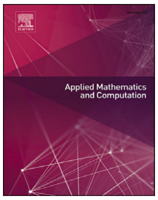
Applied Mathematics and Computation 474 (2024) 128689

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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



Full Length Article

A spectral approach to Hebbian-like neural networks

Elena Agliari [a,b,*], Alberto Fachechi [a,b], Domenico Luongo [a]

[a] *Dipartimento di Matematica "Guido Castelnuovo", Sapienza Università di Roma, Roma, Italy*

[b] *GNFM-INdAM, Gruppo Nazionale di Fisica Matematica, Istituto Nazionale di Alta Matematica, Italy*

ABSTRACT

We consider the Hopfield neural network as a model of associative memory and we define its neuronal interaction matrix $\boldsymbol{J}$ as a function of a set of $K \times M$ binary vectors $\{\xi^{\mu,A}\}_{\mu=1,\dots,K}^{A=1,\dots,M}$ representing a sample of the reality that we want to retrieve. In particular, any item $\xi^{\mu,A}$ is meant as a corrupted version of an unknown ground pattern ζ^{μ}, that is the target of our retrieval process. We consider and compare two definitions for $\boldsymbol{J}$, referred to as supervised and unsupervised, according to whether the class μ, each example belongs to, is unveiled or not, also, these definitions recover the paradigmatic Hebb's rule under suitable limits. The spectral properties of the resulting matrices are studied and used to inspect the retrieval capabilities of the related models as a function of their control parameters.

Contents

1. Introduction

Since its introduction, in the eighties, the Hopfield neural network has attracted a big deal of attention from a broad community of scientists at the interface of physics, mathematics and computer science [1,2]. In fact, the Hopfield network is recognized as a paradigmatic model for associative memory: if properly designed, it can store and retrieve a set of K information patterns $\xi = \{\xi^{\mu}\}_{\mu=1}^{K}$, with $\xi^{\mu} \in \{-1,+1\}^{N}$. More precisely, the model consists of a set of N binary neurons, whose configuration is denoted as

[*] Corresponding author.
E-mail address: agliari@mat.uniroma1.it (E. Agliari).

https://doi.org/10.1016/j.amc.2024.128689
Received 29 January 2024; Received in revised form 15 March 2024; Accepted 17 March 2024

Available online 27 March 2024

$\boldsymbol{\sigma} \in \{-1,+1\}^N$, interacting pairwise and symmetrically by an interaction strength encoded by the coupling matrix $\boldsymbol{J} \in \mathbb{R}^{N\times N}$, and evolving in time in such a way that any neuron σ_i gets progressively aligned with the local field $(\boldsymbol{J}\boldsymbol{\sigma}^T)_i$ acting on it and stemming from the neighboring neurons. The key point for the functioning of the model as an associative memory is therefore to design $\boldsymbol{J}$ in such a way that stored patterns are associated to attractors in the configuration space. The standard choice is inspired by Hebb's principle [3] and reads as $\boldsymbol{J} = \boldsymbol{\xi}^T\boldsymbol{\xi}/N$, which, in the case of Rademacher patterns and in the large-size limit, ensures a storage capacity of approximately $0.14N$ patterns, see e.g. [4].

In this context, determining the structure of the attraction basins is a paramount goal in order to understand the information processing principles lying behind the associative memory functionalities and possibly to highlight qualitatively-different working regimes of the model corresponding to different parameter settings. Such analysis can be naturally framed by means of the statistical mechanics of spin glasses, as pioneered by Amit, Gutfreund and Sompolinsky [5,6]. On the mathematical side, several results have been derived for the Hopfield model, by exploiting different techniques – ranging from large deviation analysis [7,8] to Guerra's interpolation [9–11] – and leading to bounds on the storage capacity [12–19]. The analytical investigations underlying these results have significantly benefited from the simple expression of the Hebbian interaction matrix. On the other hand, the cost for this simplicity is a limited capacity of the network: in the limit of large size N, a symmetric neural network can store up to N patterns [20], that is much higher than the aforementioned $0.14N$. In fact, as the number of stored patterns is relatively large, the related attraction basins tend to overlap, giving rise to frustration and, consequently, to a plethora of spurious attractors, whose retrieval is interpreted as an error of the network. Thus, several algorithms have been developed to optimize the model coupling matrix, enhancing the attractive power of stored patterns and increasing the critical storage capacity [21–29]; in general, the core idea of these algorithms is to modify the structure of the Hebbian matrix $\boldsymbol{J}$, in order to disentangle the attraction basins of the patterns and then downsizing the harmful effect associated to the presence of spurious attractors. Among these algorithms, we recall the so-called *dreaming* kernel [30–32], which shall constitute the starting point for our work.

Beyond these variations on Hebb's theme, more recently, much attention has been devoted to a scenario in which the information supplied and used to build the interaction matrix does not correspond to *ground-truth* patterns, but rather to *examples* of theirs, namely to samples of the reality that we want to retrieve [33–35]. This modified setting allows us to develop models in which the attractiveness of the ground patterns, that are not directly accessible and therefore are not stored in the coupling matrix, emerges as a consequence of the coalescence of attraction basins associated to examples related to the same ground pattern. This kind of phenomenon is responsible for the generalization capabilities of the model. The ways examples can be combined into the coupling matrix mimic the two training protocols: *i.* in the *supervised* setting, we know *a priori* the organization of the examples in, say, P classes, so that – assuming noise in the examples is uncorrelated – the empirical mean, say $\bar{\boldsymbol{\xi}}^\mu$ for the μ-th class, is a good approximation of the reality; in this way, we can promote this mean as a representative of the given μ-th class, and store this, namely use the set $\{\bar{\boldsymbol{\xi}}^\mu\}_{\mu=1,\dots,P}$ to build $\boldsymbol{J}$; *ii.* in the *unsupervised* setting, in which there is no *a priori* distinction between examples belonging to different classes, the only possible way to store information in $\boldsymbol{J}$ is to treat all of them as distinct information patterns.

As a matter of fact, crucial properties of these Hopfield-like models are entirely encoded in the structure of the (random) coupling matrix, and, in particular, in its spectral properties, see for example [36–38] for recent investigations. Moreover, the strong relation between random-matrix theory tools and spin-glass models constitutes a long-standing research topic, see for instance [39–42], and also [43–53] for applications in machine learning and information theory. In this paper, we consider a set of Rademacher ground patterns and we obtain a sample of examples by randomly flipping a certain fraction of their entries. With these samples, we build the dreaming kernel, distinguishing between a supervised and an unsupervised version and for both we derive the exact eigenvalue distribution in the limit of a large size N. By relying on such a knowledge, we inspect the generalization capabilities of the model, as a function of its parameters.

The path that we pursue is the following: first, we present the model and the related definitions (Sec. 2), next, we state the main analytical results (Sec. 3), and we apply them to investigate the information-processing capabilities of the model (Sec. 4); finally, we summarize and discuss our findings (Sec. 5). The proofs and the technical details are collected in the Appendices.

2. The framework: models, methods and quantities

Given a set of patterns $\boldsymbol{\xi}^\mu \in \{-1,+1\}^N$, with $\mu = 1,\dots,K$, the reference coupling matrix for the following analysis is given by

$$J^{\xi}_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu=1}^{K} \xi^\mu_i \left(\frac{1+t}{\boldsymbol{1}+t\boldsymbol{C}}\right)_{\mu\nu} \xi^\nu_j, \tag{2.1}$$

where $t \in \mathbb{R}^+$ and $\boldsymbol{C}$ is the pattern correlation matrix (*vide infra*). The matrix $\boldsymbol{J}^\xi$ was introduced in [30] and can be derived from Hebb's one by implementing consolidation and remotion mechanisms inspired by those occurring in mammal's brain during sleep. Thus, the resulting model is referred to as "dreaming Hopfield model" and t, which tunes the extent of such mechanisms, as "dreaming time", see also the recent related works [36,54–57]. Notably, the matrix $\boldsymbol{J}^\xi$ includes paradigmatic cases: by setting $t = 0$ we recover the Hebbian coupling and, in the limit $t \to \infty$, we recover Kohonen's projection matrix [58]; the latter is known to reach the storage-capacity upper-bound, that is, a number $K = N$ of patterns can be successfully stored and retrieved. Moreover, the coupling matrix (2.1) turns out to emerge as the solution of the minimization of a L_2-regularized loss-function where a cost is shaped whenever the configuration corresponding to one of the stored patterns is not stable and where the regularization parameter is mapped into the dreaming time [37]. In fact, the dreaming time controls the overlap between different attraction basins: the higher t, the lower the attractive power of spurious configurations [30–32].

Before proceeding, it is worth introducing the following notation $x \sim \mathrm{Rad}(p)$, with $p \in [-1,+1]$, that, in the following, shall define a binary random variable x, drawn from the distribution $\mathcal{P}(x) = \frac{1-p}{2}\delta_{x,-1} + \frac{1+p}{2}\delta_{x,+1}$, in such a way that, when $p = 0$, x is a standard Rademacher variable, while, when $p \neq 0$, x is a biased binary random variable with expectation p. We are now ready to describe the three settings that we are inspecting in the next sections:

a) In the basic *storing* setting, we have $K = P$ patterns[1] $\{\xi^\mu\}_{\mu=1}^P$, each made of N Rademacher entries: $\xi_i^\mu \sim \mathrm{Rad}(0)$ for any $\mu = 1,\dots,P$ and $i = 1,\dots,N$. The correlation matrix in (2.1) reads as $C_{\mu\nu} = \frac{1}{N}\sum_i \xi_i^\mu \xi_i^\nu$.

b) In the *supervised* setting, we have P ground patterns $\{\zeta^\mu\}_{\mu=1}^P$, each made of N entries and, from these, we generate $K = P \times M$ examples, denoted as $\{\xi^{\mu,A}\}_{\mu=1,\dots,P}^{A=1,\dots,M}$, by randomly flipping the entries of the related ground-pattern. Specifically, we choose Rademacher ground-patterns, that is $\zeta_i^\mu \sim \mathrm{Rad}(0)$, and uncorrelated noise for examples, that is,

$$\xi_i^{\mu,A} = \chi_i^{\mu,A}\zeta_i^\mu, \quad \text{with } \chi_i^{\mu,A} \sim \mathrm{Rad}(r), \quad r \in [0,1], \tag{2.2}$$

for any $\mu = 1,\dots,P$, $i = 1,\dots,N$, and $A = 1,\dots,M$. In this supervised setting, since the class of each example is unveiled, we can calculate the empirical mean of examples in each class as

$$\bar{\xi}_i^\mu := \frac{1}{M}\sum_{A=1}^M \xi_i^{\mu,A} = \frac{1}{M}\sum_{A=1}^M \chi_i^{\mu,A}\zeta_i^\mu =: \bar{\chi}_i^\mu \zeta_i^\mu.$$

Then, the coupling matrix is defined as

$$J_{ij}^s(t) = \frac{1}{N}\sum_{i,j=1}^N \sum_{\mu,\nu=1}^P \zeta_i^\mu \bar{\chi}_i^\mu \left(\frac{1+t}{\mathbf{1}+t\mathbf{C}^s}\right)_{\mu\nu} \bar{\chi}_j^\nu \zeta_j^\nu,$$

where

$$C_{\mu\nu}^s = \frac{1}{N}\sum_{i=1}^N \zeta_i^\mu \bar{\chi}_i^\mu \bar{\chi}_i^\nu \zeta_i^\nu$$

is the correlation matrix of the empirical means of the examples.

c) In the *unsupervised* setting, the examples $\{\xi^{\mu,A}\}_{\mu=1,\dots,P}^{A=1,\dots,M}$ are generated precisely as in the previous setting b), but, in this case, there is no preassigned label distinguishing between classes. As a consequence, we store all the examples as information patterns, i.e. in (2.1) we replace ξ_i^μ with $\xi_i^{\mu,A}$ and the sum over μ is replaced with the sum over (μ, A), namely the sum is performed over all the $P \times M$ examples. The coupling matrix is

$$J_{ij}^u(t) = \frac{1}{NM}\sum_{\mu,\nu=1}^P \sum_{A,B=1}^M \xi_i^{\mu,A}\left(\frac{1+t}{\mathbf{1}+t\mathbf{C}^u}\right)_{(\mu A),(\nu B)} \xi_i^{\nu,B},$$

where

$$C_{(\mu A),(\nu B)}^u = \frac{1}{NM}\sum_{i=1}^N \xi_i^{\mu,A}\xi_i^{\nu,B},$$

is the dataset correlation matrix.

