The *relativistic* Hopfield network: Rigorous results

Elena Agliari (iD), Adriano Barra, and Matteo Notarnicola

View Online   Export Citation   CrossMark
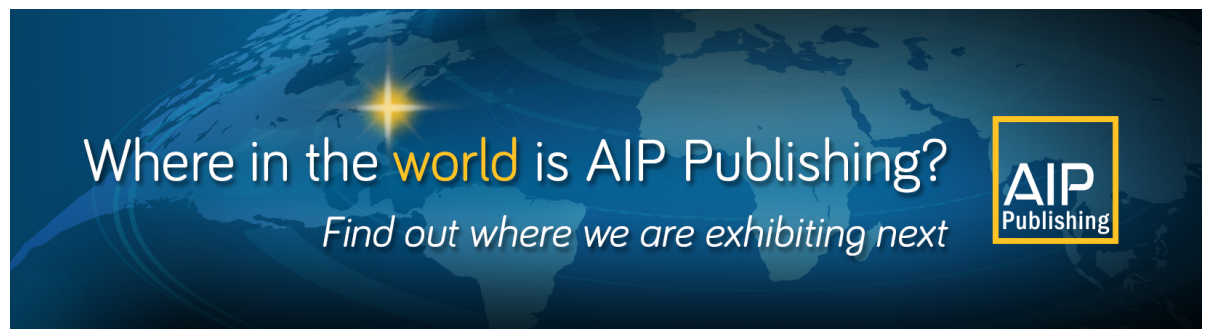
The *relativistic* Hopfield network: Rigorous results

View Online   Export Citation   CrossMark

Elena Agliari,[1] (iD) Adriano Barra,[2] and Matteo Notarnicola[1,2]

AFFILIATIONS

[1] Dipartimento di Matematica, Sapienza Università di Roma, Rome, Italy
[2] Dipartimento di Matematica e Fisica Ennio De Giorgi, Università del Salento, Lecce, Italy

ABSTRACT

The relativistic Hopfield model constitutes a generalization of the standard Hopfield model that is derived by the formal analogy between the statistical-mechanic framework embedding neural networks and the Lagrangian mechanics describing a fictitious single-particle motion in the space of the tuneable parameters of the network itself. In this analogy, the cost-function of the Hopfield model plays as the standard kinetic-energy term and its related Mattis overlap (naturally bounded by one) plays as the velocity. The Hamiltonian of the relativisitc model, once Taylor-expanded, results in a *p*-spin series with alternate signs: the attractive contributions enhance the information-storage capabilities of the network, while the repulsive contributions allow for an easier unlearning of spurious states, conferring overall more robustness to the system as a whole. Here, we do not deepen the information processing skills of this generalized Hopfield network, rather we focus on its statistical mechanical foundation. In particular, relying on Guerra's interpolation techniques, we prove the existence of the infinite-volume limit for the model free-energy and we give its explicit expression in terms of the Mattis overlaps. By extremizing the free energy over the latter, we get the generalized self-consistent equations for these overlaps as well as a picture of criticality that is further corroborated by a fluctuation analysis. These findings are in full agreement with the available previous results.

Published under license by AIP Publishing. https://doi.org/10.1063/1.5077060

I. INTRODUCTION: A GLANCE AT THE *RELATIVISTIC* HOPFIELD NETWORK

Recent advances, in hardware [mainly due to the novel generation of GPU (graphics processing unit) computing architectures over the standard CPU (central processing unit) ones[30]] as well as in software (mainly due to the novel generation of algorithmic prescriptions overall termed *Deep Learning*[19]), have made neural networks pervasive in every-day life, and this obviously raised the quest for stronger mathematical frameworks where these algorithms and models can be suitably analyzed, controlled, or even *understood*.[27]

To this goal, statistical mechanics has proved to be a particularly convenient tool, as evidenced by the seminal work by Hopfield[23] and the successive analysis carried out by Amit-Gutfreund-Somponlinsky,[7,16,17] and it will constitute the main tool exploited in this paper too.

In the remaining of this introductory section, we briefly revise the classical and the relativistic Hopfield models, pointing out their capabilities as neural networks within a statistical mechanical setting. Then, in the next two sections, we focus on the relativistic generalization, and, in particular, in Sec. II, we prove the existence of the thermodynamic limit for its free energy,[36] while in Sec. III, we give an explicit expression for such a quantity (that turns out to coincide sharply with the one obtained with mechanical techniques in Ref. 10) in terms of the natural order parameters of the theory, namely, the Mattis overlaps; moreover, by extremizing the free energy over the Mattis overlaps, we obtain the related self-consistent expressions, and by studying the fluctuations of the (rescaled and centered) Mattis overlaps, we inspect ergodicity breaking and the critical behavior of the system. Finally, Sec. IV is left for our conclusions and outlooks.

A. The *classical* Hopfield network in a nutshell

Consider N Boolean (i.e., Ising) spins (or *neurons*) $\sigma_i = \pm 1$, $i \in (1, \ldots, N)$ and P patterns ξ^μ, $\mu \in (1, \ldots, P)$ of length N whose entries are identically and independently drawn from

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

Note that we are taking the pattern entries completely at random: this choice may sound poorly realistic, yet it ensures, by a Shannon compression argument, that if the network is able to deal with P entirely random patterns, it will be likewise able to deal with, at least, P structured ones.[37]

Definition 1. *The cost function (or Hamiltonian to keep a physical jargon) of the classical Hopfield model can be written as*

$$H_N^{(c)}(\boldsymbol{\sigma}|\boldsymbol{\xi}) := -\frac{1}{2N} \sum_{i,j}^{N,N} \left(\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \right) \sigma_i \sigma_j. \tag{2}$$

In order to have an intuition about the *spontaneous associative memory* capabilities of this model, one quantifies the retrieval pertaining to each pattern and this is accomplished by the following order parameters.

