Topological and dynamical structures induced by Hebbian learning in random neural networks

Benoit Siri
Alchemy, INRIA, Orsay, France
Hugues Berry
Alchemy, INRIA, Orsay, France, hugues.berry@inria.fr
Bruno Cessac
Institut Non Lineaire de Nice, CNRS UMR6618 & UNSA, Nice, France
Bruno Delord
ANIM, Inserm U742 & UPMC, Paris, France
Mathias Quoy
ETIS, CNRS UMR8051 & UCP-ENSEA, Cergy-Pontoise, France

We study the consequences of the coupling between neuron dynamics and synaptic weight changes on the dynamics and architecture of random recurrent neural networks (RRNNs) using a simple implementation of Hebb's learning rule. Due to the this coupling, learning shapes the network dynamics, topology and function. Indeed we evidence that the modifications of the dynamics can be related to changes in the local loop content. We further show that, because of these local structural alterations, the global network topology changes as well. Under the influence of learning, the network becomes highly clustered while its mean-shortest path remains low, so that this learning rule organizes the network as a small-world one. Hence, these findings raise the hypothesis that small-worldness in natural neural networks may be a spontaneous consequence of the learning scheme governing the links. Moreover, we show that pattern recognition task emerges from this mutual coupling, thus questioning the relevance of small-world architectures for storing and processing information.

1.1 Introduction

In recent years, a vast amount of work concerning dynamical systems interacting on complex networks has focused on the influence of network topology on global dynamics[**?**, **?**, **?**]. In this framework, neural networks are particularly interesting because the evolutions of the neuron dynamics and the weights of the synaptic links that bind them, are interrelated. This dynamic-structure shaping mechanism remains largely obscure.
Here, we study the consequences of such a coupling on dynamics and architecture. To this end, we investigate the influence of learning on the topology of random recurrent neural networks, which exhibit learning and dynamical behaviors yielding associative memory properties that mimic those observed in the olfactory bulb. Indeed, experimental works (especially by Freeman) have shown that spontaneous neuron dynamics in the olfactory bulb is chaotic. Though, when a known odor is recognized, the dynamics reduces to a simpler attractor (a limit cycle)[**?**, **?**]. This behavior can be mimicked by random recurrent neural networks (RRNNs) using a classical Hebbian learning rule[**?**].
The spontaneous behavior (i.e. in the absence of learning) of RRNNs has been thoroughly

studied. In particular, methods from statistical physics (mean-field approaches) and dynamical systems theory have shown that, when the neuron nonlinearity or the variance of the inputs increase, the system undergoes a quasi-periodicity route that leads from stable fixed-point regime to stable limit cycles, quasiperiodicity and eventually chaos[**?**]. Simulations have shown that, starting from a chaotic spontaneous behavior, learning usually induces the reverse transition, i.e. a reduction of the dynamics complexity by a reverse quasiperiodicity route (dynamics reduction phase)[**?**]. However, because mean-field arguments are ineffective here (learning introduces correlations between the synaptic strengths which forbids the use of a mean-field approach), our understanding of the impact of learning on RRNNs dynamics and structure remains limited. Recently, we have reported that learning in RRNNs comes together with an increased organization of the strong synaptic weights[**?**]. However these effects were only sensible much after the reduction phase so that their relations with the dynamics were not obvious.

In the present paper, we introduce new analysis approaches that allow us to tackle the complex couplings between dynamics, global structure and function during learning. We first present the model and the learning rule we use (§1.2). The second part presents our results. Section §1.3.1 shows how the dynamics reduction induced by learning can be interpreted as variations in the local loop content. These variations are responsible for the increased organization of the global network structure, that we describe section §1.3.2. We then illustrate in section §1.3.3 how pattern recognition emerges from the coupling between dynamics and structure. Finally, we give some conclusions and directions for future work in §1.4.

