that the conditional expectation F_N^k of the extensive free energy over two consecutive subsets can be used to estimate the fluctuations in the intensive free energy. Finally, exploiting the martingale nature of the sequence $(F_N^k, \mathscr{R}_k^N)$, we can bound fluctuations and show that they are vanishing in the thermodynamic limit.
This method has been widely exploited in the past (see e.g., [53, 57, 58]) as it can be applied to different examples of neural networks, provided that patterns are independent and the probability of the conditioning subsets are non null (we refer to [63] for a more extensive discussion on this method).

Let us now describe the proof in more details. The sequence of subsets $\mathscr{R}_k^N$ with $k = 1,...,N$, defined in (62), constitutes a decreasing family of σ-algebras, over which we define the conditional expectation

$$\mathbb{E}_{<k} \doteq \frac{1}{2^{P(k-1)}} \sum_{\{\xi_i^\mu = \pm 1\}_{i<k}^{\mu=1,...,P}}. \tag{63}$$

In particular, the conditional expectation over $\mathscr{R}_k^N$ of the extensive free energy F_N is referred to as F_N^k and reads as

$$F_N^k = \mathbb{E}(F_N|\mathscr{R}_k^N) = \frac{1}{2^{P(k-1)}} \sum_{(\xi_i^\mu = \pm 1)_{i<k}^{\mu=1,...,P}} [\log Z_{P,N}(\beta|\xi)]. \tag{64}$$

Notice that in the left hand side the conditioning is with respect to any event of the σ-algebra $\mathscr{R}_k^N$ (namely with respect to any choice of the random variables ξ_i^μ, with $i \geq k$) and in the right hand side the average is performed over the non-conditioned events (namely over all possible weigths ξ_i^μ, with $i < k$). Otherwise stated, the expectation $\mathbb{E}(\cdot|\mathscr{R}_k^N)$ provides the average over a subset of the quenched variables, corresponding to $i < k$, while any particular realization of weights in $\mathscr{R}_k^N$ can be taken as a "boundary condition". Remarkably, the sequence constituted by the stochastic variables F_N^k and by the σ-algebra $\mathscr{R}_k^N$ fulfill the martingale property [52, 53]

$$\mathbb{E}(F_N^k|\mathscr{R}_l^N) = \begin{cases} F_N^k & if \quad l < k \\ F_N^l & if \quad l \geq k \end{cases}. \tag{65}$$

Now, by setting $k = 1$ and $k = N+1$ in Eq. 64, we get

$$F_N^1 = \log Z_{P,N}(\beta|\xi), \quad F_N^{N+1} = \mathbb{E}(\log Z_{P,N}(\beta|\xi)), \tag{66}$$

namely, we recover the extensive free-energy and quenched extensive free-energy, respectively. Let also introduce the relative increment Ψ_k as

$$\Psi_k = F_N^k - F_N^{k+1}, \tag{67}$$

in such a way that we can write the linear term inside eq. (29) as an arithmetic average of the incremental free-energy Ψ_k over the N visible neurons, that is

$$A_{P,N} - \mathbb{E}(A_{P,N}) = \frac{1}{N} \sum_{k=1}^{N} \Psi_k. \tag{68}$$

The free-energy fluctuations therefore reads as

$$\mathbb{E}(A_{P,N} - \mathbb{E}(A_{P,N}))^2 = \frac{1}{N^2} \sum_{k=1}^{N} \mathbb{E}\Psi_k^2 + \frac{2}{N^2} \sum_{k=1}^{N} \sum_{l=k+1}^{N} \mathbb{E}\Psi_k \Psi_l, \tag{69}$$

where in the right hand side we split the diagonal and the off-diagonal contributions. Actually, only the former matters since the latter is null as shown in the following:

$$\begin{aligned} \mathbb{E}(\Psi_k \Psi_l) &= \mathbb{E}(\mathbb{E}(\Psi_k \Psi_l|\mathscr{R}_l^N)) \\ &= \mathbb{E}(\Psi_l \mathbb{E}(\Psi_k|\mathscr{R}_l^N)) \\ &= \mathbb{E}(\Psi_l \mathbb{E}(F_N^k - F_N^{k+1}|\mathscr{R}_l^N)) \\ &= \mathbb{E}(\Psi_l \mathbb{E}(F_N^l - F_N^l)) \\ &= 0. \end{aligned} \tag{70}$$

