MULTI-CHANNEL PATTERN RECONSTRUCTION THROUGH L-DIRECTIONAL ASSOCIATIVE MEMORIES

Elena Agliari, Paulo Duarte Mourão & Alberto Fachechi
Department of Mathematics
Sapienza Università di Roma
Piazzale Aldo Moro 5, 00185 Rome, Italy
{elena.agliari,alberto.fachechi,paulo.duartemourao}@uniroma1.it

Andrea Alessandrelli
Department of Informatics
Università di Pisa
Largo Bruno Pontecorvo 3, 56127 Pisa, Italy
{andrea.alessandrelli}@phd.unipi.it

ABSTRACT

We consider L-directional associative memories, composed of L Hopfield networks, displaying imitative Hebbian intra-network interactions and anti-imitative Hebbian inter-network interactions, where couplings are built over a set of hidden binary patterns. We evaluate the model's performance in reconstructing the whole set of hidden binary patterns when provided with mixtures of noisy versions of these patterns. Our numerical results demonstrate the model's high effectiveness in the reconstruction task for structureless and structured datasets.

1 INTRODUCTION AND RELATED WORKS

The Hopfield model (Hopfield, 1982) is a cornerstone in the investigation of artificial neural networks, the main reason for such an importance lying in the crucial intuition that functionalities of artificial neural networks can be framed, from a physical point of view, as emerging collective properties much as like the thermodynamic properties of particle systems. Since its introduction, and especially after the solution by Amit, Gutfreund and Sompolinsky (Amit et al., 1987), the Hopfield model – and related models of associative memory – has attracted a continuously growing attention and today we have a clear picture of its working principles, including issues that may impair its pattern-reconstruction functionalities. Among these, spurious attractors have been examined in detail and several modifications have been proposed in order to reduce their attractiveness, retaining the pairwise interaction structure between the units (e.g., Dotsenko et al. (1991); Fachechi et al. (2019)) or extending the interaction order as in the dense associative memories (e.g., Krotov & Hopfield (2016)).

Remarkably, in recent years, pattern reconstruction and variations on the theme of the Hopfield model have gained broad significance and found applications in various fields. For instance, from a purely numerical perspective, they have been employed in matrix (and possibly tensor) factorization through decimation schemes (see, for example, (Camilli & Mézard, 2023) and references therein). Further, autonomous pattern reconstruction has today become one of the key aspects in modern Machine Learning theory, as it allows to shed light on the ability of neural networks to extract patterns from set of data and enable feature learning (Bengio et al., 2012; Aiudi et al., 2025), as well as investigating generalization in simplified settings (Negri et al., 2023; Kalaj et al., 2024; Agliari et al., 2024).

In this work, we explore the possibility to reconstruct binary hidden patterns by means of L-directional associative memories, assuming that the Hebb coupling matrix built on these patterns is given, along with additional information in terms of mixtures of corrupted versions of the same hidden patterns. We present numerical results across various settings, demonstrating strong performance for both structureless and structured datasets.

2 THE MODEL: L-DIRECTIONAL ASSOCIATIVE MEMORY

The L-directional generalization of the Hopfield model proposed in Agliari et al. (2025) is an energy-based model made up of an assembly of L Hopfield networks, each referred to as a layer, whose neuronal configurations are denoted as $\boldsymbol{\sigma}^a \in \{-1,+1\}^N$ with $a = 1, ..., L$. The model exhibits both intra- and inter-layer interactions. Specifically, given a realization of patterns $\boldsymbol{\xi}^\mu \in \{-1,+1\}^N$, with $\mu = 1, ..., K$, the energy function reads as $E = -N \sum_{a,b=1}^{L} g_{a,b} m_\mu^a m_\mu^b$, where $m_\mu^a = N^{-1} \sum_{i=1}^{N} \xi_i^\mu \sigma_i^a$ is the overlap between the a-th layer configuration and the μ-th pattern, while $g_{a,b}$ is chosen in such a way that $g_{a,a} = 1$ – hence reproducing the usual Hopfield energy function within each layer – and $g_{a,b} = -\lambda$ for $a \neq b$, with $\lambda \in \mathbb{R}_+$ being a tunable hyper-parameter – hence discouraging the retrieval of the same pattern by different layers. As shown in Agliari et al. (2025) focusing on the case $L = 3$, this network is able to disentangle mixtures of patterns, like the notorious spurious states $\boldsymbol{x} = \text{sgn}(\sum_{\nu=1}^{L} \boldsymbol{\xi}^\nu)$, in a wide region of the parameter space, that is, by supplying $\boldsymbol{x}$ as input configuration on each layer, the system can relax to the target configuration $(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, \boldsymbol{\sigma}^3) = (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \boldsymbol{\xi}^3)$, or any suitable permutation that ensures the retrieval of each single pattern in the original mixture [1]. However, it was also noticed that the energy function is invariant under a global spin-flip of all layers, but it is not invariant if layer configurations are reversed individually, namely $\boldsymbol{\sigma}^a \to -\boldsymbol{\sigma}^a$ for some $a = 1, \ldots, L$. As a consequence, beyond the target configuration $(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, ..., \boldsymbol{\sigma}^L) = (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^L)$, also configurations such as $(\boldsymbol{\xi}^1, \ldots, -\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^1)$ can exhibit strong attractive power for the neural dynamics, thus impairing the disentangling capabilities of the model. One way to prevent these undesired attractors and reduce their attraction basins, is to break the quadratic nature of the energy function by considering the square of inter-layer contributions in the energy function. Also, an external field $\boldsymbol{h}^a$ (modulated by a field strength H) driving the dynamics during evolution can be applied on each layer. Putting all pieces together and denoting with $\boldsymbol{\sigma}$ the overall configuration of the composite network, the resulting energy function reads:

$$E_{N,\boldsymbol{\xi}}(\boldsymbol{\sigma}) = -N \sum_{a=1}^{L} \sum_{\mu=1}^{K} (m_\mu^a)^2 + N\lambda \sum_{a \neq b=1}^{L} (\sum_{\mu=1}^{K} m_\mu^a m_\mu^b)^2 - H \sum_{a=1}^{L} \sum_{i=1}^{N} h_i^a \sigma_i^a. \tag{1}$$

