A mathematical analysis of the effects of Hebbian learning rules on the dynamics and structure of discrete-time random recurrent neural networks

Benoit Siri, Hugues Berry, Bruno Cessac, Bruno Delord, Mathias Quoy

A mathematical analysis of the effects of Hebbian learning rules on the dynamics and structure of discrete-time random recurrent neural networks

Benoît Siri,[1] Hugues Berry,[1,*] Bruno Cessac,[2,3,4] Bruno Delord,[5] and Mathias Quoy[6]

[1] *Team Alchemy,*
INRIA,
Parc Club Orsay Université,
4 rue J Monod,
91893 Orsay Cedex - France

[2] *Team Odyssee,*
INRIA,
2004 Route des Lucioles,
06902 Sophia Antipolis,
France

[3] *Université de Nice,*
Parc Valrose,
06000 Nice,
France

[4] *Institut Non Linéaire de Nice,*
UMR 6618 CNRS,
1361 route des Lucioles,
06560 Valbonne,
France

[5] *ANIM,*
U742 INSERM - Université P.M. Curie,
9 quai Saint-Bernard,
75005 Paris,
France

[6] *ETIS,*
UMR 8051 CNRS-Université de Cergy-Pontoise-ENSEA,

6 avenue du Ponceau,
BP 44,
95014 Cergy-Pontoise Cedex,
France

Abstract

We present a mathematical analysis of the effects of Hebbian learning in random recurrent neural networks, with a generic Hebbian learning rule including passive forgetting and different time scales for neuronal activity and learning dynamics. Previous numerical works have reported that Hebbian learning drives the system from chaos to a steady state through a sequence of bifurcations. Here, we interpret these results mathematically and show that these effects, involving a complex coupling between neuronal dynamics and synaptic graph structure, can be analyzed using Jacobian matrices, which introduce both a structural and a dynamical point of view on the neural network evolution. Furthermore, we show that the sensitivity to a learned pattern is maximal when the largest Lyapunov exponent is close to 0. We discuss how neural networks may take advantage of this regime of high functional interest.

*Corresponding author, `hugues.berry@inria.fr`

I. INTRODUCTION

The mathematical study of the effects of synaptic plasticity (or more generally learning) in neural networks is a difficult task because the dynamics of the neurons depends on the synaptic weights network, that itself evolves non trivially under the influence of neuron dynamics. Understanding this mutual coupling (and its effects on the computational efficiency of the neural network) is a key problem in computational neuroscience and necessitates new analytical approaches.

In recent years, the related field of dynamical systems interacting on complex networks has attracted vast interest. Most studies have focused on the influence of network structure on the global dynamics (for a review, see (Boccaletti *et al.*, 2006)). In particular, much effort has been devoted to the relationships between node synchronization and the classical statistical quantifiers of complex networks (degree distribution, average clustering index, mean shortest path, motifs, modularity...) (Grinstein and Linsker, 2005; Lago-Fernández *et al.*, 200; Nishikawa *et al.*, 2003). The core idea was that the impact of network topology on global dynamics might be prominent, so that these structural statistics may be good indicators of global dynamics. This assumption proved however largely wrong and some of the related studies yielded contradictory results (Hong *et al.*, 2002; Nishikawa *et al.*, 2003). Actually, synchronization properties cannot be systematically deduced from topology statistics but may be inferred from the spectrum of the network (Atay *et al.*, 2006). Most of these studies have considered diffusive coupling between the nodes (Hasegawa, 2005). In this case, the adjacency matrix has real nonnegative eigenvalues, and global properties, such as stability of the synchronized states (Barahona and Pecora, 2002) can easily be inferred from its spectral properties (see also (Atay. *et al.*, 2006; Volchenkov and Blanchard, 2007) and (Chung, 1997) for a review on mathematically rigorous results). Unfortunately, the coupling between neurons (synaptic weights) in neural networks is rarely diffusive, the corresponding matrix is not symmetric and may contain positive and negative elements. In addition, the synaptic graph structure of a neural network is usually not fixed but evolves with time, which adds another level of complexity. Hence, these results are not directly applicable to neural networks.

Discrete-time random recurrent neural networks (RRNNs) are known to display a rich variety of dynamical behaviors, including fixed points, limit cycle oscillations, quasi periodicity

and deterministic chaos (Doyon *et al.*, 1993). The effect of hebbian learning in RRNN, including pattern retrieval properties, has been explored numerically by Daucé and some of us (Dauce *et al.*, 1998). It was observed that Hebbian learning leads to a systematic reduction of the dynamics complexity (transition from chaos to fixed point by an inverse quasi-periodicity route). This property has been exploited for pattern retrieval. After a suitable learning phase the presentation of a learned pattern induces a bifurcation (e.g. from chaos to a simpler attractor such as a limit cycle). This effect is inherited via learning (it does not exist before learning), is robust to a small amount of noise, and selective (it does not occur for drastically different patterns). These effects were however neither analyzed nor really understood in (Dauce *et al.*, 1998). This work was extended to sequence learning and expoited on a robotic platform in (Daucé *et al.*, 2002).

More recently, Echo State Networks (ESN) (Jaeger and Haas, 2004) have been developed, where, as in our case, the network acts as a reservoir of resonant frequencies. However, learning only affects output links in ESN networks, while the weights within the reservoir are kept constant. Tsuda's chaotic itinerancy is an alternative way for linking different attractors with different inputs (Tsuda, 2001). In this model, weights are initially fixed in a Hopfield-like manner (and are thus symmetric) and a chaotic dynamics successively explores the different fixed point attractors. In this scheme, each input constitutes an different initial condition that leads to one attractor of the *same* dynamical system, whereas in (Dauce *et al.*, 1998), each (time-constant) input leads to a *different* dynamical system.

In the current state of the art, there is a relatively large number of models, observations and applications of Hebbian learning effects in neural networks, but considerably less mathematical results. Mathematical analysis is however necessary to classify the many variants of Hebbian learning rules according to the effects they produce. The present paper is one step further towards this aim. Using methods from dynamical systems theory, we analyze the effects of a generic version of Hebbian learning proposed in (Hoppensteadt and Izhikevich, 1997) on the neural network model numerically studied in (Dauce *et al.*, 1998) with spontaneous (i.e. before learning) chaotic dynamics.

We essentially classify the effects into three families:

(i) Topological: the structure of the synaptic weight network evolves, implying prominent (e.g. cooperative) effects on the dynamics.

(ii) Dynamical: the dynamical complexity (measured e.g. by the maximal Lyapunov ex-

ponent or the Kolmogorov-Sinai entropy) reduces during Hebbian learning. This effect is mathematically analyzed and interpreted. Especially, we provide a rigorous upper bound on the maximal Lyapunov exponent and identify two major causes for this reduction: the decay of the norm of the synaptic weight matrix and the saturation of neurons.

(iii) Functional: Focusing on the network response to a learned pattern, we show that there is a learning stage at which the response is maximal, in the sense that it generates a drastic change of the neuronal dynamics (i.e. a bifurcation). This stage precisely corresponds to vanishing of the maximal Lyapunov exponent.

Some of these results may appear neither "new" nor "surprising" for the neural networks community. For example, (ii) and (iii) have already been reported in (Dauce *et al.*, 1998). However, the results were mainly numerical while the present paper proposes a mathematical framework and formal tools to analyze them. Moreover, a direct consequence of (iii) is that the response of the neural network to a learned pattern is maximal at the "edge of chaos" (where the maximal Lyapunov exponent vanishes).

The claim that the neural network response is maximal close to a bifurcation is common in the neural network community (Langton, 1990). Similarly, (Hoppensteadt and Izhikevich, 1997) already pointed out the necessity for some neurons to lie close to a bifurcation point in order to have relevant computational capacities. As a matter of fact, an analysis of the effects of a Hopfield-Hebb rule was performed in this book with neurons close to codimension one *fixed-point* bifurcations.

We go a step further in the present paper and show that a similar conclusion holds for a neural network in a *chaotic regime*. Conceptually, the analysis of (Hoppensteadt and Izhikevich, 1997) could be extended to chaotic systems [1] (Cessac and Samuelides, 2007). However, the analytic treatment of the chaotic case is really challenging. Hence, bifurcation analysis of fixed points (or periodic orbits) uses a linear analysis via Jacobian matrices, which is usually

[1] A cornerstone of the analysis in (Hoppensteadt and Izhikevich, 1997) is the use of Hartman-Grobman theorem, and its consequence, namely that neural networks have non trivial properties only if some neurons are close to a bifurcation point. In some sense, this analysis can be extended to uniformly hyperbolic dynamical systems, a small subset of chaotic systems (though it has never been done). In addition, it is absolutely not guaranteed that chaotic RRNNs are uniformly hyperbolic, since one does not control the spectrum of the Jacobian matrices. The main difficulty is to characterize this spectrum on the ω-limit set (and not in the whole phase space). As a matter of fact, we do not know of any mathematical result with regard to this aspect.

considered non-applicable to chaotic systems where nonlinear effects and initial conditions sensitivity are prominent. Nevertheless, recent results by Ruelle (Ruelle, 1999) on linear response theory, formally extended to chaotic neural networks (Cessac and Sepulchre, 2006, 2007), show that a linear analysis is indeed possible if one uses an average of the Jacobian matrix along its chaotic trajectory. The associated linear response operator provides a deep insight into the links between topology and dynamics in chaotic neural networks. Incidentally, it shows that the relevant matrix is not the weight matrix (as would be expected), but the linear response matrix, which reduces, in the present context, to the ergodic average of the Jacobian matrix along its trajectory [2].

Though the main results in this paper are mathematical, we also use some numerical simulations. They were necessary because mathematical results are obtained using a limit where time goes to infinity, which is not operational in numerical situations. Moreover, the central rigorous results we obtain provide upper bounds, whose quality had to be checked numerically.

The paper is organized as follows. We first present the model and the generic framework for neuronal dynamics and learning rules in section II. The following sections are devoted to the analysis of the model. In section III, we present analytical results explaining the evolution of dynamics during learning using mathematical tools from dynamical systems and graph theory. These analytical results are confirmed by extensive numerical simulations. Section IV focuses on functional effects related to network sensitivity to the learned pattern. We finally discuss our results in the last section (V).

II. GENERAL FRAMEWORK

A. Model description

We consider firing-rate recurrent neural networks with N point neurons and discrete-time dynamics, where learning may occur on a different (slower) time scale than neuron dynamics. Synaptic weights are thus constant for $\tau \geq 1$ consecutive dynamics steps, which defines a "learning epoch". The weights are then updated and a new learning epoch begins.

