PERSPECTIVE

Supervised Hebbian learning

To cite this article: Francesco Alemanno *et al* 2023 *EPL* **141** 11001

A Letters Journal Exploring the Frontiers of Physics

January 2023

EPL, **141** (2023) 11001 www.epljournal.org
doi: 10.1209/0295-5075/aca55f

Perspective

Supervised Hebbian learning

Francesco Alemanno[1,2], Miriam Aquaro[3,4], Ido Kanter[5], Adriano Barra[1,2(a)] and Elena Agliari[3,4]

[1] *Dipartimento di Matematica e Fisica, Università del Salento - Campus Ecotekne, via Monteroni, Lecce 73100, Italy*
[2] *Istituto Nazionale di Fisica Nucleare, Sezione di Lecce - Campus Ecotekne, via Monteroni, Lecce 73100, Italy*
[3] *Dipartimento di Matematica, Sapienza Università di Roma - P.le A. Moro 5, 00185, Rome, Italy*
[4] *Istituto Nazionale d'Alta Matematica (GNFM) "F. Severi" - P.le A. Moro 5, 00185, Rome, Italy*
[5] *Department of Physics, Bar-Ilan University - Ramat-Gan, 52900, Israel*

received 8 September 2022; accepted in final form 23 November 2022
published online 4 January 2023

Abstract – In neural network's literature, *Hebbian learning* traditionally refers to the procedure by which the Hopfield model and its generalizations *store* archetypes (*i.e.*, definite patterns that are experienced just once to form the synaptic matrix). However, the term *learning* in machine learning refers to the ability of the machine to extract features from the supplied dataset (*e.g.*, made of blurred examples of these archetypes), in order to make its own representation of the unavailable archetypes. Here, given a sample of examples, we define a supervised learning protocol based on Hebb's rule and by which the Hopfield network can infer the archetypes. By an analytical inspection, we detect the correct control parameters (including size and quality of the dataset) that tune the system performance and we depict its phase diagram. We also prove that, for structureless datasets, the Hopfield model equipped with this supervised learning rule is equivalent to a restricted Boltzmann machine and this suggests an optimal and interpretable training routine. Finally, this approach is generalized to structured datasets: we highlight an ultrametric-like organization (reminiscent of replica-symmetry-breaking) in the analyzed datasets and, consequently, we introduce an additional *broken-replica hidden layer* for its (partial) disentanglement, which is shown to improve MNIST classification from $\sim 75\%$ to $\sim 95\%$, and to offer a new perspective on deep architectures.

Forty years have elapsed since Hopfield's seminal work, yielding a model for biological information processing [1]; meanwhile, we have witnessed a striking development of artificial machine learning (see, *e.g.*, [2–4]) and we are finally in a stage where ideas, techniques and results stemming from biological and artificial sides can be fruitfully compared (see, *e.g.*, [5–9]). Here we leverage their analogies to unveil the internal mechanisms of a learning machine, focusing on two paradigmatic models, that is, respectively, the Hopfield neural network (HNN) and the restricted Boltzmann machine (RBM). In order for this comparison to be exhaustive, we first need to profoundly revise the assumptions underlying the theory developed by Amit, Gutfreund and Sompolinsky (AGS) [10], who, in the eighties, gave a pioneering statistical-mechanical treatment of the HNN based on spin glasses [11]. The point is that, in the AGS theory, the HNN actually does not learn, rather it stores definite patterns —hereafter called *archetypes*— by the so-called Hebb rule (or countless variations on the theme); on the other hand, in standard machine learning the network has to infer these archetypes by solely experiencing (a finite number of) their noisy versions —hereafter called *examples*— while the original archetypes remain unknown. Hence, in order to match biological and artificial information processing, we must supply the HNN with examples rather than directly archetypes and therefore turn Hebb's rule into a genuine learning rule. Despite some notable contributions in the late eighties and nineties, during the first wave of formalization of neural networks via statistical mechanics, see, *e.g.*, [12–15], this particular aspect has remained overlooked. In the following we will reach such a framework, whence we will show that standard machine learning rules based on contrastive divergence algorithms collapse onto Hebb's learning rule, and we will highlight quantitative control parameters whose tuning determines the learning machine failure or success. These results are obtained analytically by statistical-mechanics

(a) E-mail: adriano.barra@gmail.com (corresponding author)

tools for random, unstructured datasets, where we can also establish a direct connection between the number of archetypes and the number of hidden neurons in the RBM. As for structured datasets, the robustness of these results is checked numerically for the MNIST and the fashion-MNIST datasets [16,17] and we also generalize the connection between the size of the hidden layer(s) and the intrinsic complexity of the dataset, exploiting an iterative rule, reminiscent of the replica-symmetry-breaking (RSB) paradigma [11].