Remark 1. The parameter r, underlying the statistics of the random variable χ appearing in Eq. (2.2), is related to the fraction of pixels that are expected to be flipped in any example, say $\xi^{\mu,A}$, with respect to the ground ζ^μ. In particular, when $r = 0$, each example is, in the average over the entry-flipping probability, orthogonal to the related archetype, while, when $r = 1$, each example is a perfect copy of the related archetype. Thus, r and M can be interpreted as, respectively, a measure of the *quality* and of the *quantity* of the available dataset.

3. Algebraic properties of the coupling matrices

The retrieval capabilities of the models described in the previous section can be addressed by relying on the eigenvalue distributions of the related coupling matrices.[2] Also, by comparing their spectra we can assess to what extent the models encoded by $\boldsymbol{J}^s$ and

[1] In the basic storing setting, we use P to denote the number of orthogonal patterns for homogeneity with the other scenarios, where P is the number of classes; in any case, in the thermodynamic limit, we pose $\alpha = P/N$.

[2] A derivation of these coupling matrices from statistical inference can be found in [59] for the standard Hebbian model and in [60] for the dreaming Hopfield model.

$\boldsymbol{J}^u$ differ from the model built on ground patterns and therefore the effectiveness of their definitions. This motivates the aim of this section, that is, determining the spectral properties for the random matrices under consideration.

Definition 1 *(Thermodynamic limit)*. The thermodynamic limit (TDL) is defined as $N, P \to \infty$ with $P = P(N)$ and $\lim_{N\to\infty} P/N = \alpha$, with $0 < \alpha \leq 1$. When dealing with ground patterns (i.e., setting *a*), this coincides with the so-called high-storage regime of the Hopfield model.

In the following, unless it is explicitly specified, we will denote the coupling matrix as $\boldsymbol{J}(t)$, regardless of the setting under consideration. In fact, by denoting with $\boldsymbol{X}$ the matrix made of the information vectors (ground patterns or examples) on the rows – in the random pattern and supervised cases, it is a $P \times N$ matrix with entries resp. $X_{\mu i} = \xi_i^\mu$ and $X_{\mu i} = \frac{1}{M}\sum_A \chi_i^{\mu,A}\zeta_i^\mu$, while in the unsupervised case it is a $MP \times N$ matrix with entries $X_{(\mu,A),i} = \chi_i^{\mu,A}\zeta_i^\mu$ where the double index (μ, A) labels each example in the dataset – any of the coupling matrices introduced above can be written as

$$\boldsymbol{J}(t) = \frac{1}{D_N}\boldsymbol{X}^T\left(\frac{1+t}{\mathbf{1}+t\boldsymbol{C}}\right)\boldsymbol{X}, \tag{3.1}$$

with $\boldsymbol{C} = \frac{1}{D_N}\boldsymbol{X}\boldsymbol{X}^T$ and D_N is a normalization factor that reads as $D_N = N$ for the basic storing and the supervised case, or $D_N = NM$ for the unsupervised case.

Lemma 1. *The following results hold:*

1. *The coupling matrix $\boldsymbol{J}(t)$ satisfies the differential equation*

$$\dot{\boldsymbol{J}}(t) = \frac{1}{1+t}[\boldsymbol{J}(t) - \boldsymbol{J}(t)^2]. \tag{3.2}$$

2. *Given the eigenvalues λ_α^0 of the coupling matrix $\boldsymbol{J}^0 := \boldsymbol{J}(0)$, then the eigenvalues $\lambda_\alpha(t)$ of $\boldsymbol{J}(t)$ are in bijective correspondence with λ_α^0 through the relation*

$$\lambda_\alpha(t) = \frac{1+t}{1+t\lambda_\alpha^0}\lambda_\alpha^0. \tag{3.3}$$

3. *The eigenspaces of $\boldsymbol{J}^0$ are stable under dreaming flow, that is, if $\{v_\alpha^1, \dots, v_\alpha^m\}$ are the eigenvectors of $\boldsymbol{J}^0$ associated to an m-degenerate eigenvalue λ_α^0, then $Span(\{v_\alpha^1, \dots, v_\alpha^m\})$ is the eigenspace of $\boldsymbol{J}(t)$ associated to the eigenvalue $\lambda_\alpha(t)$.*

The proof of this lemma is detailed in App. A.

Remark 2. Lemma 1 states that the dreaming interaction matrix $\boldsymbol{J}(t)$ defined in (2.1) results from the evolution (3.2), regardless of the underlying setting. In other words, whether $\boldsymbol{J}(t)$ stems from a basic storing or by the combination of corrupted examples (either labeled or not), that is, whether $\boldsymbol{J}(t)$ is meant for storing or for generalization, it still results from the process represented by (3.2) which encodes for a consolidation (positive term in the square brackets in Eq. (3.2)) and a remotion (negative term) mechanism.

Having established these basic properties of the coupling matrices in all the three settings under consideration at finite N, P and M, we are now able to study their relevant spectral properties in the thermodynamic limit. In that limit, for the supervised and unsupervised settings, we also pose $M \to \infty$, regardless of N; this condition is also referred to as the big-data regime. The main results of the section are summarized in the following

Theorem 1. *In the thermodynamic limit $N \to \infty$ and, for the supervised and unsupervised settings, in the infinite sample-size limit $M \to \infty$, the following results hold:*

1. *The empirical spectral distribution $\mu_N^0 = \frac{1}{N}\sum_\alpha \delta_{\lambda_\alpha^0}$ of $\boldsymbol{J}^0$ (with δ_λ being the Dirac delta measure at λ) converges in weak topology $\mu_N^0 \to \mu^0$, where*

$$d\mu^0(\lambda) = (1-\alpha)\delta(\lambda - \hat{\lambda}^0)d\lambda + \alpha d\mu_{MP}(\lambda), \tag{3.4}$$

with the measure $d\mu_{MP}(\lambda)$ being a shifted Marchenko-Pastur distribution $MP(\alpha, \sigma^2)$, i.e.

$$d\mu_{MP}(\lambda) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+^0 - \lambda)(\lambda - \lambda_-^0)}}{\alpha(\lambda - \hat{\lambda}^0)}d\lambda, \tag{3.5}$$

and $\lambda_\pm^0 = \sigma^2(1 \pm \sqrt{\alpha})^2 + \hat{\lambda}^0$. The parameters σ^2 and $\hat{\lambda}^0$ depend on the setting under consideration;

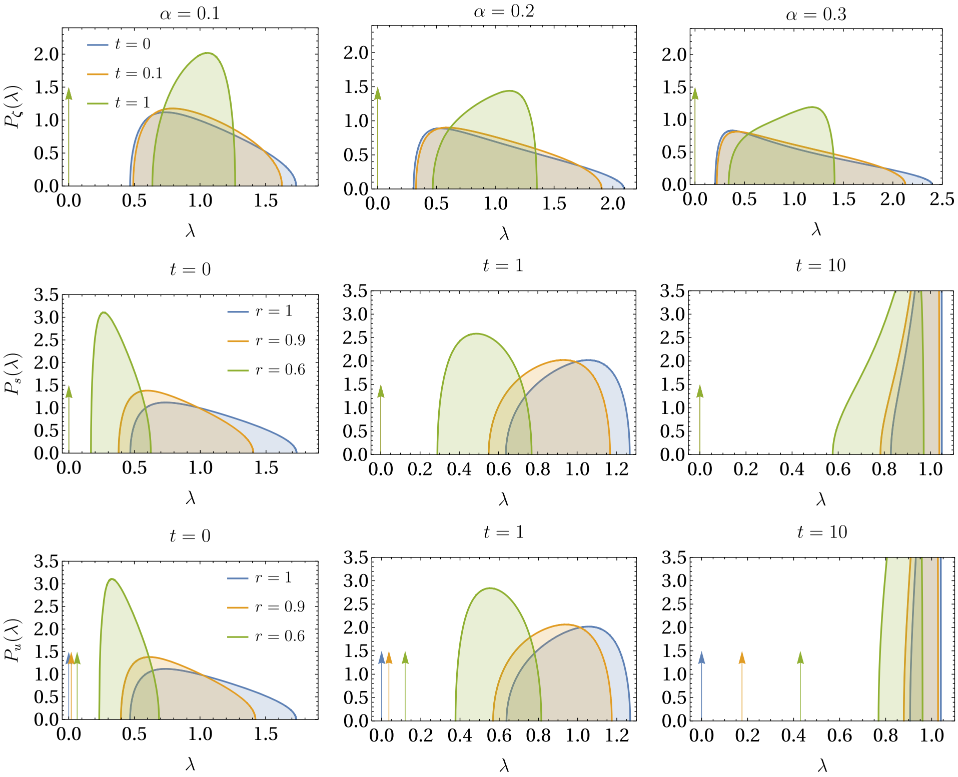


Fig. 1. Limiting spectral distributions of the couplings matrix. The figure shows the probability distribution $P(\lambda) = \frac{d\mu}{d\lambda}$ (3.7) in the three settings under consideration: basic storing (first row), supervised (second row) and unsupervised (third row) cases. In the first row, we plotted the spectral distribution for various values of α and t, while in the supervised and unsupervised setting we fixed $\alpha = 0.1$ and vary t and r. The vertical arrows (whose heights are arbitrary) refer to the location of the δ-peak: in the basic storing and supervised cases, the location is at $\lambda = 0$ (as $\hat{\lambda}^0 = 0$), while in the unsupervised case it depends on α, t and r, as predicted by Theorem 1.

2. *The empirical spectral distribution* $\mu_N^t = \frac{1}{N}\sum_\alpha \delta_{\lambda_\alpha(t)}$ *of the coupling matrix* $\boldsymbol{J}(t)$ *converges in weak topology* $\mu_N^t \to \mu^t$*, where*

$$d\mu^t(\lambda) = d\mu^0\left[\frac{\lambda}{1+t(1-\lambda)}\right], \tag{3.6}$$

i.e.

$$d\mu^t(\lambda) = (1-\alpha)\delta\left[\lambda - \frac{(1+t)\hat{\lambda}^0}{1+t\hat{\lambda}^0}\right]d\lambda + \alpha d\mu^t_{bulk}, \tag{3.7}$$

where the bulk distribution is

$$d\mu^t_{bulk}(\lambda) = \frac{1+t}{2\pi\sigma^2}\frac{\sqrt{(1+t\lambda^0_-)(1+t\lambda^0_+)}}{[1+t(1-\lambda)]^2}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\alpha[(1+t\hat{\lambda}^0)\lambda-(1+t)\hat{\lambda}^0]}d\lambda, \tag{3.8}$$

with $\lambda_\pm = \frac{(1+t)\lambda^0_\pm}{1+t\lambda^0_\pm}$*, and* σ^2 *and* $\hat{\lambda}^0$ *are the same parameters of the previous point depending on the setting under consideration.*

The proof of this theorem is provided in App. B.

Remark 3. Theorem 1 fully characterizes the spectral properties of the coupling matrix in all the three settings. Below, we discuss point by point its content, and refer to Fig. 1 for a visual representation of the spectral distributions for various values of the tunable parameters t, r and α.

- Point 1. of the theorem states that, at $t = 0$ (that is where the Hebbian structure $\boldsymbol{J}^0 \propto \boldsymbol{X}^T\boldsymbol{X}$ is recovered), the spectral distribution consists in a delta-peak located at $\hat{\lambda}^0$ and a shifted Marchenko-Pastur bulk with mass, respectively, $1-\alpha$ and α. The parameters of the distribution depend on the setting (see App. B); in particular, for ground patterns and supervised settings, the location of the delta peak is at $\hat{\lambda}^0 = 0$, while in the unsupervised setting $\hat{\lambda}^0 = \alpha(1-r^2)$. This is consequence of the fact that, unlike the other cases, in the unsupervised setting the coupling matrix is full-rank in the large dataset limit, thus all the eigenvalues are positive;

- When $t > 0$, point 2. of the theorem states that the spectral distribution is accordingly deformed under the dreaming flow (3.3). In this case, the shifted Marchenko-Pastur bulk will move and progressively concentrates around the limiting eigenvalue $\lambda = 1$. On the other side, the delta-peak will always be located in $\lambda = 0$ for the ground patterns and supervised settings, while in the unsupervised one, as t is increased, it will move as well towards $\lambda = 1$, so that at $t \to \infty$ the whole spectrum concentrates around the limiting eigenvalue, and the coupling matrix converges to the identity.

Before going further, we stress that the unsupervised setting is a rather peculiar scenario if compared to basic storing and supervised cases, as it is characterized by two different regimes. If $MP \geq N$, the coupling matrix is full-rank, and eigenvalues are all strictly positive (this is the regime in which Theorem 1 is derived, as we are interested in the large dataset limit). If $MP < N$, the coupling matrix is low-rank, with a fraction $1 - MP/N$ of vanishing eigenvalues; in this regime, the positive component of the spectrum exhibits a different distribution, with the K largest eigenvalues, well-separated from the continuous bulk at low t, and the bulk ultimately collapsing to $\lambda = 1$ in the $t \to \infty$ limit. In that case, the emerging generalization performance exhibits specific features and we refer to [37] for an extensive discussion.