Definition 2. *For each stored pattern of information ξ^μ, we introduce the Mattis overlap m_μ defined as*

$$m_\mu := \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i \in [-1 + 1]. \tag{3}$$

Notice that the Hopfield Hamiltonian can then be written as $H_N^{(c)}(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -N \sum_{\mu=1}^{P} m_\mu^2$. This expression highlights that, if the neural configuration $\boldsymbol{\sigma}$ is uncorrelated with any of the P patterns, the scalar product of the vector state $(\sigma_1, \sigma_2, \ldots, \sigma_N)$ with any of these patterns, say, $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu)$, would vanish as $N^{-1/2}$; hence, its contribution to lower the cost function (namely, the *energy of the system*) would be rather marginal; on the other hand, if the neural configuration is highly correlated with one of the patterns, then its corresponding Mattis overlap would be $O(1)$ and this would significantly decrease the energy of the system: as the Hamiltonian is parabolic in the Mattis overlap, this observation candidates the P patterns to act as *attractors* for any reasonable network's dynamics.[16] Therefore, if the network is fed with partial information concerning one pattern (for instance, a corrupted pattern is presented to the network[38]), it will autonomously denoise the supplied information and find out the correct (pure) reference pattern (if the noise level is not too high[7]). It is worth noticing that, according to (1), patterns become orthogonal in the infinite-volume limit in such a way that the system can retrieve patterns only sequentially (see Refs. 1, 3, and 34 for examples of neural networks able to perform in a parallel way).

An extensive mathematical formalization of these statements requires the introduction of several concepts and goes beyond the scope of the present paper: we refer to excellent textbooks for deepening the associative memory capabilities of the Hopfield network (see, e.g., Refs. 7, 16, and 17) and to further readings for understanding the mathematical complexity behind these models.[4,12–15,32,33]

An important remark is that, once introduced the *storage* of the network as the limiting ratio $\alpha := \lim_{N \to \infty} P/N$, we distinguish between two different regimes: the low-storage case, where $\alpha = 0$, and the high-storage case, where $\alpha > 0$. As for the latter, at a rigorous level, several questions still need to be answered (e.g., a proof of the existence of the infinite-volume limit for its free energy at present is still lacking, not to mention a full, clear replica-symmetry-breaking picture for this type of *spin glass*); thus, generalizations of the high-storage case may be still premature. In fact, the relativistic generalization of the Hopfield model was introduced as a low-storage model in Ref. 10, and here, we provide a rigorous treatment of the generalization still at $\alpha = 0$.

B. The *relativistic* Hopfield network in a nutshell

As discussed and investigated in Refs. 2 and 5, the analysis of the Hopfield model (and suitable extensions) can be mapped into a mechanical problem. More precisely, the free-energy associated with the model can be shown to fulfill a Hamilton-Jacobi equation describing a fictitious single-particle motion in a P-dimensional space, where the Mattis overlap plays as the particle velocity, which therefore displays an intrinsic bound ($m_\mu \leq 1, \forall \mu$). When studying the associative memory capabilities of the Hopfield model, one is particularly interested in regimes where the velocity is approaching its upper bound in such a way that the correct framework to embed the problem is the relativist one and the related generalized Hopfield model is described hereafter.

Consider N Boolean (i.e., Ising) spins (or *neurons*) $\sigma_i = \pm 1$, $i \in (1, \ldots, N)$ and P patterns ξ^μ, $\mu \in (1, \ldots, P)$ of length N whose entries are independently and identically drawn from

$$P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = 1/2.$$

Definition 3. *The cost function (or Hamiltonian to keep a physical jargon) of the relativistic Hopfield model can be written as*

$$H_N^{(r)}(\boldsymbol{\sigma}|\boldsymbol{\xi}) := -N \sqrt{1 + \sum_{\mu=1}^{P} m_\mu^2} := -N\sqrt{1 + \mathbf{m}^2}. \tag{4}$$

Note that this cost function (4) can be expanded in an alternate-sign series as

$$\frac{H_N^{(r)}(\boldsymbol{\sigma}|\boldsymbol{\xi})}{N} \sim -1 - \frac{1}{2N}\sum_{ij}^{N,N}(\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j)\sigma_i\sigma_j + \frac{1}{8\,N^4}\sum_{i,j,k,l}^{N,N,N,N}(\boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_j)(\boldsymbol{\xi}_k \cdot \boldsymbol{\xi}_l)\sigma_i\sigma_j\sigma_k\sigma_l + O(\sigma^6). \tag{5}$$

Thus, focusing on the attractive contributions (beyond the classical pairwise model $p = 2$), it is enriched by p-spin terms (with $p = 6, 10, \ldots$) that yield to further synaptic couplings where information can be stored (as recently suggested by Hopfield himself[26]), while focusing on the repulsive contributions, it also displays p-spin terms (with $p = 4, 8, \ldots$) that favour network's pruning (as suggested, in the past, by Hopfield himself and several other authors[18,24,28,29,35] to erase spurious states).

The analysis of the information processing skills of this network has been accomplished elsewhere;[6,10] we summarize it by Fig. 1, referring to the original papers for further algorithmic details, while hereafter, we deepen the mathematical aspects of its statistical mechanical foundation.

C. The statistical mechanical setting

Having defined a suitable Hamiltonian playing as a cost function [see (2) and (4)], the next step in the statistical-mechanics approach requires the introduction of a parameter $\beta \in \mathbb{R}^+$ to account for the noise level in the system (i.e., the *inverse temperature* in physical jargon) in such a way that for $\beta \to 0$, the network evolution is completely random, while in the $\beta \to \infty$ limit, its evolution becomes deterministic (and the Hamiltonian plays as a Lyapounov function[16]). Then, we can state the next.

Definition 4. *Defined $\mathbb{E}$ as the average over the patterns, we introduce the partition function $Z_N^{(c,\,r)}(\beta, \boldsymbol{\xi})$ and the intensive free energy[39] $\alpha_N^{(c,\,r)}(\beta)$ as*

$$Z_N^{(c,\,r)}(\beta, \boldsymbol{\xi}) := \sum_{\{\sigma\}}^{2^N} e^{-\beta H_N^{(c,\,r)}(\boldsymbol{\sigma}|\boldsymbol{\xi})}, \tag{6}$$

$$\alpha_N^{(c,\,r)}(\beta) := \frac{1}{N}\mathbb{E}\ln Z_N^{(c,\,r)}(\beta, \boldsymbol{\xi}). \tag{7}$$