Let us start with the theoretical approach and introduce the information the network has to deal with: we define K archetypes denoted with $\boldsymbol{\xi}^\mu, \mu \in \{1,\dots,K\}$, as binary vectors of length N and whose entries are i.i.d. variables drawn from

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

for any $i \in \{1,\dots,N\}$ and $\mu \in \{1,\dots,K\}$, then, for each of them we generate M examples $\boldsymbol{\eta}^{\mu a}$, $a \in \{1,\dots,M\}$, that we obtain by corrupting the archetype flipping its digits randomly as

$$\eta_i^{\mu a} = \xi_i^\mu \chi_i^{\mu a}, \tag{2}$$

$$\mathcal{P}(\chi_i^{\mu a}) = \frac{1+r}{2}\delta(\chi_i^{\mu a} - 1) + \frac{1-r}{2}\delta(\chi_i^{\mu a} + 1), \tag{3}$$

for any i, μ, a, being $r \in (0,1]$ a parameter tuning the *quality* of the sample. We now feed the HNN on the dataset $\mathcal{S} = \{\boldsymbol{\eta}^{\mu a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$ and, for this operation to be unambiguous, we also need to specify *how* these examples are presented to the network, mirroring supervised and unsupervised learning. In fact, the HNN Hamiltonian reads as $\mathcal{H}^{(\text{HNN})}(\boldsymbol{\sigma}|\boldsymbol{J}) = -\sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j$, where $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1,\dots,N} \in \{-1,+1\}^N$ are N binary neurons and the synaptic connections J_{ij}'s incorporate the accessible information: in the original setting, where archetypes are available, the Hebbian (storing) rule reads as $J_{ij} \propto \sum_\mu \xi_i^\mu \xi_j^\mu$, while here $J_{ij} = J_{ij}(\mathcal{S})$ and we envisage the following protocols.

Supervised Hebbian learning: A teacher discloses the example labels and they can therefore be combined as

$$J_{ij}^{\text{sup}} \propto \sum_{\mu=1}^{K}\left(\sum_{a=1}^{M}\eta_i^{\mu a}\right)\left(\sum_{b=1}^{M}\eta_j^{\mu b}\right). \tag{4}$$

Unsupervised Hebbian learning: Without a teacher that tells how to cluster examples, we mix them up obtaining

$$J_{ij}^{\text{unsup}} \propto \sum_{\mu=1}^{K}\sum_{a=1}^{M}\eta_i^{\mu a}\eta_j^{\mu a}. \tag{5}$$

Clearly, when $r = 1, M$ becomes a dummy variable because examples coincide with the related archetype and we recover the classical Hebbian rule in both cases. Here we focus on the former (4), while we refer to the Supplementary Material `Supplementarymaterial.pdf` (SM)[1] for a discussion on the latter (5).

A convenient control parameter to assess the information content in $\mathcal{S}$ is $\rho := \frac{1-r^2}{Mr^2}$. To see this, let us focus on the μ-th pattern and the i-th digit, whose related block is $\boldsymbol{\eta}_i^\mu = (\eta_i^{\mu 1}, \eta_i^{\mu 2}, \dots, \eta_i^{\mu M})$; the error probability for any single entry is $\mathcal{P}(\chi_i^{\mu a} = -1) = (1-r)/2$ and, by applying the majority rule on the block, it is reduced to $\mathcal{P}(\text{sgn}(\sum_a \chi_i^{\mu a}) = -1) \underset{M\gg 1}{\approx} [1 - \text{erf}(1/\sqrt{2\rho})]$, thus, the conditional entropy $H(\xi_i^\mu|\boldsymbol{\eta}_i^\mu)$, that quantifies the amount of information needed to describe the original message ξ_i^μ, given the related M-length block $\boldsymbol{\eta}_i^\mu$, is monotonically increasing with ρ, saturating to 1 bit. Hence, in order for the dataset to retain information on the original archetypes, ρ must be finite, that is, Mr^2 must be non-vanishing.

