Bollettino dell'Unione Matematica Italiana
https://doi.org/10.1007/s40574-026-00538-2

Pattern disentanglement and reconstruction by modular Hebbian networks

Elena Agliari[1,3] · Shahed Akter[1] · Andrea Alessandrelli[2,3] · Alberto Fachechi[1,3]

Received: 31 January 2026 / Accepted: 1 July 2026

Abstract
We address the problem of reconstructing a set of K binary patterns $\boldsymbol{\Xi} = (\boldsymbol{\xi}^1, ..., \boldsymbol{\xi}^K) \in \{-1, +1\}^{N\times K}$, starting from their covariance matrix $\boldsymbol{\Xi}\,\boldsymbol{\Xi}^T$ and from a set of mixtures of the form $\boldsymbol{\sigma}_M = \text{sign}(\sum_{\mu\in S_M\subseteq\{1,...,K\}} \boldsymbol{\xi}^\mu)$, with $|S_M| = M$. To this aim, we engage a modular Hebbian network, specified by the interaction matrix $\boldsymbol{J}^H = \frac{1}{N}\boldsymbol{\Xi}\,\boldsymbol{\Xi}^T$, where $\boldsymbol{\sigma}_M$ is taken as the initial state, repeated in each of the constituting L modules and the stable state resulting from a Gibbs evolution rule is taken as an L-tuple of candidate patterns. Finally, by properly handling $\boldsymbol{J}^H$ we can derive the projector matrix $\boldsymbol{J}^K = \boldsymbol{\Xi}(\boldsymbol{\Xi}^T\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^T$ by which we can determine whether these candidate patterns are a good estimate for the ground patterns.

Contents

✉ Elena Agliari
agliari@mat.uniroma1.it

Alberto Fachechi
alberto.fachechi@uniroma1.it

[1] Dipartimento di Matematica, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy

[2] Dipartimento di Matematica e Fisica, Università del Salento, via Arnesano, 73100 Lecce, Italy

[3] Gruppo Nazionale per la Fisica Matematica (GNFM-INdAM), Rome, Italy

Published online: 28 July 2026

1 Introduction

Since the seminal works by Little, Amari, Pastur and Figotin, and Hopfield [9, 16, 19, 23], associative memory models have provided a powerful framework to describe how collective interactions among simple units can give rise to robust pattern retrieval and error correction capabilities. Beyond their biological motivation, these networks have found applications in a wide range of fields, including machine learning, signal processing, and inference problems, where the task often consists in reconstructing hidden structures from partial or noisy cues [2, 10, 12, 17, 18, 22, 25, 28].

In this work, we consider a class of associative memories consisting of L interacting modules (also referred to as *species* in the literature, see e.g. [1]), where the interaction between units belonging to the same module follows the usual Hebbian prescription, while the mutual influence of neurons in different modules is anti-imitative[1]. The key feature of these systems is that, while usual Hebbian contributions enable pattern retrieval, inter-module interaction explicitly disfavors the simultaneous retrieval of identical patterns across different modules. This mechanism makes these networks well-suited for retrieving distinct pattern in parallel, thereby addressing the specific task of *spurious mixture disentanglement*, rather than general retrieval or reconstruction capabilities. Indeed, as explored in [3, 4, 6, 8], given a set of binary patterns

$$\boldsymbol{\Xi} = (\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^K) \in \{-1, +1\}^{N \times K}, \tag{1.1}$$

and an input of the form

$$\text{sign}(\sum_{\mu \in \mathcal{S}} \boldsymbol{\xi}^\mu), \text{ with } \mathcal{S} \subseteq \{1, ..., K\}, \tag{1.2}$$

supplied to each module and representing a mixture of patterns, the network can spontaneously relax to configurations in which different modules align with different patterns composing the mixture.

This disentangling capability can possibly be leveraged for matrix factorization tasks (see e.g., [24]). In this setting the K binary patterns $\boldsymbol{\Xi}$ are assumed to be hidden, and the goal is to *infer* each of them given limited information. Specifically, we assume access to their covariance matrix $\boldsymbol{\Xi}\boldsymbol{\Xi}^T$ and to a collection of observations of the form (1.2) which represent nonlinear superpositions of a certain, unknown, number of patterns. Such mixtures may naturally arise in situations where measurements aggregate multiple sources and only thresholded outputs are available, rendering direct inversion infeasible. A crucial aspect of this approach is the post-processing of the output configurations. We show that, by suitably manipulating the Hebbian interaction matrix $\boldsymbol{J}^H = \boldsymbol{\Xi}\boldsymbol{\Xi}^T/N$, it is possible to reach the Kohonen projector matrix $\boldsymbol{J}^K = \boldsymbol{\Xi}(\boldsymbol{\Xi}^T\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^T$, which enables a quantitative assessment of the quality of the inferred patterns. In particular, this projection allows us to determine whether a retrieved configuration constitutes a reliable reconstruction of a ground pattern. As empirically shown in [5], this approach is able (at least in the so-called *low storage* regime) to reconstruct all the hidden patterns starting solely from their covariance, namely the Hebbian matrix $\boldsymbol{J}^H$, with a proper combination of hyper-parameter tuning and energy-based acceptance criterion. In this paper, we contribute to give a solid mathematical ground to such a procedure.

[1] Although this negative interaction between modules closely resembles the repulsive terms naturally emerging in inverse problems for pattern reconstruction [7, 11, 14, 21, 26], in this setting it is explicitly designed to force orthogonality of modules activity once the Hebbian correlation matrix is given.

2 The modular Hebbian network and the reconstruction algorithm

Let us consider a set of K binary vectors $\{\boldsymbol{\xi}^\mu\}_{\mu=1,...,K}$, collected as the columns of the $N \times K$ matrix $\boldsymbol{\Xi}$ (1.1). In order to allow for an analytical treatment, we will consider i.i.d. Rademacher entries, namely

$$P(\xi_i^\mu) = \frac{1}{2}\left(\delta_{\xi_i^\mu,-1} + \delta_{\xi_i^\mu,+1}\right), \text{ for any } i = 1, ..., N \text{ and for any } \mu = 1, ..., K. \tag{2.1}$$

Next, let us consider a set of L Hebbian networks, whose individual configurations are N-dimensional binary vectors denoted as $\boldsymbol{\sigma}^a = (\sigma_1^a, \ldots, \sigma_N^a)$, $a = 1, \ldots, L$, so that the whole state collecting the activity of each neuron of the system will read as $\boldsymbol{\sigma} \in \{-1, +1\}^{N\times L}$. As standard in attractor neural networks, we introduce the Mattis magnetizations as order parameters assessing the quality of the associative memory functionality of the system:

$$m_\mu^a(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i^a, \text{ for } \mu = 1, ..., K$$

or, in compact form, $\boldsymbol{m}(\boldsymbol{\sigma}) = \frac{\boldsymbol{\Xi}^T\boldsymbol{\sigma}}{N} \in [-1, +1]^{K\times L}$. Based on this definition, the Hamiltonian of the model studied here and introduced in [4, 5] reads as

$$\mathcal{E}_{N,\boldsymbol{\Xi}}(\boldsymbol{\sigma}) = -N\sum_{a,\mu=1}^{L,K}(m_\mu^a(\boldsymbol{\sigma}))^2 + N\lambda\sum_{\substack{a,b=1\\a\neq b}}^{L}\Big(\sum_{\mu=1}^{K} m_\mu^a(\boldsymbol{\sigma})m_\mu^b(\boldsymbol{\sigma})\Big)^2 - H\sum_{a,i=1}^{L,N} h_i^a\sigma_i^a =$$

$$= -\sum_{a=1}^{L}\sum_{i,j=1}^{N} J_{ij}^H\sigma_i^a\sigma_j^a + \frac{\lambda}{N}\sum_{\substack{a,b=1\\a\neq b}}^{L}\sum_{i,j,k,l=1}^{N} J_{ij}^H J_{kl}^H\sigma_i^a\sigma_j^b\sigma_k^a\sigma_l^b - H\sum_{a,i=1}^{L,N} h_i^a\sigma_i^a, \tag{2.2}$$

where $\boldsymbol{J}^H = \frac{\boldsymbol{\Xi}\cdot\boldsymbol{\Xi}^T}{N}$ is the Hebbian matrix, $\lambda \in [0, (L-1)^{-1})$ is the interaction strength between modules,[2] and $H \in \mathbb{R}_+$ is the magnitude of the external field. We emphasize that the squared form of the inter-module interaction is precisely chosen to penalize redundant retrieval of the same pattern across different modules, independently of the relative sign[3]. Furthermore, although the inter-layer interaction engages four neurons simultaneously, the related coupling strength can still be written in terms of Hebb's matrix entries, in contrast with classical dense associative memories where p-body interactions are typically tied to p-point correlations of the stored patterns.

Exploiting the mean-field structure of the model, the Hamiltonian can be conveniently recast as $\mathcal{E}_{N,\boldsymbol{\Xi}}(\boldsymbol{\sigma}) = -\sum_{a=1}^{L}\sum_{i=1}^{N}\hat{h}_i^a\sigma_i^a$ with the effective field $\hat{\boldsymbol{h}}^a$ acting on neurons in the layer a being the sum of three contributions:

$$\hat{\boldsymbol{h}}^a(\boldsymbol{\sigma}) = \boldsymbol{h}^{a\to a}(\boldsymbol{\sigma}) + \sum_{\substack{b=1\\b\neq a}}^{L}\boldsymbol{h}^{b\to a}(\boldsymbol{\sigma}) + H\,\boldsymbol{h}^a, \tag{2.3}$$

[2] The upper bound of λ ensures that $\mathcal{E}_{N,\boldsymbol{\Xi}}(\boldsymbol{\sigma})$ is negative definite, see also [6].

[3] By contrast, with a linear term $\sum_{\mu=1}^{K} m_\mu^a(\boldsymbol{\sigma})m_\mu^b(\boldsymbol{\sigma})$, configurations where two modules retrieve the same pattern with opposite signs would become energetically favorable; for instance, states of the form $(\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^\mu, -\boldsymbol{\xi}^\mu)$ would gain from the inter-module contribution, despite being undesirable for disentanglement purposes.

respectively, the intra-module (auto-associative), inter-module (anti-imitative) and external fields. The definitions of the first two contributions come directly from the expression (2.2), that is

$$\boldsymbol{h}^{a\to a}(\boldsymbol{\sigma}) = \boldsymbol{J}^{\mathrm{H}} \cdot \boldsymbol{\sigma}^{a}, \tag{2.4}$$

$$\boldsymbol{h}^{b\to a}(\boldsymbol{\sigma}) = -\frac{\lambda}{N}\big(\boldsymbol{J}^{\mathrm{H}} \cdot \boldsymbol{\sigma}^{b}\big)\big((\boldsymbol{\sigma}^{b})^{T}\boldsymbol{J}^{\mathrm{H}}\boldsymbol{\sigma}^{a}\big). \tag{2.5}$$

We now set up a dynamics that evolves configurations toward lower-energy states. Allowing for stochastic noise, controlled by the inverse temperature $\beta = T^{-1} \in \mathbb{R}_{+}$, the neuronal configuration $\boldsymbol{\sigma}(t)$ at (discrete) time t is updated synchronously according to

$$\boldsymbol{\sigma}^{a}(t+1) = \mathrm{sign}[\tanh(\beta \hat{\boldsymbol{h}}^{a}(\boldsymbol{\sigma}(t))) + \boldsymbol{u}^{a}(t)], \tag{2.6}$$

where $\hat{\boldsymbol{h}}^{a}$ is the a-th layer effective field, computed at each time-step t according to Eqs. (2.3-2.5), and $\boldsymbol{u}^{a}(t) \underset{i.i.d}{\sim} \mathcal{U}([-1,1]^{N})$ for all $a = 1, \ldots, L$.

The hyper-parameters of the model are i) the level of thermal noise β, ii) the strength λ of the anti-imitative interactions between modules, and iii) the strength H of the external field.

Using the neural dynamics (2.6) and preparing the network in a mixture state

$$\boldsymbol{\sigma}^{a}(0) = \boldsymbol{\sigma}_{M} = \mathrm{sign}\Big(\sum_{\mu\in\mathcal{S}_{M}} \boldsymbol{\xi}^{\mu}\Big), \ \text{for } a = 1, ..., L, \tag{2.7}$$

where $\mathcal{S}_{M} \subseteq \{1, ..., K\}$ is a set of indices containing M different pattern labels, i.e., $|\mathcal{S}_{M}| = M$, pattern disentanglement emerges from the competition between imitative intra-module couplings, which stabilize alignment with patterns, and anti-imitative inter-module couplings, which penalize the retrieval of the same pattern across modules, as empirically shown in [5].

Remarkably, the input of this disentangling algorithm consists solely of the matrix $\boldsymbol{J}^{H}$, together with a set of observations of the form $\boldsymbol{\sigma}_{M}$, in such a way that this system appears to be well suited also for reconstructing hidden patterns in a parallel way, given such input as the only accessible information[4]. In fact, we can design a reconstruction algorithm as follows (see also the pseudo-code in Appendix A):

- We initialize all modules in the network on a given input data $\boldsymbol{\sigma}_{M}$, and run the neural dynamics (2.6) upon convergence (or, due to the stochasticity of the process, once that some *a priori* defined stopping criterion is met); we repeat the process for each available input data, and collect the final configurations of each module in a set $\mathcal{V}$ of candidate reconstructed patterns.
- For each couple of candidates $\boldsymbol{\zeta}^{1}, \boldsymbol{\zeta}^{2} \in \mathcal{V}$, we compute the mutual overlap $q(\boldsymbol{\zeta}^{1}, \boldsymbol{\zeta}^{2}) = \frac{1}{N}\sum_{i=1}^{N} \zeta_{i}^{1}\zeta_{i}^{2}$, and we retain only those with sufficiently low overlap in order to discard duplicates. The optimal threshold on mutual overlap is a function of the hyper-parameters, and its determination can be carried out using a statistical-mechanical approach as detailed in [5].
- For any configuration surviving the overlap criterion, we only accept those which have a large projection on the vector space spanned by the patterns. In particular, denoting with $\boldsymbol{J}^{K}$ the projector on $\mathrm{Span}(\{\boldsymbol{\xi}^{1}, ..., \boldsymbol{\xi}^{K}\})$, we accept candidates $\boldsymbol{\zeta} \in \mathcal{V}$ with $\frac{1}{N}\boldsymbol{\zeta}^{T}\boldsymbol{J}^{K}\boldsymbol{\zeta} \geq \tau$ for some acceptance threshold τ. This step avoids misreconstruction of

[4] In fact, in this context, the input needs not be restricted to configurations of the form $\boldsymbol{\sigma}_{M}$: any configuration exhibiting sufficiently high correlation with the hidden patterns can work, for instance $\mathrm{sign}(\sum_{\mu} c_{\mu}\boldsymbol{\xi}^{\mu})$ with non-random coefficients c_{μ}.

the hidden patterns that would harm the performance of the algorithm. The projector $\boldsymbol{J}^K$ can be uniquely determined as the solution of an iterative mechanism, also referred to as *dreaming algorithm* [15], whose starting condition is the Hebbian coupling matrix $\boldsymbol{J}^H$.

The investigation of this protocol was initiated in [5] and here, we give rigorous basis on its functioning. In particular, in Sec. 3, we prove that an appropriate combination of hyperparameters leads to the destabilization of spurious mixture states while stabilizing the stored patterns, thereby avoiding the intrinsic loophole of the Hopfield model whereby spurious states remain stable at low temperatures. Next, in Sec. 4, we give a renewed and refined proof for the convergence of the dreaming algorithm to the projecting kernel $\boldsymbol{J}^K$. Then, in Sec. 5, we discuss proper choices of the acceptance threshold τ. Finally, in Sec. 6, we provide further numerical analysis to support our theoretical claims.

3 Insights from signal-to-noise analysis

In this section, we present the results obtained from a signal-to-noise analysis. Although this approach is restricted to the *one-step* behavior of the neural dynamics – namely, the evolution of the network after a single Monte Carlo (MC) update starting from a prescribed initial configuration – it nevertheless yields valuable insights into the retrieval capabilities of the system. In particular, it allows us to assess the stability of candidate retrieval states as the hyper-parameters are varied, as well as to estimate their attraction basins. Our investigation is carried out in the thermodynamic limit, namely $N \to \infty$ with $K/N \to \alpha \in [0, 1]$. Throughout the discussion we will mostly consider the low-storage regime, corresponding to $\alpha = 0$, in which no glassy phenomena are expected.