Definition 5. *Note that $\exp[-\beta H_N^{(c,\,r)}(\boldsymbol{\sigma}|\boldsymbol{\xi})]/Z^{(c,\,r)}(\beta, \boldsymbol{\xi})$ defines a probability measure, whose average (Boltzmann average) we indicate with $\langle \cdot \rangle_{(c,r)}$, i.e.,*

$$\langle f(\boldsymbol{\sigma})\rangle_{(c,\,r)} := \frac{\sum_{\{\sigma\}}^{2^N} f(\boldsymbol{\sigma})e^{-\beta H_N^{(c,\,r)}(\boldsymbol{\sigma}|\boldsymbol{\xi})}}{Z_N^{(c,\,r)}(\beta, \boldsymbol{\xi})}. \tag{8}$$
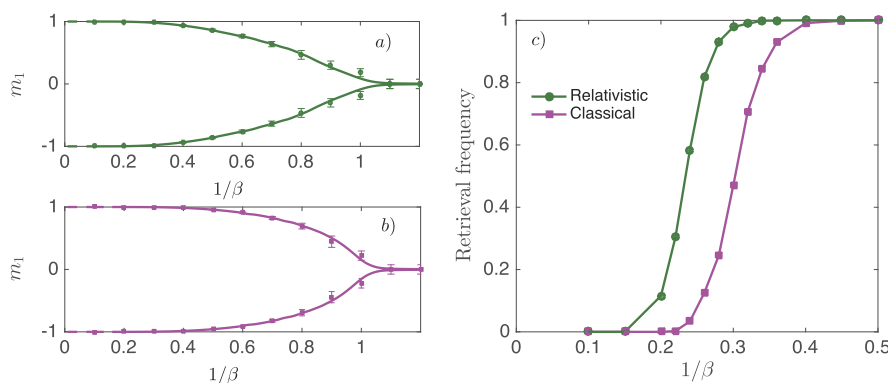

FIG. 1. Left: Mattis overlap m_1 related to the retrieved pattern versus the noise level $1/\beta$ for the relativistic [panel (a)] and the classical [panel (b)] Hopfield model. Symbols represent Monte Carlo runs for a network made of $N = 400$ neurons and $P = 3$ patterns, while lines represent the numerical solution of the self-consistencies, namely, Eqs. (28) and (11). Right [panel (c)]: Numerical comparison of the *retrieval frequency*, defined as the fraction of the runs that (starting from a spurious state) end up into a pure state attractor, versus the noise level in the network for the relativistic (◯) and classical (□) Hopfield model (again, $N = 400$, $P = 3$). Notice that, in both cases, at low temperatures, the spurious states are stable and that for $\beta^{-1} > 0.15$ the relativistic model (respectively, $\beta^{-1} > 0.22$ for the classical model) starts to retrieve and at $\beta^{-1} = 0.25$ (respectively, $\beta^{-1} = 0.32$) roughly half of the runs end up into a pure state.

Remark 1. In order to highlight when observables are evaluated in the thermodynamic limit, we omit their subscript N. For instance, for the free-energy, we write

$$\alpha^{(c,r)}(\beta) := \lim_{N \to \infty} \alpha_N^{(c,r)}(\beta). \tag{9}$$

For the sake of completeness, we recall that for the standard Hopfield model in the infinite-volume limit, the free energy and the related self-consistency equations for the P Mattis overlaps, respectively, read as[7,16,17]

$$\alpha^{(c)}(\beta) = \ln 2 + \langle \ln \cosh(\beta \boldsymbol{\xi} \cdot \langle \mathbf{m} \rangle_{(c)}) \rangle_{\xi} + \frac{\beta}{2} \langle \mathbf{m}^2 \rangle_{(c)}, \tag{10}$$

$$\langle m_\mu \rangle_{(c)} = \langle \xi^\mu \tanh(\beta \boldsymbol{\xi} \cdot \langle \mathbf{m} \rangle_{(c)}) \rangle_{\xi}, \tag{11}$$

with the abbreviation $\langle f(\boldsymbol{\xi}) \rangle_\xi = N^{-1} \sum_i f(\xi_i)$.

In the following, we will focus solely on the relativistic generalization; thus, in order to lighten the notation, we will drop the super- and sub-scripts (r).

II. EXISTENCE OF THE THERMODYNAMIC LIMIT OF THE FREE ENERGY

In this section, we prove the existence of $\alpha(\beta)$ for the relativistic Hopfield model. The underlying idea is to adapt the Guerra-Toninelli scheme, originally developed for a quadratic cost-function, to the Hamiltonian (4) featuring a square root. Following the standard scheme, we consider a system made of N neurons and two other, independent, systems made of N_1 and N_2 neurons such that $N = N_1 + N_2$. Then, we need to prove that the extensive free energy of the former is strictly smaller than the sum of those pertaining to the two subsystems (hence, we have its sub-additivity) so that applying the Fekete Lemma,[31] we get the result.[22] A way to compare the free energies of these two limiting cases is by interpolation, namely, defining an interpolating free energy whose extrema reproduce the free energies in the two cases of interest and then showing that its derivative—with respect to the interpolating parameter—has a semi-definite sign. Here, the main adaptation will consist in a proof by reduction to absurd assuming the free energy to be super-additive (see the Appendix for the definitions of sub-super-additive successions and functions and for the main statement of Fekete Lemma).

For the sake of simplicity (and without loss of generality), we fix $\beta = 1$ just in this Section.

Let us consider, beyond the Mattis overlaps related to the P patterns of the original N-neuron's model, also two further Mattis overlaps related to the two aforementioned smaller systems made of, respectively, N_1 and N_2 neurons. We can write

$$\mathbf{m}_{N_1} := \frac{1}{N_1} \sum_{i=1}^{N_1} \xi_i \sigma_i, \quad \mathbf{m}_{N_2} := \frac{1}{N_2} \sum_{i=1}^{N_2} \xi_i \sigma_i$$

and notice that

$$\mathbf{m}_N = \frac{1}{N} \sum_{i=1}^{N} \xi_i \sigma_i = \frac{1}{N} \sum_{i=1}^{N_1+N_2} \xi_i \sigma_i = \frac{1}{N} \left(\sum_{i=1}^{N_1} \xi_i \sigma_i + \sum_{j=1}^{N_2} \xi_j \sigma_j \right)$$

$$= \frac{1}{N} (N_1 \mathbf{m}_{N_1} + N_2 \mathbf{m}_{N_2}) = \rho_1 \mathbf{m}_{N_1} + \rho_2 \mathbf{m}_{N_2},$$

once introduced the relative densities ρ_1, ρ_2 as

$$\rho_1 := \frac{N_1}{N} \quad \text{and} \quad \rho_2 := \frac{N_2}{N}.$$

Definition 6. Let us introduce an interpolating parameter $t \in [0, 1]$ that we use to define an interpolating free energy $\alpha_N(\beta, t)$ as follows:

$$\alpha_N(\beta, t) := \frac{1}{N} \mathbb{E} \ln \sum_{\{\sigma\}}^{2^N} \exp\left[tN\sqrt{1 + \mathbf{m}_N^2} + (1-t)\left(N_1\sqrt{1 + \mathbf{m}_{N_1}^2} + N_2\sqrt{1 + \mathbf{m}_{N_2}^2} \right) \right]. \tag{12}$$