[2] This result, which may a posteriori appear obvious to readers familiar with dynamical systems theory is in fact highly non trivial and requires Ruelle's linear response theory to be properly justified.

We denote by t the update index of neuron states (neuron dynamics) inside a learning epoch, while T indicates the update index of synaptic weights (learning dynamics). Call $x_i^{(T)}(t) \in [0,1]$ the mean firing rate of neuron i, at time t within the learning epoch T. Set $\mathbf{x}^{(T)}(t) = \left[x_i^{(T)}(t)\right]_{i=1}^{N} \in [0,1]^N$. Denote by $\mathbf{F}$ the function $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^N$ such that $F_i(\mathbf{x}) = f(x_i)$ where f is a sigmoidal transfer function (e.g. $f(x) = (1 + \tanh(gx)/2)$). Let $\mathcal{W}^{(T)}$ be the matrix of synaptic weights at the T-th learning epoch. Then the discrete time neuron dynamics writes:

$$\mathbf{x}^{(T)}(t+1) = \mathbf{F}\left[\mathbf{u}^{(T)}(t)\right] = \mathbf{F}\left[\mathcal{W}^{(T)}\mathbf{x}^{(T)}(t) + \boldsymbol{\xi}\right], \tag{1}$$

$\mathbf{u}^{(T)}(t)$ is called "the local field (or the synaptic potential), at neuron time t and learning epoch T". The output gain g tunes the nonlinearity of the function and mimics the reactivity of the neuron. The vector $\boldsymbol{\xi} = (\xi_i)_{i=1}^N$ is the "pattern" to be learned. The initial weight matrix $\mathcal{W}^{(1)}$ is randomly and *independently* sampled from a Gaussian law with mean 0 and variance $1/N$. Hence, the synaptic weights matrix $\mathcal{W}^{(T)} = \left(W_{ij}^{(T)}\right)_{i,j=1}^{N}$ typically contains positive (excitation), negative (inhibition) or null (no synapse) elements and is asymmetric ($W_{ij}^{(T)} \neq W_{ji}^{(T)}$).

The network can display different dynamical regimes (chaos, (quasi-) periodicity, fixed point), depending on these parameters (Dauce *et al.*, 1998). In the present study, the parameters were set so that the spontaneous dynamics (i.e. the network dynamics at $T = 1$) was chaotic. At the end of every learning epoch, the neuron dynamics indices are reset, and $x_i^{(T+1)}(0) = x_i^{(T)}(\tau), \forall i$.

The learning rules we study conform to Hebb's postulate (Hebb, 1948). Specifically, we define the following generic formulation (Hoppensteadt and Izhikevich, 1997):

$$\mathcal{W}^{(T+1)} = \lambda \mathcal{W}^{(T)} + \frac{\alpha}{N}\Gamma^{(T)} \tag{2}$$

where α is the learning rate and $\Gamma^{(T)}$ a Hebbian function (see below). The first term in the right-hand side (RHS) member accounts for passive forgetting, i.e. $\lambda \in [0,1]$ is the forgetting rate. If $\lambda < 1$ and $\Gamma_{ij} = 0$ (i.e. both pre- and postsynaptic neurons are silent, see below), eq. (2) leads to an exponential decay of the synaptic weights (hence passive forgetting), with a characteristic rate $\frac{1}{|\log(\lambda)|}$ (see discussion, section V). Note that there is no forgetting when $\lambda = 1$. The second term in the RHS member generically accounts for activity-dependent plasticity, i.e. the effects of the pre- and postsynaptic neuron firing rates. We focus here on

learning rules where this term depends on the *history* of activities[3], i.e.

$$\Gamma_{ij}^{(T)} = h(\tilde{x}_i^{(T)}, \tilde{x}_j^{(T)}) \tag{3}$$

where $\tilde{x}_i^{(T)} = \left\{ x_i^{(T)}(t) \right\}_{t=1}^{\tau}$ is the trajectory of neuron i firing rate. In the present paper, as a simple example, we shall associate to the history of neuron i rate an activity index $m_i^{(T)}$:

$$m_i^{(T)} = \frac{1}{\tau} \sum_{t=1}^{\tau} (x_i^{(T)}(t) - d_i) \tag{4}$$

where $d_i \in [0, 1]$ is a threshold and h is a function of $m_i^{(T)}$ and $m_j^{(T)}$.

The neuron is considered active during learning epoch T whenever $m_i^{(T)} > 0$, and silent otherwise. d_i does not need to be explicitly defined in the mathematical study. In numerical simulations however, we set it to 0.50, $\forall i$. Definition (4) actually encompasses several cases. If $\tau = 1$, weight changes depend only on the instantaneous firing rates, while if $\tau \gg 1$, weight changes depend on the mean value of the firing rate, averaged over a time window of duration τ in the learning epoch. In many aspects the former case can be considered as plasticity, while the latter may be related to meta-plasticity (Abraham and Bear, 1996). In this paper, we set $\tau \to \infty$ for the mathematical analysis. We chose a value of $\tau = 10^4$ in numerical simulations, which corresponds to the time scale ratio between neuronal dynamics (ms) and synaptic plasticity (10 s) (see (Delord *et al.*, 2007)). Importantly, note that other values of τ (including $\tau = 1$) have been tested in simulations and did not lead to any qualitative change in the network behavior, although some integration lag effects were observed for very small values. Therefore, the exact value of τ has no impact on the major conclusions of the present paper.

The explicit definition of the function h in eq.(3) is constrained by Hebb's postulate for plasticity. This postulate is somewhat loosely defined, so that many implementations are possible in our framework. Our choice is guided by the following points (Hoppensteadt and Izhikevich, 1997):

---

[3] As a matter of fact, note that $\Gamma_{ij}^{(T)}$ is a function of the trajectories $\tilde{x}_i^{(T)}, \tilde{x}_j^{(T)}$, which depend on $\mathcal{W}^{(T)}$, which in turn depends on $\Gamma_{ij}^{(T-1)}$... Hence, the set of synaptic weights at time $T+1$ and the dynamics of the corresponding neurons are functions of the *whole history* of the system. In this respect, we address a very untypical and complex type of dynamical systems where the flow at time t is a function of the past *trajectory* and not only a function of the previous state. (In the context of stochastic processes, such systems are called "chains with complete connections" by opposition to (generalized) Markov processes). This induces rich properties such as a wide learning-induced *variability* in the network response to a given stimulus, with the same set of initial synaptic weights, simply by changing the initial conditions.

1. $h > 0$ whenever post-synaptic (i) and pre-synaptic (j) neurons are active, as in long-term potentiation (LTP).

2. $h < 0$ whenever i is inactive and j is active, corresponding to homosynaptic long-term depression (LTD).

3. $h = 0$ whenever j is inactive. This point is often considered as a corollary to Hebb's rule (Hoppensteadt and Izhikevich, 1997). Moreover, it renders the learning rule asymmetric and excludes the possibility that dynamics changes induced by learning could be due to weight symmetrization. This hypothesis however formally excludes heterosynaptic LTD (Bear and Abraham, 1996), which would correspond to $h < 0$ for i active and j inactive. However, most of the results presented herein remain valid in the presence of heterosynaptic LTD (see section V for a discussion).

Although these settings are sufficient for mathematical analysis, h has to be more precisely defined for numerical simulations. Hence, for the simulations, we set an explicit implementation of $\Gamma^{(T)}$ such that :

$$\mathcal{W}^{(T+1)} = \lambda \mathcal{W}^{(T)} + \frac{\alpha}{N} \mathbf{m}^{(T)} \left[\mathbf{m}^{(T)} H(\mathbf{m}^{(T)}) \right]^{+} \tag{5}$$

where $\mathbf{m}^{(T)} = \left[m_i^{(T)} \right]_{i=1}^{N}$, $H(x)$ is the Heaviside function, $H(\mathbf{m}^{(T)}) = \left[H(m_i^{(T)}) \right]_{i=1}^{N}$, $\mathbf{m}^{(T)} H(\mathbf{m}^{(T)})$ is the vector of components $m_i^{(T)} H(m_i^{(T)})$ and $+$ denotes the transpose. Finally, in the simulations, we forbid weights to change their sign, and self-connections $W_{ii}^{(T)}$ stay to 0 (note however that these settings do not influence qualitatively the results presented here).

For the purpose of the present paper, the exact value of this input pattern $\boldsymbol{\xi}$ is not very important, as soon as its maximal amplitude remains small with respect to the neuron maximal firing rate. Here, we used $\xi_i = 0.010 \sin(2\pi i/N) \cos(8\pi i/N)$, $\forall i = 1 \dots N$ in all numerical simulations. The main rationale for this choice is that this pattern is easily identified by eyes when the ξ_is are plotted against i, which is particularly helpful when interpreting alignment results, such as in fig. 3.

Equations (1) & (5) define a dynamical system where two distinct processes (neuron dynamics and synaptic network evolution) interact with distinct time scales. This results in a complex interwoven evolution where neuronal dynamics depends on the synaptic structure

and synapses evolve according to neuron activity. On general grounds, this process has a memory that is *a priori* infinite and the state of the neural network depends on the past history.

B. Analysis tools

One possible approach to topology and dynamics interactions in neural networks consists in searching structural cues in the synaptic weight matrix that may be informative of specific dynamical regimes. The weight matrix is expected to carry information about the *functional* network. However, it can be easily shown that the synaptic weight matrix is not sufficient to analyze the relationship between topology and dynamics in neural networks such as (1). A standard procedure for the analysis of nonlinear dynamical systems starts with a *linear analysis*. This holds e.g. for stability and bifurcation analysis but also for the computation of indicators such as Lyapunov exponents. The key object for this analysis is the Jacobian matrix. In our case, it writes:

$$D\mathbf{F}_{\mathbf{X}} = \Lambda(\mathbf{u})\mathcal{W}, \tag{6}$$

with:

$$\Lambda_{ij}(\mathbf{u}) = f'(u_i)\delta_{ij}. \tag{7}$$

Interestingly enough, the Jacobian matrix generates a graph structure that can be interpreted in causal terms (see Appendix F for more details). Applying a small perturbation δ_j to x_j, the induced variation on x_i is given, to the linear order, by $f'(u_i)W_{ij}\delta_j$. Therefore, the induced effect, on neuron i, of a small variation in the state of neuron j is not only proportional to the synaptic weight W_{ij}, *it also depends on the state of neuron i via f'*. For example, if $|u_i|$ is very large (neuron "saturation"), f' is very close to 0 and the perturbation on any x_j has no effect on x_i.

