Available online at www.sciencedirect.com
ScienceDirect
Journal of Physiology - Paris xxx (2007) xxx–xxx
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Effects of Hebbian learning on the dynamics and structure of random networks with inhibitory and excitatory neurons

Benoît Siri [a], Mathias Quoy [b], Bruno Delord [c], Bruno Cessac [d,e], Hugues Berry [a,*]

[a] INRIA – Futurs Research Centre, Project-Team Alchemy, 4 rue J Monod, 91893 Orsay Cedex, France
[b] ETIS, UMR 8051 CNRS-Université de Cergy-Pontoise-ENSEA, 6 avenue du Ponceau, BP 44, 95014 Cergy-Pontoise Cedex, France
[c] ANIM, U742 INSERM – Université P.M. Curie, 9 quai Saint-Bernard, 75005 Paris, France
[d] INLN, UMR 6618 CNRS-Université de Nice, 1361 route des Lucioles, 06560 Valbonne, France
[e] INRIA – Sophia Antipolis – Méditerranée Research Centre, Project-Team Odyssee, 2004 Route des Lucioles, 06902 Sophia Antipolis, France

Abstract

The aim of the present paper is to study the effects of Hebbian learning in random recurrent neural networks with biological connectivity, i.e. sparse connections and separate populations of excitatory and inhibitory neurons. We furthermore consider that the neuron dynamics may occur at a (shorter) time scale than synaptic plasticity and consider the possibility of learning rules with passive forgetting. We show that the application of such Hebbian learning leads to drastic changes in the network dynamics and structure. In particular, the learning rule contracts the norm of the weight matrix and yields a rapid decay of the dynamics complexity and entropy. In other words, the network is rewired by Hebbian learning into a new synaptic structure that emerges with learning on the basis of the correlations that progressively build up between neurons. We also observe that, within this emerging structure, the strongest synapses organize as a small-world network. The second effect of the decay of the weight matrix spectral radius consists in a rapid contraction of the spectral radius of the Jacobian matrix. This drives the system through the "edge of chaos" where sensitivity to the input pattern is maximal. Taken together, this scenario is remarkably predicted by theoretical arguments derived from dynamical systems and graph theory.

Keywords: Random recurrent neural networks; Hebbian learning; Network structure; Chaotic dynamics

1. Introduction

Neural networks show amazing abilities for information storage and processing, and stimulus-dependent activity shaping. These capabilities are mainly conditioned by the mutual coupling relationships between network structure and neuron dynamics. Actually, learning in neural networks implies that activity guides the way synapses evolve; but the resulting connectivity structure in turn can raise new dynamical regimes. This interaction becomes even more complex if the considered basic architecture is not feed-forward but includes recurrent synaptic links, like in cortical structures. Understanding this mutual coupling between dynamics and topology and its effects on the computations made by the network is a key problem in computational neuroscience, that could benefit from new approaches.

In the related field of dynamical systems interacting via complex coupling networks, a large amount of work has recently focused on the influence of network topology on global dynamics (for a review, see Boccaletti et al., 2006). In particular, much effort has been devoted to understanding the relationships between node synchronization and the classical statistical quantifiers of complex networks (degree distribution, average clustering index, mean shortest path, modularity…) (Grinstein and Linsker, 2005; Nishikawa et al., 2003; Lago-Fernández et al., 2000). The main idea was that the impact of network topology on the global

[*] Corresponding author.
E-mail address: hugues.berry@inria.fr (H. Berry).

0928-4257/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.jphysparis.2007.10.003

dynamics might be prominent, so that these structural statistics may be good indicators of the global dynamics. This assumption proved however largely wrong so that some of the related studies yielded contradictory results (Nishikawa et al., 2003; Hong et al., 2002). Actually, synchronization properties cannot be systematically deduced from topology statistics but may be inferred from the spectrum of the network (Atay et al., 2006). Accordingly, many studies have considered diffusive coupling of 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.

In this perspective, neural networks can be considered as mere examples of these complex systems, with the particularity that the dynamics of the network nodes (neurons) depends on the network links (synaptic weights), that themselves vary over time as a function of the node dynamics. Unfortunately, the coupling between neurons (synaptic weight) is rarely diffusive, so that the corresponding matrix is not symmetric and may contain positive and negative elements. Hence the mutual coupling between neuron dynamics and network structure remains largely to be understood.

Our general objective is to shed light on these interactions in the specific case of random recurrent neural networks (RRNNs). These network models display a rich variety of dynamical behaviors, including fixed points, limit cycle oscillations, quasiperiodicity and deterministic chaos (Doyon et al., 1993), that are suspected to be similar to activity patterns observed in the olfactory bulb (Skarda and Freeman, 1987; Freeman, 1987). It is known that the application of biologically-plausible local learning rules (Hebbian rules) reduces the dynamics of chaotic RRNNs to simpler attractors that are specific of the learned input pattern (Dauce et al., 1998). This phenomenon endows RRNNs with associative memory properties, but remains poorly understood.

Our previous work showed that the evolution of the network structure during learning can be tracked in numerical simulations via the classical topological statistics from “complex networks approaches” (Berry and Quoy, 2006). In a companion paper (Siri et al., 2007), we devise a mathematical framework for the effects of Hebbian learning on the dynamics, topology and some functional aspects of RRNNs. This theoretical approach is shown to explain the effect of learning in a “canonical” RRNN, i.e. a completely connected network where a neuron projects both excitatory and inhibitory synapses. The major advantage of this simplification is that it allows mathematical analysis. But this network type remains poorly realistic from a biological point of view. The aim of the present paper is thus to study the effects of learning with a more biological connectivity, for which the precision of the theoretical tools developed in our companion paper is not guarantied *a priori*.

In particular, we segregate the neurons into two distinct populations, namely excitatory (projecting only excitatory synapses) and inhibitory (projecting only inhibitory synapses) neurons. Furthermore, the network is sparsely connected and the overall connectivity parameters are fixed to emulate local circuitry in the cortex. We show that the application of Hebbian learning leads to drastic changes in the network dynamics and structure. We also demonstrate that the mathematical arguments mentioned above remain a very useful unifying framework to understand the effects of learning in this system.

2. The model

2.1. Connectivity

We consider networks with a total of $N = 500$ neurons and random connectivity. Each neuron is either inhibitory (with probability p_{I}) or excitatory (with probability $p_{\mathrm{E}} = 1 - p_{\mathrm{I}}$) and projects to $p_{\mathrm{c}}N$ randomly chosen postsynaptic neurons (independently of their excitatory or inhibitory nature). Probabilities are taken uniform on $[0, 1]$ for the network connectivity p_{c} and fraction of inhibitory neurons p_{I}. From a network point of view, this means that the connectivity network is (uncorrelated) random. For instance, the number of synapses between an excitatory and an inhibitory neuron will be proportional to $p_{\mathrm{E}}p_{\mathrm{I}}p_{\mathrm{c}}$. In the neurophysiology literature, this principle, known as “Peter’s rule” is usually considered a valid approximation of the neocortical microcircuitry (Binzegger et al., 2004). We fixed p_{I} and p_{c} so as to account for the neural circuitry of a typical neocortical column. Experimental quantification of the fraction of inhibitory neurons in the cat primary visual cortex showed limited variations from one layer to the other (with the exception of layer 1), with an average of 0.21 (Gabbott and Somogyi, 1986). Here we used $p_{\mathrm{I}} = 0.25$. The connectivity p_{c} of pyramidal neurons in rat somatosensory cortex has been experimentally estimated to ~0.10, (Markram et al., 1997; Kalisman et al., 2005). Geometrical analysis based on experimental data from the mouse cortex however yielded higher values, ~0.26 (Stepanyants et al., 2002). Here, we used $p_{\mathrm{c}} = 0.15$.

The initial weight of each synapse between a postsynaptic neuron i and a presynaptic neuron j, $W_{ij}^{(1)}$, is drawn at random, according to a Gamma distribution, whose parameters depend on the nature of the presynaptic neuron j. If j is inhibitory, $W_{ij}^{(1)} \sim \mathrm{Gamma}(-\mu_w/n_i, \sigma_w/n_i)$, where $\mathrm{Gamma}(m, s)$ denotes the Gamma distribution with mean m and standard deviation s, and $n_i = p_{\mathrm{I}}p_{\mathrm{c}}N$. If j is excitatory, then $W_{ij}^{(1)} \sim \mathrm{Gamma}(\mu_w/n_e, \sigma_w/n_e)$ where $n_e = p_{\mathrm{E}}p_{\mathrm{c}}N$. Using Gamma distributions (instead of Gaussian ones, for instance) allows to ensure that inhibitory (excitatory) neurons project only negative (positive) synapses, whatever the values of μ_w and σ_w. Thanks to the normalization terms (n_e and n_i), the total excitation received by a postsynaptic neuron is *on average* equal to the total inhibition it receives. Hence, in their initial setups (i.e. before learning) our networks are guarantied to conserve the excitation/inhibition balance (on average).