This scaling, arising from an information theory perspective, is recovered and sharpened in the neural network framework. We start with the signal-to-noise analysis on the HNN to check for local stability of the archetype retrieval configurations in the noiseless limit, that is, we study the conditions under which the internal field $h_i(\boldsymbol{\sigma}) = \sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}(\mathcal{S})\sigma_j$, namely the post-synaptic potential experienced by the neuron i, is aligned with the neural activity σ_i while $\boldsymbol{\sigma} = \boldsymbol{\xi}^\mu$, for any arbitrary μ and i. This is usually inspected by checking that the "signal" (*i.e.*, the expectation of $h_i(\boldsymbol{\xi}^\mu)\xi_i^\mu$ over the realization of archetypes and examples) is larger than the "noise" (*i.e.*, the standard deviation of $h_i(\boldsymbol{\xi}^\mu)\xi_i^\mu$). As detailed in the SM, this condition can be recast into the requirement that the system configuration after one Monte Carlo step exhibits at least a fraction $(1+1/\sqrt{2})/2 \approx 0.85$ of spins aligned with $\boldsymbol{\xi}^\mu$ and this, in turn, returns $\frac{K}{N}(1+\frac{1-r^2}{Mr^2})^2 + \frac{1-r^2}{Mr^2} \lesssim 1$. This relation advises on the suitable rescaling of the dataset size ($M \gtrsim r^{-2}$), as the dataset quality is impaired ($r \to 0$), in order to preserve network's abilities; note that power-law scalings were already evidenced in the machine-learning context, see, *e.g.*, [18]. To achieve a quantitative picture and control of the network behavior, we work out a statistical-mechanics investigation and we start by introducing the Boltzmann-Gibbs measure for the system,

$$\mathcal{P}_\beta(\boldsymbol{\sigma}|\mathcal{S}) = \frac{1}{Z_\beta^{(\text{HNN})}(\mathcal{S})}e^{-\beta\mathcal{H}^{(\text{HNN})}(\boldsymbol{\sigma}|\boldsymbol{J}(\mathcal{S}))}, \tag{6}$$

where $Z_\beta^{(\text{HNN})}$ is the partition function and $\beta := 1/T \in \mathbb{R}^+$ tunes the distribution broadness; β along with the load $\alpha := \lim_{N\to\infty} K/N$ and the dataset "entropy" $\rho = (1-r^2)/Mr^2$, make up the set of control parameters. Further, we introduce the macroscopic observables (order parameters) useful to describe the system behavior, namely

$$m := \frac{1}{N}\sum_{i=1}^{N}\xi_i^1\sigma_i, \tag{7}$$

[1]Details concerning calculations and numerical simulations can be found in the SM.

$$n := \frac{1}{r(1+\rho)} \frac{1}{NM} \sum_{i,a=1}^{N,M} \eta_i^{1a} \sigma_i, \tag{8}$$

$$q_{12} := \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{(1)} \sigma_i^{(2)}, \tag{9}$$

where we defined, respectively, the Mattis magnetization of the archetype (eq. (7)), the typical magnetization of the example (eq. (8)), and the two-replica overlap (eq. (9)); for m and n we referred to $\mu = 1$ without loss of generality.

Under the replica-symmetry (RS) ansatz, all the order parameters do not fluctuate in the thermodynamic limit, *i.e.*, being $\mathcal{P}(x)$ the probability distribution for the observable $x = (m, n, q_{12})$ and $\langle x \rangle$ its expectation, then $\lim_{N\to\infty} \mathcal{P}(x) = \delta(x - \langle x \rangle)$. These expectation values can be obtained by extremizing the quenched free energy of the model with respect to the order parameters and, as explained in the SM, for $N \to \infty$ and $M \gg 1$, we obtain the following set of self-consistent equations:

$$\langle m \rangle = \mathbb{E}_z \tanh \left\{ \beta \langle n \rangle + z\beta \sqrt{\langle n \rangle^2 \rho + \frac{\alpha \langle q \rangle}{\left[1 - \beta\left(1 - \langle q \rangle\right)\right]^2}} \right\}, \tag{10}$$

$$\langle n \rangle = \frac{\langle m \rangle}{(1+\rho) - \rho\beta(1 - \langle q \rangle)}, \tag{11}$$

$$\langle q \rangle = \mathbb{E}_z \tanh^2 \left\{ \beta \langle n \rangle + z\beta \sqrt{\langle n \rangle^2 \rho + \frac{\alpha \langle q \rangle}{\left[1 - \beta\left(1 - \langle q \rangle\right)\right]^2}} \right\}, \tag{12}$$

where $\mathbb{E}_z$ denotes the average with respect to the standard Gaussian variable z. The inspection of eqs. (10)–(12) provides a quantitative picture of the system behavior in the space of the control parameters as reported in fig. 1(a), (b). In particular, like in the classical HNN, we recognize the emergence of an ergodic region corresponding to large values of T and a retrieval region for relatively small values of α and T, yet, the Hebbian learning rule (4) makes the phenomenolgy much richer: here we have an additional tuneable parameter ρ which controls the width of the retrieval region. Denoting with $\alpha_c(T, \rho)$ the first-order transition line between the spin-glass phase and the retrieval phase, we show that $\alpha_c(T = 0, \rho)$ is a decreasing function of ρ and, as expected, $\alpha_c(T = 0, \rho = 0) \approx 0.138$, consistently with the AGS theory. Signatures of this transition are also found by means of finite-size Monte Carlo (MC) simulations as shown in fig. 1(c), (d). Further, looking at eqs. (10)–(12) and requiring a non-vanishing magnetization $\langle m \rangle$, we derive that ρ must be finite and therefore we recover the scaling $M \sim r^{-2}$; also, in the zero fast-noise limit $T \to 0$, these equations can be treated to get explicit expressions as achieved in the SM.

