LETTER Communicated by Alexander Mozeika

From Pavlov Conditioning to Hebb Learning

Elena Agliari
elena.agliari@uniroma1.it
Miriam Aquaro
miriam.aquaro@uniroma1.it
Sapienza University of Rome, Department of Mathematics, 00185, Rome, Italy, and Istituto Nazionale d'Alta Matematica, 00185, Rome, Italy

Adriano Barra
adriano.barra@unisalento.it
Istituto Nazionale d'Alta Matematica, 00185, Rome, Italy, and University of Salento, Lecce 73100, Italy

Alberto Fachechi
alberto.fachechi@uniroma1.it
Chiara Marullo
chiara.marullo@uniroma1.it
Sapienza University of Rome, Department of Mathematics, 00185, Rome, Italy, and Istituto Nazionale d'Alta Matematica, 00185, Rome, Italy

Hebb's learning traces its origin in Pavlov's classical conditioning; however, while the former has been extensively modeled in the past decades (e.g., by the Hopfield model and countless variations on theme), as for the latter, modeling has remained largely unaddressed so far. Furthermore, a mathematical bridge connecting these two pillars is totally lacking. The main difficulty toward this goal lies in the intrinsically different scales of the information involved: Pavlov's theory is about correlations between concepts that are (dynamically) stored in the synaptic matrix as exemplified by the celebrated experiment starring a dog and a ringing bell; conversely, Hebb's theory is about correlations between pairs of neurons as summarized by the famous statement that *neurons that fire together wire together*. In this letter, we rely on stochastic process theory to prove that as long as we keep neurons' and synapses' timescales largely split, Pavlov's mechanism spontaneously takes place and ultimately gives rise to synaptic weights that recover the Hebbian kernel.

1 Introduction

The cell assembly theory, introduced in 1949 by Donald Hebb in his milestone *The Organization of Behaviour* (Hebb, 1949), looks at learning

Neural Computation **35, 930–957** (2023)
https://doi.org/10.1162/neco_a_01578

processes as collective features of neurons and synapses and can be effectively summarized by the famous prescription, "Neurons that fire together wire together." Since then, Hebb's learning has become a prolific cornucopia for modeling neural networks, (Coolen et al., 2005; Nishimori, 2001). In particular, the Hopfield model of neural network (Amit et al., 1985), where the Hebbian scheme is implemented mathematically, constitutes a paradigmatic model that has been attracting much attention since the mid-1980s, following the seminal investigation by Amit et al. (1985) based on statistical-mechanics tools. Indeed, after the breakthroughs by Parisi (1979) and coworkers (Mézard et al., 1984, 1987) as the 1970s turned into the 1980s, statistical mechanics of disordered systems was established as a theoretical reference where information processing capabilities of neural networks could be fruitfully inspected. Along the way, important achievements have been made (see Amit, 1989; Coolen et al., 2005), and now, Hebb's rule is usually the starting point when getting acquainted with (bio-inspired) neural networks and related modeling.

Hebb's rule did not appear out of the blue; rather, it was an enlightened synthesis of decades of meticulous inspections in neurophysiology and psychology, whose origins can be traced back to the systematic studies of Ivan Pavlov on classical conditioning,[1] which greatly inspired Donald Hebb.[2] Despite this closeness, the two theories—Hebb's learning based on neural activity correlation and Pavlov's learning based on conditioning by stimuli—had quite distant fates, at least in the context of theoretical modeling: the former has been by far much more represented in the mathematics and physics literature than the latter (one should still recall significant methodological contributions to classical conditioning by Estes (1950), Hull and Spence (Vogel et al., 2004), and modern extensions toward a generalized conditioning by Rescorla-Wagner (1972) and Mackintosh (1975); however, these are behavioral models rather than effective mathematical models).

We aim to provide a mathematical description of classical conditioning (and a glance at its generalized extension), whence one could appreciate how Pavlov implies Hebb: this bridge is the main focus of this work.[3] Establishing such a mathematical connection between Pavlov's and Hebb's theories requires a smooth interpolation from correlations at the neural level up to correlations among concepts, thus merging neurophysiological with behavioral perspectives. In fact, as anticipated, Hebb's learning rule suggests that if two adjacent neurons are emitting spikes, in a temporally correlated

[1] It is worth noting that in roughly the same years, Thorndike (1911) and Twitmyer (1905) understood the core idea behind classical conditioning, but they drew only minor attention from the scientific community at that time.

[2] It is reported that Boris Babkin, a student of Pavlov, spurred Donald Hebb to deepen classical conditioning.

[3] Inspiring ideas about this connection between Pavlov and Hebb from a modeling perspective are due to Francesco Guerra (Agliari, Barra, Gervasi Vidal, et al., 2012; Guerra & D'Autilia, 1996).

manner, the wire connecting them should be strengthened; Pavlov's learning rule, however, connects concepts, somehow preserving the same meaning: if the two stimuli are simultaneously presented for a long enough time (e.g., for a time window larger than the characteristic synaptic timescale), then the synapses (linking the involved neurons) rearrange in order to store this correlation (like food presentation and bell ringing in the famous dog example). Thus, in our framework, we ultimately need to interpolate different degrees of information, from simple bits managed by single neurons (Hebb) to patterns of bits, codifying for concepts as handled by neural circuits (Pavlov).

In general, this procedure is not unique and can be achieved in countless ways. Here we present a simple scheme whose strength is its versatility as it can be broadly adaptable to different contexts. We think of the neurons as the nodes of a network whose weighted links represent the synapses: the fact that solely by assuming that neural and synaptic timescales are quite spread apart suffices to obtain classical conditioning as a spontaneous phenomenon makes this model a rather general one, not confined to neural assemblies. This observation also suggests that large, frustrated[4] networks of interacting units can exhibit computational capabilities, as typically happens broadly in the biological world, from the intercellular level (e.g., lymphocyte networks within the immune system; Parisi, 1990) to the intracellular level (e.g., gene regulatory networks within the cell; Osella et al., 2011), for invertebrates (Canew et al., 1981) as well as vertebrates (Gormezano et al., 1962), and even for prokaryotes as well as eucaryotes (Kramar & Karen, 2021).

The letter is structured as follows. In section 2, we present the theoretical framework to embed our model (we derive a set of evolution equations for the mean neural activities and the synaptic arrangement), and we outline the numerical setup of the experiments that we use to prove our claim. In section 3, we study the related dynamics recovering classical conditioning for "simple" concepts" (*vide infra*) and show that the synaptic matrix accounting for the Pavlov mechanism actually displays a Hebb-like shape. Section 4 offers discussion and outlooks, and technical details and further insights are provided in the appendixes.

2 Neural and Synaptic Dynamics Derived from Statistical Mechanics

2.1 Theoretical Setting. We consider a network made up of N binary neurons $\sigma_i = \pm 1$ with $i = 1, \ldots, N$ and interacting each other through symmetric 2-body couplings of strength J_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, N$.

[4] With the term *frustrated* we mean that the nodes in the network can be cohesive but also competitive as the links connecting them can assume both positive and negative values, as, for instance, excitatory versus inhibitory synapses in neural networks, eliciting or suppressive cytokines in lymphocyte networks (Agliari et. al., 2013; Germain, 2001; Parisi, 1990).

The space of neural configurations is denoted as $\Sigma_N = \{-1, +1\}^N$, and each configuration of the system is represented by the corresponding point $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_N\} \in \Sigma_N$. We denote with the matrix $\boldsymbol{J} \in \mathbb{R}^{N\times N}$ the synaptic configuration and with $\mathcal{J}_N$ the related space, namely, the space of symmetric matrices with null diagonal and size $N \times N$. We assume that the network cost function $H_N(\boldsymbol{\sigma}, \boldsymbol{J})$ admits a two-body Hamilton representation:

$$H_N(\boldsymbol{\sigma}, \boldsymbol{J}) = -\frac{1}{2}\sum_{i\neq j=1}^{N,N} J_{ij}\sigma_i\sigma_j - u\sum_{i=1}^{N} h_i\sigma_i, \tag{2.1}$$

where $h_i \in \{-1, 0, +1\}$ is the bias (or *firing threshold* in more biological terms) acting on each neuron and u is a global amplification factor (the strength of the bias).

Once the Hamilton function is introduced, we can equip the configuration space $\Sigma_N \times \mathcal{J}_N$ with a Boltzmann-Gibbs measure,

$$\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}) = \frac{1}{Z_{N,\beta}} \exp\big[-\beta H_N(\boldsymbol{\sigma}, \boldsymbol{J})\big], \tag{2.2}$$

where

$$Z_{N,\beta} = \sum_{\boldsymbol{J}\in\mathcal{J}_N}\sum_{\boldsymbol{\sigma}\in\Sigma_N} \exp\big[-\beta H_N(\boldsymbol{\sigma}, \boldsymbol{J})\big] \tag{2.3}$$

is the associated partition function and where we introduced the parameter β to rule its broadness, namely, the noise level affecting the neuron dynamics. As $\beta \to 0$, the distribution tends to be flat, and each neuron is randomly oriented, while as $\beta \to \infty$, the distribution is sharply peaked at the configurations, which minimize the Hamiltonian. Note that while in the statistical mechanical formalization of neural networks (e.g., when performing pattern recognition (Amit, 1989; Coolen et al., 2005)), the synapses are quenched and therefore the sum over $\boldsymbol{J} \in \mathcal{J}_N$ does not appear in the definition, equation 2.3, of the partition function $Z_{N,\beta}$. Here we are interested in the coupled dynamical evolution of neurons $\boldsymbol{\sigma}$ and synapses $\boldsymbol{J}$. As a consequence, couplings can not be assumed as quenched variables, and the partition function also has to involve the summation in the related space, assuming the above annealed expression (Mézard et al., 1987).

Hereafter we derive the evolution equations for the expectation values of σ_i and J_{ij} over time by exploiting the master equation of their stochastic process to show that neural dynamics driven by (coupled) external stimuli gives rise to classical conditioning in the synaptic space and the long-term relaxation for synapses matches the mathematical formulation of Hebb's learning rule.