2.2. Dynamics

We consider firing-rate neurons with discrete-time dynamics and take into account that learning may occur on a different (slower) time scale than neuron dynamics. Indeed, synaptic plasticity is known to implicate intracellular sensors of the neuron average voltage activity, most notably intracellular calcium. Modifications of conductances in response to changes in intracellular calcium concentrations are however much slower than the dynamics of neuron membrane potential. The properties of such two-time scale systems have been exploited in the modelling literature to propose explanations to the tuning of rhythmic motor patterns (Soto-Trevino et al., 2001) or long-term storage of plastic modifications (Delord et al., 2007), for instance.

In the present model, we wished to capture this aspect while keeping the model as simple as possible. Hence, synaptic weights are kept constant for $\tau \geqslant 1$ consecutive dynamics steps, which defines a “learning epoch”. The weights are then updated and a new learning epoch begins. We denote by $t \geqslant 0$ the update index of neuron states (neuron dynamics) inside a learning epoch, while $T \geqslant 1$ indicates the update index of synaptic weights (learning dynamics).

Let $x_i^{(T)}(t) \in [0,1]$ be the mean firing rate of neuron i, at time t within the learning epoch T. Let $\mathscr{W}^{(T)}$ be the matrix of synaptic weights at the Tth learning epoch and $\boldsymbol{\xi}$ the vector $(\xi_i)_{i=1}^N$. Then the discrete time neuron dynamics (1) writes:

$$x_i^{(T)}(t+1) = f\left(\sum_{j=1}^{N} W_{ij}^{(T)} x_j^{(T)}(t) + \xi_i\right). \tag{1}$$

Here f is a sigmoidal transfer function ($f(x) = 1/2(1+\tanh(gx))$). The output gain g tunes the nonlinearity of the function and mimics the excitability of the neuron. ξ_i is an external input applied to neuron i and the vector $\boldsymbol{\xi}$ is the “pattern” to be learned (see below). $W_{ij}^{(T)}$ represents the weight of the synapse from presynaptic neuron j to postsynaptic neuron i during learning epoch T. Finally, at the end of one learning epoch, the neuron dynamics indices are reset: $x_i^{(T+1)}(0) = x_i^{(T)}(\tau)\ \forall i$.

2.3. Input pattern

The pattern to be learned by the network consists in the (time constant) external input ξ_i applied to each neuron i at each update step (Eq. (1)). For the purpose of the present paper, the exact value of this pattern is not very important, as soon as its maximal amplitude remains small with respect to the neuron maximal firing rate. Here, we use $\xi_i = 0.010\sin(2\pi i/N)\ \cos(8\pi i/N)\ \ \forall i = 1,\ldots,N$. 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. 6.

2.4. Learning

In the present work, we used the following Hebbian learning rule:

$$W_{ij}^{(T+1)} = \lambda W_{ij}^{(T)} + s_j \frac{\alpha}{N} m_i^{(T)} m_j^{(T)} \Theta(m_j^{(T)}), \tag{2}$$

where α is the learning rate, $s_j = +1$ if j is excitatory and -1 if it is inhibitory and Θ denotes the Heaviside step function ($\Theta(x) = 0$ if $x < 0$, 1 otherwise). 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 m_i or $m_j = 0$ (i.e. the pre or postsynaptic neurons are silent, see below), Eq. (2) leads to an exponential decay of the synaptic weights (hence passive forgetting). Another important consequence of this rule choice is that if $\lambda < 1$, the weights are expected to converge to stationary values. Hence $\lambda < 1$ also allows avoiding divergence of the synaptic weights. Note that there is no forgetting when $\lambda = 1$.

The second term in the RHS member of Eq. (2) generically accounts for activity-dependent plasticity, i.e. the effects of the pre and postsynaptic neuron firing rates. In our model, this term depends on the *history* of activities through the time-average of the firing-rate:

$$m_i^{(T)} = \frac{1}{\tau} \sum_{t=1}^{\tau} x_i^{(T)}(t) - d_i, \tag{3}$$

where $d_i \in [0,1]$ is a threshold that represents the maximal spontaneous firing rate of the neuron, i.e. the boundary between spontaneous and presynaptically evoked activity. In the present study, we set $d_i = 0.10$, $\forall i$. Hence, a neuron i will be considered active during learning epoch T whenever $m_i^{(T)} > 0$ (i.e. whenever its average firing rate has been >10% of the maximal value), and silent otherwise.

The definition of this learning rule was guided by the following tradeoff between biological knowledge and our simplified computational model. For excitatory synapses (i.e. $s_j = +1$), the rule Eq. (2) captures long-term potentiation (LTP) and homosynaptic long-term depression (LTD). LTP is evoked by the association of strong presynaptic and postsynaptic activities and derives from the molecular properties of NMDA channels (Bliss and Collingridge, 1993). In the rule, this phenomenon is reflected by an increase of the second term in the RHS member of Eq. (2) whenever $m_j > 0$ and $m_i > 0$. Concerning LTD, we neglected in the present study possible plastic interactions between synapses (i.e. heterosynaptic forms) (Tao et al., 2000; Nishiyama et al., 2000) and took into account only homosynaptic LTD. Hence, the second term in the RHS member of Eq. (2) decays for $m_j > 0$ and $m_i < 0$, which translates the biological conditions for homosynaptic LTD (strong presynaptic activity, low postsynaptic activity). Finally, in the absence of presynaptic activity, the NMDA receptors and downstream signalling pathways are not activated, so that no plastic change occurs (at least if heterosynaptic forms are neglected). This is accounted for in the model by the Heaviside term $\Theta(m_j^{(T)})$.

From a functional point of view, the rule Eq. (2) for excitatory synapses is a Hebbian rule, both in the sense that it is associative (depends on both pre and postsynaptic activities) and that it is anti-homeostatic (i.e. leads to diverging neuron activity).

In contrast to inhibitory synapses, there is much fewer experimental documentation of long-term plastic modifications for inhibitory synapses. LTP and LTD at GABAergic synapses have however been reported in several regions of the vertebrate brain, including the cerebellum and visual cortex (Kano, 1995). As in the case of excitatory synapses, heterosynaptic plastic interactions can occur (e.g. Nugent et al. (2007)). Here again, we focused on the homosynaptic forms only. To keep the learning rule as simple as possible, we have chosen to endow inhibitory synapses with a behavior that is symmetric to that defined for excitatory ones (i.e. we only change the value of s_j to $s_j = -1$), in agreement with some of the available experimental results. For instance, in rat cerebellum, the inhibitory synapse from Purkinje cell (PC) to the deep cerebellar nuclei (DCN) neurons undergoes LTP when both the pre and postsynaptic neurons are active, while homosynaptic LTD is observed with high PC activity and silent DCN neurons (Aizenman et al., 1998). This is accounted for in Eq. (2) by the decrease of the (negative) synaptic weight (increase of its absolute value) for $m_j > 0$ and $m_i > 0$, and its increase (decreasing absolute value) for $m_j > 0$ and $m_i < 0$.

For these inhibitory synapses, the learning rule Eq. (2) is thus still Hebbian in the sense that it remains associative. However, in contrast to its influence on excitatory synapses, it is homeostatic. For instance, if the presynaptic inhibitor and the postsynaptic neurons are both active, the inhibitory influence of the synapse will increase, thus contributing to a decrease of postsynaptic activity.

The resulting rule shares common features with classical learning rules. For instance, it includes time-averages of the neuron activities, as well as passive forgetting, like in the original Bienenstock–Cooper–Munro (BCM) rule (Bienenstock et al., 1982). However, in contrast to the latter, our rule does not consider sliding threshold mechanisms separating LTD from LTP. Furthermore, thresholding concerns presynaptic activities as well as postsynaptic ones. Finally, it is based on two time scales, i.e. the update rate for the neuron firing rate is different from the update rate for the synaptic weights. In fact, Eq. (2) is similar to the so-called "covariance" (or Linsker's) rule (Linsker, 1986; Montague and Sejnowski, 1994), for which the activity dependent plastic term (the second term in the RHS member of Eq. (2)) is proportional to $(x_i(t) - \theta_{\text{post}})\,(x_j(t) - \theta_{\text{pre}})$, where θ_{post} and θ_{pre} are two thresholds. Hence, our learning rule can be considered a two-time scale covariance rule with passive forgetting.

Note that definition (3) 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 genuine plasticity, while the latter may be related to meta-plasticity (Abraham and Bear, 1996). In this paper, we used $\tau = 10^4$. Finally, weights cannot change their sign. Note however that this setup does not have a significant impact on the present results.

3. Results

3.1. Spontaneous dynamics

We first present simulation results on the spontaneous dynamics of the system, i.e. the dynamics Eq. (1) *in the absence of learning*. The phase diagrams in Fig. 1 locate regions in the parameter space for which chaotic dynamics are observed in simulations. Each panel shows the isocurve $L_1 = 0$ (where L_1 is the largest Lyapunov exponent[1]) that represents the boundary between chaotic ($L_1 > 0$) and non chaotic ($L_1 < 0$) dynamics.

It is clear from this figure that chaotic dynamics are found for large parts of the parameter space. Generally speaking, chaotic behaviors disappear when the average weight μ_w increases, which may be related to an increase of the average neuron saturation. A more surprising observation is that chaotic dynamics tends to *disappear* when the gain of the transfer function g is increased. This behavior is in opposition to the behavior observed with classical random recurrent networks with homogeneous population (where each neuron has both excitatory and inhibitory projections). In the latter models (and even in related two-populations models, see Daucé et al. (2002)), chaotic dynamics usually appear for increasing values of g (see e.g. Cessac and Samuelides (2007)).

