

# Dreaming neural networks: forgetting spurious memories and reinforcing pure ones.

---

**Alberto Fachechi<sup>a,b,c</sup> Elena Agliari<sup>d,e</sup> Adriano Barra<sup>a,b,c</sup>**

<sup>a</sup>*Dipartimento di Matematica e Fisica Ennio De Giorgi, Università del Salento, Italy*

<sup>b</sup>*GNFM-INdAM Sezione di Lecce, Italy*

<sup>c</sup>*INFN, Istituto Nazionale di Fisica Nucleare, Sezione di Lecce, Italy*

<sup>d</sup>*Dipartimento di Matematica, Sapienza Università di Roma, Italy*

<sup>e</sup>*GNFM-INdAM Sezione di Roma, Italy*

*E-mail:* [alberto.fachechi@le.infn.it](mailto:alberto.fachechi@le.infn.it), [elena.agliari@uniroma1.it](mailto:elena.agliari@uniroma1.it),  
[adriano.barra@unisalento.it](mailto:adriano.barra@unisalento.it)

**ABSTRACT:** The standard Hopfield model for associative neural networks accounts for biological Hebbian learning and acts as the *harmonic oscillator* for pattern recognition, however its maximal storage capacity is  $\alpha \sim 0.14$ , far from the theoretical bound for symmetric networks, i.e.  $\alpha = 1$ .

Inspired by sleeping and dreaming mechanisms in mammal brains, we propose an extension of this model displaying the standard on-line (awake) learning mechanism (that allows the storage of external information in terms of patterns) and an off-line (sleep) unlearning&consolidating mechanism (that allows spurious-pattern removal and pure-pattern reinforcement): this obtained *daily prescription* is able to saturate the theoretical bound  $\alpha = 1$ , remaining also extremely robust against thermal noise.

Both neural and synaptic features are analyzed both analytically and numerically. In particular, beyond obtaining a phase diagram for neural dynamics, we focus on synaptic plasticity and we give explicit prescriptions on the temporal evolution of the synaptic matrix. We analytically prove that our algorithm makes the Hebbian kernel converge with high probability to the projection matrix built over the pure stored patterns. Furthermore, we obtain a sharp and explicit estimate for the “sleep rate” in order to ensure such a convergence.

Finally, we run extensive numerical simulations (mainly Monte Carlo sampling) to check the approximations underlying the analytical investigations (e.g., we developed the whole theory at the so called *replica-symmetric* level, as standard in the Amit-Gutfreund-Sompolinsky reference framework) and possible finite-size effects, finding overall full agreement with the theory.

**KEYWORDS:** Unlearning, Reinforcement learning, Statistical Mechanics, Sleep&Dream

---

## Contents

<b>1</b>	<b>Introduction: the starting points</b>	<b>1</b>
<b>2</b>	<b><i>Unlearning&amp;Consolidating: Focusing on Neurons</i></b>	<b>7</b>
2.1	Model's definition, free energy and self-consistency equations	7
2.1.1	Replica-symmetric scenario through statistical mechanics	8
2.2	Remotion <i>or</i> Reinforcement: a separate analysis	9
2.3	Zero-temperature (noise-less) critical capacity	10
2.4	Replica symmetric phase diagram	12
2.5	Numerical results	14
2.5.1	Checking the Replica Symmetric assumption	14
2.5.2	Fields distributions in retrieved states	15
2.5.3	Retrieval frequency for noisy inputs: on the attraction basins	16
<b>3</b>	<b><i>Unlearning&amp;Consolidating: Focusing on Synapses</i></b>	<b>18</b>
3.1	Time evolution of the synaptic matrix	18
3.1.1	The continuous algorithm	18
3.1.2	The discrete algorithm	18
<b>4</b>	<b>Conclusions</b>	<b>20</b>
<b>A</b>	<b>Calculations to obtain the replica symmetric solution</b>	<b>22</b>
<b>B</b>	<b>Convergence of the discrete algorithm: the analytical proof</b>	<b>23</b>

---

## 1 Introduction: the starting points

An *intelligent* machine must be able to *learn* new patterns of information and to *retrieve* previously learnt ones as a response to external stimuli: these two intimately related concepts are the main aspects of cognition in Artificial Intelligence (AI). More sophisticated machines also exhibit the ability to *reinforce* relevant memories (e.g. pure states) and to *remove* irrelevant ones (e.g. mixture states), allowing a smarter storage of information. In this work, keeping the paradigmatic [Hopfield model](#) as the awake reference, we equip it with reinforcement and remotion features (able to work simultaneously, during the network *sleep*, as inspired by real sleeping and dreaming mechanisms in mammal brains: oversimplifying, a sleeping session can be split in two different modes: *rapid eye movement sleep* (REM sleep) and *slow wave sleep* (SW sleep); the former yields to erasure of unnecessary memories, the latter to consolidation of the important ones [2, 25, 58, 63]. Usually these two stages of sleep alternate during the night and, of relevance for synaptic homeostasis, the former is particular important in order to globally reduce synaptic strength (and its relative consumption of energy and tissue, an idea in agreement with the original Parisi proposal on forgetting neural networks [47, 52]), while the latter is more dedicated to consolidation of relevant memories through some sort of off-line reinforcement learning [23, 57].

[In the Literature on Artificial Intelligence](#), reinforcement (of pure states) and remotion (of spurious states) are typically addressed separately (see *e.g.* [36, 67] for the former and [40, 64] for the latter). Here, instead, we propose a unified framework for synaptic plasticity where simultaneously

reinforcement *and* remotion take place. As we will see, the combined effect of these mechanisms determines a larger retrieval region, where retrieval is stable against both the fast and the slow noise. To our knowledge, the resulting associative network outperforms other models (with symmetric interactions) appeared in the Literature. In the remainder of this Section, we provide a short description of the state of the art focusing on those aspects that are mostly related to our work.

Since the seminal work by John J. Hopfield in the eighties [35], associative neural networks have become the standard model to capture collective capabilities spontaneously shown by networks of interacting neurons. In a nutshell, a Hopfield network is made of  $N$  units mimicking binary neurons, whose state (spiking/quiescent) is described by an Ising spin ( $\sigma = \pm 1$ ). Units interact pairwise through weighted links mimicking synaptic connections, whose magnitude is defined according to Hebb's rule for learning, namely, given  $P$  patterns of information  $\{\xi^\mu\}_{\mu=1,\dots,P}$  of length  $N$ , the coupling  $J_{ij}$  between the neuron  $i$  and the neuron  $j$  reads

$$J_{ij} \equiv \frac{1}{N} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu, \quad i, j = 1, \dots, N. \quad (1.1)$$

Typically, one takes Boolean patterns with entries identically and independently drawn with equal probability, *i.e.*  $P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = 1/2$ . Moreover, the set of patterns is taken as static<sup>1</sup> and are thus called *quenched*. In order to assess the retrieval of the  $\mu^{th}$  pattern, one introduces the so-called Mattis overlaps

$$m_\mu \equiv \frac{1}{N} \sum_{i=1}^N \xi_i^\mu \sigma_i, \quad \mu = 1, \dots, P, \quad (1.2)$$

in such a way that, when the neuronal configuration  $\{\sigma_i\}_{i=1,\dots,N}$  is aligned with  $\xi^\mu$ , then  $m_\mu = 1$ ; this configuration is interpreted as the retrieval of the pattern  $\xi^\mu$ . The Hopfield model is formally described by a cost-function (or *energy*, or *Hamiltonian*, to keep a physical jargon)  $H_{N,P}(\sigma|\xi)$  defined as

$$H_{N,P}(\sigma|\xi) \equiv - \sum_{i<j}^{N,N} J_{ij} \sigma_i \sigma_j \sim - \frac{1}{2N} \sum_{i,j=1}^{N,N} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j = - \frac{N}{2} \sum_{\mu=1}^P m_\mu^2. \quad (1.3)$$

Here, the first sum runs over all possible pairs of neurons, while in the second passage we neglected  $O(N^{-1})$  terms and implemented the Hebb coupling (1.1), and lastly we used the definition (1.2). Once that the cost-function  $H_{N,P}(\sigma|\xi)$  is given, exploiting the mean-field nature of the model, a neural dynamics can be easily constructed [9, 20]. To this goal, we introduce the *fast noise* (*i.e.* standard white noise, or *temperature* in physical jargon) whose magnitude is tuned by a parameter  $T \equiv 1/\beta$  (such that as  $T \rightarrow \infty$  the neural update is entirely random, while as  $T \rightarrow 0$  it reduces to a deterministic evolution [9, 20]) and the internal field  $h_i$  acting on the  $i^{th}$  neuron. In this way, the Hopfield cost-function (1.3) can be written as

$$H_{N,P}(\sigma|\xi) = - \sum_{i=1}^N h_i \sigma_i, \quad h_i = \frac{1}{2N} \sum_{j=1}^N \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \sigma_j, \quad (1.4)$$

and the stochastic neural update rule can be written as

$$P(\sigma_1(\tau+1), \dots, \sigma_N(\tau+1)) = \prod_{i=1}^N \left\{ \frac{1}{2} [1 + \sigma_i(\tau) \cdot \tanh(\beta h_i(\tau))] \right\}, \quad (1.5)$$

where the parameter  $\tau$  identifies a suitable neural-update timescale. This dynamics ensures that detailed balance holds and the neural configuration eventually converges to the Boltzmann-Gibbs

<sup>1</sup>The time scale for neuronal dynamics is much shorter than the time scale for synaptic (and therefore pattern) dynamics, in such a way that, when focusing on retrieval tasks one can take synapsis as static, see *e.g.* [9, 20].

distribution associated to the Hamiltonian (1.3), the latter playing as a Lyapunov function  $\dot{E} = 0$ . This system can be addressed via sophisticated techniques stemming from the statistical mechanics of disordered systems as pioneered by Amit-Gutfreund-Sompolinsky (AGS) [10, 11]. Before moving to that, we sketch a heuristic argument, due to Hopfield and Tank [37], to see the retrieval capabilities of the network. Since patterns are randomly generated, for an arbitrary vector state  $\{\sigma_i\}_{i=1,\dots,N}$  the related Mattis magnetizations would vanish as  $O(N^{-1/2})$  and the corresponding contribution to the cost-function (1.3) is negligible. On the other hand, for a vector state that is (partially) aligned with a given pattern, the contribution to the energy would be  $O(N)$ , and it therefore occurs to be a convenient (stable) state for the system. This suggests that the model displays energy minima at each of the assigned memories. In order to strengthen this picture, statistical mechanics definitions and tools are now essential.

The (intensive) free energy associated to the cost-function (1.3) is defined as

$$A_{N,P}(\beta) \equiv -\frac{1}{\beta N} \mathbb{E} \ln Z_{N,P}(\beta|\xi), \quad (1.6)$$

where  $Z_{N,P}(\beta|\xi) \equiv \sum_{\{\sigma\}} \exp[-\beta H_{N,P}(\sigma|\xi)]$  is called the *partition function* and  $\mathbb{E}$  denotes the average over the patterns, also termed *slow noise*. In the following, we will mainly focus on the thermodynamic limit of the free energy, namely  $A(\alpha, \beta) \equiv \lim_{N \rightarrow \infty} A_{N,P}(\beta)$ , where  $\alpha \equiv \lim_{N \rightarrow \infty} P/N$  is referred to as the *load* (or *storage capacity*) of the system.

In a statistical-mechanics approach, one aims to express the free energy explicitly in terms of the *order parameters*, namely simple functions of the state of the system under study giving informations about the behavior of the system itself. In this context, the  $P$  Mattis overlaps  $\mathbf{m} = (m_1, \dots, m_P)$  work as order parameters since, according to their value, one can infer whether the network is retrieving ( $\exists \mu | m_\mu \neq 0$ ) or not ( $m_\mu = 0, \forall \mu$ ). Further, the so-called Edward-Anderson overlap  $q_{ab} \equiv N^{-1} \sum_i \sigma_i^a \sigma_i^b$  is likewise useful as it detects the *spin-glass* regime (namely a region where “structured disorder” prevails, as will be explained in more details later) [9, 20]. In the replica symmetric approximation provided by AGS theory, the free energy of the Hopfield model expressed in terms of the Mattis and Edward-Anderson overlaps ( $q_{ab} \equiv q$  for simplicity) reads as

$$A(\alpha, \beta) = \frac{1}{2} \mathbf{m}^2 - \frac{1}{\beta} \int_{-\infty}^{+\infty} d\mu(x) \left\langle \ln \cosh \left\{ \beta \left( \mathbf{m} \cdot \boldsymbol{\xi} + x \frac{\sqrt{\alpha q}}{[1 - \beta(1 - q)]^2} \right) \right\} \right\rangle_{\boldsymbol{\xi}} \quad (1.7)$$

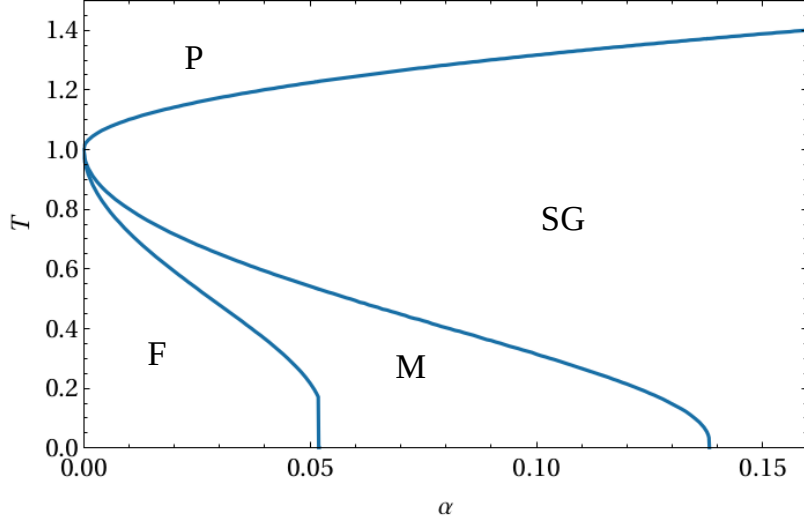
$$+ \frac{\alpha}{2\beta} \ln [1 - \beta(1 - q)] - \frac{\alpha}{2} \frac{q}{1 - \beta(1 - q)} + \frac{\alpha}{2} \frac{q}{[1 - \beta(1 - q)]^2}.$$

Recalling that, in its original interpretation in Thermodynamics [9], the free energy equals the difference between the energy (*i.e.* the expectation value of the cost-function) and the entropy (related to the probability of observing a configuration  $\{\sigma_i\}_{i=1,\dots,N}$ ), the extremization of the free energy over the order parameters ensures simultaneously the minimum energy and the maximum entropy principles. Thus, the exploration of the free-energy landscape (as the noise level  $\beta$  and the load  $\alpha$  are tuned) allows us to inspect the system thermalization and equilibria. Remarkably, as pointed out by Jaynes [39], this route has a clear meaning also from a statistical inference perspective, much closer in spirit to Machine Learning: minimizing the free-energy equals searching for the minima of the cost-function under the constraint of Maximum Entropy (see also [61]).