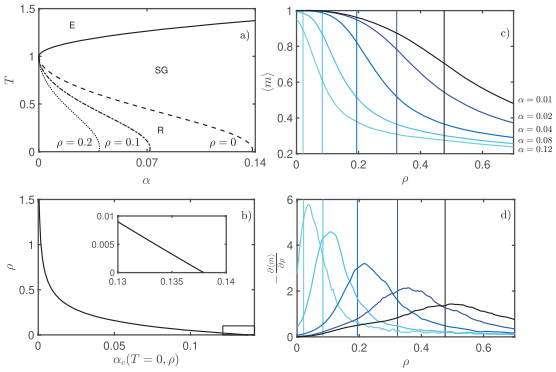

Fig. 1: Behaviour of the supervised HNN as the control parameters are varied. (a) Phase diagram highlighting the ergodic (E), the spin-glass (SG) and the retrieval (R) phase *vs.* T and α; the transition line between the SG phase and the R phase depends on ρ and three cases are shown: $\rho = 0$ (dashed line, corresponding to AGS theory), $\rho = 0.1$ (dashed-dotted line), and $\rho = 0.2$ (dotted line). (b) Critical load α_c obtained for $T = 0$ and as a function of ρ. (c) Estimate of the Mattis magnetization *vs.* ρ by MC simulations for systems of size $N = 5000$; different loads are considered and plotted in different colors (brighter nuances correspond to larger values of α, as reported on the right); the vertical lines represent the transition points predicted analytically. (d) From data presented in panel (c) we derive the susceptibility with respect to ρ and notice that the peaks approximately match the transition points (by a finite-size scaling we checked that the match gets closer as N is made larger).

We now bridge this theory with the machine learning counterpart. We consider a RBM made of two layers, a visible one endowed with N binary neurons $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1,\dots,N} \in \{-1, +1\}^N$, and a hidden one built of K real-valued neurons $\boldsymbol{z} = \{z_\mu\}_{\mu=1,\dots,K} \in \mathbb{R}^K$ with a Gaussian prior, and whose Hamiltonian reads as $\mathcal{H}^{(\mathrm{RBM})}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{W}) = -\sum_{i,\mu}^{N,K} W_{i,\mu} \sigma_i z_\mu$. We choose the length of the hidden layer to match the number of archetypes in such a way that, as we will see, we can assign to each hidden neuron the recognition of a unique archetype. The Boltzmann-Gibbs distribution associated to $\mathcal{H}^{(\mathrm{RBM})}$ is

$$\mathcal{P}_\beta(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{W}) = \frac{1}{Z_\beta^{(\mathrm{RBM})}(\boldsymbol{W})} e^{-\beta \mathcal{H}^{(\mathrm{RBM})}(\boldsymbol{\sigma}, \boldsymbol{z}|\boldsymbol{W}) - \beta \frac{\boldsymbol{z}^2}{2}}. \tag{13}$$

Now, the goal is to find the weight setting such that this measure mimics the target one, referred to as $\mathcal{Q}$, which generated the examples in $\mathcal{S}$. Focusing on a classification task, we adopt the so-called grandmother cell scheme: during training, the generic input-output pair is $(\boldsymbol{\eta}^{\nu a}, \boldsymbol{z}^{(\nu)})$, where $\boldsymbol{z}^{(\nu)}$ is the one-hot vector whose ν-th entry is the single non-null entry [19,20]. Thus, the target distribution reads as

$$\mathcal{Q}(\boldsymbol{\sigma}, \boldsymbol{z}) = \sum_{\mu,a} \delta(\boldsymbol{\eta}^{\mu a} - \boldsymbol{\sigma}) \delta(\boldsymbol{z}^{(\mu)} - \boldsymbol{z}), \tag{14}$$

and, if training is successful, we expect that, initializing the visible layer as a test example $\tilde{\boldsymbol{\eta}}^\nu$ of the ν-th archetype