When dealing with examples of unavailable, ground patterns, either in a supervised or unsupervised setting, it is natural to question whether our empirical models accounts for a good representation of the reality, namely whether $\boldsymbol{J}^s$ and $\boldsymbol{J}^u$ are close to $\boldsymbol{J}^\zeta$ where we directly store the ground-truths as information patterns. In order to assess the validity of our models, we consider the squared error between the empirical coupling matrices and the one built with the ground-truths, as the parameter α, r and t are tuned.

Definition 2. The Squared Error (SE) between empirical and ground-truth coupling matrices is defined as

$$\mathcal{E}_M^{s,u}(\alpha, r, t) = \frac{1}{N} \| \boldsymbol{J}^\zeta(t) - \boldsymbol{J}^{s,u}(t) \|_F^2 \tag{3.9}$$

where the superscripts s, u label the supervised or unsupervised setting, $\boldsymbol{J}^\zeta(t)$ is the coupling matrix built with the ground-truths, and $\|\cdot\|_F$ is the Frobenius norm between matrices. We denote $\mathcal{E}^{s,u}(\alpha, r, t) = \lim_{M\to\infty} \mathcal{E}_M^{s,u}(\alpha, r, t)$.

Proposition 1. *In the thermodynamic limit, and for $M \to \infty$, the SE can be expressed as*

$$\mathcal{E}^{s,u}(\alpha, r, t) = \int \left[\lambda - f_{r,t}^{s,u}(\lambda)\right]^2 d\mu_\zeta^t(\lambda), \tag{3.10}$$

where μ_ζ^t is the limiting spectral distribution of $\boldsymbol{J}^\zeta$, and

1. *in the supervised setting:*

$$f_{r,t}^s(\lambda) = \frac{\lambda r^2 (t+1)}{\lambda \left(r^2 - 1\right) t + t + 1};$$

2. *in the unsupervised setting:*

$$f_{r,t}^u(\lambda) = \frac{(t+1)\left\{\lambda r^2 + \alpha \left(r^2 - 1\right)[(\lambda - 1)t - 1]\right\}}{\lambda \left(r^2 - 1\right) t(\alpha t + 1) - \left[\alpha \left(r^2 - 1\right)(t+1)t\right] + t + 1}.$$

Again, we refer to the Appendices and, specifically, to App. C for the complete proof.

The exact SE $\mathcal{E}^{s,u}(\alpha, r, t)$, obtained by evaluating (3.10), is plotted versus r in Fig. 2 where several values of α and t are considered. Also, these theoretical results, valid in the limit $M \to \infty$, are compared with the numerical evaluation of $\mathcal{E}_M^{s,u}(\alpha, r, t)$, as reported in (3.9), at finite sample size M. In general, there is a strong agreement between the numerical results and the theoretical predictions, and the accuracy of theoretical results gets better by increasing the dataset size M. Also, as expected, the empirical versions of the coupling matrix do converge to the basic storing setting as r approaches 1. However, the interesting point is to analyze the role of the dreaming parameter t. In the supervised setting, increasing the dreaming time results in a fast convergence of the coupling matrix towards the basic storing setting (see e.g. the case $t = 10$ in the first row of Fig. 2). The reason lies in the fact that, as t gets larger, the coupling matrix approaches the projector model [23] and, at the leading order, the empirical means $\bar{\xi}^\mu \sim r\zeta^\mu$, while the empirical correlation matrix $\boldsymbol{C}^s = \frac{1}{N}\bar{\xi}\bar{\xi}^T \sim \frac{r^2}{N}\zeta\zeta^T$. Then, in the $t \to \infty$ limit, $\boldsymbol{J}^s \to \frac{1}{N}\bar{\xi}^T C_s^{-1}\bar{\xi} \sim \frac{1}{N}\zeta^T \boldsymbol{C}_\zeta^{-1}\zeta$ for any $r > 0$, while at $t = 0$ one has $\boldsymbol{J}^s \to r^2 \boldsymbol{J}^{0,\zeta}$. Therefore, increasing t, the supervised coupling matrix gets more and more insensitive to the quality of the examples, and $\boldsymbol{J}^s$ approaches the basic storing setting $\boldsymbol{J}^\zeta$ with the ground patterns.

In the unsupervised setting, the dreaming mechanism works in the same way but this time the patterns that are stored are the single examples rather than their empirical mean. Thus, as t gets larger, provided that the overall number of examples is not too large, each single example can be a fixed point. This way, increasing t too much would prevent the convergence of $\boldsymbol{J}^u$ to $\boldsymbol{J}^\zeta$, unless the examples are perfect realizations of the ground-truths, i.e. $r = 1$.

Another way to see the qualitative difference between supervised and unsupervised cases is by noticing the emergence of a different interplay between M and r. In a nutshell, and focusing on the Hebbian case for simplicity (the general case with $t > 0$ can

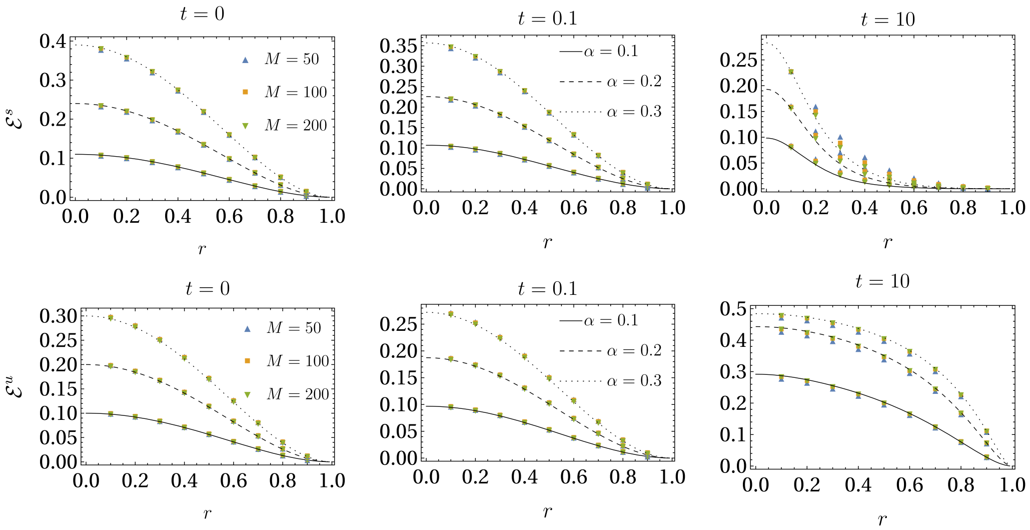


Fig. 2. Squared error for supervised and unsupervised settings. The figure shows the comparison between numerical results for the SE (3.9) at finite M and the theoretical prediction for $M \to \infty$ in the thermodynamic limit as a function of r for various values of α and t (Proposition 1). The first row refers to the supervised setting, while the second line shows the results for the unsupervised case. For fixed t, each plot exhibits the results for $\alpha = 0.1$ (solid line), $\alpha = 0.2$ (dashed line) and $\alpha = 0.3$ (dotted line), while the markers refer to $M = 50, 100, 200$. The network size is fixed to $N = 1000$ in all cases.

be recovered by leveraging the results of Lemma 1), in the supervised case, the variance of the random variable $\bar{\xi}^\mu$ over all possible realizations of ground-truths and related examples is $\mathbb{E}(\bar{\xi}^\mu_i)^2 = r^2(1+\rho)$ with $\rho = \frac{1-r^2}{Mr^2}$, in such a way that finite-M corrections in the distribution (3.8) are simply captured by replacing $\sigma^2 = r^2$ (holding in the $M \to \infty$ limit) with $\sigma^2 = r^2(1+\rho)$; conversely, in the unsupervised case, n-point correlation functions of the χ variables contribute to the moments of the coupling matrix, and the spectral distribution is accordingly deformed in a non-trivial way.

4. Spectral tools at work: an application to retrieval

The Hopfield model and its variations are nothing but spin-glasses with a Hebbian-like prescription for the interactions, and the structure of the quenched disorder encoded in the coupling matrix is known to govern the thermodynamic behavior of the statistical-mechanical model, see for instance [40,53,61–64]. Thus, it is reasonable to expect that the properties of Hopfield-like models ultimately stem from the spectral details of the interaction matrix $\boldsymbol{J}$. In this section, we aim to provide details about the functioning of these models by applying the results derived so far. Let us start from the (deterministic) parallel dynamics for the neuronal configuration, which reads as

$$\boldsymbol{\sigma}^{(n+1)} = \text{sign}\,[\boldsymbol{J}(t)\cdot\boldsymbol{\sigma}^{(n)}]. \tag{4.1}$$

We stress that, here, the evolution time is represented by the integer n, while t is the dreaming time that is retained fixed: synaptic weights are quenched during the neural dynamics. We are primarily interested in the stability of specific configurations, that is, in the probability that the system in a configuration $\boldsymbol{\sigma}^{(0)}$ at time n will be in the same configuration at time $n+1$. In order to analyze the stability of a given initial configuration $\boldsymbol{\sigma}^{(0)}$, we consider the 1-step update of the neural network:

$$\sigma_i^{(1)} = \text{sign}\Big[\sum_{j=1}^N J_{ij}(t)\sigma_j^{(0)}\Big], \tag{4.2}$$

and say that the configuration $\boldsymbol{\sigma}^{(0)}$ is stable at the neuron-index i if $\sigma_i^{(1)} = \sigma_i^{(0)}$. Although a 1-step horizon may appear rather limited, as we will see, it is enough to understand important properties about the evolution of the models under consideration. Also, we incidentally notice that this is a standard time span in machine-learning training algorithms like CD-1 [65] and that checking the stability in the 1-step dynamics can be recast in checking the stability by signal-to-noise-techniques [35].

By multiplying both sides of Eq. (4.2) by $\sigma_i^{(0)}$ and by exploiting the binary nature of $\sigma_i^{(0)}$, we get

$$\sigma_i^{(1)}\sigma_i^{(0)} = \text{sign}\Big[\sum_{j=1}^N J_{ij}(t)\sigma_j^{(0)}\sigma_i^{(0)}\Big], \tag{4.3}$$

and notice that, if the argument of the sign function is positive (negative), $\sigma_i^{(0)}$ is stable (unstable). In fact, $\frac{1}{2}[N - \sum_i \sigma_i^{(1)}\sigma_i^{(0)}]$ represents the overall number of neurons that change their state in the first step of the dynamics, namely the Hamming distance $d_H(\boldsymbol{\sigma}^{(0)}, \boldsymbol{\sigma}^{(1)})$ between $\boldsymbol{\sigma}^{(0)}$ and $\boldsymbol{\sigma}^{(1)}$. Thus, we introduce the following (see also [4,36,62])

Definition 3. Given a configuration $\boldsymbol{\sigma}$, the *stability* of its i-th neuron is

$$\Delta_i(\boldsymbol{\sigma}) = \sum_{j=1}^{N} J_{ij}(t)\sigma_j\sigma_i. \tag{4.4}$$

Besides the notion of stability, the width of the attraction basins plays a crucial role as for retrieval (when dealing with ground patterns) or generalization (when dealing with examples) capabilities [37]. In order to get access to these properties, we focus on a target configuration $\boldsymbol{x}$ (i.e., $\boldsymbol{x} = \boldsymbol{\xi}^\mu$ in the basic storing setting or $\boldsymbol{x} = \boldsymbol{\zeta}^\mu$ in the (un)supervised setting) and on an initial configuration $\boldsymbol{\sigma}^{(0)}$ lying on the boundary of the Hamming ball $\mathcal{B}_R(\boldsymbol{x})$ centered in $\boldsymbol{x}$ and with radius R. These boundaries can be realized by perturbing $\boldsymbol{x}$ as $x_i \to x'_i = \eta_i x_i$, with $\eta_i \sim \mathrm{Rad}(p)$, in such a way that, $d_H(\boldsymbol{x}',\boldsymbol{x}) = \frac{1}{2}\sum_i (1-\eta_i) \underset{N\gg 1}{\approx} \frac{N}{2}(1-p) = R$. Then, being $\boldsymbol{\sigma}^{(0)} = \boldsymbol{x}' = \boldsymbol{\eta} \odot \boldsymbol{x}$ and multiplying both sides of (4.2) by x_i, we obtain $\sigma_i^{(1)} x_i = \mathrm{sign}\left[\sum_{j=1}^N J_{ij}(t) x_j \eta_j x_i\right]$: a positive argument of the sign function means that, in a single step, the state of the i-th neuron either remains equal to the target x_i or changes from $-x_i$ to x_i. This motivates the following

Definition 4. Given two configurations $\boldsymbol{x}$ and $\boldsymbol{x}' = \boldsymbol{\eta} \odot \boldsymbol{x}$, the attractiveness of the i-th neuron in $\boldsymbol{x}$ w.r.t. $\boldsymbol{x}'$ is the random variable

$$\Delta_i(\boldsymbol{x},\boldsymbol{\eta}) = \sum_{j=1}^{N} J_{ij}(t) x_j \eta_j x_i, \tag{4.5}$$

where $\eta_i \underset{i.i.d.}{\sim} \mathrm{Rad}(p)$. Trivially, when $p = 1$, $\Delta_i(\boldsymbol{x},\boldsymbol{\eta})$ recovers the stability $\Delta_i(\boldsymbol{x})$ of x_i.