More precisely, in the first passage we exploit the fact that $\mathbb{E}(\mathbb{E}(\cdot|\mathscr{R}_l^N)) = \mathbb{E}(\cdot)$, in the second passage we exploit the fact that the conditioning is effective only on Ψ_k (recalling that $k < l$), in the third passage we use the definition (67), and in the fourth passage we use Eq. 65 (recalling that $k+1 \leq l$).
Thus, in order to get the convergence in probability of the infinite volume limit of the free energy of the RBM (22) we only need to show that

$$\mathbb{E}\Psi_k^2 \leq C, \tag{71}$$

for some constant C. To this aim, let us define the following interpolating functions

$$\begin{array}{lcl} H_k & = & -\frac{1}{\sqrt{N}}\sum_{i=1,i\neq k}^{N}\sum_{\mu=1}^{P}\xi_i^\mu\sigma_i z_\mu, \\ R_k & = & -\frac{1}{\sqrt{N}}\sum_{\mu=1}^{P}\xi_k^\mu\sigma_k z_\mu, \\ \Phi_k(t) & = & H_k + tR_k, \\ g_k(t) & = & \log Z_{P,N}(\Phi_k(t)) - \log Z_{P,N}(\Phi_k(0)). \end{array} \tag{72}$$

Notice that H_k describes the original system but devoid of the k-th neuron (i.e., a "cavity" [47]), R_k represents the interaction term related to the k-th neuron, in such a way that the interpolating function $\Phi_k(t)$ returns the original Hamiltonian when $t = 1$, and H_k when $t = 0$. Finally, the function $g_k(t)$ vanishes when $t = 0$, while when $t = 1$ it provides the contribution to the (extensive) free energy given by the k-th spin. By applying the conditional average (63) to $g_k(1)$ and comparing with the definition (67) we have the following identity

$$\Psi_k = \mathbb{E}_{<k}\left(g_k(1)\right) - \mathbb{E}_{<k+1}\left(g_k(1)\right). \tag{73}$$

Next, by applying the Cauchy-Schwarz inequality and by averaging we get

$$\mathbb{E}\Psi_k^2 \leq 2\mathbb{E}(g_k(1))^2. \tag{74}$$

Further, from the convexity of the partition function we have also

$$\frac{d^2}{dt^2}g_k(t) \geq 0. \tag{75}$$

Exploiting the last expression, along with the fact that $g_k(0) = 0$, we can get

$$g_k'(0) \leq g_k(1) \leq g_k'(1). \tag{76}$$

The left inequality in (76) is obtained by a Taylor expansion of $g_k(1)$ around $t = 0$, that is

$$g_k(1) = g_k(0) + g_k'(0) + g_k''(\xi)$$

where $0 \leq \xi \leq 1$, from which it follows that $g_k(1) \geq g_k'(0)$ since $g_k(0) = 0$ and $g_k''(\xi) \geq 0$. Analogously, for the right inequality in (76) we Taylor expand $g_k(0)$ around $t = 1$, that is

$$g_k(0) = g_k(1) + (0-1)g_k'(1) + g_k''(\xi),$$

where $0 \leq \xi \leq 1$, thus

$$g_k(1) = g_k'(1) - g_k''(\xi)$$

from which it follows that $g_k(1) \leq g_k'(1)$ since $g_k''(\xi) \geq 0$.
In the following we will use the right inequality of (76) and, as for $g_k'(1)$, we notice that it is just the thermal average of the term R_k and, exploiting the fact that the ξ are uncorrelated and that P/N is finite in the thermodynamic limit, one can state that

$$\frac{\mathbb{E}\Psi_k^2}{2} \leq \mathbb{E}\left(g_k(1)^2\right) \leq \mathbb{E}\left(g_k'(1)^2\right) \leq C. \tag{77}$$

Now, recalling eqs. (69)-(70) and applying the bound in eq. (77), we finally get

$$\mathbb{E}\left(A_{P,N} - \mathbb{E}\left(A_{P,N}\right)\right)^2 = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Psi_k^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}C = \frac{NC}{N^2} \to 0. \tag{78}$$

□

It is worth mentioning that there are other techniques which may be used to prove Theorem 1, such as Talagrand's techniques on concentration of measure [62]; a clear outline on this perspective can be found in [21].