As anticipated, the order parameters for the Hopfield model are the  $P$  Mattis overlaps  $m_\mu$  and the Edward-Anderson overlap  $q$  and, by extremizing the free-energy (1.7) with respect to such variables, we get the following self-consistent equations:

$$\frac{dA(\alpha, \beta)}{d\mathbf{m}} = 0 \Rightarrow \mathbf{m} = \int_{-\infty}^{+\infty} d\mu(x) \left\langle \boldsymbol{\xi} \tanh \left[ \beta \left( \mathbf{m} \cdot \boldsymbol{\xi} + x \frac{\sqrt{\alpha q}}{1 - \beta(1 - q)} \right) \right] \right\rangle_{\boldsymbol{\xi}}, \quad (1.8)$$

$$\frac{dA(\alpha, \beta)}{dq} = 0 \Rightarrow q = \int_{-\infty}^{+\infty} d\mu(x) \left\langle \tanh^2 \left[ \beta \left( \mathbf{m} \cdot \boldsymbol{\xi} + x \frac{\sqrt{\alpha q}}{1 - \beta(1 - q)} \right) \right] \right\rangle_{\boldsymbol{\xi}}, \quad (1.9)$$



**Figure 1. Phase diagram of the Hopfield network.** The phase diagram lives in the  $(\alpha, \beta)$  plane. In the upper region (P) the network behaves randomly while in the top-right region (SG) it is frozen in a spin-glass phase. The working regions are solely the two down-left where patterns are global minima of the free energy (F) or relative minima of the free energy (M).

where the bracket  $\langle \cdot \rangle_{\xi}$  means the average over the quenched patterns.

By studying the solutions of these equations, one can obtain a *phase diagram* for the Hopfield network, namely, in the  $(\alpha, \beta)$  plane one can distinguish three phases characterized by different solutions of Eqs. (1.8, 1.9) and qualitatively different behaviors of the system. In our opinion, this is the greatest reward by the statistical-mechanics approach: the concept of phase diagram allows researchers to predict the network response as a function of the tunable parameters, and this can be a fundamental information in the modern theoretical foundation of AI. The phase diagram<sup>2</sup> for the Hopfield model is shown in Fig. 1, and one can detect:





- *Ergodic phase*: in the high-temperature limit, the fast noise in the system is too strong for the neurons to reciprocally feel each other, therefore the system behaves randomly and no emergent collective properties of neurons can be appreciated. This region is characterized by  $\mathbf{m} = 0$  and  $q = 0$ .
- *Spin-glass phase*: in the high-load limit, the slow-noise is too large for the neurons to correctly handle the whole set of patterns, and, again, the system fails to retrieve information. This region is characterized by  $\mathbf{m} = 0$  but  $q \neq 0$ .
- *Retrieval phase*: when both fast and small noise are relatively small, the system behaves as an associative neural network and neural collective capabilities spontaneously appear. The phase is characterized by  $\mathbf{m} \neq 0$  and  $q \neq 0$ . This region can be further split in a pure retrieval region (where pure states are global minima) and in a mixed retrieval region (where pure states are local minima, yet their attraction basin is large enough for the system to end there if properly stimulated).

<sup>2</sup>Despite almost four decades has elapsed since Hopfield's seminal work, a rigorous control of the entire phase diagram is still beyond the current mathematical technologies as the low temperature analysis of such disordered systems is notoriously difficult [26, 65]. In neural networks we usually rely on the so-called *replica symmetric approximation* for the description of the free-energy landscape, as originally outlined by AGS [9–11]. More details on replica symmetry will be presented in Sec. 2.1.



As is clear from the phase diagram of the Hopfield model, the Hebbian prescription (1.1) implies a limitation in the number of patterns that the network can correctly handle. The network can, at most, manage a number of patterns  $P$  that grows linearly in the number of neurons  $N$  available, *i.e.*  $P = \alpha N$  [11]. Once a critical threshold  $\alpha_c \sim 0.14$  is reached, the network experiences a plethora of unpleasant symptoms, ranging from the worst scenario (the abrupt transition to the spin-glass phase, sometimes called *blackout catastrophe*, affecting solely fully connected models), to milder confusional states where network’s performances are sensibly reduced, if not lost at all.<sup>3</sup>

The reason underlying this impasse is that the free-energy landscape of the system is characterized by pure-state minima (corresponding to pure pattern retrieval) but also by spurious-state minima (corresponding to mixtures of patterns that are interpreted as errors); as long as the network is fed by a linear increment (in the neural volume  $N$ ) of patterns to be stored (*i.e.*  $P \propto N$ ), there is an unavoidable, combinatorial (roughly exponential in  $N$ ) proliferation of mixed patterns which may work as “traps” for the system state [9, 20].

In the late eighties, Elisabeth Gardner found, by general arguments, that the maximal theoretical capacity for symmetric networks<sup>4</sup> is  $\alpha_c = 1$  , so the Hopfield model threshold  $\alpha_c \sim 0.14$  has always been looked at as rather poor: indeed, along the decades, scientists tried to improve the maximal capacity by implementing some extensions and variations on theme (*e.g.* keeping the network out of equilibrium [21, 24], allowing the network to process multiple tasks at once [3, 6, 7]). In this context, particularly appealing works were inspired by Crick and Mitchinson’s paper [22], where it was argued that the REM phase of sleep in mammals may serve to delete all the (involuntarily stored) irrelevant information (in order to save memory and avoid overloading catastrophes). Further evidences toward this hypothesis were found both on the empirical [2, 25, 48, 58, 63] and the theoretical [13, 27, 38, 50, 51, 54, 67] level.

A first crucial contribution to frame this idea in AI was achieved by Hopfield himself (with Feinstein and Palmer [36]), and can be summarized as follows: the spin-glass transition occurring as  $\alpha$  increases beyond  $\alpha_c$  ultimately originates from the fact that the number of spurious states are exponentially more abundant than the number of pure states (regardless their depth in the free energy landscape), such that, making a quench from infinite to zero temperature, the system would get trapped with higher probability in one of these mixtures. By sampling a number of these final configurations, one can measure the average pairwise correlation  $\langle \sigma_i \sigma_j \rangle_{\text{mix}}$ , for any  $i$  and  $j$ . Next, one updates the coupling matrix performing an *inverse Hebbian rule* so that these mixture states are effectively removed: .

$$J_{ij} \rightarrow J_{ij} - \frac{\epsilon}{N} \langle \sigma_i \sigma_j \rangle_{\text{mix}} = \frac{1}{N} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu - \frac{\epsilon}{N} \langle \sigma_i \sigma_j \rangle_{\text{mix}}, \quad (1.10)$$

where  $\epsilon$  is a tunable (but small) parameter called *unlearning strength*. Such a procedure should be iterated in order to progressively clean the free-energy landscape from these traps.<sup>5</sup> The minus sign in (1.10) is responsible for the so-called *unlearning* process. A further motivation for the analogy with the REM phase  is supplied by the fact that, during these REM phases, dreams are not entirely uncorrelated  with the experiences we actually lived during the wakefulness state and there are (possibly weird) correlations between dreams and these experiences; a similar scenario happens

<sup>3</sup>It is worth pointing however that, despite the severe criticism - see *e.g.* [29] - initially raised against the *connectionist perspective* this approach belongs to (as far as the Hopfield blackout scenario is concerned), in diluted models this abrupt transition gets smoothed and switches from first-order to second-order (in the Ehrenfest notation), thus cross-talks effects in more realistic networks certainly are present (and strong), but the discontinuous lost of the whole information is just a chimera of fully connected models [4].

<sup>4</sup>The maximal theoretical capacity reaches  $\alpha_c = 2$  for asymmetric networks.

<sup>5</sup>We stress that the normalization factor  $N^{-1}$  in front of the term  $\langle \sigma_i \sigma_j \rangle_{\text{mix}}$  is appropriately chosen if the unlearning algorithm is iterated  $O(N)$  times. We will deepen this point (the amplitude of the unlearning or consolidating rates) in Section 3 and in the Appendix B.

in the artificial side as spurious states are just mixtures of patterns<sup>6</sup> that unavoidably implies short-length correlations with the pure patterns. In the same spirit as Hopfield’s proposal, Plakhov and Semenov [54] realized an unlearning algorithm by replacing the pure pairwise correlations between spins with correlations between inner fields, namely

$$J_{ij} \rightarrow J_{ij} - \epsilon \langle h_i h_j \rangle, \quad (1.11)$$

where the average  $\langle \cdot \rangle$  is performed on a sample of randomly selected states in the configuration space with internal fields  $h_i$ . The main result is that, with a suitable choice of the unlearning strength, this algorithm is ensured to converge (up to scaling factors) to the projector (or pseudo-inverse) matrix

$$J_{ij}^p = \frac{1}{N} \sum_{\mu, \nu=1}^{P, P} \xi_i^\mu (C^{-1})_{\mu, \nu} \xi_j^\nu, \quad (1.12)$$

where

$$C_{\mu, \nu} = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu \xi_i^\nu, \quad (1.13)$$

is the pattern correlation matrix. Notably, in this model, similar in spirit to the original Kohonen idea [44] but closer in its statistical mechanical construction to the model introduced and studied by Kanter and Sompolinsky in [41], the storage capacity reaches  $\alpha_c = 1$ . A closely related model, studied by Dotsenko and coworkers [27, 28], is based on the following coupling matrix<sup>7</sup>

$$J_{ij}(t) = \frac{1}{N} \sum_{\mu, \nu=1}^{P, P} \xi_i^\mu (\mathbb{I} + tC)^{-1}_{\mu, \nu} \xi_j^\nu, \quad (1.14)$$

where  $\mathbb{I}$  is the identity matrix and  $t \in \mathbb{R}^+$  is a tuneable parameter. This model emerges as a continuous time limit (*i.e.*  $\epsilon \sim dt$ ) of the unlearning rule (1.11). Most remarkably, it turns out that the maximal storage capacity increases as  $t$  gets large.<sup>8</sup> However, as  $t$  gets larger and larger (which is the interesting limit in order to see the maximal capacity), the coupling matrix identically vanishes. As a result, on one side the retrieval region (see Fig. 2, right panel) is stretched toward higher values in  $\alpha$  with respect to the Hopfield reference (see Fig. 1), but on the other side it is also confined to smaller values of  $T$ . This effect gets more pronounced as  $t$  is increased, resulting in the total disappearance of the retrieval region.

One of the main contributions of the present work is to extend these unlearning approaches by simultaneously allowing also for reinforcement of the pure states [40, 64]. As we will see, this confers an extra-stability of these states against the fast noise, finally resulting in a sensibly enlarged and more robust retrieval region (with respect to the Hopfield reference and all the past extensions). This result also suggests that, for a smart storage of information, remotion alone does not suffice: a suitable reinforcement is also in order.

<sup>6</sup>The typical example is given by the symmetric mixtures of three patterns, that is  $\sigma_i = \text{sign}(\xi_i^1 + \xi_i^2 + \xi_i^3)$ .

<sup>7</sup>While quite marginal in AI, it is still worth stressing that such a learning rule is no longer *local*, like the Hebbian prescription, in fact, the coupling between neurons  $i$  and  $j$  now depends on pattern entries related to all the neurons making up the system. In the biological world this point constitutes a modeling weakness, however Dotsenko and coworkers have shown how to bypass it in order to obtain roughly the same results [27]. Another local algorithm able to converge to the projector matrix is the called Adeline learning rule (see [42, 67] for an overview.)

<sup>8</sup>Actually, the critical threshold found by [27] is approximately 1.07. This is not to be meant as a violation of Garner’s bound: the overflow is due to the underlying replica-symmetry approximation.



## 2 *Unlearning&Consolidating*: Focusing on Neurons

### 2.1 Model's definition, free energy and self-consistency equations

Our investigation is based on the works by Personnaz, Guyon, Dreyfus [53], by Kanter and Sompolinsky [41], and by Dotsenko et al. [27, 28]. Along the same lines, we consider a network composed by  $N$  neurons  $\{\sigma_i\}_{i=1,\dots,N}$ , with  $\sigma_i \in \{-1, +1\} \forall i$ , and  $P$  patterns  $\{\xi^\mu\}$ , with  $\xi_i^\mu \in \{-1, +1\} \forall i, \mu$ . Denoting with  $t \in \mathbb{R}^+$  the sleep extent, we propose the following

**Definition 1.** *The reinforcement&removal algorithm we propose has the following Hamiltonian representation:*<sup>9</sup>

$$H_{N,P}(\sigma|\xi, t) = -\frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N \sum_{\mu=1}^P \sum_{\nu=1}^P \xi_i^\mu \xi_j^\nu \left( \frac{1+t}{\mathbb{I}+tC} \right)_{\mu,\nu} \sigma_i \sigma_j, \quad (2.1)$$

where the  $P$  patterns  $\{\xi^\mu\}_{\mu=1,\dots,P}$ , have  $N$  binary entries  $\xi_i^\mu \in \{-1, +1\}$ , with  $i \in (1, \dots, N)$ , drawn from

$$P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = \frac{1}{2},$$

and the correlation matrix is defined as

$$C_{\mu,\nu} \equiv \frac{1}{N} \sum_{i=1}^N \xi_i^\mu \xi_i^\nu.$$

Note that the interpretation of  $t$  as the sleep extent is clear: for  $t = 0$  the system reduces to the standard Hopfield model, while for  $t \rightarrow \infty$  the system approaches the pseudo-inverse matrix model (see the Appendix B for the analytical proof). Remarkably, during the sleeping session, both reinforcement and remotion take place. In fact, in the generalized kernel appearing in 2.1, the denominator (*i.e.*, the term  $\propto (1+tC)^{-1}$ ) yields to the remotion of unwanted mixture states, while the numerator (*i.e.*, the term  $\propto 1+t$ ) reinforces the memories. We refer to Secs. 2.2 and 3 for a more extensive discussion.

In this Section, we are instead interested in obtaining the phase diagram of our model, thus to compute explicitly -and extremize over the order parameters- the model's free energy (in the thermodynamic limit and under the replica symmetric assumption). The partition function of such a model is

$$Z_{N,P}(\sigma|\xi, t) = \sum_{\{\sigma\}} e^{-\beta H_{N,P}(\sigma|\xi, t)} = \sum_{\{\sigma\}} \exp \left[ \frac{\beta}{2N} \sum_{i,j=1}^{N,N} \sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu \left( \frac{1+t}{\mathbb{I}+tC} \right)_{\mu,\nu} \sigma_i \sigma_j \right]. \quad (2.2)$$

by which we can introduce the main observable, namely

**Definition 2.** *The infinite volume limit of the intensive free energy  $A(\alpha, \beta, t)$  associated to the model (2.1) is defined as*

$$A(\alpha, \beta, t) = - \lim_{N \rightarrow \infty} \frac{1}{\beta N} \mathbb{E} \ln Z_{N,P}(\sigma|\xi, t). \quad (2.3)$$

**Remark 1.** *The “temporal variable”  $t$  within an (equilibrium) statistical mechanical theory may look weird, yet it should be noticed that the timescale for a sleeping session is much longer than the typical time scale for neuronal dynamics.*<sup>10</sup>

<sup>9</sup>As a matter of notation, we stress that the denominator  $1/(\mathbb{I}+tC)$  in the generalized kernel is intended as the inverse matrix  $(\mathbb{I}+tC)^{-1}$ .

<sup>10</sup>The latter, at least within a biological context, is fixed around  $O(10^2)$  Hertz, namely the typical spiking time (considering also the absolute refractory period of a biological neuron).