Recalling that $\boldsymbol{x}'$ is taken as the initial configuration and $\boldsymbol{x}$ is a target configuration, we can restate these concepts by introducing the configuration overlaps

$$m^{(0)}(\boldsymbol{x},\boldsymbol{\eta}) = \frac{1}{N}\sum_{i=1}^{N} x_i \sigma_i^{(0)} = \frac{1}{N}\sum_{i=1}^{N} \eta_i \tag{4.6}$$

$$m^{(1)}(\boldsymbol{x},\boldsymbol{\eta}) = \frac{1}{N}\sum_{i=1}^{N} x_i \sigma_i^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{sign}[\Delta_i(\boldsymbol{x},\boldsymbol{\eta})], \tag{4.7}$$

and say that $\boldsymbol{x}$ is attracting $\boldsymbol{\sigma}^{(0)}$ if $m^{(1)} > m^{(0)}$. When $\boldsymbol{x}$ coincides with a pattern, the overlaps above are also known as Mattis magnetizations (related to that pattern) evaluated at time steps $n = 0, 1$.

The above-defined stability and attractiveness (which we denote in general as Δ_i, as the meaning shall be clear from the context) are simple tools, but general enough to be handled in any scenario we are interested in. In order to further simplify the computations and obtain closed-form expressions for the 1-step Mattis magnetization, we will carry out our computations under the following

Working assumption *(Gaussian approximation).* Within the Gaussian approximation (GA), we assume that the quantities Δ_i are

- i.i.d. random variables;
- Gaussian distributed.

Clearly, this is not valid in general, however, in all the settings under examination, this assumption leads to a good approximation of the numerical results (see also App. D).

Within the GA and in the thermodynamic limit, the 1-step Mattis magnetization w.r.t. the reference configuration $\boldsymbol{x}$ is given by

$$m^{(1)}(\boldsymbol{x},\boldsymbol{\eta}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{sign}\left[\Delta_i(\boldsymbol{x},\boldsymbol{\eta})\right] \xrightarrow[TDL-GA]{} 2\,\mathrm{Prob}[\Delta \geq 0] - 1 = \mathrm{erf}\left[\frac{\mu_1}{\sqrt{2(\mu_2 - \mu_1^2)}}\right] = m^{(1)}(\boldsymbol{x},p), \tag{4.8}$$

where the argument of the error function is the usual signal-to-noise ratio [4], with μ_1 and μ_2 the first and (non-centered) second moments of the attractiveness, that is, following the GA, $\Delta_i \sim \mathcal{N}(\mu_1, \mu_2 - \mu_1^2)$; as we will show in the following, $\mu_{1,2}$ can be expressed in terms of integrals over the Marchenko-Pastur law. To avoid confusion, from now on we will denote with μ_ξ^t the limiting spectral measure of the basic-storing coupling matrix $\boldsymbol{J}^\xi$, while with μ_s^t and μ_u^t resp. those of $\boldsymbol{J}^s$ and $\boldsymbol{J}^u$.

Let us first focus on the basic storing case. In this setting, we are both interested in the stability and attractiveness of the patterns. Then, the following proposition holds:

Proposition 2. *In the basic storing setting, within the GA and in the thermodynamic limit:*

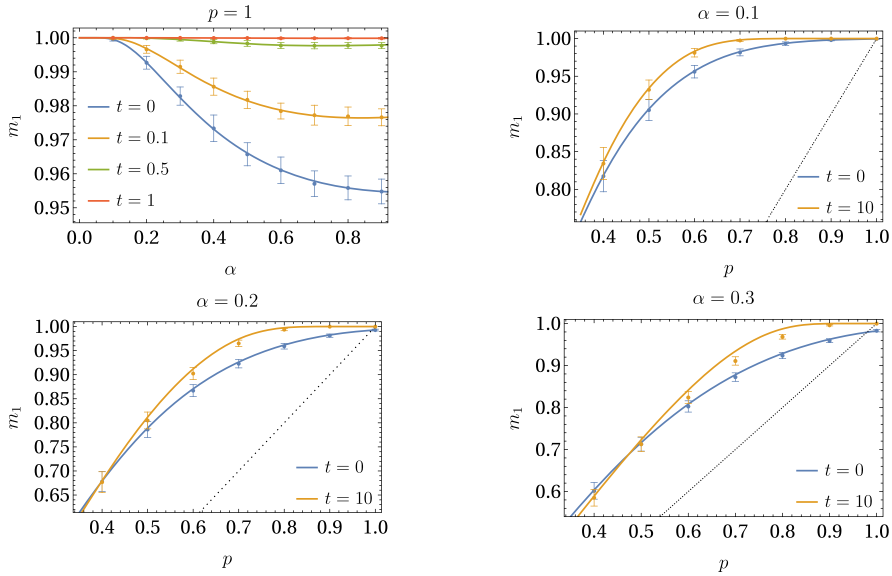


Fig. 3. Stability and attractiveness of patterns in the basic storing setting. The figure shows a comparison between the theoretical predictions of stability (upper left plot) and attractiveness (other plots). In the former, the 1-step magnetization $m^{(1)}$ starting from one of the patterns ($p = 1$) is given as a function of α, while for the attractiveness we fixed $\alpha = 0.1, 0.2, 0.3$ and $t = 0, 10$ (resp. Hebbian and large dreaming time limit) and consider $m^{(1)}$ as a function of the noise level p of the starting configuration. In these plots, the dashed line is the identity function $m^{(1)}(\xi, p) = m^{(0)}(\xi, p) = p$. In the numerical simulations, we averaged over 100 different realizations of the patterns, for systems with fixed size $N = 5000$. In these plots, m_1 stands for $m^{(1)}$.

1. *The first and second moments of the pattern stability are, respectively,*

$$\mu_1 = \frac{1}{\alpha} \int \frac{\lambda^2}{1 + t(1-\lambda)} d\mu_\xi^t(\lambda), \tag{4.9}$$

$$\mu_2 = \frac{1}{\alpha} \int \frac{\lambda^3}{1 + t(1-\lambda)} d\mu_\xi^t(\lambda). \tag{4.10}$$

2. *The first and second moments of the pattern attractiveness are, respectively,*

$$\mu_1 = \frac{p}{\alpha} \int \frac{\lambda^2}{1 + t(1-\lambda)} d\mu_\xi^t(\lambda), \tag{4.11}$$

$$\mu_2 = (1-p^2) \int \lambda^2 d\mu_\xi^t(\lambda) + \frac{p^2}{\alpha} \int \frac{\lambda^3}{1 + t(1-\lambda)} d\mu_\xi^t(\lambda). \tag{4.12}$$

The proof of this proposition is provided in App. D along with details on the validity of the GA.

Once the first two moments are estimated, we can predict the 1-step Mattis magnetization according to Eq. (4.8). In particular, at $t = 0$, the results reported Eqs. (4.11)-(4.12) lead to $\mu_1 = 1 + \alpha$ and $\mu_2 = \alpha^2 + 3\alpha + 1$, so that

$$m^{(1)}(\xi, 1) = \text{erf}\left(\frac{1+\alpha}{\sqrt{2\alpha}}\right),$$

which recovers the well-known expression for the expected magnetization in the Hopfield model [4].[3] At $t \gg 1$, we get $\mu_1 = 1 - \frac{\alpha}{(\alpha-1)t^2} + \mathcal{O}(t^{-3})$ and $\mu_2 = 1 - \frac{3\alpha}{(\alpha-1)t^2} + \mathcal{O}(t^{-3})$, whence

$$m^{(1)}(\xi, 1) \underset{t \gg 1}{=} \text{erf}\left(\frac{\sqrt{1-\alpha t}}{\sqrt{2\alpha}} + \mathcal{O}(t^0)\right) \approx 1 - \frac{1}{t} \frac{\exp(-\frac{1-\alpha}{2\alpha} t^2)}{\sqrt{\pi \frac{1-\alpha}{2\alpha}}}.$$

Similar results can also be obtained for $m^{(1)}(\xi, p)$. These theoretical predictions and the relative comparison with numerical results are shown in Fig. 3 for different values of the tunable parameters α, p, t. As expected, the stability of a pattern is impaired by α, but dreaming can mitigate this effect (see the upper left panel in Fig. 3). Dreaming can also enhance the attractiveness of a pattern, yielding a large overlap $m^{(1)}(\xi, p) \approx 1$ for a relatively large range of noise values p (see the upper right and lower panels in Fig. 3).

[3] The factor $1 + \alpha$ at the numerator in the error function is due to the fact that we are also including self-interactions. If $J_{ii} = 0$, instead, we would have $\text{erf}(1/\sqrt{2\alpha})$ [66].

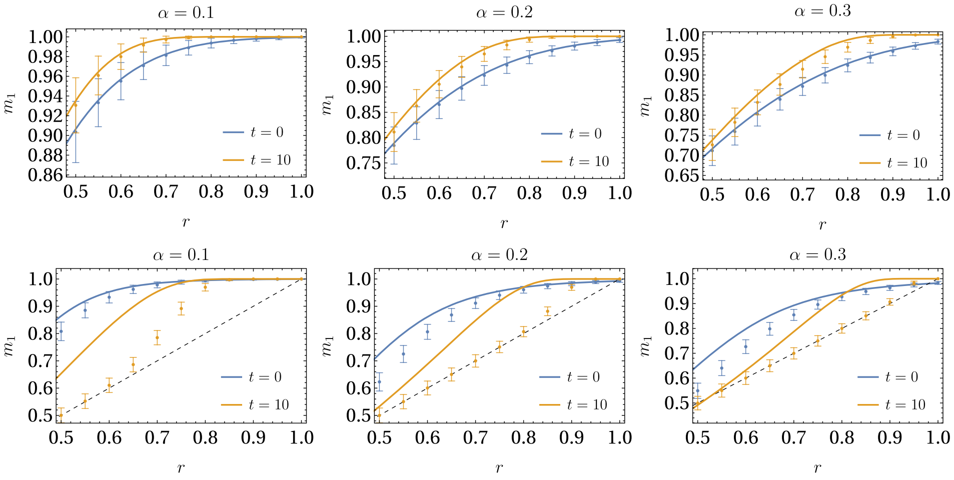


Fig. 4. Attractiveness of ground-truths in the supervised and unsupervised settings. The plot shows a comparison between the theoretical (following the results in Proposition 3) and the numerical estimates of $m^{(1)}(\zeta, r)$ for the supervised (first row) and unsupervised (second row) settings. Numerical results are averaged over 100 different realization of $M = 1000$ examples by varying α and r, and 100 different realization of the initial conditions. In this case, initial conditions are testing examples, i.e. examples with the same statistics as the training points, but which are not stored as fixed points. The system size is fixed to $N = 1000$. In the plots, m_1 stands for $m^{(1)}$.

In the supervised and unsupervised settings, rather than in the attracting power of stored (training) examples, we are interested in the *generalization* capabilities of the model, that is in the attractiveness of the ground-truths underlying the training dataset. Specifically, in these settings we consider a starting *test* configuration $\bar{\boldsymbol{\sigma}} = \boldsymbol{\chi} \odot \boldsymbol{\zeta}^\mu$ for some $\mu = 1, \dots, P$, being $\chi_i \sim \text{Rad}(r)$ for any i, namely, we consider a starting configuration that lays at an expected distance $R = N(1-r)/2$ from the target and that displays the same statistics of the stored examples; of course, training and testing items are independently drawn. Then, we are interested in the probability that, after a 1-step update, the neural configuration is aligned with the ground-truth $\boldsymbol{\zeta}^\mu$ and the related attractiveness reads as

$$\Delta_i(\boldsymbol{\zeta}^\mu, \boldsymbol{\chi}) = \sum_{j=1}^{N} J_{ij}(t)\zeta_i^\mu \bar{\sigma}_j = \sum_{j=1}^{N} J_{ij}(t)\chi_j \zeta_j^\mu \zeta_i^\mu. \tag{4.13}$$

Then, the following proposition holds.