Appendix B. Proof of Proposition 2, Theorem 3 and Corollary 1

Proof The proof works in five stages. First (*i*) we evaluate explicitly the Cauchy condition $\phi_{P,N}(\beta,n,t=0)=\frac{1}{Nn}\ln\mathbb{E}\left(Z_0^n\right)$. Next (*ii*) we calculate $d\phi_{P,N}(\beta,n,t)/dt$, which we integrate back for $t\in[0,1]$; in this calculation (*iii*) we will explicitly make use of the assumption of replica symmetry. Then (*iv*), we construct the n-quenched replica symmetric free energy and we extremize the latter over the order parameters to get the related self-consistent equations. Finally (*v*), such equations are expanded to get the transition line.

i. Cauchy condition.

$$\begin{aligned}\phi_{P,N}(\beta,n,t=0) &= \frac{1}{Nn}\ln\mathbb{E}Z_0^n \qquad (79)\\ &= \frac{1}{Nn}\ln\mathbb{E}\prod_{\gamma=1}^{n}\left[\sum_{\sigma_i^\gamma}\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu^\gamma)e^{\sqrt{\alpha\beta\bar{p}}\sum_i^N J_i\sigma_i^\gamma+\sqrt{\beta\bar{q}}\sum_\mu^P J_\mu z_\mu^\gamma+\frac{\beta(1-\bar{q})}{2}\sum_\mu^P(z_\mu^\gamma)^2}\right]\\ &= \frac{1}{Nn}\ln\left\{\mathbb{E}_{J_i}\sum_{\sigma_i^\gamma}e^{\sqrt{\alpha\beta\bar{p}}\sum_i^N J_i\sigma_i^\gamma}\cdot\mathbb{E}_{J_\mu}\int_{-\infty}^{+\infty}\prod_{\mu=1}^{P}d\mu(z_\mu^\gamma)e^{\sqrt{\beta\bar{q}}\sum_\mu^P J_\mu z_\mu^\gamma+\frac{\beta(1-\bar{q})}{2}\sum_\mu^P(z_\mu^\gamma)^2}\right\}\\ &= P_1+P_2, \qquad (80)\end{aligned}$$

where

$$P_1 = \ln 2+\frac{1}{n}\ln\int_{-\infty}^{+\infty}d\mu(x)\cosh^n\left(x\sqrt{\alpha\beta\bar{p}}\right) \qquad (81)$$

$$P_2 = \frac{1}{Nn}\ln\mathbb{E}_{J_\mu}\left\{\frac{1}{[1-\beta(1-\bar{q})]^{\frac{nP}{2}}}e^{\frac{\bar{q}\beta nP}{2[1-\beta(1-\bar{q})]}J_\mu^2}\right\}=-\frac{\alpha}{2}\ln[1-\beta(1-\bar{q})]+\frac{\alpha\beta}{2}\frac{\bar{q}}{[1-\beta(1-\bar{q})]}. \qquad (82)$$

ii. t-streaming of $\phi_{P,N}(\beta,n,t)$.
The derivative of the n-quenched free energy reads as

$$\dot{\phi}_{P,N}(\beta,n,t)\equiv\frac{d\phi_{P,N}(\beta,n,t)}{dt}=\frac{d}{dt}\frac{1}{Nn}\ln\mathbb{E}\left(Z_t^n\right)=\frac{1}{N}\frac{1}{\mathbb{E}\left(Z_t^n\right)}\mathbb{E}\left(Z_t^n\frac{\dot{Z}_t}{Z_t}\right) \qquad (83)$$

and, in particular,

$$\begin{aligned}\frac{\dot{Z}_t}{Z_t} &\equiv \frac{1}{Z_t}\frac{dZ_t}{dt}\\ &= \frac{1}{Z_t}\frac{d}{dt}\sum_\sigma\int_{-\infty}^{+\infty}\prod_\mu^P d\mu(z_\mu)e^{\sqrt{t}\left(\frac{\sqrt{\beta}}{\sqrt{N}}\sum_{i,\mu}^{N,P}\xi_i^\mu\sigma_i z_\mu\right)+\sqrt{1-t}\left(\sqrt{\alpha\beta\bar{p}}\sum_i^N J_i\sigma_i+\sqrt{\beta\bar{q}}\sum_\mu^P J_\mu z_\mu\right)+(1-t)\beta(1-\bar{q})\sum_\mu^P\frac{z_\mu^2}{2}}\\ &= \tilde{\mathscr{A}}+\tilde{\mathscr{B}}+\tilde{\mathscr{C}}+\tilde{\mathscr{D}}, \qquad (84)\end{aligned}$$