Let $w_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J} \to \boldsymbol{\sigma}', \boldsymbol{J}')$ be the transition rate between two system states $(\boldsymbol{\sigma}, \boldsymbol{J})$ and $(\boldsymbol{\sigma}', \boldsymbol{J}')$, and $\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t)$ the probability for the system to be in the configuration $(\boldsymbol{\sigma}, \boldsymbol{J})$ at time t. Then the master equation that describes the evolution of the probability $\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t)$ over time is

$$\frac{d}{dt}\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t) = \sum_{\boldsymbol{\sigma}' \in \Sigma_N} \sum_{\boldsymbol{J}' \in \mathcal{J}_N} \big[\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}) - P_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t) w_{N,\beta}(\boldsymbol{J}, \boldsymbol{\sigma} \to \boldsymbol{J}', \boldsymbol{\sigma}')\big]. \tag{2.4}$$

Under parallel dynamics (where neural activities and entries of the synaptic matrix are updated simultaneously), the transition rate $w(\boldsymbol{\sigma}', \boldsymbol{J}' \to \boldsymbol{\sigma}, \boldsymbol{J})$ can be factorized as

$$w_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}' \to \boldsymbol{\sigma}, \boldsymbol{J}) = \prod_{l,m} \frac{\exp(\beta J_{lm}\sigma'_l\sigma'_m)}{2\cosh\left(\beta\sigma'_l\sigma'_m\right)} \prod_i \frac{1}{2}\left[1 + \sigma_i \tanh\left(\beta \sum_j J'_{ij}\sigma'_j + \beta h_i\right)\right]. \tag{2.5}$$

By using the following relation,

$$\frac{\exp(\beta J_{ij}\sigma'_i\sigma'_j)}{2\cosh\left(\beta\sigma'_i\sigma'_j\right)} = \frac{1}{2}\left[1 + \tanh\left(\beta J_{ij}\sigma'_i\sigma'_j\right)\right]. \tag{2.6}$$

equation 2.5 can be rewritten as

$$w_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}' \to \boldsymbol{\sigma}, \boldsymbol{J}) = \prod_{l,m} \frac{1}{2}\left[1 + \tanh\left(\beta J_{lm}\sigma'_l\sigma'_m\right)\right] \prod_i \frac{1}{2}\left(1 + \sigma_i \tanh\left(\beta \sum_j J'_{ij}\sigma'_j + \beta h_i\right)\right). \tag{2.7}$$

Notice that the transition rate is thus a product of single-variable conditional probabilities.

2.1.1 Neural Dynamics. We introduce the mean neural activities as

$$\bar{\sigma}_i := \langle \sigma_i \rangle = \sum_{\boldsymbol{\sigma} \in \Sigma_N} \sum_{\boldsymbol{J} \in \mathcal{J}_N} \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t)\sigma_i, \text{ for } i = 1, \ldots, N, \tag{2.8}$$

and we look for the explicit dynamical equation ruling its evolution. To this aim, we set the typical timescale for neural dynamics to be τ, and we multiply the master equation 2.4, by σ_i. Then we sum over all possible neural and synaptic configurations to obtain the following stochastic process:

$$\tau \frac{d\bar{\sigma}_i}{dt} = + \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i [\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma})] - \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i [\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t) w_{N,\beta}(\boldsymbol{J}, \boldsymbol{\sigma} \to \boldsymbol{J}', \boldsymbol{\sigma}')]. \tag{2.9}$$

We observe that, by definition,

$$\sum_{\boldsymbol{\sigma}' \in \Sigma_N} \sum_{\boldsymbol{J}' \in \mathcal{J}_N} w_{N,\beta}(\boldsymbol{J}, \boldsymbol{\sigma} \to \boldsymbol{J}', \boldsymbol{\sigma}') = 1, \tag{2.10}$$

thus,

$$\tau \frac{d\bar{\sigma}_i}{dt} = + \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}) - \sum_{\boldsymbol{J} \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma} \in \Sigma_N} \sigma_i \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t). \tag{2.11}$$

By using the definition of $\bar{\sigma}_i$ the previous equation turns into

$$\tau \frac{d\bar{\sigma}_i}{dt} = -\bar{\sigma}_i + \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}), \tag{2.12}$$

and by making explicit use of the transition rate, the second term of the right-hand side of equation 2.12 reads as

$$\sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}) = = \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \sigma_i \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) \prod_{l,m} \frac{1}{2} \left[1 + \tanh\left(\beta J_{lm} \sigma'_l \sigma'_m\right)\right] \cdot \cdot \prod_k \frac{1}{2} \left[1 + \sigma_k \tanh\left(\beta \sum_j J'_{kj} \sigma'_j + \beta h_k\right)\right]. \tag{2.13}$$

Now, since

$$\prod_{l,m} \sum_{J_{lm}=\pm 1} \frac{1}{2} \left[1 + J_{lm} \tanh\left(\beta \sigma'_l \sigma'_m\right)\right] = 1 \tag{2.14}$$

and

$$\prod_{k\neq i}\sum_{\sigma_k=\pm 1}\frac{1}{2}\left[1+\sigma_k\tanh\left(\beta\sum_j J'_{kj}\sigma'_j+\beta h_k\right)\right]=1, \tag{2.15}$$

equation 2.13 becomes

$$\begin{aligned}
&\sum_{\boldsymbol{J},\boldsymbol{J}'\in\mathcal{J}_N}\sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'\in\Sigma_N}\sigma_i\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}',\boldsymbol{J}',t)w_{N,\beta}(\boldsymbol{J}',\boldsymbol{\sigma}'\to\boldsymbol{J},\boldsymbol{\sigma})=\\
&=\sum_{\boldsymbol{\sigma}'\in\Sigma_N}\sum_{\boldsymbol{J}'\in\mathcal{J}_N}\sum_{\sigma_i=\pm 1}\frac{1}{2}\left[\sigma_i+\tanh\left(\beta\sum_j J'_{ij}\sigma'_j+\beta h_k\right)\right]=\\
&=\sum_{\boldsymbol{\sigma}'\in\Sigma_N}\sum_{\boldsymbol{J}'\in\mathcal{J}_N}\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}',\boldsymbol{J}',t)\tanh\left(\beta\sum_j J'_{ij}\sigma'_j+\beta h_i\right),
\end{aligned} \tag{2.16}$$

such that, in turn, equation 2.12 gets

$$\tau\frac{d\bar{\sigma}_i}{dt}=-\bar{\sigma}_i+\left\langle\tanh\left(\beta\sum_j J_{ij}\sigma_j+\beta h_i\right)\right\rangle. \tag{2.17}$$

Finally, by using the mean field approximation, equation 2.17 can be rewritten as

$$\tau\frac{d\bar{\sigma}_i}{dt}=-\bar{\sigma}_i+\tanh\left(\beta\sum_j\langle J_{ij}\rangle\bar{\sigma}_j+\beta h_i\right). \tag{2.18}$$

Note that while at the single-neuron level, neural firing is typically described by Langevin equations (Tuckwell, 1988), at this averaged level of description, the evolution of the mean firing rate is ruled by an ODE: the random terms are averaged away, yet the presence of the noise β remains such that for $\beta\to 0$, neither the postsynaptic potentials nor the external stimuli can be perceived, and the solely leakage term $-\bar{\sigma}_i$ in equation 2.18 ensures the mean firing rate to rest to zero.

2.1.2 Synaptic Dynamics. We set the typical timescale for synaptic dynamics as $\tau'>\tau$. In order to derive the dynamical equation for evolution of the mean synaptic weights,

$$\bar{J}_{ij}:=\langle J_{ij}\rangle=\sum_{\boldsymbol{\sigma}\in\Sigma_N}\sum_{\boldsymbol{J}\in\mathcal{J}_N}\mathcal{P}_{N,\beta}(\boldsymbol{\sigma},\boldsymbol{J},t)J_{ij},$$

we multiply the master equation by J_{ij} and sum over all possible neural configurations and synaptic matrices, obtaining

$$\tau' \frac{d\bar{J}_{ij}}{dt} = + \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} J_{ij} \left[\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}) \right] - \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} J_{ij} \left[\mathcal{P}_{N,\beta}(\boldsymbol{\sigma}, \boldsymbol{J}, t) w_{N,\beta}(\boldsymbol{J}, \boldsymbol{\sigma} \to \boldsymbol{J}', \boldsymbol{\sigma}') \right], \tag{2.19}$$

which can be rewritten as

$$\tau' \frac{d\bar{J}_{ij}}{dt} = -\bar{J}_{ij} + \sum_{\boldsymbol{J},\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \Sigma_N} \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) w_{N,\beta}(\boldsymbol{J}', \boldsymbol{\sigma}' \to \boldsymbol{J}, \boldsymbol{\sigma}) J_{ij}. \tag{2.20}$$

After simple calculations very similar to those used in the previous case, we get

$$\tau' \frac{d\bar{J}_{ij}}{dt} = -\bar{J}_{ij} + \sum_{\boldsymbol{J}' \in \mathcal{J}_N} \sum_{\boldsymbol{\sigma}' \in \Sigma_N} \mathcal{P}_{N,\beta}(\boldsymbol{\sigma}', \boldsymbol{J}', t) \tanh(\beta)\, \sigma_i' \sigma_j', \tag{2.21}$$

which, in the mean field approximation, becomes

$$\tau' \frac{d\bar{J}_{ij}}{dt} = -\bar{J}_{ij} + \tanh(\beta)\, \bar{\sigma}_i \bar{\sigma}_j. \tag{2.22}$$

Equations 2.18 and 2.22 form the system of coupled ODEs whose behavior we inspect.

2.2 Experimental Setting. To see this theoretical picture at work on the full neural network, we perform extensive simulations and, for transparency, we report the discretized version of equations 2.18 and 2.22 for numerical implementation:

$$\bar{\sigma}_i^{(n+1)} = \bar{\sigma}_i^{(n)} \left(1 - \frac{\delta t}{\tau}\right) + \frac{\delta t}{\tau} \tanh\left(\beta \sum_{i \neq j}^{N} \bar{J}_{ij}^{(n)} \bar{\sigma}_j^{(n)} + u\beta h_i^{(n)}\right), \tag{2.23}$$

$$\bar{J}_{ij}^{(n+1)} = \bar{J}_{ij}^{(n)} \left(1 - \frac{\delta t}{\tau'}\right) + \frac{\delta t}{\tau'} \bar{\sigma}_j^{(n)} \bar{\sigma}_i^{(n)} \tanh\beta, \tag{2.24}$$

where δt identifies the time unit for the system relaxation and n represents the number of time steps that have elapsed (iterations), such that $t = n\delta t$; we also stressed the dependence of the fields $h_i^{(n)}$ on the time step n, since in general, the external stimulus is presented for a limited temporal window.