1.2 The Model

1.2.1 Dynamics

We consider a set of N fully connected neurons (here we use $N = 500$). Each neuron $i = 1 \dots N$ is associated with a continuous state $x_i \in [0, 1]$ that reflects the firing rate of the neuron. State dynamics is given by

$$x_i(t+1) = f\left(\sum_{j=1}^{N} w_{ij}(t)x_j(t) + I_i\right) \ \forall i = 1 \dots N \tag{1.1}$$

where $f(x)$ is a sigmoidal function with slope g at $x = 0$ (we use $f(x) = (1 + \tanh(gx))/2$ with $g = 10$), w_{ij} is the strength of the $j \to i$ synapse and I_i is the (time independent) input applied to neuron i. In random recurrent neural networks (RRNNs), the input to the network consists of the N dimensional input vector $\mathbf{I} = \{I_i\}$. Furthermore, the synaptic strengths w_{ij} are asymmetric ($w_{ij} \neq w_{ji}$) and may be positive, negative or null (excitatory, inhibitory or ineffective synapses, respectively). Their initial values $w_{ij}(0)$ are randomly drawn according to a Gaussian distribution with zero mean and variance J^2/N (for which we use $J = 1$). In this paper, we use a constant input $\mathbf{I}$ (see §1.2.2 below) and, unless otherwise stated, average the results over 50 realizations of the initial weights $w_{ij}(0)$.

Learning in neural networks means that synaptic strengths evolve over time according to a given learning rule. Here, we consider a learning rule inspired by the postulate of Hebb [**?**] that has been corroborated by neurophysiological observations [**?**], and which is usually translated into the statement that the connection between two neurones increases when they are co-active. In the present paper, we use a straightforward formulation of this postulate :

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \cdot x_i(t+1)x_j(t) \tag{1.2}$$

with $\alpha = 10^{-2}$. Furthermore, we add the constraint that the synaptic strengths w_{ij} cannot change their sign. Note that this rule implies that the synaptic strengths cannot decrease and may grow unboundedly.

1.2.2 Pattern presentation and learning

Because we ultimately wish to compare the input pattern to the network configuration, we feed the network with a structured pattern. The input pattern to be learnt consists in a $(22 \times 22 + 16)$ binary cross (see Fig. 1.4A) that is straightforwardly mapped to each of the N inputs (i.e. the input of neuron number 0, I_0, is defined by the first image pixel; I_1 by the second pixel and so forth). For black pixels $I_i = 0.5$, while $I_i = 0$ for white ones.
One learning epoch consists in the two following steps : 1) The network dynamics evolves spontaneously (Eq.(1.1)) without synaptic modifications, towards the network attractor for 20 time steps, 2) the synaptic strengths are modified according to the learning rule (Eq.(1.2)).