From this very simple argument we come to the conclusion that the Jacobian matrix displays more information than the synaptic weight matrix:

1. The "causal" graph induced by the Jacobian matrix leads to the notion of cooperative systems, introduced by Hirsch in (Hirsch, 1989) and widely studied in the field of genetic networks (Gouzé, 1998; Thomas, 1981). This notion is also useful in the present context (see appendix F).

2. The Jacobian matrix allows to perform local bifurcation analysis. In our case, this provides information about the effect of pattern presentation before and after learning (section IV).

3. The Jacobian matrix allows to define Lyapunov exponents, which are used to measure the degree of chaos in a dynamical system.

4. The Jacobian matrix allows to define the notion of linear response in chaotic systems (Cessac and Sepulchre, 2006, 2007; Ruelle, 1999), which extends the notion of causal graph to nonlinear systems with chaotic dynamics (see in section IV).

III. DYNAMICAL VIEWPOINT

As explained in the introduction and reported in (Dauce *et al.*, 1998), Hebbian learning rules can lead to reduction of the dynamics complexity from chaos to quasiperiodic attractor, limit cycle and fixed point, due to the mutual coupling between weights evolution and neuron dynamics. The aim of this section is to provide a theoretical interpretation of this reduction of complexity for a more general class of Hebbian learning rules than those considered in (Dauce *et al.*, 1998).

A. Entropy reduction.

1. Evolution of the weight matrix.

From eq. (2) it is easy to show by recurrence that:

$$\mathcal{W}^{(T+1)} = \lambda^T \mathcal{W}^{(1)} + \frac{\alpha}{N} \sum_{n=1}^{T} \lambda^{T-n} \Gamma^{(n)}. \tag{8}$$

The evolution of the weight matrix under the influence of the generic learning rule eq.(2) originates from two additive contributions. If $\lambda < 1$, the "direct" contribution of $\mathcal{W}^{(1)}$ to $\mathcal{W}^{(T+1)}$ (the first term in the RHS member) decays exponentially fast. Hence the effect of λ is that the initial synaptic structure is progressively forgotten, offering the possibility to entirely "rewire" the network in a time scale proportional to $\frac{1}{|\log(\lambda)|}$. The second RHS term of eq. (8) corresponds to the new synaptic structure emerging with learning and replacing

the initial one (which fades away exponentially fast). Importantly, this second term includes contributions from each previous matrices $\Gamma^{(n)}, \quad \forall n \leq T$ (with an exponentially decreasing contribution λ^{T-n}). Hence, the emerging weights structure depends on *the whole history of the neuronal dynamics*.

If $\lambda < 1$, one expects to reach a stationary regime where synaptic weights do not evolve anymore: both matrices $\mathcal{W}^{(T)}$ and $\Gamma^{(T)}$ are expected to stabilize at long learning epochs to constant values ($\lim_{T\to\infty} \mathcal{W}^{(T)} = \mathcal{W}^{(\infty)}$ and $\lim_{T\to\infty} \Gamma^{(T)} = \Gamma^{(\infty)}$). This means that, if $\lambda < 1$, the dynamics settle at long learning epochs onto a stable attractor that is not modified by further learning of a given stimulus. The existence of such a stationary distribution is provided by the sufficient condition:

$$\mathcal{W}^{(\infty)} = \frac{\alpha}{N(1-\lambda)}\Gamma^{(\infty)}. \tag{9}$$

We show in appendix B that, assuming moderate hypotheses on h (eq. 3), $\|\Gamma^{(T)}\|$ can be upper-bounded, $\forall T$, by a constant NC, so that $\|\mathcal{W}^{(\infty)}\| \leq \alpha C/(1-\lambda)$. From eq.(8), an upper bound for the norm of $\mathcal{W}^{(T)}$ is trivially found:

$$\|\mathcal{W}^{(T+1)}\| \leq \lambda^T \|\mathcal{W}^{(1)}\| + \frac{\alpha}{N}\sum_{n=1}^{T} \lambda^{T-n}\|\Gamma^{(n)}\|, \tag{10}$$

where $\|\|$ is the operator norm (induced e.g. by Euclidean norm). Hence,

$$\|\mathcal{W}^{(T+1)}\| \leq \lambda^T \|\mathcal{W}^{(1)}\| + \alpha C\frac{1-\lambda^T}{1-\lambda} \leq \lambda^T \|\mathcal{W}^{(1)}\| + \alpha C\frac{1}{1-\lambda}. \tag{11}$$

This result shows that the major effect of the Hebbian learning rule we study may consist in an exponentially fast contraction of the norm of the weight matrix, which is due to the term λ, i.e. to passive forgetting ($\lambda < 1$). Note also that if $\lambda = 1$, this term may diverge, leading to a divergence of $\mathcal{W}^{(T)}$. Therefore, in this case, one has to add an artificial cut-off to avoid this unphysical divergence.

These analytical results need not to be "confirmed" by numerical simulations, as they are rigorous. However, they only provide an upper bound that can be rough, while simulations allows to evaluate how far from the exact values these bounds are.

Let $s_i^{(T)}$ be the eigenvalues of $\mathcal{W}^{(T)}$, ordered such that $|s_1^{(T)}| \geq |s_2^{(T)}| \geq \ldots \geq s_i^{(T)} \geq \ldots$. Since $|s_1^{(T)}|$, the spectral radius of $\mathcal{W}^{(T)}$, is smaller than $\|\mathcal{W}^{(T)}\|$ one has from eq.(11):

$$|s_1^{(T+1)}| \leq \lambda^T \|\mathcal{W}^{(1)}\| + \alpha C\frac{1}{1-\lambda}. \tag{12}$$

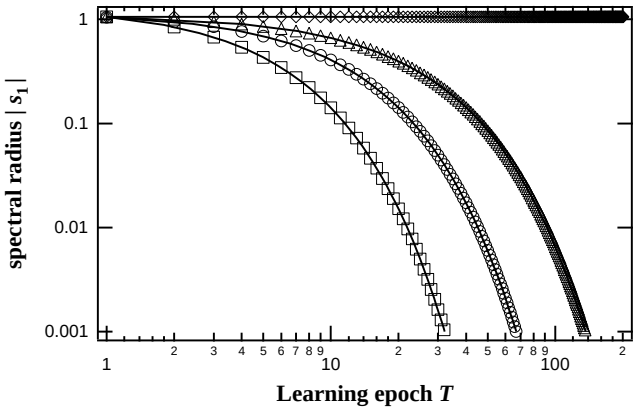

FIG. 1 The Hebbian learning rule eq.(5) contracts the spectral radius of $\mathcal{W}$. The evolution during learning of the norm of $\mathcal{W}$ largest eigenvalue, $|s_1^{(T)}|$ is plotted on a log-log scale for, from bottom to top, $\lambda = 0.80$ (squares), 0.90 (circles), 0.95 (triangles) or 1.00 (diamonds). Each value is an average over 50 realizations with different initial conditions (initial weights and neuron states). Standard deviations are smaller than the symbols. Black full lines are plots of exponential decreases with equation $g(T) = |s_1^{(1)}|\lambda^T$.

This equation predicts a bound on the spectral radius that contracts exponentially fast with time, under the control of the forgetting rate λ. Figure 1 shows the evolution of the spectral radius of $\mathcal{W}^{(T)}$ for different values of λ during numerical simulations (open symbols). The results show that the spectral radius indeed decays exponentially fast. Moreover, we also plot on this figure (full lines) exponential decays according to the first RHS member of eq.(12), i.e. $g(T) = |s_1^{(1)}|\lambda^T$. The almost perfect agreement with the measurements tells us that the bound obtained in eq.(12) actually represents a very good estimate of the value of $|s_1^{(T)}|$.

2. Jacobian matrices.

Let $\mathbf{x} \in [0,1]^N$. A bound for the spectral radius of $D\mathbf{F}_{\mathbf{X}}^{(T)}$ can easily be derived from 11 and 6. Call $\mu_i^{(T)}(\mathbf{x})$ the eigenvalues of $D\mathbf{F}_{\mathbf{X}}^{(T)}$ ordered such that $|\mu_1^{(T)}(\mathbf{x})| \geq |\mu_2^{(T)}(\mathbf{x})| \geq \ldots \geq |\mu_i^{(T)}(\mathbf{x})| \geq \ldots$. One has, $\forall \mathbf{x}$:

$$|\mu_1^{(T)}(\mathbf{x})| \leq \|D\mathbf{F}_{\mathbf{X}}^{(T)}\| \leq \|\Lambda(\mathbf{u}^{(T)})\|\|\mathcal{W}^{(T)}\|. \tag{13}$$

Since $\|\Lambda(\mathbf{u}^{(T)})\| = \max_i f'(u_i^{(T)})$ (Λ is diagonal and $f' > 0$), one finally gets

$$|\mu_1^{(T)}(\mathbf{x})| \leq \max_i f'(u_i^{(T)})\|\mathcal{W}^{(T)}\|. \tag{14}$$

Therefore, we obtain a bound on the spectrum of $D\mathbf{F}_{\mathbf{X}}^{(T)}$ that can be contracted by two effects: the contraction of the spectrum of $\mathcal{W}^{(T)}$ and/or the decay of $\max_i f'(u_i)$ related to

the saturation of neuronal activity. Indeed, $f'(u_i)$ is small if x_i is saturated to 0 or 1 (i.e. $|u_i|$ is large), but large whenever $|u_i|$ is intermediate, i.e. falls into the central, pseudo-linear part of the sigmoid $f(u_i)$. We have already evidenced above that $\lambda < 1$ yields to a decrease of $\|\mathcal{W}^{(T)}\|$. Note that even if $\lambda = 1$ (no passive forgetting) and $\mathcal{W}^{(T)}$ diverges, then $\mathbf{u}^{(T)}$ diverges as well, leading $\max_i f'(u_i^{(T)})$ to vanish, thus decreasing the spectral radius of the Jacobian matrix. Hence, if the initial value of $|\mu_1^{(T)}(\mathbf{x})|$ is larger than 1 and the bound in eq.(14) represents an accurate estimate of $|\mu_1^{(T)}(\mathbf{x})|$, eq.(14) predicts that the latter may decrease down to a value < 1. We are dealing here with discrete time dynamical systems, so that the value $|\mu_1^{(T)}(\mathbf{x})| = 1$ locates a *bifurcation* of the dynamical system. Hence, eq.(14) opens up the possibility that learning drives the system through bifurcations. Again, simulations (fig. 4) show that the bound obtained in eq. 14 is indeed very close to the actual value of the Jacobian matrix spectral radius. As will be shown later (section IV), this point is of great importance from a functional viewpoint.