To appreciate conditioning at work and how the latter implies Hebbian learning, we conduct a simple series of numerical experiments whose structure is preserved through the letter: starting from a network whose memory is a tabula rasa, during the first two parts of the experiment, we let the network learn separately the first and second patterns (one pattern per part); then, once these have been stored in the synaptic matrix, in the third part of the experiment, we present both patterns simultaneously (for a time window larger than the typical synaptic timescale), forcing the network to conditioning. We expect the correlation between these patterns to be captured by the system such that later, in the fourth part of the experiment, by presenting solely one of these patterns, the network retrieves both of them.

The core idea for proving that the mathematics underlying Palvov's conditioning is the same underlying Hebb's learning (and, thus, that the two are ultimately the same phenomenon on large enough timescales) is to repeat this experiment by enlarging the amount of information the network has to deal with. We start with two patterns, each encoding a single bit; thus, we are working at the single-neuron level (hence, we expect to see that neurons that fire together wire together). Then, we let the network handle two concepts; patterns are taken as vectors, each encoding $O(N)$ bits, in such a way that, preserving the same structure of the numerical experiments, we expect to see that concepts also get correlated by this mechanism (hence, showing that classical conditioning obeys the same mathematical structure too).

We now report on the first numerical experiment. We take $N = 2$ (hence, we work at the single-neuron level), and the stimuli presented to the neurons are $\xi^1 \in \{-1, 1\}$ for neuron 1 and $\xi^2 \in \{-1, 1\}$ for neuron 2. The field $\boldsymbol{h}^{(n)}$ inserted in equations 2.23 and 2.24 takes the form

$$\boldsymbol{h}^{(n)} = \begin{cases} \boldsymbol{h}^1 := (\xi^1, 0) & \text{for } n \in \mathcal{I}_1 \\ \boldsymbol{h}^2 := (0, \xi^2) & \text{for } n \in \mathcal{I}_2 \\ \boldsymbol{h}^3 := (\xi^1, \xi^2) & \text{for } n \in \mathcal{I}_3 \end{cases}, \tag{2.25}$$

where $\mathcal{I}_i$ is the time interval where the neuron i occurs to be stimulated, whereas $\mathcal{I}_3$ is the time interval where both neurons occur to be stimulated. Note that in the fourth segment of the experiments, only one pattern is presented; hence, the structure of its field is either $\boldsymbol{h}^1$ or $\boldsymbol{h}^2$.

To check the alignment between the neural configuration $\boldsymbol{\sigma}^{(n)}$ and the stimulus $\boldsymbol{h}^{(n)}$, we also introduce the Mattis magnetization $\boldsymbol{m}^{(n)}$,

$$m_i^{(n)} = \frac{\boldsymbol{\sigma}^{(n)} \cdot \boldsymbol{h}^i}{\left\| \boldsymbol{h}^i \right\|_1} = \sigma_i \xi^i, \text{ where } \left\| \boldsymbol{h}^i \right\|_1 = \sum_{j=1}^{2} |h_j^i|, \quad i \in \{1, 2\}, \tag{2.26}$$

which quantifies the overlap of the neuron state $\boldsymbol{\sigma}^{(n)}$ with the stimulus $\boldsymbol{h}^i$.

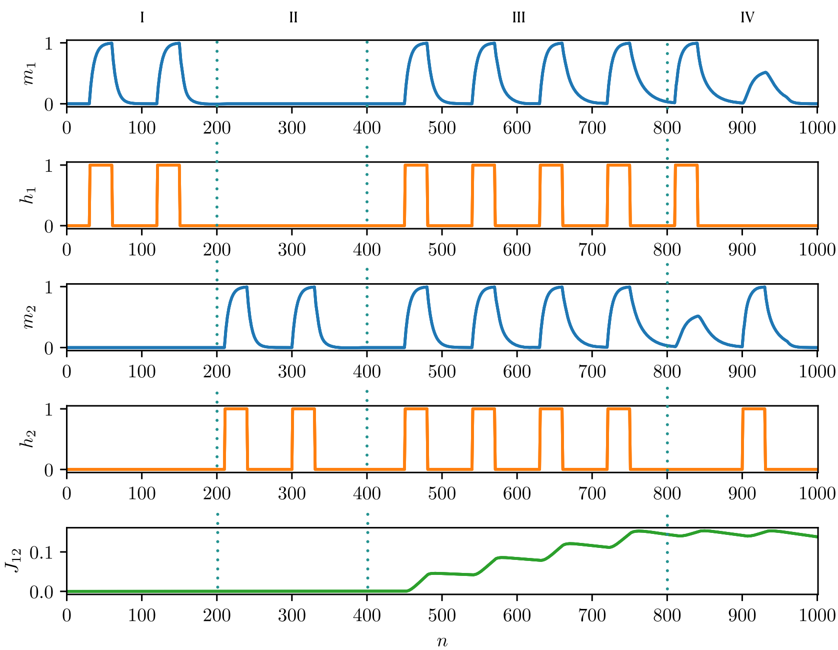

Figure 1: Classical conditioning on single couples of neurons: Neuron that fire together wire together. These five panels report the evolution of the neural activities (blue lines) through $m_i^{(n)}$ with $i = 1, 2$ (see equation 2.26) of the stimuli (orange lines) presented to the network, being $h_i = \boldsymbol{h}^{(n)} \cdot \boldsymbol{h}^i$ with $i = 1, 2$, (see equation 2.25) and of the synaptic matrix J_{12} (green line). During the first 400 time steps, the experiment crosses the first two phases, where the stimuli are presented separately. In the third phase of the experiment, corresponding to the next 400 steps, stimuli are simultaneously presented, and their correlation starts to get stored in the relative coupling of the synaptic matrix (the green curve increases from zero). As a result, in the fourth phase of the experiment (corresponding to $n \geq 800$), while we present solely the first stimulus, both the magnetizations m_1 and m_2 respond, confirming that classical conditioning has successfully taken place. In this simulation, the parameters appearing in equations 2.23 and 2.24 are chosen as follows: $\beta = 4$, $N = 2$, $\delta t = 0.16\tau$, $\tau/\tau' = 0.012$ and $u = 1$. Furthermore, the duration of each stimulation is $n_s = 30$ stimulus is presented for a time $t \geq \tau$ which in terms of number of iterations corresponds to $n_s \delta t \geq \tau$, and since in the simulation $\delta t = 0.16\tau$, n_s must be $\gtrsim 10$.

Results for the time evolution of $\boldsymbol{h}^{(n)}$, $\boldsymbol{m}^{(n)}$, and $J_{12}^{(n)}$ are shown in Figure 1. Note that in the first two parts of the experiment, the two neurons are stimulated separately, and each single stimulation session corresponds to an alignment of the related neuron, while the other neuron does not exhibit any persistent orientation, giving rise to a null magnetization and, consistently, the synapse J_{12} also remains neutral. In the third part of the experiment, where the two neurons are simultaneously stimulated, both react accordingly. This time, the synapse J_{12} is also affected and stores its correlation (if the external stimuli last for a time window $t \geq \tau'$). This information is retained by the system even when both fields are switched off in such a way that when only one neuron is stimulated (in the fourth part of the experiment), the second one also reacts. The system has therefore learned to relate the stimuli: solely the presence of one pattern, out of the two, is sufficient to prompt the retrieval of both ξ^1 and ξ^2.

3 From Pavlov's Classical Conditioning to Hebb's Learning

Once Pavlov's conditioning at the single-synapse level is understood, we repeat our experimental protocol by enlarging the information the network has to cope with: we move from two stimuli made of simple bits toward two stimuli made of concepts as represented by patterns whose length is $O(N)$ bits, therefore employing a large number of neurons and synapses. Clearly, here we are performing a very simplistic assumption by identifying a concept with a generic pattern of information; however, as these patterns can effectively code for thoughts, images, objects (e.g., the food and the bell in Pavlov's celebrated experiments on dog salivation), and so on, this working assumption is actually very versatile.

The simplest path to move from Pavlov to Hebb consists of considering two stimuli affecting different neural areas as standard in classical conditioning. This will be the first scenario addressed in section 3.1; in section 3.2, we generalize this setting by considering several stimuli at once toward modern versions of generalized conditioning like those studied by Rescorla-Wagner (1972) and Mackintosh (1975). Next, in section 3.3, we show that the long-time limit of the synaptic matrix subjected to different stimuli displays a Hebbian structure. Finally, in section 3.4, we speculate on phenomena like obsessions and unlearning (Benedetti et al., 2022; Fachechi et. al., 2019; Hopfield et al., 1983; Marinari, 2019).

3.1 Emergence of Concept's Correlations by Classical Conditioning. We preserve the schedule of the numerical experiment. In the first two parts of the experiment, to make the network learn two uncorrelated patterns, we present them separately (one pattern per part), checking the absence of mutual correlations (when we re-present one of them after learning, solely that pattern is retrieved, as expected in Hebb's learning). Then, in the third phase, we present both patterns simultaneously and persistently for a time window larger than the synaptic timescale. To inspect how these patterns get correlated within the synaptic matrix, in the fourth part of the experiment, by presenting solely one of them, we check that the network actually retrieves both of them, confirming Pavlov's conditioning successfully.

We start by introducing two stimuli $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ as two vectors (whose length is $O(N)$ bits) to be applied to the system. We choose a particular structure for stimuli that follows from the original Pavlov setting, where stimuli come by different senses involving different neural regions in the brain (e.g., visual for the food and acoustic for the ringing bell). Thus, we assume that the two stimuli involve different neurons, and in the simplest setting, both concepts to learn involve exactly $N/2$ neurons.[5] Specifically, $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ are arranged as

$$\boldsymbol{\xi}^1 = \left(\xi_1^1, \xi_2^1, \ldots, \xi_{\frac{N}{2}}^1, 0, 0, \ldots, 0\right), \tag{3.1}$$

$$\boldsymbol{\xi}^2 = \left(0, 0, \ldots, 0, \xi_{\frac{N}{2}+1}^2, \xi_{\frac{N}{2}+2}^2, \ldots, \xi_N^2\right). \tag{3.2}$$

Note that now, $\boldsymbol{\xi}^{1,2} \in \{-1, 0, +1\}^N$.