1.3 Results

1.3.1 Reduction of the Dynamics

We first begin with an analysis of the synaptic strength evolution during learning. To understand the effects of these modifications on the network dynamics, we study the evolution of local loops induced by learning. In the context of RRNNs, a local n-loop (or circuit) is a sequence of n nonzero synapse strengths, linking n connected neurons $(1, 2 \dots n)$, given by $\{w_{21}, w_{32}, w_{43}, \dots, w_{n(n-1)}, w_{1n}\}$. For instance, a network of three fully connected neurons i, j and k contains two 3-loops: $\{w_{ji}, w_{kj}, w_{ik}\}$ and $\{w_{ki}, w_{jk}, w_{ij}\}$ (and three trivial 2-loops). A loop is positive (negative) if its product, i.e. the product of its synaptic strengths $(w_{21} \times w_{32} \times \dots \times w_{1n})$, is positive (negative). A great deal of work has been devoted to studies of the influence of the network composition in terms of these local loops, on the network global behavior[**?**, **?**]. We will only retain the general claim that positive n-loops tend to increase stability in the network, while negative ones tend to induce local oscillatory behaviors. For instance, a continuous time version of the dynamics (1.1), considered on a network constituted by only one negative loop, generically undergoes a Hopf bifurcation, giving rise to oscillations, when the synaptic strength J (or the gain g) increases. When several negative loops are competing together and when J (or g) increases, one is first expecting synchronization phases leading to quasiperiodicity and frequency locking, and then chaos, by the Ruelle-Takens scenario[**?**]. This is exactly what is observed in the dynamical system (1.1)[**?**]. On the other hand, positive loops lead to so-called cooperative systems[**?**], which are convergent (they have only fixed points and no more complex attractors). We are thus expecting that a decay in the negative loop content naturally leads to a reduction of the dynamics complexity.
With the parameters used in this article, the network dynamics is chaotic before learning (see Fig. 1.1A, bottom), as attested by the positive value of the maximal Lyapounov exponent ($\lambda_{max}(0) = 0.293 \pm 0.032$) that uncovers sensibility to initial conditions in the system, a hallmark of chaotic behaviors. The learning rule, Eq.(1.2), gives a straightforward example of our analysis. Consider two neurons i and j, that are active at some time t ($x_i(t) > 0$ and $x_j(t) > 0$). According to the rule, the strength of the $i \to j$ synapse will increase at time $t + 1$. Hence inhibitory synapses between active neurons vanishes, while excitatory ones increase. The net result is to increase excitation between active neurons. As a consequence, a decrease (increase) of the average loop product for negative (positive) loops is predicted. Fig. 1.1C shows that the product of positive (negative) 2-loops indeed slightly increases (decreases). This trend is however more obvious when observing 3-loops. Fig. 1.1D shows that the product of the negative 3-loops with all weights negative decrease significatively during learning, while the products of positive 3-loops slightly increase. Note however that the product of negative 3-loops with a single negative weight tend to increase slightly due to the increase of their two positive weights (not shown).
This indicates that Hebbian learning would tend to decrease the oscillatory dynamics at the benefit of enhanced stability. In agreement with this analysis, the dynamics becomes

progressively more regular and less chaotic (Fig. 1.1A, middle), as stated by the decrease and sign change of the maximal Lyapunov exponent (Fig. 1.1B), down to a global stable fixed point at the end of learning (Fig. 1.1A, top). Hence, taken together, the results we obtain Fig. 1.1 suggest that the Hebbian learning rule Eq.(1.2), induces a loss in the product of negative loops that provokes a reduction of the dynamics complexity from oscillatory chaotic to more regular oscillations and ultimately to fixed-point stability regime.