This energy function results in a larger portion of the parameter space where the system successfully disentangle spurious states (Agliari et al., 2025). In the present paper, we show that this model can be employed even for more challenging tasks, as detailed in the following section. Before proceeding, we explicit the neuronal dynamics applied to the system: allowing for the presence of stochastic noise, tuned by the thermal parameter $\beta \in \mathbb{R}_+$, the neuronal configuration is synchronously updated as

$$\boldsymbol{\sigma}^a(t+1) = \text{sgn}[\tanh(\beta \tilde{\boldsymbol{h}}^a(t)) + \boldsymbol{u}^a(t)], \tag{2}$$

with t being the discrete time, $\boldsymbol{u}^a(t) \sim \mathcal{U}([-1,+1]^N)$ i.i.d. providing the source of noise, and $\tilde{h}_i^a$ being the net field acting on the spin i in the a-th layer. This can be expressed as

$$\tilde{\boldsymbol{h}}^a(t) = \boldsymbol{h}^{(a \to a)}(t) + \sum_{b \neq a} \boldsymbol{h}^{(b \to a)}(t) + H\boldsymbol{h}^a. \tag{3}$$

where, denoting with $\boldsymbol{J} = N^{-1}\boldsymbol{\xi}\boldsymbol{\xi}^T$ the Hebbian matrix, $\boldsymbol{h}^{(a \to a)}(t) = \boldsymbol{J} \cdot \boldsymbol{\sigma}^a(t)$ and $\boldsymbol{h}^{(b \to a)} = -\lambda \boldsymbol{h}^{(b \to b)}(t)(\boldsymbol{\sigma}^b(t) \cdot \boldsymbol{h}^{(a \to a)}(t))$ are, respectively, the intra- and inter-layer internal fields at time t, acting on the layer a.

3 TASKS AND RESULTS: MULTI-CHANNEL PATTERN RECONSTRUCTION

Given the ability of the model (1) to disentangle spurious states, it is worth investigating whether it can reconstruct patterns also from more general combinations. Specifically, we provide the model with a fixed number m of inputs of the form $\boldsymbol{x}^\gamma = \text{sgn}(\sum_{\mu=1}^{K} c_\mu^\gamma \boldsymbol{\xi}^\mu)$, with c_μ^γ for $\mu = 1, ..., K$

[1] The scheme here adopted can be interpreted as a parametric algorithm to achieve Independent Component Analysis (ICA) where data are available in a random feature setting (Negri et al., 2023; Kalaj et al., 2024). Notice, however, that the proposed scheme only gives the source vectors (the hidden patterns) involved in the mixture combinations but not the associated coefficients, whose determination requires additional procedures.

and $\gamma = 1, ..., m$ to be particularized according to the setting [2]. Next, we run the dynamics (2) and check whether the final configuration[3] $\bar{\boldsymbol{\sigma}} = \{\bar{\boldsymbol{\sigma}}^1, ..., \bar{\boldsymbol{\sigma}}^L\}$ has reached the target configuration $(\boldsymbol{\xi}^1, ..., \boldsymbol{\xi}^L)$, or any proper permutation. We emphasize that, in fact, there is no guarantee that the system relaxes to a disentangled representation of the inputs; thus, we should include specific quality checks for candidate reconstructed patterns. Remarkably, since the patterns $\{\boldsymbol{\xi}\}_{\mu=1}^K$ are not available, a direct comparison between $\bar{\boldsymbol{\sigma}}$ and $\boldsymbol{\xi}^\mu$ is not feasible and, as explained in the following, these checks leverage the algebraic properties of a suitable transformation of $\boldsymbol{J}$.

Let us start with the following setting: assume that the ground patterns are Rademacher, namely each entry is extracted as $\mathcal{P}(\xi_i^\mu = \pm 1) = 1/2$ for all $i = 1, \dots, N$ and $\mu = 1, \dots, K$, and hidden, while we have access to the mixtures $\boldsymbol{x}^\gamma$, $\gamma = 1, ..., m$ as defined above with $c_\mu^\gamma \sim \mathcal{N}(0, 1)$ i.i.d. for $\mu = 1, ..., K$ and $\gamma = 1, ..., m$. For each combination γ, we set $\boldsymbol{h}^a = \boldsymbol{x}^\gamma$ for all $a = 1, \dots, L$ and let the system evolve under neural dynamics (2), whence we collect the $L \cdot m$ final configurations $\{\bar{\boldsymbol{\sigma}}^l\}_{l=1}^{Lm}$ as candidate reconstructed pattern; clearly, if we want to recover the whole set of hidden patterns we need $Lm \geq K$. At this point, we notice that: $i)$ there could be duplicate candidates, *i.e.* configurations in $\{\bar{\boldsymbol{\sigma}}^l\}_{l=1}^{Lm}$ with high mutual overlap, and $ii)$ configurations stacked in some spurious state. To address point $i)$, we compute the mutual overlap $q_{lk} = N^{-1} \sum_{i=1}^N \bar{\sigma}_i^l \bar{\sigma}_i^k$, and discard duplicates if $q_{lk} > 0.5$ (a sufficiently high threshold for the random pattern setting). Regarding the point $ii)$, we recall that the true patterns $\boldsymbol{\xi}^\mu$ are eigenvectors (with a degenerate eigenvalue 1) of the pseudo-inverse coupling matrix $J_{ij}^K = N^{-1} \sum_{i,j}^N \sum_{\mu,\nu=1}^L \xi_i^\mu C_{\mu,\nu}^{-1} \xi_j^\nu$, with $C_{\mu,\nu} = N^{-1} \sum_{i=1}^N \xi_i^\mu \xi_i^\nu$ being the pattern correlation matrix (Kohonen & Ruohonen, 1973; Personnaz et al., 1985; Kanter & Sompolinsky, 1987). We can obtain the latter coupling matrix as fixed point of the iterative algorithm (Fachechi et al., 2019)