Proposition 3. *Under the GA and in the thermodynamic limit, the first and second moments of the attractiveness* (4.13) *read*

1. *In the supervised setting:*

$$\mu_1 = \frac{1}{\alpha r}\int \frac{\lambda^2}{1+t(1-\lambda)} d\mu_s^t(\lambda), \tag{4.14}$$

$$\mu_2 = (1-r^2)\int \lambda^2 d\mu_s^t(\lambda) + \frac{1}{\alpha}\int \frac{\lambda^3}{1+t(1-\lambda)} d\mu_s^t(\lambda); \tag{4.15}$$

2. *In the unsupervised setting:*

$$\mu_1 = \frac{1}{\alpha r}\int \frac{\lambda^2}{1+t(1-\lambda)} d\mu_u^t(\lambda) - \frac{1-r^2}{r}\int \lambda d\mu_u^t(\lambda), \tag{4.16}$$

$$\mu_2 = \frac{1}{\alpha}\int \frac{\lambda^3}{1+t(1-\lambda)} d\mu_u^t(\lambda). \tag{4.17}$$

The proof of this proposition can be found in App. E.

These results can now be plugged into Eq. (4.8) to obtain an analytical estimate of $m^{(1)}(\zeta, r)$; in Fig. 4 we show a comparison between such theoretical predictions, and numerical simulations.

In particular, in the supervised scenario we highlight a positive role of dreaming for any load α, while in the unsupervised scenario the effects of dreaming do not have an obvious outcome. In fact, at relatively low load, a large t is detrimental for retrieving ground patterns; on the other hand, when α is relatively large, increasing t can be slightly beneficial, at least as long as the dataset is not too corrupted. Finally, we emphasize that, in Figs. 3-4, the deviation of the numerical results w.r.t. the theoretical predictions has to be ascribed to the break-down of the GA as the initial condition is progressively moved away from the target pattern.

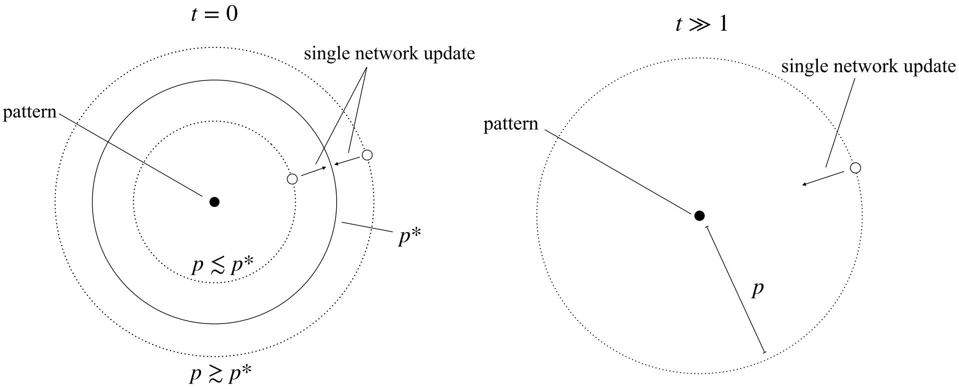


Fig. 5. Schematic representation of attractors in the basic storing setting. The figure shows a pictorial representation of fixed points in the Hopfield model (left) and dreaming model (right) at large dreaming time $t \gg 1$. For large α (above the critical storage capacity $\alpha_c \approx 0.14$ for the Hopfield model), fixed points are the balls centered in the pattern with Hamming radius $R(p^*) = \frac{N}{2}(1-p^*)$, while in the dreaming model for large t patterns are fixed points.

5. Discussion of the results

To conclude this work, we comment on the (1-step) retrieval and generalization capabilities of the Hebbian-like models in the settings under examination.

First, let us consider the attractiveness of patterns in the basic storing setting, so we refer to Fig. 3. Looking at the upper right plot ($\alpha = 0.1$, just below the critical storage capacity of the Hopfield model $\alpha_c \approx 0.14$), we see that, in both cases $t = 0$ (Hopfield model) and $t = 10$ (essentially the projector case), stored patterns exhibit a strong attractiveness w.r.t. noisy initial conditions. In particular, for all values of p, $m^{(1)}(\xi, p) \geq m^{(0)}(\xi, p)$, the only difference between the two extreme values of the dreaming time consists in the range of p where the 1-step update leads to $m^{(1)} \approx 1$: when $t \gg 1$, retrieval is by far more robust versus noise in the initial condition. However, as α is increased above the Hopfield model critical storage capacity (see e.g. $\alpha = 0.2$ and $\alpha = 0.3$), the situation is different. Indeed, in the $t = 0$ case, a non-trivial solution of the equation $m^{(1)}(\xi, p) = p$ appears (this is more evident in the $\alpha = 0.3$ case), corresponding to a specific value of the noise in the initial condition, say p^*, for which the network update does not lead to a higher magnetization: the network is stacked on the boundary of the ball $\mathcal{B}_R(\xi^\mu)$ centered in the pattern ξ^μ with radius $R(p^*) = \frac{N}{2}(1-p^*)$. Further, for $p > p^*$, after the network update, the final Mattis magnetization is *lower* than p, meaning that the system is getting farther from the pattern, while for $p < p^*$, the magnetization $m^{(1)}$ increases. Although this is a 1-step result, it evidences that, in the high-load Hopfield model, the patterns are no longer stable configurations under neural dynamics, while the fixed points are on the boundary of balls with a non-zero radius ($R(p^*)$ in the 1-step case), see the left picture in Fig. 5 for a schematic representation. This is in agreement with the results reported in [67]. Increasing the dreaming time t can fix this behavior: even at large α, the dreaming model displays $m^{(1)}(\xi, p) > p$, and a relatively wide range of p for which $m^{(1)} \approx 1$. Thus, in the basic storing setting, the dreaming model always exhibits better retrieval capabilities than the standard Hopfield model, and, for large t, patterns are attractive, see the right picture in Fig. 5.

The supervised setting shares the same features of the basic storing case, as the empirical, class-wise means are – for dataset with multiplicative, uncorrelated noise – a good prototype of the corresponding ground-truths. In this case, the dreaming mechanism allows the coupling matrix to get more and more insensitive to the quality of the examples as t is increased, as we already noticed in Fig. 2. Thus, the above discussion for the basic storing setting holds in this case too, as can be checked by the first row in Fig. 4.

Conversely, in the unsupervised setting, the situation is quite different. First, as can be seen in the second row of Fig. 4, the theoretical predictions exhibit large deviations w.r.t. the numerical results; specifically, the results derived by spectral tools always overestimate the numerical results. This signals a statistical dependence among the terms contributing to the attractiveness (4.5), so that the GA gets weaker. Despite this deviation, our results capture the qualitative behavior of the setting. In particular, for a relatively-high noise-level in the training and testing sets (i.e., for small r), the Hopfield model performs better than the dreaming model. This is due to the fact that, at low t, the attraction basins associated to training points corresponding to the same class are wide enough to merge, in such a way that their center of mass, (approximately) corresponding to the ground-truth, results to be an attracting point [37]. Yet, by increasing t, the attraction basins of the single training items shrink, so that they no longer overlap (especially if r is low) and, consequently, the system can only retrieve training examples (in fact, at low r, the 1-step magnetization settles on the identity line $m^{(1)}(\zeta, r) = r$), hence loosing its generalization capabilities. Increasing α, the Hopfield model undergoes the same behavior as in the supervised setting: ground-truths are no longer fixed points, no matter how accurate the training points are, that is, even for $r = 1$; on the other hand, implementing dreaming mechanisms can lead to appreciable results, as long as the sample quality is very high, that is $r \approx 1$, otherwise we still tend to retrieve training examples.

In conclusion, the spectral results derived in this paper have been applied to investigate the retrieval properties of Hopfield-like models, and, even in the worse scenario (where our working assumptions break down), we were able to give a qualitative picture of the processes taking place in associative neural networks while relaxing to fixed points for the neural dynamics.

Data availability

No data was used for the research described in the article.

Acknowledgements

E.A. and A.F. acknowledge financial support from PNRR MUR project PE0000013-FAIR and from Sapienza University of Rome (RM12117A8590B3FA, RM12218169691087, AR2221815D7192C1). A.F. has been fully supported by PNRR MUR project PE0000013-FAIR. A.F. acknowledges CUP_E53C22001930001 from Gruppo Nazionale di Fisica Matematica.

E.A. and A.F. acknowledge the stimulating research environment provided by the Alan Turing Institute's Theory and Methods Challenge Fortnights event "Physics-informed Machine Learning".

Appendix A. Proof of Lemma 1

1. Differentiating (3.1) w.r.t. the dreaming time, we have

$$\dot{\boldsymbol{J}} = \frac{1}{D_N}\boldsymbol{X}^T \frac{d}{dt}\Big(\frac{1+t}{\boldsymbol{1}+t\boldsymbol{C}}\Big)\boldsymbol{X} = \frac{1}{D_N}\boldsymbol{X}\Big(\frac{1}{\boldsymbol{1}+t\boldsymbol{C}} - (1+t)\frac{1}{\boldsymbol{1}+t\boldsymbol{C}}\boldsymbol{C}\frac{1}{\boldsymbol{1}+t\boldsymbol{C}}\Big)\boldsymbol{X}^T =$$
$$= \frac{1}{1+t}\Big(\frac{1}{D_N}\boldsymbol{X}\frac{1+t}{\boldsymbol{1}+t\boldsymbol{C}}\boldsymbol{X}^T - \frac{1}{D_N}\boldsymbol{X}\frac{1+t}{\boldsymbol{1}+t\boldsymbol{C}}\frac{\boldsymbol{X}\boldsymbol{X}^T}{D_N}\frac{1+t}{\boldsymbol{1}+t\boldsymbol{C}}\boldsymbol{X}^T\Big) =$$
$$= \frac{1}{1+t}(\boldsymbol{J}(t) - \boldsymbol{J}(t)^2).$$

2. Let us start from the eigenvalue problem $\boldsymbol{J}\boldsymbol{v}_\alpha = \lambda_\alpha \boldsymbol{v}_\alpha$, and differentiate w.r.t. the dreaming time:

$$(\boldsymbol{J}\dot{\boldsymbol{v}}_\alpha) = \dot{\boldsymbol{J}}\boldsymbol{v}_\alpha + \boldsymbol{J}\dot{\boldsymbol{v}}_\alpha = \frac{1}{1+t}(\boldsymbol{J}-\boldsymbol{J}^2)\boldsymbol{v}_\alpha + \boldsymbol{J}\dot{\boldsymbol{v}}_\alpha = \frac{1}{1+t}(\lambda_\alpha - \lambda_\alpha^2)\boldsymbol{v}_\alpha + \boldsymbol{J}\dot{\boldsymbol{v}}_\alpha.$$

On the other hand, we have

$$(\boldsymbol{J}\dot{\boldsymbol{v}}_\alpha) = \dot{\lambda}_\alpha \boldsymbol{v}_\alpha + \lambda_\alpha \dot{\boldsymbol{v}}_\alpha.$$

Combining the previous equations, we have

$$\dot{\lambda}_\alpha \boldsymbol{v}_\alpha + \lambda_\alpha \dot{\boldsymbol{v}}_\alpha = \frac{1}{1+t}(\lambda_\alpha - \lambda_\alpha^2)\boldsymbol{v}_\alpha + \boldsymbol{J}\dot{\boldsymbol{v}}_\alpha.$$

Moving the terms proportional to $\boldsymbol{v}_\alpha$ in the l.h.s. and moving those involving $\dot{\boldsymbol{v}}_\alpha$ in the r.h.s., we have

$$\big(\dot{\lambda}_\alpha - \frac{1}{1+t}\lambda_\alpha + \frac{1}{1+t}\lambda_\alpha^2\big)\boldsymbol{v}_\alpha = (\boldsymbol{J} - \lambda_\alpha)\dot{\boldsymbol{v}}_\alpha.$$

Multiplying on the left by $\boldsymbol{v}_\alpha^T$, which is a left eigenvector of $\boldsymbol{J}$, the r.h.s. is zero, and – due to the fact that $\boldsymbol{v}_\alpha \neq 0$, we have

$$\dot{\lambda}_\alpha - \frac{1}{1+t}\lambda_\alpha + \frac{1}{1+t}\lambda_\alpha^2 = 0. \tag{A.1}$$

The solution of the differential equation is

$$\lambda_\alpha(t) = \frac{1+t}{1+t\lambda_\alpha^0}\lambda_\alpha^0,$$

where λ_α^0 is the generic eigenvalue of the Hebbian coupling matrix $\boldsymbol{J}^0 = \frac{1}{D_N}\boldsymbol{X}^T\boldsymbol{X}$.