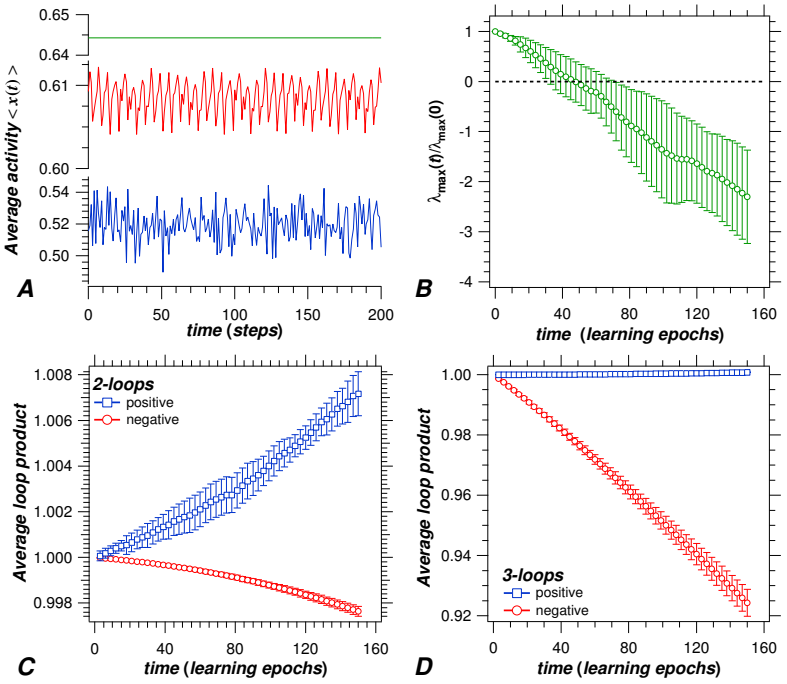

Figure 1.1: Dynamics and local loop content changes during learning with rule Eq.(1.2). (**A**) Dynamics of the RRNN before learning (bottom), after the first learning epoch yielding a negative maximal Lyapounov exponent (that shows the network leaves chaotic behavior) (middle), and after 150 learning epochs (top). Plotted is the average state of the network at the corresponding time $\langle x(t)\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ for a single realization of the initial synaptic strengths. (**B**) Decrease of the maximal Lyapunov exponent λ_{max} during learning. Values are normalized by the exponent before learning, $\lambda_{max}(0)$. The dotted line indicates $\lambda_{max} = 0$. (**C**) Evolution of the average loop product for positive (squares) and negative (circles) 2-loops. (**D**) Evolution of the average loop product for positive (squares) and negative (circles) 3-loops. Note that the values for negative loops correspond to loops with 3 negative weights. In (**B-D**), the variables shown are normalized by their value before learning and are averages over 50 realizations of the initial synaptic strengths $w_{ij}(0)$. Bars represent the standard deviation.

1.3.2 Structure changes

As the learning rule is applied, some synapses are thus strengthened while some others vanish. We now turn to the study of the way strengthened synapses distribute over the RRNN - or equivalently - to the structure of the RRNN. To this aim, we apply a threshold on the absolute value of the synaptic strength matrix $\mathbf{W}(t) = \{w_{ij}(t)\}$, to obtain a thresholded matrix $\mathbf{S}(t)$, whose elements define as : $s_{ij}(t) = \Theta\left(w_{ij}(t) - \epsilon\right)$, where ϵ is the threshold and $\Theta(.)$ the Heavyside step function ($\Theta(x) = 1$ if $x \geq 0$, and 0 else). Hence, $s_{ij}(t) = 1$ denotes that the absolute value of the $j \to i$ synapse at time t is stronger than ϵ. $\mathbf{S}$ thus reflects the network formed by the strongest synapses only. To quantify the structure of this network and its evolution during learning, we compute two classical observables from complex networks analysis: the clustering index $< C >$ and the mean shortest path $< l >$ (for definitions of these quantities, see [**?**]).

Fig. 1.2A & B show the evolution of these quantities for several threshold values. Before