We initialize the system in a starting condition $\boldsymbol{\sigma}(0)$, compute the local effective field $\hat{\boldsymbol{h}}^a$ at time $t = 0$ for each module, and estimate the overlap of the network state after a single update with some target configuration $\boldsymbol{x} \in \{-1, +1\}^N$. To do this, let us first notice that the Markov process

$$\sigma_i^a(t+1) = \sigma_i^a(t)\,\mathrm{sign}[\hat{h}_i^a(\boldsymbol{\sigma}(t))\,\sigma_i^a(t) + T\,\tanh^{-1}(u_i^a(t))], \tag{3.1}$$

with $u_i^a \sim \mathcal{U}([-1, 1])$ is equivalent to the neural dynamics given in Eq. (2.6), in the sense that the law of $\boldsymbol{\sigma}(t+1) \mid \boldsymbol{\sigma}(t)$ is the same for both the processes. Focusing on $t = 0$, we define the overlap between the updated state and the target configuration

$$q(\boldsymbol{\sigma}^a(1), \boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N x_i\,\sigma_i^a(1) = \frac{1}{N}\sum_{i=1}^N \mathrm{sign}[\hat{h}_i^a(\boldsymbol{\sigma}(0))\,x_i + T\,\tanh^{-1}(u_i^a(0))\,x_i\,\sigma_i^a(0)], \tag{3.2}$$

whose conditional expectation given the patterns is

$$\mathbb{E}_{\boldsymbol{u}}[q(\boldsymbol{\sigma}^a(1), \boldsymbol{x})|\,\boldsymbol{\Xi}] = \frac{1}{N}\sum_{i=1}^N \tanh\left(\frac{\hat{h}_i^a(\boldsymbol{\sigma}(0))x_i}{T}\right).$$

Since our primary focus here is to analyze the disentanglement capabilities of the machine, we consider the initial condition (2.7), or some partially disentangled state, meaning that D out of L modules ($0 \le D < \min(M, L)$) may already be aligned with some of the patterns involved in the mixture $\boldsymbol{\sigma}_M$. The relevant target states are:

i) The initial mixture $\boldsymbol{\sigma}_M$; in this case, $\bar{m}^a_{mix,M} = \mathbb{E}_{\boldsymbol{u}}[q(\boldsymbol{\sigma}^a(1), \boldsymbol{\sigma}_M)|\boldsymbol{\Xi}]$ measures the stability of the mixture: the lower $\bar{m}^a_{mix,M}$ and the farther the network configuration gets from the initial condition, therefore signaling that these spurious states are not fixed points.
ii) A stored pattern involved in the combination $\boldsymbol{\sigma}_M$; in this case, $\bar{m}^a_\mu = \mathbb{E}_{\boldsymbol{u}}[q(\boldsymbol{\sigma}^a(1), \boldsymbol{\xi}^\mu)|\boldsymbol{\Xi}]$ measures the capability of the network to reconstruct patterns within the nonlinear combination $\boldsymbol{\sigma}_M$, thus probing the disentangling power of the machine: the larger $\bar{m}^a_\mu$ (with $\mu \in \mathcal{S}_M$) and the closer the update moves the network to the manifold of relevant disentangled states.

Before stating the main results of this section, it is useful to report the following

Proposition 1 *Let $m_\mu(\boldsymbol{\sigma}_M) = \frac{1}{N}\sum_{i=1}^N \xi_i^\mu \operatorname{sign}(\sum_{\nu\in\mathcal{S}_M}\xi_i^\nu)$ be the overlap of the mixture state $\boldsymbol{\sigma}_M$ with the pattern $\boldsymbol{\xi}^\mu$, and $M \in 2\mathbb{N}+1$. Then, in the thermodynamic limit, $m_\mu \to \mathcal{C}_M 1_{\mu\in\mathcal{S}_M}$ $\boldsymbol{\Xi}$-almost surely, and $\sqrt{N}(m^a_\mu - \mathcal{C}_M 1_{\mu\in\mathcal{S}_M}) \xrightarrow{d} \mathcal{N}(0, 1-\mathcal{C}_M^2 1_{\mu\in\mathcal{S}_M})$, where*

$$\mathcal{C}_M = \frac{1}{2^{M-1}}\binom{M}{\frac{1}{2}(M-1)}.$$

The proof works by joining the strong law of large numbers, the central limit theorem and $\mathbb{E}\ \xi^\mu \operatorname{sign}(\sum_{\mu=1}^M \xi^\mu) = \mathcal{C}_M 1_{1\le\mu\le M}$, see Appendix B for further details. The reason for restricting our attention to nonlinear combinations of an odd number of pure components is that, as in the Hopfield setting, the even ones are already unstable states. With these ingredients, we are able to state the following

Theorem 1 *Consider a L-modular Hebbian network with size N and K stored Rademacher patterns, and let $\boldsymbol{\sigma}_M = \operatorname{sign}(\sum_{\mu\in\mathcal{S}_M}\boldsymbol{\xi}^\mu)$ with $|\mathcal{S}_M| = M \in 2\mathbb{N}+1$ be a nonlinear mixture of an odd number of stored patterns. Then:*

i) Let $\boldsymbol{\sigma}^a(0) = \boldsymbol{\sigma}_M$ and set the external fields as $\boldsymbol{h}^a = \boldsymbol{\sigma}_M$ for all $a = 1, \dots, L$. In the thermodynamic limit $N \to \infty$ with $K/N \to 0$, the 1-step stability of the state $\boldsymbol{\sigma}_M$ is

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\Xi}} \bar{m}^a_{mix,M} = \mathbb{E}_{X_M} \tanh\left(\frac{H + A_\infty \mathcal{C}_M |X_M|}{T}\right), \tag{3.3}$$

with $A_\infty = 1-\lambda(L-1)M\mathcal{C}_M^2$ and $X_M = \sum_{l=1}^M \varepsilon_l$ be a random walk with $\varepsilon_l \underset{i.i.d.}{\sim} Rad(I)$.

ii) Let $I \subset \mathcal{S}_M$, with $|I| = D \le L-1$, and let $\boldsymbol{\sigma}^a(0) = \boldsymbol{\xi}^{\nu_a}$, $\nu_a \in I$ for $a = 1, \dots, D$, and $\boldsymbol{\sigma}^a(0) = \boldsymbol{\sigma}_M$ otherwise. In the thermodynamic limit $N \to \infty$ with $K/N \to 0$ and $D > 0$, the 1-step magnetization of the generic non-aligned layer ($b > D$) with any pattern $\boldsymbol{\xi}^\mu$ is

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\Xi}} \bar{m}^b_\mu = \begin{cases} 0 & \mu \notin \mathcal{S}_M \\ \mathbb{E}_{X_M} \frac{V}{D} \tanh(\frac{1}{T}A) & \mu \in I \\ \mathbb{E}_{X_M} \frac{U-V}{M-D} \tanh(\frac{1}{T}A) & \mu \in \mathcal{S}_M \backslash I \end{cases}, \tag{3.4}$$

with $X_M = X_D^{(1)} + X_{M-D}^{(2)}$, $X_s^a = \sum_{l=1}^s \varepsilon_l^{(a)}$ being a random walk with $\varepsilon_l^{(a)} \sim_{i.i.d.} Rad(I)$, $U = |X_M|$ and $V = \operatorname{sign}(X_M)X_D^{(1)}$, $A = H + A_\infty \mathcal{C}_M U - \lambda \mathcal{C}_M V$ and $A_\infty = 1 - \lambda M \mathcal{C}_M^2 (L-D-1)$. Conversely, if $D = 0$ (i.e., modules are all prepared in $\boldsymbol{\sigma}_M$) the 1-step magnetization w.r.t. any pattern $\boldsymbol{\xi}^\mu$ is

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\Xi}} \bar{m}^b_\mu = \begin{cases} 0 & \mu \notin \mathcal{S}_M \\ \mathbb{E}_{X_M} \frac{|X_M|}{M} \tanh\left(\frac{H + A_\infty \mathcal{C}_M |X_M|}{T}\right) & \mu \in \mathcal{S}_M \end{cases}. \tag{3.5}$$

with $A_\infty = 1 - \lambda(L-1)MC_M^2$.

Remark 1 The results of Thm. 1 are intrinsically symmetric under pattern relabeling within a given class ($\mathcal{S}_M$, I or $\mathcal{S}_M \backslash I$), as it considers all possible trajectories in the configuration space of the system, and thereby averages over all those converge to the same disentangled configurations, irrespective of permutations of the modules. In actual neural evolutions achieving disentanglement, a layer initially prepared in a spurious mixture $\boldsymbol{\sigma}_M$ will magnetize with a specific pattern in that combination (thus breaking the pattern-exchange symmetry), while loosing correlation with the others because of orthogonality. To match numerics with theory, one should further average the empirical magnetization over patterns within the same class.

The proof of the theorem is reported in Appendix B. These results suggest how to properly tune the strength λ of the anti-imitative interaction between modules in such a way that stored patterns ($M = 1$) are stable states, while spurious ones ($M \geq 3$) are made unstable. Thus, it is natural to compare $\bar{m}^a_{mix,1}$[5] with the envelope $\max_{M\geq 3, M\text{odd}}(\bar{m}^a_{mix,M})$, and look for regions where the former is close to 1, and the latter is as low as possible. In Fig. 1 we report such a comparison for $H = 0$[6] as is clear, $|\bar{m}_{mix,M}|$ displays similar profiles for various values of M, with a sharp valley concentrated at the value of λ yielding maximal instability for the related configuration. Also, as M increases, the minimum is located at a value that converges to a finite value of λ, and the width of the valley gets larger. Remarkably, the valley related to $M = 1$ and $M \geq 3$ are well-separated, so that the minimum of the envelop $\max_{M\geq 3, M\text{odd}}(\bar{m}^a_{mix,M})$ falls in a region where the stability of the patterns is very high. This behavior also holds for larger L. Thus, we can define an optimal value of the anti-imitative interaction strength λ_{opt} corresponding to the minimum value of the envelop $\max_{M\geq 3, M\text{odd}} |\bar{m}^a_{mix,M}|$, where any spurious mixture is largely unstable, while the pure patterns are indeed stable fixed points. The dependence on L of this optimal value is reported in the right plot of Fig. 1, and is compatible (at $\alpha = 0$) with a scaling law $\lambda_{opt} \approx a(\beta)L^{-b(\beta)}$.

We now inspect the 1-step magnetization $\bar{m}^a_\mu$ and, in Fig. 2, we collect the theoretical predictions starting from an initial condition as given by Thm. 1, point ii). Specifically, the plots are obtained setting $D = L - 1$, meaning that only a single module is set on the mixture state $\boldsymbol{\sigma}_M$, while the remaining are already aligned with one of the patterns making up the mixture: this is the simplest scenario for disentangling, where the objective is to reconstruct the remaining pattern in the mixture which is not used in the initial condition (in Sec. 6 we will numerically address more challenging scenarios). Further, we differentiate if the target pattern $\boldsymbol{\xi}^\mu$ belongs to I, namely, one of the modules is fixed on $\boldsymbol{\xi}^\mu$ (left plot), or whether $\boldsymbol{\xi}^\mu \in \mathcal{S}_M \backslash I$, that is, no layer is prepared in $\boldsymbol{\xi}^\mu$ (right plot). In the left panel, we see that, for low λ, the module initialized in the mixture state $\boldsymbol{\sigma}_M$ remains trapped there (which is confirmed with the fact that $m_I \approx C_M$), and therefore these spurious states are stable. Tuning λ results in a different scenario in which the correlation with the initial condition is progressively lost (and in particular $m_I = 0$ around $\lambda = 1$). In contrast, as shown on the right, the patterns

[5] Despite the fact that, for $M = 1$, the subscript mix is not meaningful, we keep it to provide a consistent notation within the Section.

[6] In particular, we consider the modulus of the stability, since the transformation $\boldsymbol{\sigma}^a \to -\boldsymbol{\sigma}^a$ is a symmetry of the model as neural dynamics is performed in parallel. This could give at most 2-cycles as a result of parallel dynamics; this happens when the 1-step stability is very close to -1. Signatures of the emergence of 2-cycles are present in the left panel of Fig. 1; at $\lambda = \lambda_{opt}$: the non-monotonic behavior of the 1-step stabilities testifies a sign change for high λ; in this scenario, the dynamics of a given layer can be trapped in a 2-cycle of the form $\boldsymbol{\xi}^1 \to -\boldsymbol{\xi}^1 \to \boldsymbol{\xi}^1$, while correlation with the input condition is gradually downsized when the network is fed with spurious mixtures. For reconstruction purposes starting solely from the Hebbian matrix, however, this is not a problem, since patterns can be succesfully reconstructed apart for a global sign.

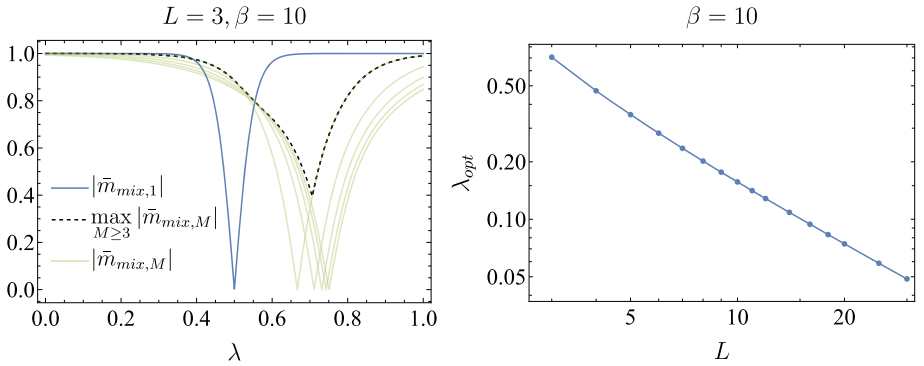

Fig. 1 Stability of pure patterns and mixed combinations $M \geq 3$. The left plot shows the 1-step stability for $L = 3$ and $\beta = 10$ as a function of λ for the stored patterns (blue curve) and mixed combinations (transparent green curves) with $M = 3, 5, \ldots, 15$; the black dashed line is the max-envelop of the latter curves. The right plot reports the value of λ yielding the minimum value of the envelop as a function of the number L of modules in the network. Results are reported for $L = 3$, $\beta = 10$ and $H = 0$

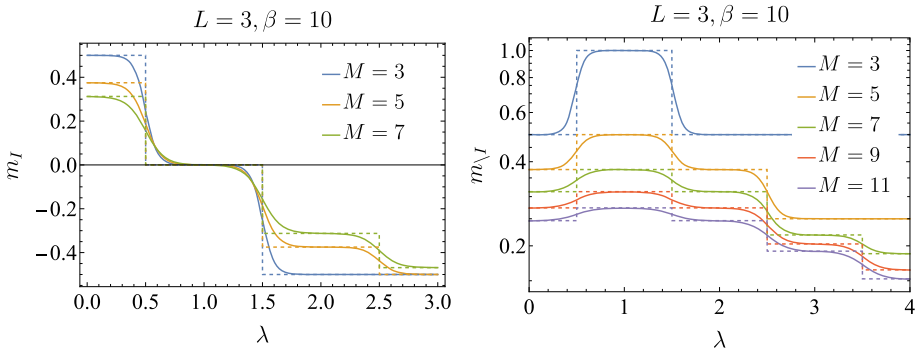

Fig. 2 1-step magnetizations of patterns. The left plot shows the 1-step magnetization of non-aligned layers ($b > D$) with patterns already retrieved in other layers ($\mu \in I$) for various values of M – denoted as m_I in the figure, while the right plot reports the 1-step magnetization of non-aligned layers with patterns $\mu \in \mathcal{S}_M \setminus I$. The theoretical results are reported for different numbers of patterns involved in $\boldsymbol{\sigma}_M$. In both cases, we fixed $H = 0$, $L = 3$, $\beta = 10$ (solid curves) and $\beta = +\infty$ (dashed lines), and $D = L - 1$

in $\mathcal{S}_M \setminus I$ are gradually reconstructed, and the system retrieves the remaining pattern in that module rather than aligning with other layers. Even though the magnetization $m_{\setminus I}$ strongly reduces as M increases, this remains a remarkable phenomenon because of the 1-step nature of the result: even if the single update leads to a limited reconstruction power, the system experiences a net drift towards disentangled configurations. Also, it is interesting to notice the behavior at zero temperature (dashed lines) and $M > 3$, evidencing different transitions associated to various steps in the associated magnetization. This underscores the rich retrieval phenomenology of modular Hebbian networks.