3. A bound on the maximal Lyapunov exponent.

Eq. (14) depends on $\mathbf{x}$ and cannot provide information on the *typical* behavior of the dynamical system. This information is provided by the computation of the largest Lyapunov exponent (see appendix A for definitions). In the present setting, the largest Lyapunov exponent, $L_1^{(T)}$ depends on the learning epoch T. It can be computed exactly before learning in the thermodynamic limit $N \to \infty$, because W_{ij}'s are i.i.d. random variables (Cessac, 1995) and it can be showed that it is positive provided g is sufficiently large[4]. However, because the weights deviate from i.i.d. random distribution under the influence of Hebbian learning, the evolution of $L_1^{(T)}$ cannot be computed analytically as soon as $T > 1$. Nevertheless, the following theorem (proven in appendix C) yields a useful upper-bound of $L_1^{(T)}$:

Theorem 1

$$L_1^{(T)} \leq \log(\|\mathcal{W}^{(T)}\|) + \left\langle \log(\max_i f'(u_i)) \right\rangle^{(T)} . \tag{15}$$

[4] In the limit $N \to \infty$ and for random i.i.d. weights with 0 mean and variance $\frac{1}{N}$, $|\mu_1^{(T)}(\mathbf{x})|$ converges almost surely to a value proportional to g, the proportionality factor depending on the explicit form of f (Cessac, 1994; Girko, 1984)

where $\langle \log(\max_i f'(u_i)) \rangle^{(T)}$ *denotes the time average of* $\log(\max_i f'(u_i))$*, in the learning epoch* T *(see appendix for details).*

This theorem emphasizes the two main effects that may contribute to a decrease of $L_1^{(T)}$. The first term in the RHS member states that the upper bound on $L_1^{(T)}$ decreases if the norm of the weights matrix $\|\mathcal{W}^{(T)}\|$ decreases during learning. The second term is related to the saturation of neurons. However, the main difference with eq. (14) is that we now have an information on how saturation effects act *on average* on dynamics, via $\log(f')$. The second term in the RHS member is positive if some neurons have an average $\log(f')$ larger than 1 (that is, they are mainly dominated by amplification effects corresponding to the central part of the sigmoid) and becomes negative when all neurons are saturated on average.
In any case, it follows that if learning increases the saturation level of neurons or decreases the norm of the weights matrix $\|\mathcal{W}^{(T)}\|$, then the result can be a decay of $L_1^{(T)}$ (if the bound is a good estimate), thus a possible transition from chaotic to simpler attractors. A canonical measure of dynamical complexity is the Kolmogorov-Sinai (KS) entropy which is bounded from above by the sum of positive Lyapunov exponents. Therefore, if the largest Lyapunov exponent decreases, KS entropy and the dynamical complexity decrease. On numerical grounds we observe the following. Fig. 2A shows measurements of $L_1^{(T)}$ during numerical simulations with different values of the passive forgetting rate λ. Its initial value is positive because we start our simulations with chaotic networks ($L_1^{(1)} \approx 0.21 \pm 0.10$). The Hebbian learning rule eq.(5) indeed leads to a rapid decay of $L_1^{(T)}$, whose rate depends on λ. Hence $L_1^{(T)}$ shifts quickly to negative values, confirming the decrease of the dynamical complexity that could be inferred from visual inspection of temporal traces of the network averaged activity (fig. 2B).
To conclude, our mathematical framework indicates a systematic decay of $L_1^{(T)}$ induced by passive forgetting and/or increased neuronal saturation. This decay explains the decreasing dynamical complexity from chaos to steady state that is observed numerically.

B. Neuron activity.

We now present analytical results concerning the evolution of individual neuron activity. Application of the learning rule eq.(2) changes the structure of the attractor from one learning epoch to the other. The magnitude of this change can be measured by changes

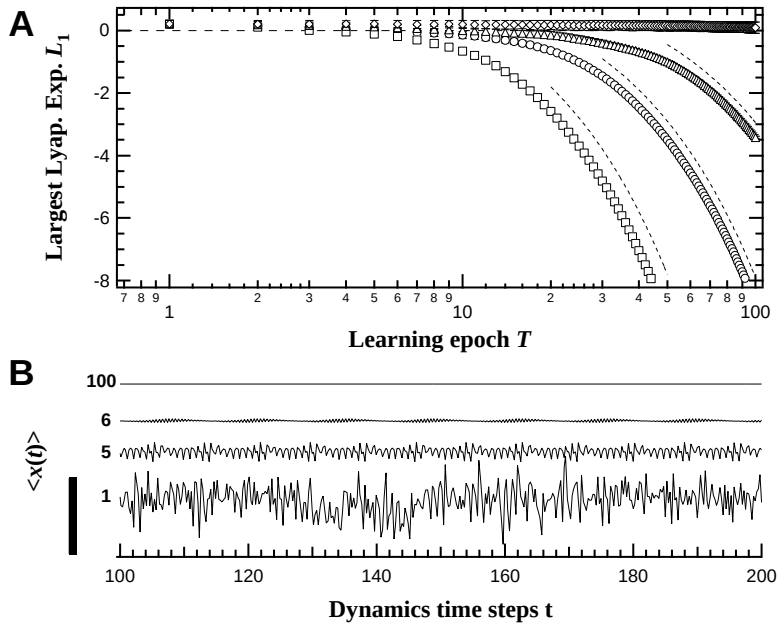

FIG. 2 The Hebbian learning rule eq.(5) induces reduction of the dynamics complexity from chaotic to periodic and fixed point. (*A*) Evolution of the largest Lyapunov exponent L_1 during 100 learning epochs for, from bottom to top, $\lambda = 0.80$ (squares), 0.90 (circles), 0.95 (triangles) or 1.00 (diamonds). Each value is an average over 50 realizations with different initial conditions (initial weights and neuron states). Bars are standard deviations (and are mostly smaller than symbol size). The dashed lines illustrate decays of the form $g(T) \propto T\log(\lambda)$ (see text). (*B*) Examples of network dynamics when learning is stopped at epoch (from bottom to top) $T = 1$ (initial conditions, chaos), 5 (limit cycle), 6 (simpler limit cycle) or 100 (fixed point). These curves show the network-averaged state $\langle x^{(T)}(t)\rangle = 1/N \sum_{i=1}^{N} x_i^{(T)}(t)$ and are shifted on the y-axis for clarity. The height of the vertical bar represents an amplitude of 0.1. $N = 100$ and all other parameters are as in fig. 1.

in the average value of some relevant observable such as neuron activity (more generally, learning induces a variation in the SRB measure $\rho^{(T)}$, see appendix A). Let $\delta\rho^{(T+1)}(\mathbf{x})$ be the variation of the average activity $\mathbf{x}$ between learning epoch T and $T+1$. By definition (see appendix A):

$$\delta\rho^{(T+1)}(\mathbf{x}) = \langle \mathbf{x} \rangle^{(T+1)} - \langle \mathbf{x} \rangle^{(T)} . \tag{16}$$

We show in appendix D that the average value of the neuron local field, $\mathbf{u}$, at learning epoch T depends on four additive terms:

$$\langle \mathbf{u} \rangle^{(T+1)} = \lambda^T \langle \mathbf{u} \rangle^{(1)} + (1-\lambda^T)\boldsymbol{\xi} + \lambda \sum_{n=1}^{T} \lambda^{T-n} \mathcal{W}^{(n)} \delta\rho^{(n+1)}(\mathbf{x}) + \frac{\alpha}{N} \sum_{n=1}^{T} \lambda^{T-n} \Gamma^{(n)} \langle \mathbf{x} \rangle^{(n+1)} . \tag{17}$$

Provided that $\lambda < 1$, as $T \to +\infty$, time averages of observables converge to a constant. So that $\delta\rho^{(T)}(\mathbf{x}) \to 0$ and $\lim_{T\to+\infty} \langle \mathbf{x} \rangle^{(T)} = \langle \mathbf{x} \rangle^{(\infty)}$. Therefore, asymptotically:

$$\langle \mathbf{u} \rangle^{(\infty)} = \boldsymbol{\xi} + \mathbf{H}^{(\infty)}, \tag{18}$$

where:

$$\mathbf{H}^{(\infty)} = \mathcal{W}^{(\infty)} \langle \mathbf{x} \rangle^{(\infty)} = \frac{\alpha}{N(1-\lambda)} \Gamma^{(\infty)} \langle \mathbf{x} \rangle^{(\infty)} . \tag{19}$$

Therefore, the asymptotic local field ($\langle \mathbf{u} \rangle^{(\infty)}$) is the sum of the *stimulus* (input pattern) plus an additional vector $\mathbf{H}^{(\infty)}$ which accounts for the history of the system. Note that equations (18), (19) characterize the asymptotic regime $T \to \infty$ which usually corresponds to a fixed-point (see fig 2) with limited dynamical and functional interest (see e.g. fig. 4). On intermediate time scales, eq. (17) must be considered. It shows that the local field $\mathbf{u}$ contains a constant component (the input pattern) as well as additional (history-dependent) terms whose relative contribution cannot systematically be predicted.