where

$$\tilde{\mathscr{A}} = \frac{\sqrt{\beta}}{2\sqrt{Nt}}\sum_{i,\mu}^{N,P}\xi_i^\mu\omega(\sigma_i z_\mu), \qquad (85)$$

$$\tilde{\mathscr{B}} = -\frac{\sqrt{\alpha\beta\bar{p}}}{2\sqrt{1-t}}\sum_i^N J_i\omega(\sigma_i), \qquad (86)$$

$$\tilde{\mathscr{C}} = -\frac{\sqrt{\beta\bar{q}}}{2\sqrt{1-t}}\sum_\mu^P J_\mu\omega(z_\mu), \qquad (87)$$

$$\tilde{\mathscr{D}} = -\frac{\beta}{2}(1-\bar{q})\sum_\mu^P z_\mu^2. \qquad (88)$$

Thus, posing

$$\mathscr{A} = \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\mathbb{E}\left(Z_t^n \tilde{\mathscr{A}}\right), \tag{89}$$

$$\mathscr{B} = \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\mathbb{E}\left(Z_t^n \tilde{\mathscr{B}}\right), \tag{90}$$

$$\mathscr{C} = \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\mathbb{E}\left(Z_t^n \tilde{\mathscr{C}}\right), \tag{91}$$

$$\mathscr{D} = \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\mathbb{E}\left(Z_t^n \tilde{\mathscr{D}}\right). \tag{92}$$

we can recast $\dot{\phi}_{P,N}(\beta,n,t)$ as

$$\dot{\phi}_{P,N}(\beta,n,t) = \mathscr{A} + \mathscr{B} + \mathscr{C} + \mathscr{D}. \tag{93}$$

Let us start with the evaluation of $\mathscr{A}$:

$$\begin{aligned}
\mathscr{A} &= \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\mathbb{E}\left(Z_t^n \frac{\sqrt{\beta}}{2\sqrt{Nt}}\sum_{i,\mu}^{N,P}\xi_i^\mu \omega(\sigma_i z_\mu)\right) && (94)\\
&= \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\frac{\sqrt{\beta}}{2\sqrt{tN}}\sum_{i,\mu}^{N,P}\mathbb{E}\xi_i^\mu Z_t^n \omega(\sigma_i z_\mu) && (95)\\
&= \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\frac{\sqrt{\beta}}{2\sqrt{tN}}\sum_{i,\mu}^{N,P}\mathbb{E}\partial_{\xi_i^\mu}\left(Z_t^n \omega(\sigma_i z_\mu)\right) && (96)\\
&= \frac{1}{N}\frac{1}{\mathbb{E}(Z_t^n)}\frac{\sqrt{\beta}}{2\sqrt{tN}}\sum_{i,\mu}^{N,P}\mathbb{E}\left(nZ_t^{(n-1)}\frac{dZ_t}{d\xi_i^\mu}\omega(\sigma_i z_\mu) + Z_t^n \partial_{\xi_i^\mu}\omega(\sigma_i z_\mu)\right) && (97)\\
&= \mathscr{A}_1 + \mathscr{A}_2, && (98)
\end{aligned}$$

where

$$\begin{aligned}
\mathscr{A}_1 &= n\cdot\frac{P\beta}{2N}\langle q_{12}p_{12}\rangle_n, && (99)\\
\mathscr{A}_2 &= \frac{\beta}{2N}\sum_\mu^P\langle z_\mu^2\rangle_n - \frac{P\beta}{2N}\langle q_{12}p_{12}\rangle_n. && (100)
\end{aligned}$$