4 From Hebb's matrix to Kohonen's projector

As remarked in Sec. 2, for any input configuration $\boldsymbol{\sigma}(0)$, the output read in a given module is interpreted as a sound reconstruction of one of the hidden patterns as long as an acceptance

criterion (requiring a large projection on the span of the stored patterns) is satisfied. Before presenting this passage, some preliminary results are in order and discussed hereafter, while the related proofs are detailed in Appendix C.

We start recalling that, by basic algebra, the projector associated to the basis $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^K$ is given by

$$\boldsymbol{J}^K = \frac{1}{N}\boldsymbol{\Xi}\boldsymbol{C}^{-1}\boldsymbol{\Xi}^T,$$

where $\boldsymbol{C} = \frac{1}{N}\boldsymbol{\Xi}^T\boldsymbol{\Xi}$ is the correlation matrix. The factor N is needed since the $\boldsymbol{\xi}^\mu$ have norm $\sqrt{N}$. Following the analysis performed in [15], the iterative procedure to construct the projector $\boldsymbol{J}^K$ starting from the Hebbian coupling matrix $\boldsymbol{J}^H$ is

$$\boldsymbol{J}^K = \lim_{k\to\infty} \boldsymbol{J}(k), \tag{4.1}$$

$$\boldsymbol{J}(k+1) = \Big(1 + \frac{\varepsilon}{1+\varepsilon k}\Big)\boldsymbol{J}(k) - \frac{\varepsilon}{1+\varepsilon k}\boldsymbol{J}(k)^2, \tag{4.2}$$

$$\boldsymbol{J}(0) = \boldsymbol{J}^H, \tag{4.3}$$

with $k \in \mathbb{Z}_{\geq 0}$, $\varepsilon > 0$ being the update step, and $\|\cdot\|$ being the operator norm. We stress that constructing the projector through this algorithm requires neither explicit pattern knowledge nor the correlation matrix $\boldsymbol{C}$ as the Hebbian matrix alone is sufficient. Such an iterative algorithm can be further simplified by noticing that both the initial condition and the final coupling matrix are of the form $\frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}\boldsymbol{\Xi}^T$. Then, assuming that this holds at all $k \geq 0$, and introducing

$$q_k \doteq \frac{1+\varepsilon k}{1+\varepsilon(k+1)},$$
$$p_k \doteq \frac{\varepsilon}{1+\varepsilon(k+1)},$$

the algorithm (4.1–4.3) can be rewritten as $\boldsymbol{J}^K = \lim_{k\to\infty}\frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}(k)\boldsymbol{\Xi}^T$ where

$$\boldsymbol{G}(k+1) = q_k^{-1}\boldsymbol{G}(k) - p_k q_k^{-1}\boldsymbol{G}(k)\boldsymbol{C}\boldsymbol{G}(k), \tag{4.4}$$

$$\boldsymbol{G}(0) = \mathbf{1}_K. \tag{4.5}$$

We stress that the former formulation is used in practical scenario to compute numerically the projector matrix, while the latter is used to prove the following

Theorem 2 (Convergence) *If* $0 < \varepsilon < \varepsilon_c \doteq \frac{1}{\|\boldsymbol{C}\|-1}$ *and* $\boldsymbol{G}(k)$ *is defined according to Eqs. (4.4-4.5), then* $\frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}(k)\boldsymbol{\Xi}^T$ *converges to* $\boldsymbol{J}^K$ *in operator norm, namely*

$$\lim_{k\to\infty} \|\boldsymbol{J}^K - \frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}(k)\boldsymbol{\Xi}^T\| = 0.$$

We recall that, since the pattern correlation matrix $\boldsymbol{C}$ is a symmetric positive-definite (for $K \leq N$) matrix, its operator norm coincides with its largest eigenvalue which, in the thermodynamic limit $N \to \infty$ and $K/N \to \alpha \in [0, 1]$, is exactly $(1+\sqrt{\alpha})^2$ according to Marchenko-Pastur (MP) theorem. At finite size N, fluctuations around this value for the largest eigenvalue are of the order $\mathcal{O}(N^{-2/3})$ (see e.g., [20]), then for large matrices the MP reference is accurate enough. Clearly, if only the Hebbian coupling matrix is known, the number of hidden patterns can be deduced by computing $\frac{1}{N}\,\mathrm{Tr}\,\boldsymbol{J}^H$ due to the Rademacher

nature of the pattern entries. Thus, a good choice for the update step is

$$\varepsilon = \frac{\eta}{(1+\sqrt{\frac{1}{N}\operatorname{Tr} \boldsymbol{J}^H})^2 - 1}, \tag{4.6}$$

with $\eta \lesssim 1$. With respect to [15], here we obtain a more accurate error control in terms of the step size ε and the load $\alpha \geq 0$. This is the content of the following

Proposition 2 *For N and k large enough and $\alpha \in [0, 1)$, the approximation error*

$$\mathcal{R}_k(\varepsilon, \alpha) \doteq \| \frac{1}{N} \boldsymbol{\Xi}\, \boldsymbol{G}(k+1)\, \boldsymbol{\Xi}^T - \boldsymbol{J}^K \|,$$

of the algorithm (4.4-4.5) with the step size ε can be upper bounded as

$$\mathcal{R}_k(\varepsilon, \alpha) \leq \Big(\frac{1+\sqrt{\alpha}}{1-\sqrt{\alpha}}\Big)^4 \Big(1 + \frac{(1+\varepsilon)^2}{\varepsilon}\Big) \frac{\log k}{k} + \mathcal{O}(k^{-1}).$$

In particular, for $\alpha \in (0, 1)$ and with the update step ε as prescribed in Eq. (4.6), we have

$$\mathcal{R}_k(\eta, \alpha) \leq \Big(\frac{1+\sqrt{\alpha}}{1-\sqrt{\alpha}}\Big)^4 \Big(1 + \frac{(\alpha + 2\sqrt{\alpha} + \eta)^2}{\eta(\alpha + 2\sqrt{\alpha})}\Big) \frac{\log k}{k} + \mathcal{O}(k^{-1}).$$

Thus, at $\alpha = 0$, the approximation error is minimized (at leading order) for $\varepsilon = 1$, which leads to $\mathcal{R}_k(\eta, 0) \lesssim 5 \log k / k$.

5 Acceptance criterium

Having selected a family of configurations as candidate patterns through thermalization of the L-modular associative memory, we are left with checking their quality. Indeed, the disentanglement power of this network holds only probabilistically: even if we set the hyper-parameters in such a way that a disentangled configuration is the most energetically convenient, given the stochastic nature of the dynamics and possible finite-size effects, there still could be runs ending in spurious combinations, or some module retrieving patterns and others remaining stuck in mixed states. Thus, an acceptance criterion is needed. We recall that our theoretical results for the disentanglement capabilities of the L-directional associative memories are primarily derived in the low-storage regime ($\alpha = 0$), where equilibrium states display a macroscopic overlap with the stored vectors. Consequently, it is reasonable to restrict our attention to the family of spurious configurations $\{\boldsymbol{\sigma}_M\}_{M \in 2\mathbb{N}+1}$, and to focus on the quantity[7]

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} = \frac{1}{N} (\boldsymbol{\sigma}_M)^T \boldsymbol{J} \boldsymbol{\sigma}_M, \tag{5.1}$$

which is twice the inverse sign of the intensive energy of the state $\boldsymbol{\sigma}_M$, retaining only intra-module interactions (in the notation we made explicit the dependence on the coupling matrix $\boldsymbol{J}$ and the network size N). We choose this observable because , in the large N limit, the energy levels are well-separated and, in particular, their value decreases with M, hence allowing us to effectively discriminate pure patterns from spurious combinations. Although in this paper we primarily restrict to the low-storage case, the probabilistic properties of the random variables $\hat{\mathcal{E}}_{M,\boldsymbol{J},N}$ can be tackled analytically at any $\alpha \in [0, 1]$ as we do in this section. Our objective

[7] Since the states $\boldsymbol{\sigma}_M$ considered here are function of the (random) patterns alone, convergence theorems reported in this section should be read always with respect to the law of the patterns.

is therefore to determine a threshold $\tau_{opt}(\alpha, N)$ such that $\hat{\mathcal{E}}_{M,\boldsymbol{J},N} \geq \tau_{opt}(\alpha, N)$ successfully identifies patterns with high probability. The proofs of the results stated in this section are reported in Appendix D. The following is a general result (see e.g., [13]), which we now give in rigorous form[8]

Proposition 3 *In the thermodynamic limit $N \to \infty$ with $K/N = \alpha > 0$ and finite $M \in 2\mathbb{N}+1$, setting $\boldsymbol{J} = \boldsymbol{J}^H$:*

$$\lim_{N\to\infty} \hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N} = MC_M^2 + \alpha, \quad a.s. \tag{5.2}$$

Further

$$\sqrt{N}\left(\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N} - (MC_M^2 + \alpha)\right) \xrightarrow{d} S, \tag{5.3}$$

with $S \sim \mathcal{N}(0, \hat{\sigma}_M^2)$, and

$$\hat{\sigma}_M^2 = 4\, MC_M^2(1 - MC_M^2) + 2\alpha.$$

In Fig. 3, we report the comparison between the theoretical predictions and the numerical simulation, showing a perfect agreement between the two. From the expression of the asymptotic limit of $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N}$ (see Eq. 5.3), we see that the gap between the $M = 1$ peak and the other modes does not depend on α at large N; further, the related variances rapidly go to zero as N grows. Now, since the asymptotic distributions of $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N}$ are centered around values whose magnitude decreases with M, we can check that the configuration $\boldsymbol{\sigma}^*$ sampled from the L-directional associative memory has energy compatible with the distribution of $\hat{\mathcal{E}}_{M=1,\boldsymbol{J}^H,N}$ by estimating the overlap between the tails of the $M = 1$ distribution and its closest alternative, namely the $M = 3$ case. Based on this observation, we can define the decision boundary $\tau_{opt}^H(\alpha, N)$ as the energy value such that $P(M = 1|\tau_{opt}^H) = P(M = 3|\tau_{opt}^H)$, with $P(M|\hat{\mathcal{E}}^*)$ being the probability that the state $\boldsymbol{\sigma}^*$ with energy $\hat{\mathcal{E}}^*$ is a combination of M patterns. Exploiting the results in Prop. 3, the optimal threshold is given by the following

Lemma 1 *Given a configuration $\boldsymbol{\sigma}^* \in \{-1, +1\}^N$, for N sufficiently large, the optimal selection threshold is $\hat{\mathcal{E}}_{\boldsymbol{J}^H,N}(\boldsymbol{\sigma}^*) := \frac{1}{N}(\boldsymbol{\sigma}^*)^T \boldsymbol{J}^H \boldsymbol{\sigma}^* \gtrsim \tau_{opt}(\alpha, N)$, where*

$$\tau_{opt}^H(\alpha, N) \approx \frac{1}{6}(10\alpha - \sqrt{2\alpha(8\alpha+3)} + 6) + \frac{\sqrt{2\alpha(8\alpha+3)}\log\frac{\alpha}{\alpha+3/8}}{N}. \tag{5.4}$$

This acceptance criterion is effective at large N and relatively low α, as the probability of misclassifications (in particular false positive, namely mixed states which are accepted as reconstructed patterns) is very low. However, when dealing with a number of patterns comparable with N, the onset of glassy phenomena, which roughen the energy landscape, can impair this criterion. Further, in practical situations, this criterion may provide a non-negligible probability to accept spurious configurations as patterns. For instance, for $N = 500$ and $\alpha = 0.5$ (so that $K = 250$), one has $P(\hat{\mathcal{E}} \geq \tau_{opt}^H|M = 3) \approx 0.01$, meaning that 1 configuration out of 100 trials is a false positive. In problems requiring high confidence on pattern reconstruction (such as, for instance, matrix factorization) this would be a major limitation. For all these reasons, one should figure out alternative strategies. The projector-based criterion is indeed more robust, as the associated random variables $\hat{\mathcal{E}}_{M,\boldsymbol{J}^K,N}$ have tighter distributions; further, patterns are eigenvectors with maximal eigenvalues: any binary configuration has lower $\hat{\mathcal{E}}_{M,\boldsymbol{J}^K,N}$ compared to the patterns.

[8] Since even mixture of patterns are generally unstable, we only focus on odd M. Also, it is convenient to recall here that $C_M = \frac{1}{2^{M-1}}\binom{M-1}{\frac{1}{2}(M-1)}$, as it plays an important role in many of the following computations.:

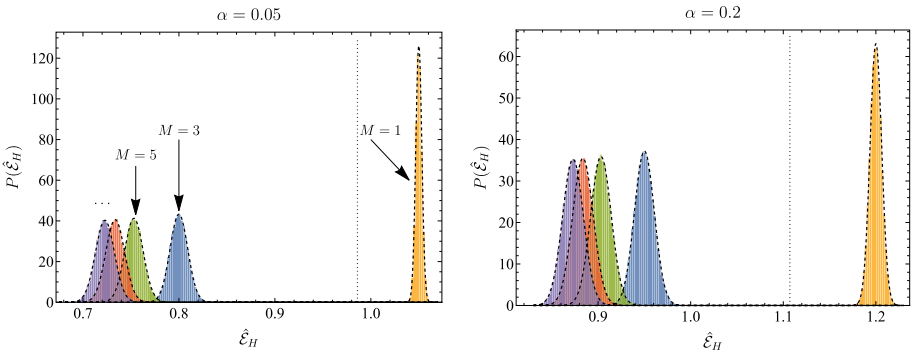

Fig. 3 Comparison between the empirical distributions and theoretical predictions of the quantity $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N}$ at large N. The two plots report the histograms of $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N} = \hat{\mathcal{E}}_H$ for different values of $M(= 1, 3, 5, 7, 9$ from the right to the left) and for two choices of the storage capacity, $\alpha = 0.05$ (left) and $\alpha = 0.2$ (right). The plot are realized by randomly selecting 1000 different realizations of mixed combinations for fixed M. The vertical dotted lines correspond to the optimal threshold per separating $M = 1$ and $M = 3$, see Lem. 1. Dashed lines are the theoretical prediction for the asymptotic behavior of $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N}$. The network size is fixed to $N = 10^4$

Proposition 4 *In the thermodynamic limit $N \to \infty$ with $K/N = \alpha > 0$ and finite $M \in 2\mathbb{N} + 1$, setting $\boldsymbol{J} = \boldsymbol{J}^K$:*

$$\lim_{N\to\infty} \hat{\mathcal{E}}_{M,\boldsymbol{J}^K,N} = \alpha + (1-\alpha)M\mathcal{C}_M^2, \quad a.s. \tag{5.5}$$

Controlling the fluctuations of the random variable $\hat{\mathcal{E}}$ requires precise estimates of the correlation between the vectors $\boldsymbol{\sigma}_M$ and the pattern correlation matrix $\boldsymbol{C}$, which enters in the energy in a non-trivial way. However, it is numerically found that the typical variance of the projector case is much lower than the Hebbian case. In particular, we numerically find that a good approximation for the $M = 3$ distribution is the asymptotic behavior $\sqrt{N}[\hat{\mathcal{E}}_{3,\boldsymbol{J}^K,N} - (\alpha + (1-\alpha)M\mathcal{C}_M^2)] \xrightarrow{d} \mathcal{N}(0, \hat{\sigma}_3^2)$ with

$$\hat{\sigma}_3^2 = 2\alpha(1-\alpha)(1-3\mathcal{C}_3^2)^2, \tag{5.6}$$

with $\mathcal{C}_3 = 1/2$, see Fig. 4, left plot. This should be compared to the case $M = 3$ for the Hebbian setting, that is, $\hat{\sigma}_3 = 4 \cdot 3\mathcal{C}_3^2(1-3\mathcal{C}_3^2) + 2\alpha$. It is clear that, in the projector case, the α enters through a multiplicative factor, rather than an additive constant as in the Hebbian case. Thus, the variance at large N is far lower in the former case. In particular, for $N = 100$, one finds that $\hat{\sigma}_3^2 \approx 10^{-5} \div 10^{-4}$ for the projector, while in the Hebbian case typically $\hat{\sigma}_3^2 \sim 10^{-2}$.