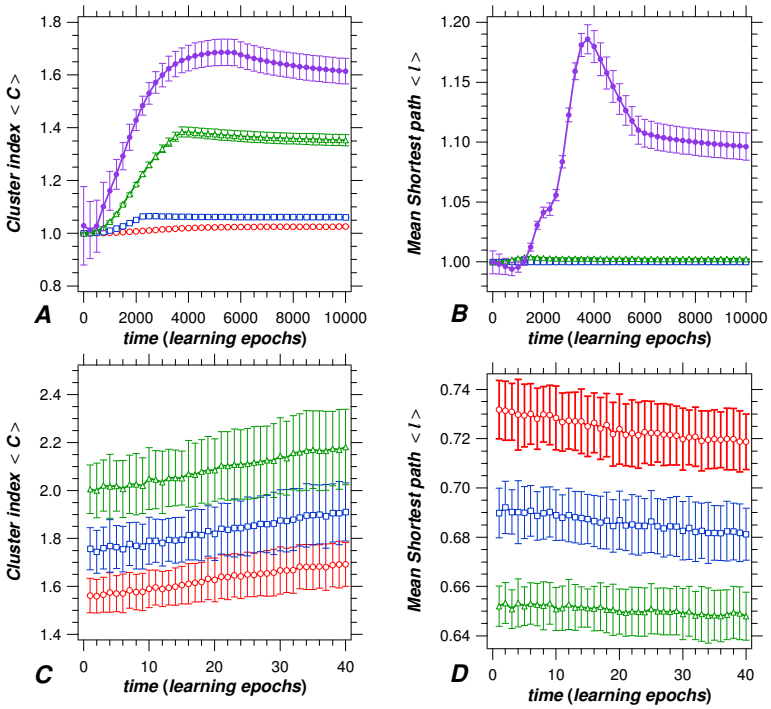

Figure 1.2: Network structure changes during learning with rule Eq.(1.2). Evolution of clustering coefficient (**A**) or mean shortest path (**B**) of the network of strong synapses at long learning times. Thresholds $\epsilon = 0.01$ (open circles), 0.05 (squares), 0.08 (triangles) or 0.12 (full circles). (**C**) and (**D**) present the same measurements, applied to the network of fast learning synapses ($\delta\mathbf{W}$) during reduction of the dynamics (see Fig. 1.1). Here, thresholds are $\epsilon = 10^{-9}$ (circles), 10^{-8} (squares) or 10^{-7} (triangles). All values are normalized by the value that would be obtained with a comparable random graph (i.e. a random graph with the same number of neurons and synapses). Note that computation of the mean shortest path in (**B & D**) is obtained by averaging over neuron pairs that can be connected by a path. Bars present one standard deviation, as estimated from 50 realizations of the initial synapse strength distribution.

learning, synaptic strengths are randomly (homogeneously) distributed over the network whatever the applied threshold, so that $< l >$(Fig. 1.2B) as well as $< C >$(Fig. 1.2A) remain identical to their values in comparable random networks. At long learning times (i.e. long after the dynamics reaches the fixed-point regime), increasing thresholds uncover a slight increase (less than 10%) of the MSP (Fig. 1.2B) which indicates that the average degree of separation between two neurons increases only slightly. However, $< C >$ increases up to 60%, compared to purely random networks, see Fig. 1.2A. Thus, at long learning times, the distribution of large synaptic weights over the network is no more random. In particular, the probability that two neurons i and j are connected through a strong synapse w_{ij} is much more likely if i and j have a third common neighbor k to which they are connected *via* two strong synapses (w_{ik} and w_{jk}). Hence the probability to find triangular circuits between neurons i, j and k where all implied synapses are strong is increased by learning. However, this distribution guarantees that the number of strong synapses separating any two neurons in the network remains very low. In other words the Hebbian learning rule Eq.(1.2) organizes the strong synapse network as a "small-world" network (in the sense of Watts & Strogatz's model[**?**]).

The above structural changes cannot be detected at short learning times because the synapse strength matrix $\mathbf{W}$ has not been modified enough to be distinguished from a random network by our analysis. To circumvent this problem, we applied similar analysis to the matrices of synaptic strength modifications (*increments*) $\delta\mathbf{W}(t)$ whose elements are given by $\delta w_{ij}(t) = w_{ij}(t) - w_{ij}(t-1)$. Here again, after thresholding of the absolute values, $\delta w_{ij}(t) = 1$ indicates

that the increase *rate* (in absolute value) of the $j \to i$ synapse between $t-1$ and t has been greater than the threshold value. $\delta\mathbf{W}(t)$ thus characterizes the network of rapidly increasing synaptic strengths. Fig. 1.2C & D show the structure of this matrix, *at the same time scale as that of the reduction of the dynamics*. Clearly, during these early steps, the learning rule organizes rapidly strengthening synapses as a small-world network. At longer learning times, iterated applications of this structural scheme will ultimately be perceptible at the level of the synaptic strength matrix $\mathbf{W}$ itself, resulting in its small-world organization.
Of course, these observations can be related to the modifications of the local loop content described previously (§1.3.1). Indeed, as the learning rule preferentially strengthens positive local loops, excitatory synapses interconnecting small clusters of neurons tend to strengthen. In particular, positive 3-loops made of excitatory synapses, tend to be favored and all of their synapses will be more likely to increase above the threshold value used to compute the structural observables. The network of strong synapses will thus be enriched with triangular motives, which explains the observed increase of the clustering coefficient.