Notice that to get from (95) to (96) we applied Wick's theorem, namely $\mathbb{E}[x_i F(x)] = \sum_k A_{ik}^{-1}\mathbb{E}[\partial F(x)/\partial x_k]$, where $A_{ik}^{-1} = \mathbb{E}(x_i x_k)$, and, since our variables are i.i.d. from $\mathscr{N}(0,1)$, we have $A_{ik}^{-1} = \delta_{ik}$.
Performing analogous calculations overall we obtain

$$\begin{aligned}
\mathscr{A} &= (n-1)\frac{\alpha\beta}{2}\langle q_{12}p_{12}\rangle_n + \frac{\alpha\beta}{2}\frac{1}{P}\sum_\mu^P\langle z_\mu^2\rangle_n, && (101)\\
\mathscr{B} &= -(n-1)\frac{\alpha\beta}{2}\bar{p}\langle q_{12}\rangle_n - \frac{\alpha\beta}{2}\bar{p}, && (102)\\
\mathscr{C} &= -(n-1)\frac{\alpha\beta}{2}\bar{q}\langle p_{12}\rangle_n - \frac{\alpha\beta}{2}\bar{q}\frac{1}{P}\sum_\mu^P\langle z_\mu^2\rangle_n, && (103)\\
\mathscr{D} &= -\frac{\alpha\beta}{2}(1-\bar{q})\frac{1}{P}\sum_\mu^P\langle z_\mu^2\rangle_n. && (104)
\end{aligned}$$

Putting all the terms together according to (93) and noticing that the nasty terms containing the factor $\frac{1}{P}\sum_\mu^P\langle z_\mu^2\rangle_n$ cancel out, we have

$$\dot{\phi}_{P,N}(\beta,n,t) = (n-1)\frac{\alpha\beta}{2}\langle q_{12}p_{12}\rangle_n - (n-1)\frac{\alpha\beta}{2}\bar{p}\langle q_{12}\rangle_n - \frac{\alpha\beta}{2}\bar{p} - (n-1)\frac{\alpha\beta}{2}\bar{q}\langle p_{12}\rangle_n. \tag{105}$$

iii. Integration.
Recalling that, in the replica symmetric regime and in the thermodynamic limit,

$$0 = \langle (q_{12} - \bar{q})\,(p_{12} - \bar{p}) \rangle_n,$$

we can use the above relation to verify that

$$\lim_{N\to\infty} \dot{\phi}_{P,N}(\beta, n, t) = -\frac{\alpha\beta}{2}\bar{p}\,[1 + (n-1)\bar{q}]\,. \tag{106}$$

Its integration back in t simply coincides with the multiplication by one:

$$\lim_{N\to\infty} \int_0^1 dt\, \dot{\phi}_{P,N}(\beta, n, t) = -\frac{\alpha\beta}{2}\bar{p}\,[1 + (n-1)\bar{q}] \cdot 1, \tag{107}$$

and this closes the calculations of the various contributions to $\dot{\phi}_{P,N}(\beta, n)$, in the asymptotic limit.

iv. Extremization.
Recalling eqs. (38), (42) and (44), $\phi_{P,N}(\beta, n, t = 1) = (Nn)^{-1} \log \mathbb{E}(Z^n) = \phi_{P,N}(\beta, n)$ and, by a trivial application of the Fundamental Theorem of Calculus, $\phi_{P,N}(\beta, n, t = 1) = \phi_{P,N}(\beta, n, t = 0) + \int_0^1 \dot{\phi}_{P,N}(\beta, n, t)dt$. In particular, in the thermodynamic limit, summing the Cauchy condition (79) to the r.h.s. of eq. (107) we obtain $\lim_{N\to\infty} \phi_{P,N}(\beta, n, t = 1) = \lim_{N\to\infty} \phi_{P,N}(\beta, n)$ as

$$\begin{aligned} \lim_{N\to\infty} \phi_{P,N}(\beta, n) \;&=\; \ln 2 + \frac{1}{n} \ln \int_{-\infty}^{+\infty} d\mu(x) \cosh^n \left(x\sqrt{\alpha\beta\bar{p}} \right) - \frac{\alpha}{2} \ln\left[1 - \beta(1-\bar{q})\right] \\ &+\; \frac{\alpha\beta}{2} \frac{\bar{q}}{1 - \beta(1-\bar{q})} - \frac{\alpha\beta}{2}\bar{p}\,[1 + (n-1)\bar{q}]\,. \end{aligned}$$

whose extremization with respect to $\bar{q},\ \bar{p}$ returns

$$\frac{\partial \phi_{P,N}(\beta, n)}{\partial \bar{q}} \;=\; 0 \Rightarrow \bar{p} = \frac{\beta\bar{q}}{(1-n)\left[1 - \beta\,(1-\bar{q})\right]^2}, \tag{108}$$

$$\frac{\partial \phi_{P,N}(\beta, n)}{\partial \bar{p}} \;=\; 0 \Rightarrow \bar{q} = \int d\mu(x) \tanh^2 \left(x\sqrt{\alpha\beta\bar{p}} \right). \tag{109}$$