This is an interesting property of the spontaneous dynamics in our model, whose understanding is however out of the scope of the present paper and is left for future work. In the framework of the present study, these phase diagrams mainly allow locating suitable parameters for the initial conditions of our networks. We wish the initial dynamics to provide a wide range of possible dynamical regimes, a large (KS) entropy and self-sustaining dynamics. For these reasons, we set our initial dynamics inside the chaotic region, and fix $\mu_w = 50$, $\sigma_w = 1.0$, $g = 10$ and $N = 500$. The initial weight distribution will thus consist in a Gamma distribution with effective average -2.67 and s.d. 0.053 for inhibitory synapses, and 0.89 and 0.018, respectively, for excitatory ones.

3.2. Structure modifications

In this section, we want to study what changes are induced in the network structure by the learning rule Eq. (2). The quantification of the structure of random weighted

[1] The Lyapunov exponents have been numerically computed by the Eckmann–Ruelle method (Eckmann and Ruelle, 1985) using QR decomposition optimized by Von Bremen et al. (1985).

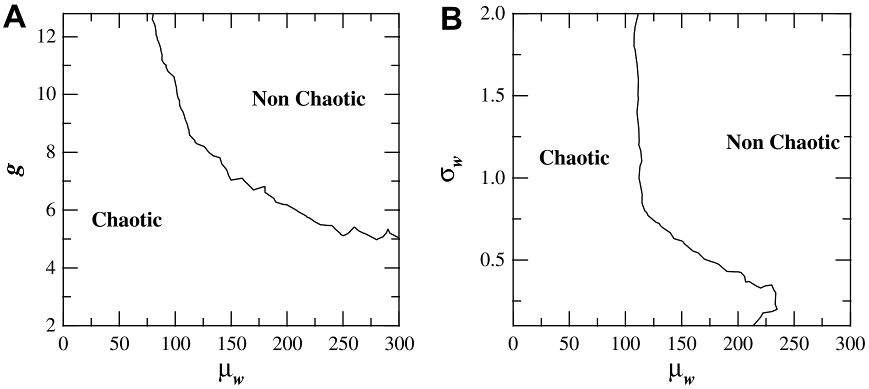

Fig. 1. Phase diagram for the spontaneous dynamics Eq. (1). The full line represents the boundary between chaotic and non chaotic dynamics (i.e. the isocurve $L_1 = 0$ where L_1 is the largest Lyapunov exponent). Shown are projection in (A) the (g, μ_w) plan with $\sigma_w = 1.0$ or (B) the (σ_w, μ_w) parameter plan with $g = 10.0$. Other parameters were: $\xi_i = 0.0\ \forall i = 1, \ldots, N$, $p_c = 0.15$, $p_I = 0.25$ and $N = 500$.

networks such a those obtained here is however not trivial. In the following, we adopt two different approaches. We first use quantifiers from the so-called "complex networks" approaches that mainly apply to the adjacency matrix. Albeit these quantifiers uncover important structural changes, we show they do not provide explanations for the evolution of the dynamics during learning. We then approach the problem from the perspective of the Jacobian matrix and show it actually provides useful tools to link structural to dynamical changes.

3.2.1. Adjacency matrix

3.2.1.1. Definitions. The major goal of the "complex networks" approaches is to develop quantitative tools for the characterization of random networks with complex (i.e. neither purely uncorrelated not purely regular) structures (for a review, see Boccaletti et al. (2006)). These quantifiers are usually defined on the adjacency matrix of the network, i.e. the matrix $\mathscr{A}$ whose elements $a_{ij} = 1$ if a synapse exists between i and j, and 0 otherwise.This matrix can be extracted from the weight matrix $\mathscr{W}$ by thresholding and binarization. Here, we applied a simple relative thresholding method that consists in keeping only the absolute values of the θ percent highest weights (again, in absolute value) from the nonzero connections in $\mathscr{W}$. Hence gradual decrease of θ enables to progressively isolate the adjacency network formed by the strongest weights only. The resulting matrix is then binarized and symmetrized, yielding the adjacency matrix $\mathscr{A}(\theta)$ whose elements $a_{ij}(\theta)$ indicate whether i and j are connected by a synapse with a large ($>\theta$) weight (either inhibitory or excitatory), compared to the rest of the network[2].

To characterize the structure of these matrices, the two main quantifiers are the clustering index and the mean shortest path (see Siri et al. (2007) for formal definitions). The clustering index C reflects the degree of "cliquishness" or local clustering in the network (Watts and Strogatz, 1998). It expresses the probability that two neurons connected to a third one are also connected together and thus can be interpreted as the density of triangular subgraphs in the network. The mean shortest path (MSP) is the average, over all nonidentical neurons pairs (i,j), of the smallest number of synapses one must cross to reach i from j. Note that these two quantifiers are informative only when compared to the same measures obtained from reference random networks, C_{rand} and MSP_{rand}[3].

3.2.1.2. Results. Fig. 2A and B shows simulation results for the evolution of the relative clustering index $C^{(T)}(\theta)/C^{(T)}_{\text{rand}}(\theta)$ and $\text{MSP}^{(T)}(\theta)/\text{MSP}^{(T)}_{\text{rand}}(\theta)$ during learning. The distribution of the initial weights over the network being totally random, one expects $C^{(1)}(\theta)/C^{(1)}_{\text{rand}}(\theta) \approx 1$ and $\text{MSP}^{(1)}(\theta)/\text{MSP}^{(1)}_{\text{rand}}(\theta) \approx 1\ \forall\theta$. This is confirmed in Fig. 2.

For $T \gtrsim 100$, the relative MSP remains essentially 1 for all thresholds θ (less than 4% variation, Fig. 2B). Hence, the average minimal number of synapses linking any two neurons in the network remains low, even when only large synapses are considered. Conversely, the clustering index (Fig. 2A) increases at $T > 100$ for the stronger synapses and reaches a stable value that is up to almost two twofold the value found in the reference random networks. Hence, if one considers the strong synapses at long learning epochs, the probability that the neighbors of a given neuron are themselves interconnected is almost twofold higher than if these strong synapses were laid at random over the

[2] We limit the range of θ values to ensure that not more that 10% of the neurons get disconnected from the network by the thresholding process.

[3] Here, to build reference networks, we start with the weight matrix at learning epoch T, $\mathscr{W}^{(T)}$ and rewire it at random but preserving the inhibitory/excitatory nature of the neurons. Hence for each element $W^{(T)}_{ij}$, we choose (uniformly) at random another element $W^{(T)}_{kl}$ *with the same sign*, and exchange their values. We then compute the clustering index and mean shortest path of the resulting rewired network, and average the obtained values over 15 realizations of the rewiring process, yielding the reference values $C_{\text{rand}}(\theta)$ and $\text{MSP}_{\text{rand}}(\theta)$.

Please cite this article in press as: Siri, B. et al., Effects of Hebbian learning on the dynamics and structure of ..., J. Physiol. (2007), doi:10.1016/j.jphysparis.2007.10.003

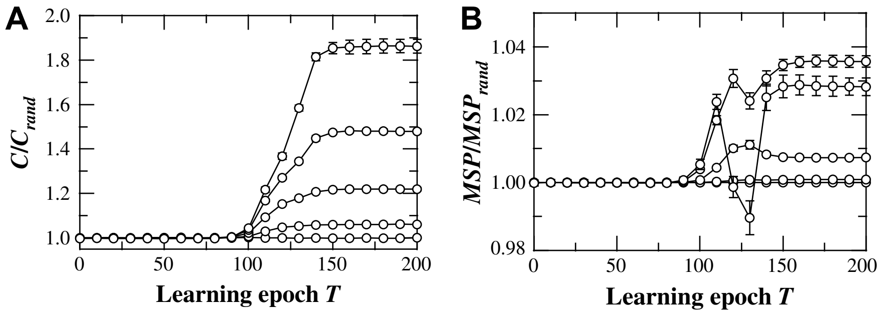

Fig. 2. Evolution of the normalized structural statistics during learning with rule Eq. (2). Values are averages over 20 different realizations of the network (random initial firing rates and synaptic weights). The values of the threshold θ are, from bottom to top in each panel, 100%, 87%, 73%, 60% and 47%. (*A*) Normalized clustering index $C^{(T)}(\theta)/C^{(T)}_{\text{rand}}(\theta)$. (*B*) Normalized mean-shortest path $\text{MSP}^{(T)}(\theta)/\text{MSP}^{(T)}_{\text{rand}}(\theta)$. Bars are ± 1 standard deviation. Other parameters were: $\lambda = 0.90$, $\alpha = 5 \times 10^{-3}$, $g = 10$, $\xi_i = 0.010\ \sin(2\pi i/N)\ \cos(8\pi i/N)\ \forall i = 1,\ldots,N$, $\mu_w = 50$, $\sigma_w = 1.0$ and $N = 500$.

network. In other terms, the learning rule yields correlations among the largest synapses at long learning epochs.

In the literature related to "complex networks", networks with a larger clustering index but a similar MSP with respect to a comparable reference random network, are referred to as *small-world* networks. Hence, the learning rule Eq. (2) organizes strong synapses as a "small-world" network.