### 2.1.1 Replica-symmetric scenario through statistical mechanics


The replica symmetric assumption means that, in the thermodynamic limit  $N \rightarrow \infty$ , the order parameters self-average over their averages (denoted hereafter by a bar), *i.e.*  $\lim_{N \rightarrow \infty} P(q) = \delta(q - \bar{q})$  and  $\lim_{N \rightarrow \infty} P(\mathbf{m}) = \delta(\mathbf{m} - \bar{\mathbf{m}})$ , in such a way that the related fluctuations can be discarded. Although this is a *reasonable* assumption, we actually know that in mean-field spin-glasses replica symmetry is broken at low temperatures. However, the effects of replica symmetry breaking are expected to be mild in associative neural networks [9] and replica symmetry is the standard level of approximation in the statistical mechanical analysis of these models.

Using the standard approach of replica technique (*i.e.* the so-called *replica trick* [9, 20]), we write the large  $N$  free-energy  $A(\alpha, \beta, t)$

$$A(\alpha, \beta, t) = - \lim_{N \rightarrow \infty} \frac{1}{\beta N} \mathbb{E} \log Z_{N,P}(\sigma|\xi, t) = - \lim_{\substack{n \rightarrow 0 \\ N \rightarrow \infty}} \frac{\mathbb{E} Z_{N,P}(\sigma|\xi, t)^n - 1}{\beta n N}. \quad (2.4)$$

Throughout the paper, we shall assume that the candidate pattern to be retrieved is  $\xi^1$  and  $\xi^\mu$  for  $\mu \geq 2$  contribute to the slow noise, therefore here  $\mathbb{E}$  is the average over the  $P - 1$  not-retrieved patterns. The quenched average of the replicated partition function can be represented in Gaussian integral form as

$$\begin{aligned} \mathbb{E} Z_{N,P}(\sigma|\xi, t)^n &= \mathbb{E} \prod_{n=1}^{\alpha} \mathcal{C} \sum_{\{\sigma^1\}} \cdots \sum_{\{\sigma^n\}} \int \left( \prod_{\mu, \alpha=1}^{P,n} D z_{\mu}^{\alpha} \right) \left( \prod_{i, \alpha=1}^{N,n} D \phi_i^{\alpha} \right) \\ &\cdot \exp \left( \sqrt{\frac{\beta}{N}} (t+1) \sum_{\mu, i, \alpha=1}^{P, N, n} z_{\mu}^{\alpha} \xi_i^{\mu} \sigma_i^{\alpha} + i \sqrt{\frac{t}{N}} \sum_{\mu, i, \alpha=1}^{P, N, n} z_{\mu}^{\alpha} \xi_i^{\mu} \phi_i^{\alpha} \right), \end{aligned} \quad (2.5)$$

where  $\mathcal{P}(z_{\mu}^{\alpha}) = \mathcal{P}(\phi_i^{\alpha}) = \mathcal{N}(0, 1)$  and  $\mathcal{C} = \det^{1/2}(\mathbb{I} + tC)$  is a normalization constant compensating the prefactors of the Gaussian integrations and trivially contributing to the free energy. This model strongly resembles [27], with the only difference that - in the first term - here we have  $\beta(1+t)$  (instead of  $\beta$ ) realizing an optimal tuning between the two 2-body couplings of the relevant variables. As we will see, this scaling is crucial to keep the thermodynamic properties of the model stable, since it ensures that the critical temperature at zero load stays fixed at  $\beta_c = 1$  as  $t$  is tuned  thus, this interpolation between the Hopfield model and the pseudo-inverse one automatically prevents the collapse of the retrieval region on the horizontal axis in the phase diagram. In the Appendix A, we report in details the calculations of the free energy  $A(\alpha, \beta, t)$  of the model, while here we provide just the explicit expression (ignoring trivial contributions):

$$\begin{aligned} A(\alpha, \beta, t) &= \frac{1}{2n(1+t)} \sum_{\alpha=1}^n (m_1^{\alpha})^2 + \frac{\alpha\beta}{2n} \sum_{\alpha, \beta=1}^{n,n} p_{\alpha\beta} q_{\alpha\beta} + \frac{\alpha}{2n\beta} \log \det [\mathbb{I} - \beta(1+t)\hat{q}] \\ &- \frac{1}{n\beta} \mathbb{E} \ln \sum_{\{\sigma\}} \int \left( \prod_{\alpha=1}^n D \phi^{\alpha} \right) \exp \left[ \beta \sum_{\alpha=1}^n m_1^{\alpha} \xi^1 \left( \sigma^{\alpha} + i \sqrt{\frac{t}{\beta(1+t)}} \phi^{\alpha} \right) \right. \\ &\left. + \frac{\alpha\beta^2}{2} \sum_{\alpha, \beta=1}^{n,n} p_{\alpha\beta} \left( \sigma^{\alpha} + i \sqrt{\frac{t}{\beta(1+t)}} \phi^{\alpha} \right) \left( \sigma^{\beta} + i \sqrt{\frac{t}{\beta(1+t)}} \phi^{\beta} \right) \right], \end{aligned} \quad (2.6)$$

where  $m_1^{\alpha}$  is the M Mattis magnetization (of the  $\alpha$ -th replica) associated to the pattern  $\xi^1$  to be retrieved,  $\hat{q}$  is the overlap matrix whose element  $q_{\alpha\beta}$  is the generalized overlap between the replicas of the system (labeled with  $\alpha$  and  $\beta$ ), and it is defined as [27]

$$q_{\alpha\beta} = \frac{1}{N} \sum_i \left( \sigma^{\alpha} + i \sqrt{\frac{t}{\beta(1+t)}} \phi^{\alpha} \right) \left( \sigma^{\beta} + i \sqrt{\frac{t}{\beta(1+t)}} \phi^{\beta} \right), \quad (2.7)$$

with  $p_{\alpha\beta}$  its conjugate variables [20].

Imposing replica symmetry

$$m_1^\alpha = m \quad \forall \alpha, \quad (2.8a)$$

$$q_{\alpha\beta} = Q\delta_{\alpha\beta} + q(1 - \delta_{\alpha\beta}), \quad (2.8b)$$

$$p_{\alpha\beta} = P\delta_{\alpha\beta} + p(1 - \delta_{\alpha\beta}), \quad (2.8c)$$

after straightforward computations, we can finally state the next

**Proposition 1.** *The infinite volume limit of the replica-symmetric free energy for the model (2.1), expressed in terms of the order parameters  $m$  and  $q$ , reads as*

$$\begin{aligned} A(\alpha, \beta, t) = & \frac{m^2}{2(1+t)} \left(1 + \frac{t}{\Delta}\right) + \frac{(1+t)(\Delta-1)}{2t} Q + \frac{\alpha\beta}{2} p(Q-q) \\ & + \frac{\alpha}{2\beta} \left( \log[1 - \beta(1+t)(Q-q)] - \frac{q\beta(1+t)}{1 - \beta(1+t)(Q-q)} \right) + \frac{(1+t)(1-\Delta)}{2t\Delta} \\ & + \frac{\log \Delta}{2\beta} + \frac{\alpha p t}{2(1+t)\Delta} - \frac{1}{\beta} \int Dx \log \cosh \left[ \frac{\beta}{\Delta} (m + \sqrt{\alpha p} x) \right] - \frac{\log 2}{\beta}, \end{aligned} \quad (2.9)$$

where  $Dx$  is the Gaussian measure and  $\Delta = 1 + \alpha\beta t(1+t)^{-1}(P-p)$ .

The self-consistency equations for the model 2.1 are derived by imposing the extremal condition for the free energy 2.9 with respect to the five order parameters, so we arrive at the following

**Proposition 2.** *The self-consistency equations read*

$$m = \frac{1+t}{\Delta+t} \int Dx \tanh \left[ \frac{\beta}{\Delta} (m + \sqrt{\alpha p} x) \right], \quad (2.10a)$$

$$p = \frac{q(1+t)^2}{[1 - \beta(1+t)(Q-q)]^2}, \quad (2.10b)$$

$$\Delta = 1 + \frac{\alpha t}{1 - \beta(1+t)(Q-q)}, \quad (2.10c)$$

$$q = Q + \frac{t}{\beta(1+t)\Delta} - \frac{1}{\Delta^2} \int Dx \cosh^{-2} \left[ \frac{\beta}{\Delta} (m + \sqrt{\alpha p} x) \right], \quad (2.10d)$$

$$Q\Delta^2 = 1 - \frac{t\Delta}{\beta(1+t)} + \frac{\alpha p t^2}{(1+t)^2} - \frac{m^2 t(t+2\Delta)}{(1+t)^2} - \frac{2\alpha\beta p t}{(1+t)\Delta} \int Dx \cosh^{-2} \left[ \frac{\beta}{\Delta} (m + \sqrt{\alpha p} x) \right]. \quad (2.10e)$$

By studying these equations it is possible to derive the phase diagram related to the cost-function (2.1). This point will be achieved in Sec. 2.4.

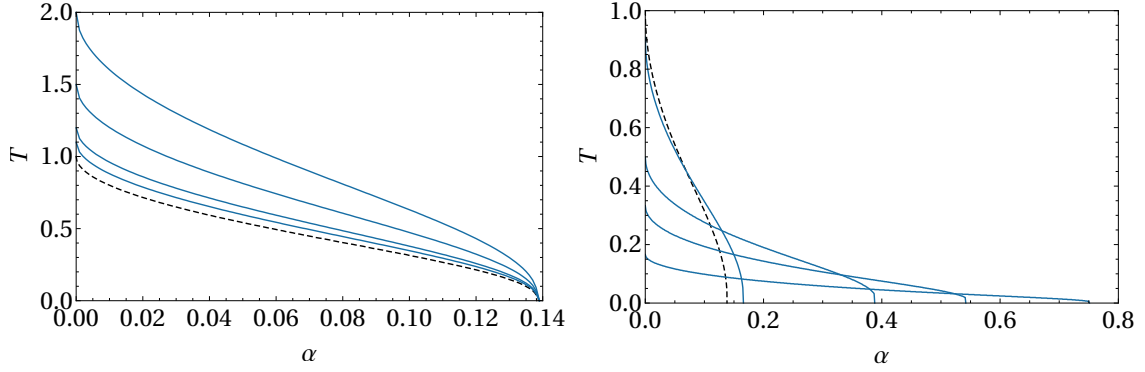
**Remark 2.** For  $t \rightarrow 0$ , both the free energy (2.9) and the self-consistency equations (2.10) reduces to the AGS ones as they should.

## 2.2 Remotion or Reinforcement: a separate analysis

In order to better analyze the structure of our model, we split the whole Hamiltonian (2.1) in two by considering separately the contributions coming from the numerator (reinforcement) and the denominator (remotion) in the generalized kernel. In other words, we take into account the following cost-functions separately to show that, when isolated, none of them constitutes a major breakthrough, that appears solely when these two features are left to work together (as we will prove later).

$$H_{N,P}^{(1)} = -\frac{1}{2} \sum_{\mu} \sum_{ij} \xi_i^{\mu} \xi_j^{\mu} (1+t) \sigma_i \sigma_j, \quad (2.11a)$$

$$H_{N,P}^{(2)} = -\frac{1}{2} \sum_{\mu\nu} \sum_{ij} \xi_i^{\mu} \xi_j^{\nu} (\mathbb{I} + tC)_{\mu,\nu}^{-1} \sigma_i \sigma_j. \quad (2.11b)$$



**Figure 2. Reinforcing and unlearning models.** Left: the plot shows the retrieval regions for the reinforcing model  $H^{(1)}$  for  $t = 0$  (Hopfield), 0.1, 0.2, 0.5 and 1. The critical temperature in the zero-capacity limit is  $T_c = (1 + t)$  and this trivial shift in the critical temperature is the solely novelty of this model. Right: the plot shows the retrieval regions for the Dotsenko model as also discussed in [27]. The critical temperature grows with  $t$ , by the critical temperature in the zero-capacity limit decreases as  $T_c = (1 + t)^{-1}$ , so that the retrieval regions are mashed on the horizontal axes.

- Concerning the first cost-function, due to reinforcement, it is evident that the only net effect it may induce (when playing along) is to stretch the minima landscape by amplifying the energetic gaps by a factor  $(1 + t)$ . The model is formally identical to the Hopfield one with a rescaled thermal noise  $\tilde{\beta} = \beta(1 + t)$ : this implies that the zero-capacity critical temperature is given by  $\tilde{T}_c = \tilde{\beta}_c^{-1} = 1$ , namely  $T_c = (1 + t)$ . See Figure 2 (left panel).
- Concerning the second cost-function, due to mixtures removal, this is precisely the coupling matrix (1.14) whose statistical mechanics has been deeply analyzed in [27] (in the standard replica-symmetric regime). This model emerges as a continuous-time limit of the unlearning procedure analyzed by Plakov and Semenov [55] and it is thus natural to link this model to unlearning features. An important point is that the zero-capacity critical temperature for the transition between the retrieval and spin-glass phases is  $T_c = (1 + t)^{-1}$ , therefore in the large unlearning time limit the former is mashed on the  $\alpha$  axes and, actually, no robustness is retained. See Figure 2 (right panel).

With these ideas in mind, it is also reasonable to expect that - in the full model (2.1) - the mashing effect of unlearning can be compensated by the rescaling of the thermal noise, therefore giving an optimal balance between the Reinforcement and the Removal features. The evaluation of the phase diagram for our model is presented in the next Section.