For this reason, it is reasonable to consider the variance negligible in the projector case, and select the optimal selection threshold as the intermediate value between the expectations of $\hat{\mathcal{E}}_{M=1,\boldsymbol{J}^K,N}$ and $\hat{\mathcal{E}}_{M=3,\boldsymbol{J}^K,N}$, namely

$$\tau_{opt}^K(\alpha) = \frac{(\alpha + (1-\alpha)M\mathcal{C}_M^2)|_{M=1} + (\alpha + (1-\alpha)M\mathcal{C}_M^2)|_{M=3}}{2} = \frac{7+\alpha}{8}. \tag{5.7}$$

With this selection threshold, one can thus compare the probability of false negatives in the two settings. Specifically, one computes the probability that the energy of $M = 3$ mixed states $\hat{\mathcal{E}}_{M=3}$ is higher than the selection threshold. The comparison between the Hebbian

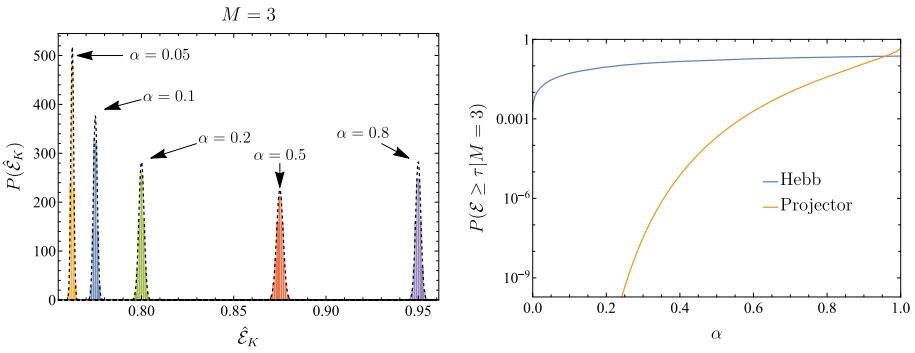

Fig. 4 Energy distributions in the projector case and false positive probability. The left histogram shows the distribution of the quantity $\hat{\mathcal{E}}_{M,\boldsymbol{J}^K,N} = \hat{\mathcal{E}}_K$ for $N = 10^4$, $M = 3$ and different values of the storage α. The black dashed lines are Gaussian fits with mean value given by Eq. (5.5) and variance σ_3^2/N as in Eq. (5.6). On the right plot, we report the probability of false positive $P(\mathcal{E}_K \geq \tau | M = 3)$, that is, the probability that a mixed state $M = 3$ has energy higher than the selection threshold τ in both cases. In the Hebbian case we used $N = 100$; at higher N the two criteria have comparable effectiveness, with the projector one always performing better (except for α close to 1)

and the projector settings is reported in Fig. 4, right plot, showing that – for a large range of storage α – the criterion (5.7) leads to a probability of false positives which is several order of magnitudes lower than its Hebbian counterpart. Then, this criterion is in general more effective, and robust w.r.t. noise in the realization of the projector through algorithm (4.1-4.2), and should be preferred in practical situations.

6 Numerical checks

In this section, we aim to corroborate the theoretical predictions derived in the signal-to-noise analysis through two complementary numerical protocols. In all our simulations, the numerical procedure follows the same overall pipeline, as detailed in Appendix A.

1-step MC *versus* signal-to-noise analysis. We test whether the probabilistic, theoretical predictions, which are expected to hold after a single MC synchronous update, are supported by extensive numerical simulations. In particular, we compute both the stability of patterns ($M = 1$) and odd mixed states ($M \geq 3$), as well as the 1-step magnetizations with the patterns when preparing $L - D$ modules on the mixed state $\boldsymbol{\sigma}_M$ and the remaining D on different patterns which are involved in this combination. The results are reported in Fig. 5, where theoretical predictions given by Thm. 1 perfectly match the numerical results.

Beyond 1-step MC. We investigate whether the parameter region predicted by the signal-to-noise analysis – in which, after one dynamical step, the layer configurations move closer to a disentangling retrieval state – indeed leads, after full thermalization, to a stable pure configuration. More precisely, we verify whether the dynamics converges to a state in which each layer aligns with one of the patterns contained in the initial spurious input mixture fed to the network.[9] A summary of (both short- and long-term relaxations of the) MC Markov

[9] Specifically, we run dynamics Eq. 2.6 until different stopping times. The largest one ($t = 10^4$) is high enough to ensure a proper convergence to fixed point, which we checked numerically by looking at the stationarity of the mean magnetizations.

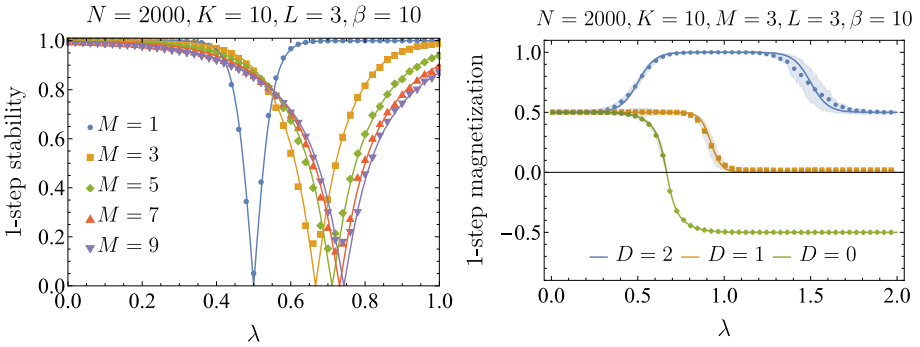

Fig. 5 1-step Monte Carlo simulations *versus* signal-to-noise. The plots show the numerical results for the 1-step MC updates for a network with $N = 2000$ spin per module, $L = 3$ layers, $K = 10$ stored patterns. The temperature is fixed to $\beta = 10$, the external field $H = 0$ and λ is tuned in the range $[0, 2]$. In the left plot, we report the modulus of the 1-step stability starting from a mixed combination $\boldsymbol{\sigma}_M$ with $M = 1, 3, 5, 7, 9$; the right plot instead reports the result for the 1-step magnetization where $L - D$ layers are prepared in $\boldsymbol{\sigma}_M$ and D to patterns involved in such combinations. In this case, data points refer to the magnetization of free modules w.r.t. pattern not presented at the machine in the initial condition. Data points are computed averaging over a batch of 100 different realizations of the patterns in the following way: for each run, we focus on a fixed layer, and compute mean the overlap with patterns in $\mathcal{S}_M \setminus I$. Solid curves are the 1-step theoretical predictions given in Thm. 1, while shaded regions represent the interquartile range for fixed λ

Chains used for disentangling is reported in Fig. 6, where we focus on the $D = 0$ case (that is, all layers are prepared in the configuration $\boldsymbol{\sigma}_M$). From the upper left plot, we see that the mean overlap of each layer with the patterns involved in the mixture has the same qualitative behavior of the 1-step case (blue points for the numerical simulations, black dashed curves for the theoretical predictions), the main difference consisting in the more pronounced region of instability of the mixture state. We stress that the lower *mean* magnetization of each layer with the patterns in the mixture does not necessarily imply a failure in disentanglement task. Indeed, if modules in the network specialize on different patterns, the empirical mean of the magnetization m_μ^a over $\mathcal{S}_M$ results in a lower mean overlap $\langle |\bar{m}| \rangle$. For instance, for $M = 3$ and all layers in $\boldsymbol{\sigma}_M$, the mean magnetization is $\approx \mathcal{C}_3 = 0.5$; conversely, if $(\boldsymbol{\sigma}^*)^a = \boldsymbol{\xi}^a$ with $a = 1, 2, 3$, we would have $m_1^1 = 1$ and $m_{2,3}^1 \approx 0$, then the mean overlap would be $\langle |\bar{m}| \rangle \approx 0.33$ (see Rem. 1). The fact that this behavior is compatible with retrieval is confirmed by upper right and lower left plots. In the former, for each layer we compute the highest magnetization with *all* the patterns (without discriminating if they are in $\mathcal{S}_M$ or not, e.g. a pure *retrieval* setting), and then average over different runs, also computing the retrieval frequency at the equilibrium (black dashed curve).[10] We stress that the magnetizations in the figures are not meant to be averaged also over the modules, as in each run we focus on a fixed module. Numerical evidences highlight that, in the region of the hyperparameters space the mean magnetization is lower w.r.t. the $\lambda = 0$ case, retrieval is possible with relatively high frequency.

The lower left plot is a consistency check of retrieval setting for the long-run case with the projector-based acceptance criterion presented in Sec. 5: a large fraction of final states of the network in our numerical experiments are compatible with the pattern reconstruction hypothesis. In the lower right plot we instead focus precisely on the disentangling scenario: maximum magnetizations (orange points) as well as the success frequency (dashed black curve) are computed only w.r.t. patterns involved in the combination $\boldsymbol{\sigma}_M$.

[10] Here, for simplicity we consider a success a final overlap higher than 0.8.

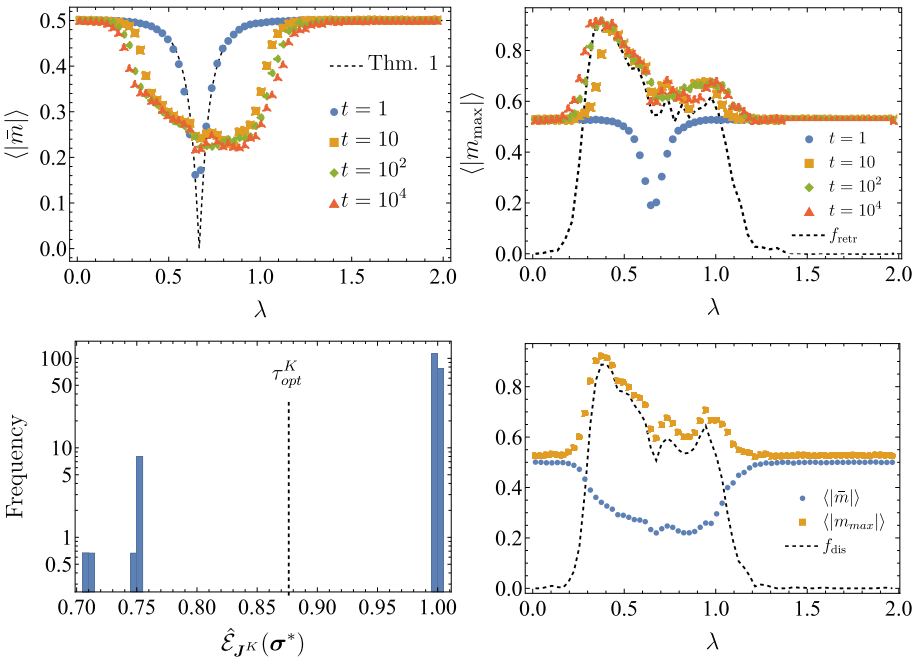

Fig. 6 Summary of Monte Carlo simulations. (Upper left plot) mean magnetizations of modules with patterns in $\mathcal{S}_M$ for different stopping times $t = 1, 10, 10^2, 10^4$. Black dashed line is the prediction of Thm. 1. (Upper right plot) highest magnetizations of layer with all the K stored patterns for different stopping times $t = 1, 10, 10^2, 10^4$. Black dashed line is the retrieval frequency computed upon thermalization. (Lower left plot) Comparison of the $\hat{\mathcal{E}}_{J^K}(\boldsymbol{\sigma}^*)$, where $\boldsymbol{\sigma}^*$ are the fixed points collected among the states of each module upon thermalization. Black dashed line is the comparison with projector-based acceptance criterion. (Lower right plot) Mean (blue points) and maximum (orange points) magnetizations of the layers with patterns involved in $\boldsymbol{\sigma}_M$. Black dashed curve is the disentanglement frequency. Empirical points are computed by averaging over a batch of 100 different realizations of the patterns. Network details are: $N = 1000$, $L = 3$, $K = 10$, $M = 3$, $\beta = 10$

Discussion of numerical results: to summarize, the numerical results reported here contribute to strengthening the evidence for the disentangling capabilities of modular Hebbian networks with anti-imitative interactions between layers. These findings provide insights into the information-processing principles of such networks when fed with partial information in the form of majority states arising from symmetric combinations of patterns, which in the standard Hopfield model correspond to stable fixed points at low thermal noise and therefore impair pattern reconstruction performance. Furthermore, we show that a 1-step analysis is sufficient to yield a qualitative picture of the long-term relaxation of the system and to identify the range of hyperparameter values for which disentanglement is most effective. In this scenario, the instability of mixed states plays a crucial role in enabling the recovery of their constituent hidden patterns. In particular, the *nearly* optimal strength of the anti-imitative interaction can be identified *a priori* through closed-form one-step expressions (Theorem 1), together with its scaling behavior with the number of network modules (see Fig. 1).

7 Conclusions and outlooks

In this work, we analyzed a modular neural network capable of disentangling and reconstructing multiple stored patterns in parallel. By combining imitative interactions within modules with anti-imitative couplings across modules, the network can suppress spurious mixture states while favoring the retrieval of relevant patterns. Remarkably, since the algorithm relies solely on second-order statistics of the data and on a set of pattern mixtures, it can also be employed as an inference procedure for hidden patterns, provided that their covariance matrix is known and that configurations exhibiting non-vanishing correlations with the patterns are available.

In the low-storage regime, we obtain an exact characterization of the one-step dynamics, which provides clear indications that appropriate choices of the hyperparameters stabilize the desired retrieval states while destabilizing spurious ones. Although our analysis is restricted to the early-time behavior of the dynamics, it already captures the key mechanisms underlying pattern disentanglement and parallel reconstruction, as corroborated by numerical simulations.

These results open the way to a more detailed investigation of the high-storage regime, as well as of the role played by noise and temperature. More broadly, the present framework suggests a viable route toward the design of neural architectures capable of scalable and robust parallel retrieval through structured interactions. In particular, the inclusion of higher-order correlations is expected to further enhance the performance of the system and to enable reliable reconstruction in applications such as matrix factorization. Natural evolutions of this work include applying our disentangling framework to empirical settings where direct estimation of the patterns is not straightforward and more general structure of the covariance matrix or beyond the pure Rademacher setting, as well as the analysis of robustness of our scheme – as well as the acceptance criterion – with respect to noise in the estimation of the Hebbian coupling matrix.

Acknowledgements EA and AF acknowledge support from PNRR MUR project PE0000013-FAIR, Sapienza University of Rome (RM120172B8066CB0, RM123188F3CD7763, RM12218169691087). The authors acknowledge the use of the Lagrange Multi-GPU Server at the Department of Mathematics, Sapienza University of Rome, for computational resources supporting this work.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

A Pseudo code for the reconstruction algorithm

B Signal-to-noise analysis: proofs

In this Section, we provide all the tools needed to state Thm. 1.

Algorithm 1: Pattern reconstruction pipeline

Parameters: *Number of layers L, number of neurons per layer N, number of parallel dynamic steps N_p; thermal noise T; acceptance threshold τ; overlap threshold δ*

Input: Hebbian coupling matrix $\{J_{ij}^H\}_{i,j=1,\cdots,N}$; input set $\{\boldsymbol{\sigma}_M^{(m)}\}_{m=1,\ldots,P}$.