$$\boldsymbol{J}_{k+1} = \boldsymbol{J}_k + \frac{\epsilon}{1 + \epsilon k}(\boldsymbol{J}_k - \boldsymbol{J}_k^2),$$

with $\epsilon < (\|\boldsymbol{C}\| - 1)^{-1}$ being the unlearning strength and the initial condition being Hebb's matrix: $\boldsymbol{J}_0 = \boldsymbol{J}$. Thus, in order to solve $ii)$ and discard spurious states, we require $\bar{\boldsymbol{\sigma}}^l \boldsymbol{J}^K \bar{\boldsymbol{\sigma}}^l / N > 0.8$.

Out of the mL collected final configurations, we now select those that fulfill the last inequality and are distinct as prescribed in point $i)$. The items of this subset are denoted as $\boldsymbol{\xi}_R^\ell$, $\ell = 1, ..., K_R$ to emphasize that they provide a reconstruction of the hidden patterns; the cardinality K_R represents the number of the reconstructed hidden patterns. We stress that this outcome is reached by simply exploiting the knowledge of the Hebbian matrix and the set of m mixtures. Finally, to assess the quality of the reconstruction achieved by $\boldsymbol{\xi}_R^\ell$ we compute the quantity $m_\ell = \max_\nu [N^{-1} \boldsymbol{\xi}_R^\ell \cdot \boldsymbol{\xi}^\nu]$. Based on this procedure, we performed extensive Monte Carlo simulations and evaluated the expectation of K_R and the quality of reconstruction $N^{-1} \boldsymbol{\xi}_R \cdot \boldsymbol{\xi}$. The results of the algorithm described here are presented in Fig. 1. In the left plot, we report the average number K_R of reconstructed patterns as a function of the number of channels L for various values of K; clearly, the higher the complexity of the machine, the more effective the pattern extraction. In particular, as the number of patterns K to be extracted increases, the complexity required to successfully accomplish the task also rises. This is evident from the inset of the same plot, reporting the fractions of reconstructed patterns as a function of K for $L = 3, 10$. In any case, the individual quality of the reconstructed patterns is high and slightly improves by increasing L, as shown by the histograms on the right.

In the second setting we address a more realistic situation, where the accessible mixtures of hidden patterns are replaced by mixtures of noisy versions of the hidden patterns, referred to as examples. These are denoted as $\{\boldsymbol{\xi}^{\mu,A}\}_{\mu,A=1}^{K,M}$, with μ labeling the class and A distinguishing different items associated to the same pattern. Moreover, in the unsupervised scenario there is no *a priori* distinction of the examples in classes, that is, the label μ is unknown. To mimic this setting, we produce a synthetic dataset in the following way: first, extract the (hidden) patterns $\boldsymbol{\xi}^\mu$ as before, then we generate the examples by applying a multiplicative Bernoulli noise with quality parameter $r \in (0, 1)$,

[2] The application $\xi_i^\mu \to x_i^\gamma = \text{sgn}(\sum_{\mu=1}^K c_\mu^\gamma \xi_i^\mu)$ can be interpreted as a (non-linear) random mapping of the K-dimensional vectors $\boldsymbol{\xi}_i$ onto a space with dimension m, or, equivalently, as the response of a perceptron with K inputs and m outputs, with the spin index i labeling data points.

[3] This is reached after a time t long enough to ensure the stationarity of the temporal average of the overlaps m_μ^a over a sufficiently wide time window.