The fields stemming from these stimuli that are inserted in equations 2.23 and 2.24 read as

$$h_i^{(n)} = \begin{cases} \xi_i^{(n)} & \text{if } |\xi_i^{(n)}| = 1 \\ \mathcal{U}[-1, +1] & \text{otherwise} \end{cases} \tag{3.3}$$

with

$$\boldsymbol{\xi}^{(n)} = \begin{cases} \boldsymbol{\xi}^1 & \text{for } n \in \mathcal{I}_1 \\ \boldsymbol{\xi}^2 & \text{for } n \in \mathcal{I}_2 \end{cases}, \tag{3.4}$$

where $\mathcal{I}_\mu$ with $\mu = 1, 2$ represents the time interval where the stimulus $\boldsymbol{\xi}^\mu$ is active and $\mathcal{U}[-1, +1]$ is a uniform random variable supplying noise when no information is present. This mimics that in the absence of stimuli, the neurons are still receptive, but they experience just noise, uniformly distributed over the same amplitude ruled by β when information is present (see equations 2.18 and 2.22).

As anticipated, in the early learning stage, we present to the network these patterns separately in the form prescribed in equation 3.3 until these are learned. At the end of this trivial exercise, the resulting synaptic matrix of the network convergences to a modular Hebbian network (Agliari et al., 2014; Sollich et al., 2014), whose blocks share the same size $\frac{N}{2} \times \frac{N}{2}$ and the off-diagonal blocks are null:

$$\boldsymbol{J}(\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\}) = \frac{1}{2} \times \begin{pmatrix} \boldsymbol{J}^1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}^2 \end{pmatrix}, \tag{3.5}$$

[5] The general case of patterns of arbitrary length is deepened in appendix B.

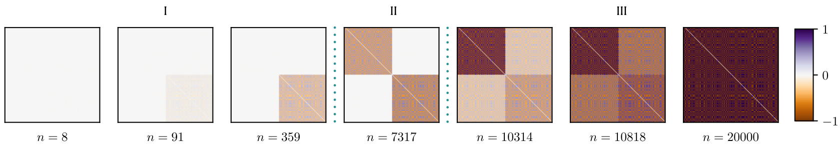

Figure 2: Evolution of the synaptic matrix when two stimuli encoding concepts are involved. In the first three panels (for $n \le 10^5$), the synaptic matrix evolves while storing the single patterns (pure learning mode of single stimuli) while the last three images (for $n > 10^5$) show how classical conditioning forces the off-diagonal blocks to have non-null entries accounting for the Pavlovian correlation among concepts. These results are obtained by setting the parameters in equations 2.23 and 2.24 as follows: $N = 200$, $\beta = 10$, $\tau/\tau' = 0.012$, $\delta t = 0.10\tau$ and the strength of the stimuli is $u = 200$.

namely,

$$J_{ij}(\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\}) = \frac{1}{2} \times \begin{cases} \xi_i^1 \xi_j^1 & (i,\, j) \in B_1 \\ \xi_i^2 \xi_j^2 & (i,\, j) \in B_2 \\ 0 & (i,\, j) \vee (j, i) \in B_{mix}, \end{cases} \tag{3.6}$$

where $B_1 = [1, N/2] \times [1, N/2]$, $B_2 = [N/2+1, N] \times [N/2+1, N]$, and $B_{mix} = [1, N/2] \times [N/2+1, N]$ (see the left panels of Figure 2). In order to understand why we expect the synaptic matrix to converge precisely to the structure coded in equation 3.6, we use the result obtained in section A.2. At the end of the training, the synaptic matrix converges to the temporal mean of the stimuli $J_{ij}(t) = \frac{1}{t}\int_0^t h_i(\tau)h_j(\tau)d\tau$, plus a fluctuation at depends on the noise and the ratio τ/τ' (see equation A.11). There are now two fields given as stimuli with equal frequency during the stimulation sessions (i.e., $\boldsymbol{h}^{(n\in\mathcal{I}_1)}$ and $\boldsymbol{h}^{(n\in\mathcal{I}_2)}$). Thus, the temporal mean of the stimuli is equal to $\frac{1}{2}\sum_{k=1}^{2}\xi_i^k\xi_j^k$, which corresponds to equation 3.6.

As for the off-diagonal blocks, since there are no correlations among the related neurons, they contain no information. Next, we present both stimuli simultaneously: the overall stimulation provided to the network is

$$\boldsymbol{\xi}^{1+2} = \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2 = (\xi_1^1, \xi_2^1, \ldots, \xi_{\frac{N}{2}}^1, \xi_{\frac{N}{2}+1}^2, \xi_{\frac{N}{2}+2}^2, \ldots, \xi_N^2), \tag{3.7}$$

and it is retained long enough to allow synapses to be plastic. The resulting synaptic matrix converges to

$$\boldsymbol{J}(\boldsymbol{\xi}^{1+2}) = \begin{pmatrix} \boldsymbol{J}^1 & \boldsymbol{J}^{\text{mix}} \\ \boldsymbol{J}^{\text{mix}} & \boldsymbol{J}^2 \end{pmatrix}, \tag{3.8}$$

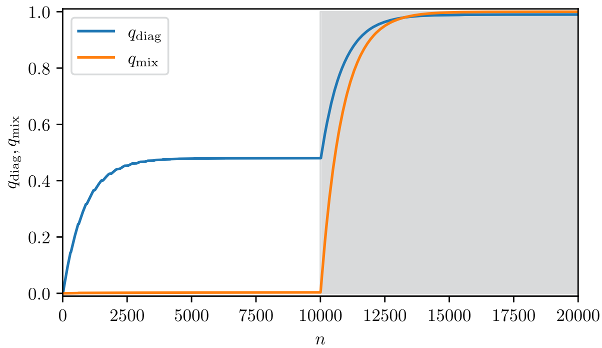

Figure 3: Evolution of the overlaps $q_{\rm diag}$, $q_{\rm mix}$ versus the simulation time steps n. On the left side of the plot (white background), the two stimuli are presented separately and alternately to the network (each stimulus is presented consecutively for $n_s = 300$ iterations), whereas on the right side of the plot (gray background) the two stimuli are jointly presented. Note that on the left side of the plot (for $n \leq 10^5$), only standard learning of patterns takes place; instead, on the right side of the plot (for $n > 10^5$) classical conditioning takes place too. The parameters of the simulation are the same as in Figure 2.

namely,[6]

$$J_{ij}(\boldsymbol{\xi}^{1+2}) = \begin{cases} \xi_i^1 \xi_j^1 & (i,\, j) \in B_1 \\ \xi_i^2 \xi_j^2 & (i,\, j) \in B_2 \\ \xi_i^1 \xi_j^2 & (i,\, j) \in B_{mix} \\ \xi_j^1 \xi_i^2 & (j,\, i) \in B_{mix}. \end{cases} \tag{3.9}$$

In order to corroborate this picture numerically, we inspect the overlap between the blocks of the matrix $\boldsymbol{J}(\boldsymbol{\xi}^{1+2})$ and the time-dependent blocks of the matrix $\boldsymbol{J}^{(n)}$ emerging from our numerical experiment. To this purpose, we introduce the (planted) overlaps

$$q_{\rm diag}^{(n)} := \frac{1}{2}\sum_{k=1}^{2} \frac{\sum_{(i,j)\in B_k} J_{ij}^k J_{ij}^{(n)}}{\sum_{(l,m)\in B_k} J_{lm}^k J_{lm}^k}, \quad q_{\rm mix}^{(n)} := \frac{\sum_{(i,j)\in B_{\rm mix}} J_{ij}^{\rm mix} J_{ij}^{(n)}}{\sum_{(l,m)\in B_{\rm mix}} J_{lm}^{\rm mix} J_{lm}^{\rm mix}}. \tag{3.10}$$

Their evolution versus time is represented in Figure 3. As long as the two concepts ($\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$) are presented separately (white background), only the

[6]Note that the 1/2 factor is no longer present (see, e.g., equation A.4). This is because only one pattern is given as stimulus (in the form $\boldsymbol{\xi}^{1+2}$), and the network learns only this one.

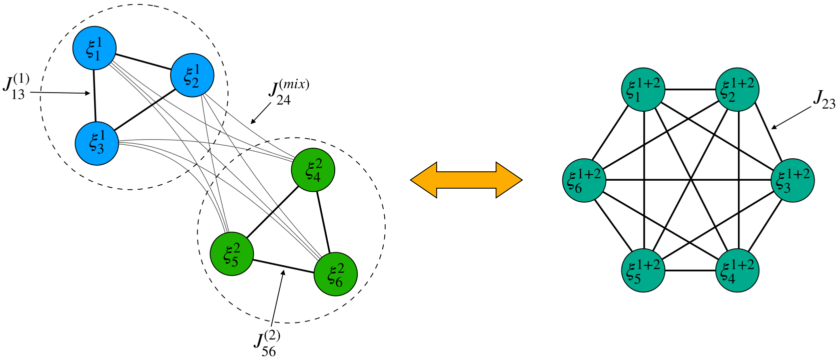

Figure 4: Stylized representation of the classical conditioning within the neural network. Before conditioning (left panel), the whole system, which has already learned the two patterns $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$, is split into two cliques (blue and green subfamilies of neurons, respectively): half of the neurons account for the pattern $\boldsymbol{\xi}^1$ and half of the neurons account for pattern $\boldsymbol{\xi}^2$. Accordingly, the synaptic matrix is naturally split into four blocks: $\boldsymbol{J}^{(1)}$ (accounting for the pattern $\boldsymbol{\xi}^1$) and $\boldsymbol{J}^{(2)}$ (accounting for the pattern $\boldsymbol{\xi}^2$) are the intrafamily connections storing information on the single patterns, while the two off-diagonal blocks $\boldsymbol{J}^{(mix)}$ connect neurons of different subfamilies. Their synapses grow solely when driven by classical conditioning and result in a fully connected network, the Hopfield model (Hopfield, 1982) (right panel).

synaptic blocks related to the single stimuli increase (i.e., blocks $\boldsymbol{J}^1$ and $\boldsymbol{J}^2$), and, accordingly, q_{diag} saturates to 1/2. On the other hand, while the two concepts are presented simultaneously ($\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$, gray background), all four synaptic blocks are increasing, and both q_{diag} and q_{mix} saturate to one. Otherwise stated, the presentation of $\boldsymbol{\xi}^{1+2}$ implies classical conditioning among $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ (as evidenced by the growth of q_{mix}), and the synaptic matrix reaches the form given by equation 3.9. In Figure 2, we show the evolution of the synaptic matrix as the experiment flows: while the diagonal blocks increase from zero when stimuli are individually presented (see the left panels), the off-diagonal blocks increase from zero just when both stimuli are simultaneously presented (see the right panels), and eventually all the entries in the matrix take values ± 1. Finally, Figure 4 provides a qualitative description of the conditioning phenomenon at the network's level.