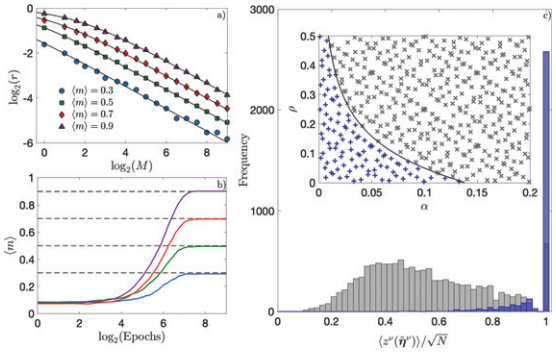

Fig. 2: Comparison between HNN and RBM performances. In (a) we fix a certain value for the expected magnetization $\langle m\rangle$ and we derive from eq. (10), obtained theoretically for the HNN, how r and M should be tuned in order to retain this value constant (solid line); an analogous analysis is repeated numerically for the RBM where now m is evaluated as the overlap between the visible layer and a given archetype (symbols); different values of magnetization are considered and represented with different symbols. (b) Expected value of the RBM magnetization versus the training time and for given values of r and M, under on-line contrastive divergence (CD-1) [21]; the long-time value corresponds to the theoretical estimate obtained for the HNN for the same choice of r and M (horizontal lines). In (c) we sampled 1.5×10^4 couples $(\alpha,\rho)\in(0,0.2)\times(0,0.5)$ by Sobol's low-discrepancy sequence; for each extraction (represented by a cross in the inset) we build a RBM of size $N=5000$ and $K=\alpha N$, we generate a set $\mathcal{S}$ of examples and we set the machine weights as $\boldsymbol{W}=\bar{\boldsymbol{\eta}}$. Then, we initialize the visible layer as a test example $\tilde{\boldsymbol{\eta}}^\nu$, we run MC simulations and we evaluate $\langle z_\nu\rangle$, whose histogram is depicted in the main plot, distinguishing between cases inside (blue) and outside (grey) the retrieval region. Since $\langle z_\nu\rangle\propto\langle m_\nu\rangle$ (see eq. (16)), the delta-like shape of its histogram is a signature of retrieval.

and letting the neurons evolve freely up to thermalization, the hidden layer will provide the estimated class as argmax $[\langle\boldsymbol{z}(\tilde{\boldsymbol{\eta}}^\nu)\rangle]$.

The learning rule can be derived by a gradient descent on the Kullback-Leibler (KL) cross entropy $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})$ between the distributions $\mathcal{Q}$ and $\mathcal{P}$, that is, $W^{n+1}_{i,\mu}=W^n_{i,\mu}-\epsilon\frac{\mathrm{d}D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})}{\mathrm{d}W_{i,\mu}}$, where n accounts for training iterations and ϵ is the learning rate; recalling (14) this yields

$$W^{n+1}_{i,\mu}=W^n_{i,\mu}+\epsilon\left(\langle\sigma_i z_\mu\rangle_{\boldsymbol{\sigma}\&\boldsymbol{z}}-\langle\sigma_i z_\mu\rangle\right), \tag{15}$$

where the brackets denote the expectation under the Boltzmann-Gibbs measure (13) and the bracket subscript specifies the clamped variables.

In the case of orthogonal patterns, the configuration where weight entries are set as the empirical average of example entries, *i.e.*, $W_{i\mu}=\bar{\eta}_{i\mu}:=\frac{1}{M}\sum_{a=1}^M\eta_i^{\mu a}$, is a fixed point for the contrastive divergence and therefore compatible with a trained machine (see the SM and [19,22]. Further, with this choice we can prove that the RBM is equivalent, in distribution, to the HNN with supervised Hebbian rule; in fact, by a Gaussian integration,

$$\begin{aligned}Z^{(\mathrm{RBM})}_\beta(\boldsymbol{W}=\bar{\boldsymbol{\eta}})&=\sum_{\boldsymbol{\sigma}}^{2^N}\int e^{\frac{\beta}{\sqrt{N}}\sum_\mu(\sum_i\sigma_i\bar{\eta}_{i\mu})z_\mu)-\frac{\beta z_\mu^2}{2}}\\&=\sum_{\boldsymbol{\sigma}}^{2^N}e^{\frac{\beta}{2N}\sum_\mu\sum_{ij}\sigma_i\bar{\eta}_{i\mu}\bar{\eta}_{j\mu}\sigma_j}=Z^{(\mathrm{HNN})}_\beta(\mathcal{S}).\end{aligned} \tag{16}$$

This equivalence implies that the phase diagram outlined for the HNN (see fig. 1(a), (b)) also applies to the RBM, as confirmed in fig. 2. In particular, the retrieval region corresponds to a parameter setting where the trained RBM relaxes to configurations such that the overlap between the visible layer and the archetype are close to one. Remarkably, this is consistent with the usual performance and score values [14] or error-based measures as in the Vapnik-Chervonenkis learning theory [23] where one aims to minimize the distance between the output and the instances of a test set. In fact, $-(\boldsymbol{\sigma}-\boldsymbol{\xi}^1)^2\propto\boldsymbol{\sigma}\cdot\boldsymbol{\xi}^1=m$ and $-\overline{(\boldsymbol{\sigma}-\boldsymbol{\eta}^1)^2}\propto n$: whenever the network is in the retrieval region, for some archetype μ it is minimizing one of these Loss functions $L^\mu_\pm=(1/2N)||\xi^\mu\pm\sigma||^2=1\pm m_\mu$ as the Hopfied Hamiltonian can be written as

$$\mathcal{H}^{(\mathrm{HNN})}(\boldsymbol{\sigma}|\boldsymbol{J})=-N\sum_\mu^K(1-L^\mu_+L^\mu_-),$$

where the term $L^\mu_+L^\mu_-$ guarantess that it learns both the pattern ξ^μ and its gauge symmetric copy $-\xi^\mu$.

