Journal of Statistical Mechanics:
Theory and Experiment

PAPER • **OPEN ACCESS**

Eigenvector dreaming

To cite this article: Marco Benedetti *et al J. Stat. Mech.* (2024) 013302

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Disordered systems, classical and quantum

Eigenvector dreaming

Marco Benedetti[1,*], **Louis Carillo**[1,2], **Enzo Marinari**[1,3,4] **and Marc Mézard**[5]

[1] Dipartimento di Fisica, Sapienza Università di Roma, Roma, Italy
[2] Département de physique, ENS Paris-Scalay, Gif-sur-Yvette, France
[3] CNR Nanotec, Roma, Italy
[4] INFN Sezione di Roma, Roma, Italy
[5] Department of Computing Sciences, Bocconi University, Milano, Italy
E-mail: marco.benedetti@uniroma1.it

Received 18 September 2023
Accepted for publication 14 November 2023
Published 17 January 2024

Online at stacks.iop.org/JSTAT/2024/013302
https://doi.org/10.1088/1742-5468/ad138e

Abstract. Among the performance-enhancing procedures for Hopfield-type networks that implement associative memory, Hebbian unlearning (HU) (or dreaming) strikes for its simplicity and lucid biological interpretation. However, it does not easily lend to a clear analytical understanding. Here, we show how HU can be efficiently described in terms of the evolution of the spectrum and the eigenvectors (EVs) of the coupling matrix. That is, we find that HU barely changes the EVs of the coupling matrix, whereas the benefits of the procedure can be ascribed to an intuitive evolution of the spectrum. We use these ideas to design novel dreaming algorithms that are effective from a computational point of view and are analytically far more transparent than the original scheme.

Keywords: learning theory, neuromorphic models, synaptic plasticity

*Author to whom any correspondence should be addressed.

Contents

1. Introduction

Hopfield-type neural networks are a ubiquitous framework for modeling associative memory [1]. The task at hand is to reconstruct an extensive number $P = \alpha N$ of binary patterns $\{\xi_i^\mu\} = \pm 1$, $\mu \in [1, \ldots, P]$, called *memories*, based on noise corrupted inputs. This outcome is achieved through a dynamical process, dictating the evolution in time of a collection of N binary neurons $\{S_i = \pm 1\}$, $i \in [1, .., N]$

$$S_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} J_{ij} S_j(t)\right), \qquad i = 1, .., N, \tag{1}$$

where J is the coupling matrix of the network. The dynamics can be run either in parallel (i.e. *synchronously*) or in series (i.e. *asynchronously* in a predetermined or in a random order) over the i indices. The reconstruction process is based on initializing the network dynamics into a configuration similar enough to one of the memories and iterating equation (1) asynchronously until a fixed point is reached. The network performs well if such asymptotic states are similar enough to the memories. Whether this is the case depends on the number of patterns one wants to store and on the choice of the coupling

J. Stat. Mech. (2024) 013302

matrix J. We will focus on i.i.d. memories generated with a probability $P(\xi_i^\mu = \pm 1) = 1/2$. In this setting, Hebb's learning prescription [2]

$$J_{ij}^H = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu , \qquad J_{ii}^H = 0 \tag{2}$$

used in [1] allows retrieving memories up to a critical capacity $\alpha_c^H \sim 0.14$ [3]. Even when $\alpha < \alpha_c^H$ memories are not perfectly recalled, the system's state always presents a small finite fraction of misaligned spins.

Several techniques have been developed to build more performing coupling matrices, that is, to reduce the retrieval errors and increase the critical capacity and the size of the basins of attraction to which the memories belong [4–8]. One such technique is Hebbian unlearning (HU). Inspired by the brain functioning during REM sleep [9], the unlearning algorithm [10–13] is a training procedure for the coupling matrix J to facilitate error-free retrieval and increased critical capacity in a symmetric neural network. More than 40 years after its inception, the mechanism underlying HU efficacy is still not completely understood. In this work, we first provide a novel characterization of HU in terms of the evolution of the spectrum of the coupling matrix. Intuition gathered from this analysis is then used to design two novel and efficient dreaming algorithms whose performance, contrary to HU, can be characterized analytically.

2. HU

The coupling matrix is built according to the following iterative procedure:

Algorithm 1. Hebbian unlearning.

Initialize J using Hebb's rule equation (2)
for $d = 1$ to $D_{\max}$ **do**
  Initialize network to a random state S.
  Follow dynamics equation (1) to a stable point S^*.
  for $i \neq j$ **do**
    $J_{ij} \leftarrow J_{ij} - \frac{\epsilon}{N} S_i^* S_j^*$
  end for
end for

The learning rate ϵ and the number of dreams $D_{\max}$ are free parameters of the algorithm. Algorithm 1 does not change the diagonal elements of the coupling matrix, which are fixed to $J_{ii} = 0$. This is an example of a two-phase learning process: the first phase explicitly exploits the information contained in the dataset through Hebb's rule. The iterative part of the training process, on the other hand, can be considered *unsupervised* in the sense that the network does not have access to the data but uses only the information implicitly encoded in the Hebb's rule to improve its performance.