3. Let us first consider the random pattern case. In this setting, the Hebbian coupling matrix $\boldsymbol{J}^0 = \frac{1}{N}\sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu$ is a positive semidefinite random matrix with rank $P \leq N$, so it has P positive eigenvalues. From the theory of Wishart matrices [68,69], it is known that positive eigenvalues are distinct with probability 1, so $\lambda_1^0 > \cdots > \lambda_P^0$, and the eigenvalue $\lambda^0 = 0$ has degeneracy $N - P$. Since the application $\lambda_\alpha^0 \to \lambda_\alpha(t)$ given by Eq. (3.3) is injective, the algebraic multiplicity of eigenvalues in the spectrum is preserved for all $t > 0$ finite (at $t \to \infty$, all positive eigenvalues concentrate around $\lambda = 1$). Let us now consider eigenvectors of $\boldsymbol{J}(t)$ with positive eigenvalue. Starting from

$$\dot{\lambda}_\alpha \boldsymbol{v}_\alpha + \lambda_\alpha \dot{\boldsymbol{v}}_\alpha = \frac{1}{1+t}(\lambda_\alpha - \lambda_\alpha^2)\boldsymbol{v}_\alpha + \boldsymbol{J}\dot{\boldsymbol{v}}_\alpha,$$

given in the previous point, and using the differential equation (A.1), we get

$$(\boldsymbol{J} - \lambda_\alpha)\dot{\boldsymbol{v}}_\alpha = 0,$$

so that $\dot{\boldsymbol{v}}_\alpha$ is also eigenvector of $\boldsymbol{J}$ with the same eigenvalue of $\boldsymbol{v}_\alpha$. Since for positive eigenvalues, the associated eigenspace is one-dimensional, it follows that $\dot{\boldsymbol{v}}_\alpha = c(t)\boldsymbol{v}_\alpha$, whose solution is

$$\boldsymbol{v}_\alpha(t) = \boldsymbol{v}_\alpha(0) \cdot \exp \int^t dt' c(t'). \tag{A.2}$$

Then, the only effect of dreaming time is the rescaling of the eigenvectors norm. If we take $\boldsymbol{v}_\alpha(t)$ to be the normalized eigenvectors with positive eigenvalues, it follows that $\boldsymbol{v}_\alpha(t) = \boldsymbol{v}_\alpha(0)$, i.e. they do not depend on t. For the eigenvalue $\lambda = 0$, its eigenspace is $N - P$-dimensional. Denoting $\boldsymbol{v}_0^{(1)}(t), \dots, \boldsymbol{v}_0^{(N-P)}(t)$ at finite dreaming time $t \geq 0$ the associate eigenvectors, we can write

$$\boldsymbol{v}_0^{(n)}(t) = \sum_{m=1}^{N-P} U_{n,m}(t) \boldsymbol{v}_0^{(m)}(0), \quad n = 1, \dots, N - P,$$

which is nothing but a change of basis in the $\lambda = 0$ eigenspace. Since we are free to map eigenspaces in themselves without altering the structure of coupling matrix, we can choose $\boldsymbol{U}(t) = \mathbf{1}$ for all $t > 0$.

For the supervised setting the situation is analogous, the only difference being the different structure of the information vectors, whose relevant details are directly encoded in the $t = 0$ limit of the coupling matrix $\boldsymbol{J}^0$. For the unsupervised setting, the situation is different. First of all, it is clear that the rank of the matrix will be $\min(N, PM) \equiv N \min(1, \alpha M)$. However, we are mostly interested in the case where the number of examples per class is sufficiently high, i.e. $M \gg 1$ regardless of the value of α: in this case, the rank of the coupling matrix $\boldsymbol{J}^0$ would be N, then the matrix is full-rank, and all of the eigenvalues would be positive and distinct (for the same reasons of the random pattern case). Thus, positivity and non-degeneracy of the eigenvalues, relation (3.3) and stability of eigenvectors w.r.t. dreaming mechanism trivially follows also in this case. □

Appendix B. Proof of Theorem 1

1. Let us start again with the basic storing case, where $\boldsymbol{X} = \boldsymbol{\xi}$. As we already said, in this setting the eigenvalue $\lambda = 0$ has degeneracy $N - P$, thus the limiting spectral distribution would have a delta around 0 with mass $1 - \alpha$. We now focus on positive eigenvalues and start again with the eigenvalue problem $\boldsymbol{J}^0 \boldsymbol{v}_\alpha = \frac{1}{N} \boldsymbol{\xi}^T \boldsymbol{\xi} \boldsymbol{v}_\alpha = \lambda_\alpha^0 \boldsymbol{v}_\alpha$. Multiplying on the left by $\boldsymbol{\xi}$, we have

$$\frac{1}{N} \boldsymbol{\xi} \boldsymbol{\xi}^T \boldsymbol{\xi} \boldsymbol{v}_\alpha = \boldsymbol{C} \boldsymbol{\xi} \boldsymbol{v}_\alpha = \lambda_\alpha^0 \boldsymbol{\xi} \boldsymbol{v}_\alpha.$$

Thus, positive eigenvalues of the coupling matrix $\boldsymbol{J}^0$ are exactly the eigenvalues of the usual correlation matrix, and the corresponding eigenvector is

$$\boldsymbol{e}_\alpha = \frac{1}{\sqrt{\lambda_\alpha^0 N}} \boldsymbol{\xi} \boldsymbol{v}_\alpha, \tag{B.1}$$

where the prefactor is needed to ensure the normalization $\|\boldsymbol{e}_\alpha\| = 1$. By universality arguments holding for centered patterns with finite variance [70], positive eigenvalues of the correlation matrix will be Marchenko-Pastur-distributed with $\mathrm{MP}(\alpha) = \mathrm{MP}(\alpha, 1)$, since $\mathbb{E}(\xi_i^\mu)^2 = 1$. Since positive eigenvalues have mass α, it trivially follows that the empirical spectral distribution $\mu_N^0(\lambda)$ will converge in weak topology to μ^0, in such a way that

$$d\mu^0(\lambda) = (1 - \alpha)\delta(\lambda) d\lambda + \alpha d\mu_{\mathrm{MP}}(\lambda), \tag{B.2}$$

with

$$d\mu_{\mathrm{MP}}(\lambda) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+^0 - \lambda)(\lambda - \lambda_-^0)}}{\alpha \lambda} d\lambda,$$

and $\lambda_\pm^0 = (1 \pm \sqrt{\alpha})^2$. Thus, in the random pattern case, $\hat{\lambda}^0 = 0$ and $\sigma^2 = 1$.

For the supervised case, the situation is similar. Indeed, also in this case the spectral distribution will have a delta peak at $\lambda = 0$ with mass $1 - \alpha$. For the bulk distribution, the relation between positive eigenvalues of the coupling matrix $\boldsymbol{J}^0$ and the correlation matrix (which is now computed with the empirical means of examples in each class), still holds provided that we replace $\xi_i^\mu \to \bar{\xi}_i^\mu = \frac{1}{M} \sum_A \xi_i^{\mu,A}$. By strong law of large numbers, $\bar{\xi}_i^\mu \overset{a.s.}{\to} \mathbb{E}_\chi \bar{\xi}_i^\mu = r\zeta_i^\mu$; this means that $J_{ij}^0 \overset{a.s.}{\to} \frac{1}{N} \sum_\mu \mathbb{E}_\chi \bar{\xi}_i^\mu \mathbb{E}_\chi \bar{\xi}_j^\mu = \frac{1}{N} \sum_\mu (r\zeta_i^\mu)(r\zeta_j^\mu) = r^2 \boldsymbol{J}_{ij}^{0,\zeta}$, where $J_{ij}^{0,\zeta}$ is the Hebbian matrix in the random pattern case built with the ground-truths features, i.e.

$$\boldsymbol{J}_{ij}^{0,\zeta} = \frac{1}{N} \sum_\mu \zeta_i^\mu \zeta_j^\mu.$$

Now, notice that $\mathbb{E}_\zeta(\zeta_i^\mu) = 0$ and $\mathbb{E}_\zeta(\zeta_i^\mu)^2 = 1$ in the $M \to \infty$ limit, eigenvalues of the correlation matrix

$$\boldsymbol{C} = \frac{1}{D_N} \boldsymbol{X} \boldsymbol{X}^T,$$

are distributed according to the Marchenko-Pastur law $\mathrm{MP}(\alpha, r^2)$. Thus, in the supervised case, the empirical spectral distribution of the coupling matrix $\boldsymbol{J}^0$ will converge in weak topology to μ^0, with measure

$$d\mu^0(\lambda) = (1-\alpha)\delta(\lambda)d\lambda + \alpha d\mu_{\text{MP}}(\lambda), \tag{B.3}$$

with

$$d\mu_{\text{MP}}(\lambda) = \frac{1}{2\pi r^2}\frac{\sqrt{(\lambda_+^0-\lambda)(\lambda-\lambda_-^0)}}{\alpha\lambda}d\lambda, \tag{B.4}$$

and $\lambda_\pm^0 = r^2(1\pm\sqrt{\alpha})^2$; thus, in the supervised setting, $\hat{\lambda}^0 = 0$ and $\sigma^2 = r^2$.

In the unsupervised case, by strong law of large number, for $i \neq j$ we have $J_{ij}^0 \overset{a.s.}{\to} \mathbb{E}_\chi J_{ij}^0$ as $M \to \infty$, while $J_{ii}^0 = \alpha$. Thus, in this case we can safely replace the coupling matrix $\boldsymbol{J}^0$ with its noise-independent version:

$$\boldsymbol{J}^0 \overset{a.s.}{\to} \alpha(1-r^2)\mathbf{1} + r^2\boldsymbol{J}^{0,\zeta}, \tag{B.5}$$

where $J_{ij}^{0,\zeta}$ is again the Hebbian matrix in the random pattern case built with the ground-truths features. Translating Eq. (B.5) for the eigenvalues of $\boldsymbol{J}^0$, we see that the quantity

$$\frac{\lambda^0 - \alpha(1-r^2)}{r^2}$$

has the same distribution of the eigenvalues of the random pattern case. Then, it follows that, as $M \to \infty$ and in the thermodynamic limit, the empirical spectral distribution μ_N^0 of the unsupervised coupling matrix converges in weak topology μ^0, with

$$d\mu^0(\lambda) = (1-\alpha)\delta(\lambda-\alpha(1-r^2))d\lambda + \alpha d\mu_{\text{MP}}(\lambda),$$

with

$$d\mu_{\text{MP}}(\lambda) = \frac{1}{2\pi r^2}\frac{\sqrt{(\lambda_+^0-\lambda)(\lambda-\lambda_-^0)}}{\alpha(\lambda-\alpha(1-r^2))}d\lambda,$$

with $\lambda_\pm^0 = r^2(1\pm\sqrt{\alpha})^2 + \alpha(1-r^2)$. Thus, in the unsupervised case, we have $\hat{\lambda}^0 = \alpha(1-r^2)$ and $\sigma^2 = r^2$.