1.3.3 Function

Here, our aim is to illustrate how the dynamics of the system sustains the network function. Fig. 1.3 shows the evolution of the dynamics attractor for the network (obtained by plotting $\langle x(t+1)\rangle$ vs $\langle x(t)\rangle$) at different stages of a simulation. Fig. 1.3A shows the strange attractor underlying chaotic dynamics, in the absence of input pattern and without learning. Fig. 1.3B shows the attractor obtained when the input pattern is fed to the same network. Clearly, the attractor is only slightly modified, yielding chaotic dynamics that remains very close to the spontaneous case (no input). Hence, in the absence of learning, the application of an input pattern does not really modify the network dynamics. Fig. 1.3C then shows how the initial attractor is restored when the input pattern is withdrawn. We then presented the input pattern in the presence of learning (i.e. applying Eq.(1.2)) for 81 learning epochs. The resulting attractor is shown Fig. 1.3D. As already outlined, learning resulted in a reduction of the dynamics to a (multiply folded) limit cycle, yielding regular oscillations. At that point, we turned off learning to compare the dynamics in the absence or presence of the input pattern. Withdrawing the input pattern settled the network to a different attractor, which, in that case, happened to be simpler than the one with input. This attractor (Fig. 1.3E) is clearly not chaotic, but the important point is that the system displayed a different behavior depending on the presence of the input pattern. Presenting again the same input pattern in the absence of learning brought (Fig. 1.3F) the network almost immediately to the same attractor as in Fig. 1.3D. Hence this very limit cycle attractor is specific of the input that the system has memorized and learned to recognize. Generally speaking these features define associative memory properties, here in the sense of a pattern-attractor association. Furthermore, learning is obtained by reduction of a chaotic dynamics to limit cycle oscillations, which is very close to the behavior observed upon odor recognition in the olfactory bulb[**?**, **?**].

Finally, we wished to investigate how the structure of the input pattern impacts on the state of the network. To this aim, we considered the difference between the neuron states x_is and the average network state $\langle x(t)\rangle$ at long learning times and display these differences using the reverse mapping from that was used for mapping the input cross to the neurons. Fig. 1.4B shows the results obtained after averaging over 50 realizations of the initial synapse strengths $w_{ij}(0)$, using the same input pattern. Clearly the state of the network Fig. 1.4B does not match the input pattern Fig. 1.4A, which indicates that input storage is largely distributed over the network in a way that depends on the interplay between the random initial synapse strengths and the input pattern.