J. Stat. Mech. (2024) 013302

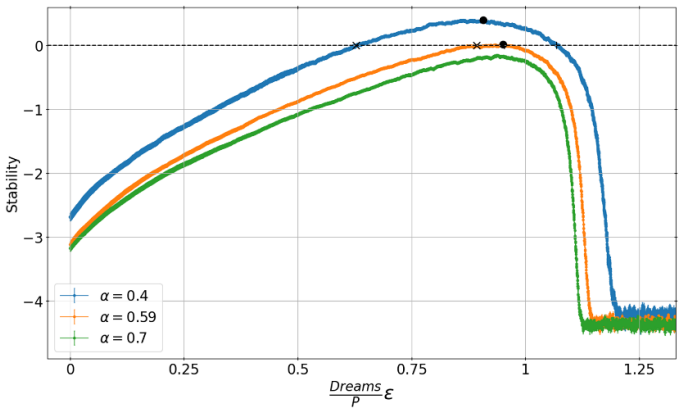

Figure 1. The minimum stability $\Delta_{\min}$ as a function of the normalized number of dreams for different values of α. The threshold $\Delta = 0$ is indicated with the gray dotted line. For $\alpha < 0.59$, $\Delta_{\min}$ crosses zero at D_{in}, peaks at $D = D_{\text{top}}$ and then becomes negative again at $D = D_{\text{fin}}$. The three appropriate relevant amounts of dreams are indicated as follows: $D = D_{\text{in}}$ by 'x', $D = D_{\text{top}}$ by a dot, $D = D_{\text{fin}}$ by a '+'. All measurements are averaged over 50 realizations of the network. $N = 400$, $\epsilon = 10^{-2}$.

One way to benchmark a learning algorithm is by tracking the evolution of the minimum stability $\Delta_{\min} \equiv \min_{i,\mu}\{\Delta_i^\mu\}$, where the *stability* Δ_i^μ is defined by:

$$\Delta_i^\mu = \frac{\xi_i^\mu}{\sqrt{N}\sigma_i} \sum_{j=1} J_{ij}\xi_j^\mu, \qquad \sigma_i = \sqrt{\sum_{j=1}^{N} J_{ij}^2/N}. \tag{3}$$

The stability's value tells us if a given pattern is aligned or not to its memory field. As soon as $\Delta_{\min} > 0$, memories themselves become fixed points of the dynamics [14], allowing error-free retrieval when the dynamics is initialized close enough to one of them. For sufficiently small values of the learning rate, below the critical load $\alpha < \alpha_c^{\text{HU}} \sim 0.6$, the evolution of $\Delta_{\min}$ in HU follows a non-monotonic curve as a function of $D_{\max}$, as illustrated in algorithm 1. The number of dreams $D = D_{\text{in}}$ marks the point where $\Delta_{\min}$ crosses 0. Here, all the memories are fixed points of the dynamics. Two other points, $D = (D_{\text{top}}, D_{\text{fin}})$, are shown in the plot, corresponding to the maximum of $\Delta_{\min}$ and the point where $\Delta_{\min}$ becomes negative again. The scaling of $(D_{\text{in}}, D_{\text{top}}, D_{\text{fin}})$ with N, ϵ, α was studied in [13].

In addition to error-free retrieval, when $\alpha < \alpha_c^{\text{HU}}$, dreaming creates large basins of attraction around the memory. This can be measured in terms of the *retrieval map:*

$$m_{\text{f}}(m_0) \equiv \overline{\left\langle \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu S_i^\mu(\infty) \right\rangle}, \tag{4}$$

where $\vec{S}^\mu(\infty)$ is the stable fixed point reached when the dynamics is initialized to a configuration $\vec{S}^\mu(0)$ having overlap m_0 with a given memory $\vec{\xi}^\mu$. The symbol $\overline{\cdot}$ denotes

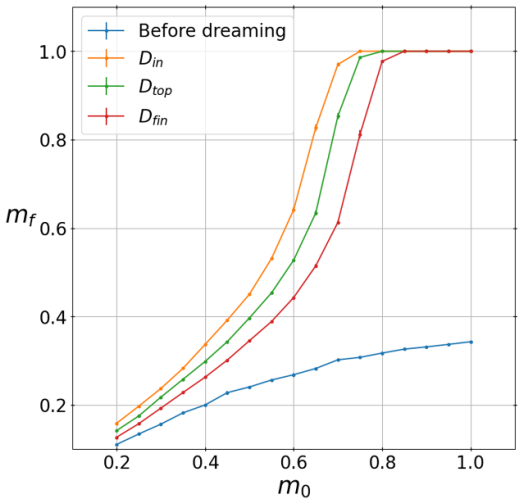

Figure 2. Retrieval map $m_{\mathrm{f}}(m_0)$ for the unlearning algorithm at the three relevant steps indicated in figure 1, and before unlearning. All measurements are averaged over 10 realizations of the network. $N = 1000$, $\alpha = 0.4$, $\epsilon = 10^{-2}$. The performance of the algorithm is maximal at $D = D_{\mathrm{in}}$.

the average over different realizations of the memories and $\langle\cdot\rangle$ the average over different realizations of $\vec{S}^{\mu}(0)$. Figure 2 shows the retrieval map for $N = 1000$ and $\alpha = 0.4$. The HU's performance is the best at $D = D_{\mathrm{in}}$. Interestingly, as discussed in [13], the curve relative to Gardner's optimal symmetric perceptron [4, 14] and to unlearning at $D = D_{\mathrm{in}}$ coincide with good accuracy.

