Neural Networks 173 (2024) 106174

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Hebbian dreaming for small datasets

Elena Agliari [a], Francesco Alemanno [d], Miriam Aquaro [a], Adriano Barra [d,*], Fabrizio Durante [c], Ido Kanter [b]

[a] *Department of Mathematics of Sapienza Università di Roma, Rome, Italy*
[b] *Department of Physics of Bar-Ilan University, Ramat Gan, Israel*
[c] *Department of Economic Sciences of Università del Salento, Lecce, Italy*
[d] *Department of Mathematics and Physics of Università del Salento, Lecce, Italy*

ARTICLE INFO

Keywords:
Hebbian learning
Sleeping phenomena
Statistical mechanics
Hopfield model

ABSTRACT

The dreaming Hopfield model constitutes a generalization of the Hebbian paradigm for neural networks, that is able to perform on-line learning when "awake" and also to account for off-line "sleeping" mechanisms. The latter have been shown to enhance storing in such a way that, in the long sleep-time limit, this model can reach the maximal storage capacity achievable by networks equipped with symmetric pairwise interactions. In this paper, we inspect the minimal amount of information that must be supplied to such a network to guarantee a successful generalization, and we test it both on random synthetic and on standard structured datasets (*i.e.*, MNIST, Fashion-MNIST and Olivetti). By comparing these minimal thresholds of information with those required by the standard (*i.e.*, always "awake") Hopfield model, we prove that the present network can save up to $\sim 90\%$ of the dataset size, yet preserving the same performance of the standard counterpart. This suggests that sleep may play a pivotal role in explaining the gap between the large volumes of data required to train artificial neural networks and the relatively small volumes needed by their biological counterparts. Further, we prove that the model Cost function (typically used in statistical mechanics) admits a representation in terms of a standard Loss function (typically used in machine learning) and this allows us to analyze its emergent computational skills both theoretically and computationally: a quantitative picture of its capabilities as a function of its control parameters is achieved and consistency between the two approaches is highlighted.

The resulting network is an associative memory for pattern recognition tasks that learns from examples on-line, generalizes correctly (in suitable regions of its control parameters) and optimizes its storage capacity by off-line sleeping: such a reduction of the training cost can be inspiring toward sustainable AI and in situations where data are relatively sparse.

1. Introduction

The investigations led in this paper are guided by a central question in Machine Learning: why do artificial neural networks require many more training examples than biological neural networks do in order to form their own representations and thus correctly generalize? Given the significant footprint of extensive AI training (Hao, 2019; Strubell, Ganesh, & McCallum, 2019), this theoretical question has also practical implications: understanding how to design networks and algorithms that consume less and less, still preserving performances, is a nowadays priority. Further, beyond this ethic challenge, there can be a number of real scenarios where large datasets are not available at all, hence the need for neural networks that can be trained by minimal amounts of information.

Here, we approach this issue by tools pertaining to the *statistical-mechanics of neural networks* (Agliari, Barra, Sollich, & Zdeborova, 2020; Amit, 1992; Carleo et al., 2019; Coolen, Kühn, & Sollich, 2005; Engel & Van den Broeck, 2001; Kirkpatrick, Gelatt, & Vecchia, 1983; Marino, Parisi, & Ricci-Tersenghi, 2016; Seung, Sompolinsky, & Tishby, 1992) and looking for inspiration in the mechanisms that make *biological neural networks* particularly effective (Andrillon, Pressnitzer, Leger, & Kouider, 2018; Crick & Mitchinson, 1983; Diekelmann & Born, 2010; Maquet, 2001; McGaugh, 2000; Paton, Belova, Morrison, & Salzman, 2006). This way, we are able to show that, by implementing "sleeping" mechanisms (Agliari, Alemanno, Barra, & Fachechi, 2019; Fachechi, Agliari, & Barra, 2019; Fachechi, Barra, Agliari, & Alemanno, 2022) (a

[*] Corresponding author.
E-mail addresses: agliari@mat.uniroma1.it (E. Agliari), adriano.barra@unisalento.it (A. Barra).

https://doi.org/10.1016/j.neunet.2024.106174
Received 20 July 2023; Received in revised form 2 January 2024; Accepted 9 February 2024
Available online 12 February 2024

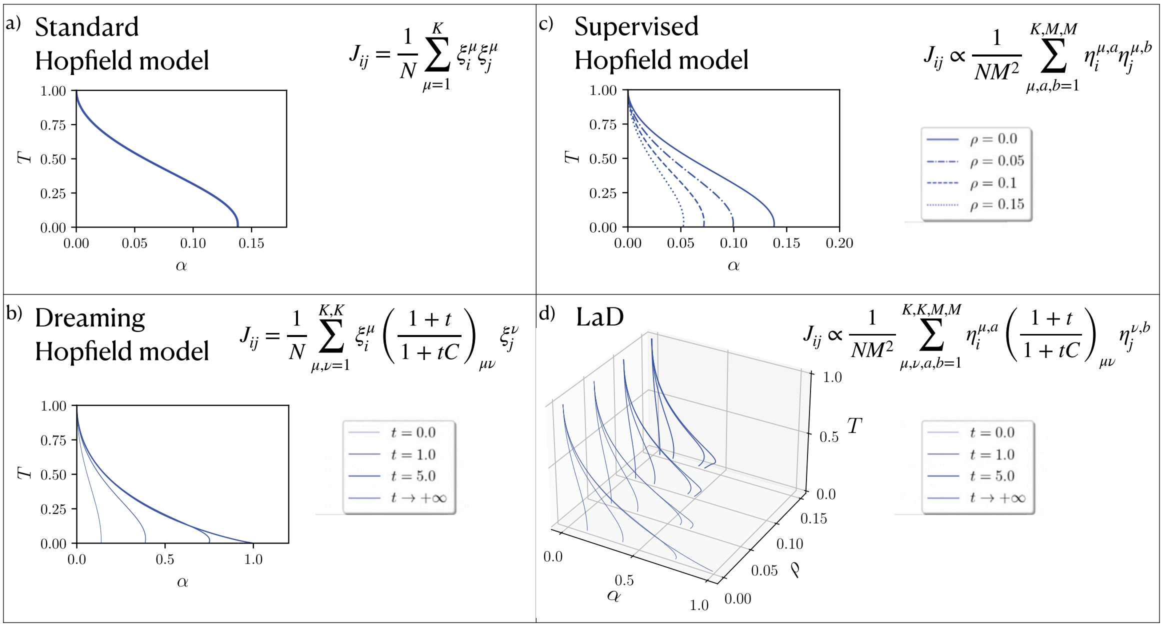

Fig. 1. In this scheme we outline the *standard Hopfield model* (panel *a*), the two extensions corresponding to, respectively, the so-called *dreaming Hopfield model* (panel *b*) and the *supervised Hopfield model* (panel *c*), as well as the LaD model under investigation (panel *d*) that stems from merging the last two. In each case, we report the expression for the interaction strength J_{ij} between the generic neurons i and j, which basically defines the model, along with the main results in terms of phase diagram. Specifically, for the standard Hopfield model, the phase diagram is drawn in the space (α, T) of the tuneable parameters, being α the network load and T the degree of noise: below (above) the transition line the network is (not) able to retrieve; for its extensions an additional control parameter is introduced: respectively, the sleeping time t and the dataset entropy ρ, thus, retaining the plane (α, T), several transition lines are drawn corresponding to different choices of the additional parameter. Notice that, by increasing t, the retrieval phase is broadened and, by increasing ρ, the retrieval phase shrinks. Finally, for the model under investigation, the control parameters are four and, accordingly, the phase diagram accounts for the ability of the system to retrieve as (α, T, t, ρ) are tuned.

biological essential function) also in artificial neural networks, the minimal size of the training set that guarantees a secure learning diminishes up to $\sim 90\%$, with consequent huge computational savings. This result is obtained by merging two extensions of the Hopfield paradigm (see *e.g.*, Amit, Gutfreund, and Sompolinsky (1985), Coolen et al. (2005) and several variations on theme (Kermiche, 2020; Kobayashi, 2013; Pu, Yi, & Zhou, 2017; Tanaka et al., 2020)) that have been recently developed (see Fig. 1):

In the first extension, named *dreaming Hopfield model*, the standard Hebbian coupling among neurons is revised to mimic consolidation (of pure memories) and remotion (of spurious mixtures) mechanisms occurring during sleep (Agliari, Acquaro, Alemanno, & Fachechi, 2023; Agliari et al., 2019; Fachechi et al., 2019, 2022). More specifically, we introduce a scalar parameter t, interpreted as "sleep time", which tunes the effective correlation between the patterns to be retrieved: the resulting network is able to store patterns on-line and optimize their memory allocation during off-line rearrangement of synapses (*i.e.* weights), in such a way that the number of storable patterns shifts from 0.14 pattern per neurons (standard Hopfield) to 1 pattern per neuron (the theoretical upper bound for symmetric networks (Amit, 1992)).

The second extension, named *supervised Hopfield model,* revises the Hebbian scenario in order to work with noisy versions of patterns (referred to as *examples*) instead of the original patterns (referred to as *archetypes*) (Agliari, Alemanno, Barra, & De Marzo, 2002; Alemanno et al., 2023; Fontanari, 1990): in this setting the focus is on the network ability to generalize the experienced information, namely on the ability to infer the archetypes out of the supplied examples rather than on the ability to store. Thus, a machine-learning framework is recovered: while the original Hopfield model stores definite patterns of information, in machine learning we typically have datasets for training the network and making it able to infer the patterns of information encoded in the datasets.

Here, by merging these extensions we obtain an outperforming model, referred to as "Learning and Dreaming" (LaD) network, that is able to form its own representation of the sampled reality, perfectly reconstructing the archetypes if the supplied information is *enough* – where *enough* is made explicit in terms of the quality and the size of the supplied dataset – and that can "take some rest" to better re-organize the storage of what has been learnt during its awake activities.