### 2.3 Zero-temperature (noise-less) critical capacity

The first point we would like to analyze is the critical capacity in the vanishing temperature limit ( $\beta \rightarrow \infty$ ). As standard in this case, it is convenient to introduce  $c \equiv \beta(Q - q)$  quantifying the difference between diagonal and non-diagonal replica overlaps. From Eq. 2.10c it is easy to verify that this quantity satisfies the self-consistency equation

$$c = \frac{\beta}{\Delta^2} \int Dx \cosh^{-2} \left[ \frac{\beta}{\Delta} (m + \sqrt{\alpha p x}) \right] - \frac{t}{(1 + t)\Delta}. \quad (2.12)$$

Using the equation for  $\Delta$  in the zero temperature limit, with simple arguments it is easy to check that  $c$  is finite and, consequently,  $q \rightarrow Q$  as  $T \rightarrow 0$ . Now, since the hyperbolic tangent in (2.10a)

tends to the error function, after some rearrangement we end with the simplified set of equations

$$m = \frac{1+t}{\Delta+t} \operatorname{erf} \left( \frac{m}{\sqrt{2\alpha p}} \right), \quad (2.13a)$$

$$p = \frac{Q(1+t)^2}{[1-(1+t)c]^2}, \quad (2.13b)$$

$$\Delta = 1 + \frac{\alpha t}{1-(1+t)c}, \quad (2.13c)$$

$$c = \frac{1}{\Delta} \sqrt{\frac{2}{\pi\alpha p}} \exp \left( -\frac{m^2}{2\alpha p} \right) - \frac{t}{\Delta(1+t)}, \quad (2.13d)$$

$$Q\Delta^2 = 1 + \frac{\alpha p t^2}{(1+t)^2} - \frac{m^2 t(t+2\Delta)}{(1+t)^2} - \frac{2\alpha t}{1+t} \sqrt{\frac{2}{\pi\alpha p}} \exp \left( -\frac{m^2}{2\alpha p} \right). \quad (2.13e)$$

Introducing the quantities  $\mu = m(2p)^{-1/2}$ ,  $\Pi = p^{-1/2}$  and eliminating  $Q$ , we recast the original set of equations as

$$\mu = \frac{\Pi}{\sqrt{2}} \frac{1+t}{\Delta+t} \operatorname{erf} \left( \frac{\mu}{\sqrt{\alpha}} \right), \quad (2.14a)$$

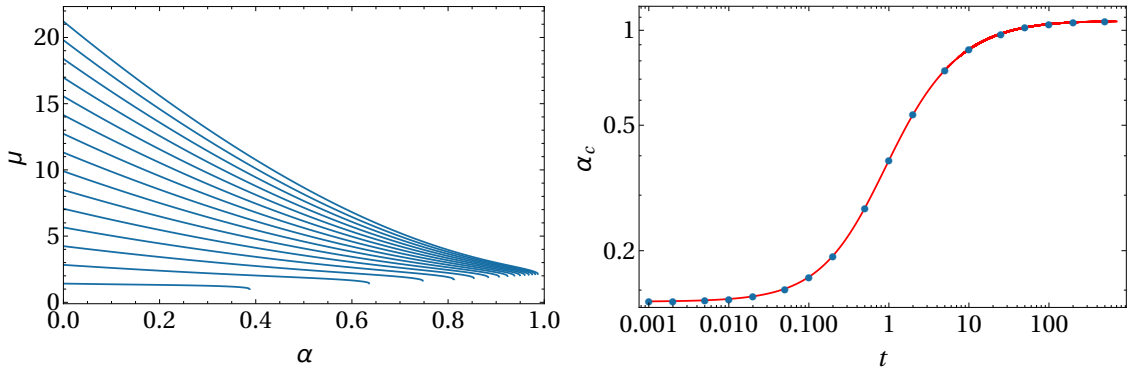
$$\Delta = 1 + \frac{\alpha t}{1-(1+t)c}, \quad (2.14b)$$

$$c = \frac{\Pi}{\Delta} \sqrt{\frac{2}{\pi\alpha}} \exp \left( -\frac{\mu^2}{\alpha} \right) - \frac{t}{\Delta(t+1)}, \quad (2.14c)$$

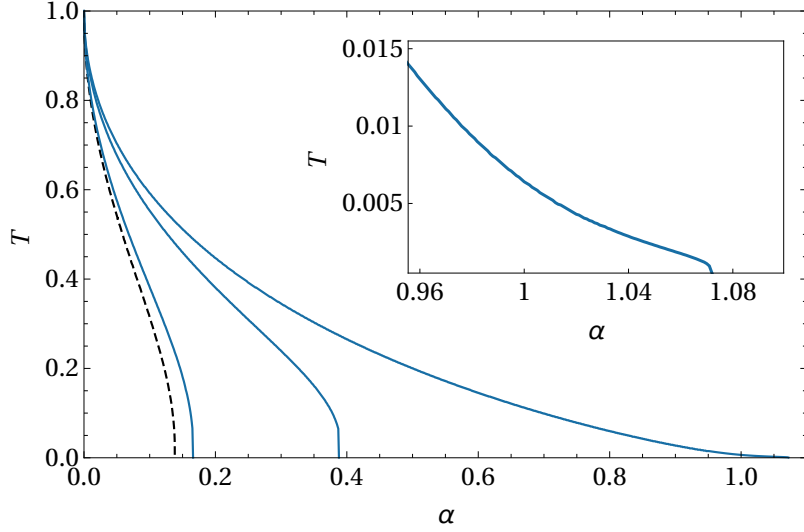
$$\Delta^2 [1-(1+t)c]^2 = \Pi^2 (1+t)^2 + \alpha t^2 - 2\mu^2 t(t+2\Delta) - 2\alpha t(1+t)\Pi \sqrt{\frac{2}{\pi\alpha}} \exp \left( -\frac{\mu^2}{\alpha} \right). \quad (2.14d)$$

Since  $\mu$  is proportional to  $m$  and since  $p$  is finite for any  $t$ , solutions with  $\mu \neq 0$  correspond to retrieval solutions (since  $m \neq 0$ ). Searching for solutions of (2.14) with  $\mu \neq 0$  for given  $t$  is therefore equivalent to determine the upper bound for the storage capacity  $\alpha_c(t)$ . We solved these equations numerically<sup>11</sup> for  $t \in (1, 1000)$  and we reported the solutions in Fig. 3 (left panel). The end point of each curve separates the  $\alpha$  axis in the regions with respectively  $\mu \neq 0$  and  $\mu = 0$ ,

<sup>11</sup>Note that, for  $\alpha \sim 0$ , we have the behaviors  $\Delta \sim 1$ ,  $c \sim -t/(t+1)$ ,  $\mu \sim 2^{-1/2}(1+t)$  and  $\Pi \sim 1+t$ .



**Figure 3. Zero-temperature analysis of the critical capacity.** Left panel: numerical solutions for  $\mu$  of the self-consistency equations in the zero temperature limit (2.14) for several unlearning times:  $t = 1, 3, \dots, 29$ . Right panel: temporal dependence of the critical capacity at zero temperature. The blue dots represent the storage capacity beyond which the only possible solution has  $\mu = 0$ , *i.e.* the end-points of the curves in the left plot). The red curve is the fit given by  $y = x/(x+a)$ , with  $a = 2.84 \pm 0.01$  obtained by first normalizing data in  $[0, 1]$ , namely  $\alpha_c \rightarrow [\alpha_c - \min(\alpha_c)]/[\max(\alpha_c) - \min(\alpha_c)]$ .



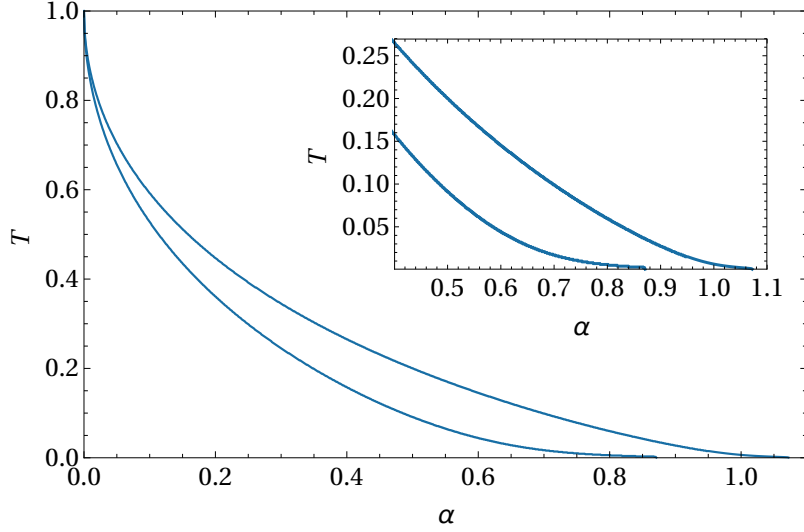
**Figure 4.** Critical line for the transition between retrieval and spin-glass phases for various values of the unlearning time. From the left to the right:  $t = 0$  (Hopfield, black dashed line), 0.1, 1 and 1000. The inner plot on the top-right corner shows the tail of the critical curve for  $t = 1000$ .

therefore identifying the critical capacity for each fixed  $t$  value. We report the critical capacity  $\alpha_c$  as a function of the sleep extent  $t$  in the right plot (blue dots) of Fig. 3. The  $t \rightarrow 0$  limit provides the critical capacity  $\alpha_c(t = 0) \sim 0.138$ , that correctly recovers the standard Hopfield result [10], while in the opposite limit  $t \rightarrow \infty$  we have the upper bound  $\alpha_c \sim 1.07$ , in perfect agreement with [27]. It is worth stressing that the gap between the critical capacity obtained here ( $\alpha_c \sim 1.07$ ) and the maximal critical capacity according to Gardner’s theory ( $\alpha_c = 1.00$ ) should be ascribed to the replica symmetry breaking expected to take place in this model (this was already pointed out in [27]). Interestingly, the critical capacity displays a log-sigmoidal growth in  $t$ . This suggest that the intrinsic scale for  $t$  is logarithmic: relatively small values of  $t$  already provide a critical threshold  $\alpha_c$  close to 1; more precisely,  $\alpha_c(t = 1) \approx 0.4$  and  $\alpha_c(t = 5) \approx 0.8$ . The log-sigmoidal shape also gives hints for convenient choice of the sleep extent: assuming that we want the best possible capable machine, increasing  $t$  is somehow expensive (*e.g.* in terms of time), then the region corresponding to the flex of the curve (approximately  $t = 1$ ) is where a small increase in  $t$  determines the largest return in terms of capacity.

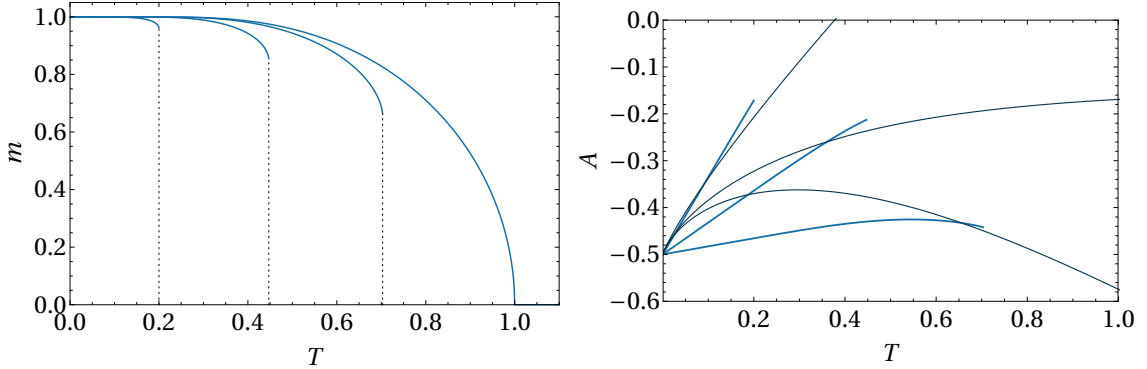
## 2.4 Replica symmetric phase diagram

We solved the set (2.10) of five self-consistency equations numerically for different values of  $t$ . We used these solutions to build the phase diagram depicted in Figs. 4 and 5. A comparison with the phase diagram from AGS theory and Dotsenko *et al.* (Figs. 1 and 2, respectively) shows that our model displays the same qualitative behaviors (*i.e.* spin-glass, mixed retrieval and pure retrieval phases), but their boundary lines significantly depend on  $t$ . In particular, the retrieval region gets wider with  $t$ .<sup>12</sup> More precisely, we distinguish the following transition lines:

<sup>12</sup>On the contrary, in the model studied by Dotsenko *et al.* [27], the area of the retrieval region decreases as  $t$  grows, and vanishes in the large unlearning time limit. This is clear by noticing that the critical capacity at zero thermal noise level and  $t \rightarrow \infty$  reaches the fixed value  $\alpha_c \sim 1.07$ , while the critical temperature at zero capacity is  $T_c = (1 + t)^{-1}$ , so it vanishes in the  $t \rightarrow \infty$ . Since the critical curve characterizing the phase transition to the spin-glass phase is (from thermodynamics argument) a monotonous decreasing as a function  $T(\alpha)$ , it immediately follows that the retrieval region area becomes smaller and smaller for increasing  $t$ .



**Figure 5.** Phase diagram in the large unlearning time limit ( $t = 1000$ ). The two curves trace the boundary of the maximal retrieval regions where patterns are global free energy minima (inner boundary) or local free energy minima (outer boundary). The inner plot on the top-right corner shows the tails of both the critical curves. We stress that, as already pointed out by Dotsenko&Tirozzi, the extension of the retrieval region in the low-temperature regime up to  $\alpha_c \sim 1.07$  is just a chimera of the replica symmetric approximation, while in the true RSB phase  $\alpha_c \rightarrow 1.00$ , according to Gardner’s theory [30].



**Figure 6.** Mattis magnetization and free-energy for  $t = 1000$ . Left: the plot shows the Mattis magnetization  $m$  as a function of the temperature for various storage capacity values ( $\alpha = 0, 0.05, 0.2$  and  $0.5$ , going from the right to the left). The vertical dotted lines indicates the jump discontinuity identifying the critical temperature  $T_c(\alpha)$  which separates the retrieval region from the spin-glass phase. Right: the plot shows the free-energy as a function of the temperature for various storage capacity values ( $\alpha = 0.05, 0.2$  and  $0.5$ , going from the bottom to the top) in the retrieval (thicker light blue lines) and spin-glass (thinner dark blue lines) states.

- *Spin-glass versus Mixed retrieval region.* Here, we focus on the transition between the retrieval region and the spin glass phase, therefore searching for the critical curve  $T_c(\alpha)$  beyond which the only possible solution has  $\mathbf{m} = 0$  with  $q \neq 0$ . The situation we found is formally similar to the original Hopfield model: in the low-storage limit (*i.e.*  $\alpha = 0$ ), the replica-symmetric free-energy is continuous everywhere and differentiable *almost* everywhere (with the only exception being the critical point  $T_c = 1$  as expected, where we have a second-order phase transition in the standard Ehrenfest classification). For higher values of the capacity  $\alpha > 0$ , the phase

transition turns out instead to be of the first kind, with a discontinuity taking place at the critical temperature  $T_c(\alpha)$ . Left plot in Fig. 6 shows an example of this behavior for various values of the storage capacity  $\alpha$  and  $t = 1000$ . The jumps of the magnetization versus  $T$  take place on the critical line separating the retrieval region from the spin glass phase, so that we can study the occurrences of these jumps to reconstruct the phase boundary of the retrieval region. The results have been collected in Figure 4 for various sleep extents. By inspecting the plot, it clearly emerges that the critical storage capacity  $\alpha$  effectively increases with the sleeping session, with the zero-capacity critical temperature  $T_c(\alpha = 0)$  being stable to 1. Thus, our interpolation scheme effectively leads to an increase of the retrieval performances offering a working tradeoff between unlearning spurious memories and consolidating pure ones.

- *Mixed retrieval versus Pure retrieval region.* The region where the pure states are global minima for the free-energy is identified by solving the self-consistency equations (with fixed  $\alpha$  and  $T$ ) for both retrieval ( $m \neq 0$ ) and spin-glass ( $m = 0$ ) states and then comparing the values of the corresponding free-energies. Right plot in Fig. 6 shows the behavior of free-energy for both these solutions for various storage capacity. The intersection point between the corresponding curves identifies the critical temperature  $T_R(\alpha)$  below which the pure states (globally) minimize the free-energy. We performed this analysis for  $t = 1000$ . The result is the phase diagram depicted in Fig. 5.

A visual comparison between the Hopfield phase diagram (Fig. 1) and the present one (Fig. 5) immediately evidences the crucial role of sleeping for improving the network performances.

## 2.5 Numerical results

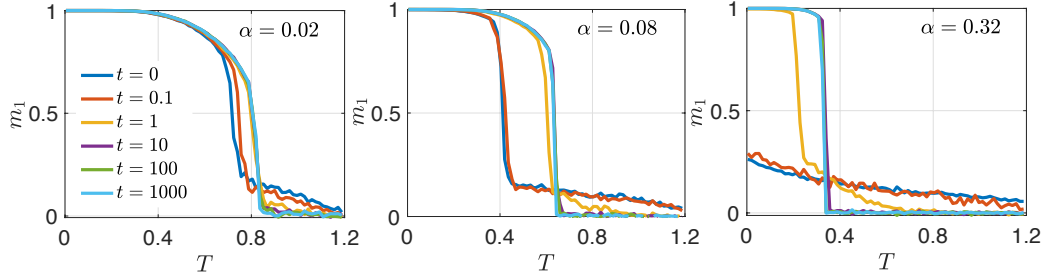
In this Section, we inspect numerically some aspects of the exposed theory which are too difficult to control analytically. In particular, we want to check that our replica-symmetric ansatz is reasonable, by comparing its predictions with Monte Carlo simulations (where no assumptions are made). Then, we want to analyze the field distributions  $h_i$  and the robustness of the attraction basins of the pure minima.