Although the HU has served as a source of inspiration for many interesting training algorithms, some of which can be studied analytically [8, 15–17], the fundamentally non-linear sampling process for the dreaming configurations S^* makes it analytically untreatable. The next section highlights this apparently obscure procedure.

3. Two novel dreaming algorithms

An interesting interpretation of the HU algorithm emerges while analyzing the evolution of the spectrum and of the eigenvectors (EVs) of the coupling matrix J during the dreaming procedure. Before dreaming, the spectrum of J is of the Marchenko–Pastur type [18], and the N-dimensional vector space is split between a degenerate $N - P$ dimensional eigenspace orthogonal to all the memories, and a P dimensional space spanned by the memories splits into non-degenerate eigenspaces. Figure 3 focuses on the evolution under dreaming of the ranked spectrum of J. The evolution of the ranked spectrum indicates that HU is targeting, and reducing, the largest eigenvalues of the coupling matrix, whereas all other eigenvalues are increased by a constant amount at every dream, maintaining a traceless coupling. This leads to a plateau at the high end of the ranked spectrum. In figure 4, we qualify the evolution of the EV $\vec{\zeta}$ of the coupling matrix J as a function of the dreaming number. For each D, the eigenvalues are ranked from 1 to N. For each rank, we measured the overlap $\omega(\vec{\zeta}(D), \vec{\zeta}(D-1))$

J. Stat. Mech. (2024) 013302

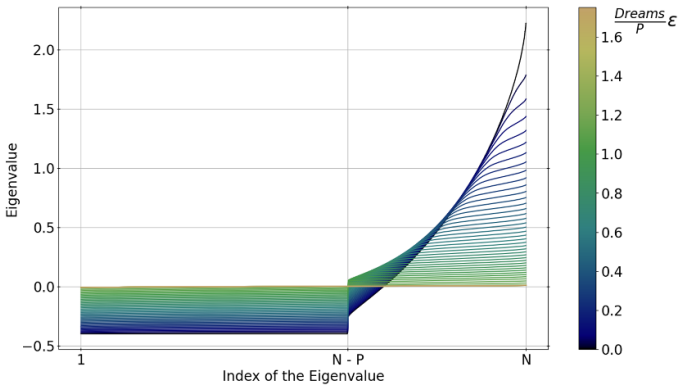

Figure 3. The y-axis presents the value of the eigenvalues; the x-axis presents their ranking. Curves of different colors correspond to the measures of the ranked spectrum taken after different amounts of dreams. Before dreaming, the spectrum is of the Marchenko–Pastur type. HU progressively flattens the high portion of the ranked spectrum.

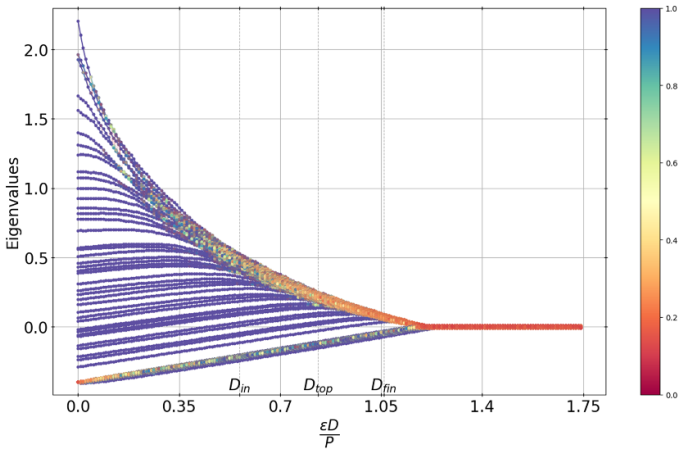

Figure 4. The x-axis presents the normalized number of steps of the dreaming algorithm. The y-axis presents the eigenvalues of the coupling matrix, for one sample, $N = 100$. Eigenvalues at different steps of the algorithm are connected by colored lines. Darker colors indicate a high overlap between the corresponding eigenvectors. Only lines corresponding to overlaps larger than 0.1 are shown. The overlap among subsequent eigenvectors is high, except for the highest and lowest parts of the ranked spectrum, where the eigenvalues are effectively degenerate.

between the corresponding EVs at step D and at step $D-1$. Eigenvalues of the same rank at different dreaming steps are connected by a continuous line, and colored with a color code connected to ω. For clarity, only lines corresponding to overlaps larger than 0.1 are shown. As the dreaming procedure unfolds, the majority of the EVs do not change much (blue lines), and the lines do not cross. This means that EVs evolve continuously, while the corresponding EVs barely change. The highest and lowest parts of the ranked spectrum, on the other hand, show some crossing of lines, and low values

J. Stat. Mech. (2024) 013302

of the overlaps (in red). This is attributed to the eigenvalues becoming almost equal, leading to an effectively degenerate eigenspace, corresponding to a plateau in figure 3.