As a sideline note, we stress that the statistical–mechanical analysis accomplished here allows us to obtain *phase diagrams*, namely plots in the space of the control parameters where different operational modes of the network under study are represented as regions split by computational phase transitions (see Fig. 1) much as like the different phases of water are split in ice, liquid and vapor by physical phase transitions in its phase diagram. Remarkably, this knowledge can drive the data scientists to design suitable settings *a priori.*

The paper is structured into a main text, where we report the major computational and analytical findings and an extensive Supplementary Material (SM) where we collect all the technical details and long mathematical proofs.

2. The learning and dreaming neural network

Before presenting the LaD model and the observables useful to assess its performance, it is convenient to introduce the synthetic dataset that shall be considered in our analytical investigation.

We define K binary patterns, each of length N and denoted as $\xi^\mu \in \{-1,+1\}^N$ for $\mu = 1, \dots, K$, whose entries are drawn i.i.d. from a Rademacher distribution, *i.e.*,

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

for any $i = 1, \dots, N$ and any $\mu = 1, \dots, K$. These patterns play as archetypes and, for each of them, we generate M corrupted examples, obtained by flipping randomly the archetype pixels, that is, the ith entry of the ath example related to the μth archetype is denoted as $\eta_i^{\mu,a}$ and is defined as $\eta_i^{\mu,a} = \xi_i^\mu \chi_i^{\mu,a}$, being $\chi_i^{\mu,a} \in \{-1, 1\}$ a Bernoullian random variable parameterized by $r \in (0, 1]$, *i.e.*,

$$\mathbb{P}(\chi_i^{\mu,a} = x) = \left(\frac{1+r}{2} \delta(x-1) + \frac{1-r}{2} \delta(x+1) \right), \tag{2}$$

for any $i \in (1, \dots, N)$, $\mu \in (1, \dots, K)$, and $a = (1, \dots, M)$. Note that the average of $\chi_i^{\mu,a}$ is r and the latter tunes the quality of the dataset: if $r = 1$, each example coincides with the related archetype, whereas, if $r \to 0$, each example gets, on average (over χ), orthogonal to the related archetype.

The dataset parameters M and r can be properly combined into $\rho := (1 - r^2)/(M r^2)$ that, with a slight abuse of language, shall be referred to as *dataset entropy*. In fact, intuitively, for $M \gg 1$, we can approximate the mean of the examples belonging to the μth class as

$$\frac{1}{M} \sum_{a=1}^{M} \xi_i^\mu \chi_i^{\mu,a} \sim \sqrt{\Gamma} X \text{ with } X \sim \mathcal{N}(0, 1), \tag{3}$$

where $\Gamma := r^2 + \frac{1}{M}(1 - r^2)$; the Shannon differential entropy $\mathcal{H}$ of the variable $\sqrt{\Gamma} X$ is

$$\mathcal{H}(\sqrt{\Gamma} X) = \ln(\sqrt{2\pi\Gamma}) + \frac{1}{2} = \frac{1}{2} \ln\left[2\pi r^2 (1+\rho)\right] + \frac{1}{2}, \tag{4}$$

to be compared with the differential entropy of a perfect dataset, corresponding to setting $M \to +\infty$ or $r \to 1$ in such a way that $\rho \to 0$, that reads as

$$\lim_{\rho \to 0} \mathcal{H}(\sqrt{\Gamma} X) = \frac{1}{2} \ln(2\pi r^2) + \frac{1}{2}. \tag{5}$$

Then, by evaluating the difference between these expressions we get

$$\Delta\mathcal{H} = \frac{1}{2} \log(1 + \rho) \tag{6}$$

which is a measure of network's ignorance on archetypes given the available datasets. Thus, when $r \to 0$, $\Delta\mathcal{H}$ remains finite only if M grows at a rate $\frac{1}{r^2}$, while if $r \to 0$ and M grows at a faster rate or if $r = 1$, then $\Delta\mathcal{H}$ is vanishing.[1] An alternative discussion on the role of ρ, via Hoeffding's inequality, is provided in Section 1 of the SM.

Let us now turn to the network, that is made of N nodes, fully-connected; each node represents a neuron whose state is binary and denoted by σ_i for $i = 1, \dots, N$. Then, we introduce an *energy function* (or Cost function, or Hamiltonian) that maps any neural configuration $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N) \in \{-1, +1\}^N$ onto a real number, intuitively representing some "cost" associated with that configuration and which is given by

Definition 1. The Cost function of the LaD model is

$$\mathcal{E}_{N,M,K}(\boldsymbol{\sigma}|\boldsymbol{\eta}, t) = -\frac{1}{2N} \sum_{i,j=1}^{N,N} J_{ij}(\boldsymbol{\eta}, t)\, \sigma_i \sigma_j \tag{7}$$

where the synaptic matrix $\boldsymbol{J}$ is symmetric (*i.e.*, $J_{ij} = J_{ji}$) with entries

$$J_{ij}(\boldsymbol{\eta}, t) = \frac{1}{\Gamma M^2} \sum_{a,b=1}^{M,M} \sum_{\mu,\nu=1}^{K,K} \eta_i^{\mu,a} \left(\frac{1+t}{1+tC} \right)_{\mu\nu} \eta_j^{\nu,b}, \tag{8}$$

$t \in \mathbb{R}^+$ is the sleeping time, and C is the correlation matrix with entries

$$C_{\mu\nu}(\boldsymbol{\eta}) := \frac{1}{N} \frac{1}{\Gamma} \sum_{i=1}^{N} \left(\frac{1}{M} \sum_{a=1}^{M} \eta_i^{\mu,a} \right) \left(\frac{1}{M} \sum_{b=1}^{M} \eta_i^{\nu,b} \right). \tag{9}$$

The standard Hopfield model (see *e.g.*, Amit et al. (1985), Coolen et al. (2005)) is recovered when $r = 1$ (*i.e.*, archetypes and examples do coincide and are generically referred to as *patterns*) and $t = 0$ (*i.e.*, sleeping mechanisms are not at work). The supervised Hopfield model (Alemanno et al., 2023), where the Hebbian learning is built over examples instead of archetypes, is recovered for $t = 0$. The dreaming Hopfield model (Agliari et al., 2019; Fachechi et al., 2019), where archetypes are available and the correlation matrix was built over archetypes, corresponds to $r = 1$ and t finite. We can also recover Kohonen's decorrelation rule (see *e.g.*, Kanter and Sompolinsky (1987), Kohonen (1984), Personnaz, Guyon, and Dreyfus (1985)) meant to diagonalize patterns (and therefore reduce their interference so to improve the network capacity), by setting $r = 1$ and $t \to \infty$[2]: it is instructive to inspect these limits as reported in Fig. 1.

Remark 1. The interaction matrix (8) can be looked at as the result of the following evolution rule

$$\begin{cases} \frac{d\boldsymbol{J}}{dt} = \frac{\boldsymbol{J} - \boldsymbol{J}^2}{1+t} \\ J_{ij}(0) = \frac{1}{\Gamma M^2} \sum_{a,b=1}^{M,M} \sum_{\mu=1}^{K} \eta_i^{\mu,a} \eta_j^{\mu,b}, \quad 1 \le i, j \le N \end{cases} \tag{10}$$

where the synaptic time scale is meant to be much larger than that characterizing neuronal dynamics (consistently with the fact that synaptic plasticity is a relatively slow process, see *e.g.* Amit (1992)). In the previous equation, the evolution of the coupling matrix results from the interplay of consolidation and remotion mechanisms (corresponding to the positive and negative contribution in the evolution equation, respectively) that are inspired by analogous mechanisms occurring in mammal's brain during sleep. For this reason, the resulting model is referred to as "dreaming Hopfield model" and t as the "sleeping time". The underlying presence of consolidation and remotion effects also appears neatly by looking at the solution of (10), that is, by looking directly at (8), where, as the sleeping time increases, the numerator in the kernel $\left(\frac{1+t}{1+tC}\right)_{\mu\nu}$ plays a role in consolidating retrieval states, while the denominator tends to remove spurious memories: we refer to the original papers (Agliari et al., 2019; Fachechi et al., 2019) (and references therein) for an in-depth explanation of this system and its relation with actual sleeping and dreaming mechanisms in mammals.

The introduction of the Cost function (7) can be justified on different grounds (see *e.g.*, Amit (1992), Coolen et al. (2005) and Section 2 in the SM for a derivation based on the maximum entropy statistical inference) and it rules the dynamics of the neural configuration as described in Algorithm 1. Under this dynamics the system eventually reaches a stationary state characterized by the Boltzmann–Gibbs measure

$$\mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{\eta}, t) = \frac{\exp[-\beta \mathcal{E}_{N,M,K}(\boldsymbol{\sigma}|\boldsymbol{\eta}, t)]}{\mathcal{Z}_{N,M,K}(\beta|\boldsymbol{\eta}, t)}, \tag{11}$$

where $\mathcal{Z}_{N,M,K}(\beta|\boldsymbol{\eta}, t)$ is a normalization factor, also called *partition function* and defined as

$$\mathcal{Z}_{N,M,K}(\beta|\boldsymbol{\eta}, t) = \sum_{\{\sigma\}}^{2^N} e^{-\beta \mathcal{E}_{N,M,K}(\sigma|\boldsymbol{\eta}, t)} \tag{12}$$

where $\beta := 1/T \in \mathbb{R}^+$ tunes the degree of stochasticity (or thermal noise) in the network: for $\beta \to 0$ (infinite noise limit) the Boltzmann–Gibbs measure becomes uniformly distributed over the neural configurations, while in the opposite limit $\beta \to \infty$ (zero fast noise) the probability distribution peaks at the cost-function minima.

The neural network described so far can be used for reconstruction tasks: being $\boldsymbol{\sigma}^{(0)}$ the initial configuration corresponding to some perturbed version of, *e.g.*, $\boldsymbol{\xi}^\mu$, we say that the system is able to reconstruct the archetype if the neural dynamics reaches a stationary state $\boldsymbol{\sigma}^* = \boldsymbol{\xi}^\mu$

[1] The monotonic relation highlighted for $\Delta\mathcal{H}$ and ρ, that allows us to refer to ρ as the dataset entropy, was proved for structureless datasets defined according to Eqs. (1)–(2) and, in the analytical investigations carried on for this dataset, it turns out to be a key control parameter. In principle, the same quantities can be evaluated also for structured datasets, although the relation (6) would not hold in general.