### 2.5.1 Checking the Replica Symmetric assumption

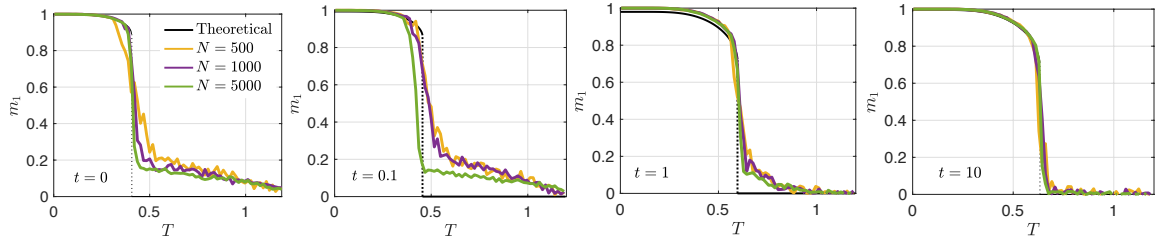
We performed Monte Carlo simulation to mimic the evolution of a finite-size network made of  $N$  neurons and  $P$  patterns. For a given realization of patterns  $\xi_i^\mu$ ,  $i = 1, \dots, N$  and  $\mu = 1, \dots, P$ , for a given temperature  $T = 1/\beta$  and for a given sleeping time  $t$ , we let the system evolve by a single spin-flip Glauber dynamics and, once the equilibrium state is reached,<sup>13</sup> we measure the thermal average of the Mattis magnetization, referred to as  $\bar{\mathbf{m}}$ . This is repeated for  $M$  realizations of the patterns over which thermal averages are accordingly averaged. The resulting value provides our numerical estimate for the Mattis overlaps. Different parameters  $(N, P, \beta, t)$  are considered and, for each choice, the same procedure applies. A sample of results is shown in Fig. 7, where one can check that, as  $t$  increases, the Mattis magnetization  $m_1$  corresponding to the retrieved pattern  $\xi^1$  vanishes at larger values of  $T$  and  $\alpha$  (with a slight abuse of notation here we mean  $\alpha = P/N$ ). Remarkably, these results are also quantitatively consistent with those presented in Fig. 4. Since these results were obtained from simulations at finite size and without asking for replica symmetry, this check strongly corroborates the analytical findings. Further, in Fig. 8 we compare outlines pertaining to systems of different sizes  $N$ , but same choice of  $\beta, t, \alpha$ ; the theoretical expression found in Eq. 2.10 is also depicted. Finite-size effects tend to overestimate the magnetization at temperatures just above the critical one. However, for a system with size  $N = 1000$  the curve is already pretty well overlapped with the theoretical one.

<sup>13</sup>This can be checked by evaluating the stability of observables and the width of their fluctuations





**Figure 7. Results from Monte Carlo simulations.** These panels report the results from Monte Carlo simulations run for different choices of the parameters  $(P, \beta, t)$  and fixing  $N = 5000$  and  $M = 10$ . More precisely,  $1/\beta = T$  ranges from 0 to 1.2,  $P/N = 0.02$  in the leftmost panel,  $P/N = 0.08$  in the middle panel and  $P/N = 0.32$  in the rightmost panel. Also, we considered  $t = 0, 0.1, 1, 10, 100, 1000$ , which are depicted in different colors, as explained in the legend.



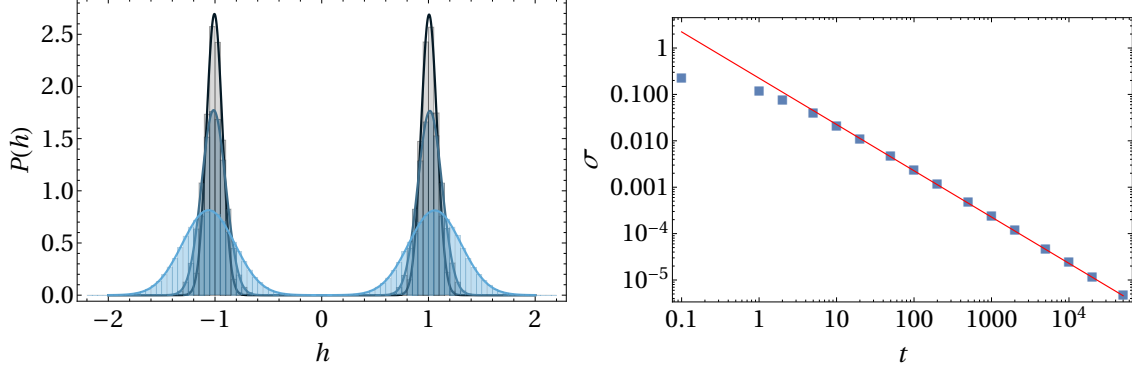
**Figure 8. Finite size scaling.** Average values for the Mattis magnetization  $m_1$  corresponding to the retrieved pattern  $\xi^1$  obtained from numerical simulations for fixed  $\alpha = 0.08$  and  $M = 10$ . Different sizes ( $N = 500, 1000, 5000$ ) are considered and presented in different colors, as explained by the common legend, and are also compared to the theoretical solution reported in Eq. 2.10 and obtained in the thermodynamic limit. Moreover, each panel correspond to a different choice of  $t$ , as reported.

### 2.5.2 Fields distributions in retrieved states

In order to study the internal field distributions characterizing the retrieval mode, we perform extensive Monte Carlo simulations at fixed network size  $N$  and for various sleep extents  $t$ . Since we want to examine the effects of reinforcement and remotion in the retrieval regime, it is convenient to let the network evolve in a point of the tuneable parameters, where retrieval is certainly feasible (namely where pure states dominate the free energy landscape): in the following we will focus on the case  $N = 1000$  and  $P = 50$  with a ratio  $P/N$  well below the theoretical (Hopfield) critical threshold.

A numerical observation is that, as we expect our *unlearning&consolidating* algorithm to clean the free energy landscape from metastabilities, we could be able to avoid sophisticated thermalization techniques (*e.g.*, simulated annealing [43]). Rather, we aim to check directly if already with rudimental minimizers available for the dynamical update of the neurons, the network is still able to reach a global minimum: these simulations are thus carried with standard Glauber dynamics in the  $\beta \rightarrow \infty$  limit, with the expressions for the fields acting over the neurons as prescribed by eq. (2.16). We start the simulations from random initial configurations and simple check that the dynamics ends in a retrieval state. Taking advantage of the mean-field nature of the model, the expression for these fields can be extracted by representing the cost-function (2.1) as

$$H_{N,P}(\sigma|\xi, t) = -\frac{1}{2} \sum_{i=1}^N h_i \sigma_i, \quad (2.15)$$



**Figure 9. Internal fields probability densities for various unlearning time.** Numerical results (histograms) of the Monte Carlo simulations for the internal fields configuration and comparison with best-fitting Gaussian distributions (smooth curves). The values of the unlearning time here considered are  $t = 0$  (standard Hopfield case, in light blue),  $t = 1$  (dark blue) and  $t = 2$  (light gray). The statistics used in numerical simulations consists in 20 different stochastic evolutions (with different random initial conditions) and 20 different realizations of the stored patterns. **Dependence of standard deviation of internal fields distribution on the unlearning time.** The plot shows the standard deviation of the (best-fitting) Gaussian distribution of the internal fields configuration as a function of the unlearning time obtained by the previously described Monte Carlo simulations. The results are average on 20 different stochastic evolutions (with different random initial conditions) and 20 different realizations of the stored patterns for each unlearning time value. The fit returns a power-law scaling as  $\sigma(t) \sim 0.224 \cdot t^{-0.998}$ .

with the internal fields defined as

$$h_i = \frac{1}{N} \sum_{j=1}^N \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu (1+t)(1+tC)_{\mu\nu}^{-1} \sigma_j. \quad (2.16)$$

As stated above, to analyze the internal field configurations, we adopt a standard Glauber dynamics at zero thermal noise level, *i.e.* calling  $\tau$  the neural update time, with the (parallel) update rule

$$\sigma_i(\tau + 1) = \text{sign}[h_i(\tau)], \quad (2.17)$$

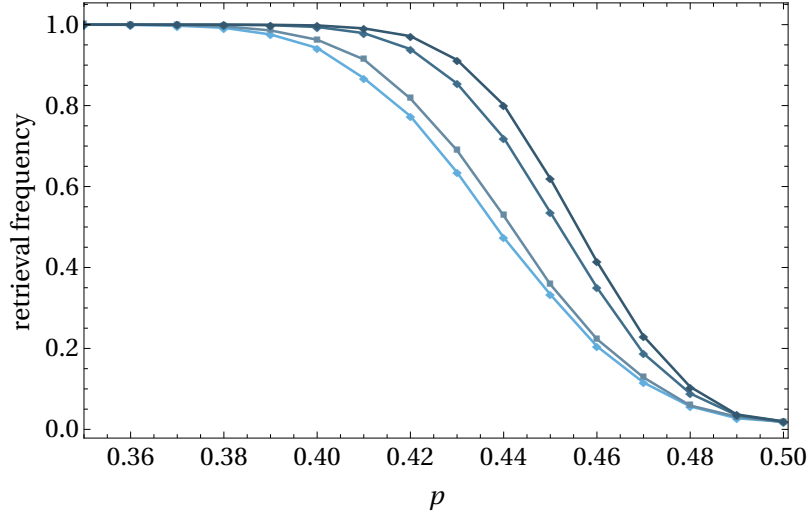
so that the simulations stop when all spins are aligned to the internal fields.<sup>14</sup> From the internal field configurations, we estimated numerically the probability density function  $P(h)$  and compared it to a standard Gaussian distribution: the results are reported in Fig. 9. Remarkably, the fields distribution  $P(h)$  become more and more narrow and peaked as the sleep extent increases. Indeed, the standard deviation  $\sigma_{P(h)}$  scaling as a power law of the dreaming time, *i.e.*  $\sigma_{P(h)} \sim 1/t$ , suggesting that dreaming acts, in this picture, as a regularizer in the internal field distributions. The results supporting this picture are shown in Fig. 9.

### 2.5.3 Retrieval frequency for noisy inputs: on the attraction basins

As a natural successive step, we need to (partially) reintroduce the noise in the network and use it to analyze the depth of the free energy pure minima: the underlying idea is to present some noisy inputs to the network (at various noise intensities  $p$ ) and check the proper signal reconstruction. Otherwise stated, check if, once supplied a noisy input, the network is still able to find its path to the related global free energy minimum accounting for the correct pattern.

<sup>14</sup>Due to detailed balance, convergence of this kind of algorithm is guaranteed for symmetric synaptic couplings [9, 20].

Also in this case, the parameters are fixed to  $N = 1000$  and  $P = 50$ .<sup>15</sup> The procedure we adopted is standard: we prepare the network  $\{\sigma_i\}_{i=1,\dots,N}$  aligning it to the first pattern  $\xi^1$ , then we flip each neuron ( $\sigma_i \rightarrow -\sigma_i$ ) with probability  $p$ . In this way, we construct the initial state of the network. Then, we let the network evolve according to the classical zero-noise Glauber dynamics (see eq. (2.17)) and count how many times the signal is properly reconstructed (in other words, we check that the overlap of the network state with the candidate pattern  $\xi^1$ , *i.e.*  $m_1$  is the maximal Mattis magnetization). The results are plotted in Fig. 10, and show that our algorithm has the effect of making the basins of the pure attractors more stable with the sleeping session: this is intuitively in agreement with the observation that - increasing the sleep extent - the retrieval region becomes larger (w.r.t. the Hopfield reference).



**Figure 10. Analysis of attraction basins.** The plots shows the retrieval frequency as a function of the spin-flip probability for  $t = 0, 0.1, 1$  and  $1000$  (from the left to the right). These results are obtained with 200 different stochastic evolutions for each of the 200 pattern realizations.

<sup>15</sup>We checked that analogous results hold also for (randomly selected) different configurations (whose results we do not report).

### 3 *Unlearning&Consolidating: Focusing on Synapses*

#### 3.1 Time evolution of the synaptic matrix

The Hebbian paradigm can be interpreted as the adiabatic collection of memories sequentially learnt. Along the same line, we show that the coupling (2.1) encoding for reinforcement and removal can be seen as the result of a sequential synaptic updating. To this goal it is convenient to first look at the evolution of the coupling  $J$  as a continuous dynamic process. Later, we will show how to make it discrete and suitable for an iterative implementation.

##### 3.1.1 The continuous algorithm

By construction, the interpolating coupling matrix

$$J_{ij}(t) = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu \left( \frac{1+t}{1+tC} \right)_{\mu\nu}, \quad (3.1)$$

has as limiting cases  $J(0) = J$  and  $J(\infty) = J^p$ . Exploiting the identity

$$\frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu (C^n)_{\mu\nu} = \sum_k J(0)_{ik} (J(0)^n)_{kj}, \quad (3.2)$$

we can recast the time-dependent matrix  $J(t)$  as

$$J(t) = (1+t)J(0)(1+tJ(0))^{-1}. \quad (3.3)$$

Upon differentiating (3.3) with respect to  $t$ , we end with the evolution equation

$$\dot{J} = \frac{1}{1+t}(J - J^2). \quad (3.4)$$

Comparing this matrix ODE with standard unlearning process in the Literature [54–56], we notice two main difference. First, we have a non-trivial dependence on the sleep time through the prefactor  $(t+1)^{-1}$ . Second, there are two contributions in the evolution equation, associated to different scaling (respectively, linear and quadratic in  $J$ ) and opposed signs (mimicing the two opposite features of the algorithm, *i.e.* consolidation and remotion).