Emerging experimental evidence shows that numerous brain anatomical and functional connectivity networks at several length scales indeed display a common small-world connectivity (for a recent review, see Bassett and Bullmore (2006)). These include quantifications of the physical (Shefi et al., 2002) or functional (Bettencourt et al., 2007) connectivity of neuronal networks grown in vitro; quantifications of the anatomical connectivity of *Caenorhabditis elegans* full neural system (Watts and Strogatz, 1998) or, at larger scale, cortico-cortical area connectivity maps in macaque, cat (Sporns and Zwi, 2004) and more recently human (He et al., 2007); and quantitative studies of functional human brain connectivity based on MEG (Stam, 2004), EEG (Micheloyannis et al., 2006) or fMRI data (Achard et al., 2006; Eguiluz et al., 2005).

An hypothesis for this frequent observation of small-world connectivity in real biological networks could be that small-world networks are emerging properties of neural networks subject to Hebbian learning. In favor of this possibility, small-world connectivity has recently been shown to arise spontaneously from spiking neuron networks with STDP plasticity and total connectivity (Shin and Kim, 2006) or with correlation-based rewiring (Kwok et al., 2007). Hence our present findings tend to strengthen this hypothesis.

Unfortunately, these indicators give no obvious clue about the mutual coupling between global dynamics and the network structure. Hence, in our case at least, the classical statistics of the "complex networks" do not provide causal explanation for the dynamical effects of learning. For instance, it does not help understand why dynamics complexity systematically decreases during learning. The adjacency matrix is however not the only viewpoint from which the network structure can be observed (see Siri et al. (2007) for a discussion). In the following, we propose as an alternative to examine the structure at the level of the Jacobian matrices.

3.2.2. Jacobian matrices

3.2.2.1. Definitions. Denote by $\mathbf{F}$ the function $\mathbf{F}: \mathbb{R}^N \to \mathbb{R}^N$ such that $F_i(\mathbf{x}) = f(x_i)$. In our case, the components of the *Jacobian matrix* of $\mathbf{F}$ at $\mathbf{x}$, denoted by $D\mathbf{F}_\mathbf{x}$ are given by

$$\frac{\partial F_i}{\partial x_j} = f'\left(\sum_{k=1}^{N} W_{ik}x_k + \xi_i\right)W_{ij} = f'(u_i)W_{ij}. \qquad (4)$$

Thus it displays the following specific structure:

$$D\mathbf{F}_\mathbf{x} = \Lambda(\mathbf{u})\mathscr{W}, \qquad (5)$$

with

$$\Lambda_{ij}(\mathbf{u}) = f'(u_i)\delta_{ij}. \qquad (6)$$

Note that $D\mathbf{F}_\mathbf{x}$ depends on $\mathbf{x}$, contrarily to $\mathscr{W}$. Generally speaking, $D\mathbf{F}_\mathbf{x}$ gives the effects of perturbations at the linear order. To each Jacobian matrix $D\mathbf{F}_\mathbf{x}$ one can associate a graph, called "the graph of linear influences". To build this graph, one draws an oriented link $j \to i$ iff $\frac{\partial f(u_i)}{\partial x_j} \neq 0$. The link is positive if $\frac{\partial f(u_i)}{\partial x_j} > 0$ and negative if $\frac{\partial f(u_i)}{\partial x_j} < 0$. A detailed presentation of the properties of the graph of linear influences can be found in Cessac and Samuelides (2007) and Siri et al. (2007). We just recall here that this graph contains *circuits or feedback loops*. If e is an edge, we denote by $\text{o}(e)$ the origin of the edge and $t(e)$ its end. Then a feedback loop is a sequence of edges $e_1,\ldots,e_k$ such that $\text{o}(e_{i+1}) = t(e_i)\ \ \forall i = 1,\ldots,k-1$, and $t(e_k) = \text{o}(e_1)$. A feedback loop is said positive (negative) if the product of its edges is positive (negative).

In general, positive feedback loops are expected to promote fixed-point stability (Hirsch, 1989) whereas negative loops usually generate oscillations (Thomas et al., 1981; Gouzé, 1998). In a model such as Eq. (1) the weight of a loop $k_1, k_2, \ldots, k_n, k_1$ is given by $\prod_{l=1}^{n} W_{k_{l+1}k_l} f'(u_{k_l})$, where $k_{n+1} = k_1$. Therefore, the weight of a loop is the product of a "topological" contribution $(\prod_{l=1}^{n} W_{k_{l+1}k_l})$ and a dynamical one $(\prod_{l=1}^{n} f'(u_{k_l}))$.

Please cite this article in press as: Siri, B. et al., Effects of Hebbian learning on the dynamics and structure of ..., J. Physiol. (2007), doi:10.1016/j.jphysparis.2007.10.003

3.2.2.2. Results. We measured the evolution of feedback loops during learning via the weighted-fraction of positive circuits in the Jacobian matrix, $R_n^{(T)}$, that we defined as

$$R_n^{(T)} = \frac{\sigma_n^{+(T)}}{|\sigma_n^{+(T)}| + |\sigma_n^{-(T)}|}, \tag{7}$$

where $\sigma_n^{+(T)}$ (resp. $\sigma_n^{-(T)}$) is the sum of the weights of every positive (resp. negative) feedback loops of length n in the Jacobian network at learning epoch T. Hence $R_n^{(T)} \in [0, 1]$. If its value is >0.5, the positive feedback loops of length n are stronger (in total weight) than the negative ones in the network. We computed the weighted-fraction of positive feedback loops for length $n = 2$ and $n = 3$ (i.e. $R_2^{(T)}$ and $R_3^{(T)}$).

The evolutions of $R_2^{(T)}$ and $R_3^{(T)}$ are presented in Fig. 3A. During the first ≈20 learning epochs, the time course of these quantities are highly noisy (and the corresponding standard deviation very large), so that we could not interpret them conclusively. However, a $T \approx 25$ learning epochs, $R_2^{(T)}$ stabilizes to values <0.5 ($R_2^{(1)} \approx 0.47$), indicating a slight imbalance in favor of negative feedback loops over positive ones. According to the above theoretical considerations, this indicates a trend toward complex oscillatory dynamics. This viewpoint may be considered another perspective to explain the initial chaotic dynamics. Note however that the initial imbalance in circuits of length-3 is much more modest, $R_3^{(1)} \approx 0.497$.

When $25 < T < 50$, $R_2^{(T)}$ increases and converges to ≈0.50. A dynamical interpretation would be that the corresponding dynamics attractors become progressively less chaotic and more periodic. This is exactly the behavior observed in the simulations (see Fig. 5B). Hence in spite of the huge fluctuations observed at the beginning, the study of the feedback loops in the Jacobian matrix offers a useful interpretation to the reduction of dynamics induced by learning at short learning epochs.

Upon further learning, $R_2^{(T)}$ and $R_3^{(T)}$ remain constant at 0.5 up to $T \approx 100$ learning epochs. Thus, these quantities do not detect variations in the balance between positive and negative feedback loops for $50 < T < 100$. However, at longer times ($T > 100$), $R_2^{(T)}$ and $R_3^{(T)}$ both increase abruptly and rapidly reach ≈0.62 for $R_2^{(T)}$ and ≈0.56 for $R_3^{(T)}$. Hence, at long learning epochs, the system switches to a state where positive feedback loops hold a significantly larger weight as compared to negative ones. Note that the time course of these indicators for $T > 25$ closely follows the time course of the relative clustering index (Fig. 2A). The causal relation between these two phenomena is however not obvious.

Because of the particular form of the Jacobian matrix in our system, the sign of a feedback loop is given by the sign of the weights along it (see above). We thus proceeded (Fig. 3B) to the computation of the evolution of the weighted-fraction for feedback loops computed in $\mathcal{W}$, i.e. we compute here the weight of a feedback loop $e_1, \ldots, e_k$ as the product of the *synaptic* weights of its edges, thus independently of the neuron state. The evolution of the weighted-fraction of positive feedback loops in $\mathcal{W}$ does not account for the initial imbalance observed in the feedback loops of $D\mathbf{F}$. However, its evolution at long times is remarkably identical to that measured in $D\mathbf{F}$. Thus, the weighted-fraction of positive feedback loops in $\mathcal{W}$ is able to account for at least part of the evolution of the dynamics and represents a link between purely structural and purely dynamical aspects. However, more information can be extracted by a more dynamical approach.

3.3. Dynamical perspective

As shown above, studying the evolution of the network structure during learning can yield valuable information about general characteristics such as oscillatory or fixed-point regimes. It is however not enough to explain the major dynamics changes. In the following, we inspect the system from a purely dynamical perspective, and show that related theoretical tools allow a quantitative explanation of the network evolution due to learning.