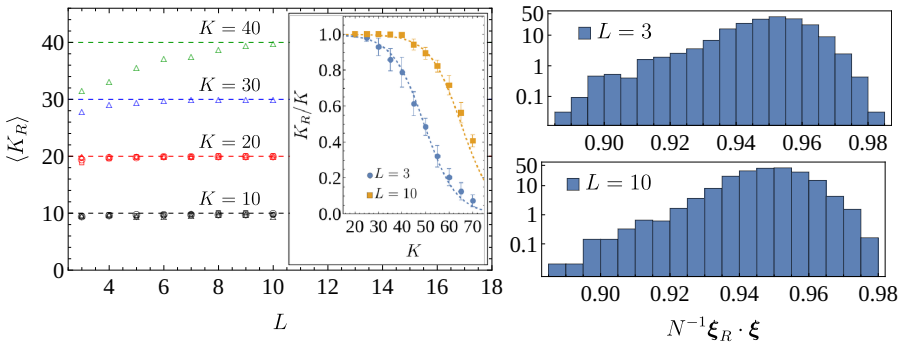

Figure 1: Summary of results for pattern reconstruction by general combinations $\text{sgn}(\boldsymbol{c}^\gamma \boldsymbol{\xi})$. In the left plot, we present the average number of reconstructed patterns as a function of L for various values of K. For $K = 10, 20$, we reported the results starting with $m = 10, 20, 30, 40, 50$ combinations shown by different symbols (as they lead to the same values of K_R symbols are collapsed), while for $K \geq 30$ only the results for $m = 50$ are shown. In the inset of the same plot, we reported the fraction of reconstructed patterns as a function of K for $L = 3$ (low-complexity machine) and $L = 10$ (high-complexity scenario). The dashed lines represents a fit of the form $K_R = K/[1 + \exp(\frac{1}{\kappa}(K_R - K_c))]$. In particular, for $L = 3$ we have $K_c \approx 50$, while for $L = 10$ the critical number of patterns is $K_c \approx 65$. The numerical results are averaged over 10 different realizations of the patterns $\boldsymbol{\xi}^\mu$ and the matrix $\boldsymbol{c}$. In the right plots, we present the aggregated results for the overlap between reconstructed patterns and the hidden ones: the histograms are realized by collecting all the results with fixed $L = 3$ and $L = 10$ (that is, for all the values of K and m). The network size is fixed to $N = 2000$, while $\beta = 2$, $\lambda = 0.2$, $H = 0.1$.

specifically $\xi_i^{\mu,A} = \chi_i^{\mu,A}\xi_i^\mu$, with $\mathcal{P}(\chi_i^{\mu,A} = \pm 1) = \frac{1\pm r}{2}$.[4] Taking a mini-batch of size n at random from the dataset, we can construct combinations of the form $x_i = \text{sgn}(\sum_{p=1}^n \xi_i^{\mu_p,A_p})$ mixing examples in different classes (thus, in this setting, the coefficients $c_{\mu,A}^\gamma$ are 1 if the corresponding item lies in the mini-batch, 0 otherwise). For large enough n, we would also have a large number of examples belonging to the same class, so that (denoting with n_μ the multinomial random variable representing the number of examples belonging to the class μ in a specific mini-batch) by virtue of the central limit theorem $\sum_{p=1}^{n_\mu} \xi^{\mu,A_p} = \xi_i^\mu \sum_{p=1}^{n_\mu} \chi_i^{\mu,A_p} \sim r^2 \xi_i^\mu (1 + \sqrt{\rho_\mu} z_i^\mu)$, with z_i^μ normally distributed and $\rho_\mu = (1 - r^2)/(n_\mu r^2)$. Thus, in this regime, we get $\text{sgn}(\sum_{p=1}^n \xi_i^{\mu_p,A_p}) \approx \text{sgn}(\sum_{\mu=1}^n \xi_i^\mu)$, resulting again in a spurious combination of patterns. We use configurations of the form $\boldsymbol{x}^\gamma$ (where now γ labels the m different realizations of the mini-batch) as input configurations for the model in (1) and reconstruct patterns with the same procedure as before. Our findings are reported in Fig. 2. Again, high-complexity machines have better extraction capabilities. Notably, in all situations the extraction procedure appears to be very robust w.r.t. to intrinsic noise in the dataset (even for high values of the mini-batch entropy ρ), as clearly shown by the weak dependence on r of the fraction K_R/K. In fact, as explained above, employing combinations of data points filters out the intrinsic noise, with these states being – at finite r – almost indistinguishable from usual spurious configurations of patterns. Therefore, the machine is expected to work nicely for the task under consideration.

As a last experiment, we test the procedure on a structured (but still simple) dataset. We take as patterns a synthetic realizations of the first 4 digits, we realize the dataset again with multiplicative noise, and consider vectors $x_i^\gamma = \text{sgn}(\sum_{p=1}^n \xi_i^{\mu_p,A_p})$ built by m mini-batches of size n. Then, we perform the pattern extraction procedure.[5] As we have shown in the previous experiment, the

[4] The role of the parameter r as the quality of the dataset is clear since, for $r = 1$, the examples are perfect copies of the hidden pattern, while for $r = 0$ examples are just random vectors carrying no information about the hidden patterns.

[5] Since, in the structured dataset, intrinsic features would have a higher mutual correlation w.r.t. the random case, we relax the eligibility condition of final configurations by considering duplicates two states with mutual overlap $q_{lk} > 0.9$. The "almost eigenvectors" criterion for the Kohonen kernel is left unchanged.

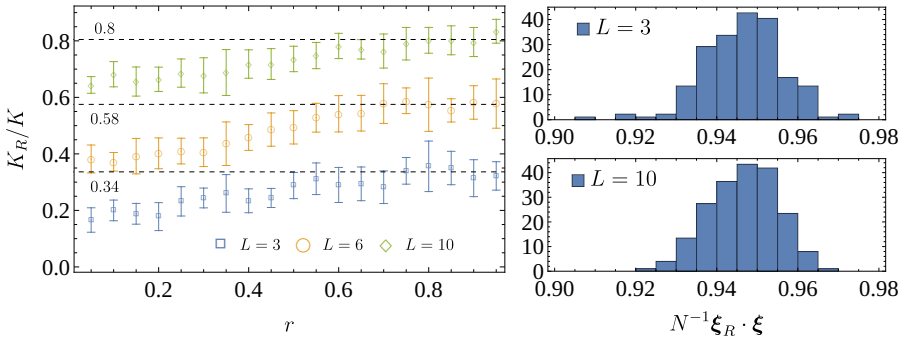

Figure 2: Summary of the results for pattern reconstruction with unsupervised combinations of examples. The left plot shows the dependence on the dataset quality r of the fraction of reconstructed patterns (here, $K = 50$) for different complexity of the machines: $L = 3, 6, 10$. The horizontal dashed lines stand for the asymptotic values of K_R/K at $r = 1$. The results are averaged over 10 different realizations of the patterns and the associated dataset. On the right side, we reported the histograms of the overlap of reconstructed patterns with the true ones. The combinatinations of examples are $m = 50$, the number of training examples (the mini-batches used to generate them) is fixed to $n = 25$, the number of examples per class is $M = 500$. The network size is $N = 2000$, while $\beta = 2$, $\lambda = 0.2$, $H = 0.1$.

pattern reconstruction procedure is robust against data noise. In the case under consideration, the dataset is indeed generated with very poor quality ($r = 0.2$). The final results are reported in Fig. 3. Even starting with visually unrecognizable samples, taking spurious combinations of examples filters out the noise, so that the system is able to effectively reconstruct the hidden patterns. The average quality of overlap between the reconstructed patterns and the true ones is very high, that is $\langle N^{-1}\boldsymbol{\xi}_R \cdot \boldsymbol{\xi} \rangle \approx 0.98$.