[2] In the present case, as archetypes are unavailable to the network, the matrix C is built over examples, however, in the regime of large number of examples $M \gg 1$, we have $\frac{1}{M} \sum_{a=1}^{M} \chi_i^{\mu,a} \xi_i^\mu \approx r\xi_i^\mu$, where we approximated $\frac{1}{M} \sum_{a=1}^{M} \chi_i^{\mu,a}$ with the mean of $\chi_i^{\mu,a}$, thus, in this limit, the kernel $(\frac{1+t}{1+tC})$ is effectively decorrelating the archetypes.

Algorithm 1 Sequential dynamic

```
Input: Couplings $\boldsymbol{J} \in \mathbb{R}^{N\times N}$, input $\boldsymbol{\sigma}^{(0)} \in \{-1,1\}^N$, number of dynamic steps $N_s$, noise $T$
Output: Final neuronal configuration $\boldsymbol{\sigma}^*$
Remove the diagonal terms from $\boldsymbol{J}$
Set $i = 0$
repeat
    sample a random integer $n$ uniformly in the set $\{1, 2, \dots, N\}$
    sample a random variable $\zeta$ from the distribution $\frac{1}{2}(1 - \tanh^2 \zeta)$
    update the nth spin $\sigma_n$ according to $\sigma_n = sign(\sum_{j=1}^N J_{nj}\sigma_j + T\zeta)$
    $i = i + 1$
until $i = N_s$
```

Table 1
Operative implementations of the Mattis overlap.

Observable	Description
m_μ	Mattis overlap: standard definition with respect to the archetype (see eq. (13))
$\hat{m}_\mu$	Mattis overlap: empirical estimate over the examples (see eq. (14))
m^*_μ	Mattis overlap: numerical evaluation (see algorithm (2))
$\bar{m}_\mu$	Mattis overlap: analytical estimate (see eq. (19))

(or, at least, $\boldsymbol{\sigma}^* \approx \boldsymbol{\xi}^\mu$, with some tolerance threshold); the symmetric configuration $\boldsymbol{\sigma}^* = -\boldsymbol{\xi}^\mu$ is also retained as a retrieval. The Mattis overlap, defined as (Amit et al., 1985; Coolen et al., 2005)

$$m_\mu(\boldsymbol{\sigma}) := \frac{1}{N}\sum_{i=1}^N \xi_i^\mu \sigma_i, \tag{13}$$

quantifies how close a neural configuration $\boldsymbol{\sigma}$ is to a given archetype $\boldsymbol{\xi}^\mu$. However, it should be recalled that the system is not aware of the archetypes, in fact, its cost function (7) and, in particular, the weights appearing therein, are built over the set of examples. One could therefore define an empirical Mattis overlap, referred to an empirical estimate $\hat{\boldsymbol{\xi}}^\mu$ of the μth archetype, as

$$\hat{m}_\mu(\boldsymbol{\sigma}) := \frac{1}{N}\sum_{i=1}^N \hat{\xi}_i^\mu \sigma_i := \frac{1}{N}\sum_{i=1}^N \Big(\frac{1}{M}\sum_{a=1}^M \eta_i^{\mu,a}\Big)\sigma_i, \tag{14}$$

in such a way that $\mathcal{E}_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\eta}, t) = -N/\Gamma \sum_{\mu,\nu} \hat{m}_\mu (\frac{1+t}{1+tC})_{\mu\nu} \hat{m}_\nu$.

The definition of the weights $\boldsymbol{J}$ given in (8) can then be interpreted as a learning rule, the set of examples $\{\boldsymbol{\eta}^{\mu,a}\}_{a=1,\dots,M}^{\mu=1,\dots,K}$ as a training set, and the sample of initial configurations $\boldsymbol{\sigma}^{(0)}$ as a test set (these can be taken as corrupted versions of the archetypes, not belonging to the training set). Thus, we can look for the conditions under which the training underlying this network is sufficient to ensure the task accomplishment, by introducing a Loss function that measures to what extent the task is accomplished. Since we want the network to reconstruct the archetypes $+\boldsymbol{\xi}^\mu$, or their symmetric copies $-\boldsymbol{\xi}^\mu$, a natural L_2 Loss function is

$$\mathcal{L}_\pm^\mu = \frac{1}{2N}\left\|\boldsymbol{\xi}^\mu \pm \boldsymbol{\sigma}^*\right\|^2 = \frac{1}{2N}\|\boldsymbol{\xi}^\mu\|^2 + \frac{1}{2} \pm m^*_\mu, \tag{15}$$

for all $\mu \in (1, \dots, K)$. In this expression the dependence on the weight arrangement and, therefore, on the training dataset is implicit in $\boldsymbol{\sigma}^*$: this can be seen theoretically – the $\boldsymbol{\sigma}^*$'s are drawn from (11) – or numerically – the $\boldsymbol{\sigma}^*$'s result from the dynamics in Algorithm 1. Now, it is convenient to rotate the archetypes into $\tilde{\boldsymbol{\xi}}^\mu = \sum_\nu \left(\sqrt{\frac{1+t}{1+tC}}\right)_{\mu\nu} \boldsymbol{\xi}^\nu$, whence the related Mattis overlap $\tilde{m}^*_\mu = \frac{1}{N}\sum_i \tilde{\xi}_i^\mu \sigma_i^*$, and the related loss

$$\tilde{\mathcal{L}}_\pm^\mu = \frac{1}{2N}\left\|\tilde{\boldsymbol{\xi}}^\mu \pm \boldsymbol{\sigma}^*\right\|^2 = \frac{1}{2N}\left\|\tilde{\boldsymbol{\xi}}^\mu\right\|^2 + \frac{1}{2} \pm \tilde{m}^*_\mu, \quad \mu = 1, \dots, K, \tag{16}$$

follow. Remarkably, this rotation does not affect learning – since $\text{sign}(\tilde{\boldsymbol{\xi}}^\mu) = \text{sign}(\boldsymbol{\xi}^\mu)$, the optimal configuration (in the domain $\{-1,1\}^N$) remains $\boldsymbol{\sigma}^* = \pm\boldsymbol{\xi}^\mu$ – but it results particularly convenient because it highlights a direct relation between the Loss function (typically used in a machine-learning framework) and the Cost function $\mathcal{E}_{N,K,M=1}(\boldsymbol{\sigma}|\boldsymbol{\xi}, t)$ evaluated at the archetypes (typically used in a pattern-retrieval framework) as

$$\begin{aligned}\mathcal{E}_{N,K,M=1}(\boldsymbol{\sigma}|\boldsymbol{\xi}, t) &= -\frac{1}{2N}\sum_{i,j=1}^N \sum_{\mu,\nu=1}^{K,K} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{1+tC}\right)_{\mu\nu} \sigma_i\sigma_j = \\ &= -\frac{1}{2N}\sum_{i,j=1}^N\sum_{\mu=1}^K \tilde{\xi}_i^\mu \tilde{\xi}_j^\mu \sigma_i\sigma_j = \text{const}(t) + \frac{N}{2}\sum_{\mu=1}^K \tilde{\mathcal{L}}_+^\mu \tilde{\mathcal{L}}_-^\mu,\end{aligned} \tag{17}$$

with $\text{const}(t) = -\frac{N}{8}\sum_{\mu=1}^K (\frac{1}{N}\|\tilde{\boldsymbol{\xi}}^\mu\|^2 + 1)^2$. For both $\tilde{\mathcal{L}}_\pm^\mu$ and $\mathcal{E}_{N,K,M=1}(\boldsymbol{\sigma}|\boldsymbol{\xi}, t)$, the extremization is reached when the Mattis overlap m_μ is maximal. Replacing $\boldsymbol{\xi}^\mu$ in (16) and (17) with its empirical estimate $\hat{\boldsymbol{\xi}}^\mu$ we recover $\mathcal{E}_{N,K,M}(\boldsymbol{\sigma}|\boldsymbol{\eta}, t)$. In the large dataset scenario $M \gg 1$, the empirical estimate $\hat{\boldsymbol{\xi}}^\mu$ over examples belonging to the μth class can be approximated as $\hat{\boldsymbol{\xi}}^\mu \approx r\boldsymbol{\xi}^\mu$, thus, both $\mathcal{L}_\pm^\mu$ and its empirical version $\hat{\mathcal{L}}_\pm^\mu = \frac{1}{2N}\left\|\hat{\boldsymbol{\xi}}^\mu \pm \boldsymbol{\sigma}^*\right\|^2$ are minimal when $m_\mu = 1$. Further, one could notice that, being strongly related and fully consistent with the mean-squared-error (MSE), the Mattis overlap can also be related to other measures of performance like the peak signal-to-noise ratio (PSR), being PSNR $:= 20\log_{10}\left(\frac{2}{\sqrt{\text{MSE}}}\right)$ (Horé & Ziou, 2010). We refer to Section 3 in the SM for further details on the mapping between Cost and Loss functions.

To summarize, the conditions yielding to a large Mattis overlap ensure retrieval as well as learning, and, in this setting, these are just two faces of the same medal. With this framework in mind, in the following we investigate the LaD model from two different perspectives: first we inspect for the minimal dataset-size M_c which permits correct retrieval, checking whether and, if so, to what extent, sleeping mechanisms are helpful in reducing this minimal size; next, we look for phase diagrams which overall summarize the network performance versus its control parameters. Before proceeding it is also worth recalling that, in the remaining of this paper, we will settle on the so-called *high-load* regime, identified by a number of archetypes scaling linearly with the number of neurons, *i.e.*, $K = \alpha N$, with $\alpha \in \mathbb{R}^+$.

3. Numerical results

For the structureless dataset introduced in Section 2, analytical investigations are feasible in the limit $N \to \infty$ and the related details are collected in Section 4. Here we focus on the Mattis overlap (see Table 1 for its various operative implementations) to explore how the dataset size M impacts on LaD's capabilities (subsect. 3.1) and to outline a phase diagram (sub Section 3.2). Numerical experiments on benchmark datasets are also performed to check the robustness of our theoretical results when relaxing the structureless hypothesis.