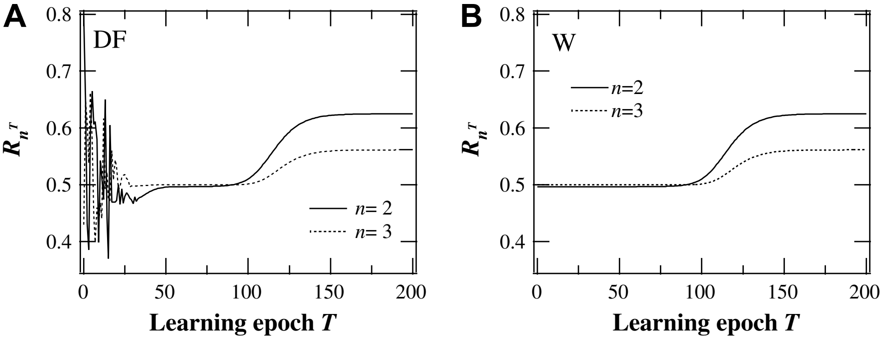

Fig. 3. Evolution of the weighted-fraction of positive feedback loops $R_n^{(T)}$ for loops in $D\mathbf{F}$ (A) and $\mathcal{W}$ (B) and circuit length $n = 2$ (thick line) and $n = 3$ (dotted line). The weighted-fraction of positive feedback loops is computed according to $R_n^{(T)} = \frac{\sigma_n^{+(T)}}{|\sigma_n^{+(T)}|+|\sigma_n^{-(T)}|}$ where $\sigma_n^{+(T)}$ (resp. $\sigma_n^{-(T)}$) is the sum of the weights of every positive (resp. negative) feedback loops of length n in the network at learning epoch T. The loops are taken either in $D\mathbf{F}$ (A) or in $\mathcal{W}$ (B). Values are averages over 20 different networks using $\lambda = 0.90$. Standard deviations are omitted for readability purpose. See text for definitions. All other parameters as in Fig. 2.

Please cite this article in press as: Siri, B. et al., Effects of Hebbian learning on the dynamics and structure of ..., J. Physiol. (2007), doi:10.1016/j.jphysparis.2007.10.003

Q1 ~~*Dynamical regimes*~~

Starting from spontaneous chaotic dynamics, application of the Hebbian learning rule (2) in our sparse two-populations model leads to dynamics simplification, as in the case of completely-connected, one-population random recurrent neural networks (Dauce et al., 1998). Fig. 5B shows the network-averaged neuron dynamics obtained at different learning epochs. The dynamics, initially chaotic ($T = 1$), gradually settles onto a periodic limit cycle ($T = 270$), then on a fixed point attractor at longer learning epochs (see e.g. $T = 290$ in this figure). This evolution of the global dynamics is a typical example of the reduction of the attractor complexity due to the mutual coupling between weight evolution and neuron dynamics.

In Siri et al. (2007), we developed a theoretical approach derived from dynamical systems and graph theory and evidenced that it explains this reduction of complexity in homogenous (single population) recurrent neural networks. We shall show thereafter that it also provides a useful framework for the present model. Below, we first summarize the main results obtained from this mathematical analysis (for details, see Siri et al. (2007)).

3.3.1. Main theoretical results

The first prediction of our approach is that Hebbian learning rules contract the norm of the weight matrix $\mathscr{W}$. Indeed, we could compute the following upper bound:

$$\|\mathscr{W}^{(T+1)}\| \leqslant \lambda^T \|\mathscr{W}^{(1)}\| + \frac{\alpha}{N}\frac{1}{1-\lambda}C, \tag{8}$$

where $\|\|$ is the operator norm (induced e.g. by Euclidean norm) and C a constant depending on the details of the rule. Hence the major effect of the learning rule is expected to be an exponentially fast contraction of the norm (or equivalently the spectral radius) of the weight matrix, which is due to the term λ, i.e. to passive forgetting ($\lambda < 1$).

The next prediction concerns the spectral radius of the Jacobian matrix. Starting from the specific form of the Jacobian matrix in our case, Eq. (5) and noting that $|\mu_1^{(T)}(\mathbf{x})| \leqslant \|D\mathbf{F}_{\mathbf{x}}^{(T)}\|$, one can easily derive a bound for the spectral radius of $D\mathbf{F}_{\mathbf{x}}^{(T)}$:

$$|\mu_1^{(T)}(\mathbf{x})| \leqslant \max_i f'(u_i^{(T)})\|\mathscr{W}^{(T)}\|, \tag{9}$$

where $\max_i$ denotes the maximum over the N neurons. This equation predicts a contraction of the spectrum of $D\mathbf{F}_{\mathbf{x}}^{(T)}$ that can arise via two effects: either the contraction of the spectrum of $\mathscr{W}^{(T)}$ and/or the decay of $\max_i f'(u_i)$, which arises from saturation in neuron activity. Indeed, $f'(u_i)$ is small when x_i is saturated to 0 or 1, but large whenever its synaptic inputs are intermediate, i.e. fall into the central part of the sigmoid $f(u_i)$. We emphasize that when $\lambda = 1$, $\mathscr{W}^{(T)}$ and $\mathbf{u}^{(T)}$ can diverge and lead $\max_i f'(u_i^{(T)})$ to vanish. Hence the spectral radius of the Jacobian matrix can decrease even in the absence of passive forgetting. In all cases, if the initial value of $|\mu_1^{(T)}(\mathbf{x})|$ is larger than 1, Eq. (9) predicts that the spectral radius may decrease down to a value <1. Note that in discrete time dynamical systems the value $|\mu_1^{(T)}(\mathbf{x})| = 1$ locates a *bifurcation* of the dynamical system.

According to our present setting, the largest Lyapunov exponent, $L_1^{(T)}$ depends on the learning epoch T. We were able to show that:

$$L_1^{(T)} \leqslant \log(\|\mathscr{W}^{(T)}\|) + < \log(\max_i f'(u_i)) >^{(T)}, \tag{10}$$

where $<\log(\max_i f'(u_i)) >^{(T)}$ denotes the time average of $\log(\max_i f'(u_i))$, in the learning epoch T (see Siri et al. (2007) for formal definitions). The second term in the RHS member is related to the saturation of neurons. The first one states that $L_1^{(T)}$ will decrease if the norm of the weight matrix $\|\mathscr{W}^{(T)}\|$ decreases during learning, resulting in a possible transition from chaotic to simpler attractors.

Let $u_i^{(T)}(t) = \sum_{j=1}^{N} W_{ij}^{(T)} x_j^{(T)}(t) + \xi_i$, the local field (or membrane potential) of neuron i at dynamics step t within learning epoch T. Our theoretical work also showed that provided $\lambda < 1$, the vector $\mathbf{u} = (u_i)_{i=1}^{N}$ converges to a fixed point as $T \to +\infty$:

$$\langle \mathbf{u} \rangle^{(\infty)} = \boldsymbol{\xi} + \mathbf{H}^{(\infty)}, \tag{11}$$

where the elements of the vector $\mathbf{H}^{(\infty)}$ are given by

$$H_i^{(\infty)} = \frac{\alpha}{N(1-\lambda)} m_i^{(\infty)} \sum_j m_j^{(\infty)} \Theta(m_j^{(\infty)}) x_j^{(\infty)}. \tag{12}$$

Therefore, the asymptotic local field is predicted to be the sum of the input pattern plus an additional term $\mathbf{H}^{(\infty)}$, which accounts for the *history* of the system and can be weak or not, depending on the exact learning rule and system history.

In Siri et al. (2007), we studied the effects of Hebbian learning in a completely connected ($p_c = 1$) one-population network (i.e. where each neuron can project inhibitory (negative) and excitatory (positive) synapses) and showed that these analytical arguments explain and describe results of the system simulation with a very good accuracy.

While the model studied in the present work is much more compatible with our knowledge of biological neural networks, it is very different from the model studied in Siri et al. (2007). In the present model, the connectivity is (severely) sparse and the neurons are segregated in two distinct groups, with distinct synaptic properties. Furthermore, the learning rule Eq. (2) is also more complex. Hence, it is not clear whether the above theoretical arguments account for the current case. In particular, these arguments mainly provide upper bounds, whose quality is not guarantied. In the following sections, we present simulation results about the influence of learning on the network dynamics and function, using our theoretical framework as an oracle.

3.3.2. Dynamics evolution in the sparse 2-populations model

Fig. 4 shows the evolution of the spectral radius of $\mathscr{W}$, $|s_1^{(T)}|$ for $\lambda = 0.90$ or 0.99 in simulations of our sparse two-populations model with dynamics Eq. (1) and learning rule Eq. (2). Let $s_i^{(T)}$ be the eigenvalues of $\mathscr{W}^{(T)}$, ordered

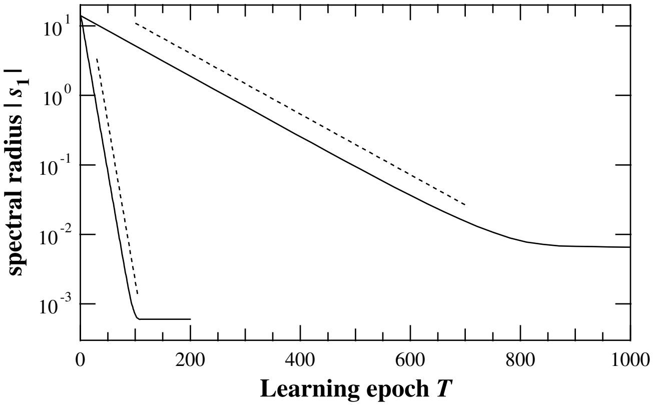

Fig. 4. Contraction of the spectral radius of $\mathscr{W}$. The evolution during learning of the norm of $\mathscr{W}$ largest eigenvalue, $|s_1^{(T)}|$ (thick lines) is plotted on a log-lin scale for $\lambda = 0.90$ (bottom) or 0.99 (top). Each curve is an average over 20 realizations with different initial conditions (initial weights and neuron states). For clarity, standard deviations are omitted but are always <20% of the average. Dashed thin lines are plots of exponential decreases with equation $g(T) \propto \lambda^T$. All other parameters as in Fig. 2.

such that $|s_1^{(T)}| \geqslant |s_2^{(T)}| \geqslant \cdots \geqslant s_i^{(T)} \geqslant \cdots$ Since $|s_1^{(T)}|$, the spectral radius of $\mathscr{W}^{(T)}$, is smaller than $\|\mathscr{W}^{(T)}\|$ one has from Eq. (8):

$$|s_1^{(T+1)}| \leqslant \lambda^T \|\mathscr{W}^{(1)}\| + \frac{\alpha}{N} \frac{1}{1-\lambda} C. \tag{13}$$

It is clear from this figure that in both cases the spectral radius decreases exponentially fast, with a rate that is very close to the prediction of the theory (i.e. $\propto \lambda^T$). Hence, the decay predicted by our analytical approach (Eq. (8)) is obviously observed in the simulations. Note that the clear trend in the simulation results for a decay proportional to λ^T, even tells us that the bound in (13) is indeed very good.