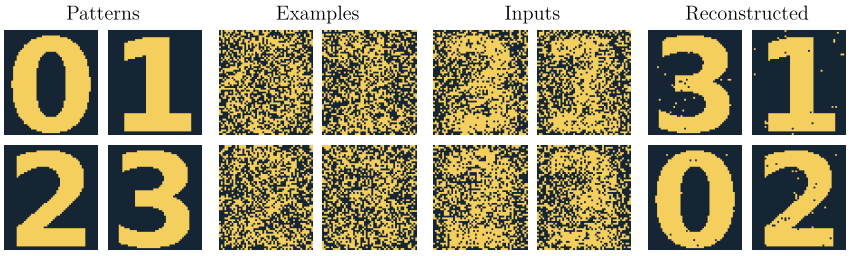

Figure 3: Summary of results for the pattern reconstruction by unsupervised structured examples. In the left block, we report the hidden patterns we want to reconstruct, starting from a very noisy dataset ($r = 0.2$) a sample of which is presented in the second block from the left. The number of examples per class is $M = 5000$, from which we generate $m = 50$ different mini-batches of size $n = 10$, which are used to generate the input configurations. In the right column, we reported the results of the pattern reconstruction. The network size is $N = 3016$ (images have size 58×52), the parameters are $\beta = 4$, $\lambda = 0.2$, $H = 0.05$ and $L = 4$.

4 Conclusions

We presented a procedure to reconstruct hidden patterns starting from partial information, namely Hebb's coupling matrix and additional information in terms of spurious combinations of the patterns. We extensively used the L-direction associative memories, allowing for a parallel retrieval of the patterns by disentangling such spurious states. We analyze the procedure in three settings, namely random patterns, synthetic and structured noisy datasets, always leading to high-quality reconstruction of the hidden features. We intend to deepen the results here reported in order to extend the possibility to known higher-order spatial moments of the patterns by suitably modifying the

energy function (for instance, adding dense contributions) as well as hyper-parameter fine-tuning (possibly by means of a statistical-mechanical approach), and applying the procedure to realistic datasets.

ACKNOWLEDGMENTS

EA and AF acknowledge financial support from PNRR MUR project PE0000013-FAIR and from Sapienza University of Rome (RM120172B8066CB0, AR2221815D7192C1, AR1221815EA97525).

REFERENCES

E Agliari, F Alemanno, M Aquaro, and A Fachechi. Regularization, early-stopping and dreaming: a hopfield-like setup to address generalization and overfitting. *Neural Networks*, 177:106389, 2024.

E Agliari, A Alessandrelli, A Barra, MS Centonze, and F Ricci-Tersenghi. Networks of neural networks: more is different. *arXiv preprint arXiv:2501.16789*, 2025.

R Aiudi, R Pacelli, P Baglioni, A Vezzani, R Burioni, and P Rotondo. Local kernel renormalization as a mechanism for feature learning in overparametrized convolutional neural networks. *Nature Communications*, 16(1):568, 2025.

DJ Amit, H Gutfreund, and H Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of physics*, 173(1):30–67, 1987.

Y Bengio, AC Courville, and P Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1(2665):2012, 2012.

F Camilli and M Mézard. Matrix factorization with neural networks. *Physical Review E*, 107(6): 064308, 2023.

VS Dotsenko, ND Yarunin, and EA Dorotheyev. Statistical mechanics of hopfield-like neural networks with modified interactions. *Journal of Physics A: Mathematical and General*, 24(10):2419, 1991.

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

JJ Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

S Kalaj, C Lauditi, G Perugini, C Lucibello, EM Malatesta, and M Negri. Random Features Hopfield Networks generalize retrieval to previously unseen examples. *arXiv preprint arXiv:2407.05658*, 2024.

I Kanter and Haim Sompolinsky. Associative recall of memory without errors. *Physical Review A*, 35(1):380, 1987.

T Kohonen and M Ruohonen. Representation of associated data by matrix operators. *IEEE Transactions on Computers*, 100(7):701–702, 1973.

D Krotov and JJ Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

M Negri, C Lauditi, G Perugini, C Lucibello, and E Malatesta. Storage and learning phase transitions in the random-features hopfield model. *Physical Review Letters*, 131(25):257301, 2023.

L Personnaz, I Guyon, and G Dreyfus. Information storage and retrieval in spin-glass like neural networks. *Journal de Physique Lettres*, 46(8):359–365, 1985.

A Details on numerical computations

Experiments are conducted by initializing each layer with a generic spurious observation $\boldsymbol{x}^\gamma$, and then evolving the system according to the dynamics described in Eq. 2, using a parallel update scheme (i.e., all neurons across the entire network are updated simultaneously). The dynamics are run for a sufficiently long time to ensure thermalization toward a fixed point. Unless otherwise specified, the total number of parallel updates is set to 5000. Numerical simulations were performed using TensorFlow 2.11 with CUDA Toolkit 11.7 and cuDNN 8.5, on an NVIDIA GeForce RTX 4070 Ti GPU.