Figure 3 shows numerical simulations of the evolution of the local field $\mathbf{u}$ during learning. Clearly, while the initial values are random, the local field (thin full line) shows a marked tendency to converge to the input pattern (thick dashed line) after as soon as 10 learning epochs. The convergence is complete after $\approx$ 60 learning epochs. An additional term corresponding to $\mathbf{H}^{(\infty)}$ is observed numerically (but is hardly visible in the normalized representations of fig. 3). This last term has an interesting structure in the case of the learning rule (3). Indeed, in this case:

$$\mathbf{H}^{(\infty)} = \frac{\alpha}{N(1-\lambda)} \mathbf{m}^{(\infty)} \left[\mathbf{m}^{(\infty)} H(\mathbf{m}^{(\infty)}) \right]^{+} \langle \mathbf{x} \rangle^{(\infty)} ,$$

so that:

$$H_i^{(\infty)} = \frac{\alpha}{N(1-\lambda)} \eta m_i^{(\infty)} \tag{20}$$

where :

$$\eta = \sum_{j, m_j^{(\infty)} > 0} m_j^{(\infty)} x_j^{(\infty)} = \sum_{j, x_j^{(\infty)} > d_j} (x_j^{(\infty)} - d_j) x_j^{(\infty)}, \tag{21}$$

can be interpreted as an *order parameter*. A large positive η means that neurons are mainly saturated to 1, while a small η corresponds to neuron whose average activity is close to d_i. Note that η is related to a set of self-consistent equations. Indeed, since $x_i = f(u_i)$ one has:

$$< u_i >^{(\infty)} = \xi_i + \frac{\alpha}{N(1-\lambda)} \eta \left[\langle f(u_i) \rangle^{(\infty)} - d_i \right] \tag{22}$$

In the case where this constant asymptotic attractor is a fixed point (i.e. the attractor with smallest complexity), one has:

$$u_i^* = \xi_i + \frac{\alpha}{N(1-\lambda)} \eta (f(u_i^*) - d_i), \tag{23}$$

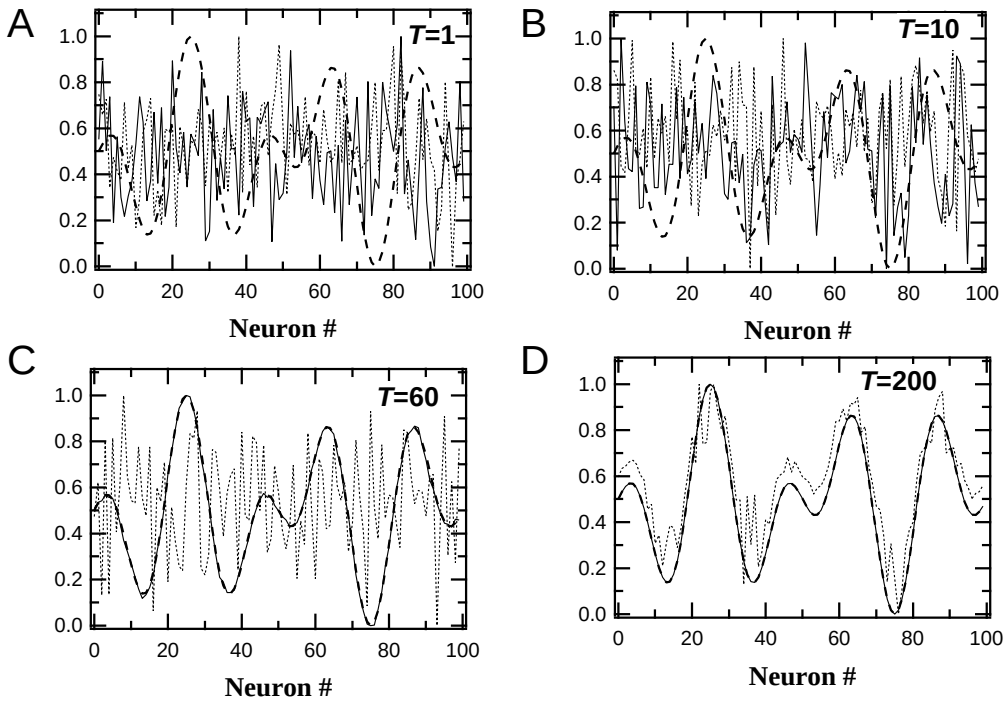

FIG. 3 The local field $\mathbf{u} = \boldsymbol{\xi} + \mathcal{W}\mathbf{x}$ (thin full line) and the real part of the first eigenvector of the Jacobian matrix (thin dotted line) converge to the input pattern $\boldsymbol{\xi}$ (thick dashed line) at intermediate-to-long learning epochs. Snapshot are presented at $T = 1$ (A, initial conditions), $T = 10$ (B), $T = 60$ (C) and $T = 200$ (D) learning epochs. Each curve plots averages over 50 realizations (standard deviations are omitted for clarity), vectors have been normalized to $[0, 1]$ for clarity. All other parameters as in fig. 1

where $\mathbf{u}^*$ and $\mathbf{x}^*$ denote the values of $\mathbf{u}$ and $\mathbf{x}$, respectively, on the fixed point attractor. Here, the set of N nonlinear self-consistent equations (22) includes both a local (u_i^∞) and a global term (the order parameter η). Assume that we slightly perturb the system, for example by removing the stimulus ξ_i for some neurone i. If the system (22) is away from a bifurcation point, this perturbation is expected to result in only a slight change in u_i^*. Alternatively, if a bifurcation occurs, a dramatic change in u_i^* can take place. This local modification of activity may in turn yield a big change in η, which corresponds to a global (i.e. network-wide) modification of activity, through a some avalanche-like mechanism. On practical grounds this means that presentation or removal of some parts of the input pattern may induce a drastic change of the dynamics of the network.

IV. FUNCTIONAL VIEWPOINT

Pattern recognition is one of the functional properties of RRNNs. In our terms, a pattern is "learned" when its presentation (or removal) induces a bifurcation [5]. Moreover, this effect must be *acquired* via learning, selective (i.e. only the presented pattern is learned) and robust (i.e. a noisy version of the learned pattern should lead to an attractor similar to the one reached after presentation of the learned pattern). We now proceed to an analysis of the effect of pattern removal, as a simple indicator of the functional properties of the network. A deeper investigation of the functional properties of the network is out of the scope of the present study and will be the subject of future works.

Label by $\mathbf{x}$ (resp. $\mathbf{u}$) the neuron firing rate (resp. local field) obtained when the (time constant) input pattern $\boldsymbol{\xi}$ is applied to the network (see eq. 1) and by $\mathbf{x}'$ (resp. $\mathbf{u}'$) the corresponding quantities when $\boldsymbol{\xi}$ is removed ($\boldsymbol{\xi} = 0$). The removal of $\boldsymbol{\xi}$ modifies the attractor structure and the average value of any observable ϕ (though the amplitude of this change depends on ϕ). More precisely call:

$$\Delta^{(T)}[\phi] = \langle \phi(\mathbf{x}') \rangle^{(T)} - \langle \phi(\mathbf{x}) \rangle^{(T)} \tag{24}$$

where $\langle \phi(\mathbf{x}') \rangle^{(T)}$ is the (time) average value of ϕ without $\boldsymbol{\xi}$ and $\langle \phi(\mathbf{x}) \rangle^{(T)}$ the average value in the presence of $\boldsymbol{\xi}$. Two cases can arise.

In the first case, the system is away from a bifurcation point and removal results in a variation of $\Delta^{(T)}[\phi]$ that remains proportional to $\boldsymbol{\xi}$ provided $\boldsymbol{\xi}$ is sufficiently small (remember here that the present network admits a single attractor at a given learning epoch). Albeit common for non-chaotic dynamics, we emphasize that this statement still holds for chaotic dynamics. This has been rigorously proven for uniformly hyperbolic systems, thanks to the linear response theory developed by Ruelle (Ruelle, 1999). In the present context, the linear response theory predicts that the variation of the average value of $\mathbf{u}$ is given by (Cessac and Sepulchre, 2006, 2007):

$$\Delta^{(T)}[\mathbf{u}] = -\chi^{(T)}\boldsymbol{\xi} \tag{25}$$

[5] This idea, as well as the preceding works of the authors on this topic was deeply influenced by Freeman's work (Freeman, 1987; Freeman *et al.*, 1988).

where

$$\chi^{(T)} = \sum_{n=0}^{\infty} \langle D\mathbf{F}^n \rangle^{(T)} \tag{26}$$

is a matrix[6], [7] whose entries can be written:

$$\chi_{ij}^{(T)} = \mathcal{I} + \sum_{n=1}^{+\infty} \sum_{\gamma_{ij}(n)} \prod_{l=1}^{n} W_{k_l k_{l-1}} \left\langle \prod_{l=1}^{n} f'(u_{k_{l-1}}(l-1)) \right\rangle^{(T)} \tag{27}$$

where the sum $\sum_{\gamma_{ij}(n)}$ holds on every possible path $\gamma_{ij}(n)$ of length n, connecting neuron $k_0 = j$ to neuron $k_n = i$, in n steps.

Note therefore that $\Delta^{(T)}[\mathbf{u}] = -\boldsymbol{\xi} - M^{(T)}\boldsymbol{\xi}$ where the matrix $M^{(T)} = \sum_{n=1}^{\infty} \langle D\mathbf{F}^n \rangle^{(T)}$ integrates dynamical effects. A slight variation of u_i at $t = 0$ implies a reorganization of the dynamics which results in a complex formula for the variation of $\langle \mathbf{u} \rangle^{(T)}$, even if the dominant term is $\boldsymbol{\xi}$, as expected. More precisely, as emphasized several times above, one remarks that each path in the sum $\sum_{\gamma_{ij}(n)}$ is weighted by the product of a *topological* contribution depending only on the weights W_{ij} and on a *dynamical* contribution. The weight of a path γ_{ij} depends on the average value of $\left\langle \prod_{l=1}^{n} f'(u_{k_{l-1}}(l-1)) \right\rangle^{(T)}$ thus on *correlations* between the state of saturation of the units $k_0, \ldots, k_{n-1}$ at times $0, \ldots, n-1$.

Eq. 25 shows how the effects of pattern removal are complex when dealing with a chaotic dynamics. However, the situation is much easier mathematically in the simplest case where dynamics have converged to a stable fixed point $\mathbf{u}^{*(T)}$ (resp. $\mathbf{x}^{*(T)}$). In this case, eq. (25) reduces to:

$$\Delta^{(T)}[\mathbf{u}] = -\sum_{n=0}^{\infty} \left(\mathcal{W}^{(T)} \Lambda(\mathbf{u}^*) \right)^n \boldsymbol{\xi} \tag{28}$$

Calling $\lambda_k, \mathbf{v}_k$ the eigenvalues and eigenvectors of $\mathcal{W}^{(T)}\Lambda(\mathbf{u}^{*(T)})$, ordered such that $|\lambda_N| \leq |\lambda_{N-1}| \leq |\lambda_1| < 1$ one obtains:

$$\Delta^{(T)}[\mathbf{u}] = -\sum_{k=1}^{N} \frac{(\mathbf{v}_k, \boldsymbol{\xi})}{1 - \lambda_k} \mathbf{v}_k \tag{29}$$

[6] The convergence of this series is discussed in (Cessac and Sepulchre, 2004, 2006; Ruelle, 1999). Note that a similar formula can be written for an arbitrary observable ϕ, but is more cumbersome.