It is crucial to observe that in the two limits of $t \to 1$ and $t \to 0$, we recover, respectively,

$$\alpha_N(\beta, 1) = \frac{1}{N} \mathbb{E} \ln \sum_{\{\sigma\}}^{2^N} \exp\left(N\sqrt{1 + \mathbf{m}_N^2} \right) = \alpha_N(\beta), \tag{13}$$

while

$$\alpha_N(\beta, 0) = \frac{1}{N} \mathbb{E} \ln \sum_{\{\sigma\}}^{2^N} \exp\left(N_1 \sqrt{1 + \mathbf{m}_{N_1}^2} + N_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \right)$$

$$= \frac{1}{N} \mathbb{E} \ln(Z_{N_1} Z_{N_2}) = \frac{1}{N} \mathbb{E}(\ln Z_{N_1} + \ln Z_{N_2})$$

$$= \rho_1 \alpha_{N_1}(\beta) + \rho_2 \alpha_{N_2}(\beta). \tag{14}$$

Therefore, by varying t, we interpolate between the original system and the sum of the two smaller subsystems, properly weighted by their relative densities. By the fundamental theorem of calculus, we can write

$$\alpha_N(\beta, 1) = \alpha_N(\beta, 0) + \int_0^1 \frac{d\alpha_N(\beta, t)}{dt} dt, \tag{15}$$

and since the *integral* operator is monotonous (and therefore it must respect inequalities), it will be sufficient to prove that the derivative of the interpolating free energy with respect to t, i.e., $\dot{\alpha}_N(\beta, t) := d_t \alpha_N(\beta, t)$, has a negative semi-definite sign.

Denoting with $\tilde{Z}$ the interpolating partition function coupled to the interpolating free energy (12), we can write

$$\frac{d\alpha_N(\beta, t)}{dt} = \frac{\partial}{\partial t} \frac{1}{N} \mathbb{E} \ln \sum_{\{\sigma\}}^{2^N} \exp\left[tN\sqrt{1 + \mathbf{m}_N^2} + (1 - t)\left(N_1 \sqrt{1 + \mathbf{m}_{N_1}^2} + N_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \right) \right]$$

$$= \frac{1}{N} \sum_{\{\sigma\}}^{2^N} \mathbb{E} \frac{1}{\tilde{Z}} \left(N\sqrt{1 + \mathbf{m}_N^2} - N_1 \sqrt{1 + \mathbf{m}_{N_1}^2} + N_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \right) \exp(\phi), \tag{16}$$

where we called

$$\phi := tN\sqrt{1 + \mathbf{m}_N^2} + (1 - t)\left(N_1 \sqrt{1 + \mathbf{m}_{N_1}^2} + N_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \right).$$

Remark 2. We observe that Eq. (12) allows us to generalize the partition function $Z(\beta, \xi)$ to $Z(\beta, \xi, t)$ such that Eq. (16) can be written in terms of its t-generalized Boltzmann average $\langle \cdot \rangle_t$ as

$$\frac{d\alpha_N(\beta, t)}{dt} = \left\langle \sqrt{1 + \mathbf{m}_N^2} - \rho_1 \sqrt{1 + \mathbf{m}_{N_1}^2} - \rho_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \right\rangle_t. \tag{17}$$

Now, we must prove that the expression averaged in the r.h.s. of Eq. (17) is less or equal to zero, as stated by the next.

Proposition 1. The t-derivative of the interpolating free energy (12) is semi-definite negative, namely,

$$\sqrt{1 + \mathbf{m}_N^2} - \rho_1 \sqrt{1 + \mathbf{m}_{N_1}^2} - \rho_2 \sqrt{1 + \mathbf{m}_{N_2}^2} \le 0. \tag{18}$$

Proof. We first observe that, as $\rho_2 = 1 - \rho_1 = 1 - \rho$, the parameters effectively involved in Eq. (17) are solely $\mathbf{m}_{N_1}$, $\mathbf{m}_{N_2}$, and ρ_1, which from now on we will call simply ρ, in such a way that $\mathbf{m}_N = \rho_1 \mathbf{m}_{N_1} + \rho_2 \mathbf{m}_{N_2} = \rho \mathbf{m}_{N_1} + (1 - \rho)\mathbf{m}_{N_2}$. Then, we can write the l.h.s. of Eq. (18) as $\sqrt{1 + (\rho \mathbf{m}_{N_1} + (1 - \rho)\mathbf{m}_{N_2})^2} - \rho\sqrt{1 + \mathbf{m}_{N_1}^2} - (1 - \rho)\sqrt{1 + \mathbf{m}_{N_2}^2}$. Let us now suppose, by contradiction, that

$$\sqrt{1 + (\rho \mathbf{m}_{N_1} + (1 - \rho)\mathbf{m}_{N_2})^2} - \rho\sqrt{1 + \mathbf{m}_{N_1}^2} - (1 - \rho)\sqrt{1 + \mathbf{m}_{N_2}^2} > 0.$$

With some algebra, we obtain

$$2\rho(1 - \rho) > 2\rho(1 - \rho)\left(\sqrt{\left(1 + \mathbf{m}_{N_1}^2\right)\left(1 + \mathbf{m}_{N_2}^2\right)} - \mathbf{m}_{N_1}\mathbf{m}_{N_2} \right),$$

and as $2\rho(1 - \rho) > 0$, we have $1 > \sqrt{\left(1 + \mathbf{m}_{N_1}^2\right)\left(1 + \mathbf{m}_{N_2}^2\right)} - \mathbf{m}_{N_1}\mathbf{m}_{N_2}$, whence

$$\sqrt{\left(1 + \mathbf{m}_{N_1}^2\right)\left(1 + \mathbf{m}_{N_2}^2\right)} < 1 + \mathbf{m}_{N_1}\mathbf{m}_{N_2}.$$