3.1. Dataset saving

We first look at the behavior of the Mattis overlap m_μ versus the number M of examples per archetype, at zero noise $T = 0$. As shown in Fig. 2 (where analytical results obtained in the $N \to \infty$ limit are compared with results obtained from simulations at finite size), this grows monotonically and there exists a threshold M_c beyond which $\bar{m}_\mu$ starts raising from zero; the detachment from zero is steeper and steeper as the size N is increased and it is discontinuous in the $N \to \infty$ limit (and, in general, M_c depends on the dataset quality r, the network size N, the load α and the sleeping time t). This behavior makes clear the reward of the sleeping mechanism: the training-set size ensuring a satisfactory retrieval decreases exponentially with t and the effect is even more significant when the load is larger.

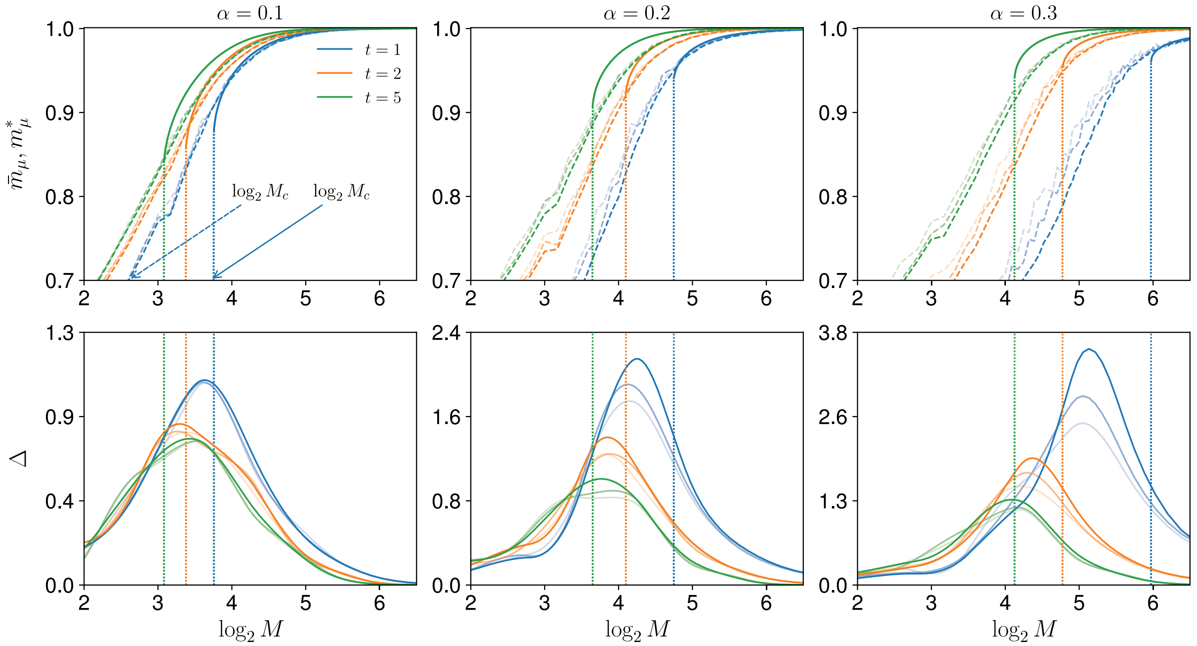

Fig. 2. Expectation of the Mattis overlap ($\bar{m}_\mu$ analytical and m^*_μ numerical) with respect to the Boltzmann–Gibbs distribution (11) (upper panels) and its numerical derivative $\Delta := dm_\mu/dM$ (lower panels), versus the logarithm of the dataset size M, in the limit $\beta \to +\infty$, for different values of the sleeping time t (shown in different colors) and of the load α (shown in different panels), keeping the quality of the dataset fixed at $r = 0.5$. Analytical predictions (solid lines) obtained for infinite network size are compared with computational simulations (dashed lines) obtained at finite size: the arrows in the first panel highlight the thresholds M_c, numerical vs analytical, for $t = 1$ as an example. In the simulations, for any value of t considered, we also varied the network size (from brighter to darker nuances, $N^2\alpha$ ranges in $[5, 10, 80] \times 10^4$), to highlight a finite-size scaling; further, the neuronal evolution is sequential, as specified in Algorithm 1, and the initial configuration is chosen as $\sigma^{(0)} = \xi^\mu$, for consistency with the analytical picture (however, we checked that results are only slightly quantitatively affected if the initial configuration is just close to the archetype, that is, if a small fraction of pixels are flipped, see Section 4). The vertical, dotted lines correspond to the theoretical estimate of M_c; in the upper (lower) panels the curves exhibit a flex (peak) that, as N grows, gets closer and closer to M_c.

In order to quantify this dataset saving, we introduce the quantity

$$S(\alpha, r, t) := \left(1 - \frac{M_c(\alpha, r, t)}{M_{\max}(\alpha, r)}\right) \tag{18}$$

where $M_{\max}(\alpha, r) = \max_{t \in \mathbb{R}^+}\{M_c(\alpha, r, t)\}$. Then, if M_c displays poor variability with respect to t, S will be close to 0; vice versa, if there exist values of t able to significantly reduce M_c, S will be close to 1. The computational estimate of this quantity is obtained by initializing the system in a configuration $\sigma^{(0)} = \xi^\mu$, then, by letting the system evolve as explained in Algorithm 1, we evaluate m^*_μ, for different choices of t, α, and M, finally, we determine the lowest value of M leading to $m^*_\mu > 0$. The procedure is also reported in Algorithm 2. The results obtained in this way are shown in Fig. 3, where, again, we get a good agreement between analytical and numerical outcomes.[3] In particular, we notice that S grows monotonically with t and saturates to a value that increases with the dataset quality r and decreases with the load α. Specifically, when $\alpha = 0.1$ and $r = 0.5$ (*i.e.*, 1/4 of the archetype pixels are flipped on average) S can reach values even larger than 0.9. In other words, when setting t without any strategy, the minimum number M_c of examples needed for letting the system retrieve can be, in the worst case, $M_{\max}$, while, by setting t properly large, this number can be a factor 10 smaller, that is, by wisely setting t the system can generalize from examples even by relying on a relatively small dataset.

We check the robustness of these results on structured datasets, such as MNIST (Deng, 2012), Fashion-MNIST (Xiao, Rasul, & Vollgraf, 2017), and Olivetti (Phillips & O'Toole, 2014). In these cases, a data pre-processing is in order, as they are originally provided in a gray-scale format, in fact, we binarized the dataset by exploiting Otsu's method[4] (Otsu, 1979). Next, we identified the archetypes according to the nature of the dataset itself: for MNIST and Fashion-MNIST datasets, where there are 10 categories, the chosen archetypes are the averages

Algorithm 2 Algorithm for assessing the dataset saving

Input: *Archetypes matrix* $\xi \in \{-1, +1\}^{N \times K}$, *threshold for successful retrieval* $m_\times$, *vector of sleeping-time values* T_{vett}.

Output: *Saving corresponding to the input sleeping-time values*

```
for $t$ in $T_{\text{vett}}$ do
    $\frac{1}{K}\sum_{\mu=1}^{K} m^*_\mu = 0,\ M_c = 0$
    while $\frac{1}{K}\sum_{\mu=1}^{K} m^*_\mu \le m_\times$ do
        $M_c = M_c + 1$
        sample the examples matrix $\boldsymbol{\eta} \in \{-1, +1\}^{N \times K \times M}$
        evaluate the coupling matrix $J_{ij}(\boldsymbol{\eta}, t) = \frac{1}{\Gamma M^2}\sum_{a,b=1}^{M,M}\sum_{\mu,\nu=1}^{K,K}\eta_i^{\mu,a}\left(\frac{1+t}{1+tC}\right)_{\mu\nu}\eta_j^{\nu,b}$
        for $\mu$ in $(1, \dots, K)$ do
            update the neural configuration $\boldsymbol{\sigma}^{(0)} = \xi^\mu$ by Algorithm 1 until convergence to $\boldsymbol{\sigma}^*$
            evaluate the Mattis overlap $m^*_\mu = \sum_{i=1}^{N}\sigma^*_i \xi^\mu_i / N$
        compute $\frac{1}{K}\sum_{\mu=1}^{K} m^*_\mu$
Calculate $M_{\max} = \max_t M_c(t)$
Compute the saving $S(t) = (1 - M_c(t)/M_{\max}) \times 100$
```

[3] The analytical results show the existence of a threshold M_c above which $\bar{m}_\mu$ discontinuously detaches from zero. This abrupt phenomenon occurs in the thermodynamic limit $N \to \infty$, while, in simulations run at finite size, the magnetization increases continuously from zero. Consequently, in the numerical experiments, we choose a threshold $m_\times \sim 1 - 1/\mathcal{O}(N)$ on the magnetization to estimate the critical value M_c: see Algorithm 2.

[4] Briefly, Otsu's algorithm returns an intensity threshold that separates pixels of the image into two groups. This threshold is image-dependent and it is determined by minimizing the intra-group intensity variance, or equivalently, by maximizing the inter-group variance.