These observations suggest the following alternative algorithm.

3.1. Eigenvector dreaming

Algorithm 2. EVdreaming.

Initialize J using Hebb's rule equation (2)
for $D = 1$ to D_{max} **do**
1-Find an orthonormal basis of eigenvectors ζ^{μ} of J.
2-Select the eigenvector ζ^{u_D} with the largest absolute eigenvalue.
3-Update $J_{ij} \leftarrow J_{ij} - \epsilon \zeta_i^{u_D} \zeta_j^{u_D}$.
4-Reset diagonal terms to zero $J_{ii} \equiv 0$
end for

In this algorithm, the update of the couplings reduces the value of the highest eigenvalue by an amount ϵ, leaving the EVs unchanged. Resetting the diagonal to zero, on the other hand, increases the value of every eigenvalue by a stochastic amount (section 3.2), and also modifies the EVs. Each step of this algorithm is based on the spectrum of the current coupling matrix. It is noteworthy that this algorithm can be implemented using purely local rules, by iterating a synchronous update

$$S^{t+1} = f\left(JS^t\right); \quad f(x) = \frac{x}{||x||_2}, \tag{5}$$

which converges toward the EV of J with the largest eigenvalue.

3.2. Initial Eigenvector dreaming (IEV)

An even simpler dreaming procedure, which does reproduce the qualitative features of HU (specifically, the centrality of the spectrum evolution and the marginality of the eigenspaces evolution), is obtained by modifying the coupling matrix on the basis of the EVs of the *initial* coupling matrix J^H, as listed in algorithm 3. We call this procedure *IEV* (dreaming).

Algorithm 3. IEVdreaming.

1-Initialize J using Hebb's rule equation (2)
2-Find an orthonormal basis of eigenvectors ζ^{μ} of the initial coupling matrix.
for $D = 1$ to D_{max} **do**
3-Consider the most recent coupling matrix J^{D-1}, and select the eigenvector ζ^{u_D} with the largest absolute eigenvalue.
4-Update $J_{ij} \leftarrow J_{ij} - \epsilon \zeta_i^{u_D} \zeta_j^{u_D}$.
5-Remove the average value of the diagonal elements of J: $J_{ii} \leftarrow J_{ii} - \frac{\epsilon}{N}$.
end for

This algorithm is simple enough that it can be analyzed in some detail.

J. Stat. Mech. (2024) 013302

3.3. A first analysis of IEV dreaming

As a first approach, imagine removing step 5 of the iterative process, and simply setting the diagonal to zero after the cycle. The resulting J reads

$$\begin{aligned} J_{ij}^{D} &= \sum_{\mu=1}^{N} \zeta_i^{\mu}\zeta_j^{\mu}\left(\lambda_\mu - \epsilon\sum_{d=1}^{D}\delta_\mu^{u_d}\right) + \epsilon\sum_{d=1}^{D}\left(\zeta_i^{u_d}\right)^2\delta_{ij} \\ &= \sum_{\mu=1}^{N} \zeta_i^{\mu}\zeta_j^{\mu}\left(\lambda_\mu - \epsilon\sum_{d=1}^{D}\delta_\mu^{u_d}\right) + \epsilon\sum_{d=1}^{D}\langle\left(\zeta_i^{u_d}\right)^2\rangle\delta_{ij} \\ &\quad + \epsilon\sum_{d=1}^{D}\left[\left(\zeta_i^{u_d}\right)^2 - \langle\left(\zeta_i^{u_d}\right)^2\rangle\right]\delta_{ij}\,, \end{aligned} \tag{6}$$

where the average $\langle(\zeta_i^{u_d})^2\rangle$ is computed over the statistics generated by the choice of the EV u_D to be dreamed at each step, given the realization of disorder (i.e., the value of the EVs ζ_i^{μ}). Since the EVs of a Wishart matrix are isotropically distributed on the $(N-1)$-dimensional sphere, one has that $\langle(\zeta_i^{u_d})^2\rangle = 1/N$. The result is then

$$J_{ij}^{D} \simeq \sum_{\mu=1}^{N}\zeta_i^{\mu}\zeta_j^{\mu}\left(\lambda_\mu - \epsilon\, d_\mu\right) + \epsilon\frac{D}{N}\delta_{ij} + \eta_{ij}\,, \tag{7}$$

where $d_\mu = \sum_{D=1}^{D}\delta_\mu^{u_D}$ and η_{ij} is a diagonal random matrix

$$\eta_{ij} \equiv \epsilon\sum_{d=1}^{D}\left[\left(\zeta_i^{u_d}\right)^2 - \langle\left(\zeta_i^{u_d}\right)^2\rangle\right]\delta_{ij}\,. \tag{8}$$