The r.h.s. term of the above expression is certainly non-negative (as $\mathbf{m}_{N_1}, \mathbf{m}_{N_2} \in [-1, 1]$): even in the worst scenario where $\mathbf{m}_{N_1}$ and $\mathbf{m}_{N_2}$ have opposite signs, their product will never be smaller than -1; hence, by quadrature, the inequality remains unchanged. We are then allowed to write

$$\left(1 + \mathbf{m}_{N_1}^2\right)\left(1 + \mathbf{m}_{N_2}^2\right) < \left(1 + \mathbf{m}_{N_1}\mathbf{m}_{N_2}\right)^2$$

by which we get

$$\left(\mathbf{m}_{N_2}^2 - \mathbf{m}_{N_1}^2\right)^2 < 0$$

that is obviously wrong. □

Remark 3. We also highlight that the function $f : x \mapsto \sqrt{1 + x^2}$ is convex, namely, for $\lambda \in [0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

If we identify

$$\lambda := \rho, \qquad x_1 = \mathbf{m}_{N_1}, \qquad x_2 = \mathbf{m}_{N_2}$$

as $\mathbf{m}_N = \rho\mathbf{m}_{N_1} + (1 - \rho)\mathbf{m}_{N_2}$, then

$$\sqrt{1 + \mathbf{m}_N^2} = \sqrt{1 + (\rho_1\mathbf{m}_{N_1} + \rho_2\mathbf{m}_{N_2})^2} \leq \rho_1\sqrt{1 + \mathbf{m}_{N_1}^2} + \rho_2\sqrt{1 + \mathbf{m}_{N_2}^2}.$$

Proposition 1 allows us to state that

$$\alpha_N(\beta, 1) - \alpha_N(\beta, 0) = \int_0^1 \frac{\partial \alpha_N(\beta, t)}{\partial t} dt \leq 0,$$

namely,

$$N\alpha_N(\beta) \leq N_1\alpha_{N_1}(\beta) + N_2\alpha_{N_2}(\beta), \tag{19}$$

such that we can finally state the next.

Theorem 1. *The infinite-volume limit of the intensive free energy defined by the relativistic Hopfield cost function* (4) *exists, and it equals its infimum, that is,*

$$\exists \lim_{N \to \infty} \alpha_N(\beta) = \inf_{N \in \mathbb{N}}\{\alpha_N(\beta)\} = \alpha(\beta).$$

Proof. The proof is a straightforward consequence of Proposition 1 coupled to the Fekete Lemma. □

III. EXPRESSION OF THE THERMODYNAMIC LIMIT OF THE FREE ENERGY

In this section, we give an explicit expression for the infinite-volume limit of the free energy related to the relativistic Hopfield cost-function (4) in terms of the Mattis overlaps. Again, we use a suitable Guerra's interpolation scheme in such a way that one extremum of the interpolation recovers the original model to be solved and the other recovers a case whose solution is straightforward. In particular, as we know that all the mean-field models (even the generalized ones) have a product space structure, namely, their probability distribution $P(\sigma_1, \sigma_2, \ldots, \sigma_N)$ in the thermodynamic limit factorizes (i.e., $\lim_{N \to \infty} P(\sigma_1, \sigma_2, \ldots, \sigma_N) = \prod_{i=1}^N P(\sigma_i)$), we can use this information to construct the "easy" extremum of the interpolation. In other words,

Definition 7. Being the scalar $t \in [0, 1]$ an interpolating parameter, we define a novel interpolating free energy $\alpha_N(\beta, t)$ as

$$\alpha_N(\beta, t) := \frac{1}{N}\mathbb{E}\ln\sum_{\{\sigma\}}^{2^N} \exp\left(t\beta N\sqrt{1 + \sum_{\mu=1}^P m_\mu^2} + (1 - t)\beta \sum_{\mu=1}^P \psi^\mu \sum_{i=1}^N \xi_i^\mu \sigma_i\right), \tag{20}$$

where ψ^μ, $\mu \in (1, \ldots, P)$, are P fields that are functions of the patterns.

It is worth stressing that each neuron experiences the simultaneous action of P fields $\{\psi^\mu\}_{\mu=1,\ldots,P}$ linearly combined, unlike the interpolating structure working in the ferromagnetic case where each spin just experiences the action of one single field.[8] The particular choice of the fields ψ^μ will be discussed later.

Remark 4. Given a smooth function $F(\boldsymbol{\sigma}|\boldsymbol{\xi})$ of the neurons and of the patterns, we observe that the interpolating free-energy (20) implicitly defines the extended averages $\langle F(\boldsymbol{\sigma}|\boldsymbol{\xi})\rangle_t$ as

$$\langle F(\boldsymbol{\sigma}|\boldsymbol{\xi})\rangle_t := \frac{\sum_{\{\sigma\}}^{2^N} F(\boldsymbol{\sigma}|\boldsymbol{\xi}) \exp\left(t\beta N\sqrt{1 + \sum_{\mu=1}^P m_\mu^2} + (1-t)\beta \sum_{\mu=1}^P \psi^\mu \sum_{i=1}^N \xi_i^\mu \sigma_i\right)}{\sum_{\{\sigma\}}^{2^N} \exp\left(t\beta N\sqrt{1 + \sum_{\mu=1}^P m_\mu^2} + (1-t)\beta \sum_\mu^P \psi^\mu \sum_{i=1}^N \xi_i^\mu \sigma_i\right)}. \tag{21}$$

We now must identify the extrema of $\alpha_N(\beta, t)$: at $t = 1$, we get the original model we aim to solve, while at $t = 0$, we get a system with one-body terms. The latter can be handled easily, and its contribution to the free energy reads out as

$$\alpha_N(\beta, t = 0) = \ln 2 + \left\langle \ln \cosh\left(\beta \sum_{\mu=1}^P \psi^\mu \langle m_\mu\rangle\right)\right\rangle_\xi. \tag{22}$$

To obtain the free energy of the model $\alpha_N(\beta) = \alpha_N(\beta, t = 1)$, we use again the fundamental theorem of calculus and write

$$\alpha(\beta) = \lim_{N\to\infty} \alpha_N(\beta, t = 1) = \lim_{N\to\infty}\left(\alpha_N(\beta, t = 0) + \int_0^1 \dot\alpha_N(\beta, t')dt'\right), \tag{23}$$

where, as usual, $\dot\alpha_N(\beta, t) := d_t \alpha_N(\beta, t)$ is the t-derivative of the interpolating free energy. The latter reads as

$$\frac{d\alpha_N(\beta, t)}{dt} = \beta\left\langle\left(\sqrt{1 + \langle \mathbf{m}_N^2\rangle} - \sum_{\mu=1}^P \psi^\mu \langle m_\mu\rangle\right)\right\rangle_\xi. \tag{24}$$