Output: Accepted reconstructed patterns $\mathcal{Z} = (\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^{\hat{K}})$

Build $\widehat{\boldsymbol{J}}^K$ via unlearning (see Sec. 4)

Set

$$\varepsilon = \frac{\eta}{\left(1 + \sqrt{\mathrm{Tr}\, \boldsymbol{J}^H / N}\right)^2 - 1} \qquad \text{with} \qquad 0 < \eta < 1,$$

initialize $\boldsymbol{J}(0) = \boldsymbol{J}^H$, and iterate

$$\boldsymbol{J}(k+1) = \left(1 + \frac{\varepsilon}{1+\varepsilon k}\right) \boldsymbol{J}(k) - \left(\frac{\varepsilon}{1+\varepsilon k}\right) \boldsymbol{J}(k)^2,$$

until convergence; denote the limit by $\widehat{\boldsymbol{J}}^K$;

Collect candidate reconstructed patterns

$\mathcal{V} \leftarrow \emptyset$;

for $m \leftarrow 1$ **to** P **do**

  Initialize all layers/modules of the model on input $\boldsymbol{\sigma}_M^{(m)}$;

  $t \leftarrow 0$;

  while $t < N_p$ **do**

    Update the network state according to the neural dynamics Eq. (2.6) using noise T;

    $t \leftarrow t + 1$;

  Extract the final layer configuration(s) as reconstructed candidate(s) $\boldsymbol{\zeta}$;

  $\mathcal{V} \leftarrow \mathcal{V} \cup \{\boldsymbol{\zeta}\}$;

Discard duplicates using pairwise mutual overlap

$\widetilde{\mathcal{V}} \leftarrow \emptyset$;

foreach $\boldsymbol{\zeta} \in \mathcal{V}$ **do**

  `is_duplicate` $\leftarrow$ `false`;

  foreach $\boldsymbol{\zeta}' \in \widetilde{\mathcal{V}}$ **do**

    $q(\boldsymbol{\zeta}, \boldsymbol{\zeta}') \leftarrow \frac{1}{N} \sum_{i=1}^N \zeta_i \zeta_i'$;

    if $q(\boldsymbol{\zeta}, \boldsymbol{\zeta}') \geq \delta$ **then**

      `is_duplicate` $\leftarrow$ `true`;

      break;

  if $\neg$ *is_duplicate* **then**

    $\widetilde{\mathcal{V}} \leftarrow \widetilde{\mathcal{V}} \cup \{\boldsymbol{\zeta}\}$;

Accept candidates using spectral criterion

$\mathcal{Z} \leftarrow \emptyset$;

foreach $\boldsymbol{\zeta} \in \widetilde{\mathcal{V}}$ **do**

  $p(\boldsymbol{\zeta}) \leftarrow \frac{1}{N} \boldsymbol{\zeta}^T \widehat{\boldsymbol{J}}^K \boldsymbol{\zeta}$;

  if $p(\boldsymbol{\zeta}) \geq \tau$ **then**

    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{\boldsymbol{\zeta}\}$;

return $\mathcal{Z}$;

Proof of Prop. 1 Without loss of generality, we choose $\mathcal{S}_M = \{1, \ldots, M\}$. Let $\eta^\mu \underset{i.i.d.}{\sim} Rad(1)$ and consider the random variable $T = \eta^\nu \,\mathrm{sign}(\sum_{\mu=1}^M \eta^\mu)$. If $\nu > M$, due to statis-

tical independence of patterns we have $\mathbb{E}T = \mathbb{E}\eta^\nu \cdot \mathbb{E}\,\mathrm{sign}(\sum_{\mu=1}^{M}\eta^\mu) = 0$ and $\mathrm{Var}(T) = 1$. Conversely, if $\nu \leq M$, we can write $T = X\,\mathrm{sign}\,(X+S)$, with $X = \eta^\nu$ and $S = \sum_{\mu\neq\nu}^{M}\eta^\mu$. By law of total expectation $\mathbb{E}T = \mathbb{E}[\mathbb{E}[X\,\mathrm{sign}\,(X+S)|S]]$. Since M is odd, S is the sum of an even number of Rademacher variables; hence

$$S \in \{-M+1, -M+3, \ldots, M-1\},$$

so that $P(S = \pm 1) = 0$. Now, for fixed S, we have

$$\mathbb{E}[X\mathrm{sign}(X+S)|S] = \tfrac{1}{2}\big(\mathrm{sign}(S+1) - \mathrm{sign}(S-1)\big) = 1_{\{S=0\}},$$

so that $\mathbb{E}T = P(S=0)$, and the event is realized if and only if exactly half of the η^μ with $\mu = 1, \ldots, M$ are +1 and the other half is -1, which has probability

$$\mathcal{C}_M = \frac{1}{2^{M-1}}\binom{M-1}{\frac{1}{2}(M-1)}.$$

Along the same lines, $\mathrm{Var}(T) = 1 - \mathcal{C}_M^2$. Now, the magnetization of $\boldsymbol{\sigma}_M$ with the patterns can be written as $m_\mu^a(\boldsymbol{\sigma}_M) = \frac{1}{N}\sum_{i=1}^{N} T_i$ with T_i being of the form above and i.i.d. Then, by strong law of large numbers, $m_\mu^a(\boldsymbol{\sigma}_M) \xrightarrow{a.s.} \mathbb{E}T = \mathcal{C}_M$, while by virtue of the central limit theorem, $\sqrt{N}(m_\mu^a(\boldsymbol{\sigma}_M) - \mathcal{C}_M) \xrightarrow{d} \sqrt{1-\mathcal{C}_M^2}\,\mathcal{N}(0,1)$. This completes the proof. □

Proof of Thm. 1 Without loss of generality, we choose $\mathcal{S}_M = \{1, \ldots, M\}$. Set the initial condition of the network $\boldsymbol{\sigma}^a(0) = \boldsymbol{\sigma}_M$ for all $a = 1, \ldots, L$ and the external fields $h_i^a = \sigma_{M,i}$, and the magnetization of the mixture $\boldsymbol{\sigma}_M$ with the patterns is denoted as $m_\mu^a = m_\mu^a(\boldsymbol{\sigma}_M)$. In terms of these overlaps, the net field acting on each layer is rewritten as

$$\hat{h}_i^a(\boldsymbol{\sigma}_M) = \sum_{\mu=1}^{K}\big(m_\mu - \lambda(L-1)m_\mu\sum_{\nu=1}^{K}m_\nu^2\big)\xi_i^\mu + H\sigma_{M,i}.$$

i) The 1-step stability is

$$\bar{m}_{mix,M}^a = \mathbb{E}_{\boldsymbol{u}}[q(\boldsymbol{\sigma}^a(1), \boldsymbol{\sigma}_M)|\boldsymbol{\Xi}] = \frac{1}{N}\sum_{i=1}^{N}\tanh\Big(\frac{\hat{h}_i^a(\boldsymbol{\sigma}_M)\sigma_{M,i}}{T}\Big).$$

We can now rewrite $\hat{h}_i^a(\boldsymbol{\sigma}_M)\sigma_{M,i} = H + AR_i$ with

$$A \doteq 1 - \lambda(L-1)\sum_{\nu=1}^{K}m_\nu^2,$$

$$R_i \doteq \sum_{\mu=1}^{K}m_\mu\xi_i^\mu\sigma_{M,i},$$

For the A term, we recall that almost surely – by Prop. 1 – $m_\mu \to \mathcal{C}_M$ if $\mu \leq M$ and $m_\mu \to 0$ otherwise. Then, using the hypothesis that $K/N \to 0$ in the thermodynamic limit, we have $\sum_{\nu=1}^{K}m_\nu^2 = \sum_{\nu=1}^{M}m_\nu^2 + \sum_{\nu>M}^{K}m_\nu^2 \to M\mathcal{C}_M^2$ almost surely, so that $A \to A_\infty = 1 - \lambda(L-1)M\mathcal{C}_M^2$. Similarly, we decompose $R_i = \sum_{\mu=1}^{M}m_\mu\xi_i^\mu\sigma_{M,i} + \sum_{\mu>M}^{K}m_\mu\xi_i^\mu\sigma_{M,i}$. For the first contribution, we have:

$$\sum_{\mu=1}^{M}m_\mu\xi_i^\mu\sigma_{M,i} = \sum_{\mu=1}^{M}(m_\mu - \mathcal{C}_M)\xi_i^\mu\sigma_{M,i} + \mathcal{C}_M\sum_{\mu=1}^{M}\xi_i^\mu\sigma_{M,i}.$$

But

$$\sum_{\mu=1}^{M} \xi_i^\mu \sigma_{M,i} = \sigma_{M,i} \sum_{\mu=1}^{M} \xi_i^\mu = \text{sign}(\sum_{\mu=1}^{M} \xi_i^\mu) \cdot \sum_{\mu=1}^{M} \xi_i^\mu = \Big| \sum_{\mu=1}^{M} \xi_i^\mu \Big| \doteq |X_M^{(i)}|,$$

where we defined the random walk $X_M^{(i)} = \sum_{\mu=1}^{M} \xi_i^\mu$. Hence, using the fact that M is fixed in the thermodynamic limit and Prop. 1, we get

$$\sum_{\mu=1}^{M} m_\mu \xi_i^\mu \sigma_{M,i} \to \mathcal{C}_M |X_M^{(i)}|.$$

For the rest, we fix i and $\mu > M$ and adopt the decomposition

$$m_\mu = \frac{1}{N} \sum_{j \neq i} \xi_j^\mu \sigma_{M,j} + \frac{1}{N} \xi_i^\mu \sigma_{M,i} = m_\mu^{(i)} + \frac{1}{N} \xi_i^\mu \sigma_{M,i},$$

so that $m_\mu \xi_i^\mu \sigma_{M,i} = m_\mu^{(i)} \xi_i^\mu \sigma_{M,i} + \frac{1}{N} (\xi_i^\mu \sigma_{M,i})^2 = m_\mu^{(i)} \xi_i^\mu \sigma_{M,i} + \frac{1}{N}$. Summing over $\mu = M+1, \dots, K$:

$$\sum_{\mu=M+1}^{K} m_\mu \xi_i^\mu \sigma_{M,i} = \sum_{\mu=M+1}^{K} m_\mu^{(i)} \xi_i^\mu \sigma_{M,i} + \frac{K-M}{N}.$$

By construction, we get $\mathbb{E}(m_\mu^{(i)} \xi_i^\mu \sigma_{M,i}) = 0$, then by law of large numbers, using independence over μ and $K/N \to 0$ in the thermodynamic limit, we obtain

$$\sum_{\mu=M+1}^{K} m_\mu^{(i)} \xi_i^\mu \sigma_{M,i} \to 0.$$

By using independence between the signal block (depending on $\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^M$) and the noise block (depending on $\boldsymbol{\xi}^{M+1}, \dots, \boldsymbol{\xi}^K$), we join the results to get

$$R_i \to C_M |X_M^{(i)}|.$$

Putting everything together, we have $\hat{h}_i^a(\boldsymbol{\sigma}_M)\sigma_{M,i} \to H + A_\infty \mathcal{C}_M |X_M^{(i)}|$ almost surely. Fluctuations around this thermodynamic value are of order $\mathcal{O}(N^{-1/2})$. Under exchangeability of components, continuity and boundedness of the hyperbolic tangent and linearity of the expected value, the empirical average over i is equal to the expectation of a generic component, we can therefore write

$$\lim_{N \to \infty} \mathbb{E}_{\boldsymbol{\xi}} \bar{m}_{mix,M}^a = \mathbb{E}_{X_M} \tanh\Big(\frac{H + A_\infty \mathcal{C}_M |X_M|}{T} \Big).$$

ii) The proof for the 1-step magnetization works in an analogous way. The only further ingredient needed in this case is the following. Given a fixed positive odd integer M and $\varepsilon_l \sim_{i.i.d.} Rad(1)$ for $l \in \mathcal{S}_M$, let $X = \sum_{\ell=1}^{M} \varepsilon_\ell$, $\sigma = \text{sign}(X)$, and $I \subset \{1, \dots, M\}$ such that $0 \leq |I| \doteq D \leq M-1$, $S_I = \sum_{\ell \in I} \varepsilon_\ell$, and finally $U = |S|$, $V = \sigma S_I$. Then, for every measurable function $g(U, V)$ we get

$$\mathbb{E}(\sigma \varepsilon_\mu g(U,V)) = \begin{cases} \mathbb{E} \dfrac{V}{D} g(U,V) & \mu \in I \\ \\ \mathbb{E} \dfrac{U-V}{M-D} g(U,V) & \mu \notin I \end{cases}.$$

Then, using the asymptotic behavior of $\hat{h}_i^a(\boldsymbol{\sigma}(0))\sigma_{M,i}$ with $a > D$ and $\boldsymbol{\sigma}(0)$ as defined in the statement of the theorem, and boundedness and continuity of the hyperbolic tangent one gets again the thesis.

□

Remark 2 Similar computations can be performed in the more general case with $\alpha > 0$, or specializing the system at $T = 0$. For instance, for the 1-step stability one obtains:

- in the thermodynamic limit $N \to \infty$ with $K/N \to \alpha \in (0, 1)$ and $T > 0$, the 1-step stability of the state $\boldsymbol{\sigma}_M$ is

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\xi}} \bar{m}^a_{mix,M} = \mathbb{E}_{X_M,Z} \tanh\Big(\frac{H + A_\infty\big(C_M|X_M| + \alpha + Z\big)}{T}\Big);$$

- in the thermodynamic limit $N \to \infty$ with $K/N \to \alpha \in (0, 1)$ and $T = 0$, the 1-step stability of the state $\boldsymbol{\sigma}_M$ is

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\xi}} \bar{m}^a_{mix,M} = \mathbb{E}_{X_M,Z} \text{sign}[H + A_\infty(C_M\,|X_M| + \alpha + Z)],$$

Here, again $X_M = \sum_{l=1}^M \varepsilon_l$ with $\varepsilon_l \underset{i.i.d.}{\sim} Rad(1)$, and $Z \sim \mathcal{N}(0, \alpha)$.

C Convergence of dreaming algorithm: proofs

The algorithm (4.1–4.3) emerges as discrete version of the matrix ODE [15]

$$\frac{d\boldsymbol{J}(t)}{dt} = \frac{1}{1+t}(\boldsymbol{J}(t) - \boldsymbol{J}(t)^2),$$

with initial condition $\boldsymbol{J}(0) = \boldsymbol{J}^H$, whose unique solution is the dreaming kernel $\boldsymbol{J}(t) = \frac{1+t}{N}\boldsymbol{\Xi}(\mathbf{1}_K + t\boldsymbol{C})^{-1}\boldsymbol{\Xi}^T$, namely a Tichonov regularization of the Hebbian coupling matrix. Specifically, the projector $\boldsymbol{J}^K$ is the $t \to \infty$ limit of the above ODE. For $K \leq N$, the matrix $\boldsymbol{C}$ is SPD with large probability, thus the coupling matrix $\boldsymbol{J}(t)$ is well-defined for all $t \geq 0$, and $\|\boldsymbol{C}\| \geq 1$. By induction, it is possible to prove the following

Proposition 5 *For every $k \in \mathbb{N}$ and $i, j = 1, \ldots, N$:*

$$\frac{1}{N}\sum_{\mu,\nu=1}^{K} \xi_i^\mu \xi_j^\nu (C^k)_{\mu\nu} = \sum_{l=1}^{N} J(0)_{il}(J(0)^k)_{lj}.$$

Notice that $0 \leq t < K$ where $K = \min\{\|\boldsymbol{C}\|^{-1}, \|\boldsymbol{J}(0)\|^{-1}\}$ we have:

$$(\mathbf{1}_K + t\boldsymbol{C})^{-1} = \sum_{k=0}^{+\infty}(-t)^k\boldsymbol{C}^k,$$

therefore, thanks to the proposition given above and the Neumann series:

$$\boldsymbol{J}(t) = \frac{1+t}{N}\boldsymbol{\Xi}\sum_{k=0}^{\infty}(-t)^k\boldsymbol{C}^k\boldsymbol{\Xi}^T = \frac{1+t}{N}\sum_{k=0}^{\infty}(-t)^k\boldsymbol{\Xi}\boldsymbol{C}^k\boldsymbol{\Xi}^T =$$
$$= (1+t)\sum_{k=0}^{+\infty}(-t)^k\boldsymbol{J}(0)^{k+1} = (1+t)\boldsymbol{J}(0)(\mathbf{1}_N + t\boldsymbol{J}(0))^{-1} \doteq \boldsymbol{J}^*(t).$$

This argument, although flawless from a theoretical point of view, has a problem: the equality is verified only for certain values in an interval of the time variable t. On the other hand, since we are considering only nonnegative values of t and both $\boldsymbol{J}(0)$ and $\boldsymbol{C}$ are resp. positive semidefinite and definite matrices, $\mathbf{1}_K + t\boldsymbol{C}$ and $\mathbf{1}_N + t\boldsymbol{J}(0)$ are positive definite matrices, and therefore invertible for every $t \geq 0$. Indeed, let $\sigma(\boldsymbol{C}) = \{\lambda_1^{\boldsymbol{C}}, \ldots, \lambda_K^{\boldsymbol{C}}\}$ be the spectrum of $\boldsymbol{C}$. Since it is symmetric and positive definite, there exists an orthogonal matrix $\boldsymbol{O}$ such that $\boldsymbol{C} = \boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^T$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1^{\boldsymbol{C}}, \ldots, \lambda_K^{\boldsymbol{C}})$, consequently the polynomial $p(t) = \det(\mathbf{1}_K + t\boldsymbol{C}) = \prod_{\mu=1}^{K}(1 + t\lambda_\mu^{\boldsymbol{C}})$ never vanishes for $t \geq 0$, and the same argument holds for $\mathbf{1}_N + t\boldsymbol{J}(0)$. Therefore it is legitimate to expect that the two representations may be equal for any nonnegative t. Below, we will prove the equivalence of the two formulations.