The scenario just described is preserved in the case of patterns involving a different number of neurons, say, pN for $\boldsymbol{\xi}^1$ and $(1-p)N$ for $\boldsymbol{\xi}^2$, with $p \in (0, 1)$, as discussed in section A.1.

To conclude, we have shown how classical conditioning naturally extends from single couples of neurons up to the case of assemblies of neurons (see Figure 5). To verify this universality of the Pavlovian mechanism, it is instructive to compare Figure 1 (obtained with single bits of information

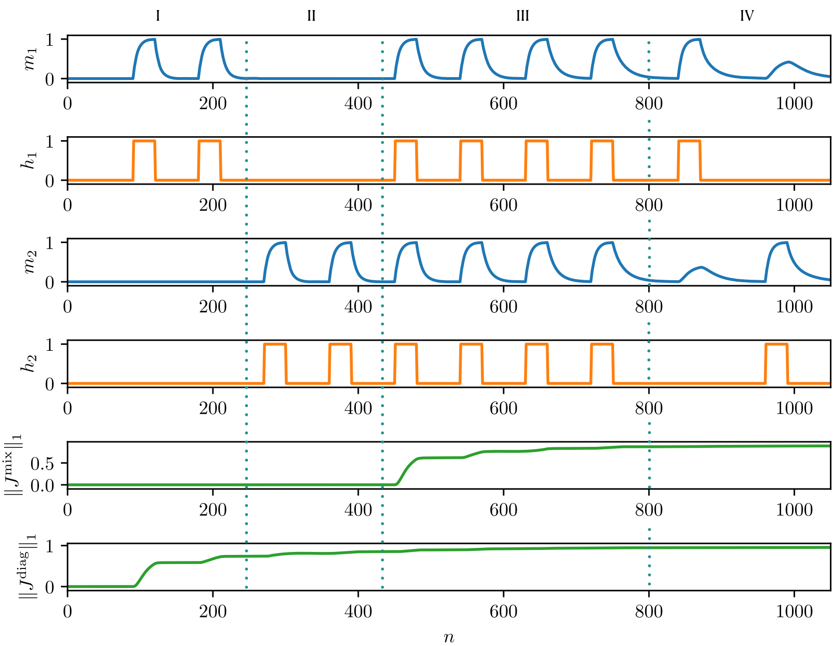

Figure 5: Classical conditioning among patterns, involving cliques of neurons. Panels 1 and 3 report the evolution of the neural activities through $m_i^{(n)}$ (whose definition is a direct generalization of equation 2.26), panels 2 and 4 report the values of the stimuli presented to the network through $h_{1,2}$, and panels 5 and 6 report the evolution of the blocks of the synaptic matrix $\boldsymbol{J}$ through $\left\|J^{\mathrm{mix}}\right\|_1$ and $\left\|J^{\mathrm{diag}}\right\|_1$. During the first 400 iterations, the stimuli are presented separately, and $\left\|J^{\mathrm{diag}}\right\|_1$ increase from zero. In the next 400 iterations, the stimuli are simultaneously presented to the network and also $\left\|J^{\mathrm{mix}}\right\|_1$ increases from zero. As a result when, at $n = 850$, we present solely the first stimulus, both the magnetization m_1 and m_2 respond, providing Pavlov's mechanism at the level of extended patterns. The parameters used for this simulation, which were inserted equations 2.23 and 2.24, are the following: $N = 20$, $\beta = 4$, $u = 1$, $\tau/\tau' = 0.0004$, and $\delta t = 0.16\tau$.

and resulting in the Hebbian prescription that neurons that fire together wire together, as confirmed by its last line) with Figure 5 (obtained with entire patterns of information, that is, *concept*, and giving rise to Pavlov's classical conditioning, as well as to a synaptic matrix that obeys the Hebbian prescription).

3.2 Generalized Conditioning: Integrating Multiple Signals at Once. In this section, we move from the original Pavlovian setting, where only two

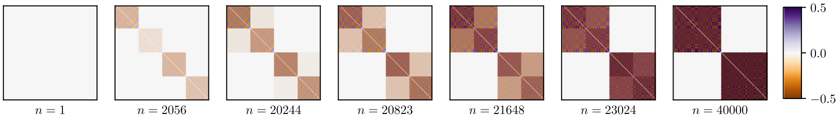

Figure 6: Evolution of the synaptic matrix under general conditioning: Symmetric case. For $n \leq 2000$ solely one pattern per time is presented, and the synaptic matrix converges to a block matrix where solely the single patterns $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \boldsymbol{\xi}^3, \boldsymbol{\xi}^4$ have been stored. For $n \geq 20{,}000$ generalized conditioning takes place as the stimuli are now presented simultaneously in couples, in the form $\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$ or $\boldsymbol{\xi}^3 + \boldsymbol{\xi}^4$, and the network correctly stores the correlations among the relative stimuli. The parameters used for the simulation (to be inserted in equations 2.23 and 2.24) are $N = 200$, $\beta = 10$, $\tau/\tau' = 0.005$, $\delta t = 0.1\tau$, and the strength of the stimuli is $u = 200$.

stimuli are involved, toward more challenging cases where larger numbers of stimuli are involved at once (Mackintosh, 1975; Rescorla & Wagner, 1972).

In particular we generalize the Pavlovian scheme to account for the following four patterns:

$$\boldsymbol{\xi}^1 = \left(\xi_1^1, \xi_2^1, \ldots, \xi_{\frac{N}{4}}^1, 0, 0, \ldots, 0\right), \tag{3.11}$$

$$\boldsymbol{\xi}^2 = \left(0, 0, \ldots, 0, \xi_{\frac{N}{4}+1}^2, \ldots, \xi_{\frac{N}{2}}^2, 0, 0, \ldots, 0\right), \tag{3.12}$$

$$\boldsymbol{\xi}^3 = \left(0, 0, \ldots, 0, \xi_{\frac{N}{2}+1}^3, \ldots, \xi_{\frac{3N}{4}}^3, 0, 0, \ldots, 0\right), \tag{3.13}$$

$$\boldsymbol{\xi}^4 = \left(0, 0, \ldots, 0, \xi_{\frac{3N}{4}+1}^4, \ldots, \xi_N^4\right). \tag{3.14}$$

As in the previous section, the patterns have no overlap ($\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu = 0$ if $\mu \neq \nu$), and each of them stimulates the same number of neurons, in this case, $N/4$ for of simplicity. The fields inserted in equations 2.23 and 2.24 and stemming from these stimuli read as equation 3.3, where, in this experiment, $\boldsymbol{\xi}^{(n)} = \boldsymbol{\xi}^\mu$ if $n \in \mathcal{I}_\mu$, for $\mu = 1, 2, 3, 4$.

The training session is scheduled analogous to the ones outlined in the previous section for classical conditioning, as we briefly recall. First, the patterns are presented and learned separately (and $\boldsymbol{J}^{(n)}$ exhibits a block shape where only diagonal blocks are non-null); next, we present two or more of them simultaneously (phase 3), and finally we check that here, the non-null diagonal blocks of $\boldsymbol{J}^{(n)}$ mirror the combination of patterns presented (phase 4; see Figures 6 and 7). Note that in this case, we have the freedom to group patterns to be presented simultaneously, and here we consider the case where they are presented in couples like $\boldsymbol{\xi}^{1+2} = \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$, $\boldsymbol{\xi}^{3+4} = \boldsymbol{\xi}^3 + \boldsymbol{\xi}^4$ (see Figure 6) and the case where they are presented in two inhomogeneous groups, like $\boldsymbol{\xi}^{1+2+3} = \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2 + \boldsymbol{\xi}^3$, $\boldsymbol{\xi}^4$ (see Figure 7).

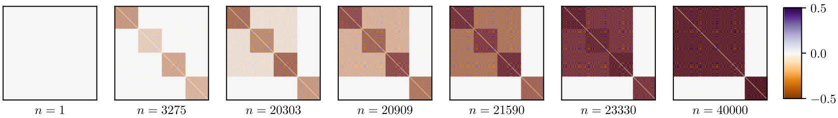

Figure 7: Evolution of the synaptic matrix under general conditioning: Asymmetric case. For $n \leq 20{,}000$ solely one pattern per time is presented and the matrix converges to a block matrix where only the single patterns $\boldsymbol{\xi}^1$, $\boldsymbol{\xi}^2$, $\boldsymbol{\xi}^3$, $\boldsymbol{\xi}^4$ have been stored. For $n \geq 20{,}000$ generalized conditioning takes place as the stimuli are now presented simultaneously in couples, in the form $\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2 + \boldsymbol{\xi}^3$ and, separately, $\boldsymbol{\xi}^4$, and the network correctly stores the correlations among the relative three stimuli that are jointly presented (hence the relative off-diagonal blocks are increased). The parameters of the simulation are the same as in Figure 6.

3.3 Convergence of the Synaptic Dynamics to the Hebbian Matrix. Up to now we have dealt with a relatively small number of patterns, each involving a subset of the neurons making up the system. In this section, we directly consider patterns that involve the whole set of neurons. They occur as $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, with $\mu = 1, \ldots, K$, where K can be very large, possibly scaling with the volume N. We define, *storage capacity* of the network as the quantity $\alpha = K/N$ (Amit et al., 1985).