Fig. 7 shows (among other curves) the evolution of $|\mu_1^{(T)}(\mathbf{x})|$ (dashed thin line). This figure confirms that the theoretical prediction about the decay of $|\mu_1^{(T)}(\mathbf{x})|$ (Eq. (9)) is also valid for this model. Hence, Eq. (9) opens up the possibility that learning drives the system through bifurcations. This aspect is studied below (Section 3.3.3).

3.3.2.1. Evolution of the dynamics complexity. We now turn to directly study how the attractor complexity changes during learning. This information is provided by the computation of the largest Lyapunov exponent. Note that another 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, the KS entropy decreases as well.

Fig. 5A shows the evolution of $L_1^{(T)}$ during numerical simulations with different values of the passive forgetting rate λ. Its initial value ($L_1^{(1)} \approx 0.94$) is positive (we start our simulations with chaotic networks). As predicted by our theoretical approach (Eq. (10)), the Hebbian learning rule Eq. (2) leads to a rapid decay of $L_1^{(T)}$. The decay rate is indeed close to $\log(\|\mathscr{W}^{(T)}\|)$ for intermediate learning epochs, in agreement with the upper bound of Eq. (10). Hence $L_1^{(T)}$ quickly shifts to negative values, confirming the decrease of the dynamical complexity that could be inferred from visual inspection of Fig. 5B.

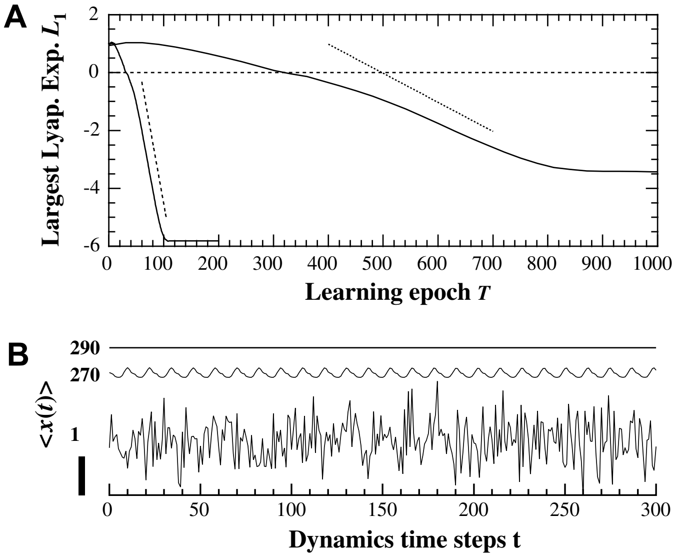

Fig. 5. Reduction of the dynamics complexity from chaotic to periodic and fixed point. (*A*) Evolution of the largest Lyapunov exponent L_1 (full thick lines) for $\lambda = 0.90$ (bottom) or 0.99 (top). Each value is an average over 20 realizations with different initial conditions (initial weights and neuron states). The thin dashed lines illustrate decays as $g(T) \propto T \log(\lambda)$. (*B*) Examples of network dynamics when learning is stopped at (from bottom to top) $T = 1$ (initial conditions), 270 or 290 and for $\lambda = 0.99$. 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 along the y-axis for readability. The height of the vertical black bar represents an amplitude of 0.1. All other parameters are as in Fig. 2.

3.3.2.2. Individual neuron activities. The former results yielded robust explanation of the evolution of global characteristics of the dynamics. Additional clue can also be obtained concerning more local aspects, such as the evolution of individual neuron activities. Fig. 6 shows the evolution of the local field $\mathbf{u}$ during learning. Clearly, the initial values are random, but the local field (thin gray line) shows a marked tendency to converge to the input pattern (thick dashed line) after as soon as 60 learning epochs. At $T = 180$, the convergence is almost complete. Hence this behavior once again conforms to the theoretical predictions Eq. (11), with $\boldsymbol{\xi} \gg \mathbf{H}^{(\infty)}$. In the results presented in this figure, we pursue the simulation up to $T = 200$, at which point we remove the pattern from the network, i.e. we set $\xi_i = 0\ \forall i$ (Fig. 2D). As a result, $\mathbf{u}$ looses its alignment from the pattern and presents a noisy aspect (note that each vector in the figure has been normalized to [0,1]). This behavior is once again in agreement with the theoretical predictions of Eq. (11), which indicates that $\langle \mathbf{u} \rangle^{(\infty)} = \mathbf{H}^{(\infty)}$ upon pattern removal.

To conclude, we have shown here that Hebbian learning in our system leads to a decrease of the attractor complexity

Please cite this article in press as: Siri, B. et al., Effects of Hebbian learning on the dynamics and structure of ..., J. Physiol. (2007), doi:10.1016/j.jphysparis.2007.10.003

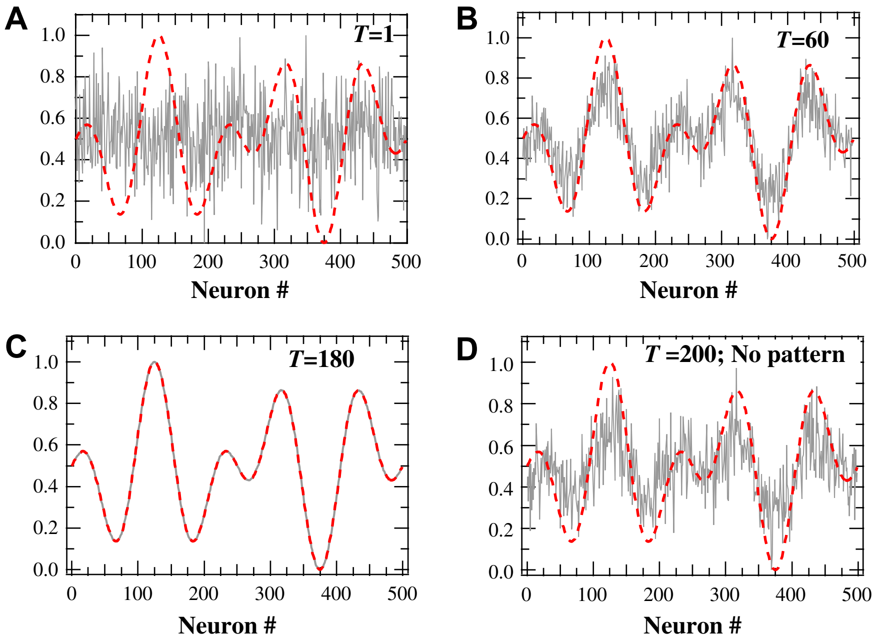

Fig. 6. Alignment of the local field $\mathbf{u} = \mathscr{W}\mathbf{x} + \xi$ (thin gray line) with the input pattern ξ (thick dashed line). Snapshot are presented at $T = 1$ (A, initial conditions), $T = 60$ (B), $T = 180$ (C) and $T = 200$ with pattern removed (D) learning epochs. Each curve plots averages over 20 realizations (standard deviations are omitted for comparison purposes), and every vector has been normalized to $[0, 1]$ for clarity. $\lambda = 0.90$ and all other parameters as in Fig. 2.

and entropy that can be induced by passive forgetting and/or an increased level of saturation of the neurons. This corresponds in details to the scenario predicted by our mathematical analysis.

3.3.3. Functional consequences

The former sections dealt with the effects of Hebbian learning on the structure and dynamics of the network but it does not say much about the links between the observed dynamics changes and the network function. We now focus on these functional aspects. The basic function of RRNNs is to learn a specific pattern ξ. In this framework, a pattern is learned when the complex (or chaotic) dynamics of the network settles onto a periodic oscillatory regime (a limit cycle) that is specific of the input pattern. This behavior emulates putative mechanisms of odor learning in rabbits that have been put forward by physiologists such as W. Freeman (Freeman, 1987; Freeman et al., 1988). An important functional aspect is that removal of the learned pattern after learning should lead to a significative change in the network dynamics. We now proceed to an analysis of this latter property.