In general, its evaluation is an hard task, yet in the thermodynamic limit (where we are focusing), calling M_μ the limiting value of the μ-th Mattis overlap (i.e., $\lim_{N\to\infty} P(\mathbf{m}) = \delta(\mathbf{m} - \mathbf{M})$), by requiring β-almost-surely that

$$\lim_{N\to\infty} \sum_{\mu=1}^P \left\langle (m_\mu - M_\mu)^2 \right\rangle = 0, \tag{25}$$

$$\lim_{N\to\infty} \left\langle\left(\sqrt{1 + \mathbf{m}^2} - \sqrt{1 + \mathbf{M}^2}\right)^2\right\rangle = 0, \tag{26}$$

namely, the self-averaging of the order parameters and of the energy of the model,[8,11,20] as $\langle m_\mu\rangle \to M_\mu$, we obtain

$$\left\langle\sqrt{1 + \mathbf{m}^2} - \frac{\mathbf{m}\cdot\mathbf{M}}{\sqrt{1 + \mathbf{M}^2}} - \frac{1}{\sqrt{1 + \mathbf{M}^2}}\right\rangle = 0.$$

Comparing the above equation with the r.h.s. of (24) and choosing $\psi^\mu = \frac{M_\mu}{\sqrt{1+\mathbf{M}^2}}$, we can write

$$\frac{d\alpha(\beta, t)}{dt} - \frac{\beta}{\sqrt{1 + \mathbf{M}^2}} = 0.$$

Merging the last result with the Cauchy condition (22), we can finally state the next.[21,25]

Theorem 2. *The infinite-volume limit of the free energy for the relativistic Hopfield cost function (4) reads, in terms of its P Mattis overlaps, as*

$$\alpha(\beta) = \ln 2 + \left\langle\ln\cosh\left(\beta\boldsymbol{\xi}\cdot\frac{\mathbf{M}}{\sqrt{1 + \mathbf{M}^2}}\right)\right\rangle_\xi + \frac{\beta}{\sqrt{1 + \mathbf{M}^2}}, \tag{27}$$

and its related self-consistency equation reads as

$$M_\mu = \left\langle\xi^\mu \tanh\left(\beta\frac{\sum_{\mu=1}^P \xi^\mu M_\mu}{\sqrt{1 + \sum_{\mu=1}^P M_\mu^2}}\right)\right\rangle_\xi, \tag{28}$$

in agreement with the previous results reported in Ref. 10.

Remark 5. We stress that the leading order expansion of the above free energy (27) *and self-consistency* (28) *correctly recovers the standard Hopfield picture coded by Eqs.* (10) *and* (11).

We can now move on to study the critical behavior of the system. To this task, it is convenient to introduce the following.

Definition 8. Taking, without loss of generality, ξ^1 as the candidate pattern to be retrieved, we introduce its centered and rescaled Mattis overlap as

$$\hat{m}_1 := \sqrt{N}(m_1 - M_1), \tag{29}$$

where, as usual, $m_1 := N^{-1}\sum_{i=1}^{N}\xi_i^1\sigma_i$ while M_1 is its thermodynamic limit, namely, $\lim_{N\to\infty}\langle m_1\rangle \to M_1$.

Notice that $\langle\hat{m}_1^2\rangle$ scales as the variance of $\langle m_1\rangle$ times the system size N. In this way, we can investigate the critical behavior of the system by studying $\langle\hat{m}_1^2\rangle$ as a function of the noise β: we look for those values β_c where the fluctuations of the order parameters diverge as this is a signature of criticality.

In order to get an explicit expression for $\langle\hat{m}_1^2\rangle$, we exploit again the interpolation scheme (20) and we write $\langle\hat{m}_1^2\rangle := \langle\hat{m}_1^2\rangle_{t=1}$ as

$$\langle\hat{m}_1^2\rangle_{t=1} = \langle\hat{m}_1^2\rangle_{t=0} + \int_0^1 \frac{d\langle\hat{m}_1^2\rangle_{t'}}{dt'}\,dt, \tag{30}$$

where the generalized average $\langle\cdot\rangle_t$ is defined in (21).

Note that our calculation starts at $t = 0$ where the system is decoupled; it is thus both natural and convenient to approach the critical line from the high noise region as in this region it is possible to use standard central-limit-theorem-like arguments to assume the probability distribution of $\hat{m}_1$ to be a Gaussian;[40] the Cauchy condition $\langle\hat{m}_1^2\rangle_{t=0}$ (that is the standard *high-temperature* value) can be easily shown to be $\langle\hat{m}_1^2\rangle_{t=0} = 1$ as $\langle\hat{m}_1^2\rangle_{t=0} = \langle N(m_1 - M_1)^2\rangle_{t=0} = 1 + (N-1)M_1^2 + NM_1^2 - 2NM_1^2 = 1 - M_1^2 = 1$ as in the ergodic region trivially $M_1^2 = 0$. We must now face the t-derivative: to this task, it is useful to state the next.

Proposition 2. Retain $\psi_\mu = M_\mu/(\sqrt{1 + \sum_{\mu=1}^{P}M_\mu^2})$ and let F be a smooth function of the Mattis overlaps, then, above and close to the critical point (namely, where the Mattis overlaps are zero or infinitesimal), the following streaming equation holds:

$$\frac{d}{dt}\langle F\rangle_t \sim \frac{\beta}{2}\big(\langle F\hat{\mathbf{m}}^2\rangle_t - \langle F\rangle\langle\hat{\mathbf{m}}^2\rangle_t\big). \tag{31}$$

Proof. The proof uses the fact that we are inspecting the critical behavior (hence we assume the Mattis overlap to move from zero continuously when reached a critical noise level) and works by direct brute force

$$\frac{d}{dt}\langle F\rangle_t = \beta N\left(\langle F\sqrt{1 + \mathbf{m}^2}\rangle_t - \sum_{\mu=1}^{P}\frac{M_\mu}{\sqrt{1 + \mathbf{M}^2}}\langle Fm_\mu\rangle_t - \langle F\rangle_t\langle\sqrt{1 + \mathbf{m}^2}\rangle_t + \frac{M_\mu}{\sqrt{1 + \mathbf{M}^2}}\langle F\rangle_t\langle m_\mu\rangle_t\right). \tag{32}$$

Expanding $\sqrt{1 + x^2} \sim 1 + x^2/2$ and $1/(\sqrt{1 + x^2}) \sim 1 - x^2/2$ in the above equation and adding and subtracting twice $(\beta N/2)\langle F\rangle_t\hat{m}^2$, we have the result. $\square$

The above proposition is the key to state the last.