The first two terms preserve the EVs of J. The η correction changes both the EVs and eigenvalues of the coupling matrix, and assuming that η is small enough, we can compute these changes perturbatively. In particular, the degenerate eigenspace corresponding to the low eigenvalue plateau will be split by corrections $\lambda \to \lambda + \delta\lambda_i$, $i = 1,\ldots,N-P$ given by the $N-P$ eigenvalues of the matrix

$$A^{\mu\nu} \equiv \zeta^{\mu\top}\eta\zeta^{\nu}, \qquad \mu,\nu = 1,\ldots,N-P\,, \tag{9}$$

where the EVs all belong to the low eigenvalue degenerate plateau (any orthonormal set of EVs is equivalent). In the thermodynamic limit, the impact of η on J becomes negligible, as shown in figure 5. The x-axis represents N. The y-axis represents the eigenvalues of the A matrix equation (9) divides by the absolute height of the low plateau. In the thermodynamic limit, all curves tend to zero, showing that the corrections become negligible compared to the low plateau value. Some insight into this behavior can be gained by considering the statistics of the diagonal element of η. Their average is zero, by definition. If the ξ_i^{μ} involved in equation (8) were a finite number, they could

J. Stat. Mech. (2024) 013302

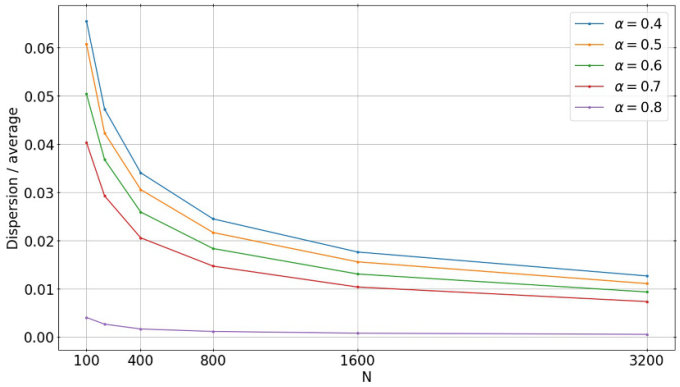

Figure 5. Dispersion of the corrections to the low plateau eigenvalues, divided by the low plateau eigenvalue, at D_{top}, as a function of N, for different values of α. As the system size is increased, the corrections become negligible compared to the low plateau eigenvalue.

be treated as i.i.d. normal variables $\mathcal{N}(0, 1/N)$, and the statistics of η could be heuristically understood as proportional to a χ^2 distribution, whose variance scales as $1/N$ (this is not exact because not every EV is dreamed the same number of times). Since we dream of an extensive number of EVs, the ξ_i^μ are not independent (for one thing, they are constrained by normalization $\sum_{\mu=1}^{N} \xi_i^\mu = 1$). Intuitively, this has the effect of reducing the variance of η_{ii}. Hence, the χ^2 distribution is an upper bound for the size of η, going to zero. Given that the dreaming procedure is described by a simple update rule

$$J_{ij}^D \simeq \sum_{\mu=1}^{N} \zeta_i^\mu \zeta_j^\mu \left(\lambda_\mu - \epsilon d_\mu\right) + \epsilon \frac{D}{N} \delta_{ij}. \tag{10}$$

This algorithm is very inexpensive from the computational point of view since one does not need to compute EVs multiple times.

Whether the correction to the diagonal elements of J is conducted at each step of the algorithm or at the end, affects the choice of the EV that gets dreamed: if the correction is performed at the end, the negative degenerate plateau will quite soon be higher in absolute value than the high plateau (we call this *inversion*). The algorithm then starts selecting EVs from the low plateau, which are orthogonal to the memories, having no effect on the stabilities. In contrast, the choice in algorithm 3 reproduces the qualitative behavior of HU in an analytically simple setting, since taking out the diagonal at each step decreases the absolute value of the low negative plateau while increasing the absolute value of the positive plateau, delaying the inversion.

J. Stat. Mech. (2024) 013302

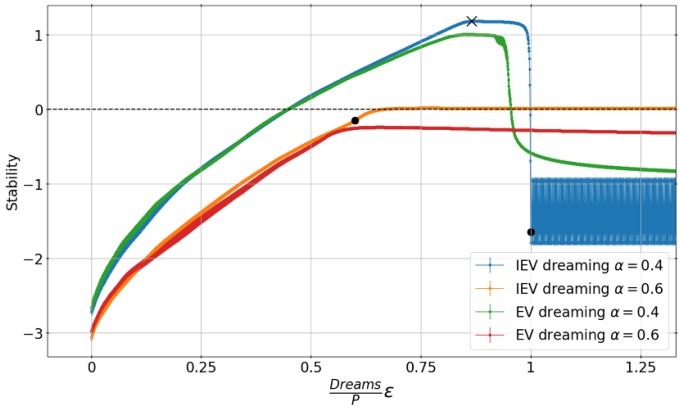

Figure 6. Evolution of $\Delta_{\min}$ while iterating different dreaming procedures, for some α values. $N = 400$, $\epsilon = 0.001$. D_{top} is indicated by a cross, and D_{inv} is indicated by a dot. The new algorithms have very similar performances before D_{inv}, indicating the IEV dreaming is indeed a good model of EV dreaming.