We are going to prove that as long as pattern entries are Rademacher, that is,

$$P(\xi_i^\mu = \pm 1) = \frac{1}{2}, \qquad \forall i = 1, \ldots, N \text{ and } \mu = 1, \ldots, K,$$

if we randomly choose one of the K patterns to stimulate the network at each iteration (and the noise is sufficiently low $\beta \gg 1$ and the neural and synaptic timescales are well split (i.e., their ratio is low $\tau/\tau' \ll 1$), the long-term limit of the evolution provided by equations 2.23 and 2.24 generates the celebrated Hebbian matrix,

$$J_{ij}^{(Hebb)} = \frac{1}{K} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu. \tag{3.15}$$

Thus, AGS statistical mechanical theory is asymptotically recovered as it should. To be more precise, in section A.2 we demonstrate that at the end of the training, the synaptic matrix converges to the temporal mean of the stored patterns plus a fluctuation (vanishing in the limit $\beta \to \infty$ and $\tau/\tau' \to 0$). Since in these simulations, at each time step n, the stimulus can be $\boldsymbol{\xi}^\mu$ with probability $1/K$, the temporal mean of the stored patterns coincides with the Hebbian prescription for the synaptic matrix as given in equation 3.15. For simplicity, we set $\delta t = \tau$ in equations 2.23 and 2.24. This means that the neurons align themselves with the stimulus instantly so we do not need to present the same stimulus for several iterations (one iteration is sufficient to produce the alignment of the neurons).[7] The learning procedure thus consists of presenting these K patterns as external stimuli to the network. To do this, we randomly choose one of the patterns, say, $\boldsymbol{\xi}^{\bar{\mu}}$, and set the external field to $\boldsymbol{h} = \boldsymbol{\xi}^{\bar{\mu}}$ for a single evolution step. This procedure is repeated as the time t goes by. In Figure 8, we reported the distance $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{(Hebb)}\|_F$, computed according to the Frobenius norm $\|A\|_F := \sqrt{\frac{1}{N^2}\sum_{ij=1}^{N} A_{ij}^2}$, between the coupling matrix $\boldsymbol{J}^{(n)}$ resulting at each time step n and the Hebbian kernel $\boldsymbol{J}^{(Hebb)}$ determined by equation 3.15. Different values of the ratio τ/τ' (panel a), of the parameters $\alpha = K/N$ (panel b), N (panel c), and β (panel d) are considered and compared. Remarkably, if the noise is low enough ($\beta \gg 1$) and if the ratio τ/τ' is low enough, the dynamics of the synaptic matrix converges to the Hebbian kernel. In fact, there is always a reconstruction error of the Hebbian kernel, and $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{(Hebb)}\|_F$ has an asymptotic value that in general is different from zero and depends on τ/τ' and β (see panels a to d). In section A.2, we study the fluctuations of the coupling matrix $\boldsymbol{J}^{(n)}$ resulting from this dynamics around the Hebbian fixed point given in equation 3.15. We find that in the case that $\beta \gg 1$ the fluctuation is vanishing like $(\tau/\tau')^{1/2}$ (see equation A.17) and perfectly reproduces the asymptotic behavior of the Frobenius norm of the difference between the synaptic matrix and the corresponding Hebbian kernel (see panel a, dashed line), whereas if $\beta \sim 1$, we also have to take into account the contribution of the fast noise to the fluctuations around the Hebbian kernel. This is given by a more general formula (see equation A.18), but also in this case, we have a perfect agreement between the theoretical estimations and the numerical simulations (panel d, dashed line). Finally, in panels b and c, we observe that the fluctuations around the Hebbian kernel do not depend on the size of the system N or on the storage capacity $\alpha = K/N$ as expected from the theoretical estimation. This implies that the learning algorithm is robust in performing the storage of the patterns in the synaptic matrix. Moreover, for values of the storage capacity larger than $\alpha_c \sim 0.14$, the magnetization no longer saturates to one, in agreement with the statistical mechanical limiting description of the blackout catastrophe provided by AGS theory (Amit et al., 1985; see Figure 9).

3.4 Persistent Retrieval, Obsessions, and Unlearning. The dynamical process depicted so far involves both neurons and synapses, and it allowed us to highlight structural similarities between Pavlov's and Hebb's

[7] In section 3 we set $\delta t = c\tau$ with $c \in (0, 1)$ because we wanted to observe the neural reaction to an external stimulus and appreciate the transient effects of the presence of an appropriate timescale for neurons (see Figure 1). Now we focus only on the dynamics of the synaptic matrix, hence the decision to set $c = 1$.

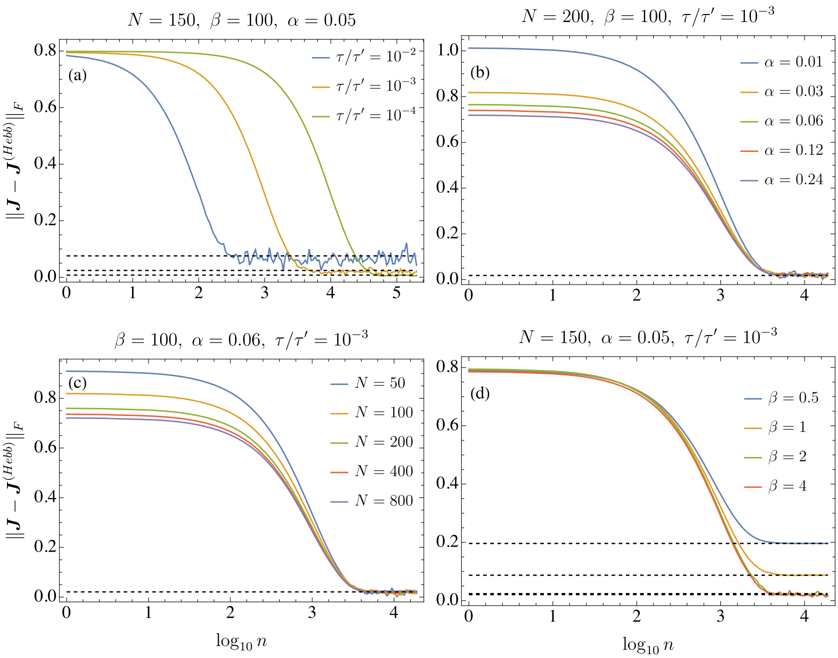

Figure 8: Generalized conditioning: Convergence to the Hebbian kernel. Results of the synaptic dynamics for various setting of the external parameters. The curves refer to the Frobenius norm of the difference between the evolving synaptic matrix and the corresponding limiting Hebbian kernel as a function of the simulation time. In panel a, we fix the network size $N = 150$, the storage capacity $\alpha = 0.05$, and the thermal noise level $\beta = 100$ and vary the ratio $\tau/\tau' = 10^{-2}, 10^{-3}$ and 10^{-4}. In panel b, we fix the network size $N = 200$ and the ratio $\tau/\tau' = 10^{-3}$ and vary the storage capacity $\alpha = 0.01, 0.03, 0.06, 0.12, 0.24$. In panel c, we fix the storage capacity $\alpha = 0.06$ and the ratio $\tau/\tau' = 10^{-3}$ and vary $N = 50, 100, 200, 400, 800$. In panel d, we fix the network size $N = 150$, the storage capacity $\alpha = 0.05$, and the ratio $\tau/\tau' = 10^{-3}$ and vary the thermal noise $\beta = 0.5, 1, 2, 4$. The horizontal dashed lines correspond to the theoretical estimate according to equation A.18. Each pattern is presented to the network as external field, the duration of the stimulus being restricted to a single time step. The pattern is selected with uniform probability in the index set $\{1, \ldots, K\}$ and the procedure is repeated for $2 \cdot 10^4$ evolution steps. Moreover in all the simulations the strength of the stimuli is equal to $u = 200$.

representations of learning. Remarkably, the interplay between neuron and synapse dynamics can give rise to interesting phenomena that cannot be captured by a statistical mechanics approach where the latter are quenched

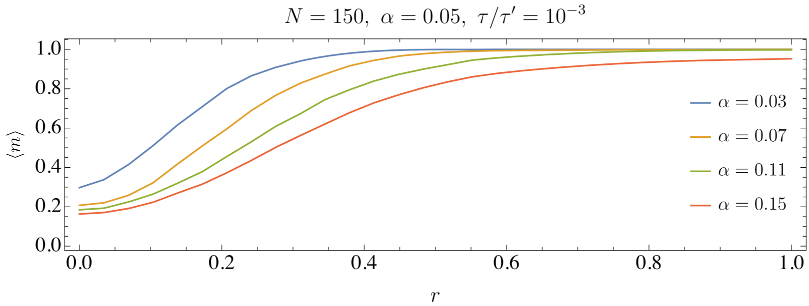

Figure 9: Hebbian-like retrieval capabilities of the network after training. For each of the K patterns stored in the network, we generate $M = 100$ noisy versions of them by flipping randomly a percentage of bits equal to $\frac{1-r}{2} \times 100$ with $r \in (0, 1]$. Then we let the system evolve for five iterations with the dynamics prescribed by equation 2.18, in the absence of stimuli $h_i = 0$ and at quenched synaptic matrix $\boldsymbol{J}$. $m_\mu = \frac{1}{M}\sum_{a=1}^{M} \boldsymbol{\sigma}_a^{\text{final}} \cdot \boldsymbol{\xi}^\mu/N$ between the M final neural configurations $\boldsymbol{\sigma}_a^{\text{final}}$ and the pattern μ. We repeat this algorithm for all the patterns, and we average over the values of m_μ obtained: $\langle m \rangle = \frac{1}{K}\sum_{\mu=1}^{K} m_\mu$. The parameters of the simulations are $N = 150$ and $\beta = 100$. Note that for $\alpha > \alpha_c \sim 0.14$ the magnetization does not saturate any longer to one, in agreement with AGS theory (Amit et al., 1985).