Theorem 3. *The centered and rescaled fluctuations of the Mattis overlap $\langle m_1\rangle$, quantifying the retrieval of pattern ξ^1, behave as*

$$\langle\hat{m}_1^2\rangle = \langle\hat{m}_1^2\rangle_{t=1} = \frac{1}{1 - \beta}; \tag{33}$$

thus, the high-noise (i.e., ergodic to keep the physical jargon) region is limited to $\beta < \beta_c = 1$.

In the low-noise (i.e., broken-symmetry) region, the Mattis overlap may assume non-null values (see Fig. 1).

Proof. Applying streaming Eq. (31) on $\langle\hat{m}_1^2\rangle$, we get the following Cauchy-problem:

$$\frac{d}{dt}\langle\hat{m}_1^2\rangle_t = \beta\langle\hat{m}_1^2\rangle_t^2, \tag{34}$$

$$\langle\hat{m}_1^2\rangle_{t=0} = 1, \tag{35}$$

whose solution is exactly Eq. (33). $\square$

IV. CONCLUSIONS AND OUTLOOKS

In the last decade, artificial intelligence has permeated our everyday lives and, as a natural consequence, the quest for novel and stronger tools to deal with it has become an urgent priority in different fields of science.[27] In particular, within the statistical-mechanics route, heuristic and rigorous results appeared steadily in the theoretical and mathematical physics communities in the last years, typically with heuristic approaches first (as in the case of the standard Hopfield model whose analysis was pioneered by Amit, Gutfreund, and Sompolinsky by replica-trick techniques[7]) and confirmed later by more rigorous techniques (see, e.g., Refs. 4, 14, and 32).

Along these lines, in this paper, relying on Guerra's interpolation schemes,[20,22] we have rigorously analyzed the statistical-mechanical picture of a new associative neural network, that is, the *relativistic Hopfield model* recently introduced in Ref. 10. More precisely, by an adaptation of the classical Guerra-Toninelli argument, we have proved that the infinite-volume limit of the free energy exists and it is well defined for this model; we also gave its explicit expression in terms of the P Mattis overlaps, namely, the natural order parameters of the theory, which perfectly matches the results of Ref. 10. By extremizing the free energy over the Mattis overlaps, we obtain the P self-consistent equations for the order parameters and their study allows us to get a picture of the critical phase transition that the model undergoes when the noise level β crosses the critical value $\beta_c = 1$. To obtain the explicit expression for the free energy, we required the self-averaging (in the infinite-volume limit) of the overlaps: this assumption has been confirmed immediately after by inspecting their (centered and rescaled) fluctuations. The divergence of these fluctuations is found to happen at $\beta_c = 1$: this result highlights a critical behavior similar to the one known for the standard Hopfield model in the low-storage regime, despite the presence of many-body contributions.

From a physical viewpoint, this may, at a first glance, sound weird as p-spin systems are known to exhibit discontinuous phase transitions (see, e.g., Ref. 9). However, an intuitive explanation for this is that, keeping the r.h.s. of Eq. (5) in mind, the pairwise-interaction provides the main driving force toward a broken phase, but the next one, i.e., the fourth-order in σ [which is small if compared to the previous order as the series (5) obviously converges], appears with the opposite sign; thus, it drives the system toward a non-magnetized state $\mathbf{M} \to 0$. Hence, we have to wait the even smaller contribution, i.e., the sixth-order in σ, to have another strengthening toward a magnetized state, but this is actually negligible.

From a mathematical viewpoint, instead, the request of the self-averaging of the order parameters (25) hints that, for the high-storage analysis, different approaches would be necessary. This can be understood with a scaling argument by observing that the request (25) breaks down if $P \propto N$ as the term inside the parenthesis scales as N^{-1}.

Finally, it is worth pointing out that the outlined mathematical scaffold stays robust against pattern's dilution, namely, for multitasking networks,[1] on whose *relativistic extension* we plan to report soon.

APPENDIX: SUB/SUPER-ADDITIVITY AND FEKETE LEMMA

Definition 9 (sub-additive, super-additive functions). *A function $f: A \to B$ is termed sub-additive (respectively, super-additive) if*

$$\forall x, y \in A, \qquad f(x + y) \le f(x) + f(y)$$

(respectively,

$$\forall x, y \in A, \qquad f(x + y) \ge f(x) + f(y) \,).$$

In the same way, we can define the concept of sub-additive succession in the following way:

Definition 10 (sub-additive, super-additive successions). *A succession $(a_n)_{n \in \mathbb{N}}$ is termed sub-additive (respectively, super-additive) if $\forall n, m \in \mathbb{N}$,*

$$a_{n+m} \le a_n + a_m$$

(respectively,

$$a_{n+m} \ge a_n + a_m \,).$$

Lemma 1. Let $(a_n)_{n \in \mathbb{N}}$ be a sub-additive succession, then the following holds:

$$a_{nk} \le k a_n, \quad \forall k \in \mathbb{N}.$$

Proof. A sketched proof is immediate by using the induction principle over k. Indeed, let $(a_n)_{n \in \mathbb{N}}$ be a sub-additive succession. If $k = 1$, then

$$a_n \leq 1 a_n = a_n.$$

Now, let us pose $k > 1$ and assume the thesis still holds for any natural up to $k - 1$ and let us prove it for k as well. To be true at $k - 1$, it must be that

$$a_{n(k-1)} \leq (k-1)a_n \tag{A1}$$

and thus that

$$a_{n(k-1)} + a_n \leq ka_n. \tag{A2}$$

Note that, thanks to the sub-additivity property of the succession, we can write

$$a_{nk-n} + a_n \geq a_{nk-n+n} = a_{nk}. \tag{A3}$$

Merging inequalities (A2) and (A3), the thesis follows trivially. □

Proposition 3 (Fekete's Lemma). *For any sub-additive succession* $(a_n)_{n \in \mathbb{N}}$,

$$\exists \lim_{n \to \infty} \frac{a_n}{n} = \inf_{n \in \mathbb{N}} \frac{a_n}{n}.$$

Similar conclusions can be drawn for super-additive successions as well, and in this case, we have that

$$\exists \lim_{N \to +\infty} \frac{a_n}{n} = \sup_{n \in \mathbb{N}} \frac{a_n}{n}.$$

REFERENCES

[1] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, "Multitasking associative networks," Phys. Rev. Lett. **109**, 268101 (2012).