Proposition 6

$$\lim_{t\to+\infty} \boldsymbol{J}^*(t) = \lim_{t\to+\infty} (1+t)\boldsymbol{J}(0)(\mathbf{1}_N + t\boldsymbol{J}(0))^{-1} = \boldsymbol{J}^K.$$

Proof Provided that $K \leq N$, we have that $\ker(\boldsymbol{J}(0)) = \mathcal{L}^\perp$, where $\mathcal{L}$ is the vector subspace of $\mathbb{R}^N$ generated by the patterns. Since $\boldsymbol{J}(0)$ is positive semidefinite and symmetric, by the spectral theorem there exists an orthogonal matrix $\boldsymbol{O}$ of order N such that $\boldsymbol{J}(0) = \boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^T$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ has on its diagonal the eigenvalues, with multiplicity, of $\boldsymbol{J}(0)$. In the coordinate system defined by $\boldsymbol{O}$, the last $N - K$ eigenvalues are relative to the eigenvectors belonging to $\mathcal{L}^\perp$, hence they are zero. Let now $\lambda \geq 0$ be any eigenvalue of $\boldsymbol{J}(0)$, then this will be eigenvalue also for $\boldsymbol{J}^*(t)$: indeed, calling $\boldsymbol{v}_\lambda$ be the associated eigenvector of $\boldsymbol{J}(0)$, then:

$$\boldsymbol{J}^*(t)v_\lambda = (1+t)\boldsymbol{J}(0)(\mathbf{1}_N + t\boldsymbol{J}(0))^{-1}\boldsymbol{v}_\lambda = \frac{(1+t)\lambda}{1+t\lambda}\boldsymbol{v}_\lambda.$$

Therefore the eigenvalues of $\boldsymbol{J}(t)$ are $\{\lambda_1(t), \ldots, \lambda_N(t)\}$ such that for every $i = 1, \ldots, n$ we have $\lambda_i(t) = \frac{(1+t)\lambda_i}{1+t\lambda_i}$ and, using the same matrix $\boldsymbol{O}$, we obtain the decomposition $\boldsymbol{J}^*(t) = \boldsymbol{O}\boldsymbol{\Sigma}(t)\boldsymbol{O}^T$ where $\boldsymbol{\Sigma}(t) = (1+t)\boldsymbol{\Sigma}(\mathbf{1}_N + t\boldsymbol{\Sigma})^{-1}$. Taking the limit we obtain:

$$\lim_{t\to+\infty} \lambda_i(t) = \begin{cases} 1 & \lambda_i > 0, \\ 0 & \lambda_i = 0. \end{cases}$$

Consequently

$$\boldsymbol{P} = \lim_{t\to+\infty} \boldsymbol{J}^*(t) = \lim_{t\to+\infty} \boldsymbol{O}\boldsymbol{\Sigma}(t)\boldsymbol{O}^T = \boldsymbol{O}\begin{pmatrix} \mathbf{1}_K & 0 \\ 0 & 0 \end{pmatrix}\boldsymbol{O}^T = \boldsymbol{O}\tilde{\boldsymbol{P}}\boldsymbol{O}^T.$$

It is easy to check that the matrix $\boldsymbol{P}$ is symmetric and idempotent, so $\boldsymbol{P}$ is an orthogonal projection. Since $\operatorname{Im}(\boldsymbol{P}) \oplus^\perp \ker(\boldsymbol{P}) = \operatorname{Im}(\boldsymbol{P}) \oplus^\perp \mathcal{L}^\perp$, we deduce that $\boldsymbol{P} = \boldsymbol{J}^K$. □

Lemma 2 *For every $t \geq 0$ we have*

$$\boldsymbol{J}(t) = \frac{1+t}{N}\boldsymbol{\Xi}(\mathbf{1}_K + t\boldsymbol{C})^{-1}\boldsymbol{\Xi}^T = (1+t)\boldsymbol{J}(0)(\mathbf{1}_N + t\boldsymbol{J}(0))^{-1}.$$

Proof We begin by noting that, differentiating with respect to t, both representations of $J(t)$ solve the following Cauchy problem:

$$\begin{cases} \frac{\mathrm{d}\boldsymbol{J}(t)}{\mathrm{d}t} = \frac{1}{1+t}(\boldsymbol{J}(t) - \boldsymbol{J}(t)^2) \quad \forall t \geq 0 \\ \boldsymbol{J}(0) = \boldsymbol{J}^H \end{cases}.$$

The claim follows by showing that the vector field $\Phi(t, \boldsymbol{J}) : A = (0, +\infty) \times \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ defined as $\Phi(t, \boldsymbol{J}) = \frac{1}{1+t}(\boldsymbol{J} - \boldsymbol{J}^2)$ satisfies the hypotheses of the Cauchy-Lipschitz existence and uniqueness theorem, hence it is continuous on A and locally Lipschitz. We immediately note that, being a composition of C^∞ functions, Φ is a C^∞ function, hence continuous on A. Let now $K \subseteq A$ be a compact subset with respect to the Euclidean topology and let t be fixed, and consider $(t, \boldsymbol{J}_1)$, $(t, \boldsymbol{J}_2) \in K$. By the triangle inequality for the operator norm we obtain:

$$\|\Phi(t, \boldsymbol{J}_1) - \Phi(t, \boldsymbol{J}_2)\| \leq \frac{1}{1+t}(\|\boldsymbol{J}_1 - \boldsymbol{J}_2\| + \|\boldsymbol{J}_1^2 - \boldsymbol{J}_2^2\|).$$

Since we can write $\boldsymbol{J}_1^2 - \boldsymbol{J}_2^2 = \boldsymbol{J}_1(\boldsymbol{J}_1 - \boldsymbol{J}_2) + (\boldsymbol{J}_1 - \boldsymbol{J}_2)\boldsymbol{J}_2$, using the submultiplicativity of the operator norm we deduce:

$$\|\boldsymbol{J}_1^2 - \boldsymbol{J}_2^2\| \leq (\|\boldsymbol{J}_1\| + \|\boldsymbol{J}_2\|)\,\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|\,.$$

We now use the following continuous map:

$$(t, \boldsymbol{J}) \in A \longmapsto \frac{1 + 2\,\|\boldsymbol{J}\|}{1+t}$$

which, restricted to K, attains a maximum thanks to the Weierstrass theorem. Denote the maximum by L_K; then local Lipschitz continuity follows from the arbitrariness of $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ for fixed t:

$$\|\Phi(t, \boldsymbol{J}_1) - \Phi(t, \boldsymbol{J}_2)\| \leq \frac{1 + \|\boldsymbol{J}_1\| + \|\boldsymbol{J}_2\|}{1+t}\,\|\boldsymbol{J}_1 - \boldsymbol{J}_2\| \leq L_K\,\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|$$

By the Cauchy-Lipschitz existence and uniqueness theorem, and since neither of the two solutions diverges as $t \to +\infty$, we can extend the solution continuously to all of $[0, +\infty)$ and thus deduce the claim, namely that the two representations are equal. □

Remark 3 Thanks to the previous theorem and Remark 6, it is interesting to note a particular behavior of the Dreaming Hopfield network: for $0 \leq t < +\infty$, the eigenvalues of $\boldsymbol{J}(t)$ are obtained from the eigenvalues of $\boldsymbol{J}(0) = \boldsymbol{J}^H$ through a continuous function and the eigenvectors remain the same. Indeed, if we denote by V the subspace generated by the eigenvectors of $\boldsymbol{J}(0)$ corresponding to positive eigenvalues, we have $V \subseteq \mathcal{L}$ and this remains true for every $0 \leq t < +\infty$. In the limit $t \to +\infty$ all the positive eigenvalues become 1 and we have $V = \mathcal{L}$.

Keeping in mind the formula for the derivative just derived, we can get into the heart of the algorithm by discretizing as in Eqs. (4.1–4.3) for some choices of $\varepsilon > 0$. Assuming that $\boldsymbol{J}(k) = \frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}(k)\boldsymbol{\Xi}^T$ for all $k \in \mathbb{Z}_{\geq 0}$ and $\boldsymbol{C} = \frac{1}{N}\boldsymbol{\Xi}^T\boldsymbol{\Xi}$, we can rewrite the algorithm as

$$\begin{cases} \boldsymbol{G}(k+1) = q_k^{-1}\boldsymbol{G}(k) - p_k q_k^{-1}\boldsymbol{G}(k)\boldsymbol{C}\boldsymbol{G}(k), \quad k \in \mathbb{Z}_{\geq 0}, \\ \\ \boldsymbol{G}(0) = \mathbf{1}_K, \end{cases}$$

which gives Eqs. (4.4-4.5) one identifying

$$q_k = \frac{1 + \varepsilon k}{1 + \varepsilon(k+1)}, \quad p_k = \frac{\varepsilon}{1 + \varepsilon(k+1)}.$$

We prove the convergence of the algorithm, namely that

$$\lim_{k\to+\infty} \boldsymbol{G}(k) = \boldsymbol{C}^{-1}.$$

Remark 4 As done in [15], it is easy to see that, for all $k \geq 0$:

1. $\boldsymbol{G}(k)$ are symmetric;
2. The matrices $\boldsymbol{C}$ and $\boldsymbol{G}(k)$ commute.

Further

Proposition 7 *If $\varepsilon > 0$ is such that $p_k < \|\boldsymbol{G}(k)\boldsymbol{C}\|^{-1}$ for every $k \geq 0$, then $\boldsymbol{X}(k) = \boldsymbol{C}\boldsymbol{G}(k)$ and $\boldsymbol{G}(k)$ are positive definite and invertible for every $k \geq 0$.*

Proof We proceed by induction. In the base case $k = 0$, this trivially follows by construction, since $\boldsymbol{G}(0) = \mathbf{1}_K$. Assume now it holds for some $k \geq 0$; we prove it for $k + 1$. Thanks to Remark 4, there exists a SPD matrix $\boldsymbol{G}^{1/2}(k)$ such that $\boldsymbol{G}^{1/2}(k)\boldsymbol{G}^{1/2}(k) = \boldsymbol{G}(k)$. Then

$$\boldsymbol{G}(k+1) = q_k^{-1}\boldsymbol{G}^{1/2}(k)\Big(\mathbf{1}_K - p_k\boldsymbol{G}^{1/2}(k)\boldsymbol{C}\boldsymbol{G}^{1/2}(k)\Big)\boldsymbol{G}^{1/2}(k).$$

Let $\boldsymbol{S}_k = p_k\boldsymbol{G}^{1/2}(k)\boldsymbol{C}\boldsymbol{G}^{1/2}(k)$. Since C is positive definite, for every $\mathbf{x} \in \mathbb{R}^K$ we have

$$\langle \boldsymbol{S}_k\mathbf{x}, \mathbf{x}\rangle = p_k \left\langle \boldsymbol{C}\boldsymbol{G}^{1/2}(k)\mathbf{x}, \boldsymbol{G}^{1/2}(k)\mathbf{x}\right\rangle \geq 0,$$

and it is equal to 0 if and only if $\mathbf{x} = 0$. Moreover, thanks to the hypothesis on p_k, we have $\|\boldsymbol{S}_k\| < 1$, hence $\mathbf{1}_K - \boldsymbol{S}_k$ is positive definite, which allows us to deduce that $\boldsymbol{G}(k+1)$ is positive definite. Since $\boldsymbol{X}(k+1)$ is given by the product of two positive definite matrices that commute, it is positive definite. Invertibility then comes trivially. □

The tools developed in the following lemma are crucial for our concerns.

Lemma 3 *For every $\varepsilon > 0$ such that $p_k < \|\boldsymbol{G}(k)\boldsymbol{C}\|^{-1}$ for every $k \geq 0$, there exists a sequence $\{c_k\}_{k\in\mathbb{N}}$ such that:*

1. *$\|\boldsymbol{G}(k)\boldsymbol{C}\| \leq c_k$ for every $k \geq 0$;*
2. *$\max_k c_k < +\infty$.*

Proof Recall that $\boldsymbol{X}(k) = \boldsymbol{G}(k)\boldsymbol{C}$ is, thanks to the hypothesis on ε, invertible for every $k \geq 0$. Consequently, we can write:

$$\boldsymbol{X}(k+1)^{-1} = q_k\left[\mathbf{1}_K - p_k\boldsymbol{X}(k)\right]^{-1}\boldsymbol{X}(k)^{-1}.$$

Using the Neumann series we obtain:

$$\left[\mathbf{1}_K - p_k\boldsymbol{X}(k)\right]^{-1}\boldsymbol{X}(k)^{-1} = \sum_{n=0}^{+\infty} p_k^n\boldsymbol{X}(k)^n\boldsymbol{X}(k)^{-1} = \sum_{n=0}^{+\infty} p_k^n\boldsymbol{X}(k)^{n-1},$$

which converges since $p_k < \|\boldsymbol{G}(k)\boldsymbol{C}\|^{-1}$ for all k. Consequently, we obtain:

$$\boldsymbol{X}(k+1)^{-1} = q_k\boldsymbol{X}(k)^{-1} + q_k p_k\mathbf{1}_K + q_k\sum_{n=2}^{+\infty} p_k^n\boldsymbol{X}(k)^{n-1}.$$

Applying this formula iteratively, and recalling that $\boldsymbol{X}(0) = \boldsymbol{C}$, we obtain:

$$\boldsymbol{X}(k)^{-1} = \frac{1}{1+\varepsilon k}\boldsymbol{C}^{-1} + N_{k-1}\mathbf{1}_K + \boldsymbol{R}(k-1),$$

where

$$\boldsymbol{R}(k-1) = \sum_{l=0}^{k-1}\frac{1+\varepsilon l}{1+\varepsilon k}\sum_{n=2}^{+\infty} p_l^n\boldsymbol{X}(l)^{n-1}, \quad N_k = \sum_{l=0}^{k} p_l\prod_{s=l}^{k} q_s.$$

It can be easily shown that

$$N_k = \frac{\varepsilon(1 + k + H_{\varepsilon^{-1}} - H_{1+k+\varepsilon^{-1}})}{1 + \varepsilon(k+1)}, \tag{C.1}$$

where H_z is the analytic continuation of the harmonic numbers, namely $H_z = \gamma + \psi(z + 1)$ with γ being the Euler-Mascheroni constant and $\psi(z)$ the digamma function. A direct computation shows that N_k is k-increasing; in particular, $N_k > N_0 = \frac{\varepsilon}{(1+\epsilon)^2}$; further $N_k \leq 1$ for all 1, and $\lim_{k\to\infty} N_k = 1$.

Now, since $\boldsymbol{X}(k)$ is symmetric and positive definite, it will have all positive eigenvalues; in particular the smallest eigenvalue of $\boldsymbol{X}(k)^{-1}$ will be the inverse of the maximum eigenvalue of $\boldsymbol{X}(k)$, hence it is $\|\boldsymbol{X}(k)\|^{-1}$. Denote by $\min\lambda(\boldsymbol{A})$ the smallest eigenvalue of the matrix $\boldsymbol{A}$; then

$$\begin{aligned}\|\boldsymbol{X}(k)\|^{-1} &= \min\lambda\left(\frac{1}{1+\varepsilon k}\boldsymbol{C}^{-1} + N_{k-1}\mathbf{1}_K + \boldsymbol{R}(k-1)\right)\\ &\geq \min\lambda\left(\frac{1}{1+\varepsilon k}\boldsymbol{C}^{-1} + N_{k-1}\mathbf{1}_K\right),\end{aligned}$$

since $\boldsymbol{R}(k)$ is positive definite as it is a sum of positive definite matrices. Using the same argument we obtain:

$$\min\lambda\left(\frac{1}{1+\varepsilon k}\boldsymbol{C}^{-1} + N_{k-1}\mathbf{1}_K\right) = \frac{1}{1+\varepsilon k}\|\boldsymbol{C}\|^{-1} + N_{k-1}.$$

Taking the inverse, we obtain the claim:

$$\|\boldsymbol{X}(k)\| \leq \left(\frac{1}{1+\varepsilon k}\|\boldsymbol{C}\|^{-1} + N_{k-1}\right)^{-1} = \frac{\|\boldsymbol{C}\|}{\frac{1}{1+\varepsilon k} + N_{k-1}\|\boldsymbol{C}\|} = c_k.$$

The fact that c_k is a bounded sequence follows from the fact that all quantities are positive and denominator is k-increasing. □

Remark 5 Concerning the $p_k\|\boldsymbol{X}(k)\| < 1$ constraint, we use the Lemma just proved we obtain:

$$p_k\|\boldsymbol{X}(k)\| \leq \frac{\varepsilon\|\boldsymbol{C}\|}{q_k^{-1} + N_{k-1}\|\boldsymbol{C}\|(1+\varepsilon(k+1))}.$$

The denominator is again a monotonically increasing function of k, hence it is sufficient to choose ε such that $p_0\|\boldsymbol{X}(0)\| < 1$ in order for the inequality to hold for all k. Since

$$p_0\|\boldsymbol{X}(0)\| = \frac{\varepsilon\|\boldsymbol{C}\|}{1+\varepsilon},$$

that is

$$\varepsilon < \varepsilon_c = \frac{1}{\|\boldsymbol{C}\| - 1}.$$

We are now ready to prove the Thm. 2.