In order to appreciate further the equivalence between HNN and RBM, we show that it can be reached from a different perspective, namely using the maximum-entropy principle, according to Jaynes' inferential interpretation [24,25]. Let us look for the least structured probability distribution $\mathcal{P}(\boldsymbol{\sigma},\boldsymbol{z})$ that is compatible with the set of data $\{(\boldsymbol{\eta}^{\mu a},\boldsymbol{z}^{(\mu)})\}^{a=1,\dots,M}_{\mu=1,\dots,K}$ to inspect which kind of correlations the machine detects in the dataset. While extensive calculations are provided in the SM, here we report the main findings: the minimal constraints needed to recover the HNN and RBM's Boltzmann-Gibbs distribution concern the variance of hidden units and the correlations between visible and hidden units —set equal to their empirical estimates $C_{z^2_\mu}$ and $C^{i,\mu}_{\sigma,z}$ for $i=1,\dots,N$ and $\mu=1,\dots,K$, respectively— beyond those for $\mathcal{P}$ to be well defined. The constrained optimization problem therefore reads: $\max_{\{\lambda_0,\lambda_1,\Lambda_{i,\mu}\}_{i,\mu}}S[\mathcal{P}]$, with

$$\begin{aligned}S[\mathcal{P}]=&-\langle\mathcal{P}\ln\mathcal{P}\rangle_{\mathcal{P}}+\lambda_0\left(\langle\mathcal{P}\rangle_{\mathcal{P}}-1\right)\\&+\lambda_1\left(\left\langle\sum_{\mu=1}^K z_\mu^2\right\rangle_{\mathcal{P}}-C_{z^2_\mu}\right)+\sum_{i,\mu}^{N,K}\Lambda_{i,\mu}\left[\langle\sigma_i z_\mu\rangle_{\mathcal{P}}-C^{i,\mu}_{\sigma,z}\right],\end{aligned} \tag{17}$$

where $\langle\cdot\rangle_{\mathcal{P}}$ denotes the expectation over $\mathcal{P}$. The solution yields the following Lagrange multipliers:

$$e^{\lambda_0-1}=\sum_{\boldsymbol{\sigma},\,\boldsymbol{z}}\mathcal{P}(\boldsymbol{\sigma},\boldsymbol{z}),\lambda_1=1,$$

$$\Lambda_{i,\mu}=\sqrt{\frac{\beta}{Nr^2(1+\rho)}}\frac{1}{M}\sum_{a=1}^M\eta_i^{\mu a}.$$

Therefore, this machine captures correlations between the two classes of neurons and, under the supervised learning protocol chosen here, these are recast into empirical averages over examples. In particular, the hidden-layer size can be interpreted as a measure of the model flexibility: a larger K allows for a larger number of degrees of freedom and for a finer inference, yet too large a flexibility can imply overfitting phenomena which in our framework are naturally recast as the emergence of a pure spin-glass phase. According to the phase diagram in fig. 1(a), the maximum flexibility allowed is $K_c = \alpha_c(\rho)N$; this estimate is successfully checked in fig. 2(c).

Up to now, we proved that, when dealing with a random, structureless dataset, the HNN with supervised Hebbian rule and the RBM trained under a grandmother cell scheme are equivalent, and that parameters that emerge naturally in a statistical mechanics framework can be related to standard quantifiers in a machine learning context. More challenging datasets can also be treated as long as the intrinsic structure is properly encoded in the system as we are going to explain. Let us denote with $\mathcal{S} = \{\boldsymbol{\zeta}^{\mu a}\}_{a=1,\dots,M}^{\mu=1,\dots,K}$ the sample of examples, where the change of notation underlines that now, in general, there is no archetype available hence $\zeta_i^{\mu a}$ cannot be obtained by flipping some pixels in the related archetype as in eq. (2). Moreover, in the structureless case, scrolling through the various examples belonging to the same class, pixels are all homogeneously subject to a flipping probability, while in the structured case some pixels turn out to be more persistent than others. This recalls the difference between ergodic and glassy configurations in spin models. In particular, glassy configurations are characterized by peculiar statistical properties (*e.g.*, lack of self-averaging) which are in turn related to an ultrametric organization. The existence of an analogous organization for dataset items may suggest effective strategies for their processing of a learning machine. In fig. 3 we show some evidence in this sense: the distribution of item overlaps —mirroring replica overlaps in spin systems— resembles the Parisi distribution [11], further (as analyzed in depth in the SM) Ghirlanda-Guerra identities [26] are numerically shown to hold.