[7] Incidentally, this equation shows once again why the synaptic weight matrix is not sufficient to capture the dynamical effects of a perturbation. Indeed, it contains a purely topological term ($\prod_{l=1}^{n} W_{k_l k_{l-1}}$) and also depends on a "purely dynamical" term $\left\langle \prod_{l=1}^{n} f'(u_{k_{l-1}}(l-1)) \right\rangle^{(T)}$ that involves an average of the derivative of the transfer functions along the orbit of the neural network.

where (,) denotes the usual scalar product. Actually, this result can easily be found without using linear response, by a simple Taylor expansion (see appendix E). The response is then proportional to $\boldsymbol{\xi}$ but becomes arbitrary large when λ_1 tends to 1 and provided that $(\mathbf{v}_1, \boldsymbol{\xi}) > 0$. This analysis can be formally extended to the general case (i.e. including chaos, eq. 26) but is delicate enough to deserve a treatment by its own and will be the scope of a forthcoming paper[8]. Here, we simply want to make the following argument. From the analysis above, we expect pattern removal to have a maximal effect at "the edge of chaos", namely when the (average) value of the spectral radius[9] of $D\mathbf{F}_{\mathbf{X}}$ is close to 1. As mentioned above, the effects are however more or less prominent according to the choice of the observable ϕ. We empirically found that the effects were particularly prominent with the following quantity:

$$\Delta^{(T)}[\Lambda] = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(\langle\Lambda_{ii}(\mathbf{u})\rangle^{(T)} - \langle\Lambda_{ii}(\mathbf{u}')\rangle^{(T)}\right)^2} \tag{30}$$

Indeed, $\Lambda_{ii} = f'(u_i)$ is maximal when the local field of i falls in the central pseudo-linear part of the transfer function, hence where neuron i is the most sensitive to its input. Hence $\Delta^{(T)}[\Lambda]$ measures how neuron excitability is modified when the pattern is removed. The evolution of $\Delta^{(T)}[\Lambda]$ during learning following rule eq.(5) is shown on fig. 4 (full lines) for two values of the passive forgetting rate λ. $\Delta^{(T)}[\Lambda]$ is found to increase to a maximum at early learning epochs, while it vanishes afterwards. Interestingly, comparison with the decay of the leading eigenvalue of the Jacobian matrix, μ_1 (dotted lines) shows that the maximal values of $\Delta^{(T)}[\Lambda]$ are obtained when $|\mu_1| = |\lambda_1|$ is close to 1. Hence, these numerical simulations confirm that sensitivity to pattern removal is maximal when the leading eigenvalue is close to 1. Therefore, *"Hebb-like" learning drives the global dynamics through*

[8] This can be achieved by formally "diagonalizing" the matrices $\langle D\mathbf{F}^n\rangle^{(T)}$ but the problem is that eigenvalues $\lambda_k(n)$ and eigenvectors $\mathbf{v}_k(n)$ now depend on the time n. Information about the time dependence of the spectrum can be found using the Fourier transform of the matrix χ and looking for its poles (Cessac and Sepulchre, 2006). These poles are closely related to the graph structure induced by the Jacobian matrices, by standard traces formula and cycle expansions (Gaspard, 1998). Essentially, we expect that, under the effect of learning, the leading resonances move toward the real axis leading to a singularity at the edge of chaos. The motion should be closely related to the reinforcement of feedback loops discussed in appendix F.

[9] There is a subtlety here. We have $D\mathbf{F}_{\mathbf{X}} = \Lambda(\mathbf{u})\mathcal{W}$, while in formula (29) we consider the eigenvalues of $\mathcal{W}\Lambda(\mathbf{u})$. However, if $\lambda_k, \mathbf{v}_k$ are eigenvalues and eigenvectors of $\mathcal{W}\Lambda(u)$ then $\Lambda(u)\mathcal{W}\Lambda(u)\mathbf{v}_k = D\mathbf{F}_{\mathbf{X}}\Lambda(u)\mathbf{v}_k = \lambda_k\Lambda(u)\mathbf{v}_k$. Therefore, $\lambda_k, \Lambda(u)\mathbf{v}_k$ are eigenvalues and eigenvectors of $D\mathbf{F}_{\mathbf{X}}$.

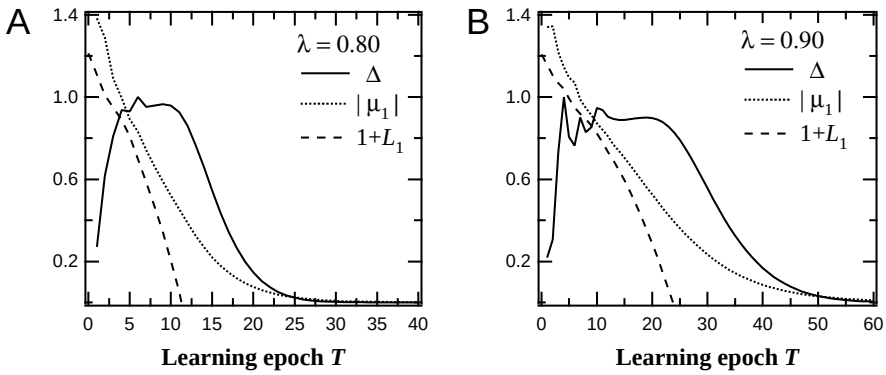

FIG. 4 The network sensitivity to the input pattern is maximal close to a bifurcation. The evolution of the average value for the spectral radius of $D\mathbf{F}_{\mathbf{x}}^{(T)}$ during learning (dotted line) is plotted together with the sensitivity measure $\Delta^{(T)}[\Lambda]$ (full line) for $\lambda = 0.80$ (A) or 0.90 (B). The panels also display the corresponding evolution of the largest Lyapunov exponent L_1, plotted as $1.0 + L_1$ for obvious comparison purpose (dashed line). The values of $\Delta^{(T)}[\Lambda]$ are normalized to the $[0-1]$ range for comparison purposes. Each value is an average over 50 realizations (standard deviations are omitted for clarity). All other parameters were as in fig. 1

a bifurcation, in the neighborhood of which sensitivity to the input pattern is maximal. This property may be crucial regarding memory properties of RRNNs, which must be able to detect, through their collective response, whether a learned pattern is present or absent. This property is obtained at the frontier where the strange attractor begins to destabilize ($|\mu_1| = 1$), hence at the so-called "edge of chaos".

We showed in section III.A that the Hebbian learning rules studied here contract the spectral radius of $D\mathbf{F}_{\mathbf{X}}, \forall \mathbf{x}$, so that the latter crosses the value 1 at some learning epoch. Thus, 1 is ensured to be an eigenvalue of $D\mathbf{F}_{\mathbf{X}}$ at some point . The evolution of v_1, the eigenvector associated to the leading eigenvalue of the Jacobian matrix μ_1, is less obvious. We plot on fig. 3 (dotted lines) the evolution of its real part during numerical simulations (actually, its imaginary part vanishes after just a couple of learning epochs). It is clear from numerical simulations that the possibility of a vanishing projection of the input pattern $\boldsymbol{\xi}$ (thick dashed line) on v_1 can be ruled out (the two vectors are not orthogonal). The tendency is even opposite, i.e. v_1 is found to align on the input pattern at long learning epochs ($T \gtrsim 100$; note that we were not able to find a satisfactory explanation for this alignment).

V. DISCUSSION

The coupled dynamical system studied in the present paper (eqs.(1) and (2)) is based on several simplifying assumptions that allowed the rigorous mathematical study we have presented. However, many of the results we obtain remain valid when some of these assumptions are relaxed to improve biological realism. Here, we give a brief overview of the related arguments. As already stated in the introduction, we do not pretend to encompass the spectrum of complexity and richness of biological learning and plasticity rules (Kim and Linden, 2007). However, the present study focuses on the major type of synaptic plasticity (i.e Hebbian plasticity), which is generally considered as the principal cellular basis of behavioral learning and memory.

The learning rule we study here eq.(2) includes a term that allows passive forgetting ($\lambda < 1$). This possibility is supported by a body of experimental data that shows that synaptic weights decay exponentially toward their baseline after LTP, in the absence of subsequent homo- or hetero-synaptic LTD, with time constants from seconds to days (Abraham *et al.*, 1994, 2002; Brager *et al.*, 2003). A plausible molecular mechanism for this passive behavior has been recently proposed, which relies on the operation of kinase and phosphatase cycles that are systematically implicated in learning and memory (Delord *et al.*, 2007). Our theoretical results predict that learning-induced reduction of dynamics complexity can still arise in the limit case of $\lambda = 1$. Indeed, numerical simulations of Hebbian learning rules devoid of passive forgetting (i.e. with $\lambda = 1$) have clearly evidenced a reduction of the attractor complexity during learning (Berry and Quoy, 2006; Siri *et al.*, 2006). In this case, the reduction of the attractor complexity is provoked by an increase of the average saturation level of the neurons, in agreement with our present analytical results. As a matter of fact, the question is not so much to know what exactly is the value of λ in real neural networks, but how the characteristic time scale $\frac{1}{|\log(\lambda)|}$ compares to other time scales in the system.

Another assumption of the generic Hebbian rule we study is that $\Gamma_{ij} = 0$ whenever the presynaptic neuron is silent. As already mentioned section II.A, an interpretation of this assumption is that this learning rule excludes heterosynaptic LTD. To assess the impact of this form of synaptic depression in the model, we ran numerical simulations using a variant of eq.(5) in which the Heavyside term (that forbids heterosynaptic LTD) was omitted. The results of these simulations (not shown) were in agreement with all the analytical results

supported here, including those on spectral radius contraction. In agreement with these numerical simulations, our analytical results on the contraction of the spectral radius are expected to remain valid when heterosynaptic LTD is accounted for, but this would require extending the model definition and further mathematical developments that are out of the scope of the present study.

The effects of Hebbian learning were studied here in a completely connected, one population (i.e. each neuron can project both excitatory and inhibitory synapses) chaotic network. While this hypothesis allows a rigorous mathematical treatment, it is clearly a strong idealization of biological neural networks. However, we have tested the analytical predictions obtained here with numerical simulations of a chaotic recurrent neural network with connectivity mimicking cortical micro-circuitry, i.e. sparse connectivity and distinct excitatory and inhibitory neuron populations. These simulations unambiguously demonstrated that our analytical results are still valid in these more realistic conditions (Siri *et al.*, 2007).

From a functional point of view, we have shown that the sensitivity to the learned pattern is maximal at the edge of chaos. Starting from chaotic dynamics, this regime is reached at intermediate learning epochs. However, longer learning times result in poorer dynamical regimes (e.g. fixed points) and the loss of sensitivity to the learned pattern. Additional plasticity mechanisms like synaptic scaling (Turrigiano *et al.*, 1998) or intrinsic plasticity ref (Daoudal and Debanne, 2003) may constitute interesting biological processes to maintain the network in the vicinity of the edge of chaos and preserve a state of high sensitivity to the learned pattern. Such possibilities are currently under investigation in our group.