B Sensitivity to hyperparameters on reconstruction performances

In this appendix, we explore how the model's reconstruction capabilities depend on the control parameters. We take a numerical approach, as a full theoretical understanding of the reconstruction regimes across the hyperparameter space would require a statistical mechanical analysis–this lies beyond the scope of the present work. For simplicity, we focus on the first setting, where the available information consists of spurious combinations of patterns, and the control parameters are β, λ, and H. To reduce the computational cost of exploring a three-dimensional hyperparameter space, we analyze two-dimensional sections by fixing one hyperparameter and varying the other two over a range of reasonable values. The results of this analysis are presented in Fig. 4. First, note that successful disentanglement of spurious pattern combinations requires the temperature to be not too high – thus avoiding an ergodic behavior –but still sufficiently high to allow the model to explore the energy minima landscape. We start by fixing $\beta = 2$ and vary λ and H. In the left plot, we see that, for the given level of thermal noise, the behavior of the reconstruction capabilities in λ is crucially dependent on H. In particular, for a low external field ($H = 0.1$), good reconstruction is achieved across a broad range of λ values ($\lambda = 0.05 \div 0.4$). This suggests that at $\beta = 2$ and $H = 0.1$ the model is relatively robust to variations in λ. A similar analysis can be carried out by fixing $\lambda = 0.2$ and varying H across different values of β. As β increases, the range of H values that yield good reconstruction performance becomes narrower and shifts toward lower values. This observation further supports the choice of $\beta = 2$ as a balanced setting for effective reconstruction. Finally, in the right plot, we perform a consistency check on the choice of temperature by fixing $\lambda = 0.2$ and varying both H and β. In all cases, the highest reconstruction performance is observed at $\beta = 2$. Our chosen setting — $\beta = 2$, $\lambda = 0.2$, and $H = 0.1$ — lies well within this favorable region. Naturally, a similar type of analysis can be carried out in the case of noisy realizations of structured patterns, which motivates the different parameter choices adopted in the third experiment.

C The acceptance criterion

As previously mentioned in the main text, the acceptance criterion for a reconstructed pattern involves a two-step verification process. First, we ensure that the final configurations of each layer exhibit low mutual overlap. This step eliminates potential duplicates in the final sample. Second, we verify that $\bar{\boldsymbol{\sigma}}^l \boldsymbol{J}^K \bar{\boldsymbol{\sigma}}^l / N > 0.8$, where $\boldsymbol{J}^K$ denotes the pseudo-inverse coupling matrix. This condition serves to filter out failed reconstructions resulting from relaxation towards spurious states. In this appendix, we further elaborate the effectiveness of the second step. Indeed, for any pattern $\boldsymbol{\xi}^\mu$, we have that

$$\sum_{j=1}^{N} J_{i,j}^K \xi_j^\mu = \sum_{j=1}^{N} \frac{1}{N} \sum_{\nu,\rho=1}^{K} \xi_i^\nu C_{\nu,\rho}^{-1} \xi_j^\rho \xi_j^\mu = \sum_{\nu,\rho=1}^{K} \xi_i^\nu C_{\nu,\rho}^{-1} C_{\rho,\mu} = \sum_{\nu=1}^{K} \xi_i^\nu \delta_{\nu\mu} = \xi_i^\mu.$$

Thus, the eigenspace associated with the eigenvalue 1 of the pseudo-inverse coupling matrix is K-dimensional and consists solely of linear combinations of the true patterns. Spurious states are thus excluded from this eigenspace due to the non-linearity of the sign function. Furthermore, by multiplying both sides of the equation by ξ_i^μ and summing over the index i, we have

$$\sum_{i,j=1}^{N} \xi_i^\mu J_{i,j}^K \xi_j^\mu = \sum_{i=1}^{N} (\xi_i^\mu)^2 = N.$$

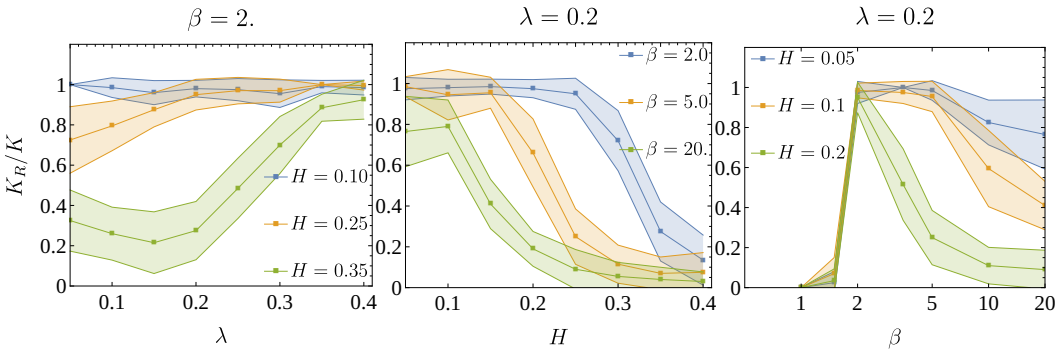

Figure 4: Sensitivity of the model's reconstruction capabilities to hyperparameters. In the three plots, we explore sections of the hyperparameter space by computing the fraction of reconstructed patterns, K_R/K, while fixing one hyperparameter and varying the other two. In the left plot, we fix $\beta = 2$ and analyze the dependence of K_R/K on λ for various values of H. In the center plot, we fix $\lambda = 0.2$ and examine how the reconstruction performance varies with H for different values of β. Finally, in the right plot, we report the dependence of K_R/K on β, fixing $\lambda = 0.2$ and varying the external field H. The shaded regions represent intervals of width two standard deviations centered around the mean. Results are averaged over 20 independent realizations of the patterns. The network size is $N = 1000$, the number of patterns is $K = 10$, and the number of layers is $L = 3$.