Proof of Thm 2 As seen above, thanks to the hypothesis on ε, we can write:

$$\boldsymbol{X}(k)^{-1} = \frac{1}{1+\varepsilon k}\boldsymbol{C}^{-1} + N_{k-1}\mathbf{1}_K + \boldsymbol{R}(k-1).$$

Note that the first term has vanishing norm as $k \to +\infty$, whereas the second converges to $\mathbf{1}_K$. If we can show that $\boldsymbol{R}(k-1) \to 0$, we obtain the claim. Following [15], we can bound

$$\begin{aligned}\|\boldsymbol{R}(k-1)\| \leq & \frac{\varepsilon}{1+\varepsilon k} \sum_{l=0}^{k-1} \frac{1+\varepsilon l}{1+\varepsilon(l+1)} \frac{p_l \|X(l)\|}{1-p_l \|X(l)\|} = \\ = & \frac{\varepsilon}{1+\varepsilon k}\Big(\frac{1}{1+\varepsilon} \frac{p_0\|\boldsymbol{X}(0)\|}{1-p_0\|\boldsymbol{X}(0)\|} + \sum_{l=1}^{k-1} \frac{1+\varepsilon l}{1+\varepsilon(l+1)} \frac{p_l \|\boldsymbol{X}(l)\|}{1-p_l \|\boldsymbol{X}(l)\|}\Big),\end{aligned}$$

where in the last equality we used the formula for the convergent geometric series since $p_l \|X(l)\| < 1$. Now, by definition $p_0 = \frac{1}{1+\varepsilon}$ and $\|\boldsymbol{X}(0)\| = \|\boldsymbol{C}\|$. On the other hand

$$\frac{p_l \|\boldsymbol{X}(l)\|}{1-p_l \|\boldsymbol{X}(l)\|} \leq \frac{\varepsilon(1+\varepsilon l)\|\boldsymbol{C}\|}{[1+\varepsilon(l+1)]\big[1+\|\boldsymbol{C}\| N_{l-1}(1+\varepsilon l)\big]}.$$

For $l \geq 1$, $N_{l-1} \geq \frac{\varepsilon}{(1+\varepsilon)^2}$, so we can directly bound the latter equality as

$$\frac{p_l \|\boldsymbol{X}(l)\|}{1-p_l \|\boldsymbol{X}(l)\|} \leq \frac{(1+\epsilon)^2}{1+\varepsilon(l+1)} \frac{\frac{\varepsilon}{(1+\varepsilon)^2}(1+\varepsilon l)\|\boldsymbol{C}\|}{1+\frac{\varepsilon}{(1+\varepsilon)^2}(1+\varepsilon l)\|\boldsymbol{C}\|} \leq \frac{(1+\varepsilon)^2}{1+\varepsilon(l+1)}.$$

This implies that

$$\|\boldsymbol{R}(k-1)\| \leq \frac{\varepsilon}{1+\varepsilon k}\Big(\frac{1}{1+\varepsilon} \frac{\|\boldsymbol{C}\|}{1+\varepsilon-\|\boldsymbol{C}\|} + (1+\varepsilon)^2 \sum_{l=1}^{k-1} \frac{1+\varepsilon l}{(1+\varepsilon(l+1))^2}\Big). \tag{C.2}$$

The latter summation is again a combination of the digamma function and its first derivative, namely:

$$\begin{aligned}F_k(\varepsilon) \doteq & \sum_{l=1}^{k-1} \frac{1+\varepsilon l}{(1+\varepsilon(l+1))^2} \\ = & \frac{\psi(1+k+\varepsilon^{-1}) - \psi(2+\varepsilon^{-1}) + \psi'(1+k+\varepsilon^{-1}) - \psi'(2+\varepsilon^{-1})}{\varepsilon}.\end{aligned} \tag{C.3}$$

Remarkably, at large k, $F_k(\varepsilon)/k$ goes to zero as $\log k/k$, so that globally

$$\lim_{k\to\infty} \|\boldsymbol{R}(k)\| = 0.$$

Multiplying on the left with $\boldsymbol{C}$ and taking the norm, using triangular inequality and that operator norm is submultiplicative, we get

$$\begin{aligned}\|\boldsymbol{G}(k)^{-1} - \boldsymbol{C}\| = & \|\frac{1}{1+\varepsilon k}\mathbf{1}_K + (N_{k-1}-1)\boldsymbol{C} + \boldsymbol{C}\boldsymbol{R}(k-1)\| \leq \\ \leq & \frac{1}{1+\varepsilon k} + (1-N_{k-1})\|\boldsymbol{C}\| + \|\boldsymbol{C}\|\|\boldsymbol{R}(k-1)\| \underset{k\to\infty}{\to} 0,\end{aligned} \tag{C.4}$$

since $\boldsymbol{C}$ is finite and $\lim_{k\to\infty} N_k = 1$, thus concluding the proof. □

The proof of Prop. 2 works with the same tools developed above.

Proof of Prop. 2 Recall the bound in Eq. C.4, namely

$$\|\boldsymbol{G}(k)^{-1} - \boldsymbol{C}\| \leq \frac{1}{1+\varepsilon k} + (1-N_{k-1})\|\boldsymbol{C}\| + \|\boldsymbol{C}\|\|\boldsymbol{R}(k-1)\|.$$

Now, using the explicit expressions given by Eqs. (C.1) and (C.2), together with Eq. (C.3), for large k we have

$$1 - N_{k-1} = \frac{\log k}{k} + \mathcal{O}(k^{-1}),$$
$$\|\boldsymbol{R}(k-1)\| \leq \frac{(1+\varepsilon)^2}{\varepsilon}\frac{\log k}{k} + \mathcal{O}(k^{-1}).$$

This implies that

$$\|\boldsymbol{G}(k)^{-1} - \boldsymbol{C}\| \leq \|\boldsymbol{C}\|\Big(1 + \frac{(1+\varepsilon)^2}{\varepsilon}\Big)\frac{\log k}{k} + \mathcal{O}(k^{-1}) \doteq \|\boldsymbol{C}\|\delta_k(\varepsilon) + \mathcal{O}(k^{-1}).$$

Let us now call $\boldsymbol{\Delta}(k+1) = \boldsymbol{G}(k+1)^{-1} - \boldsymbol{C}$, which – for k large enough – satisfies $\|\boldsymbol{\Delta}(k+1)\| \leq \|\boldsymbol{C}\|\,\delta_k(\varepsilon) + \mathcal{O}(k^{-1})$. Reverting in favor of $\boldsymbol{G}(k+1)^{-1}$, we have $\boldsymbol{G}(k+1)^{-1} = \boldsymbol{C} + \boldsymbol{\Delta}(k+1) = \boldsymbol{C}(\boldsymbol{1}_K + \boldsymbol{C}^{-1}\boldsymbol{\Delta}(k+1))$. Defining $\boldsymbol{E}(k+1) = \boldsymbol{C}^{-1}\boldsymbol{\Delta}(k+1)$, we have $\boldsymbol{G}^{-1} = \boldsymbol{C}(\boldsymbol{1}_K + \boldsymbol{E}(k+1))$, and taking the inverse by Prop. 7:

$$\boldsymbol{G}(k+1) = (\boldsymbol{1}_K + \boldsymbol{E}(k+1))^{-1}\boldsymbol{C}^{-1}.$$

The norm of the perturbation is $\|\boldsymbol{E}\| \leq \left\|\boldsymbol{C}^{-1}\right\| \|\boldsymbol{\Delta}(k+1)\| \leq \left\|\boldsymbol{C}^{-1}\right\| \|\boldsymbol{C}\|\,\delta_k(\varepsilon) + \mathcal{O}(k^{-1})$, which will be less than 1 eventually (e.g. for all $k \geq \bar{k}$ for $\bar{k}$ large enough). Then:

$$\begin{aligned}\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1} &= (\boldsymbol{1}_K + \boldsymbol{E}(k+1))^{-1}\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1} = [(\boldsymbol{1}_K + \boldsymbol{E}(k+1))^{-1} - \boldsymbol{1}_K]\boldsymbol{C}^{-1} = \\ &= -(\boldsymbol{1}_K + \boldsymbol{E}(k+1)^{-1})\boldsymbol{E}(k+1)\boldsymbol{C}^{-1}.\end{aligned} \tag{C.5}$$

Thus $\left\|\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1}\right\| \leq \left\|\boldsymbol{1}_K + \boldsymbol{E}(k+1)^{-1}\right\| \|\boldsymbol{E}(k+1)\| \left\|\boldsymbol{C}^{-1}\right\|$.
Since $\left\|(\boldsymbol{1}_K + \boldsymbol{E}(k+1)^{-1})\right\| \leq \frac{1}{1-\|\boldsymbol{E}(k+1)\|}$ whether $\|\boldsymbol{E}(k+1)\| < 1$, we have

$$\left\|\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1}\right\| \leq \frac{\|\boldsymbol{E}(k+1)\|}{1 - \|\boldsymbol{E}(k+1)\|}\left\|\boldsymbol{C}^{-1}\right\| \leq \frac{\left\|\boldsymbol{C}^{-1}\right\|^2 \|\boldsymbol{C}\|\,\delta_k(\varepsilon)}{1 - \left\|\boldsymbol{C}^{-1}\right\| \|\boldsymbol{C}\|\,\delta_k(\varepsilon)} + \mathcal{O}(k^{-1}).$$

Since $\delta_k(\varepsilon) \ll 1$, and $\mathcal{O}(\log k/k)^2$ goes to zero faster than $\mathcal{O}(k^{-1})$, we can drop the non-leading corrections of the r.h.s. of previous equation, thus writing the bound as

$$\left\|\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1}\right\| \leq \left\|\boldsymbol{C}^{-1}\right\|^2 \|\boldsymbol{C}\|\,\delta_k(\varepsilon) + \mathcal{O}(k^{-1}).$$

Now

$$\begin{aligned}\mathcal{R}_k(\varepsilon,\alpha) &\leq \left\|\frac{1}{N}\boldsymbol{\Xi}\boldsymbol{G}(k+1)\boldsymbol{\Xi}^T - \boldsymbol{J}^K\right\| = \frac{1}{N}\left\|\boldsymbol{\Xi}(\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1})\boldsymbol{\Xi}^T\right\| \leq \\ &\leq \frac{1}{N}\|\boldsymbol{\Xi}\|\left\|\boldsymbol{G}(k+1) - \boldsymbol{C}^{-1}\right\|\left\|\boldsymbol{\Xi}^T\right\| \\ &\leq \frac{1}{N}\|\boldsymbol{\Xi}\|\left\|\boldsymbol{C}^{-1}\right\|^2\|\boldsymbol{C}\|\left\|\boldsymbol{\Xi}^T\right\|\delta_k(\varepsilon) + \mathcal{O}(k^{-1}).\end{aligned} \tag{C.6}$$

By definition of operator norm, we have $\|\boldsymbol{\Xi}\| = \sqrt{N\lambda_{max}(\boldsymbol{C})} = \sqrt{N}\,\|\boldsymbol{C}\|$, $\left\|\boldsymbol{\Xi}^T\right\| = \sqrt{N\lambda_{max}(\boldsymbol{J}^H)}$. Since non-zero eigenvalues of the correlation matrix matches those of the Hebbian kernel, we have $\sqrt{N\lambda_{max}(\boldsymbol{J}^H)} = \sqrt{N}\,\|\boldsymbol{C}\|$, thus

$$\mathcal{R}_k(\varepsilon,\alpha) \leq \left\|\boldsymbol{C}^{-1}\right\|^2 \|\boldsymbol{C}\|^2\,\delta_k(\varepsilon) + \mathcal{O}(k^{-1}).$$

At large N, $\|\boldsymbol{C}\| \approx (1+\sqrt{\alpha})^2$ and $\left\|\boldsymbol{C}^{-1}\right\| \approx 1/(1-\sqrt{\alpha})^2$, then

$$\mathcal{R}_k(\varepsilon,\alpha) \lesssim \frac{(1+\sqrt{\alpha})^2}{(1-\sqrt{\alpha})^2}\Big(1+\frac{(1+\varepsilon)^2}{\varepsilon}\Big)\frac{\log k}{k} + \mathcal{O}(k^{-1}).$$

The second claim trivially comes by setting $\varepsilon = \eta/(\|\boldsymbol{C}\|-1) \approx \eta/((1+\sqrt{\alpha})^2-1)$ for $\alpha > 0$. □

D The acceptance criterion: proofs

Proof of Prop. 3 The first point to notice is that, in the Hebbian case $\boldsymbol{J} = \boldsymbol{J}^H$:

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} = \frac{1}{N^2}(\boldsymbol{\sigma}_M)^T \boldsymbol{\Xi}\,\boldsymbol{\Xi}^T \boldsymbol{\sigma}_M = \sum_{\mu=1}^{K}\Big(\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu \operatorname{sign}\Big(\sum_{\nu\in\mathcal{S}_M}\xi_i^\nu\Big)\Big)^2.$$

Let us define the random variables $X_\mu = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu \operatorname{sign}\big(\sum_{\nu\in\mathcal{S}_M}\xi_i^\nu\big)$, which is essential the magnetization of the state $\boldsymbol{\sigma}_M$ with the patterns – see also Prop. 1 – and focus on the asymptotic behavior of the energy $\hat{\mathcal{E}}$. To this aim, we exploit the decomposition

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} = \sum_{\mu\in\mathcal{S}_M} X_\mu^2 + \sum_{\mu\notin\mathcal{S}_M} X_\mu^2.$$

The evaluation of the limiting behavior of first contribution is trivial, as $X_\mu \to \mathcal{C}_M\ \boldsymbol{\xi}$-almost surely, thus $\sum_{\mu\in\mathcal{S}_M} X_\mu^2 \to M\mathcal{C}_M^2$. For the second term, it is easy to see that

$$\sum_{\mu\notin\mathcal{S}_M} X_\mu^2 = \frac{\alpha}{K}\sum_{\mu}\tilde{X}_\mu^2,$$

with $\tilde{X}_\mu = \sqrt{N}X_\mu$, and thus reduces to the empirical mean of i.i.d. random variables with mean 1 (since $\mathbb{E}X_\mu = 0$, $\mathbb{E}\tilde{X}_\mu^2 = \operatorname{Var}\tilde{X}_\mu$). As a result, by strong law of large numbers one gets $\sum_{\mu\notin\mathcal{S}_M} X_\mu^2 \to \alpha$ a.s., and combining the two pieces one easily gets Eq. (5.2).