(Amit, 1989; Coolen et al., 2005). Specifically, here we show that if the network gets stuck in retrieving persistently the same pattern (i.e., an obsession), each retrieval reinforces the same minimum in the memory landscape up to the point that the latter destroys all the other memories, and only the obsessive pattern remains stored. Note that this peculiarity of classical conditioning can be seen as a sort of (unwanted) unlearning procedure (Fachechi et. al., 2019).

In Figure 10, we report results of the following numerical experiment to confirm such a scenario. Starting from a perfect Hebbian kernel $\boldsymbol{J}^{(Hebb)}$, we always let the nework retrieve the first pattern (i.e., the field is persistently $\boldsymbol{h} = \boldsymbol{\xi}^1$), and we evaluate the evolution of both $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{(Hebb)}\|_F$ and $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{(\mu=1)}\|_F$, where $J_{ij}^{(\mu=1)} = \xi_i^1 \xi_j^1$. Remarkably, in the long time limit, the synaptic matrix collapses onto ξ^1 ($\lim_{n\to\infty} \boldsymbol{J}^{(n)} = \boldsymbol{J}^{(\mu=1)}$), while $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{(Hebb)}\|_F \to \sqrt{1 - \frac{1}{K}}$ as expected.

It is worth stressing that in classical learning theories via statistical mechanics, pattern recognition cannot alter the memory landscape, but this is not the case here. This phenomenon can also be seen as a technique for forgetting alternatives to pruning (Blalock et al., 2020) or dreaming (Fachechi et. al., 2019) mechanisms.

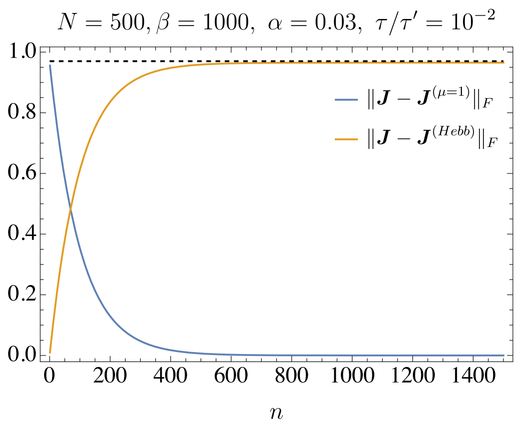

Figure 10: Persistent retrieval of the same pattern as natural memory erasure. The evolution of the Frobenius norms $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{Hebb}\|_F$ and $\|\boldsymbol{J}^{(n)} - \boldsymbol{J}^{\mu=1}\|_F$ when the first pattern is persistently retrieved proves that after enough time, the whole memory reduces to the first pattern, all the others being erased by this persistence, making the network hooked. The simulation reported in the plot has been carried by keeping $\tau/\tau' = 10^{-2},\ \alpha = 0.03,\ \beta = 10^3,\ N = 500$; the black dashed line corresponds to the theoretical estimate $\sqrt{1 - \frac{1}{K}} \approx 0.96$.

4 Conclusion

Although this research has been carried out mainly to fill a historical gap in modeling information processing by neural networks, namely, to show how Pavlov's theory on conditioning ultimately implies Hebbian learning, several observations of general interest can still be drawn.

First, we notice that the emergence of Palvov's mechanism from a two timescale dynamics and its convergence, in the long term limit, to Hebb's synaptic matrix was a (somehow expected but missing) bridge between multiscale stochastic processes and statistical mechanics of neural networks. The generality of the underlying framework suggests that this kind of information processing mechanisms is quite spread out in nature and not confined within the neural world.

Moreover, by the perspective outlined in this letter, there is no excitatory or inhibitory learning. Rather, there is just statistical learning: patterns with positive correlations would be reciprocally excitatory and vice versa for anticorrelated patterns, but the whole process of learning is entirely and solely statistically driven. At this level of modeling, we do not distinguish between primary and secondary stimuli; each pattern plays as a conditioner for each other as long as we work in a random setting.

Further, the way classical conditioning can take place is by presenting two stimuli acting on different areas (i.e., on different sets of neurons), and then correlations among the stimuli will result in the growth of synapses connecting these (previously uncorrelated) neurons or cliques of neurons. This nicely fits our framework inspired by recently introduced multitasking networks able to deal with partially blank pattern entries (Agliari, Barra, Galluzzi, et al., 2012; Agliari et al., 2014; Sollich et al., 2014).

Finally, as Pavlov's learning rule is intrinsically dynamical, there are phenomena that it describes but cannot be captured by a statistical mechanical picture of Hebbian learning. By the latter (e.g., in the Hopfield model), patterns have their stable basins of attraction, and these are static if no further storing takes place (namely, the network can retrieve persistently the same pattern, and this operation does not alter the landscape where stored memories lie); solely learning new patterns alters the memory landscape. Conversely, from Pavlov's perspective, retrieval (of one pattern) affects the amplitude and stability of (all the) minima and retrieving persistently the same pattern makes its basin of attraction more and more pronounced, to the detriment of the other pattern's basins up to the point that by constantly retrieving the same pattern (modeling obsessions), the whole memory collapses to solely that information. Thus, forgetting within the Pavlov scheme can be achieved without network pruning (Blalock et al., 2020; Molchanov et al., 2019) or removal by sleeping-induced mechanisms (Aquaro et al., 2022; Fachechi et. al., 2019).

Appendix

A.1 Classical Conditioning for Patterns of Different Lengths. We keep the assumption that we deal solely with two signals (the latter will be removed later in this section), but we relax the constraint that the length (in bits) of the patterns has to be the same for $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$. In general, patterns can be arranged such that

$$\boldsymbol{\xi}^1 = \left(\xi_1^1, \xi_2^1, \ldots, \xi_{Np}^1, 0, 0, \ldots, 0\right), \tag{A.1}$$

$$\boldsymbol{\xi}^2 = \left(0, 0, \ldots, 0, \xi_{Np+1}^2, \xi_{Np+2}^2, \ldots, \xi_N^2\right), \tag{A.2}$$

where $p \in (0, 1)$, and thus we question if, in this generalized setting, classical conditioning is preserved.

Once having presented them separately to the network (as prescribed in equation 3.3) and after enough persistence (for long enough learning dynamics such that they can be stored), the resulting synaptic matrix reads as

$$J_{ij}(\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\}) = \frac{1}{2} \times \begin{cases} \xi_i^1 \xi_j^1 & (i, j) \in B_1 \\ \xi_i^2 \xi_j^2 & (i, j) \in B_2 \\ 0 & (i, j) \vee (j, i) \in B_{mix}, \end{cases} \tag{A.3}$$

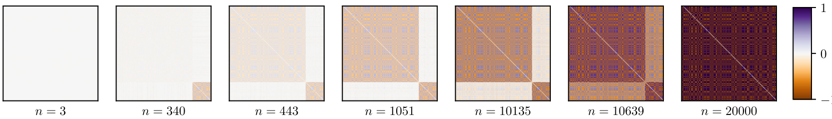

Figure 11: Evolution of the synaptic matrix $\boldsymbol{J}$. While $n \leq 10{,}000$, only single patterns are presented, while and $n > 10{,}000$, the stimulus is presented in the form $\boldsymbol{\xi}^{1+2} = \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$ and relative conditioning takes place. We can appreciate that in the first part of the experiment, off-diagonal blocks stay null, while they start to increase as the concepts are simultaneously presented and conditioning starts. In this simulation $p = 0.8$ and the other parameters are the same as Figure 3.

where $B_1 = \left[1, pN\right] \times \left[1, pN\right]$, $B_2 = \left[pN+1, N\right] \times \left[pN+1, N\right]$, and $B_{mix} = \left[1, pN\right] \times \left[pN+1, N\right]$. Thus, as expected, $\boldsymbol{J}(\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\})$ is still a block matrix, but it has obviously a rectangular shape now,

$$\boldsymbol{J}(\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\}) = \frac{1}{2} \times \begin{pmatrix} \boldsymbol{J}^1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}^2 \end{pmatrix}, \tag{A.4}$$

where $\boldsymbol{J}^1 \in \mathbb{R}^{Np \times Np}$, $\boldsymbol{J}^2 \in \mathbb{R}^{N(1-p) \times N(1-p)}$, $\boldsymbol{0} \in \mathbb{R}^{N(1-p) \times Np}$. As it can be appreciated by a glance at Figure 11, classical conditioning is robust against this kind of perturbation: this is remarkable because it states that no matter the information content of a pattern, it can still be a conditioner. Moreover, the behavior of the overlaps $q_{\text{diag}}, q_{\text{mix}}$ versus the time step turns out to be exactly the same as that of Figure 3.

A.2 Analytical Estimate of Fluctuations around the Hebbian Fixed Point. In this section, we quantify the fluctuations of the coupling matrix $\boldsymbol{J}$, resulting from the Langevin dynamics with regard to the Hebbian kernel. To this aim, we return to the continuum version of the dynamical equations and note that if the external stimulus is strong enough ($u \gg 1$) and the temperature is sufficiently low ($\beta \gg 1$), equation 2.18 can be approximated as

$$\frac{d\bar{\sigma}_i}{dt} = -\bar{\sigma}_i + \frac{1}{\tau} \tanh\left(\beta \sum_{j \neq i}^{N} \bar{J}_{ij}\bar{\sigma}_j + u\beta\xi_i(t)\right) \approx -\bar{\sigma}_j + \frac{1}{\tau}\xi_i(t), \tag{A.5}$$

where we have used $h_i = \xi_i$.