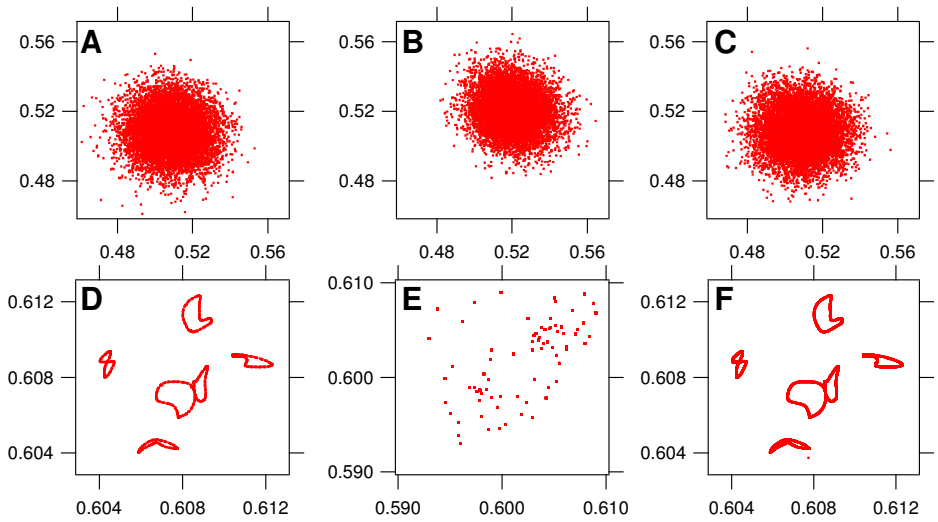

Figure 1.3: A typical simulation illustrating how the dynamics supports the network function. The panels present the dynamics attractors obtained through plotting $\langle x(t+1)\rangle$ vs $\langle x(t)\rangle$ at different stages of the simulation. See text for details.

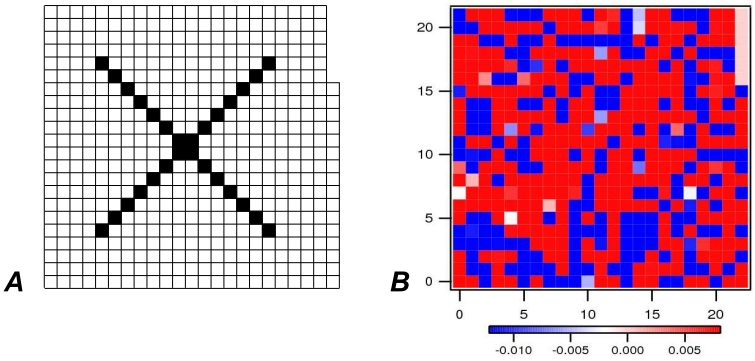

Figure 1.4: Comparison between the input pattern **A** and the network state after learning **B**. (**A**) The structured input pattern used in this study. (**B**) State of the network after learning the pattern. Each pixel i shows in color-code, the difference between the state of neuron i and the average network state, $x_i(t) - \langle x(t)\rangle$, at long learning times. Results are averaged over 50 realizations of the initial synapse strengths.

1.4 Conclusion and future work

Our aim in the present paper was to study how the structure, dynamics and function are related in RRNNs evolving with a simple Hebbian learning. We show that the learning rule Eq.(1.2) modifies the local loop content of the network, increasing the weights of positive loops and decreasing those of negative ones. This is likely to explain the reduction in the network dynamics complexity from chaos to regular limit cycle oscillations and fixed-point stability. Because the rule favors positive 3-loops, we show that the global structure of the network is progressively enriched with triplets of interconnected, strongly excitatory neurons. As a result, and because the mean shortest path length remains stable, strong synapses distribute on the RRNN structure as a small-world network. Finally, we show that this specific interplay between structure and dynamics enables the system to perform pattern recognition tasks.

A useful property of the learning rule studied (Eq.(1.2)) resides in its simplicity, which facilitates its analysis. Whether the conclusions formulated here can be generalized to other, more complex learning rules, is currently under study. One interesting rule known as the covariance rule, for instance, compares the unit activities with their average value

over a period T $(x_i(t) - \bar{x}_i)$. This rule yields decaying w_{ij} values when one neuron is active and the other inactive, and increasing ones when both neurons are inactive. While these properties are not strictly related to Hebb's law, they result in a bounded evolution for the synaptic strengths. Most of the conclusions presented in this paper for rule Eq.(1.2) (including small-world organization) are also valid with this rule. However, interpretation of the local loop content is much more difficult and will be the subject of future research. Conversely, other possible implementations behave in a very different way. For instance, the rule obtained by replacing $x_i(t+1).x_j(t)$ with $(x_i(t+1)-0.5).(x_j(t)-0.5)$ in (1.2) induces a similar reduction of the dynamics, but does not differentially modify the content in negative or positive loops, so that the small-world structure reported below is not observed. Here again, understanding the relationships between dynamics and structure modification in this case will be the subject of our further work.