The solution of these self-consistent equations provide the expectation of q and p in the thermodynamic limit and under the replica symmetry. □

v. Transition line. We expand (108) and (109) around $\bar{p} = \bar{q} = 0$ and, combining the two equations we get

$$\frac{\alpha\beta^2}{(1-\beta)^2(1-n)} = 1 \Rightarrow (T-1)^2 = \frac{\alpha}{1-n}, \tag{110}$$

and, with some algebra one recovers (50).

---

[1] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cognitive Sci. **9**.1:147-169, (1985).
[2] E. Agliari, et al., *Multitasking associative networks*, Phys. Rev. Lett. **109**, 268101, (2012).
[3] E. Agliari, et al., *Anergy in self-directed B-cells from a statistical mechanics perspective*, J. Theor. Biol. **375**, 2131, (2015).
[4] E. Agliari, et al., *A thermodynamical perspective of immune capabilities*, J. Theor. Biol. **267**, 48, (2011).
[5] E. Agliari, D. Migliozzi, D. Tantari, *Multipartite Hopfield models*, J. Stat. Phys. (2018).
[6] M. Aizenman, P. Contucci, *On the stability of the quenched state in mean-field spin-glass models*, J. Stat. Phys. **92**:765-783, (1998).
[7] D.J. Amit, *Modeling brain functions*, Cambridge Univ. Press (1989).
[8] P. Baldi, P. Sadowski, D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning*, Nature Commun. **5**, 12, (2014).
[9] A. Barra, G. Genovese, P. Sollich, D. Tantari, *Phase transitions in Restricted Boltzmann Machines with generic priors*, Phys. Rev. E **96**,042156 (2017).
[10] A. Barra et al., *Multi-species mean field spin glasses. Rigorous results*, Annales Henri Poincaré **16**(3), 691-708, (2015).
[11] A. Barra et al., *On the equivalence among Hopfield neural networks and restricted Boltzmann machines*, Neural Networks **34**, 1-9, (2012).
[12] A. Barra, F. Guerra, E. Mingione, *Interpolating the Sherrington-Kirkpatrick replica trick*, Philosophical Magazine **92**(1-3):78-97, (2012).