In the light of this result we expect that, when employing the basic, two-layered network for structured data, non-trivial correlations among hidden neurons arise, impairing the overall performance. To disentangle such correlations we conceive a routine that extends the previous grandmother cell scheme: We pre-treat each subsample $\mathcal{S}_\mu = \{\boldsymbol{\zeta}^{\mu a}\}_{a=1,\dots,M}$ to assess its intrinsic structure (*e.g.*, by principal component analysis), whence we determine K_μ disjoint and exhaustive sub-groups $\{\mathcal{S}_\mu^\ell\}_{\ell=1,\dots,K_\mu}$ and we allocate as many hidden neurons for each class, the overall size of the hidden layer therefore reads as $\hat{K} = \sum_{\mu=1}^{K} K_\mu$. The weight matrix $\boldsymbol{W} \in \mathbb{R}^{\hat{K}\times N}$ is determined by averaging over instances assigned to each sub-group $\mathcal{S}_\mu^\ell$ for $\ell = 1,\dots,K$. Classification is finally performed over this hidden layer by an additional softmax

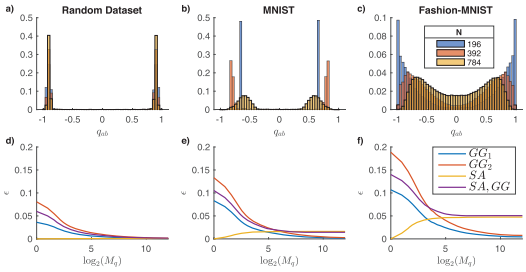

Fig. 3: Evidence of RSB in structured datasets. Upper plots: we compare the empirical overlap distribution $\mathcal{P}(q)$ obtained for the random (panel (a)), the MNIST (panel (b)), and the fashion-MNIST (panel (c)) datasets; three different item sizes are also considered, see the legend. From left to right, we move from a RS scenario where $\mathcal{P}(q)$ exhibits two peaks that get sharper as the item size increases, to a RSB scenario where $\mathcal{P}(q)$ is bimodal but with increasing broadness as the item size increases. Lower plots: we report the violation of the Ghirlanda-Guerra identities (GG_1, GG_2) and the violation of self-averaging SA as obtained for the random (panel (d)), the MNIST (panel (e)) and the fashion-MNIST (panel (f)) datasets. Again from left to right we move from a RS scenario where the self-averaging relations hold and the Ghirlanda-Guerra relations (corresponding to trivial identities) are fast vanishing, to a picture resembling RSB, where self-averaging does not hold any longer but the Ghirlanda-Guerra relations are still preserved (this time in a non-trivial manner). See the SM for further explanation.

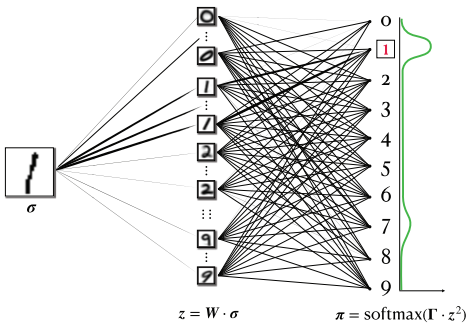

Fig. 4: Schematic representation of a three-layer RBM for the MNIST dataset based on RSB hierarchy. From left to right: visible layer $\sigma \in \{-1,+1\}^N$ receiving digits to be classified (raw data); hidden layer $\mathbf{z} \in \mathbb{R}^K$ where each node corresponds to a pseudo archetype as sketched (1-RSB effective representation); softmax layer $\pi \in [0,+1]^K$ for classification (RS effective representation).

layer $\boldsymbol{\pi} = \mathrm{softmax}[\boldsymbol{\Gamma} \cdot (\boldsymbol{W} \cdot \boldsymbol{\sigma})^2] \in [0,1]^K$, where $\boldsymbol{\Gamma}$ can again be determined by simple, algebraic operations over the training set, see fig. 4 and the SM. Notice that the determination of the weights $\boldsymbol{W}$ and $\boldsymbol{\Gamma}$ is again "one-shot" and does not require any lengthy extremization procedure.