4. Algorithm performance

Figure 6 depicts representative examples of the evolution of $\Delta_{\min}$ according to the different dreaming procedures. The newly introduced algorithms have very similar performance before the inversion point D_{inv} (marked by circles on the curves in figure 6). This also indicates that the IEV dreaming is indeed a good model of EV dreaming. They also display the same qualitative behavior as HU. In figure 6, crosses on the curves indicate when the algorithms start dreaming for the first time in the lowest eigenvalue of the highest portion of the ranked spectrum. This condition corresponds to the highest portion of the ranked spectrum becoming a plateau. In our new procedures, this instant is very close to D_{inv}. After D_{inv}, IEV and EV display a plateau in the stability curve, which lasts until the inversion point, marked by dots in the curves. After the inversion point, which experimentally happens first in EV dreaming, EV and IEV display different behaviors since the procedure becomes very sensitive to the EVs dreamt. The behavior of IEV dreaming is detailed in section 5.

Figure 7 compares the different algorithms in terms of the retrieval mapping, at $d = D_{\text{in}}$, where the performance is optimal. The quantitative differences in the $\Delta_{\min}$ profile between the algorithms are reduced to virtually no difference, when the retrieval mapping is concerned. Below the critical load wide basins of attractions are produced around the memories.

Defining the critical capacity of an algorithm α_c as the highest load such that $\Delta_{\min} > 0$ is reached before D_{inv}, we find $\alpha_c^{\text{IEVd}} \sim 0.57$ and $\alpha_c^{\text{EVd}} \sim 0.55$, to be compared with $\alpha_c^{\text{HU}} \sim 0.59$.

J. Stat. Mech. (2024) 013302

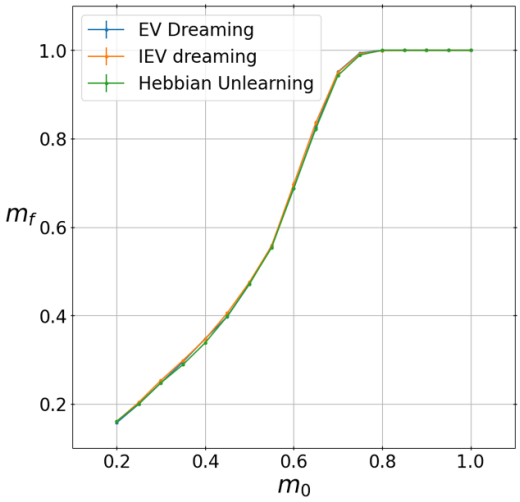

Figure 7. Retrieval mapping for the various dreaming procedures, at $D = D_{\text{in}}$, where attraction basins are the largest. $N = 400$, $\alpha = 0.4$, $\epsilon = 0.01$. Different curves coincide, suggesting that our new dreaming procedures capture the essence of HU.

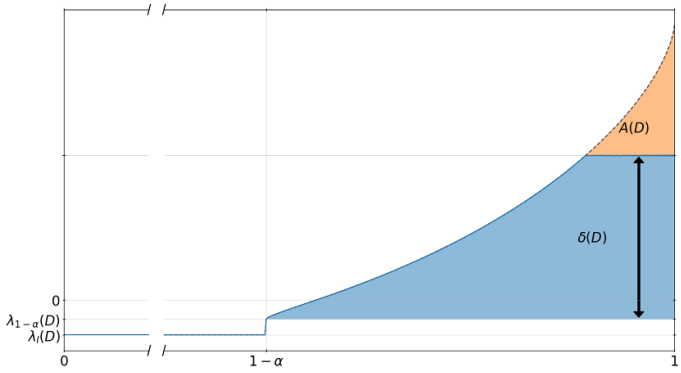

Figure 8. Evolution of the ranked spectrum during IEV dreaming.

5. Analytical characterization of IEV dreaming

In the case of IEV dreaming, both the values of D_{top} and D_{inv} can be computed analytically. Let us define by $\lambda_l(D)$ the height of the low plateau, by $\lambda_{1-\alpha}(D)$ the height of the lowest eigenvalue in the high part of the ranked spectrum, and by $\delta(D)$ the distance between the high plateau and $\lambda_{1-\alpha}(D)$ (figure 8).

Before dreaming, one has

$$\lambda_l(0) = -\alpha \tag{11}$$

$$\lambda_{1-\alpha}(0) = 1 - 2\sqrt{\alpha} \tag{12}$$

$$\delta(0) = 4\sqrt{\alpha}\,. \tag{13}$$

At each dream, the change in the ranked spectrum consists of an increase of every eigenvalue due to the resetting to zero of the diagonal elements of J, and a decrease