[13] A. Barra, F. Guerra, G. Genovese, D. Tantari, *How glassy are neural networks?*, JSTAT P07009, (2012).
[14] A. Barra, G. Genovese, F. Guerra, *Equilibrium statistical mechanics of bipartite spin systems*, J. Phys. A **44**, 245002, (2011).
[15] A. Barra, G. Genovese, F. Guerra, *The replica symmetric approximation of the analogical neural network*, J. Stat. Phys. **140**(4), 784-796, (2010).
[16] A. Barra, G. Genovese, P. Sollich, D. Tantari, *Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors*, Phys. Rev. E **97**, 022310 (2018).
[17] S. Bartolucci, A. Annibale, *A dynamical model of the adaptive immune system: effects of cells promiscuity, antigens and BB interactions*, JSTAT P08017, (2015).
[18] S. Bartolucci, A. Mozeika, A. Annibale, *The role of idiotypic interactions in the adaptive immune system: a belief-propagation approach*, JSTAT 083402, (2016).
[19] A. Bovier, V. Gayrard, P. Picco, *Gibbs states of the Hopfield model in the regime of perfect memory*, Prob. Theor. Rel. Fields **100**.3:329-363, (1994).
[20] A. Bovier, V. Gayrard, P. Picco, *Gibbs states of the Hopfield model with extensively many patterns*, J. Stat. Phys. **79**.1: 395-414, (1995).
[21] A. Bovier, *Statistical Mechanics of Disordered Systems. A Mathematical Perspective*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge (2006).
[22] B. Brown, *Martingale central limit theorems*, Annal Math. Stat. **42**:59-66, (1971).
[23] A.M. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. Le Cun, *The loss surfaces of multilayer networks*, Artificial Intelligence and Statistics, (2015).
[24] A.C.C. Coolen, R. Kuhn, P. Sollich, *Theory of neural information processing systems*, Oxford Press (2005).
[25] A. Decelle, F. Krzakala, C. Moore, L. Zdeborova, *Inference and phase transitions in the detection of modules in sparse networks*, Phys. Rev. Lett. **107**(6):065701, (2011).
[26] A. Decelle, G. Fissore, C. Furtlehner, *Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics*, submitted
[27] A. Decelle, G. Fissore, C. Furtlehner, *Spectral Learning of Restricted Boltzmann Machines*, Europhys. Lett. **119**(6): 60001, (2017).
[28] H. Englisch, V. Mastropietro, B. Tirozzi, *The B.A.M. storage capacity*, J. Phys. I France **5**, 85-96, (1995).
[29] G. Genovese, *Universality in Bipartite Mean Field Spin Glasses*, J. Math. Phys. **53**, 123304, (2012).
[30] S. Ghirlanda, F. Guerra, *General properties of overlap probability distributions in disordered spin systems. Towards Parisi ultrametricity*, J. Phys. A **31**, 9149-9155, (1998).
[31] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, M.I.T. press (2017).
[32] F. Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Comm. Math. Phy. **233**:1-12, (2003).
[33] F. Guerra, F.L. Toninelli, *The thermodynamic limit in mean field spin glass models*, Comm. Math. Phys. **230**.1:71-79 (2002).
[34] F. Guerra, F.L. Toninelli, *The infinite volume limit in generalized mean field disordered models*, arXiv preprint cond-mat/0208579, (2002).
[35] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the national academy of sciences 79.8 (1982): 2554-2558.
[36] J.J. Hopfield, D.W. Tank. *Neural computation of decisions in optimization problems*, Biol. Cybern. **52**(3):141-152, (1985).
[37] H. Huang, T. Toyoizumi, *Advanced mean-field theory of the restricted Boltzmann machine*, Phys. Rev. E **91**.5:050101, (2015).
[38] H. Huang, *Reconstructing the Hopfield network as an inverse Ising problem*, Phys. Rev. E **81**.3:036104, (2010).
[39] H.J. Kappen, F. Rodriguez, *Efficient learning in Boltzmann machines using linear response theory*, Neural Comput. **10**.5: 1137-1156, (1998).
[40] S. Kirkpatrick, et al., *Optimization by simulated annealing*, Science **220**:671-680, (1983).
[41] J. Korst, E. Aarts, *Combinatorial optimization on a Boltzmann machine*, J. Parall. & Distrib. Computing **6**.2:331, (1989).
[42] Y. Le Cun, Y. Bengio, G. Hinton, *Deep learning*, Nature **521**:436-444, (2015).
[43] N. Le Roux, Y. Bengio, *Representational power of restricted Boltzmann machines and deep belief networks*, Neural computation **20**.6:1631-1649, (2008).
[44] Lobo, D., Levin, M., *Inferring regulatory networks from experimental morphological phenotypes: a computational method reverse-engineers planarian regeneration*, PLoS Comput. Biol. **11**(6), e1004295, (2015).
[45] J. Martens, *Deep learning via Hessian-free optimization*, Proc. 27^{th} Int. Conf. Mach. Learn. (2010).
[46] M. Mezard, *Mean-field message-passing equations in the Hopfield model and its generalizations*, Phys. Rev. E **95**, 022117, (2017).
[47] M. Mezard, G. Parisi, M.A. Virasoro, *Spin glass theory and beyond*, World Scientific (1985).
[48] A. Mozeika, A.C.C. Coolen, *Statistical mechanics of clonal expansion in lymphocyte networks modelled with slow and fast variables*, J. Phys. A **50**.3:035602, (2016).
[49] A. Mozeika, A.C.C. Coolen, *Statistical mechanics of clonal expansion in lymphocyte networks modelled with slow and fast variables*, J. Phys. A **50**.3: 035602, (2016).
[50] H. Nishimori, *Statistical physics of spin glasses and information processing*, Clarendon Press (2001).
[51] D. Panchenko, *The Sherrington-Kirkpatrick model*, Springer, (2013).
[52] L. Pastur, M. Shcherbina, B. Tirozzi, *The replica-symmetric solution without replica trick for the Hopfield model*, J. Stat. Phys. **74**.5:1161-1183, (1994).
[53] L. Pastur, M. Shcherbina, B. Tirozzi, *On the replica symmetric equations for the Hopfield model*, J. Math. Phys. **40**.8:3930-3947, (1999).
[54] Ruelle, D., *Statistical Mechanics: Rigorous results*, W.A. Benjamin, Inc., New York, (1969).
[55] R. Salakhutdinov, A. Mnih, G. Hinton, *Restricted Boltzmann machines for collaborative filtering*, Proc. 24^{th} Int. Conf. Mach. Learn. ACM, (2007).
[56] R. Salakhutdinov, G. Hinton, *Deep boltzmann machines*, Artificial Intelligence and Statistics (2009).
[57] M. Shcherbina, B. Tirozzi, *The free energy of a class of Hopfield models*, J. Stat. Phys. **72**.1:113-125, (1993).
[58] M. Shcherbina, B. Tirozzi, *Rigorous solution of the Gardner problem*, Comm. Math. Phys. **234**.3:383-422, (2003).
[59] M. Talagrand, *Spin glasses: a challenge for mathematicians*, Springer, (2003).