3.3.3.1. Bifurcations and pattern sensitivity. The removal of ξ is expected to change the attractor structure and the average value of any observable ϕ (though with variable amplitude). Call $\Delta^{(T)}[\phi]$ the variation of the (time) average value of ϕ induced by pattern removal. If the system is away from a bifurcation point, removal will result in a variation of $\Delta^{(T)}[\phi]$ that remains proportional to ξ.

On the opposite, close to a bifurcation point this variation is typically not proportional to ξ and may lead to drastic changes in the dynamics. Call λ_k and $\mathbf{v}_k$ the eigenvalues and eigenvectors of $\mathscr{W}^{(T)}\Lambda(\mathbf{u}^{*(T)})$, ordered such that $|\lambda_N| \leqslant |\lambda_{N-1}| \leqslant |\lambda_1| < 1$. In the case where the dynamics has converged to a stable fixed point $\mathbf{u}^{*(T)}$ (namely, when $L_1^{(T)} < 0$, see e.g. Fig. 5), our theoretical work predicted that

$$\Delta^{(T)}[\mathbf{u}] = -\sum_{k=1}^{N} \frac{(\mathbf{v}_k, \xi)}{1 - \lambda_k}\mathbf{v}_k, \qquad (14)$$

where (,) denotes the inner product. As a matter of fact, the RHS term diverges if $\lambda_1 = 1$ and if $(\mathbf{v}_1, \xi) \neq 0$. We therefore expect pattern removal to have a maximal effect at “the edge of chaos”, namely when the value of the spectral radius of $D\mathbf{F}_{\mathbf{x}}$ is close to 1.

3.3.3.2. Simulation results. To study the effects of pattern removal in our model, we monitored the quantity

$$\Delta^{(T)}[\Lambda] = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(\langle \Lambda_{ii}(\mathbf{u})\rangle^{(T)} - \langle \Lambda_{ii}(\mathbf{u}')\rangle^{(T)}\right)^2} \qquad (15)$$

that measures how neuron excitability is modified when the pattern is removed. The evolution of $\Delta^{(T)}[\Lambda]$ during learning with rule Eq. (2) is shown on Fig. 7 (thick full lines) for two values of the passive forgetting rate λ. $\Delta^{(T)}[\Lambda]$ is found to increase to a plateau, and vanishes afterwards. Interestingly, comparison with the decay of the leading eigenvalue of the Jacobian matrix, μ_1 (thin full lines) shows that the maximal values of $\Delta^{(T)}[\Lambda]$ are obtained when $|\mu_1|$ is close to 1 and the largest Lyapunov exponent L_1 close to 0.

Hence, these numerical simulations are in agreement with the theoretical predictions that *Hebbian learning drives*

Please cite this article in press as: Siri, B. et al., Effects of Hebbian learning on the dynamics and structure of ..., J. Physiol. (2007), doi:10.1016/j.jphysparis.2007.10.003

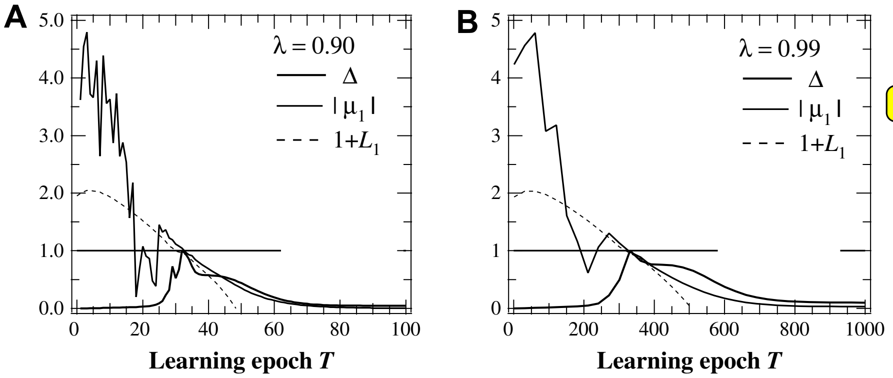

Fig. 7. Network sensitivity to the input pattern is maximal close to a bifurcation. The evolution of the average value of the spectral radius of $D\mathbf{F}_{\mathbf{x}}^{(T)}$ (thin full line) is plotted together with the sensitivity measure $\Delta^{(T)}[\Lambda]$ (thick full line) for $\lambda = 0.90$ (A) or 0.99 (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 (thin dashed line). The horizontal dashed-dotted lines locates $y = 1$. The values of $\Delta^{(T)}[\Lambda]$ are normalized to the [0, 1] range for comparison purposes. Each value is an average over 20 realizations (standard deviations are omitted for clarity). All other parameters were as in Fig. 2.

the global dynamics through a bifurcation, in the neighborhood of which sensitivity to the input pattern is maximal. Note that this property is obtained at the frontier where the chaotic strange attractor begins to destabilize ($|\mu_1| = 1$), hence at the so-called "edge of chaos". This particular dynamical regime, at the frontier between order (periodical or fixed point regimes) and disorder (chaos), has already be reported to be particularly suitable for recurrent neural networks, especially when computational power is considered (Soula et al., 2005; Bertschinger and Natschlager, 2004). The present results show that it is the optimal regime for the sensibility to the input pattern in our model. Whether this also implies improved or optimal computational performance remains however to be tested and will be the subject of future works.

It must finally be noticed that our theory predicts that pattern sensitivity should be maximal when $|\mu_1|$ is close to one. But several aspects of our simulation results are not accounted for by this theory. For instance, Fig. 7 shows that $|\mu_1|$ approaches 1 at several learning epochs. This is related to the "Arnold tongue" structure of the route to chaos. However, pattern sensibility is maximal only for the last episode, and almost zero for the former ones. This behavior is still unclear and will be the subject of future works.

4. Conclusion and future works

To conclude, we have shown in this work that Hebbian learning Eq. (2) has important effects on the dynamics and structure of a sparse two-populations RRNN. The forgetting part of the learning rule contracts the norm of the weight matrix. This effect, together with an increase in the average saturation level of the neurons, yields a rapid decay of the dynamics complexity and entropy. In other words, the network forgets its initial synaptic structure and is rewired by Hebbian learning into a new synaptic structure that emerges with learning and that depends on *the whole history of the neuron dynamics*. We have shown that the strongest synapses organize within this emerging structure as a small-world connectivity. The second effect of the decrease of the weight matrix and of the increased neuron saturation consists in a rapid contraction of the spectral radius of the Jacobian matrix. This leads the system to the edge of chaos, where sensitivity to the input pattern is maximal. This scenario is remarkably predicted by the theoretical arguments we developed in Siri et al. (2007).

In the presented simulations, most of the effects are mediated by the passive forgetting term. We believe that this term is not unrealistic from a biological point of view. Indeed, synaptic plasticity at the single synapse level is not permanent and some studies reported durations of 20 min (Volianskis and Jensen, 2003) or even 20 s (Brager et al., 2003). This would be accounted for in our model by $\lambda \ll 1$.

Nevertheless, most studies about long-term plasticity have evidenced longer cellular memory time constants, ranging from hours to days (Heynen et al., 2000; Racine et al., 1983; Doyere et al., 1996), which would correspond in our model to higher λ values. Note however that according to our mathematical analysis, most of the effects reported here are expected to occur even without passive forgetting (i.e. with $\lambda = 1$), provided the learning rule increases the average saturation of the neurons. In previous studies, we have considered such Hebbian learning rules devoid of passive forgetting but provoking increasing average saturation levels of the neurons. Numerical simulations have clearly evidenced a reduction of the attractor complexity during learning, in agreement with this suggestion (Berry and Quoy, 2006; Siri et al., 2006).

Future works will focus on the study of more detailed biological learning rules (heterosynaptic LTD, synaptic rescaling). We will also consider activity-dependent synaptic turnover (pruning/sprouting). Indeed albeit an overlooked phenomena for several decades, synaptic (or at least dendritic) turnover is now recognized as an important part of cortical networks, even in the adult (see e.g. Holtmaat et al. (2005)). Finally, one important problem with the application of RRNNs as artificial neural networks, is that it is very difficult to determine when to stop the learning process. Our results show that the effect of an input pattern is maximal at those learning epochs when the system is close to a bifurcation, but much more modest for shorter *and longer* learning times. One interesting development would thus consist in trying to find learning rules or settings that would guaranty that the system remains close to the edge of chaos, even at long learning times. As an attractive possibility, the plasticity of intrinsic properties (Daoudal and Debanne, 2003) could allow the network to stabilize its activity in this region.

Acknowledgment

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

References

Abraham, W.C., Bear, M.F., 1996. Metaplasticity: the plasticity of synaptic plasticity. Trends Neurosci. 19, 126–130.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., Bullmore, E., 2006. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J. Neurosci. 26, 63–72.

Aizenman, C., Manis, P., Linden, D., 1998. Polarity of long-term synaptic gain change is related to postsynaptic spike firing at a cerebellar inhibitory synapse. Neuron 21, 827–835.

Atay, F., Biyikoglu, T., Jost, J., 2006. Network synchronization: Spectral versus statistical properties. Physica D 224, 35–41.