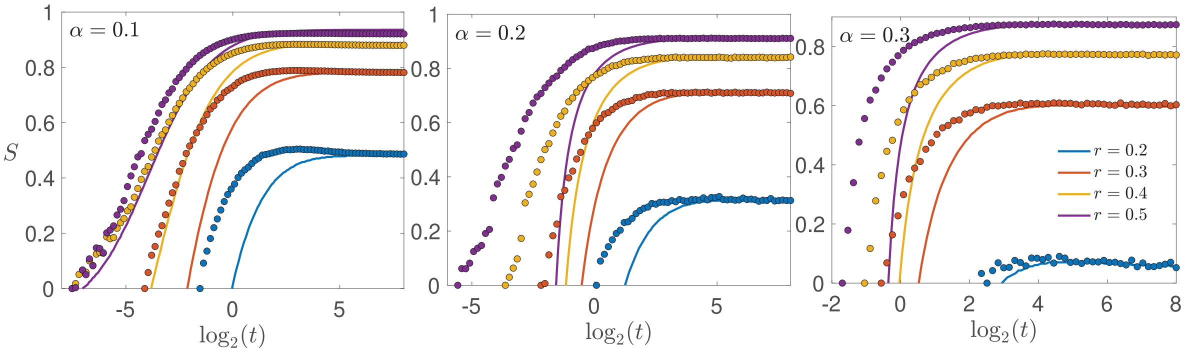

Fig. 3. Dataset saving S versus the logarithm of the sleeping time t for the random dataset. Analytical predictions (solid lines) and computational simulations (squares) are compared for different choices of α (shown in different panels), for different choices of r (as explained by the common legend) and for $\beta \to \infty$. Theoretical results are obtained by solving the equations reported in Section 4 (see Corollary 1 and Secs. V-VI in the SM) and determining M_c as the smallest value of M such that $|\bar{m}| > 0$, while numerical results are obtained by applying Algorithm 2, with a network size $N = 200$.

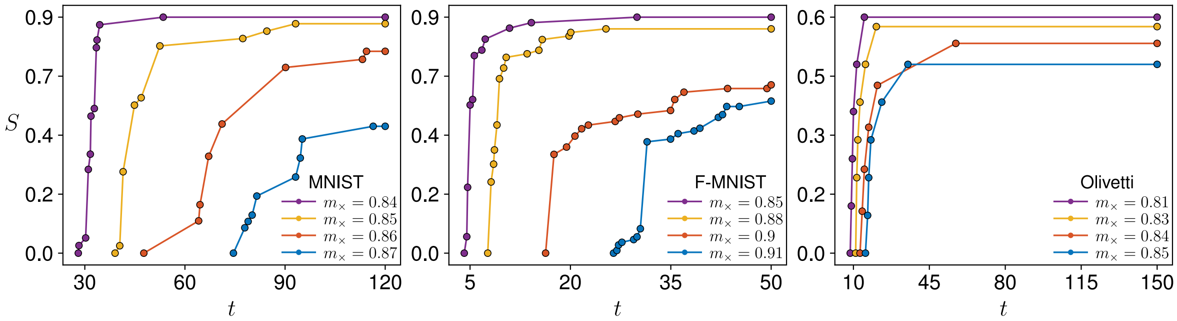

Fig. 4. Saving on the number of examples needed to reconstruct the MNIST, the Fashion-MNIST, and the Olivetti datasets, as a function of the logarithm of the sleeping time, for $\beta \to \infty$. The parameters for these datasets are $N = 784, K = 10, \alpha = 0.013$ for the MNIST and Fashion-MNIST, and $N = 4096, K = 10, \alpha = 0.01$ for the Olivetti dataset. Different choices for the threshold $m_\times$ to assess the retrieval performance are considered as explained in the legends. The smallest values of t depicted here correspond to the smallest values of t which allow the network to reach an overlap larger than $m_\times$.

of the items within the same class, so we have a total of $K = 10$ archetypes; for the Olivetti dataset, which consists of 400 photos of 40 people (10 photos for each person), we extract $K = 40$ archetypes by taking the average of the ten photos related to the same person. The training set consists of M randomly sampled examples from each of the K classes.

The first tranche of experiments is meant to obtain an estimate for S and, in analogy with the structureless case, we initialize the system in a configuration $\boldsymbol{\sigma}^{(0)} = \boldsymbol{\xi}^{\mu}$. Next, we let the system evolve as explained in Algorithm 1 and we evaluate m^*_μ; the procedure is repeated for $\mu = 1, \dots, K$ and results are averaged. Then, as explained in Algorithm 2, we compare this average with a threshold $m_\times$ and check whether it is larger (success) or smaller (failure) than $m_\times$. Unlike the structureless case, where, following analytical results, we expected a steep growth of m^*_μ from zero, here we anticipate a smooth behavior and we therefore consider several choices of $m_\times$, that is, we consider different performance thresholds. The critical dataset-size $M_c(m_\times, t)$ is then determined as the lowest number of examples that yields a successful generalization. The saving S can finally be evaluated by exploiting equation (18) and results are shown in Fig. 4. Overall these plots show that, by introducing suitably stylized sleeping mechanisms in artificial neural networks, we can retain the same performances of the "restless" counterparts, *and* save up to one order of magnitude in the required training set.

In the second tranche of experiments, in order to understand how many examples we need to ensure a good reconstruction performance, in Fig. 5 we show the Mattis overlap achieved by the system as a function of the number of examples M per archetype supplied to the network (upper panels); in particular, we start from a configuration $\boldsymbol{\sigma}^{(0)} = \boldsymbol{\eta}^{\mu}$ obtained by flipping 15% of the pixels of the μth archetype and let the system evolve as explained in Algorithm 1, then we evaluate the retrieval quality m^*_μ in terms of the overlap between the archetypes and the final configuration reached by the system. The snapshot of the reconstructed archetypes for different values of t are also provided (lower panels). By exploiting these results, we can choose the values of t that yield a good enough performance, in terms of retrieval quality as assessed by m_μ, while saving on the number M of examples, see Fig. 6.

3.2. Phase diagrams

In the previous subsection we showed that the system performance can depend qualitatively (success versus failure) on the parameter setting. While we focused on the role of M for practical reasons, similar outcomes also emerge as the load α or the noise $T = \beta^{-1}$ are tuned. For instance, as well-known for the Hopfield model (see *e.g.*, Amit et al. (1985), Coolen et al. (2005)), whenever the degree of stochasticity T affecting the neuron dynamics gets too large, the neurons are likely to be randomly oriented and the system gets useless for reconstruction tasks; similarly, there exists a critical value α_c (depending on T) such that, beyond that value, the system is no longer able to retrieve and $m^*_\mu \approx 0$ for any μ (this region is also referred as "glassy" and there stationary states correspond to mixtures between different archetypes, a signature of the fact that the stored archetypes are too many and interference between them starts to be impairing). This information can be conveniently summarized in a phase diagram, whose knowledge, as emphasized in Section 1, allows the user to set the network in the desired operational regime *a priori*, hence avoiding energy consumption for preliminary assessments. A glance at the Hopfield phase diagram (see Fig. 1, panel *a*) or Fig. 7, top-left panel, blue line) immediately reveals that it is not possible to use that network for retrieving when

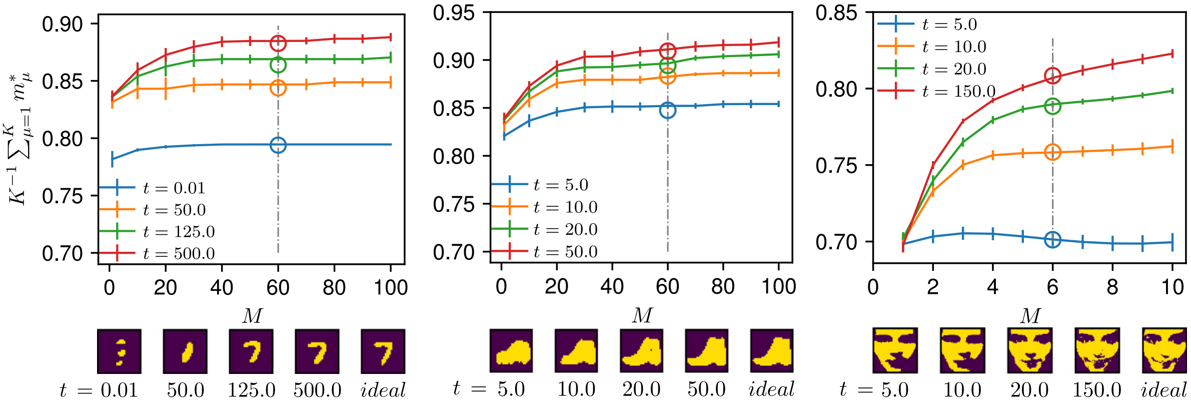

Fig. 5. Upper panels: Average Mattis overlap versus the number of examples M available for each archetype in the case of MNIST (left), Fashion-MNIST (center) and Olivetti (right) datasets. For each dataset, we initialize the system in a corrupted version of the μth archetype, obtained by flipping 15% of the pixels (*i.e.*, for $r = 0.7$), we run Algorithm 1 and evaluate m_μ^*. This procedure is repeated for $\mu = 1, \ldots, K$ (where $K = 10$ or $K = 40$, according to the dataset) and we finally report its average along with a confidence interval (corresponding to one standard deviation) as a function of M. Lower panels: Outcomes $\boldsymbol{\sigma}^*$ of the reconstruction process obtained by setting M equal to the value indicated by the vertical line in the upper panels, and for the values of t given in the corresponding legends.

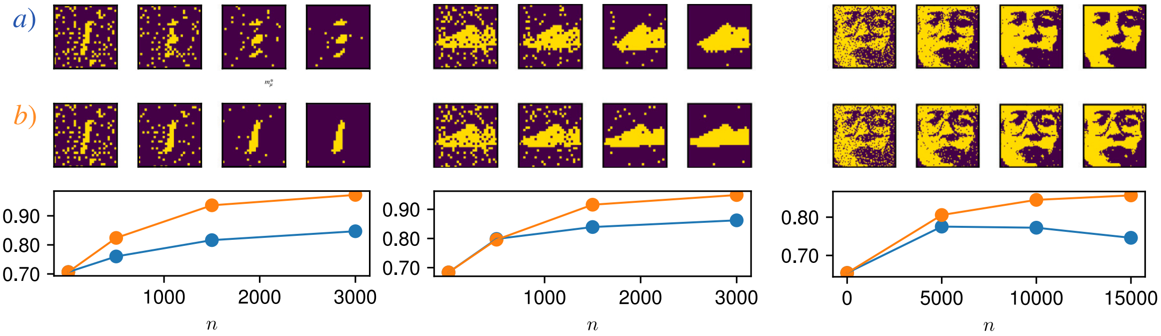

Fig. 6. Typical evolutions of the neural configurations and of the related Mattis overlaps, obtained by running Algorithm 1 for systems built over the MNIST (left), Fashion-MNIST (center) and Olivetti (right) datasets. For each dataset, we focus on two distinct cases, denoted as a and b: in case a, the parameter t is chosen equal to the minimum value shown in the legend of Fig. 5, whereas M is equal to the value indicated by the vertical line in Fig. 5; in case b, the parameter t is chosen equal to the maximum value shown in the legend of Fig. 5, whereas M is the 70% of the value chosen for case a. To be more explicit, for the MNIST dataset ($K = 10$) we have: $M = 60$, $t = 0.01$ in case a, and $M = 42$, $t = 500.0$ in case b; for the Fashion-MNIST dataset ($K = 10$) we have: $M = 60$, $t = 5.0$ in case a, and $M = 42$, $t = 50.0$ in case b; for the Olivetti dataset ($K = 40$) we have: $M = 6$, $t = 5.0$ in case a, and $M = 4$, $t = 150.0$ in case b. The result of the dynamics is shown from two complementary perspectives: in the upper panels we show a few snapshots of the neural configurations captured at different iteration times n, while in the lower panels we show the archetype overlap m_μ as a function of the number of iterations n and the instants corresponding to the snapshots above are highlighted by bullets. Overall, we notice that the reconstruction quality is significantly better in case b, although the dataset size is smaller.