Therefore, the condition $\bar{\boldsymbol{\sigma}}^l \boldsymbol{J}^K \bar{\boldsymbol{\sigma}}^l / N = 1$ would ideally fulfill the desired acceptance criterion. However, in practice, this is rarely achieved due to two main reasons: i) the candidate configurations $\bar{\boldsymbol{\sigma}}^l$ are, at best, stochastic realizations of the underlying patterns, meaning that a finite fraction of bits may be misaligned with the corresponding true pattern; and ii) the pseudo-inverse matrix $\boldsymbol{J}^K$ is itself obtained through an iterative algorithm, which may introduce numerical approximations or deviations from the exact theoretical construction. Thus, we need to relax the acceptance criterion allowing for states with $\bar{\boldsymbol{\sigma}}^l \boldsymbol{J}^K \bar{\boldsymbol{\sigma}}^l / N$ above a sufficiently high threshold. Here, this threshold is fixed to 0.8. In Fig. 5 we give numerical results supporting the validity of our criterion.

In the left column, we display histograms of the overlap $\frac{1}{N}\bar{\boldsymbol{\sigma}}^l \cdot \boldsymbol{\xi}$ between the candidate configurations and the hidden patterns. Specifically, the blue histogram corresponds to configurations that satisfy the acceptance criterion, while the yellow histogram represents those that violate the condition $\bar{\boldsymbol{\sigma}}^l \boldsymbol{J}^K \bar{\boldsymbol{\sigma}}^l / N > 0.8$. As is clear, this criterion generally succeeds in filtering out states that result from the system thermalizing into spurious combinations of the patterns. For sufficiently low values of K, a fraction of the configurations $\bar{\boldsymbol{\sigma}}^l$ satisfy the acceptance criterion, and all of these exhibit a high overlap with the hidden patterns. In contrast, the rejected configurations typically show an overlap $\frac{1}{N}\bar{\boldsymbol{\sigma}}^l \cdot \boldsymbol{\xi} \leq 0.5$, consistently with the expectation that they correspond to spurious pattern combinations. In the inset, we also report a normalized confusion matrix supporting the validity of the criterion. The structure of this matrix is the following:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \frac{TP}{TP+FP} & \frac{FP}{TP+FP} \\ \frac{FN}{TN+FN} & \frac{TN}{TN+FN} \end{pmatrix},$$

where true positives (TP) refer to configurations $\bar{\boldsymbol{\sigma}}^l$ that satisfy the acceptance criterion and exhibit a high overlap with the patterns (e.g., $\frac{1}{N}\bar{\boldsymbol{\sigma}}^l \cdot \boldsymbol{\xi} \geq 0.8$). False positives (FP) are those configurations that are accepted by the criterion but have low correlation with the ground-truth patterns (i.e., $\frac{1}{N}\bar{\boldsymbol{\sigma}}^l \cdot \boldsymbol{\xi} < 0.8$). Conversely, true negatives (TN) are configurations rejected by the criterion that indeed show low overlap, while false negatives (FN) are those that are incorrectly rejected despite exhibiting high overlap with the patterns. Although these FN cases are discarded, they do not significantly affect the overall reconstruction performance of the model. Since the fraction of true positive and true negative states is close to 1, we conclude that the acceptance criterion effectively distinguishes between accurate reconstructions and spurious combinations of the hidden patterns. As expected, increasing K leads to a larger fraction of states failing the sanity check: in this regime, pattern retrieval becomes significantly more challenging, and the reconstruction process tends to break down. This behavior is illustrated in the right-hand plot, where we show the fraction of rejected configurations as a function of $\alpha = K/N$ for $L = 3, 5, 10$, along with the corresponding