Acknowledgments

This work was supported by a grant of the French National Research Agency, project JC05_63935 "ASTICO".

APPENDIX A: Definitions.

Dealing with chaotic systems, one is faced with the necessity to defining indicators measuring dynamical complexity. There are basically two families of indicators: one is based on topological properties (e.g. topological entropy), the other is based on statistical properties

(e.g. Lyapunov exponents or Kolmogorov-Sinai entropy). The latter family can easily be accessed numerically or experimentally by *time averages* of relevant observables along typical trajectories of the dynamical system. However, to this aim, one has to assume a strong ergodic property: the time average of observables, along trajectories corresponding to initial conditions drawn at random with respect to a probability distribution having a density (with respect to the Lebesgue measure), is constant (it does not depend on the initial condition). This property is far from being evident. Actually, we are not able to prove it in the present context. On mathematical grounds, it corresponds to the following assumption.

Assumption 1 *Call μ_L is the Lebesgue measure on $[0,1]^N$ and let $\boldsymbol{F}^{*t}\mu_L$ the image of μ_L under $\boldsymbol{F}^t$. We assume that the following limit exists:*

$$\rho^{(T)} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \boldsymbol{F}^{*t}\mu_L \tag{A1}$$

where the probability measure $\rho^{(T)}$ is called "the Sinai-Ruelle-Bowen (SRB) measure at learning epoch T" (Bowen, 1975; Ruelle, 1978; Sinai, 1972). Under this assumption the following holds. Let $\phi : [0,1]^N \to \mathbb{R}^N$ be some suitable (measurable) function. Then the time average:

$$\bar{\phi}[\boldsymbol{x}^{(T)}(0)] \stackrel{\text{def}}{=} \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \phi(\boldsymbol{x}^{(T)}(t)), \tag{A2}$$

where $\boldsymbol{x}(t) = \boldsymbol{F}^t(\boldsymbol{x})$, is equal to the ensemble average:

$$\langle\phi\rangle^{(T)} \stackrel{\text{def}}{=} \int_{[0,1]^N} \phi(\boldsymbol{x})\rho^{(T)}(d\boldsymbol{x}), \tag{A3}$$

for Lebesgue-almost every initial condition $\boldsymbol{x}^{(T)}(0)$.

In other words, time average and ensemble average are identical on practical grounds. The use of $\rho^{(T)}$ is required to prove the mathematical results below while time average is what we use for numerical simulations.

Note that in doing so, we have constructed a family of probability distributions $\rho^{(T)}$ that depends on the *learning epoch* T. $\rho^{(T)}$ provides statistical information about the attractor structure. A prominent example is the maximal Lyapunov exponent. Let $\mathbf{x} \in [0,1]^N$, $\mathbf{v} \in \mathbb{R}^N$ and ρ be an SRB measure. Then, the largest Lyapunov exponent is given by:

$$L_1^{(T)} = \lim_{t\to\infty} \lim_{\|\mathbf{v}\|\to 0} \frac{1}{t} \log\left(\frac{\|D\mathbf{F}_{\mathbf{X}}^t \mathbf{v}\|}{\|\mathbf{v}\|}\right) \tag{A4}$$

Its value is constant for $\rho^{(T)}$ almost every $\mathbf{x}$. (Note indeed that the LHS does not depend on $\mathbf{x}$, while the RHS does. This is a direct consequence of the assumption that $\rho^{(T)}$ is an SRB measure).

APPENDIX B: Asymptotic behaviors

In the specific learning rule eq.(5) used in our numerical simulations, $\Gamma_{ij} = m_i m_j H(m_j)$. Thus

$$\begin{aligned}
\|\Gamma\| &= \sup_x \frac{\|\Gamma\mathbf{x}\|}{\|\mathbf{x}\|} && \text{(B1)}\\
&= \sup_x \frac{\|\mathbf{m}\left[\mathbf{m}H(\mathbf{m})\right]^+ \mathbf{x}\|}{\|\mathbf{x}\|} && \text{(B2)}\\
&\leq \|\mathbf{m}\|\| \left[\mathbf{m}H(\mathbf{m})\right]^+ \| && \text{(B3)}\\
&\leq \left(\sum_{i=1}^N m_i^2\right)^{1/2} \left(\sum_{j=1,m_j>0}^N m_j^2\right)^{1/2} && \text{(B4)}\\
&\leq \sqrt{N}\sqrt{N}\phi^{1/2} && \text{(B5)}\\
&\leq N\sqrt{\phi} && \text{(B6)}
\end{aligned}$$

where $[\mathbf{v}]^+$ denotes the transpose of vector $\mathbf{v}$, $\sum_{j=1,m_j>0}$ denotes a sum restricted to the active neurons and ϕ is the fraction of active neurons. Hence

$$\|\Gamma^{(T)}\| \leq N\sqrt{\phi^{(T)}} \tag{B7}$$

If (as observed in our numerical simulations) $\phi^{(T)}$ tends to a stationary value $\phi^{(\infty)}$ then

$$\|\Gamma^{(T)}\| \leq N\sqrt{\phi^{(\infty)}} \tag{B8}$$

Hence Γ is bounded in the specific case of eq.(5) by a constant $N\sqrt{\phi^{(\infty)}}$.

More generally, $\|\Gamma\|$ is bounded provided that the function h in (3) is bounded as well.

APPENDIX C: Proof of theorem 1

Let $\mathbf{v}, \mathbf{x} \in \mathbf{R}^N$. Denote by $\mathbf{x}(t) = \mathbf{F}^t(\mathbf{x})$, and $\mathbf{v}(t) = D\mathbf{F}_{\mathbf{x}(t)}.D\mathbf{F}_{\mathbf{X}}^{t-1}.\mathbf{v}$, $\mathbf{v}(0) = \mathbf{v}$. From the chain rule:

$$\frac{\|D\mathbf{F}^t_{\mathbf{X}}\mathbf{v}\|}{\|\mathbf{v}\|} = \frac{\|D\mathbf{F}_{\mathbf{X}(t)}\mathbf{v}(t-1)\|}{\|\mathbf{v}(t-1)\|}\frac{\|\mathbf{v}(t-1)\|}{\|\mathbf{v}\|}$$
$$= \frac{\|D\mathbf{F}_{\mathbf{X}(t)}\mathbf{v}(t-1)\|}{\|\mathbf{v}(t-1)\|}\frac{\|D\mathbf{F}_{\mathbf{X}(t-1)}\mathbf{v}(t-2)\|}{\|\mathbf{v}(t-2)\|}\cdots\frac{\|D\mathbf{F}_{\mathbf{X}(1)}\mathbf{v}\|}{\|\mathbf{v}\|}$$

Therefore:

$$L_1^{(T)} = \lim_{t\to\infty}\lim_{\|\mathbf{v}\|\to 0}\frac{1}{t}\sum_{n=1}^{t}\log\left(\frac{\|D\mathbf{F}_{\mathbf{X}(n)}\mathbf{v}(n-1)\|}{\|\mathbf{v}(n-1)\|}\right).$$

Since $\|A\mathbf{v}\| \leq \|A\|\|\mathbf{v}\|$:

$$L_1^{(T)} \leq \lim_{t\to\infty}\frac{1}{t}\sum_{n=1}^{t}\log\left(\|D\mathbf{F}_{\mathbf{X}(n)}\|\right) = \langle\log\left(\|D\mathbf{F}_{\mathbf{X}}\|\right)\rangle^{(T)} \ \ \rho^{(T)} - \text{almost surely.}$$

But since $D\mathbf{F}_{\mathbf{X}} = \Lambda(\mathbf{u})\mathcal{W}$, we have $\|D\mathbf{F}_{\mathbf{X}}\| \leq \|\mathcal{W}\|\|\Lambda(\mathbf{u})\| \leq \|\mathcal{W}\|\max_i(f'(u_i)$.

APPENDIX D: Local fields

Fix $\mathbf{x}$ and the time epoch T. Set $\mathbf{u} = \mathcal{W}^{(T)}\mathbf{x} + \boldsymbol{\xi}$. The average of $\mathbf{u}$, $\langle\mathbf{u}\rangle^{(T)}$ is defined either by the time average (A2) or by the ensemble average (A3). However, since $\mathcal{W}^{(T)}$ is constant during a given learning epoch one has:

$$\langle\mathbf{u}\rangle^{(T)} = \mathcal{W}^{(T)}\langle\mathbf{x}\rangle^{(T)} + \boldsymbol{\xi}, \quad \forall T. \tag{D1}$$

Therefore:

$$\langle\mathbf{u}\rangle^{(T+1)} = \mathcal{W}^{(T+1)}\langle\mathbf{x}\rangle^{(T+1)} + \boldsymbol{\xi} = (\lambda\mathcal{W}^{(T)} + \frac{\alpha}{N}\Gamma^{(T)})(\langle\mathbf{x}\rangle^{(T)} + \delta\rho^{(T+1)}(\mathbf{x})) + \boldsymbol{\xi},$$

where $\delta\rho^{(T+1)}(\mathbf{x}) \stackrel{\text{def}}{=} \langle\mathbf{x}\rangle^{(T+1)} - \langle\mathbf{x}\rangle^{(T)}$ is the difference of the average value of $\mathbf{x}$ between learning epochs $T+1$ and T.