$\alpha > \alpha_c \approx 0.138$ as, beyond that threshold, the network escapes the retrieval region and enters a "glassy" region (*i.e.* the so-called *blackout regime* (Amit et al., 1985)) where computational capabilities are lost (Coolen et al., 2005). This means that, if we have a Hopfield network made of, say, $N = 1000$ neurons, it is pointless to make it handle, say, $K = 500$ (random) patterns, as this would imply a value of $\alpha = 0.5 \gg \alpha_c$ and we know in advance that the network would surely fail with this load (much as like we avoid drinking water under the freezing temperature).

In the statistical mechanics framework, phase diagrams can be obtained analytically in the limit of large system-size $N \to \infty$, where transitions are associated to singularities in the free-energy function; here, for the theory to be consistent, we also need to handle the model in the regime of large dataset-size $M \gg 1$, but retaining ρ finite (see Section 4 for a detailed explanation). The control-parameter hyperspace for the LaD model is therefore given by $(\alpha, \beta, t, \rho) \in (0, 1] \times [0, \infty)^3$. As in the classical Hopfield model, while these parameters are tuned, we outline regions where the system performs qualitatively different. In particular, we focus on the transition between the so-called retrieval region – where the expected Mattis overlap is relatively close to 1 – and the blackout region – where the expected Mattis overlap is vanishing. In Figs. 7 and 8 we provide this transition line or, more precisely, its projections in, respectively, the (α, β) and (α, ρ) planes for various values of t. Remarkably, by increasing the sleeping time, the retrieval region (that is the lower region on the left) gets wider and wider and the maximum load supported by the network increases accordingly. However, by increasing the entropy in the dataset, performances decrease and, in particular, the critical storage diminishes monotonously as ρ increases, resulting in $\alpha_c(\rho = 0.00, t \to \infty, \beta \to \infty) = 1.00$ (that is the upper bound for the storage capacity for symmetric networks), $\alpha_c(\rho = 0.05, t \to \infty, \beta \to \infty) \approx 0.70$ and $\alpha_c(\rho = 0.10, t \to \infty, \beta \to \infty) \approx 0.58$.

To summarize, in the LaD model, the critical load α_c depends not only on the noise β but also on the dataset entropy ρ and on the sleeping time t; the former, as expected, is detrimental for tasks of reconstruction based on examples, while the latter turns out to be extremely beneficial.

4. Analytical results

In order to find the retrieval region in the phase diagram we need an analytical description of the computational phase transitions and, as anticipated, the theory shall be developed in the limit $N \to \infty$, which is also mathematically convenient as it allows us to neglect finite-size fluctuations. More precisely, we assume that the observables used to assess the overall state of the system (also referred to as "order parameters" and reported in Table 2, last four lines), in the limit $N \to \infty$, are no longer fluctuating and self-average around their expectation values (*i.e.*, the "replica symmetry" approximation in

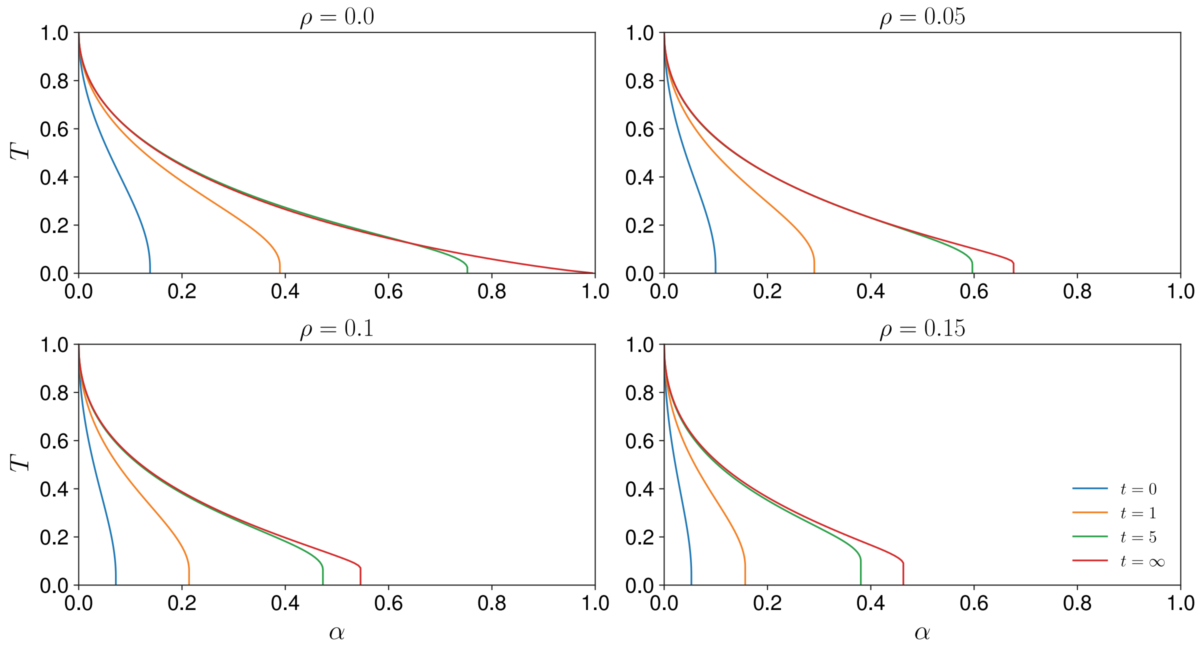

Fig. 7. Phase diagrams in the $(\alpha, T = \beta^{-1})$ plane for different values of the sleeping time t (depicted in different colors as coded in the legend) and for different values of the dataset entropy ρ (depicted in different panels as reported by titles). Specifically, we increase the entropy in the datasets from $\rho = 0$ (the full information is available, since examples and archetypes coincide and there is no learning but solely storing), to $\rho = 0.05$, $\rho = 0.10$, and finally $\rho = 0.15$. For a given choice of ρ and t, the retrieval region, where the network can retrieve archetypes, corresponds to the region underneath the related curve, the latter separates the retrieval region from the blackout scenario (glassy phase) and represents a computational phase transition. For $\rho \neq 0$ the transition line develops a cuspid when reaching the maximal storage: this is the onset of replica symmetry breaking (that will not be faced in this paper).

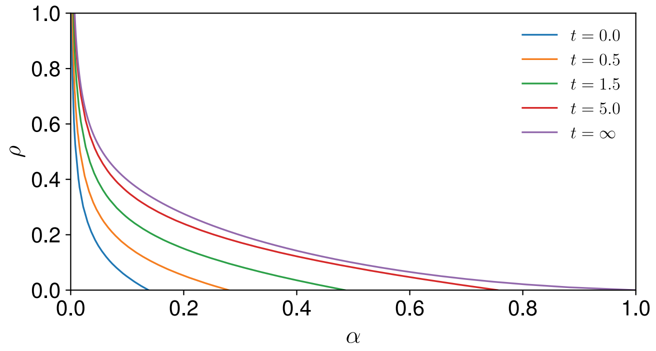

Fig. 8. Phase diagram in the (ρ, α) plane, in the zero fast-noise limit $\beta \to \infty$, for different values of the sleeping time t (shown in different colors as coded in the legend). For a given choice of t, the bottom-left region is the retrieval region, corresponding to values of α and ρ allowing the network to correctly learn and recognize the archetypes: as the sleeping time increases, this region gets wider and wider.

statistical mechanics), that shall be indicated by a bar. Thus, denoting with $\mathcal{P}(x)$ the distribution of an arbitrary observable x, we have

$$\lim_{N\to\infty} \mathcal{P}(x) = \delta(x - \bar{x}). \tag{19}$$

The key quantity for our inspection is the quenched free-energy defined as

$$\mathcal{A}_{N,M,K}(\beta, t) := \frac{1}{N} \mathbb{E} \log \mathcal{Z}_{N,M,K}(\beta | \boldsymbol{\eta}, t) \tag{20}$$

where $\mathbb{E} := \mathbb{E}_\chi \mathbb{E}_\xi$ denotes the average over the realization of the archetypes and of the examples, namely the average with respect to (1) and (2); in the $N \to \infty$ limit, recalling $\alpha := \lim_{N\to\infty} \frac{K}{N}$,

$$\mathcal{A}(\alpha, \beta, t, \rho) := \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log \mathcal{Z}_{N,M,K}(\beta | \boldsymbol{\eta}, t). \tag{21}$$

Table 2
Structural parameters, control parameters and macroscopic observables of the LaD model and of the dataset.