Controlling the fluctuations of $\hat{\mathcal{E}}_{M,\boldsymbol{J},N}$ requires a refined analysis. Let us start from the contribution of the patterns not involved in the mixture, namely $\sum_{\mu\notin\mathcal{S}_M} X_\mu^2$. This is a sum of squares of independent asymptotically Gaussian-distributed random variables (with variance $1/N$). Thus, in the large N limit, the quantity $\sum_{\mu\notin\mathcal{S}_M} X_\mu^2$ behaves as

$$\sum_{\mu\notin\mathcal{S}_M} X_\mu^2 \sim \frac{1}{N}\sum_{\mu\notin\mathcal{S}_M} z_\mu^2 = \frac{1}{N}\chi^2_{K-M},$$

with χ^2_{K-M} being a chi-squared random variable with $K-M$ degrees of freedom since $z_\mu \underset{i.i.d.}{\sim} \mathcal{N}(0,1)$. As for the first contribution, namely $\sum_{\mu\in\mathcal{S}_M} X_\mu^2$, we should also consider the fact that these are weakly correlated random variables. Indeed, it is easy to check that

$$\operatorname{Cov}[X_\mu, X_\nu] = \frac{\delta_{\mu\nu} - \mathcal{C}_M^2}{N} \doteq \Sigma_{\mu\nu}.$$

Then, the quantity $\sum_{\mu\in\mathcal{S}_M} X_\mu^2$ behaves asymptotically as the norm of a M-dimensional Gaussian vector with mean $\mathcal{C}_M\mathbf{1}$ and covariance matrix $\boldsymbol{\Sigma}$. The latter has only two possible eigenvalues, $\frac{1-M\mathcal{C}_M^2}{N}$ with multiplicity 1 (and associated non-normalized eigenvector $\boldsymbol{e} =$

$(1, \dots, 1))$ and $\frac{1-\mathcal{C}_M^2}{N}$ with multiplicity $M-1$ (and eigenvectors with total sum of entries 0. After some algebra, it can be shown that, for large N

$$\sum_{\mu \in \mathcal{S}_M} X_\mu^2 \sim M\mathcal{C}_M^2 + 2\sqrt{M}\mathcal{C}_M \sqrt{\frac{1-M\mathcal{C}_M^2}{N}} X + \mathcal{O}\Big(\frac{1}{N}\Big),$$

with $X \sim \mathcal{N}(0,1)$. Putting everything together, we thus have

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} \sim M\mathcal{C}_M^2 + \alpha + 2\sqrt{M}\mathcal{C}_M \sqrt{\frac{1-M\mathcal{C}_M^2}{N}} X + \frac{\sqrt{2(K-M)}}{N} Y + \mathcal{O}\Big(\frac{1}{N}\Big),$$

where $Y \sim \mathcal{N}(0,1)$. Recasting the finite values on the l.h.s., multiplying by $\sqrt{N}$ and taking the $N \to \infty$ limit (and provided the independence for all M and N of the separation between the signal and the noise, as well as of the limits), one finally gets convergence in distribution of the whole quantity $\hat{\mathcal{E}}_{M,\boldsymbol{J},N}$ to

$$2\sqrt{M}\mathcal{C}_M \sqrt{1-M\mathcal{C}_M^2} X + \sqrt{2\alpha} Y \sim \mathcal{N}(0, \sigma_M^2),$$

thus yielding Eq. (5.3).

□

Proof of Lem. 1 Let $\hat{\mathcal{E}}^*$ the observed energy of some configuration $\boldsymbol{\sigma}^*$ – coming for instance as fixed point of the neural dynamics in Eq. (2.6). Assuming it has probability distribution for some fixed M, the probability to observe it in the range $[\hat{\mathcal{E}}, \hat{\mathcal{E}} + d\hat{\mathcal{E}}]$ is

$$P(\hat{\mathcal{E}}^* \in [\hat{\mathcal{E}}, \hat{\mathcal{E}} + d\hat{\mathcal{E}}] | M) = \pi_M(\hat{\mathcal{E}}) d\hat{\mathcal{E}},$$

with π_M being the density of $\hat{\mathcal{E}}_{M,\boldsymbol{J}^H,N}$. By Bayes theorem, the probability that an observed value $\hat{\mathcal{E}}^*$ belongs to the distribution associated to π_M is

$$P(M|\hat{\mathcal{E}}^*) = \frac{\pi^0(M)\pi_M(\hat{\mathcal{E}})d\hat{\mathcal{E}}}{C},$$

with $\pi_M^0 = 1/2$ for $M = 1, 3$ being the prior on the population selection and C being a normalization factor. The decision boundary $\tau_{opt}^H(\alpha, N)$ will be therefore identified as the value of $\hat{\mathcal{E}}$ such that $P(M=1|\hat{\mathcal{E}}^*) = P(M=3|\hat{\mathcal{E}}^*)$: states with higher $\hat{\mathcal{E}}^*$ will be accepted as reconstructed patterns. For sufficiently large N, using Prop. 3, this criterion leads to the identification of the threshold where

$$\pi_{\mathcal{N}(M\mathcal{C}_M^2+\alpha,\sigma_M^2)}(\tau_{opt}^H)|_{M=1} = \pi_{\mathcal{N}(M\mathcal{C}_M^2+\alpha,\sigma_M^2)}(\tau_{opt}^H)|_{M=3}.$$

This constraint allows to express τ_{opt}^H as a function of α and N. Taking the $1/N$-leading contribution, one recovers Eq. (5.4). □

Proof of Prop. 4 Notice that, since $\boldsymbol{J} = \boldsymbol{J}^K$ is a projector, we have

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} = \frac{1}{N}(\boldsymbol{\sigma}_M)^T \boldsymbol{J}^K \boldsymbol{\sigma}_M = \frac{1}{N}(\boldsymbol{\sigma}_M)^T (\boldsymbol{J}^K)^2 \boldsymbol{\sigma}_M = \frac{1}{N}\|\boldsymbol{J}^K \boldsymbol{\sigma}_M\|_F^2,$$

with $\|\cdot\|_F^2$ being the Frobenius norm. Since $\|\boldsymbol{\sigma}_M\|_F^2 = N$ and $\boldsymbol{J}^K$ is symmetric, $\frac{1}{N}\|\boldsymbol{J}^K\boldsymbol{\sigma}_M\|_F^2 \leq \|\boldsymbol{J}^K\|^2 = 1$, since the projector has only 0 and 1 eigenvalues. Thus, $\hat{\mathcal{E}}_{M,\boldsymbol{J},N} \leq 1$, and in particular it saturates the bound if and only if $\boldsymbol{\sigma}_M$ is an eigenvector of $\boldsymbol{J}^K$; in other words, all the

patterns have $\hat{\mathcal{E}}_{M=1,\boldsymbol{J},N} = 1$ exactly. Thus, we only need to evaluate the norm of the projection of $\boldsymbol{\sigma}_M$ on the subspace spanned by the patterns. To do this, recall that $m_\mu(\boldsymbol{\sigma}_M) \overset{a.s.}{\to} \mathcal{C}_M$ if $\mu \in \mathcal{S}_M$, and $m_\mu \overset{a.s.}{\to} 0$ otherwise. Now, we express $\boldsymbol{\sigma}_M$ in direct sum of the projection along the subspace $V_M = \text{Span}(\boldsymbol{\xi}^\mu)_{\mu\in\mathcal{S}_M}$ and its orthogonal complement, namely:

$$\boldsymbol{\sigma}_M = \boldsymbol{J}_M^K \boldsymbol{\sigma}_M + \boldsymbol{\sigma}_M^\perp,$$

where clearly $\boldsymbol{J}_M^K$ is the projector built from the pattern belonging to $\mathcal{S}_M$. Clearly, $\boldsymbol{J}_M^K \boldsymbol{\sigma}_M^\perp = 0$. Let us also write $\boldsymbol{J}^K = \boldsymbol{J}_M^K + \boldsymbol{P}_1$, with $\boldsymbol{P}_1$ being the projector on the orthogonal complement of V_M. Then:

$$\hat{\mathcal{E}}_{M,\boldsymbol{J},N} = \frac{1}{N}(\boldsymbol{\sigma}_M)^T \boldsymbol{J}_M^K \boldsymbol{\sigma}_M + \frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp.$$

We now analyze the two contributions separately. For convenience of notation, we will call $\hat{\boldsymbol{\Xi}}_M$ the $N \times M$ matrix where columns are patterns in $\mathcal{S}_M$.

First, notice that

$$\frac{1}{N}(\boldsymbol{\sigma}_M)^T \boldsymbol{J}_M^K \boldsymbol{\sigma}_M = \frac{1}{N^2}(\boldsymbol{\sigma}_M)^T \hat{\boldsymbol{\Xi}}_M \hat{\boldsymbol{C}}_M^{-1} \hat{\boldsymbol{\Xi}}_M^T \boldsymbol{\sigma}_M,$$

where $\hat{\boldsymbol{C}}_M = \frac{1}{N}\hat{\boldsymbol{\Xi}}_M^T \hat{\boldsymbol{\Xi}}_M$ is the block of the pattern correlation matrix associated to $\mathcal{S}_M$. Since M is finite in the thermodynamic limit, the matrix $\hat{\boldsymbol{C}}_M$ has finite rank, and therefore by law of large numbers $\hat{\boldsymbol{C}}_M \overset{a.s.}{\to} \mathbf{1}_{\mathcal{S}_M}$. Since $\frac{1}{N}\hat{\boldsymbol{\Xi}}_M^T \boldsymbol{\sigma}_M \overset{a.s.}{\to} \mathcal{C}_M \boldsymbol{e}_{\mathcal{S}_M}$ with $\boldsymbol{e}_{\mathcal{S}_M}$ being the vector of ones for $\mu \in \mathcal{S}_M$ and 0 otherwise; this in turn implies $\frac{1}{N}(\boldsymbol{\sigma}_M)^T \boldsymbol{J}_M^K \boldsymbol{\sigma}_M \overset{a.s.}{\to} M\mathcal{C}_M^2$. Then, the contribution of the "spikes" is the same as in the Hebbian setting.

As for the second contribution, we fix the patterns $\boldsymbol{\xi}^\mu$, $\mu \in \mathcal{S}_M$ and consider the asymptotic behavior of $r_N = \frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp$ for large N. Thus, we consider the σ-algebra generated by $\mathcal{S}_M$, namely

$$\mathcal{F}_M = \{\boldsymbol{\xi}^\mu, \mu \notin \mathcal{S}_M | \boldsymbol{\xi}^\mu \text{ fixed for } \mu \in \mathcal{S}_M\}.$$

With respect to this algebra, the vector $\boldsymbol{\sigma}_M$ is deterministic and independent on $\mathcal{F}_M$. Since $\boldsymbol{P}_1$ is itself a projector, by standard arguments in high-dimensional probability, for some $C, c > 0$ and any $\epsilon > 0$, and denoted the interval $I_\epsilon = [(1-\epsilon)^2 \frac{K}{N}\|\boldsymbol{\sigma}_M^\perp\|_F^2, (1+\epsilon)^2 \frac{K}{N}\|\boldsymbol{\sigma}_M^\perp\|_F^2]$, we have $\mathbb{E}[\frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp | \mathcal{F}_M] = \frac{K-M}{N}\frac{\|\boldsymbol{\sigma}_M^\perp\|_F^2}{N}$, and for large N (and $K/N \to \alpha$):

$$P\Big(\frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp \notin I_\epsilon\Big) \le C \exp\big(-c\epsilon^2(K-M)\big),$$

see for example [27]. Since the r.h.s. of the inequality is summable, by Borel-Cantelli lemma we also have

$$\lim_{N\to\infty}\Big[\frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp - \frac{K-M}{N}\frac{\|\boldsymbol{\sigma}_M^\perp\|_F^2}{N}\Big] = 0,$$

almost surely. Then, it is sufficient to compute the norm of $\boldsymbol{\sigma}_M^\perp$. To do this, let us recall that $\|\boldsymbol{\sigma}_M\|_F^2 = N$; by the orthogonal decomposition of $\boldsymbol{\sigma}_M$, we also have

$$N = \|\boldsymbol{\sigma}_M\|_F^2 = \|\boldsymbol{J}^K \boldsymbol{\sigma}_M\|_F^2 + \|\boldsymbol{\sigma}_M^\perp\|_F^2.$$

Then, $\frac{\|\boldsymbol{\sigma}_M^\perp\|_F^2}{N} = 1 - \frac{1}{N}(\boldsymbol{\sigma}_M)^T \boldsymbol{J}^K \boldsymbol{\sigma}_M \overset{a.s.}{\to} 1 - M\mathcal{C}_M^2$. Then, for the second contribution we have

$$\lim_{N\to\infty} \frac{1}{N}(\boldsymbol{\sigma}_M^\perp)^T \boldsymbol{P}_1 \boldsymbol{\sigma}_M^\perp \overset{a.s.}{=} \alpha(1 - M\mathcal{C}_M^2).$$

Putting the contributions of the spikes and the bulk together yields Eq. (5.5). □

References

1. Agliari, E., Migliozzi, D., Tantari, D.: Non-convex multi-species hopfield models. J. Stat. Phys. **172**(5), 1247–1269 (2018)
2. Agliari, E., Aquaro, M., Barra, A., Fachechi, A., Marullo, C.: From Pavlov conditioning to Hebb learning. Neural Comput. **35**(5), 930–957 (2023)
3. Agliari, E., Alessandrelli, A., Barra, A., Centonze, M.S., Ricci-Tersenghi, F.: Generalized heteroassociative neural networks. J. Stat. Mech: Theory Exp. **2025**(1), 013302 (2025). https://doi.org/10.1088/1742-5468/ada918. (**ISSN 1742-5468**)
4. Agliari, E., Alessandrelli, A., Barra, A., Centonze, M.S., Ricci-Tersenghi, F.: Networks of neural networks: more is different (2025)
5. Agliari, E., Alessandrelli, A., Duarte Mourão, P., Fachechi, A.: Multi-channel pattern reconstruction through l-directional associative memories. In: New frontiers in associative memories (2025)
6. Agliari, E., Fachechi, A., Duarte Mourão, P.: The beneficial role of noises for disentanglement tasks in modular Hebbian networks. Physica A **682**, 131134 (2026). https://doi.org/10.1016/j.physa.2025.131134. (**ISSN 0378-4371**)
7. Alemanno, F., Camanzi, L., Manzan, G., Tantari, D.: Hopfield model with planted patterns: a teacher-student self-supervised learning model. Appl. Math. Comput. **458**, 128253 (2023)
8. Alessandrelli, A., Barra, A., Ladiana, A., Lepre, A., Ricci-Tersenghi, F.: Beyond disorder: Unveiling cooperativeness in multidirectional associative memories (2025)
9. Amari, S.I.: Neural theory of association and concept formation. Biol. Cybern. **26**, 175–185 (1977)
10. Ambrogioni, L.: In search of dispersed memories: generative diffusion models are associative memory networks. Entropy **26**(5), 381 (2024)
11. Barra, A., Genovese, G., Sollich, P., Tantari, D.: Phase transitions in restricted Boltzmann machines with generic priors. Phys. Rev. E **96**(4), 042156 (2017)
12. Camilli, F., Mézard, M.: Matrix factorization with neural networks. Phys. Rev. E **107**(6), 064308 (2023)
13. Coolen, A.C.C., Kuhn, R., Sollich, P.: Theory of neural information processing systems. Oxford University Press, UK (2005)
14. Decelle, A., Hwang, S., Rocchi, J., Tantari, D.: Inverse problems for structured datasets using parallel tap equations and restricted Boltzmann machines. Sci. Rep. **11**(1), 19990 (2021)
15. Fachechi, A., Agliari, E., Barra, A.: Dreaming neural networks: forgetting spurious memories and reinforcing pure ones. Neural Netw. **112**, 24–40 (2019)
16. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79**(8), 2554–2558 (1982)
17. Kalaj, S., Lauditi, C., Perugini, G., Lucibello, C., Malatesta, E.M., Negri, M.: Random Features Hopfield Networks generalize retrieval to previously unseen examples, (2024). arXiv:2407.05658 arXiv preprint
18. Krotov, D.: A new frontier for Hopfield networks. Nat. Rev. Phys. **5**, 366–367 (2023)
19. Little, W.A.: The existence of persistent states in the brain. Math. Biosci. **19**, 101–120 (1974)
20. Livan, G., Novaes, M., Vivo, P.: Introduction to random matrices. Springer, Germany (2018)
21. Manzan, G., Tantari, D.: The effect of priors on learning with restricted Boltzmann machines. Physica A **674**, 130766 (2025)
22. Marullo, C., Agliari, E.: Boltzmann machines as generalized Hopfield networks: a review on recent results and outlooks. Entropy **23**(1), 34 (2021)
23. Pastur, L.A., Figotin, A.L.: Theory of disordered spin systems. Theor. Math. Phys. **35**(2), 403–414 (1978)
24. Pourkamali, F., Barbier, J., Macris, N.: Matrix inference in growing rank regimes. IEEE Trans. Inf. Theory **70**(11), 8133–8163 (2024). https://doi.org/10.1109/TIT.2024.3422263
25. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlovic, M., Sandve, G.K., Greiff, V., Kreil, D.P., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: Hopfield networks is all you need. CoRR, arXiv:2008.02217 (2020)
26. Thériault, R., Tosello, F., Tantari, D.: Modeling structured data learning with restricted Boltzmann machines in the teacher-student setting. Neural Netw. **189**, 107542 (2025)
27. Vershynin, R.: High-dimensional probability: an introduction with applications in data science, 2nd edn. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2026)
28. Yampolskaya, M., Mehta, P.: Hopfield networks as models of emergent function in biology, (2025). arXiv:2506.13076