##### 3.1.2 The discrete algorithm

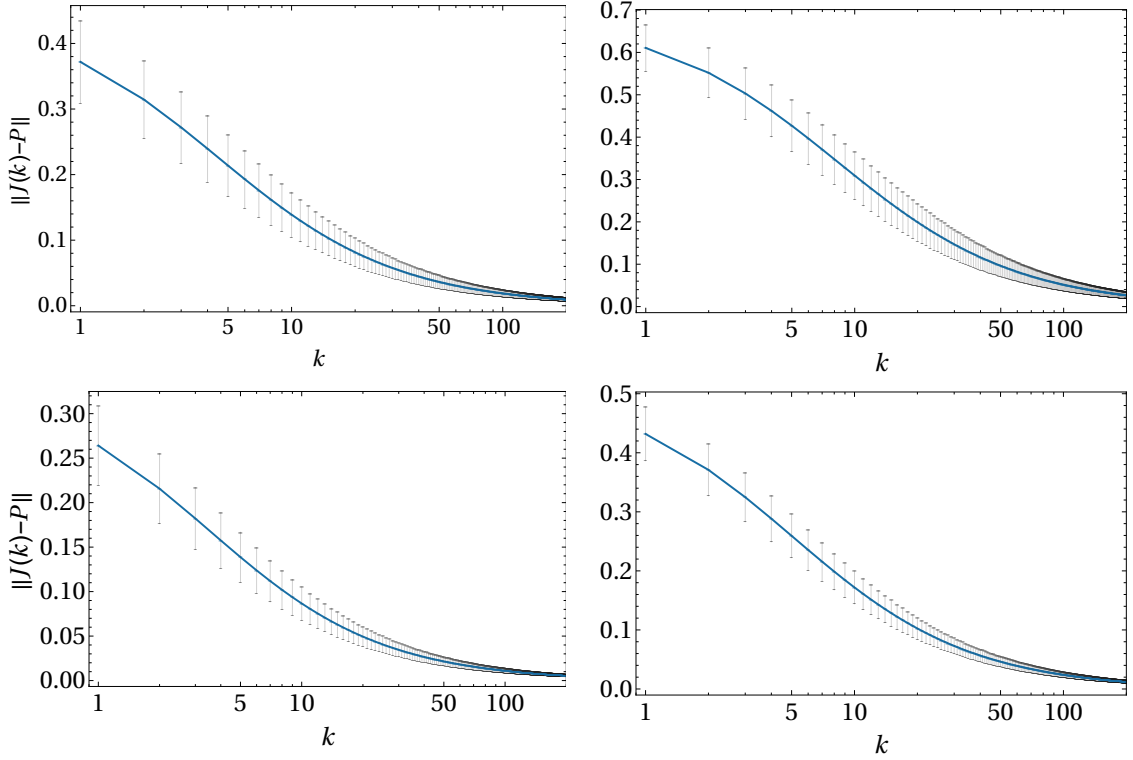
To go to the discrete picture, we have to (re)introduce a tunable parameter  $\epsilon$  (the *unlearning strength* or equivalently *effectiveness of sleep*) defining a temporal scale in which remotion and consolidation are effective. Therefore, we perform the following replacements:

$$\begin{aligned} dt &\rightarrow \epsilon, \\ t &\rightarrow k\epsilon, \\ \dot{J} &\rightarrow \epsilon^{-1}[J(k+1) - J(k)], \end{aligned} \quad (3.5)$$

where  $k$  labels the number of sleeping sessions. Thus, at a discrete level, the *reinforcement&remotion* procedure is provided by the following rule:

$$J(k+1) = J(k) + \frac{\epsilon}{1+\epsilon k}[J(k) - J(k)^2]. \quad (3.6)$$

We stress that this procedure can naturally be interpreted as an *adaptive* unlearning & consolidation scheme, as the *effective* sleep strength (*i.e.* the coefficient of the  $J$ -dependent corrections in eq. (3.6)) does depend on the sleep session  $k$ . With this prescription, the coupling matrix converges



**Figure 11. Unlearning procedures for various networks.** The four plots show the convergence (in the operator norm) of the coupling matrix with the unlearning procedure (3.6) for various network parameters: first line -  $N = 64$  ( $P = 8$  and  $P = 16$ ); second line -  $N = 128$  ( $P = 8$  and  $P = 16$ ). The temporal window shown in the plots is limited to for the first 200 cycles. The results are the average over 500 different realizations of the patterns. The parameter  $\epsilon$  is fixed to 0.5.

to the projection matrix in the limit of infinite dreams, as is clear from Fig. 11. To prove this, it is convenient to write the coupling matrix in the form

$$J_{ij}(k) = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu G_{\mu\nu}(k). \quad (3.7)$$

Then, the above unlearning&consolidation rule can be recast as

$$G(k+1) = \left(1 + \frac{\epsilon}{1 + \epsilon k}\right) G(k) - \frac{\epsilon}{1 + \epsilon k} G(k) C G(k), \quad (3.8)$$

with the initial condition  $G(0) = \mathbb{I}$ . The analytical proof of the convergence of this algorithm is reported in Appendix. B. An important point we would like to stress is that the critical value of the unlearning strength ensuring the convergence can be sharply estimated as (see Appendix B, Corollary 1)

$$\epsilon_c = \frac{1}{\|C\| - 1}. \quad (3.9)$$

This equality is very instructive, since it states that the critical strength for the synaptic update is fixed by the magnitude of the patterns' correlations: the stronger the correlations, the longer the amount of time required to the sleep for optimizing the network's free energy landscape.

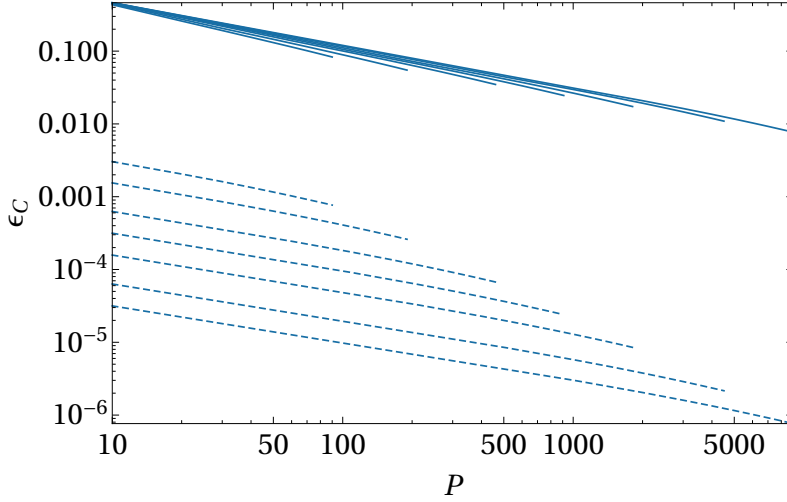
We also notice that the singular case  $\|C\| = 1$  is excluded, since it only happens if all the eigenvalues of  $C$  are 1, meaning that  $C = \mathbb{I}$  (or equivalently, all patterns are uncorrelated). In the

latter case, no unlearning is needed, since the starting point is exactly the solution of the above recurrence relation.<sup>16</sup>

It is interesting to compare these results with the unlearning procedure analyzed in [55], for which the critical unlearning strength is given by

$$\epsilon_c = (N\|J(0)\|)^{-1}, \quad (3.10)$$

where  $J(0)$  is the Hopfield coupling matrix. By inspecting at Figure 12, it is clear that, in our case, the critical unlearning strength is higher than the usual one (3.10) by many order of magnitudes, and (in the low storage case, *i.e.*  $P \ll N$ ) it is independent on  $N$ . Another important difference between the two algorithms is that, while for (3.10) and fixed  $P$  the critical strength fastly decreases as  $N$  grows, in our case it slowly increases for higher network size. Thus, our method appears to be more stable with respect to the network size.



**Figure 12. Dependence of the critical unlearning strength versus  $P$  and  $N$ .** The plot shows the values of the critical unlearning strength as functions of the number of neurons  $N$  and the number of stored patterns  $P$ . Each curve corresponds to a fixed value of  $N = 100, 200, 500, 1000, 2000, 5000$  and  $10000$ . The solid blue lines are the critical strength 3.9 of our unlearning procedure 3.6. The dashed blue lines are instead associated to the critical strength (3.10). The results are averaged on 50 different realizations of the patterns (the statistical errors are not reported since they are too small to be visualized in the log-log scale).

## 4 Conclusions

Inspired by information optimization during sleep episodes in mammal’s brains, in this paper we study Hebbian unlearning with reinforcement, namely we discuss how the Hopfield model (where the bare Hebb prescription is forecasted) can be generalized to better optimize its resources, namely, in order to have the most possible robustness w.r.t. (fast/thermal) noise and the larger possible capacity (*i.e.* the ratio among the stored patterns and the neurons available to handle them).

Oversimplifying, during human’s sleep, two main types of dreaming alternate, namely the slow-wave and the random-eye-movement phases, with two coupled -but different- purposes: while they share the final goal of achieving best possible optimization of information storage, the former contributes to the scope by consolidating important memories (that we match with the patterns in

<sup>16</sup>We stress, however, for identically distributed random boolean patterns for finite  $N$ , the correlation is always presented with a rough estimation  $\sim 1/\sqrt{N}$ .

the AI counterpart played by associative neural networks), the latter instead gets rid of the -by far more abundant- unimportant memories (that we match with spurious/mixture states in the AI counterpart played by associative neural networks).

To account for both these features at once in AI too, we proposed a novel *unlearning&consolidating* algorithm that we ideally use to stylize a dream (with the same spirit by which a neuron is reduced to a Boolean variable in mathematical modeling) and we tested it on the standard reference provided by the Hopfield model. We stress that, while in Hopfield networks just pairwise correlations are stored, the present theory can be applied in a much broader generality, e.g. for instance extending the cost-function to account for P-spin higher-order contributions [13, 45, 49] for deep nets too.

Our algorithm has solely one novel parameter, accounting for the time the network spent in dreaming, and we studied how both the neural performances as well as the synaptic couplings evolve as this parameter is tuned. As far as the neurons are concerned, as a result of our procedure, at the amount of dreams increases we obtain a significant improvement in the critical capacity that, at first (mainly thanks to discarding spurious states), in the zero fast noise limit, increases from  $\alpha_c \sim 0.14$  to  $\alpha_c \sim 1$  (that is the maximal capacity if the network is equipped with symmetric couplings, as prescribed by the Gardner theory [30, 67]), further (mainly due to the reinforcement term), pure memories remain stable even against high level of (fast) noise. Indeed we inspected how the fields acting on the neurons gets affected by the dreams and, as the dreaming time increases, the fields get better and better peaked over pattern's entries -getting rid of the noise- and, remarkably, their standard deviations  $\sigma$  have a power-law scaling with the dreaming time, i.e.  $\sigma \propto t^{-1}$ . It is also worth pointing out that network performances increase in a high non-linear way with the dreaming time (such that with a few cycles a massive optimization has already been achieved and there is no longer need to reach un-physical epochs).

This gets crystal clear when focusing on the synapses as, already an elementary glance at their dynamical evolution, suggests that the -starting with standard pairwise Hopfield- the dynamics forces the Hebbian kernel to match the projection matrix, close to the scenario pictured by Kanter and Sompolinsky [41]. We confirmed this statement both analytically and numerically and we found a sharp estimate for the optimal *dreaming rate* -the analogous of a learning or unlearning rate(s) in existing Literature.

Finally we aim to notice that there is also another important reason to investigate these improvements over the standard scenario in Hebbian machines: there is a one-to-one correspondence among Hopfield networks and restricted Boltzmann machines [3, 14] thus, as Boltzmann machines are the building blocks in modern Deep Learning architectures [46, 59], increasing efficiency in the former may imply progress even in the latter, and ultimately in Deep Learning (as we know that modern machines trained with deep learning actually do dream of electric sheeps [32]<sup>17</sup>).

## Acknowledgements

A.F. and A.B. acknowledge Salento University, MIUR (through basic funding to the Italian research) and INFN for partial support.

A.B. also acknowledges the grant *Rete Match: Progetto Pythagoras (CUP:J48C17000250006)*.

E.A. acknowledges the grant *Progetto Ateneo (RG11715C7CC31E3D)* from Sapienza University of Rome.

E.A., A.B. and A.F. are grateful to GNFM-INdAM for partial financial support.

---

<sup>17</sup>Further, this interpretation of sleep&dream raises as a stand-alone alternative against the Freudian psychoanalysis, as brilliantly pointed out by Christos in [19], but in this manuscript this point will not be deepened.



## A Calculations to obtain the replica symmetric solution

In this Appendix, we report in some detail the replica trick<sup>18</sup> calculations necessary to get an explicit expression, in terms of the order parameters, of the (replica-symmetric) free energy of the model. We start with the (quenched average of the) replicated partition function (2.5), which we rewrite here as

$$\begin{aligned} \mathbb{E} Z_{N,P}(\sigma|\xi, t)^n = & \sum_{\sigma^1} \cdots \sum_{\sigma^n} \int \left( \prod_{\alpha} D z_1^{\alpha} \right) \left( \prod_{i\alpha} D \phi_i^{\alpha} \right) \exp \left[ \sqrt{\frac{\beta(t+1)}{N}} \sum_{i\alpha} z_1^{\alpha} \xi_i^1 \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \right] \\ & \cdot \int \left( \prod_{\alpha} \prod_{\mu \geq 2} D z_1^{\alpha} \right) \mathbb{E}' \exp \left[ \sqrt{\frac{\beta(t+1)}{N}} \sum_{i\alpha} \sum_{\mu \geq 2} z_{\mu}^{\alpha} \xi_i^{\mu} \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \right]. \end{aligned} \quad (\text{A.1})$$

Note that we separated the signal term (associated to pattern  $\xi^1$ , meant to be retrieved) and the slow noise (constituted by all the other not-retrieved patterns, whose random similarities with  $\xi^1$  -i.e. the spurious correlations this paper is due to- lie at core-genesis of such a slow noise). The exponential with noisy contributions in the second line can be easily rewritten (neglecting sub-leading contributions in the large  $N$  limit) as

$$\begin{aligned} \mathbb{E} \exp \left[ \sqrt{\frac{\beta(t+1)}{N}} \sum_{i\alpha} \sum_{\mu \geq 2} z_{\mu}^{\alpha} \xi_i^{\mu} \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \right] = & \int \prod_{\alpha\beta} dq_{\alpha\beta} \frac{N dp_{\alpha\beta}}{2\pi} \\ & \cdot \exp \left[ iN \sum_{\alpha\beta} p_{\alpha\beta} q_{\alpha\beta} + \frac{\beta(t+1)}{2} \sum_{\alpha\beta} \sum_{\mu \geq 2} z_{\mu}^{\alpha} z_{\mu}^{\beta} q_{\alpha\beta} - i \sum_{\alpha\beta} p_{\alpha\beta} \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \left( \sigma_i^{\beta} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\beta} \right) \right], \end{aligned}$$

where we imposed the definition of overlap (2.7) through the insertion of a Dirac delta (in its Fourier representation, as standard [20]). We can then perform the Gaussian integration over the order parameters  $z_{\mu}^{\alpha}$  which are not associated to the retrieved pattern, so to obtain

$$\begin{aligned} \mathbb{E} Z_{N,P}(\sigma|\xi, t)^n = & \sum_{\sigma^1} \cdots \sum_{\sigma^n} \int \left( \prod_{\alpha} D z_1^{\alpha} \right) \left( \prod_{i\alpha} D \phi_i^{\alpha} \right) \left( \prod_{\alpha\beta} dq_{\alpha\beta} \frac{N dp_{\alpha\beta}}{2\pi} \right) \exp \left[ iN \sum_{\alpha\beta} p_{\alpha\beta} q_{\alpha\beta} - \frac{p}{2} \log \det(\mathbb{I} - \beta(1+t)\hat{q}) \right. \\ & \left. + \sqrt{\frac{\beta(t+1)}{N}} \sum_{i\alpha} z_1^{\alpha} \xi_i^1 \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) - i \sum_{\alpha\beta} p_{\alpha\beta} \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \left( \sigma_i^{\beta} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\beta} \right) \right]. \end{aligned} \quad (\text{A.2})$$

We replace the order parameter  $z_1^{\alpha}$  with the corresponding (replicated) Mattis magnetization by using the relation  $z_1 = \sqrt{\beta N(1+t)^{-1}} m_1^{\alpha}$  (see also [27, 28]). We also make the convenient redefinition of the conjugated overlap  $p_{\alpha\beta} \rightarrow i \frac{\alpha\beta^2}{2} p_{\alpha\beta}$ . After some trivial rearrangements, we get

$$\begin{aligned} \mathbb{E} Z_{N,P}(\sigma|\xi, t)^n = & \int \left( \prod_{\alpha} \sqrt{\frac{\beta N}{2\pi(1+t)}} dm_1^{\alpha} \right) \left( \prod_{\alpha\beta} dq_{\alpha\beta} \frac{iN\alpha\beta^2 dp_{\alpha\beta}}{4\pi} \right) \exp \left\{ -\frac{\beta N}{2} \sum_{\alpha} \frac{m_1^{\alpha 2}}{1+t} \right. \\ & - \frac{N\alpha\beta^2}{2} \sum_{\alpha\beta} p_{\alpha\beta} q_{\alpha\beta} - \frac{\alpha N}{2} \log \det(\mathbb{I} - \beta(1+t)\hat{q}) + \log \sum_{\sigma^1} \cdots \sum_{\sigma^n} \int \left( \prod_{i\alpha} D \phi_i^{\alpha} \right) \\ & \left. \cdot \exp \left[ \beta \sum_{i\alpha} m_1^{\alpha} \xi_i^1 \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) + \frac{\alpha\beta^2}{2} \sum_{\alpha\beta} p_{\alpha\beta} \left( \sigma_i^{\alpha} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\alpha} \right) \left( \sigma_i^{\beta} + i \sqrt{\frac{t}{\beta(t+1)}} \phi_i^{\beta} \right) \right] \right\}. \end{aligned} \quad (\text{A.3})$$

<sup>18</sup>A solid mathematical ground for the replica trick in the Sherrington-Kirkpatrick model for spin-glasses is already available (see e.g., [18]), while in the Hopfield model for neural networks this is only partially available (see e.g., [8]).