Parameter	Description
N	Neurons in the network
K	Archetypes to handle
M	Examples per archetype
r	Noise in the dataset
$\alpha = \lim_{N\to\infty} K/N$	Network's load
$T = \beta^{-1}$	Noise in the network
t	Sleeping time
$\rho = (1 - r^2)/Mr^2$	Dataset's entropy
$m = \frac{1}{N} \sum_i^N \xi_i^1 \sigma_i$	Mattis overlap for the archetype
$m_\eta = \frac{1}{r(1+\rho)NM} \sum_{a,i=1}^{M,N} \xi_i^1 \chi_i^{1,a} \sigma_i$	Mean Mattis overlap over the examples
$q_{12} = \frac{1}{N} \sum_i^N k_i^{(1)} k_i^{(2)}$	Spin glass order parameter I
$p_{12} = \frac{1}{K} \sum_{\mu>1}^K z_\mu^{(1)} z_\mu^{(2)}$	Spin glass order parameter II

It can be proven (see Section 5 in the SM for further details) that, in the large but finite dataset scenario $M \gg 1$ and in the thermodynamic limit $N \to \infty$, the quenched free-energy depends on the parameters M and r only through[5] ρ, thus we will write $\mathcal{A}(\alpha, \beta, t, \rho) := \lim_{N\to\infty} \mathcal{A}_{N,M,K}(\beta, t)$.

The goal now is to obtain an explicit expression for $\mathcal{A}(\alpha, \beta, t, \rho)$, whence, by a straightforward derivation, we can reach a set of self-consistent equations for the order parameters; these equations are addressable numerically and their solution allows us to check where, in the hyperspace (α, β, t, ρ), $\bar{m}$ is non-vanishing (*i.e.* the network works). Hereafter we briefly report the main results, while the underlying technical details can be found in the SM (see Secs. IV-V).

[5] More intuitively, it is sufficient to observe that the Hamiltonian of the LaD model for $M \gg 1$ depends on the dataset characteristics only through ρ.

Remark 2. By relying upon Gaussian integration,[6] the partition function $\mathcal{Z}_{N,M,K}(\beta|\boldsymbol{\eta},t)$ defined in (12) can be recast into an integral representation as

$$\mathcal{Z}_{N,M,K}(\beta|\boldsymbol{\eta},t)=\sum_{\{\sigma\}}\int\prod_{\mu=1}^{K}\left(\frac{dz_\mu}{\sqrt{2\pi}}\right)\int\prod_{i=1}^{N}\left(\frac{d\phi_i}{\sqrt{2\pi}}\right)\cdot$$

$$\cdot\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\phi_i^2-\frac{1}{2}\frac{1}{1+t}\sum_{\mu=1}^{K}z_\mu^2+\sqrt{\frac{\beta}{\Gamma N}}\frac{1}{M}\sum_{\mu,a,i=1}^{K,M,N}\xi_i^\mu\chi_i^{\mu,a}z_\mu k_i\right),\tag{22}$$

being k_i the multi-spin defined as

$$k_i:=\sigma_i+i\sqrt{\frac{t}{\beta(1+t)}}\phi_i.\tag{23}$$

In order to report the main Theorem, we have to introduce three further order parameters, a quantifier of the retrieval of the examples that is $m_\eta:=\frac{1}{r(1+\rho)}\frac{1}{NM}\sum_{a,i=1}^{M,N}\xi_i^1\chi_i^{1,a}\sigma_i=\frac{1}{r(1+\rho)}\hat{m}_1(\boldsymbol{\sigma})$ (proportional to the mean overlap between the neural configuration and the examples related to ξ^1) and two order parameters typical of the statistical mechanics of disordered systems (Coolen et al., 2005; Mézard, Parisi, & Virasoro, 1985), namely the overlap $q_{12}:=\frac{1}{N}\sum_{i=1}^{N}k_i^{(1)}k_i^{(2)}$ (that quantifies the glassiness of the landscape of the binary neurons) and the overlap $p_{12}:=\frac{1}{K}\sum_{\mu>1}^{K}z_\mu^{(1)}z_\mu^{(2)}$ (that accounts for the glassiness of the Gaussian variables z_μ appearing in the integral representation of the partition function, see Remark 2).

Given this preamble we can finally state the next

Theorem 1. *In the infinite volume limit $N\to\infty$ and large but finite dataset scenario ($M\gg1$), the free energy of the LaD model $\mathcal{A}(\alpha,\beta,t,\rho)$ is given by the following expression (to be extremized w.r.t. the order parameters):*

$$\mathcal{A}(\alpha,\beta,t,\rho)=\frac{\alpha}{2}\log(1+t)-\frac{\beta}{2\bar{D}}(1+t)\frac{\bar{m}_\eta^2}{(\bar{D}+t)}(1+\rho)+$$

$$+\frac{\alpha}{2}\left\{\frac{\beta\bar{q}(1+t)}{1-(1+t)\beta(\bar{Q}-\bar{q})}-\log\left[1-(1+t)\beta(\bar{Q}-\bar{q})\right]\right\}$$

$$+\mathbb{E}_\psi\log\cosh\left[\frac{1}{\bar{D}}\sqrt{\alpha\beta\bar{p}+\left(\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right)^2}\rho\psi+\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right]+$$

$$-\frac{\alpha\bar{p}}{2\bar{D}}\frac{t}{(1+t)}-\frac{1}{2}\beta(\bar{D}-1)\frac{1+t}{t}\bar{Q}+$$

$$+\frac{1}{2}\bar{p}\beta\alpha(\bar{q}-\bar{Q})-\frac{\beta}{2}\frac{(1-\bar{D})}{\bar{D}}\frac{1+t}{t}+\log\frac{2}{\sqrt{\bar{D}}}.$$

where the operator $\mathbb{E}_\psi$ is given by

$$\mathbb{E}_\psi g(\psi)=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp\left(-\frac{\psi^2}{2}\right)g(\psi),\tag{24}$$

and we posed $\bar{D}:=1+\alpha\frac{t}{(1+t)}(\bar{P}-\bar{p})$, with $\bar{m}_\eta,\bar{q},\bar{p},\bar{Q},\bar{P}$ being, respectively, the expectations of the example Mattis magnetization m_η, of the overlap q_{12}, of the overlap p_{12} and of their diagonal versions q_{11} and p_{11}.

Proof. The proof is based on the application of Guerra's interpolation scheme, as detailed in Section 4 of the SM. □

Corollary 1. *The expectation values of the archetype and example Mattis-magnetizations in the thermodynamic limit ($N\to\infty$) and large dataset scenario ($M\gg1$) fulfill the following self-consistent equations*

$$\bar{m}_\eta=\mathbb{E}_\psi\tanh\left[\frac{1}{\bar{D}}\sqrt{\alpha\beta\bar{p}+\left(\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right)^2}\rho\psi+\right.\tag{25}$$

$$\left.+\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right],$$

$$\bar{m}_\eta(1+\rho)=\bar{m}+(1-\hat{q})\frac{1+t}{\bar{D}+t}\beta\bar{m}_\eta\frac{\rho}{\bar{D}},\tag{26}$$

the other order parameters $\bar{D}$, $\bar{Q}$, $\bar{q}$, $\bar{p}$, $\hat{q}$ fulfill the following set of coupled equations

$$\bar{D}=1+\frac{\alpha t}{1-(1+t)\beta(\bar{Q}-\bar{q})},$$

$$\bar{p}=\frac{(1+t)^2\beta\bar{q}}{[1-(1+t)\beta(\bar{Q}-\bar{q})]^2},$$

$$\bar{D}^2\bar{Q}=1-\frac{1}{\beta}\frac{t\bar{D}}{1+t}-t\bar{m}_\eta^2(1+\rho)\frac{2\bar{D}+t}{(\bar{D}+t)^2}+$$

$$+\frac{\alpha\bar{p}}{\beta}\frac{t^2}{(1+t)^2}-(1-\hat{q})\frac{2t}{\bar{D}(1+t)}\alpha\bar{p},$$

$$\bar{D}^2(\bar{Q}-\bar{q})=1-\hat{q}-\frac{1}{\beta}\frac{t\bar{D}}{(1+t)},$$

$$\hat{q}:=\mathbb{E}_\psi\tanh^2\left[\frac{1}{\bar{D}}\sqrt{\alpha\beta\bar{p}+\left(\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right)^2}\rho\psi+\right.$$

$$\left.+\frac{\beta\bar{m}_\eta}{\bar{D}}\frac{1+t}{\bar{D}+t}\right].$$

Proof. The proof works by extremizing $\mathcal{A}(\alpha,\beta,t,\rho)$ w.r.t. the order parameters, namely by imposing $\nabla\mathcal{A}(\alpha,\beta,t,\rho)|_{\bar{m}_\eta,\bar{p},\bar{q},\bar{Q},\bar{D}}=0$ and by direct evaluation of the derivatives as detailed in Section 5 of the SM. □

By solving numerically the equations provided in Corollary 1, we can construct the phase diagrams for the LaD neural network and the main interest lies in depicting the retrieval region within these diagrams (as already provided, see Figs. 7 and 8): therein, the system – initialized in configurations corresponding to any example $\boldsymbol{\eta}^1$ or nearby configurations – spontaneously relaxes to configurations equal to or close to ξ^1; the initialization corresponds to the system input and the final, equilibrium state corresponds to the system reconstruction. Otherwise stated, this region corresponds to control parameters that yield solutions for the self-consistent equations such that $\bar{m}\sim1$. Notably, the knowledge of the phase diagram for machine retrieval allows us to inspect also the thresholds $M_c(r)$ for machine learning as – by setting the control parameters within that region – we known that, if the network has been previously supplied with enough examples $M_c(r)$ (such that learning has been properly accomplished), then, once inputted with a partial or corrupted information, it will be able to reconstruct the complete, exact information. We already get acquainted with these thresholds for learning (and, thus, for retrieval) as we have shown them in Fig. 2 in the first part of the manuscript.

5. Conclusions

We close this manuscript with a couple of remarks that stem from the reported research, then outlooks follow.

On the benefits of "sleeping" in machine learning and machine retrieval: Implementing (suitably stylized) sleeping mechanisms within artificial neural networks allows for sensible dataset-size reduction still preserving a successful learning. In particular, in the random scenario (where calculations are tractable) we proved that we can save up to 90% of the training set, retaining the same performance of "restless" networks. These results were confirmed by computational investigations also for several structured datasets: this suggests that these bio-inspired sleeping mechanisms can be pivotal for a Sustainable AI and mandatory when solely small datasets are available for training.