Thus:

$$\langle\mathbf{u}\rangle^{(T+1)} = \lambda\langle\mathbf{u}\rangle^{(T)} + (1-\lambda)\boldsymbol{\xi} + \lambda\mathcal{W}^{(T)}\delta\rho^{(T+1)}(\mathbf{x}) + \frac{\alpha}{N}\Gamma^{(T)}\langle\mathbf{x}\rangle^{(T+1)},$$

and by recurrence:

$$\langle\mathbf{u}\rangle^{(T+1)} = \lambda^T\langle\mathbf{u}\rangle^{(1)} + (1-\lambda^T)\boldsymbol{\xi} + \lambda\sum_{n=1}^{T}\lambda^{T-n}\mathcal{W}^{(n)}\delta\rho^{(n+1)}(\mathbf{x}) + \frac{\alpha}{N}\sum_{n=1}^{T}\lambda^{T-n}\Gamma^{(n)}\langle\mathbf{x}\rangle^{(n+1)} \tag{D2}$$

APPENDIX E: Proof of eq.(29)

Call $\mathbf{u}^{*(T)}$ ($\mathbf{u}'^{*(T)}$) the fixed point (for the variable $\mathbf{u}$) with (without) $\boldsymbol{\xi}$. We have:

$$\mathbf{u}'^{*(T)} = \mathcal{W}\mathbf{F}(\mathbf{u}'^{*(T)})$$

and:

$$\mathbf{u}^{*(T)} = \mathcal{W}\mathbf{F}(\mathbf{u}^{*(T)}) + \boldsymbol{\xi}$$

Therefore:

$$\mathbf{u}'^{*(T)} - \mathbf{u}^{*(T)} = \delta\mathbf{u}^{(T)} = \mathcal{W}\left[\mathbf{F}(\mathbf{u}^{*(T)} + \delta\mathbf{u}^{(T)}) - \mathbf{F}(\mathbf{u}^{*(T)})\right] - \boldsymbol{\xi}.$$

A series expansion yields, to the linear order:

$$(\mathcal{I} - \mathcal{W}\Lambda(\mathbf{u}^{(T)}))\delta\mathbf{u}^{(T)} = -\boldsymbol{\xi}$$

Decomposing on the eigenbasis $\mathbf{v}_k$ of $\mathcal{W}\Lambda(\mathbf{u}^{(T)})$ we obtain:

$$(1 - \lambda_k)(\delta\mathbf{u}^{(T)}, \mathbf{v}_k) = -(\boldsymbol{\xi}, \mathbf{v}_k) \tag{E1}$$

which corresponds to eq. (29) *provided* $|\lambda_k| < 1$ (ensuring that the matrix $\mathcal{I} - \mathcal{W}\Lambda(\mathbf{u}^{(T)})$ is invertible).

APPENDIX F: Jacobian matrix and feedback loops background

Assume that we slightly perturb at time t the state of neuron j with a small perturbation (e.g. $x_j(t) \to x_j(t) + \delta_j(t)$). Then the effect of this change on neuron i, at time $t+1$ is given by $x_i(t+1) = f\left(\sum_{k=1}^{N} W_{ik}x_k(t) + \xi_i + W_{ij}\delta_j(t)\right)$. One can perform a Taylor expansion of this expression in powers of $W_{ij}\delta_j(t)$. To the linear order the effect is given by $f'(u_i(t))W_{ij}\delta_j(t)$. To each Jacobian matrix $D\mathbf{F}_{\mathbf{X}}$ one can associate a graph, called "the graph of linear influences". such that there is an oriented edge $j \to i$ iff $\frac{\partial f(u_i)}{\partial x_j} \neq 0$. The edge is positive if $\frac{\partial f(u_i)}{\partial x_j} > 0$ and negative if $\frac{\partial f(u_i)}{\partial x_j} < 0$. An important remark is that this graph depends on the current state $\mathbf{x}$, contrarily to the weights matrix which is a constant inside a given learning epoch. This has important consequences. Indeed, in our case since

$\frac{\partial F_i}{\partial x_j} = f'(u_i)W_{ij}$, the edge $j \to i$ in the graph of linear influences can be very small even if the synaptic weight W_{ij} is large. It suffices that $|u_i|$ be large. This effect, due to the saturation of the transfer function f, is prominent in the subsequent studies.

We have now the following situation: "above" (in the tangent bundle) each point $\mathbf{x}$, there is graph. This graph contains *circuits or feedback loops*. If e is an edge, denote by $o(e)$ the origin of the edge and $t(e)$ its end. Then a circuit is a sequence of edges $e_1, ..., e_k$ such that $o(e_{i+1}) = t(e_i)$, $\forall i = 1...k-1$, and $t(e_k) = o(e_1)$. Such a circuit is positive (negative) if the product of its edges is positive (negative). A positive circuit basically yields (to the linear order) a positive feedback that induces an increase of the activity of the neurons in this circuit. Obviously, there is no exponential increase since rapidly nonlinear terms will saturate this effect. It is thus expected that positive loops enhance stability.

A particularly prominent example of this is well known in the framework of continuous time neural networks models and also in genetic networks. It is provided by so-called "cooperative systems". A dynamical system is called cooperative if $\frac{\partial f(u_i)}{\partial x_j}(\mathbf{x}) \geq 0, \forall i \neq j$. Therefore, in this case, all edges are positive edges[10], whatever the state of the neural network and all circuits are positive. Cooperative systems preserve the following partial order $\mathbf{x} \leq \mathbf{y} \Leftrightarrow x_i \leq y_i,\ i = 1 \dots N$. Thus $\mathbf{x}(\mathbf{0}) \leq \mathbf{y}(\mathbf{0}) \Rightarrow \mathbf{x}(\mathbf{t}) \leq \mathbf{y}(\mathbf{t}),\ \forall t > 0$ (this corresponds to the positive feedback discussed above). From these inequalities, Hirsch (Hirsch, 1989) proved that for a two dimensional cooperative dynamical system, any bounded trajectory converges to a fixed point. In larger dimension, one needs moreover a technical condition on the Jacobian matrix: it must be irreducible. Then Hirsch proved that the ω-limit set of almost every bounded trajectory is made of fixed points. Note that this result holds when f is nonlinear.

On the opposite, negative loops usually generate oscillations. For example, the second Thomas conjecture (Thomas, 1981), proved by Gouzé (Gouzé, 1998) under the hypothesis that the sign of the Jacobian matrix elements do not depend on the state, states that "A negative loop is a necessary condition for a stable periodic behavior". In our model, negative loop generate oscillations provided that the nonlinearity g is sufficiently large. This can be easily figured out by considering a system with 2 neurons. A necessary condition to

[10] More generally, there is a variable change which maps the initial dynamical system to a cooperative system with positive edges.

have a Hopf bifurcation giving rise to oscillations is $W_{12}W_{21} < 0$, but the bifurcation occurs only when g is large enough.

References

Abraham, B., W.C. Christe, B. Logan, P. Lawlor, , and M. Dragunow, 1994, Proc. Natl. Acad. Sci. USA **91**, 10049.

Abraham, W., B. Logan, J. Greenwood, and M. Dragunow, 2002, J. Neurosci. **22**, 9626.

Abraham, W. C., and M. F. Bear, 1996, Trends Neurosci. **19**, 126.

Atay., F., T. Biyikoglu, and J. Jost, 2006, Physica. D .

Atay, F., T. Biyiko**u**glu, and J. Jost, 2006, Physica D **224**, 35.

Barahona, M., and L. Pecora, 2002, Phys. Rev. Lett. **89**, 054101.

Bear, M., and W. Abraham, 1996, Annu. Rev. Neurosci. **19**, 437.

Berry, H., and M. Quoy, 2006, Adaptive Behavior **14**, 129.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, 2006, Physics Reports **424**, 175.

Bowen, R., 1975, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms* (Berlin: Springer-Verlag), volume 470.

Brager, D., X. Cai, and S. Thompson, 2003, Nature Neurosci. **6**, 551.

Cessac, B., 1994, J. of Physics A **27**, 927.

Cessac, B., 1995, J. de Physique **5**, 409.

Cessac, B., and M. Samuelides, 2007, EPJ Special topics: Topics in Dynamical Neural Networks **142**(1), 7.

Cessac, B., and J. Sepulchre, 2004, Phys. Rev. E **70**, 056111.

Cessac, B., and J. Sepulchre, 2006, Chaos **16**, 013104.

Cessac, B., and J. Sepulchre, 2007, Physica D **225**(1), 13.

Chung, F. R. K., 1997, *Spectral Graph Theory* (CBMS Regional Conference Series in Mathematics).

Daoudal, G., and D. Debanne, 2003, Learn. Mem. **10**, 456.

Dauce, E., M. Quoy, B. Cessac, B. Doyon, and M. Samuelides, 1998, Neural Networks **11**, 521.

Daucé, E., M. Quoy, and B. Doyon, 2002, Biol. Cybern. **87**, 185.

Delord, B., H. Berry, E. Guigon, and S. Genet, 2007, PLoS Computational Biology **3**(6), e124.

Doyon, B., B. Cessac, M. Quoy, and M. Samuelides, 1993, Int. Journ. of Bif. and Chaos **3**(2), 279.

Freeman, W., 1987, Biol. Cyber. **56**, 139.

Freeman, W., Y. Yao, and B. Burke, 1988, Neur. Networks **1**, 277.

Gaspard, P., 1998, *Chaos, scattering and statistical mechanics* (Cambridge University Press).

Girko, V., 1984, Theor. Prob. Appl **29**, 694.

Gouzé, J., 1998, Journ. Biol. Syst. **6**(1), 11.

Grinstein, G., and R. Linsker, 2005, PNAS **28**(102), 9948.

Hasegawa, H., 2005, Phys. Rev. E. **72**, 056139.

Hebb, D., 1948, *The Organization of Behaviour* (John Wiley & Sons, New-York).

Hirsch, M., 1989, Neur. Networks **2**, 331.

Hong, H., B. Kim, M. Choi, and H. Park, 2002, Phys. Rev. E **65**, 067105.

Hoppensteadt, F., and E. Izhikevich, 1997, *Weakly Connected Neural Networks* (Springer Verlag).

Jaeger, H., and H. Haas, 2004, Science , 78.

Kim, S., and D. Linden, 2007, Neuron **56**, 582.

Lago-Fernández, L. F., R. Huerta, F. Corbacho, and J. A. Sigüenza, 200, Phys. Rev. Lett. **84**, 2758.

Langton, C., 1990, Physica D. **42**.

Nishikawa, T., A. E. Motter, Y. C. Lai, and F. C. Hoppensteadt, 2003, Phys. Rev. Lett. **91**.

Ruelle, D., 1978, *Thermodynamic formalism* (Reading, Massachusetts: Addison-Wesley).

Ruelle, D., 1999, Journ. Stat. Phys. **95**, 393.

Sinai, Y. G., 1972, Lect. Notes.in Math. **27**(4), 21.

Siri, B., H. Berry, B. Cessac, B. Delord, and M. Quoy, 2006, in *International Conference on Complex Systems* (Boston).

Siri, B., M. Quoy, B. Cessac, B. Delord, and H. Berry, 2007, Journal of Physiology (Paris) **101**(1–3), 138.

Thomas, R., 1981, *On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations* (Springer-Verlag in Synergetics), chapter Numerical methods in the study of critical phenomena, pp. 180–193.

Tsuda, I., 2001, Behav. Brain Sc. **24**, 793.

Turrigiano, G., K. Leslie, N. Desai, L. Rutherford, and S. Nelson, 1998, Nature **391**, 892.

Volchenkov, D., and P. Blanchard, 2007, arXiv:0710.3566v1.