The last line can be easily handled and its terms rearranged in order to remove the site index (*i.e.* the subscript  $i$ ) from the spins  $\sigma$  and the auxiliary Gaussian fields  $\phi$ . Moreover, since in the thermodynamic limit the “hergodic” equality

$$\log \prod_i f(\xi_i^1) = \sum_i \log f(\xi_i^1) = N \mathbb{E} f(\xi), \quad (\text{A.4})$$

holds [9, 20], we can easily represent the replicated partition function in the form

$$\mathbb{E} Z_{N,P}(\sigma|\xi, t)^n = \int d\mu(m_1^\alpha, q_{\alpha\beta}, p_{\alpha\beta}) e^{-\beta N n A}, \quad (\text{A.5})$$

$d\mu$  being the measure over all the order parameters and

$$\begin{aligned} A(\alpha, \beta, t) = & \frac{1}{2n(1+t)} \sum_\alpha m_1^{\alpha 2} + \frac{\alpha\beta}{2n} \sum_{\alpha\beta} p_{\alpha\beta} q_{\alpha\beta} + \frac{\alpha}{2n\beta} \log \det(\mathbb{I} - \beta(1+t)\hat{q}) \\ & - \frac{1}{n\beta} \mathbb{E} \log \sum_\sigma \int \left( \prod_\alpha D\phi^\alpha \right) \exp \left[ \beta \sum_\alpha m_1^\alpha \xi^1 \left( \sigma^\alpha + i \sqrt{\frac{t}{\beta(1+t)}} \phi^\alpha \right) + \right. \\ & \left. + \frac{\alpha\beta^2}{2} \sum_{\alpha\beta} p_{\alpha\beta} \left( \sigma^\alpha + i \sqrt{\frac{t}{\beta(1+t)}} \phi^\alpha \right) \left( \sigma^\beta + i \sqrt{\frac{t}{\beta(1+t)}} \phi^\beta \right) \right]. \end{aligned} \quad (\text{A.6})$$

being the general free-energy of the model, see (2.6).

Imposing the replica symmetric ansatz and recalling the definition of  $\Delta$ , we can compute the replica-symmetric free energy  $A(\alpha, \beta, t)$  term by term:

- $\frac{1}{2n} \sum_\alpha \frac{m_1^{\alpha 2}}{1+t} = \frac{m_1^2}{2(1+t)},$
- $\frac{\alpha\beta}{2n} \sum_{\alpha\beta} p_{\alpha\beta} q_{\alpha\beta} = \frac{(\Delta-1)(1+t)}{2t} Q + \frac{\alpha\beta}{2} p(Q-q),$
- $\frac{\alpha}{2n\beta} \log \det[\mathbb{I} - \beta(1+t)\hat{q}] = \frac{\alpha\beta}{2} \left( \log[1 - \beta(1+t)(Q-q)] - \frac{q\beta(1+t)}{1 - \beta(1+t)(Q-q)} \right) + \mathcal{O}(n),$
- $\frac{1}{n\beta} \mathbb{E} \log \sum_\sigma \int \left( \prod_\alpha \phi^\alpha \right) [\dots] = \frac{1+t}{2t} \frac{\Delta-1}{\Delta} - \frac{1}{2\beta} \log \Delta - \frac{\alpha p t}{2\Delta(1+t)} - \frac{t}{2\Delta(1+t)} m_1^2 + \frac{1}{\beta} \log 2$   
 $+ \frac{1}{\beta} \int Dx \log \cosh \left[ \frac{\beta}{\Delta} (m_1 + \sqrt{\alpha p x}) \right].$

Putting all pieces together and taking the limit  $n \rightarrow 0$ , after some rearrangements we arrive at the free energy expression (2.9).

## B Convergence of the discrete algorithm: the analytical proof

In this Appendix, we prove the convergence of the unlearning rule (3.6) toward the inverse correlation matrix  $C^{-1}$ . In doing this, we will follow a procedure which is very close to the route paved in [54].

The norm in the matrix vector space we used is the *operator norm*, which means that

$$\|A\| = \sqrt{\max\{a | a \in \sigma(A^T A)\}}, \quad (\text{B.1})$$

where  $\sigma(A^T A)$  is the spectrum of the matrix  $A^T A$ . Note that, since we will deal only with symmetric and positive-definite matrices, this definition reduces to  $\|A\| = \max\{a | a \in \sigma(A)\}$ . In what follows, we will often use the notations

$$q_k = \frac{1 + \epsilon k}{1 + \epsilon(k+1)}, \quad p_k = \frac{\epsilon}{1 + \epsilon(k+1)}. \quad (\text{B.2})$$

**Proposition 3.** *The norm of correlation matrix is greater than one:  $\|C\| \geq 1$ .*

*Proof.* By definition, the diagonal entries of the correlation matrix are all equal to 1, so that

$$\text{Tr } C = p. \quad (\text{B.3})$$

Moreover, the correlation matrix  $C$  is symmetric and positive-definite. Then, all eigenvalues  $\gamma_\mu$  are clearly positive, and we have

$$\sum_{\mu} \gamma_{\mu} = p. \quad (\text{B.4})$$

Since the number of the eigenvalues is precisely  $p$ , it is impossible to saturate the trace equality with  $\gamma_{\mu} < 1$  for all  $\mu$ . Since the largest eigenvalue is equal to the matrix norm, it follows that  $\|C\| \geq 1$ .  $\square$

**Proposition 4.** *The matrix  $G(k)$  commutes with  $C$  for all  $k$ .*

*Proof.* We define two new matrix types by multiplying  $G(k)$  on the right and on the left with  $C$ , i.e.  $T^{(1)}(k) = G(k)C$  and  $T^{(2)}(k) = CG(k)$ . Since  $G(0) = \mathbb{I}$ , then  $T^{(1,2)} = C$ . It's easy to see from (3.8) that both  $T^{(1)}$  and  $T^{(2)}$  satisfy the same recursion relation, and since they have the same initial condition, it follows that  $T^{(1)}(k) = T^{(2)}(k)$  for all  $k$ , which means that  $G(k)C = CG(k)$  proving the statement.  $\square$

**Proposition 5.** *The matrices  $G(k)$  are invertible for all  $k \geq 0$  and  $\epsilon < \epsilon_c$  for some  $\epsilon_c$ .*

*Proof.* The statement is trivial for  $k = 0$ , since  $G(0) = \mathbb{I}$ . Then we can prove the proposition by induction. Assume that  $G(k)$  is invertible. Then we rewrite the matrix recursion relation by multiplying both side with  $C$  (by previous proposition, it's not important if on the left or on the right), then

$$\begin{aligned} T(k+1) &= \left(1 + \frac{\epsilon}{1 + \epsilon k}\right) T(k) - \frac{\epsilon}{1 + \epsilon k} T(k)^2 = \\ &= q_k^{-1} T(k) [\mathbb{I} - p_k T(k)], \end{aligned} \quad (\text{B.5})$$

with  $T(k) = CG(k)$  (therefore with the initial condition  $T(0) = C$ ). Taking the determinant of both sides and using the Binet theorem, we have

$$\det T(k+1) = q_k^{-p} \det T(k) \cdot \det [\mathbb{I} - p_k T(k)]. \quad (\text{B.6})$$

But now

$$\det [\mathbb{I} - p_k T(k)] = \exp \log \det [\mathbb{I} - p_k T(k)] = \exp \text{Tr} \log [\mathbb{I} - p_k T(k)]. \quad (\text{B.7})$$

We can expand in series the logarithm of the matrix:

$$\log [\mathbb{I} - p_k T(k)] = - \sum_{n=1}^{\infty} \frac{p_k^n}{n} T^n(k), \quad (\text{B.8})$$

which converges for  $\|p_k T(k)\| = \|p_k CG(k)\| < 1$  for all  $k$  (note that this condition imposes the constraint for  $\epsilon$  to be less than a critical value  $\epsilon_c$ , but we postpone this discussion). With the convergence ensured, it follows that

$$\exp \text{Tr} \log [\mathbb{I} - p_k T(k)] = \exp \text{Tr} \left( - \sum_{n=1}^{\infty} \frac{p_k^n}{n} T^n(k) \right) > 0. \quad (\text{B.9})$$

Since all terms on the r.h.s of (B.5) are non-vanishing it follows that  $\det T(k+1) = \det C \det G(k) \neq 0$ , which implies that  $\det G(k) \neq 0$  since  $C$  is invertible. Then, also  $G(k)$  is invertible for all  $k$ .  $\square$

**Proposition 6.** *The matrices  $T(k)$  and  $G(k)$  are positive-definite for all  $k$ .*

*Proof.* Again, the statement is trivial for  $k = 0$ . Then, we will prove the proposition inductively. Suppose that all  $G(l)$  are positive-definite for  $l = 0, \dots, k$ . Since all  $G$ s commutes with  $C$ , then also  $T(l)$  are positive-definite for  $l = 0, \dots, k$ . Since  $T(k)$  is invertible for each  $k$ , we can take the inverse of Eq. (B.5):

$$T^{-1}(k+1) = q_k [\mathbb{I} - p_k T(k)]^{-1} T(k)^{-1}. \quad (\text{B.10})$$

But now

$$[\mathbb{I} - p_k T(k)]^{-1} T(k)^{-1} = \sum_{n=0}^{\infty} p_k^n T(k)^n T(k)^{-1} = \sum_{n=0}^{\infty} p_k^n T(k)^{n-1}, \quad (\text{B.11})$$

again converging for  $\|p_k T(k)\| = \|p_k C G(k)\| < 1$  for all  $k$ . In a more transparent form we have

$$T^{-1}(k+1) = q_k T^{-1}(k) + q_k p_k \mathbb{I} + q_k \sum_{n=2}^{\infty} p_k^n T(k)^{n-1}. \quad (\text{B.12})$$

Under the hypothesis of convergence,  $T^{-1}(k+1)$  is therefore a (infinite) sum of positive-definite matrices, then it is positive-definite by itself. Then, since the inverse of a positive-definite matrix is itself positive-definite, the same result holds for  $T(k+1)$ . But  $G(k+1) = T(k+1)C^{-1}$ , and since both  $C^{-1}$  and  $T(k+1)$  are positive-definite and commute (the proof is straightforward), then also the product  $T(k+1)C^{-1}$  is positive-definite, inductively proving the proposition.  $\square$

At this point, we are ready to prove the

**Lemma 1.** *For each  $k$ , there can be found a finite real number  $c_k$  which is greater or equal to  $\|CG(k)\|$ . As a consequence, the sequence is bounded from above by  $\bar{c} = \max_k c_k$ .*

*Proof.* Applying iteratively the recurrence relation (B.12) and recalling that  $T(0) = C$ , it's easy to show that

$$T^{-1}(k) = \frac{C^{-1}}{1 + \epsilon k} + N_{k-1} \mathbb{I} + R(k-1), \quad (\text{B.13})$$

where the rest operator  $R(k)$  is defined as

$$R(k-1) = \sum_{l=0}^{k-1} \frac{1 + \epsilon l}{1 + \epsilon k} \sum_{n=2}^{\infty} p_l^n T(l)^{n-1}, \quad (\text{B.14})$$

and

$$N_k = \sum_{l=0}^k p_l \prod_{s=l}^k q_s. \quad (\text{B.15})$$

The latter is an increasing function with  $k$  with values in the range  $[0, 1]$  and such that  $N_k \underset{k \rightarrow \infty}{\sim} 1$ . Moreover, it's clear that  $N_{-1} = 0$ . Since  $T(k)$  is real and symmetric matrix, it can be diagonalized with eigenvalues  $\tau_\mu(k)$  (which are positive since it is also positive-definite). Then, the spectrum of the inverse  $T^{-1}(k)$  consists in the values  $\sigma[T^{-1}(k)] = \{\tau_\mu^{-1}(k) | \tau_\mu(k) \in \sigma[T(k)]\}$ . Then, the minimum of the spectrum of  $T^{-1}(k)$  is clearly  $\|T(k)\|^{-1}$ . Therefore, taking the minimal eigenvalue of equation (B.13), we have

$$\|T(k)\|^{-1} = \min \sigma \left[ \frac{C^{-1}}{1 + \epsilon k} + N_{k-1} \mathbb{I} + R(k-1) \right] \geq \min \sigma \left[ \frac{C^{-1}}{1 + \epsilon k} + N_{k-1} \mathbb{I} \right], \quad (\text{B.16})$$

since also the rest operator is a positive-definite operator (and as a consequence, its contribution to the spectrum is positive). With the same reasoning, the quantity on the r.h.s. is nothing but

$$\min \sigma \left[ \frac{C^{-1}}{1 + \epsilon k} + N_{k-1} \mathbb{I} \right] = \frac{\|C\|^{-1}}{1 + \epsilon k} + N_{k-1}. \quad (\text{B.17})$$

Then, by taking the inverse of the previous inequality, we get

$$\|T(k)\| \leq \left( \frac{\|C\|^{-1}}{1 + \epsilon k} + N_{k-1} \right)^{-1} = \frac{\|C\|}{\frac{1}{1 + \epsilon k} + N_{k-1}\|C\|} = c_k. \quad (\text{B.18})$$

This proves our assertion.  $\square$

A remark here is that the inequality is indeed an equality for  $k = 0$ , since  $T(k) = C$ . This is important in the following

**Corollary 1.** *The critical value of the unlearning strength  $\epsilon_c$  is fixed by the norm of correlation matrix.*

*Proof.* We recall that, in order to have a convergent algorithm, the unlearning strength  $\epsilon$  has to satisfy the criterion  $p_k \|T(k)\| < 1$  for all  $k$ . By using the previous Lemma, we see that

$$p_k \|T(k)\| \leq \frac{\epsilon \|C\|}{\frac{1 + \epsilon(k+1)}{1 + \epsilon k} + N_{k-1}[1 + \epsilon(k+1)]\|C\|}. \quad (\text{B.19})$$

It is important to notice that the denominator in this inequality is an increasing function of  $k$ . This means that, if the unlearning strength is chosen to have  $p_0 \|T(0)\| < 1$ , then it would valid for all  $k$ . But, from our previous consideration (for  $k = 0$  it is an equality)

$$p_0 \|T(0)\| = \frac{\epsilon \|C\|}{1 + \epsilon} < 1, \quad (\text{B.20})$$

meaning that the unlearning algorithm converges if and only if  $\epsilon < \epsilon_c = \frac{1}{\|C\| - 1}$ .  $\square$

Once all of these results have been proved, we will finally prove that the unlearning algorithm converges (for  $\epsilon < \epsilon_c$ ) to the desired solution  $G(k) \rightarrow C^{-1}$ . This will be proved by norm estimation in the large  $k$  limit.