Barahona, M., Pecora, L., 2002. Synchronization in small-world systems. Phys. Rev. Lett. 89, 054101.

Bassett, D., Bullmore, E., 2006. Small-world brain networks. The Neuroscientist 12, 512–523.

Berry, H., Quoy, M., 2006. Structure and dynamics of random recurrent neural networks. Adaptive Behavior 14, 129–137.

Bertschinger, N., Natschlager, T., 2004. Real-time computation at the edge of chaos in recurrent neural networks. Neural Comp. 16, 1413–1436.

Bettencourt, L., Stephens, G., Ham, M., Gross, G., 2007. Functional structure of cortical neuronal networks grown in vitro. Phys. Rev. E 75, 021915.

Bienenstock, E., Cooper, L., Munro, P., 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. J. Neurosci. 2, 32–48.

Binzegger, T., Douglas, R.J., Martin, K.A.C., 2004. A quantitative map of the circuit of cat primary visual cortex. J. Neurosci. 24, 8441–8453.

Bliss, T.V.P., Collingridge, G.L., 1993. A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361, 31–39.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U., 2006. Complex networks: Structure and dynamics. Phys. Reports 424, 175–308.

Brager, D., Cai, X., Thompson, S., 2003. Activity-dependent activation of presynaptic protein kinase c mediates post-tetanic potentiation. Nature Neurosci. 6, 551–552.

Cessac, B., Samuelides, M., 2007. From neuron to neural networks dynamics. EPJ Special Topics: Topics in Dynamical Neural Networks 142, 7–88.

Daoudal, G., Debanne, D., 2003. Long-term plasticity of intrinsic excitability: learning rules and mechanisms. Learn. Mem. 10, 456–465.

Dauce, E., Quoy, M., Cessac, B., Doyon, B., Samuelides, M., 1998. Self-organization and dynamics reduction in recurrent networks: stimulus presentation and learning. Neural Networks 11, 521–533.

Daucé, E., Quoy, M., Doyon, B., 2002. Resonant spatio-temporal learning in large random neural networks. Biol. Cybern. 87, 185–198.

Delord, B., Berry, H., Guigon, E., Genet, S., 2007. A new principle for information storage in an enzymatic pathway model. PLoS Comput. Biol. 3, e124.

Doyere, V., Errington, M., Laroche, S., Bliss, T., 1996. Low-frequency trains of paired stimuli induce long-term depression in area ca1 but not in dentate gyrus of the intact rat. Hippocampus 6, 52–57.

Doyon, B., Cessac, B., Quoy, M., Samuelides, M., 1993. Chaos in neural networks with random connectivity. Int. J. Bifurication Chaos 3, 279–291.

Eckmann, J., Ruelle, D., 1985. Ergodic theory of strange attractors. Rev. Mod. Phys. 57, 617–656.

Eguiluz, V., Chialvo, D., Cecchi, G., Apkarian, A., 2005. Scale-free brain functional networks. Phys. Rev. Lett. 94, 018102.

Freeman, W., 1987. Simulation of chaotic eeg pattern with a dynamic model of the olfactory system. Biol. Cyber. 56, 139–150.

Freeman, W., Yao, Y., Burke, B., 1988. Central pattern generating and recognizing in olfactory bulb: a correlation learning rule. Neural Networks 1, 277–288.

Gabbott, P.L.A., Somogyi, P., 1986. Quantitative distribution of GABA-immunoreactive neurons in the visual cortex (area 17) of the cat. Exp. Brain Res. 61, 323–331.

Gouzé, J., 1998. Positive and negative circuits in dynamical systems. J. Biol. Syst. 6, 11–15.

Grinstein, G., Linsker, R., 2005. Synchronous neural activity in scale-free network models versus random network models. PNAS 28, 9948–9953.

Hasegawa, H., 2005. Synchronisations in small-world networks of spiking neurons: Diffusive versus sigmoid couplings. Phys. Rev. E. 72, 056139.

He, Y., Chen, Z., Evans, A., 2007. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. Cerebral Cortex. Advance access published online January 4.

Heynen, A., Quinlan, E., Bae, D., Bear, M., 2000. Bidirectional, activity-dependent regulation of glutamate receptors in the adult hippocampus in vivo. Neuron 28, 527–536.

Hirsch, M., 1989. Convergent activation dynamics in continuous time networks. Neural Networks 2, 331–349.

Holtmaat, A., Trachtenberg, J., Wilbrecht, L., Shepherd, G., Zhang, X., Knott, G., Svoboda, K., 2005. Transient and persistent dendritic spines in the neocortex in vivo. Neuron 45, 279–291.

Hong, H., Kim, B., Choi, M., Park, H., 2002. Factors that predict better synchronizability on complex networks. Phys. Rev. E 65, 067105.

Kalisman, N., Silberberg, G., Markram, H., 2005. The neocortical microcircuit as a tabula rasa. Proc. Natl. Acad. Sci. USA 102, 880–885.

Kano, M., 1995. Plasticity of inhibitory synapses in the brain: a possible memory mechanism that has been overlooked. Neurosci. Res. 21, 177–182.

Kwok, H.F., Jurica, P., Raffone, A., van Leeuwen, C., 2007. Robust emergence of small-world structure in networks of spiking neurons. Cogn. Neurodyn. 1, 39–51.

Lago-Fernández, L.F., Huerta, R., Corbacho, F., Sigüenza, J.A., 2000. Fast response and temporal coherent oscillations in small-world networks. Phys. Rev. Lett. 84, 2758–2761.

Linsker, R., 1986. From Basic Network Principles to Neural Architecture: Emergence of Orientation-Selective Cells. Proc. Natl. Acad. Sci. USA 83, 8390–8394.

Markram, H., Lübke, J., Frotscher, M., Roth, A., Sakmann, B., 1997. Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. J. Physiol. 500, 409–440.

Micheloyannis, S., Pachou, E., Stam, C., Vourkas, M., Erimaki, S., Tsirka, V., 2006. Using graph theoretical analysis of multi channel eeg to evaluate the neural efficiency hypothesis. Neurosci. Lett 402, 273–277.

Montague, P.R., Sejnowski, T.J., 1994. The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. Learn. Mem. 1, 1–33.

Nishikawa, T., Motter, A.E., Lai, Y.C., Hoppensteadt, F.C., 2003. Heterogeneity in oscillator networks: are smaller worlds easier to synchronize ? Phys. Rev. Lett., 91.

Nishiyama, M., Hong, K., Mikoshiba, K., Poo, M., Kato, K., 2000. Calcium stores regulate the polarity and input specificity of synaptic modification. Nature 30, 584–588.

Nugent, F.S., Penick, E.C., Kauer, J.A., 2007. Opioids block long-term potentiation of inhibitory synapses. Nature 446, 1086–1090.

Racine, R., Milgram, N., Hafner, S., 1983. Long-term potentiation phenomena in the rat limbic forebrain. Brain Res. 260, 217–231.

Shefi, O., Golding, I., Segev, R., Ben-Jacob, E., Ayali, A., 2002. Morphological characterization of in vitro neuronal networks. Phys. Rev. E 66, 021905.

Shin, C.W., Kim, S., 2006. Self-organized criticality and scale-free properties in emergent functional neural networks. Phys. Rev. E 74, 045101.

Siri, B., Berry, H., Cessac, B., Delord, B., Quoy, M., 2006. Topological and dynamical structures induced by Hebbian learning in random neural networks. In: International Conference on Complex Systems. Boston.

Siri, B., Berry, H., Cessac, B., Delord, B., Quoy, M., 2007. A mathematical analysis of the effects of Hebbian learning rules on the dynamics and structure of discrete-time random recurrent neural networks. e-print: arXiv:0705.3690v1.

Skarda, C., Freeman, W., 1987. How brains make chaos in order to mahe sense of the world. Behav. Brain Sci. 10, 161–195.

Soto-Trevino, C., Thoroughman, K.A., Marder, E., Abbott, L., 2001. Activity-dependent modification of inhibitory synapses in models of rhythmic neural networks. Nat. Neurosci. 4, 297–303.

Soula, H., Alwan, A., Beslon, G., 2005. Learning at the edge of chaos: Temporal coupling of spiking neurons controller for autonomous robotics. In: AAAI Spring Symposium on Developmental Robotics. Stanford, CA, USA.

Sporns, O., Zwi, J., 2004. The small world of the cerebral cortex. Neuroinformatics 2, 145–162.

Stam, C., 2004. Functional connectivity patterns of human magnetoencephalographic recordings: a 'small-world' network? Neurosci. Lett. 355, 25–28.

Stepanyants, A., Hof, Patrick R., Chklovskii, Dmitri B., 2002. Geometry and structural plasticity of synaptic connectivity. Neuron 34, 275–288.

Tao, H., Zhang, L., Bi, G., Poo, M., 2000. Selective presynaptic propagation of long-term potentiation in defined neural networks. J. Neurosci. 20, 3233–3243.

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

Volianskis, A., Jensen, M., 2003. Transient and sustained types of long-term potentiation in the ca1 area of the rat hippocampus. J. Physiol. 550, 459–492.

Von Bremen, H., Udwadia, F., Proskuroswki, W., 1985. An efficient QR based method for the computation of lyapunov exponents. Physica D 101, 1–6.

Watts, D., Strogatz, S., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–442.