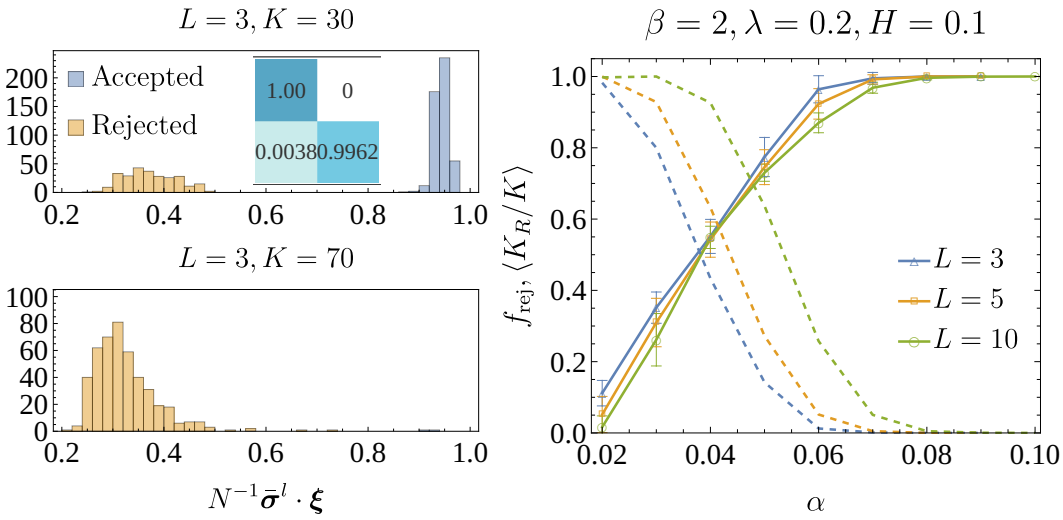

Figure 5: Effectiveness of the acceptance criterion. In the left column, we compare the fractions of the accepted final configurations (blue histogram) w.r.t. the discarded ones (yellow histogram) as a function of their overlap with the hidden patterns. For $L = 3$ and low $K = 30$ (upper left plot), the acceptance criterion is able to distinguish between reconstructed truths and their spurious combinations, and the effectiveness is high (see the confusion matrix in the inset plot). For higher values of K (upper right), the thermalization of the systems more likely ends up in spurious configurations, which are rejected in bulk, resulting in a loss of reconstruction power. In the right plot, we report the fraction of rejected configurations as a function of α for $L = 3$ (blue), 5 (yellow) and 10 (green). For the sake of completeness, in dashed lines we also reported the associated results for the average fraction of reconstructed patterns. The size of the network is $N = 1000$, the parameters are $\beta = 2$, $\lambda = 0.2$ and $H = 0.1$, the number of spurious observation is $m = 50$. Results are averaged over 20 different realizations of the hidden patterns.

average reconstruction performance K_R/K (shown as dashed lines). As the information load α increases, the likelihood that the system thermalizes into spurious states also grows, compromising the model's reconstruction accuracy. However, increasing the number of layers L improves the acceptance rate, thereby enhancing the ability to retrieve patterns even under higher storage demands. Investigating the optimal scaling relations between the hyperparameters, the number of layers, and the storage capacity is a crucial aspect of this framework. However, a thorough analysis of this problem within a statistical mechanical perspective is beyond the scope of the present work and will be addressed in future studies.

D FINITE-SIZE SCALING

As a final point, we examine the robustness of the model's reconstruction capabilities with respect to the individual layer size N. To ensure a fair comparison, networks of different sizes must operate under equivalent conditions. First, the number of stored patterns should scale with N, i.e., $K = \alpha N$. However, increasing K while keeping m fixed significantly reduces the probability of successfully reconstructing all patterns; in other words, also m should scale with N. To estimate this scaling, we considered a related problem. Suppose we have a collection of K objects, from which we uniformly sample a subset of L elements in each experiment (i.e., each object is selected with probability $1/K$). We repeat this experiment m times, replacing the extracted elements after each trial. Our goal is to compute the probability that all K patterns are observed at least once across the m trials. Consider a fixed element, say $\mu = 1$. The probability that it is not selected in a single trial is approximately $(1 - 1/K)^L \approx 1 - L/K$, assuming K is large. Therefore, the probability that this element is never observed over m independent repetitions is $(1 - L/K)^m$. From this, we can say that, for K large enough, the probability that at least one of the K elements is never observed across all trials is approximately $\approx K(1 - L/K)^m$. Since this is the complementary event to the one we

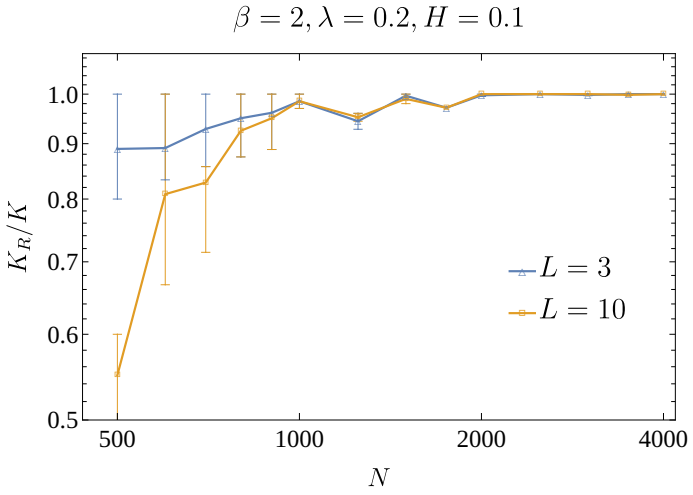

Figure 6: Finite-size scaling w.r.t. the layer size. The plot show the results of reconstruction capabilities for varying layer size N and $L = 3, 10$. Results are averaged over 20 different realizations of the patterns. The model parameters are $\beta = 2$, $\lambda = 0.2$, $H = 0.1$. The number of observation is fixed to $m = 2m_{\text{min}}(K)$. The number of dynamics updates of each network is fixed to $5 \cdot N$.

are interested in, we can conclude that the probability of extracting all of the patterns at least once is approximately

$$P(\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^K \text{ observed}) = 1 - K\Big(1 - \frac{L}{K}\Big)^m.$$

This represents an ideal scenario for our setting, in which each layer extracts exactly L distinct patterns at each step, without generating duplicates or failing to reconstruct any ground-truth. To ensure a high probability of observing all K patterns, we impose the condition $P(\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^K \text{ observed}) = 1-\epsilon$ with ϵ being the tolerance against failed experiments. Thus, we can thus set $K(1 - L/K)^m = \epsilon$ so that, expanding at the leading contribution in K, we get

$$m_{\text{min}}(K) \approx \frac{K}{L} \log \frac{K}{\epsilon}.$$

In our experiments, we fix $\alpha = 0.01$, $\epsilon = 0.01$ and $m = 2m_{min}(K)$. The results of the finite-size scaling analysis are reported in Fig. 6 for $\beta = 2$, $\lambda = 0.2$ and $H = 0.1$, with $L = 3, 10$. As evident from the plot, apart from the lower performance observed at small N – which lies outside the regime where the scaling approximation holds – the reconstruction capabilities remain consistently high. Moreover, they are robust with respect to both the layer size N and the number of layers L.