[2] E. Agliari, A. Barra, A. De Antoni, and A. Galluzzi, "Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines," Neural Networks **38**, 52 (2013).

[3] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, and F. Tavani, "Retrieval capabilities of hierarchical networks: From Dyson to Hopfield," Phys. Rev. Lett. **114**, 028103 (2015).

[4] E. Agliari, A. Barra, C. Longo, and D. Tantari, "Neural networks retrieving binary patterns in a sea of real ones," J. Stat. Phys. **168**, 1085 (2017).

[5] E. Agliari, D. Migliozzi, and D. Tantari, "Non-convex multi-species Hopfield models," J. Stat. Phys. **172**, 1247 (2018).

[6] E. Agliari, A. Barra, G. Landolfi, S. Murciano, and S. Perrone, "Complex reaction kinetics in chemistry: A unified picture suggested by mechanics in physics," Complexity **2018**, 7423297.

[7] D. J. Amit, *Modeling Brain Functions* (Cambridge University Press, 1989).

[8] A. Barra, "The mean field Ising model trough interpolating techniques," J. Stat. Phys. **132**(5), 787 (2008).

[9] A. Barra, "Notes on ferromagnetic p-spin and REM," Math. Methods Appl. Sci. **32**, 783 (2008).

[10] A. Barra, M. Beccaria, and A. Fachechi, "A new mechanical approach to handle generalized Hopfield neural networks," Neural Networks **106**, 205 (2018).

[11] A. Barra, G. Genovese, and F. Guerra, "Equilibrium statistical mechanics of bipartite spin systems," J. Phys. A: Math. Theor. **44**(24), 245002 (2011).

[12] A. Barra, G. Genovese, and F. Guerra, "The replica symmetric approximation of the analogical neural network," J. Stat. Phys. **140**(4), 784 (2010).

[13] A. Barra, F. Guerra, G. Genovese, and D. Tantari, "How glassy are neural networks?," J. Stat. Mech.: Theory Exp. **2012**, P07009.

[14] A. Bovier and V. Gayrard, "Hopfield models as generalized random mean field models," in *Mathematical Aspects of Spin Glasses and Neural Networks* (Birkhauser Press, Boston, 1998).

[15] A. Bovier, V. Gayrard, and P. Picco, "Gibbs states of the Hopfield model in the regime of perfect memory," Probab. Theory Relat. Fields **100**(3), 329 (1994).

[16] A. C. C. Coolen, R. Kuhn, and P. Sollich, *Theory of Neural Information Processing Systems* (Oxford Press, 2005).

[17] V. Dotsenko, *An Introduction to the Theory of Spin Glasses and Neural Networks* (World Scientific, 1995).

[18] A. Fachechi, E. Agliari, and A. Barra, "Dreaming neural networks: Forgetting spurious memories and reinforcing pure ones," Neural Networks **112**, 24 (2019).

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2017).

[20] F. Guerra, "Sum rules for the free energy in the mean field spin glass model," Math. Phys. Math. Phys. **30**, 161 (2001).

[21] F. Guerra and F. L. Toninelli, "The thermodynamic limit in mean field spin glass models," Commun. Math. Phys. **230**(1), 71–79 (2002).

[22] F. Guerra and F. L. Toninelli, "The infinite volume limit in generalized mean field disordered models," Markov Proc. Rel. Fields **9**(2), 195–207 (2003).

[23] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. U. S. A. **79**(8), 2554–2558 (1982).

[24] J. J. Hopfield, D. I. Feinstein, and R. G. Palmer, "Unlearning has a stabilizing effect in collective memories," Nat. Lett. **304**, 158 (1983).

[25] J. A. Horas and P. M. Pasinetti, "On the unlearning procedure yielding a high-performance associative memory neural network," J. Phys. A: Math. Gen. **31**, L463–L471 (1998).

[26] D. Krotov and J. J. Hopfield, "Dense associative memory is robust to adversarial inputs," Neural Comput. **30**, 3151 (2018).

[27] Y. Le Cun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[28] K. Nokura, "Spin glass states of the anti-Hopfield model," J. Phys. A: Math. Gen. **31**, 7447 (1998).

[29] K. Nokura, "Paramagnetic unlearning in neural network models," Phys. Rev. E **54**(5), 5571, (1996).

[30] K.-S. Oh and J. Keechul, "GPU implementation of neural networks," Pattern Recognit. **37**(6), 1311 (2004).

[31] D. Ruelle, *Statistical Mechanics: Rigorous Results* (World Scientific, 1999).

[32] M. Talagrand, "Rigorous results for the Hopfield model with many patterns," Probab. Theory Relat. Fields **110**(2), 177 (1998).

[33] M. Talagrand, "Exponential inequalities and convergence of moments in the replica-symmetric regime of the Hopfield model," Ann. Probab. **28**, 1393 (2000).

[34] J. Tubiana and R. Monasson, "Emergence of compositional representations in restricted Boltzmann machines," Phys. Rev. Lett. **118**(13), 138301 (2017).

[35] S. Wimbauer and J. L. van Hemmen, "Hebbian unlearning," in *Analysis of Dynamical and Cognitive Systems* (Springer, Berlin, 1995).

[36] Notice that, once the existence of the thermodynamic limit is proved for the free energy, it holds, in a cascade fashion, also for other various quantities of interest, as entropy and internal energy.[22]

[37] It is worth mentioning that biased patterns (e.g., where positive entries are more likely), beyond being less informative, may also lead to technical issues as spurious configurations would present large overlap with pure memories, so the attracting power of the latter is strongly downsized, see also Ref. 7.

[38] For instance, information can be supplied to the network as an *external field* to keep a physical jargon.

[39] Actually, we are committing a small abuse of notation as the "standard" free energy $\tilde{\alpha}(\beta)$ is defined as $\tilde{\alpha}(\beta) = -\beta^{-1}\alpha(\beta)$, but, as from the mathematical side $\tilde{\alpha}(\beta)$ and $\alpha(\beta)$ are (a constant apart) equivalent, we will keep our choice as it makes calculations more transparent, obviously, without affecting the results.

[40] This assumption crucially allows us to use the Wick Theorem, namely, given a Gaussian variable z we can write $\mathbb{E}(z \cdot f(z)) = \mathbb{E}(z^2)\mathbb{E}(\partial_z f(z))$.