As a side note, this contributes to explain a long-standing question about the need of a large number of examples by machine learning algorithms before generalization can take place, in contrast with biological neural-networks which require by far fewer experiences (Ghirlanda & Enquist, 2003; Ross & Kennedy, 1990; Wang, Yao, Kwok, & Ni, 2020; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018): we sleep a $\mathcal{O}(1)$ time of our life.

[6] To be sharp, it is enough to apply the Hubbard–Stratonovich transformation on the partition function defined in Eq. (12)

On the equivalence between Cost and Loss functions: At least in these simple shallow networks, it is now clear that their Cost functions (the starting point to handle them analytically via statistical mechanics) are deeply related to the corresponding Loss functions (the starting point to handle them computationally in Machine Learning). In particular, beyond possible wider and deeper interpretations, the two terms actually coincide in the present scenario and, when optimizing these functions w.r.t. the weights, the machine accomplishes learning while, when optimizing w.r.t. the neurons, it accomplishes pattern reconstruction. By this perspective, not only Cost and Loss functions look as two faces of the same medal, but also learning and retrieval appear as two particular aspects of the broader phenomenon of cognition.

As for outlooks, we stress that, at present, there are two major (and related) limitations of the theory developed here, namely, i. it works with structureless patterns and such that in a retrieval state the mean fraction of spiking neurons is forced to one half, and ii. the whole statistical–mechanical framework relies on the assumption of replica symmetry. As for the former, the existence of other kinds of Hebbian algorithms that work particularly well for correlated stimuli, such as standard covariance learning rule and its variants (Minai, 1997; Stanton & Sejnowski, 1989) highlights a path to generalize the present theory to account for more complex datasets. However, as discussed by Amit, Gutfreund and Sompolinsky (Amit, Gutfreund, & Sompolinsky, 1987) for the standard Hopfield model, while simple correlations can be easily addressable (and this would allow dropping the constraint on the fraction of spiking neurons), for general structured scenarios there is still a long way to go within the statistical–mechanical approach (Mézard, 2023). As for the latter, in its current form, the dreaming algorithm affects all memories equally as a consequence of the replica-symmetry assumption we use to make analytical progresses, while solely the most-recently acquired ones are expected to undergo consolidation (the formation of long term memory should be more realistic, see *e.g.* Atkinson and Shiffrin (1968), Shiffrin and Atkinson (1969), Squire and Alvarez (1995)); several research groups are working on a broken-replica statistical–mechanical theory for neural networks (see *e.g.*, Albanese et al. (2022), Baldassi et al. (2020, 2021)) and, once it will be ready, it should be possible to improve these biologically inspired mechanisms.

CRediT authorship contribution statement

Elena Agliari: Conceptualization, Formal analysis, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. **Francesco Alemanno:** Data curation, Formal analysis, Investigation, Methodology. **Miriam Aquaro:** Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Adriano Barra:** Conceptualization, Funding acquisition, Project administration, Writing – original draft, Writing – review & editing. **Fabrizio Durante:** Conceptualization, Validation, Writing – review & editing. **Ido Kanter:** Conceptualization, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Adriano Barra reports financial support was provided by Government of Italy Ministry of Foreign Affairs and International Cooperation. Ido Kanter reports financial support was provided by Israel Ministry of Innovation Science & Technology. Adriano Barra, Fabrizio Durante reports financial support was provided by Government of Italy Ministry of Education University and Research.

Data availability

We used standard public repositories as datasets (*e.g.* MNist, Fashion-Mnist, Olivetti).

Acknowledgments

The Authors acknowledge financial support by Ministero per gli Affari Esteri e la Collaborazione Internazionale (MAECI, Italy) and to the Ministry of Science, Technology and Space (MOST, Israel) for the shared grant for the Scientific and Technical Collaboration between Israel and Italy, *BULBUL, Brain Inspired Ultra-fast and Ultra-sharp machines for health-care*, Project n. F85F21006230001. EA and MA are in Sapienza University of Rome that does not belong to the BULBUL project and they are grateful to Sapienza University of Rome (RM120172B8066CB0, RM12117A8590B3FA) for financial support. EA acknowledges financial support from PNRR MUR project PE0000013-FAIR. FD and FA are in Salento but do not belong to the BULBUL project and they are grateful to MUR via the PRIN grants *Stochastic Methods for Complex Systems* n. 2017JFFHS and *Statistical Mechanics of Learning Machines: from algorithmic and information-theoretical limits to new biologically inspired paradigms* n. 20229T9EAT.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2024.106174.

References

Agliari, E., Acquaro, M., Alemanno, F., & Fachechi, A. (2023). Regularization, early-stopping and dreaming: a Hopfield-like setup to address generalization and overfitting. arXiv preprint arXiv:2308.01421.

Agliari, E., Alemanno, F., Barra, A., & De Marzo, G. (2002). The emergence of a concept in shallow neural networks. *Neural Networks*, *148*, 232–254.

Agliari, E., Alemanno, F., Barra, A., & Fachechi, A. (2019). Dreaming neural networks: rigorous results. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(8), Article 083503.

Agliari, E., Barra, A., Sollich, P., & Zdeborova, L. (2020). Machine learning and statistical physics: theory, inspiration, application. *Journal of Physics A: Special Issue*.

Albanese, L., et al. (2022). Replica symmetry breaking in dense hebbian neural networks. *Journal of Statistical Physics*, *189*(2), 24.

Alemanno, F., et al. (2023). Supervised hebbian learning. *Europhysics Letters*, *141*, 11001.

Amit, D. J. (1992). *Modeling brain function: the world of attractor neural networks*. Cambridge Univ. Press.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *55*, 1530–1534.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Information storage in neural networks with low levels of activity. *Physical Review A*, *35*, 2293.

Andrillon, T., Pressnitzer, D., Leger, D., & Kouider, S. (2018). Formation and suppression of acoustic memories during human sleep. *Nature Communication*, *8*, 179.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

Baldassi, C., et al. (2020). Clustering of solutions in the symmetric binary perceptron. *JSTAT*, *7*(2020), Article 073303.

Baldassi, C., et al. (2021). Unveiling the structure of wide flat minima in neural networks. *Physical Review Letters*, *127*(27), Article 278301.

Carleo, G., et al. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), Article 045002.

Coolen, A. C. C., Kühn, R., & Sollich, P. (2005). *Theory of neural information processing systems*. Oxford Univ. Press.

Crick, F., & Mitchinson, G. (1983). The function of dream sleep. *Nature*, *304*, 111.

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews. Neuroscience*, *11*, 114.

Engel, A., & Van den Broeck, C. (2001). *Statistical mechanics of learning*. Cambridge Univ. Press.

Fachechi, A., Agliari, E., & Barra, A. (2019). Dreaming neural networks: forgetting spurious memories and reinforcing pure ones. *Neural Networks*, *112*, 24–40.

Fachechi, A., Barra, A., Agliari, E., & Alemanno, F. (2022). Outperforming RBM feature-extraction capabilities by dreaming mechanism. *IEEE Transactions on Neural Networks and Learning Systems*.

Fontanari, J. F. (1990). Generalization in a Hopfield network. *Journal de Physique*, *51*(21), 2421–2430.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behavior*, *66*, 15–36.

Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. *Mitsui Technical Review*.

Horé, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *20th international conference on pattern recognition*. IEEE.

Kanter, I., & Sompolinsky, H. (1987). Associative recall of memory without errors. *Physical Review A*, *35*, 380.

Kermiche, N. (2020). Contrastive Hebbian feedforward learning for neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *31*, 2118.

Kirkpatrick, S., Gelatt, C. D., & Vecchia, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–673.

Kobayashi, M. (2013). Hyperbolic Hopfield neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *24*, 335.

Kohonen, T. O. (1984). *Self-organization and associative memory*. Berlin: Springer.

Maquet, P. (2001). The role of sleep in learning and memory. *Science*, *294*, 1048.

Marino, R., Parisi, G., & Ricci-Tersenghi, F. (2016). The backtracking survey propagation algorithm for solving random K-SAT problems. *Nature Communication*, *7*, 1–8.

McGaugh, J. L. (2000). Memory - a century of consolidation. *Science*, *287*, 248–251.

Mézard, M. (2023). Spin glass theory and its new challenge: structured disorder. *Indian Journal of Physics*, 1–12.

Mézard, M., Parisi, G., & Virasoro, M. A. (1985). *Spin glass theory and beyond*. World Sci. Comp. Press.

Minai, A. A. (1997). Covariance learning of correlated patterns in competitive networks. *Neural Computation*, *9*, 667–681.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on System, Man, and Cybernetics*, *9*, 62–66.

Paton, J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*, 865.

Personnaz, L., Guyon, I., & Dreyfus, G. (1985). Information storage and retrieval in spin-glass like neural networks. *Journal de Physique Lettres*, *46*(359), 365.

Phillips, P., & O'Toole, A. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32.

Pu, Y.-F., Yi, Z., & Zhou, J.-L. (2017). Fractional Hopfield neural networks: Fractional dynamic associative recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *28*, 2319.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*, 42.

Seung, H. S., Sompolinsky, H., & Tishby, N. (1992). Statistical mechanics of learning from examples. *Physical Review A*, *45*, 6056–6078.

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review*, *76*(2), 179.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, *5*(2), 169–177.

Stanton, P. K., & Sejnowski, T. J. (1989). Associative long-term depression in the hippocampus induced by hebbian covariance. *Nature*, *339*, 215–218.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv:1906:02243.

Tanaka, G., Nakane, R., Takeuchi, T., Yamane, T., Nakano, D., Katayama, Y., et al. (2020). Spatially arranged sparse recurrent neural networks for energy efficient associative memory. *IEEE Transactions on Neural Networks and Learning Systems*, *31*, 2162.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: a survey on few-shot learning. *ACM Computing Surveys*, *53*, 1–34.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behavior*, *2*(12), 915–924.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.