[60] M. Talagrand, *Rigorous results for the Hopfield model with many patterns*, Prob. Theor. Rel. Fields **110**.2:177-275, (1998).

[61] M. Talagrand, *Exponential inequalities and convergence of moments in the replica-symmetric regime of the Hopfield model*, Ann. Prob. 1393-1469, (2000).

[62] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Publications Mathématiques de l'Institut des Hautes Études Scientifiques, **81**(1):73205 (1995)

[63] B. Tirozzi, *Modelli matematici di reti neurali*, CEDAM (1995).

[64] E.W. Tramel, A. Dremeau, F. Krzakala, *Approximate message passing with restricted Boltzmann machine priors*, JSTAT 073401, (2016).

[65] J. Tubiana, R. Monasson, *Emergence of Compositional Representations in Restricted Boltzmann Machines*, Phys. Rev. Lett. **118**.13:138301, (2017).

[66] The symmetry of the interactions between neurons ensures the detailed balance which, in turn, ensures the relaxation to a Gibbs measure.

[67] Actually, in order to optimize the training stage, one should also properly set the internal parameters of the machine such as the ratio between the sizes of the visible and hidden layer, the kind of the neurons, etc. [9, 31].

[68] The time scale for neuronal spikes is order of 50 ms, while the time scale for synaptic rearrangement is order of hours and it takes order of weeks to consolidate.

[69] This argument can be expanded up to arbitrarily N-points correlation functions by paying the price of adding extra hidden layers and this kind of extension is a basic principle underlying Deep Learning [42].

[70] In the neural network jargon with "retrieval" we mean that the network is supposed to have learnt a set of data (also called "patterns"), say a set of pictures, and now it is given, as input, one of those pictures but corrupted or incomplete. If well-performing, the network has to provide, as output, a picture that is as close as possible, to the original one.

[71] The capability to accomplish pattern recognition, hence to perform as an associative memory, depends on several factors as, e.g., the amount of noise in the data, the amount of data with respect to the amount of neurons to deal with them, etc. [7].

[72] This picture is rather intuitive in the low-load regime (e.g., when P remains finite in the thermodynamic limit), while, in the high-load regime, one should combine a number of Curie-Weiss models that grows linearly with the system size and the landscape exhibits a glassy component.

[73] It is worth comparing the expression for the free-energy reported in Proposition 1 and the one obtained in [15] for an Hopfield model in the high-load regime. The latter displays an additional term, that is $A^A(\alpha,\beta) = \ln 2 - \frac{\alpha}{2}\ln(1-\beta) - \alpha\beta/2$ which stems from the diagonal terms $(P/2)$ in the Hamiltonian $H_N(\sigma|\xi) = -\frac{1}{N}\sum_{i<j}\sum_\mu \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j = -\frac{1}{2N}\sum_{i,j}\sum_\mu \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j + \frac{P}{2}$, which in this work are neglected.