Equation A.5 can be easily solved for σ_i:

$$\bar{\sigma}_i(t) = \frac{1}{\tau} \int_0^t \xi_i(t') e^{-\frac{t-t'}{\tau}} dt'. \tag{A.6}$$

The last previous can be rewritten as

$$\bar{\sigma}_i(t) = \frac{\int_0^t \xi_i(t')e^{-\frac{t-t'}{\tau}}dt'}{\int_0^t e^{-\frac{t-t'}{\tau}}dt'}(1 - e^{-\frac{t}{\tau}}), \tag{A.7}$$

and at this point, it is easy to see that if the same stimulus is presented for $t \gg \tau$, then equation A.7 can be approximated as follows:

$$\bar{\sigma}_i(t) \approx \xi_i(t). \tag{A.8}$$

By inserting equation A.8 into the dynamical equation for $\bar{J}_{ij}$, that is, equation 2.18, we get

$$\frac{d\bar{J}_{ij}(t)}{dt} = -\frac{1}{\tau'}\bar{J}_{ij}(t) + \frac{1}{\tau'}\xi_i(t)\xi_j(t). \tag{A.9}$$

Finally, by mapping $t \to \tau t$, the dynamical equations of the synaptic matrix becomes

$$\frac{d\bar{J}_{ij}(t)}{dt} = -\frac{\tau}{\tau'}\bar{J}_{ij}(t) + \frac{\tau}{\tau'}\xi_i(t)\xi_j(t). \tag{A.10}$$

Let us making the following ansatz for J_{ij}:

$$\bar{J}_{ij}(t) = \frac{1}{t}\int_0^t \xi_i(t')\xi_j(t')dt' + \delta_{ij}(t), \tag{A.11}$$

meaning that we separate two components in the coupling matrix (i.e., a linear integration of the external stimuli and a fluctuation contribution (the $\boldsymbol{\delta}$ matrix)). With this assumption, equation A.10 becomes an evolutive equation for the fluctuation term:

$$\begin{aligned} &\frac{d\delta_{ij}(t)}{dt} + \frac{1}{t}\xi_i(t)\xi_j(t) - \frac{1}{t^2}\int_0^t \xi_i(t')\xi_j(t')dt' \\ &= -\frac{\tau}{\tau'}\frac{1}{t}\int_0^t \xi_i(t')\xi_j(t')dt' - \frac{\tau}{\tau'}\delta_{ij}(t) + \xi_i(t)\xi_j(t). \end{aligned} \tag{A.12}$$

In the large t limit, the first two terms on the left-hand side vanish, while the integration parts reduce to the usual Hebbian kernel (since, in this case, the temporal average is equivalent to the expectation value in the pattern space because of the ergodic hypothesis). This means that in this limit, we can rewrite the entire equation in a simpler form:

$$\frac{d\delta_{ij}(t)}{dt} + \frac{\tau}{\tau'}\delta_{ij}(t) = -\frac{\tau}{\tau'}J_{ij}^{(Hebb)} + \frac{\tau}{\tau'}\xi_i(t)\xi_j(t). \tag{A.13}$$

By squaring and integrating in t the previous equation, we get

$$\int_0^{\tau'} \left[\left(\frac{d\delta_{ij}}{dt}\right)^2 + 2\frac{d\delta_{ij}}{dt}\frac{\delta_{ij}(t)\tau}{\tau'} + \left(\frac{\delta_{ij}(t)\tau}{\tau'}\right)^2\right] dt$$
$$= \int_0^{\tau'} \left[-\frac{\tau}{\tau'}J_{ij}^{(Hebb)} + \frac{\tau}{\tau'}\xi_i(t)\xi_j(t)\right]^2 dt. \tag{A.14}$$

Since $\boldsymbol{\delta}$ is a fluctuation term around zero, we can rewrite it in a transparent form as

$$\delta_{ij}(t) = 2\Delta_{ij}\cos\left(t\sqrt{\frac{\tau}{\tau'}} + \phi\right), \tag{A.15}$$

where $\Delta_{ij} \sim \Delta$ represents the fluctuation amplitude. Replacing this form directly into equation A.14 and integrating over $\phi \in [0, 2\pi]$, the oscillating terms disappear, and we achieve an estimation for the amplitude Δ:

$$2\Delta^2\frac{1+\tau'/\tau}{\tau'/\tau} = \int_0^{\tau'} \left(-\frac{\tau}{\tau'}J_{ij}^{(Hebb)} + \frac{\tau}{\tau'}\xi_i(t)\xi_j(t)\right)^2 dt. \tag{A.16}$$

Expanding the square on the right-hand side and assuming $\tau'/\tau \gg 1$, we can write the following estimation:

$$\Delta \approx \sqrt{\frac{1}{2}\frac{\tau}{\tau'}\left[1 - \left(J_{ij}^{(Hebb)}\right)^2\right]}. \tag{A.17}$$

Thus, the fluctuation amplitude around the Hebbian kernel is proportional to $(\tau/\tau')^{1/2}$. In Figure 8 we plotted using dashed lines this estimation for the fluctuations around the Hebbian kernel, and we find a perfect agreement between the theoretical prediction, equation A.17, and the numerical simulations. For general β, the same analysis allows us to quantify that

$$\Delta(\beta) \simeq \sqrt{\Delta^2\tanh^2\beta + [J_{ij}^{(Hebb)}(1-\tanh\beta)]^2}, \tag{A.18}$$

with Δ being the $\beta \to \infty$ evaluation given by equation A.17. The comparison between this prediction and the numerical analysis can be found in Figure 8 where the theoretical prediction is represented by the dashed line.

References

Agliari, E., Barra, A., Galluzzi, A., Guerra, F., & Moauro, F. (2012). Multitasking associative networks. *Physical Review Letters*, *26*(109), 268101.

Agliari, E., Barra, A., Galluzzi, A., Guerra, F., Tantari, D., & Tavani, F. (2014). Retrieval capabilities of hierarchical networks: From Dyson to Hopfield. *Physical Review Letters*, *2*(114), 028103.

Agliari, E., Barra, A., Gervasi Vidal, K., & Guerra, F. (2012). Can persistent Epstein-Barr virus infection induce chronic fatigue syndrome as a Pavlov feature of the immune response? *Journal of Biological Dynamics*, *2*(6), 740.

Agliari, E., Annibale, A., Barra, A., Coolen, C. C., & Tantari, D. (2013). Immune networks: Multitasking capabilities near saturation. *J. Phys. A.: Math. and Theor.*, *41*(46), 415003.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge University Press.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *14*(55), 123304.

Aquaro, M., Alemanno, F., Kanter, I., Durante, F., Agliari, E., & Barra, A. (2022). *Recurrent neural networks that generalize from examples and optimize by dreaming*. arXiv:2204.07954.

Benedetti, M., Ventura, E., Marinari, E., Ruocco, G., & Zamponi, F. (2022). Supervised perceptron learning vs. unsupervised Hebbian unlearning: Approaching optimal memory retrieval in Hopfield-like networks. *Journal of Chemical Physics*, *156*(10), 104107. 10.1063/5.0084219

Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2020). What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems*, *2*, 129–146.

Canew, P., Walters, E., & Kandel, E. (1981). Classical conditioning of simple withdrawal reflex in *Aplysia califonica*. *J. Neurosci.*, *1*(12), 1426–1437. 10.1523/JNEUROSCI.01-12-01426.1981, PubMed: 7320755

Coolen, A. C. C., Kühn, R., & Sollich, P. (2005). *Theory of neural information processing systems*. Oxford University Press.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *2*(57), 94.

Fachechi, A., Agliari, E., & Barra, A. (2019). Dreaming neural networks: Forgetting spurious memories and reinforcing pure ones. *Neural Networks*, *112*, 24–40. 10.1016/j.neunet.2019.01.006, PubMed: 30735914

Germain, R. N. (2001). The art of the probable: System control in the adaptive immune system. *Science*, *5528*(293), 5528.

Gormezano, I., Schneiderman, N., Deux, E., & Fuentes, I. (1962). Nictitating membrane: Classical conditioning and extinction in the albino rabbit. *Science*, *138*(3536), 33. 10.1126/science.138.3536.33

Guerra, F., & D'Autilia, R. (1996). Statistical mechanics of neural networks: Theory and applications. *Appl. Sci. Artif. Neur. Nets.*, *2*(2760), 726.

Hebb, D. (1949). *The organization of behavior*. Wiley.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. 10.1073/pnas.79.8.2554

Hopfield, J. J., Feinstein, D. I., & Palmer, R. G. (1983). "Unlearning" has a stabilizing effect in collective memories. *Nature*, *304*(5922), 158. 10.1038/304158a0

Kramar, M., & Karen, A. (2021). Encoding memory in tube diameter hierarchy of living flow network. In *Proceedings of the National Academy of Sciences*, *10*(118), 1204.

Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychol. Rev.*, *82*(4), 276–298. 10.1037/h0076778

Marinari, E. (2019). Dreaming neural networks: Forgetting spurious memories and reinforcing pure ones. *Neural Computation*, *31*, 503. 10.1162/neco_a_01162

Mézard, M., Parisi, G., Sourlas, N., Toulouse, G., & Virasoro, M. (1984). Nature of the spin-glass phase. *Physical Review Letters*, *13*(52), 1156.

Mézard, M., Parisi, G., & Virasoro, M. A. (1987). *Spin glass theory and beyond: An introduction to the replica method and its applications*. World Scientific.

Molchanov, P., Mallya, A., Tyree, S., Frosio, I., & Kautz, J. (2019). Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11264–11272).

Nishimori, H. (2001). *Statistical physics of spin glasses and information processing: An introduction*. Clarendon Press.

Osella, M., Cora, C., & Caselle, M. (2011). The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLOS Computational Biology*, *3*(7), e1001101.

Parisi, G. (1979). Infinite number of order parameters for spin-glasses. *Physical Review Letters*, *23*(43), 1754.

Parisi, G. (1990). A simple model for the immune network. In *Proceedings of the National Academy of Sciences*, *1*(87), 429.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non reinforcement. In A. H. Black & W. F. Proasky (Eds.), *Classical conditioning II*. Appleton-Century-Crofts.

Sollich, P., Tantari, D., Annibale, A., & Barra, A. (2014). Extensive parallel processing on scale-free networks. *Physical Review Letters*, *23*(113), 238106.

Thorndike, E. (1911). *Animal intelligence: Experimental studies*. Macmillan.

Tuckwell, H. C. (1988). *Introduction to theoretical neurobiology: Linear cable theory and dendritic structure* (vol. 1). Cambridge University Press.

Twitmyer, E. (1905). Knee-jerks without stimulation of the patellar tendon. *Psychol. Bull.*, *2*, 43.

Vogel, E., Castro, M., & Saavedra, M. (2004). Quantitative models of Pavlovian conditioning. *Brain Research Bulletin*, *63*, 173. 10.1016/j.brainresbull.2004.01.005

Received July 11, 2022; accepted December 23, 2022.