2. The proof works by reverting Eq. (3.3), expressing λ_α^0 as a function of $\lambda_\alpha(t)$. Thus, in the thermodynamic limit (and eventually for $M \to \infty$), the empirical spectral distribution μ_N^t will converge in weak topology to μ^t, with the latter determined by the fact that the quantity

$$\frac{\lambda}{1+t(1-\lambda)},$$

will be equal in distribution to λ^0, regardless of the setting under consideration. □

Appendix C. Proof of Proposition 1

First of all, let us notice that, since in the $M \to \infty$ limit, $\boldsymbol{J}^0 \overset{a.s.}{\to} r^2\boldsymbol{J}^{0,\zeta}$ in the supervised setting and $\boldsymbol{J}^0 \overset{a.s.}{\to} \alpha(1-r^2)\mathbf{1} + r^2\boldsymbol{J}^{0,\zeta}$ in the unsupervised one, and the fact that eigenvectors can be chosen so that they do not depend on t, $\boldsymbol{J}$ and $\boldsymbol{J}^\zeta$ (the latter being the dreaming coupling matrix in the random pattern case built with the ground-truths) have common eigenvectors a.s., so they can be simultaneously diagonalized with the transformation $\boldsymbol{J} \to \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ with the same matrix $\boldsymbol{U}$. Because of these simple relations between the (un)supervised coupling matrix and the corresponding ground-truth version, a functional relation between eigenvalues can be derived. For example, the generic eigenvalue $\lambda^s(t)$ of the supervised coupling matrix is related to the corresponding eigenvalue of $\boldsymbol{J}^0$ through Eq. (3.3). In the $M \to \infty$ limit, $\lambda_0^s \to r^2\lambda_0$, where λ_0 is the corresponding eigenvalue of $\boldsymbol{J}^{0,\zeta}$, thus

$$\lambda^s(t) = \frac{(1+t)\lambda_0^s}{1+t\lambda_0^s} = \frac{(1+t)r^2\lambda_0}{1+tr^2\lambda_0}.$$

Finally, one can re-express λ_0 in terms of $\lambda(t)$ being the eigenvalue of $\boldsymbol{J}^\zeta(t)$ the coupling matrix of the ground-truths by reverting Eq. (3.3), giving us

$$\lambda^s(t) = \frac{\lambda(t)r^2(t+1)}{\lambda(t)\left(r^2-1\right)t+t+1} = f_{r,t}^s(\lambda(t)).$$

Clearly, with the same procedure, one finds that the functional relation for the eigenvalues of the unsupervised coupling matrix is $\lambda^u(t) = f_{r,t}^u(\lambda(t))$. With these results, we find for the SE the expression

$$\mathcal{E}^{s,u}(\alpha,r,t) = \frac{1}{N}\text{Tr}(\boldsymbol{J}^\zeta - \boldsymbol{J}^{s,u}(t))^2 = \frac{1}{N}\sum_\alpha(\lambda_\alpha(t) - f_{r,t}^{s,u}(\lambda_\alpha(t)))^2 \to \int(\lambda - f_{r,t}^{s,u}(\lambda))^2 d\mu_\zeta^t(\lambda), \tag{C.1}$$

where → stands for convergence in probability in the thermodynamic limit. □

Remark 4. Because of the structure of the limiting spectral distribution of the dreaming coupling matrix $d\mu^t$, the SE can be rewritten as

$$\mathcal{E}^s(\alpha,r,t) = \alpha\int(\lambda - f^s_{r,t}(\lambda))^2 d\mu^t_{\mathrm{bulk}}(\lambda),$$

for the supervised setting, and

$$\mathcal{E}^u(\alpha,r,t) = (1-\alpha)\Big[\frac{\alpha\left(r^2-1\right)(t+1)}{\alpha\left(r^2-1\right)t-1}\Big]^2 + \alpha\int(\lambda - f^u_{r,t}(\lambda))^2 d\mu^t_{\mathrm{bulk}}(\lambda).$$

In the last expression, the constant contribution comes from the presence of the delta peak located at non-vanishing eigenvalue for the coupling matrix in the unsupervised setting.

Appendix D. Proof of Proposition 2 and details on the GA

The proof works by explicit computation of the empirical moments. Let us start with the pattern stability. Under the GA assumption in the thermodynamic limit, and since patterns are equivalent (so that we can take the average also average the index μ), we can estimate

$$\mu_1 = \frac{1}{NP}\sum_{i\mu}\Delta_i(\xi^\mu), \tag{D.1}$$

$$\mu_2 = \frac{1}{NP}\sum_{i\mu}\Delta_i(\xi^\mu)^2. \tag{D.2}$$

For the first quantity, we have

$$\mu_1 = \frac{1}{NP}\sum_{i\mu}J_{ij}(t)\xi^\mu_i\xi^\mu_j = \frac{1}{\alpha N}\sum_{ij}J_{ij}(t)J_{ij}(0) = \frac{1}{\alpha N}\mathrm{Tr}\boldsymbol{J}(t)\boldsymbol{J}(0). \tag{D.3}$$

Now, $\boldsymbol{J}(t)$ and $\boldsymbol{J}^0$ are simultaneously diagonalizable, thus

$$\mu_1 = \frac{1}{\alpha N}\sum_\alpha \lambda_\alpha(t)\lambda^0_\alpha \xrightarrow[TDL]{} \frac{1}{\alpha}\int\frac{\lambda^2}{1+t(1-\lambda)}d\mu^t_\xi(\lambda), \tag{D.4}$$

where we expressed λ^0_α as a function of $\lambda_\alpha(t)$ by reverting Eq. (3.3). Analogously, for the second moment

$$\begin{aligned}\mu_2 &= \frac{1}{NP}\sum_{i\mu jk}J_{ij}(t)J_{ik}(t)\xi^\mu_j\xi^\mu_k = \frac{1}{\alpha N}\sum_{ijk}J_{ij}(t)J_{ik}(t)J_{jk}(0) = \\ &= \frac{1}{\alpha N}\mathrm{Tr}\boldsymbol{J}(t)^2\boldsymbol{J}(0) \xrightarrow[TDL]{} \frac{1}{\alpha}\int\frac{\lambda^3}{1+t(1-\lambda)}d\mu^t_\xi(\lambda).\end{aligned} \tag{D.5}$$

As for the attractiveness, the computations follow the same lines, provided that we use (4.5) as definition, and under the GA we average the moments w.r.t. $\boldsymbol{\eta}$, and noticing that $\mathbb{E}_{\boldsymbol{\eta}}\eta_i = p$ and $\mathbb{E}_{\boldsymbol{\eta}}\eta_j\eta_k = (1-p^2)\delta_{jk} + p^2$. □

Remark 5. Notice that we can recast everything in terms of integrals of usual Marchenko-Pastur distribution with scale parameter $\alpha < 1$. Indeed, by using spectral decomposition of the coupling matrix we can write the first empirical moment of the stability as

$$\begin{aligned}\mu_1 &= \frac{1}{NP}\mathbb{E}_\xi\sum_{ij\mu}J_{ij}\xi^\mu_i\xi^\mu_j = \frac{1}{NP}\mathbb{E}_\xi\sum_{\alpha ij\mu}\lambda_\alpha v^i_\alpha v^j_\alpha\xi^\mu_i\xi^\mu_j = \frac{1}{NP}\mathbb{E}_\xi\sum_{\alpha\mu}\lambda_\alpha\Big(\sum_i v^i_\alpha\xi^\mu_i\Big)\Big(\sum_j v^j_\alpha\xi^\mu_j\Big) = \\ &= \frac{1}{NP}\sum_{\alpha\mu}\lambda_\alpha(\sqrt{\lambda^0_\alpha N}e^\mu_\alpha)^2 = \frac{1}{P}\sum_\alpha\lambda_\alpha\lambda^0_\alpha \xrightarrow[TDL]{} \int\frac{(1+t)\lambda^2}{1+t\lambda}d\mu_{\mathrm{MP}}(\lambda),\end{aligned} \tag{D.6}$$

where we used $\boldsymbol{e}_\alpha = (\lambda^0_\alpha N)^{-1/2}\boldsymbol{\xi}\boldsymbol{v}_\alpha$ such that $\sum_\mu(e^\mu_\alpha)^2 = 1$ are the eigenvectors of the correlation matrix, and the fact that $\lambda_\alpha = (1+t)\lambda^0_\alpha/(1+t\lambda^0_\alpha)$ and that the coupling matrices have only P positive eigenvalues. Similarly, for the second moment

$$\mu_2 \xrightarrow[TDL]{} \int\frac{(1+t)^2\lambda^3}{(1+t\lambda)^2}d\mu_{\mathrm{MP}}(\lambda). \tag{D.7}$$

Remark 6. In order to check the validity of the GA, we consider the third centered moment of the attractiveness, which in the thermodynamic limit can be approximated as

$$\mathbb{E}_{\boldsymbol{\eta}}(\Delta^\mu_i - \mathbb{E}_{\boldsymbol{\eta}}\Delta^\mu_i)^3 \underset{TDL}{\sim} \frac{1}{NP}\mathbb{E}_\xi\sum_{i\mu jkl}J_{ij}J_{ik}J_{il}\xi^\mu_i\xi^\mu_j\xi^\mu_k\xi^\mu_l\mathbb{E}_{\boldsymbol{\eta}}(\eta_j - p)(\eta_k - p)(\eta_l - p). \tag{D.8}$$

Noticing that $\mathbb{E}_{\boldsymbol{\eta}}(\eta_j - p)(\eta_k - p)(\eta_l - p) = 2p(1-p^2)\delta_{jk}\delta_{kl}$, it follows that

$$\mathbb{E}_{\eta}(\Delta_i^{\mu}-\mathbb{E}_{\eta}\Delta_i^{\mu})^3=\frac{2p(1-p^2)}{NP}\mathbb{E}_{\xi}\sum_{i\mu j}J_{ij}^3\xi_i^{\mu}\xi_j^{\mu}. \tag{D.9}$$

We can thus bound the third centered moment as

$$|\mathbb{E}_{\eta}(\Delta_i^{\mu}-\mathbb{E}_{\eta}\Delta_i^{\mu})^3|\le\frac{2p(1-p^2)}{NP}\mathbb{E}_{\xi}\sum_{i\mu j}|J_{ij}|^3=\frac{2p(1-p^2)}{N}\mathbb{E}_{\xi}\sum_{ij}|J_{ij}|^3\le\frac{2p(1-p^2)}{N}\mathbb{E}_{\xi}\sum_{ij}|J_{ij}|^2=$$
$$=\frac{2p(1-p^2)}{N}\mathbb{E}_{\xi}\mathrm{Tr}\boldsymbol{J}^2\to 2p(1-p^2)\int\lambda^2 d\mu_{\xi}^t(\lambda),$$

where we used the fact that $|J_{ij}|\le 1$. Then, a necessary condition for the third centered moment to be close to zero is

$$2\alpha p(1-p^2)\int\lambda^2 d\mu_{\xi}^t(\lambda)\ll 1.$$

Appendix E. Proof of Proposition 3

1. In the supervised setting, the empirical first moment of the attractiveness is

$$\mu_1=\frac{1}{NP}\mathbb{E}_{\chi}\sum_{i\mu j}J_{ij}(t)\chi_j\zeta_i^{\mu}\zeta_j^{\mu}=\frac{r}{NP}\sum_{i\mu j}J_{ij}(t)\zeta_i^{\mu}\zeta_j^{\mu}=\frac{r}{\alpha N}\sum_{ij}J_{ij}(t)J_{ij}^{0,\zeta}=\frac{r}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)\boldsymbol{J}^{0,\zeta},$$

where $\boldsymbol{J}^{0,\zeta}$ is the Hebbian matrix in the archetype-setting with ground-truths $\boldsymbol{\zeta}$. In the $M\to\infty$ limit, since $\boldsymbol{J}^0\overset{a.s.}{\to}r^2\boldsymbol{J}^{0,\zeta}$, we can safely write

$$\mu_1=\frac{r}{\alpha N}\sum_{ij}J_{ij}(t)\frac{1}{r^2}J_{ij}^0=\frac{1}{\alpha rN}\mathrm{Tr}\,\boldsymbol{J}(t)\boldsymbol{J}^0\to\frac{1}{\alpha r}\int\frac{\lambda^2}{1+t(1-\lambda)}d\mu_s^t(\lambda),$$

in the thermodynamic limit. For the empirical second moment, we have

$$\mu_2=\frac{1}{NP}\mathbb{E}_{\chi}\sum_{i\mu jk}J_{ij}(t)J_{ik}(t)\chi_j\chi_k\zeta_j^{\mu}\zeta_k^{\mu}=\frac{1-r^2}{N}\sum_{ij}J_{ij}(t)^2+\frac{r^2}{NP}\sum_{i\mu jk}J_{ij}(t)J_{ik}(t)\zeta_j^{\mu}\zeta_k^{\mu}=$$
$$=\frac{1-r^2}{N}\mathrm{Tr}\,\boldsymbol{J}(t)^2+\frac{r^2}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)^2\boldsymbol{J}^{0,\zeta}=\frac{1-r^2}{N}\mathrm{Tr}\,\boldsymbol{J}(t)^2+\frac{1}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)^2\boldsymbol{J}^0\to$$
$$\to(1-r^2)\int\lambda^2 d\mu_s^t(\lambda)+\frac{1}{\alpha}\int\frac{\lambda^3}{1+t(1-\lambda)}d\mu_s^t(\lambda),$$

in the thermodynamic limit.

2. In the unsupervised setting, we still have

$$\mu_1=\frac{r}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)\boldsymbol{J}^{0,\zeta},$$

as in the supervised case. However, in the setting under consideration, in the $M\to\infty$ limit we have

$$\boldsymbol{J}^{0,\zeta}\equiv\frac{1}{r^2}(\boldsymbol{J}^0-\alpha(1-r^2)\mathbf{1}).$$

Thus, we have

$$\mu_1=\frac{1}{\alpha rN}\mathrm{Tr}\,\boldsymbol{J}(t)\boldsymbol{J}^0-\frac{1-r^2}{rN}\mathrm{Tr}\,\boldsymbol{J}(t)\to\frac{1}{\alpha r}\int\frac{\lambda^2}{1+t(1-\lambda)}d\mu_u^t(\lambda)-\frac{1-r^2}{r}\int\lambda d\mu_u^t(\lambda),$$

in the thermodynamic limit. Analogously, for the empirical second moment we have

$$\mu_2=\frac{1-r^2}{N}\mathrm{Tr}\,\boldsymbol{J}(t)^2+\frac{r^2}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)^2\boldsymbol{J}^{0,\zeta}=\frac{1-r^2}{N}\mathrm{Tr}\boldsymbol{J}(t)^2+\frac{1}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)^2\boldsymbol{J}^0-\frac{1-r^2}{N}\mathrm{Tr}\,\boldsymbol{J}(t)^2=$$
$$=\frac{1}{\alpha N}\mathrm{Tr}\,\boldsymbol{J}(t)^2\boldsymbol{J}^0\to\frac{1}{\alpha}\int\frac{\lambda^3}{1+t(1-\lambda)}d\mu_u^t(\lambda),$$

in the thermodynamic limit. □

References

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

[2] J.J. Hopfield, D.W. Tank, “Neural” computation of decisions in optimization problems, Biol. Cybern. 52 (3) (1985) 141–152.