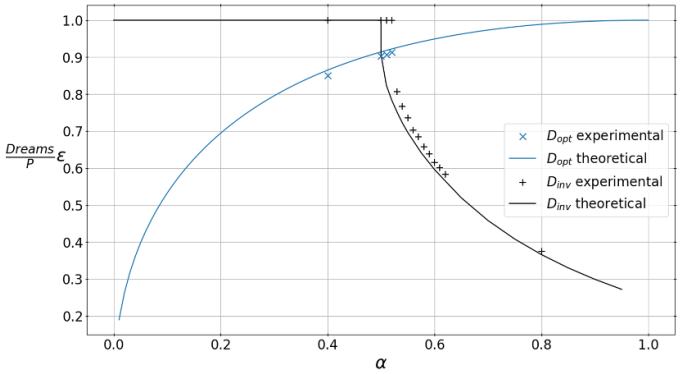

Figure 9. Comparison between analytical estimate and simulations for D_{inv} and D_{top} as a function of α. Parameters for the simulations are $N = 1000$, $\epsilon = 0.001$. The agreement is excellent, as finite size effects are already small at this size.

in the dreamed eigenvalue, as per equation (10). Before D_{top}, that is before the high part of the ranked spectrum is completely flattened into a plateau, the evolution of the spectrum can be characterized by:

$$\lambda_l(D) = \lambda_l(0) + \frac{\epsilon D}{N} \tag{14}$$

$$\lambda_{1-\alpha}(D) = \lambda_{1-\alpha}(0) + \frac{\epsilon D}{N}\,, \tag{15}$$

while $\delta(D)$ can be determined numerically, noting that the area $A(D)$ is:

$$A(D) = \frac{\epsilon D}{N}. \tag{16}$$

Similar geometrical reasoning for $D > D_{\text{top}}$ leads to even simpler equations:

$$\lambda_l(D) = \lambda_l(D_{\text{top}}) + \frac{\epsilon(D_{\text{top}} - D)}{N} \tag{17}$$

$$\lambda_{1-\alpha}(D) = \lambda_{1-\alpha}(D_{\text{top}}) + \frac{\epsilon(D_{\text{top}} - D)}{N}\left(1 - \frac{1}{\alpha}\right) \tag{18}$$

$$\delta(D) = 0\,. \tag{19}$$

Given these relations, D_{top} and D_{inv} are determined by:

$$\delta(D_{\text{top}}) = 0 \tag{20}$$

$$\left|\lambda_l(D_{\text{inv}})\right| = \left|\lambda_{1-\alpha}(D_{\text{inv}}) + \delta(D_{\text{inv}})\right|\,. \tag{21}$$

These theoretical results for D_{top} and D_{inv} are compared to the results of the numerical simulations in figure 9, with excellent agreement.

J. Stat. Mech. (2024) 013302

In IEV dreaming, the evolution of the stabilities is determined exclusively by the evolution of the spectrum of J, since the EVs do not change,

$$\Delta_i^\mu = \xi_i^\mu \frac{\sum_{\nu=1}^N \lambda_\nu \zeta_i^\nu w_\nu^\mu}{\sqrt{\sum_{\nu=1}^N \left(\lambda_\nu \zeta_i^\nu\right)^2}}, \tag{22}$$

where w_ν^μ are the coordinates of the memories in the basis of the EVs:

$$w_\nu^\mu \equiv \left(\boldsymbol{\zeta}^\nu \cdot \boldsymbol{\xi}^\mu\right) . \tag{23}$$

After D_{top}, when the spectrum is composed by two plateaus $\mathcal{P}\pm$, this expression simplifies to:

$$\Delta_i^\mu = \xi_i^\mu \frac{\sum_{\nu\in\mathcal{P}_+} \zeta_i^\nu w_\nu^\mu}{\sqrt{\sum_{\nu\in\mathcal{P}_+} \left(\zeta_i^\nu\right)^2 + \left(\frac{\lambda_l(D)}{\lambda_{1-\alpha}(D)}\right)^2 \sum_{\nu\in\mathcal{P}_-} \left(\zeta_i^\nu\right)^2}} , \tag{24}$$

which is constant (after D_{top}) as a consequence of equations (17) and (18). This explains the plateaus in figure 6.

For $\alpha < 0.5$, one has $D_{\text{inv}} = P/\epsilon$, and $\lambda_l(D_{\text{inv}}) = \lambda_{1-\alpha}(D_{\text{inv}}) = 0$. This means that at D_{inv}, we have $J = 0$. In numerical simulations, given a finite value of ϵ, this never happens. Instead, from D_{inv} the network dreams of every EV of the high plateau, making it smaller than the low plateau, and then every EV in the low plateau. Over N dreams, all EVs have been dreamt once. Thus, each eigenvalue is decreased once by $-\epsilon$ and increased N times by $\frac{\epsilon}{N}$, restoring it to the initial value. This is reflected in a periodic behavior of $\Delta_{\min}$, which oscillates (see figure 6). For $\alpha > 0.5$, the inversion occurs with well separated plateaus $\lambda_l(D_{\text{inv}}) < 0 < \lambda_{1-\alpha}(D_{\text{inv}})$. Hence, around D_{inv}, when the high plateau and the low plateau become closer than ϵ in absolute value, the network starts dreaming of one EV of the low plateau. At each dream, the corresponding eigenvalue is made even smaller, i.e. bigger in absolute value, and the network gets stuck dreaming it repeatedly. Asymptotically, this EV (orthogonal to the memories) dominates the coupling matrix, leading again to zero stability without oscillations (figure 6).