**Theorem 1** (Convergence). *The unlearning algorithm (3.8) converges to the stationary solution  $G(\infty) = C^{-1}$  in the sense defined by the operator norm.*

*Proof.* Let us start again with the equality

$$T^{-1}(k) = \frac{C^{-1}}{1 + \epsilon k} + N_{k-1} \mathbb{I} + R(k-1). \quad (\text{B.21})$$

The first two terms on the r.h.s. have simple contributions in the large  $k$  limit. In particular, the first one has a vanishing norm for  $k \rightarrow \infty$ , so it would not contribute to the final solution. We have now to evaluate the norm of the rest operator:

$$\begin{aligned} \|R(k-1)\| &\leq \sum_{l=0}^{k-1} \frac{1 + \epsilon l}{1 + \epsilon k} \left\| \sum_{n=2}^{\infty} p_l^n T(l)^{n-1} \right\| \leq \sum_{l=0}^{k-1} \frac{1 + \epsilon l}{1 + \epsilon k} \sum_{n=2}^{\infty} p_l^n \|T(l)\|^{n-1} = \\ &= \sum_{l=0}^{k-1} p_l \frac{1 + \epsilon l}{1 + \epsilon k} \sum_{n=1}^{\infty} p_l^n \|T(l)\|^n = \frac{\epsilon}{1 + \epsilon k} \sum_{l=0}^{k-1} \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|}, \end{aligned} \quad (\text{B.22})$$

since  $p_l \|T(l)\| < 1$ . To evaluate the last sum in this equation we adopt a counting argument by analyzing each factor. For the first one:

$$\frac{1 + \epsilon l}{1 + \epsilon(l+1)} \sim \mathcal{O}(l^0), \quad (\text{B.23})$$

for large enough  $l$ . For the second factor:

$$\begin{aligned}
& \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \leq \\
& \leq \epsilon \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{\|C\|}{1 + \|C\| N_{l-1}(1 + \epsilon l)} \left( 1 - \epsilon \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{\|C\|}{1 + \|C\| N_{l-1}(1 + \epsilon l)} \right)^{-1} = \\
& = \frac{\epsilon(1 + \epsilon l)\|C\|}{[1 + \epsilon(l+1)][1 + \|C\| N_{l-1}(1 + \epsilon l)]} \sim \mathcal{O}(l^{-1}).
\end{aligned} \tag{B.24}$$

Then, globally we have

$$\frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \sim \mathcal{O}(l^{-1}). \tag{B.25}$$

To evaluate the behavior for  $k \rightarrow \infty$ ,<sup>19</sup> we then take a large - but finite - integer  $\bar{l}$  such that  $1 \gg \bar{l} \gg k$  and split the sum as

$$\begin{aligned}
& \sum_{l=0}^k \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \\
& = \sum_{l=0}^{\bar{l}-1} \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} + \sum_{l=\bar{l}}^k \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|}.
\end{aligned} \tag{B.26}$$

Since  $\bar{l}$  is large, the terms in the second sum are well-approximated with  $l^{-1}$  (corrections are subleading in  $l$ ), while the first sum is a finite number:

$$\sum_{l=0}^k \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \sim \text{finite contributions} + H_k - H_{\bar{l}-1}, \tag{B.27}$$

where  $H_s = \sum_{l=1}^s l^{-1}$  is the harmonic number. Since  $\bar{l}$  is finite, the term  $H_{\bar{l}-1}$  can be incorporated in the finite contributions, leaving only with

$$\sum_{l=0}^k \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \sim \text{finite contributions} + H_k. \tag{B.28}$$

It is now well-known that the asymptotical behavior of  $H_k$  for large  $k$  is  $H_k \sim \mathcal{O}(\log k)$ . As a consequence, we find that the leading contribution goes as

$$\frac{\epsilon}{1 + \epsilon k} \sum_{l=0}^k \frac{1 + \epsilon l}{1 + \epsilon(l+1)} \frac{p_l \|T(l)\|}{1 - p_l \|T(l)\|} \sim \mathcal{O}(\log k/k) + \text{subleading contributions}, \tag{B.29}$$

and therefore  $\|R(k-1)\|$  vanishes in the  $k \rightarrow \infty$  limit. By these norm estimation and since  $N_k \underset{k \rightarrow \infty}{\sim} 1$ , we can therefore conclude that

$$T^{-1}(k) \underset{k \rightarrow \infty}{\sim} \mathbb{I}. \tag{B.30}$$

Recalling that  $T(k) = CG(k)$ , it immediately follows that  $G(k) \rightarrow C^{-1}$  in the large  $k$  limit, as claimed.  $\square$

---

<sup>19</sup>In this limit, we can replace  $k-1$  in the sum simply with  $k$ .

## References

- [1] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cognitive Sci. **9**, 147, (1985).
- [2] T. Andrillon, D. Pressnitzer, D. Leger, S. Kouider, *Formation and suppression of acoustic memories during human sleep*, Nature Comm. **8**, 179, (2018).
- [3] E. Agliari, et al., *Multitasking associative networks*, Phys. Rev. Lett. **109**, 268101, (2012).
- [4] E. Agliari, et al., *Immune networks: multitasking capabilities near saturation*, J. Phys. A **46**, 415003, (2013).
- [5] E. Agliari, et al., *Neural Networks retrieving binary patterns in a sea of real ones*, J. Stat. Phys. **168**, 1085, (2017).
- [6] E. Agliari, et al., *Multitasking attractor networks with neuronal threshold noises*, Neural Networks **49**, 19, (2013).
- [7] E. Agliari, et al., *Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines*, Neural Networks **38**, 52, (2013).
- [8] E. Agliari, A. Barra, B. Tirozzi, *Free energies of Boltzmann Machines: self-averaging, annealed and replica symmetric approximations in the thermodynamic limit*, arXiv:1810.11075.
- [9] D.J. Amit, *Modeling brain functions*, Cambridge Univ. Press (1989).
- [10] D. Amit, H. Gutfreund, H. Sompolinsky, *Spin-glass models of neural networks*, Phys. Rev. A **32**, 1007, (1985).
- [11] D. Amit, H. Gutfreund, H. Sompolinsky, *Storing infinite numbers of patterns in a spin-glass model of neural networks*, Phys. Rev. Lett. **55**, 1530, (1985).
- [12] E. Aserinsky, N. Kleitman, *Regularly occurring periods of eye motility, and concomitant phenomena, during sleep*, Science **118**, 273, (1953).
- [13] A. Barra, M. Beccaria, A. Fachechi, *A new mechanical approach to handle generalized Hopfield neural networks*, Neur. Net. **106**, 205 (2018).
- [14] A. Barra, et al., *On the equivalence among Hopfield neural networks and restricted Boltzmann machines*, Neural Networks **34**, 1-9, (2012).
- [15] A. Barra, et al., *Phase transitions of Restricted Boltzmann Machines with generic priors*, Phys. Rev. E **96**, 042156, (2017).
- [16] A. Barra, et al., *Phase Diagram of Restricted Boltzmann Machines & Generalized Hopfield Models*, Phys. Rev. E **97**, 022310, (2018).
- [17] A. Barra, G. Genovese, F. Guerra, *Equilibrium statistical mechanics of bipartite spin systems*, J. Phys. A **44**, 245002, (2011).
- [18] A. Barra, F. Guerra, E. Mingione, *Interpolating the Sherrington-Kirkpatrick replica trick*, Phil. Mag. **92**, 78 (2011).
- [19] C.A. Christos, *Investigation of the Crick-Mitchinson reverse-learning dream sleep hypothesis in a dynamical setting*, Neural Net. **9**(3):427-434, (1996).
- [20] A.C.C. Coolen, R. Kuhn, P. Sollich, *Theory of neural information processing systems*, Oxford Press (2005).
- [21] A.C.C. Coolen, D. Sherrington, *Dynamics of fully connected attractor neural networks near saturation*, Phys. Rev. Lett. **71**(23):3886, (1993).
- [22] F. Crick, G. Mitchinson, *The function of dream sleep*, Nature **304**, 111, (1983).
- [23] P. Dayan, B.W. Balleine, *Reward, motivation, and reinforcement learning*, Neuron **36**, 285, (2002).



- [24] B. Derrida, E. Gardner, A. Zippelius, *An exactly solvable asymmetric neural network model*, Europhys. Lett. **4**, 167, (1987).
- [25] S. Diekelmann, J. Born, *The memory function of sleep*, Nature Rev. Neuroscience **11**(2):114, (2010).
- [26] V. Dotsenko, *An introduction to the theory of spin glasses and neural networks*, World Scientific, (1995).
- [27] V. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, *Statistical mechanics of Hopfield-like neural networks with modified interactions*, J. Phys. A **24**, 2419, (1991).
- [28] V. Dotsenko, B. Tirozzi, *Replica symmetry breaking in neural networks with modified pseudo-inverse interactions*, J. Phys. A **24**, 5163, (1991).
- [29] R.M. French, *Catastrophic forgetting in connectionist networks*, Trends in cognitive sciences **3**, 128, (1999).
- [30] E. Gardner, *The space of interactions in neural network models*, J. Phys. A **21**, 257, (1988).
- [31] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, M.I.T. press (2017).
- [32] A. Hern, *Yes, androids do dream of electric sheep*, The Guardian, Technology and Artificial Intelligence (2015).
- [33] G. Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*, available at <http://learning.cs.toronto.edu/>, (2010).
- [34] J.A. Hobson, E.F. Pace-Scott, R. Stickgold, *Dreaming and the brain: Toward a cognitive neuroscience of conscious states*, Behavioral and Brain Sciences **23**, (2000).
- [35] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. **79**, 2554, (1982).
- [36] J.J. Hopfield, D.I. Feinstein, R.G. Palmer, *Unlearning has a stabilizing effect in collective memories*, Nature Lett. **304**, 280158, (1983).
- [37] J.J. Hopfield, D.W. Tank, *Neural computation of decisions in optimization problems*, Biol. Cybernet. **52**, 141, (1985).
- [38] J.A. Horas, P.M. Pasinetti, *On the unlearning procedure yielding a high-performance associative memory neural network*, J. Phys. A **31**, L463-L471, (1998).
- [39] E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. **106**, 620, (1957).
- [40] L.P. Kaelbling, M.L. Littman, A.W. Moore, *Reinforcement learning: A survey*, J. Artif. Intel. Res. **4**:237-285, (1996).
- [41] I. Kanter, H. Sompolinsky, *Associative recall of memory without errors*, Phys. Rev. A **35**:1:380, (1987).
- [42] W. Kinzel, M. Oppel, *Dynamics of learning*, in: E. Domany, J.L. van Hemmen, K. Schulten (Eds.) *Models of neural networks*, Springer, Berlin, 149-172 (1991).
- [43] S. Kirkpatrick, et al., *Optimization by simulated annealing*, Science **220**, 671, (1983).
- [44] T.O. Kohonen, *Self-organization and Associative Memory*, Springer, Berlin (1984).
- [45] D. Krotov, J.J. Hopfield, *Dense associative memory is robust to adversarial inputs*, arXiv:1701.00939, (2017).
- [46] Y. Le Cun, Y. Bengio, G. Hinton, *Deep learning*, Nature **521**, 436, (2015).
- [47] E. Marinari, *Forgetting Memories and their Attractiveness*, arXiv:1805.12368, (2018).
- [48] J.L. McGaugh, *Memory - a century of consolidation*, Science **287**:5451:248-251, (2000).
- [49] P. Mehta, D.J. Schwab, *An exact mapping between the variational renormalization group and deep learning*, arXiv:1410.3831, (2014).

- [50] K. Nokura, *Spin glass states of the anti-Hopfield model*, J. Phys. A **31**, 7447, (1998).
- [51] K. Nokura, *Paramagnetic unlearning in neural network models*, Phys. Rev. E **54**(5):5571, (1996).
- [52] G. Parisi, *A memory which forgets*, J. Phys. A **19**, L617, (1986).
- [53] L. Personnaz, I. Guyon, G. Dreyfus, *Information storage and retrieval in spin-glass like neural networks*, J. Phys. Lett. **46**, L-359:365, (1985).
- [54] A. Y. Plakhov, *The converging unlearning algorithm for the Hopfield neural network: optimal strategy*, IEEE Int. Conf. on Pattern Recognition Vol. 2-Conference B: Computer Vision & Image Processing (1994).
- [55] A. Y. Plakhov, S.A. Semenov, *The modified unlearning procedure for enhancing storage capacity in Hopfield network*, IEEE Trans. **242**, (1992).
- [56] A. Y. Plakhov, S.A. Semenov, I.B. Shuvalova, *Convergent unlearning algorithm for the Hopfield neural network*, IEE Comp. Soc. Press. **2**(95), 30, (1995).
- [57] J. Paton, et al., *The primate amygdala represents the positive and negative value of visual stimuli during learning*, Nature **439**:7078:865, (2006).
- [58] B. Rasch, J. Born, *About sleep's role in memory*, Physiol. Rev. **93**:681-766, (2013).
- [59] R. Salakhutdinov, G. Hinton, *Deep Boltzmann machines*, Artificial Intelligence and Statistics (2009).
- [60] R. Salakhutdinov, H. Larochelle, *Efficient learning of deep Boltzmann machines*, Proc. thirteenth int. conf. on artificial intelligence and statistics, 693, 2010.
- [61] E. Schneidman, M.J. Berry II, R. Segev, M. Bialek, *Weak pairwise correlations imply strongly correlated network states in a neural population* Nature **440**, 1007, (2006).
- [62] N. Srivastava, R. Salakhutdinov, *Multimodal learning with deep Boltzmann machines*, Adv. Neural Inform. Proc. Sys. , 2222, (2012).
- [63] R. Stickgold, J.A. Hobson, R. Fosse, M. Fosse, *Sleep, Learning and Dreams: Off-line Memory Reprocessing*, Science **294**, 1052, (2001).
- [64] R.S. Sutton, A.G. Barto, *Reinforcement learning: An introduction*, MIT press, (1998).
- [65] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models*, Springer Science & Business Media, (2003).
- [66] J. Tübiana, R. Monasson, *Emergence of Compositional Representations in Restricted Boltzmann Machines*, Phys. Rev. Lett. **118**, 138301, (2017).
- [67] S. Wimbauer, J. Leo van Hemmen, *Hebbian unlearning*, Analysis of Dynamical and Cognitive Systems, Springer, Berlin, 1995.