The rationale underlying this scheme is that we want to achieve a "simplified" representation of data that can be supplied to the classifier: each sub-sample in the structureless case displays a RS representation that allows for an identification between the class and the archetype and therefore for a direct classification; conversely, in

the MNIST and in the fashion-MNIST datasets each sub-sample exhibits an intrinsic organization, much as like there were several (pseudo) archetypes for each class in such a way that we need (at least) one extra layer to lift them before classifying them. This procedure can be iterated so to establish a connection between more and more abstract representations in deep learning layers and more and more general representations in RSB steps, hence moving from the leafs (items) toward the common ancestor (archetype). Remarkably, this interpretation offers a new perspective to understand the success of deep learning architectures [2,27] in analyzing complex datasets by relating the natural phylogenetic organization of (some) of them, see, *e.g.*, [28,29], to hierarchical clustering in the RSB-scheme. This can be particularly intriguing as, once fed the Hopfield model with real datasets, we have also established a one-to-one connection between its cost function (expected to give rise to a RSB-picture of the free energy landscape) and the typical loss functions used in machine learning.

The machine obtained in this way has been tested over the two benchmark datasets obtaining an accuracy of about 95% for MNIST and 84% for fashion-MNIST, to be compared with, respectively, 75% and 63% obtained for the simple (RS) machine, see figs. S6, S7 in the SM.

* * *

The authors are grateful to MOST (Ministry of Science, Technology and Space in Israel) and MAECI (Ministero degli Affari Esteri e della Cooperazione Internazionale in Italy) for the shared grant "BULBUL" (F85F21006230001). EA acknowledges financial support from Sapienza University of Rome (RM120172B8066CB0). FA acknowledges partial fundings by PON R&I (ARS01-00876).

Data availability statement: No new data were created or analysed in this study.

REFERENCES

[1] Hopfield J. J., *Proc. Natl. Acad. Sci. U.S.A.*, **79** (1982) 2554.
[2] LeCun Y., Bengio Y. and Hinton G., *Nature*, **521** (2015) 7553.
[3] Carleo G. *et al.*, *Rev. Mod. Phys.*, **91** (2019) 045002.
[4] Agliari E., Barra A., Sollich P. and Zdeborova L., *J. Phys. A: Math. Theor.*, **53** (2020) 500401.
[5] Barra A., Bernacchia A., Santucci E. and Contucci P., *Neural Netw.*, **34** (2012) 1.
[6] Mezard M., *Phys. Rev. E*, **95** (2017) 022117.
[7] Cocco S., Monasson R. and Sessak V., *Phys. Rev. E*, **83** (2011) 051123.
[8] Uzan H. *et al.*, *Sci. Rep.*, **9** (2019) 1.
[9] Benedetti M., Ventura E., Marinari E., Ruocco G. and Zamponi F., *J. Chem. Phys.*, **156** (2022) 104107.
[10] Amit D. J., Gutfreund H. and Sompolinsky H., *Phys. Rev. Lett.*, **55** (1985) 1530.
[11] Mezard M., Parisi G. and Virasoro M. A., *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications, World Scientific Lecture Notes in Physics* (World Scientific) 1987.
[12] Gardner E., Wallace D. J. and Stroud N., *J. Phys. A: Math. Theor.*, **22** (1989) 12.
[13] Fontanari J. F., *J. Phys. (Paris)*, **51** (1990) 2421.
[14] Seung H. S., Sompolinsky H. and Tishby S., *Phys. Rev. A*, **45** (1992) 6056.
[15] Wong M. and Sherrington D., *Phys. Rev. E*, **47** (1993) 4465.
[16] Deng L., *IEEE Signal Process. Mag.*, **29** (2012) 141.
[17] Xiao H., Rasul K. and Vollgraf R., arXiv:1708.07747 (2017).
[18] Meir Y. *et al.*, *Sci. Rep.*, **10** (2020) 1.
[19] Leonelli F. E., Agliari E., Albanese L. and Barra A., *Neural Netw.*, **143** (2021) 314.
[20] Agliari E., Alemanno F., Barra A. and De Marzo G., *Neural Netw.*, **148** (2022) 232.
[21] Hinton G. E., *Neural Comput.*, **14** (2002) 1771.
[22] Karakida R., Okada M. and Amari S., *Neural Netw.*, **79** (2016) 78.
[23] Haussler D. *et al.*, *Mach. Learn.*, **25** 21996195.
[24] Coolen A. C. C., Kühn R. and Sollich P., *Theory of Neural Information Processing Systems* (Oxford Press) 2005.
[25] Schneidman E. *et al.*, *Nature*, **440** (2006) 1007.
[26] Ghirlanda S. and Guerra F., *J. Phys. A: Math. Theor.*, **31** (1998) 9149.
[27] Mehta P. and Schwab D. J., arXiv:1410.3831 (2014).
[28] Chongli Q. and Colwell L. J., *Proc. Natl. Acad. Sci. U.S.A.*, **115** (2018) 690.
[29] Ronan T., Qi Z. and Neagle K. M., *Sci. Signal.*, **9** (2016) re6.