[3] D. Hebb, The Organization of Behavior: A Neuropsychological Theory, John Wiley & Sons, New York, NY, 1949.

[4] D. Amit, Modeling Brain Function: The World of Attractor Neural Networks, Cambridge University Press, 1989.

[5] D.J. Amit, H. Gutfreund, H. Sompolinsky, Storing infinite numbers of patterns in a spin-glass model of neural networks, Phys. Rev. Lett. 55 (Sep 1985) 1530–1533.

[6] D.J. Amit, H. Gutfreund, H. Sompolinsky, Statistical mechanics of neural networks near saturation, Ann. Phys. 173 (1) (1987) 30–67.

[7] A. Bovier, V. Gayrard, P. Picco, Large deviation principles for the Hopfield model and the Kac-Hopfield model, Probab. Theory Relat. Fields 101 (4) (1995) 511–546.
[8] A. Bovier, V. Gayrard, An almost sure large deviation principle for the Hopfield model, Ann. Probab. 24 (3) (1996) 1444–1475.
[9] A. Barra, G. Genovese, F. Guerra, D. Tantari, How glassy are neural networks?, J. Stat. Mech. Theory Exp. 2012 (07) (2012) P07009.
[10] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, Generalized Guerra's interpolation schemes for dense associative neural networks, Neural Netw. 128 (2020) 254–267.
[11] E. Agliari, L. Albanese, A. Barra, G. Ottaviani, Replica symmetry breaking in neural networks: a few steps toward rigorous results, J. Phys. A, Math. Theor. 53 (2020).
[12] A. Bovier, Sharp upper bounds on perfect retrieval in the Hopfield model, J. Appl. Probab. 36 (3) (1999) 941–950.
[13] J. Feng, M. Shcherbina, B. Tirozzi, On the critical capacity of the Hopfield model, Commun. Math. Phys. 216 (2001) 139–177.
[14] D. Loukianova, Lower bounds on the restitution error in the Hopfield model, Probab. Theory Relat. Fields 107 (2) (1997) 161–176.
[15] C.M. Newman, Memory capacity in neural network models: rigorous lower bounds, Neural Netw. 1 (3) (1988) 223–238.
[16] M. Löwe, On the storage capacity of Hopfield models with correlated patterns, Ann. Appl. Probab. 8 (4) (1998) 1216–1250.
[17] A. Bovier, V. Gayrard, Rigorous results on the thermodynamics of the dilute Hopfield model, J. Stat. Phys. 72 (1993) 79–112.
[18] P. Baldi, S.S. Venkatesh, Number of stable points for spin-glasses and neural networks of higher orders, Phys. Rev. Lett. 58 (9) (1987) 913.
[19] A. Bovier, B. Niederhauser, The spin-glass phase-transition in the Hopfield model with p-spin interactions, Adv. Theor. Math. Phys. 5 (6) (2001) 1001–1046.
[20] E. Gardner, The space of interactions in neural network models, J. Phys. A 21 (1) (1988) 257.
[21] J.J. Hopfield, D.I. Feinstein, R.G. Palmer, Unlearning has a stabilizing effect in collective memories, Nature 304 (5922) (1983) 158–159.
[22] L. Personnaz, I. Guyon, G. Dreyfus, Information storage and retrieval in spin-glass like neural networks, J. Phys. Lett. 46 (8) (1985) 359–365.
[23] I. Kanter, H. Sompolinsky, Associative recall of memory without errors, Phys. Rev. A 35 (1) (1987) 380.
[24] A. Plakhov, S. Semenov, The modified unlearning procedure for enhancing storage capacity in Hopfield network, in: [Proceedings] 1992 RNNS/IEEE Symposium on Neuroinformatics and Neurocomputers, IEEE, 1992, pp. 242–251.
[25] A.Y. Plakhov, S.A. Semenov, I.B. Shuvalova, Convergent unlearning algorithm for the Hopfield neural network, in: Proceedings of the 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, IEEE, 1995, pp. 30–33.
[26] J. Van Hemmen, Hebbian learning, its correlation catastrophe, and unlearning, Netw. Comput. Neural Syst. 8 (3) (1997) V1.
[27] J.A. Horas, P.M. Pasinetti, On the unlearning procedure yielding a high-performance associative memory neural network, J. Phys. A, Math. Gen. 31 (25) (1998) L463.
[28] V. Dotsenko, N. Yarunin, E. Dorotheyev, Statistical mechanics of Hopfield-like neural networks with modified interactions, J. Phys. A, Math. Gen. 24 (10) (1991) 2419.
[29] V. Dotsenko, B. Tirozzi, Replica symmetry breaking in neural networks with modified pseudo-inverse interactions, J. Phys. A, Math. Gen. 24 (21) (1991) 5163.
[30] A. Fachechi, E. Agliari, A. Barra, Dreaming neural networks: forgetting spurious memories and reinforcing pure ones, Neural Netw. 112 (2019) 24–40.
[31] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, Dreaming neural networks: rigorous results, J. Stat. Mech. Theory Exp. 2019 (8) (2019) 083503.
[32] A. Fachechi, A. Barra, E. Agliari, F. Alemanno, Outperforming RBM feature-extraction capabilities by "dreaming" mechanism, IEEE Trans. Neural Netw. Learn. Syst. 35 (1) (2024) 1172–1181.
[33] J. Fontanari, Generalization in a Hopfield network, J. Phys. Fr. 51 (1990) 2421–2430.
[34] E. Agliari, F. Alemanno, A. Barra, G. De Marzo, The emergence of a concept in shallow neural networks, Neural Netw. 148 (2022) 232–253.
[35] M. Aquaro, F. Alemanno, I. Kanter, A. Barra, E. Agliari, Supervised Hebbian learning, Europhys. Lett., Perspect. 141 (2023) 11001.
[36] M. Benedetti, L. Carillo, E. Marinari, M. Mézard, Eigenvector dreaming, J. Stat. Mech. Theory Exp. 2024 (1) (2024) 013302.
[37] E. Agliari, M. Aquaro, F. Alemanno, A. Fachechi, Regularization, early-stopping and dreaming: a Hopfield-like setup to address generalization and overfitting, arXiv preprint, arXiv:2308.01421, 2023.
[38] F.E. Leonelli, E. Agliari, L. Albanese, A. Barra, On the effective initialisation for restricted Boltzmann machines via duality with Hopfield model, Neural Netw. 143 (2021) 314–326.
[39] J.M. Kosterlitz, D.J. Thouless, R.C. Jones, Spherical model of a spin-glass, Phys. Rev. Lett. 36 (20) (1976) 1217.
[40] S. Galluccio, J.-P. Bouchaud, M. Potters, Rational decisions, random matrices and spin glasses, Physica A, Stat. Mech. Appl. 259 (3–4) (1998) 449–456.
[41] A. Auffinger, G.B. Arous, J. Černỳ, Random matrices and complexity of spin glasses, Commun. Pure Appl. Math. 66 (2) (2013) 165–201.
[42] L. Zhu, W.-w. Xu, The inverse eigenvalue problem of structured matrices from the design of Hopfield neural networks, Appl. Math. Comput. 273 (2016) 1–7.
[43] J. Pennington, Y. Bahri, Geometry of neural network loss surfaces via random matrix theory, in: International Conference on Machine Learning, PMLR, 2017, pp. 2798–2806.
[44] X. Mai, R. Couillet, A random matrix analysis and improvement of semi-supervised learning for large dimensional data, J. Mach. Learn. Res. 19 (1) (2018) 3074–3100.
[45] Z. Liao, R. Couillet, The dynamics of learning: a random matrix approach, in: International Conference on Machine Learning, PMLR, 2018, pp. 3072–3081.
[46] M.E.A. Seddik, C. Louart, M. Tamaazousti, R. Couillet, Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures, in: International Conference on Machine Learning, PMLR, 2020, pp. 8573–8582.
[47] J. Zhou, Z. Jiang, T. Hou, Z. Chen, K.M. Wong, H. Huang, Eigenvalue spectrum of neural networks with arbitrary Hebbian length, Phys. Rev. E 104 (6) (2021) 064307.
[48] R. Couillet, Z. Liao, Random Matrix Methods for Machine Learning, Cambridge University Press, 2022.
[49] D. Granziol, N. Baskerville, A random matrix theory approach to damping in deep learning, J. Phys. Complex. 3 (2) (2022) 024001.
[50] J. Barbier, F. Camilli, M. Mondelli, M. Sáenz, Fundamental limits in structured principal component analysis and how to reach them, Proc. Natl. Acad. Sci. 120 (30) (2023) e2302028120.
[51] L. Zdeborová, F. Krzakala, Statistical physics of inference: thresholds and algorithms, Adv. Phys. 65 (5) (2016) 453–552.
[52] E. Agliari, A. Barra, P. Sollich, L. Zdeborova, Machine learning and statistical physics: theory, inspiration, application, J. Phys. A, Math. Theor. (2020), Special volume.
[53] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, On the Marchenko–Pastur law in analog bipartite spin-glasses, J. Phys. A, Math. Theor. 52 (May 2019) 254002.
[54] P. Zanin, N. Caticha, Interacting dreaming neural networks, J. Stat. Mech. 2034 (2023) 043401.
[55] L. Serricchio, C. Chilin, D. Bocchi, R. Marino, M. Negri, C. Cammarota, F. Ricci-Tersenghi, Daydreaming Hopfield networks and their surprising effectiveness on correlated data, in: Associative Memory & Hopfield Networks in 2023, 2023.
[56] F. Camilli, M. Mézard, The decimation scheme for symmetric matrix factorization, vol. 2034, arXiv:2307.16564v1, 2023.
[57] E. Ventura, S. Cocco, R. Monasson, F. Zamponi, Unlearning regularization for Boltzmann machines 16 (2023) 1065–1095, https://arxiv.org/pdf/2311.09418.pdf.
[58] T. Kohonen, M. Ruohonen, Representation of associated data by matrix operators, IEEE Trans. Comput. (1973).
[59] L. Albanese, A. Barra, P. Bianco, F. Durante, D. Pallara, Hebbian learning from first principles, arXiv:2401.07110.
[60] E. Agliari, F. Alemanno, M. Aquaro, A. Barra, F. Durante, I. Kanter, Hebbian dreaming for small datasets, Neural Netw. 173 (2024) 106174.
[61] H.J. Sommers, A. Crisanti, H. Sompolinsky, Y. Stein, Spectrum of large random asymmetric matrices, Phys. Rev. Lett. 60 (May 1988) 1895–1898.
[62] W. Krauth, M. Mézard, J.-P. Nadal, Basins of attraction in a perceptron-like neural network, Complex Syst. 2 (4) (1988) 387–408.
[63] K. Rajan, L.F. Abbott, Eigenvalue spectra of random matrices for neural networks, Phys. Rev. Lett. 97 (Nov 2006) 188104.

[64] T. Rogers, I.P. Castillo, R. Kühn, K. Takeda, Cavity approach to the spectral density of sparse symmetric random matrices, Phys. Rev. E 78 (Sep 2008) 031116.
[65] I. Sutskever, T. Tieleman, On the convergence properties of contrastive divergence, J. Mach. Learn. Res. 9 (2010) 9.
[66] J. Rocchi, D. Saad, D. Tantari, High storage capacity in the Hopfield model with auto-interactions—stability analysis, J. Phys. A, Math. Theor. 50 (Oct 2017) 465001.
[67] R. McEliece, E. Posner, E. Rodemich, S. Venkatesh, The capacity of the Hopfield associative memory, IEEE Trans. Inf. Theory 33 (4) (1987) 461–482.
[68] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, vol. 2, Wiley, New York, 1958.
[69] A.T. James, Distributions of matrix variates and latent roots derived from normal samples, Ann. Math. Stat. 35 (2) (1964) 475–501.
[70] N.S. Pillai, J. Yin, Universality of covariance matrices, Ann. Appl. Probab. 24 (3) (2014) 935–1001.