Notice that, if one decides to always dream of the EV with the largest eigenvalue (without the absolute value), one gets rid of the inversion phenomenon. The dreaming procedure will then always lead to a spectrum with two plateaus. This spectrum corresponds to a diagonal-free version of the pseudoinverse learning rule [19], whose thermodynamics has been studied in [20]. In the present work, we decided to base dreaming on the absolute value of the EVs to preserve the idea of choosing to dream configurations on the basis of a *dynamics*, as per equation (5).

6. Conclusions

This paper unveils an interesting feature of HU, namely, the fact that EVs of the coupling matrix do not change significantly during the algorithm, and the improvement in the recognition performance is mostly due to a modification of the spectrum. Starting

J. Stat. Mech. (2024) 013302

from this observation, we have proposed two new effective unlearning algorithms: EV dreaming and IEV dreaming, which emphasize the splitting of the learning problem into a trivial EV evolution and a non-trivial spectrum evolution, respectively. IEV dreaming is the simplest algorithm, being computationally efficient and easy to control analytically. IEV dreaming turns out to give a very good description of EV dreaming, and a qualitatively good description of HU. Finally, in our new algorithm, we find a strong correlation between the moment when the lowest eigenvalues of the high plateau starts being dreamed, and the moment when the algorithm stops increasing the minimum stability $\Delta_{\min}$. This correlation, which follows from simple analytical arguments in the case of IEV dreaming, is also present, to a lesser extent, in HU.

Acknowledgments

E M acknowledges funding from the PRIN funding scheme (2022LMHTET—Complexity, disorder and fluctuations: spin glass physics and beyond) and from the FIS (Fondo Italiano per la Scienza) funding scheme (FIS783—SMaC—Statistical Mechanics and Complexity: theory meets experiments in spin glasses and neural networks) from Italian M U R (Ministery of University and Research). M M acknowledges financial support by the PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci.* **79** 2554–8
[2] Hebb D O 1949 *The Organization of Behavior* (Wiley)
[3] Amit D J, Gutfreund H and Sompolinsky H 1987 Statistical mechanics of neural networks near saturation *Ann. Phys., NY* **173** 30–67
[4] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257–70
[5] Gardner E J, Wallace D J and Stroud N 1989 Training with noise and the storage of correlated patterns in a neural network model *J. Phys. A: Math. Gen.* **22** 2019
[6] Wong K Y M and Sherrington D 1990 Optimally adapted attractor neural networks in the presence of noise *J. Phys. A: Math. Gen.* **23** 4659
[7] Dotsenko V S, Yarunin N D and Dorotheyev E A 1991 Statistical mechanics of Hopfield-like neural networks with modified interactions *J. Phys. A: Math. Gen.* **24** 2419
[8] Nokura K 1996 Unlearning in the paramagnetic phase of neural network models *J. Phys. A: Math. Gen.* **29** 3871
[9] Crick F and Mitchison G 1983 The function of dream sleep *Nature* **304** 111–4
[10] Hopfield J J, Feinstein D I and Palmer R G 1983 'Unlearning' has a stabilizing effect in collective memories *Nature* **304** 158–9
[11] van Hemmen J L, Ioffe L B, Kühn R and Vaas M 1990 Increasing the efficiency of a neural network through unlearning *Physica* A **163** 386–92
[12] van Hemmen L 1998 Hebbian learning, its correlation catastrophe and unlearning *Netw. Comput. Neural Syst.* **9** 153–153
[13] Benedetti M, Ventura E, Marinari E, Ruocco G and Zamponi F 2022 Supervised perceptron learning vs unsupervised Hebbian unlearning: approaching optimal memory retrieval in Hopfield-like networks *J. Chem. Phys.* **156** 104107

J. Stat. Mech. (2024) 013302

[14] Gardner E, Gutfreund H and Yekutieli I 1989 The phase space of interactions in neural networks with definite symmetry *J. Phys. A: Math. Gen.* **22** 1995
[15] Plakhov A Y and Semenov S A 1992 The modified unlearning procedure for enhancing storage capacity in Hopfield network *RNNS/IEEE Symp. on Neuroinformatics and Neurocomputers* pp 242–2511 (IEEE)
[16] Fachechi A, Agliari E and Barra A 2019 Dreaming neural networks: forgetting spurious memories and reinforcing pure ones *Neural Netw.* **112** 24–40
[17] Agliari E, Alemanno F, Barra A and Fachechi A 2019 Dreaming neural networks: rigorous results *J. Stat. Mech.* 083503
[18] Marčenko V A and Pastur L A 1967 Distribution of eigenvalues for some sets of random matrices *Math. USSR-Sbornik* **1** 457–83
[19] Personnaz L, Guyon I and Dreyfus G 1985 Information storage and retrieval in spin-glass like neural networks *J. Phys. Lett.* **46** 359–65
[20] Kanter I and Sompolinsky H 1987 Associative recall of memory without errors *Phys. Rev.* A **35** 380–92

J. Stat. Mech. (2024) 013302