Autonomous Dynamics in Neural networks: The dHAN Concept and Associative Thought Processes

Claudius Gros

Institute for Theoretical Physics J.W. Goethe University Frankfurt, Germany.

Abstract. The neural activity of the human brain is dominated by self-sustained activities. External sensory stimuli influence this autonomous activity but they do not drive the brain directly. Most standard artificial neural network models are however input driven and do not show spontaneous activities.

It constitutes a challenge to develop organizational principles for controlled, self-sustained activity in artificial neural networks. Here we propose and examine the dHAN concept for autonomous associative thought processes in dense and homogeneous associative networks. An associative thought-process is characterized, within this approach, by a time-series of transient attractors. Each transient state corresponds to a stored information, a memory. The subsequent transient states are characterized by large associative overlaps, which are identical to acquired patterns. Memory states, the acquired patterns, have such a dual functionality.

In this approach the self-sustained neural activity has a central functional role. The network acquires a discrimination capability, as external stimuli need to compete with the autonomous activity. Noise in the input is readily filtered-out.

Hebbian learning of external patterns occurs coinstantaneous with the ongoing associative thought process. The autonomous dynamics needs a long-term working-point optimization which acquires within the dHAN concept a dual functionality: It stabilizes the time development of the associative thought process and limits runaway synaptic growth, which generically occurs otherwise in neural networks with self-induced activities and Hebbian-type learning rules.

Keywords: cognitive system theory, autonomous systems, neural networks, associative thought processes, clique encoding
PACS: 07.05.Mh, 84.35.+i, 87.18.Sn

COGNITIVE SYSTEM THEORY

The present approach is situated within the general framework of cognitive system theory. Let us start with a general definition of a cognitive system.

> *Cognitive systems*
> A cognitive system is a continuously active complex adaptive system autonomously exploring and reacting to the environment with the capability to 'survive'.

A cognitive system is an abstract dynamical system. It might be either biological or cybernetical. Our brain, to give an example, is the physical support of the human cognitive system. A cognitive system 'dies' whenever its physical support looses functionality.

The condition for 'survival' can be phrased in a mathematical precise way. The physical support of a cognitive system remains functional only when a set of key

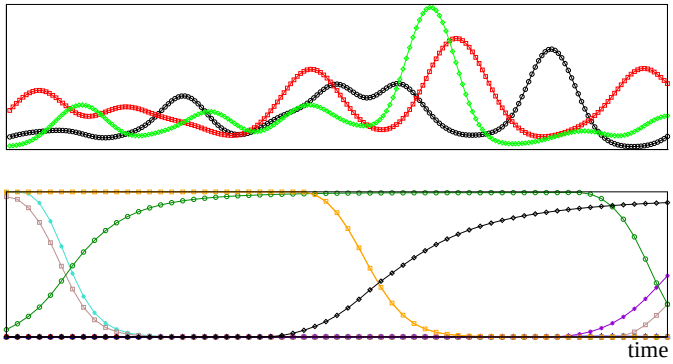

FIGURE 1. Illustration of fluctuating (top) and transient-state dynamics (bottom).

parameters remain within a given range. Examples for such 'survival parameters' are the blood-sugar level for a biological cognitive system or the battery status for an autonomous robot. The cognitive system receives information about the status of these survival parameters via appropriate sensors. It survives if its activity keeps its physical support functional via appropriate motor outputs.

It is desirable to develop artificial cognitive systems which could operate in a wide range of possible real-world or simulated environments, *viz.* cognitive systems with universal capabilities which are not tailored for specific task solutions. There are then several important points to be taken into account.

1. Autonomous dynamics
   The human brain dynamics is dominantly self-sustained. It is influenced but not driven by the sensory input [1]. It is therefore necessary to propose and to study possible organizational principles for neural networks with self-sustained dynamical activities.
2. Homeostatic principles
   The brain adapts itself autonomously to a wide range of growth conditions and injuries. It is therefore of interest to study neural-network layouts which regulate most parameters, as synaptic strength and learning rates, homeostatically.
3. Unsupervised learning
   Most learning by an autonomous cognitive system should be unsupervised - the system selects when and what to learn. Learning rules should be local, such that the system is scalable and remains functional under structural modifications.
4. Online learning
   There should be no distinct phases for learning and performance. Learning should be 'on-the-fly'.
5. Universality
   The layout principles for the cognitive system should be based, as far as possible, on universal principles. A priori knowledge about the environment can be added in a second step, if necessary, in order to boost the performance for specific tasks.

BASIC COGNITIVE SYSTEM THEORY PRINCIPLES

In addition to the rather general principles stated above one can formulate, guided by the results of neurobiological studies, several important guidelines.

- Competitive brain dynamics
  Studies of the neural correlate for conscious cognitive states suggest the formation of 'winning coalitions', also called 'critical reentrant events' [2], of competitively active neural ensembles. This competitive brain dynamics takes place in what is called a 'global workspace' [3], made-up of essential nodes [4].
- Transient-state dynamics
  Competitive dynamics naturally results in transient state dynamics, see Fig. 1, as the winning coalition of neural ensembles suppresses the activities of competing centers. A time series of semi-stable winning coalitions of computational subunits then results.
- Autonomous brain dynamics
  The spontaneously generated neural activity patterns generated in the cortex are not void of contextual information. It has been observed, that they resemble (in the visual cortex) memories of previous visual stimuli [5, 6], forming transient states [7].
- Associative thought processes
  Humans dispose over a huge commonsense database, mostly organized associatively [8, 9]. It is therefore reasonable to assume, that the autonomous dynamics reflects this fact. A possible paradigm for the self-sustained dynamics, which we will follow here, is than that of associative thought processes. Subsequent transient states then correspond to memories connected associatively.
- Sparse coding
  Neural networks with sparse coding[1] have a storage capacity orders of magnitude larger than networks with an average activity level of 50% [10]. Clique[2] encoding, an instance of a 'winners-take-all' encoding, combines competitive dynamics with the large storage capacity of sparse coding. For clique encoding, which we will consider here, the winning coalitions are constituted by mutually supporting neural activity centers. The notion of clique encoding also draws from studies in cognitive science indicating the importance of the 'chunking mechanism'[3] for human learning [11, 12].

The study of neural networks which incorporate above principles is hence a necessary step towards the eventual development of an artificial cognitive system. Here we present and study a generalized neural network which implements these requirements necessary

---

[1] Sparse coding is present in a neural network when, on the average, only a small fraction of all neurons is active simultaneously.

[2] In network theory one denotes by 'clique' a fully interconnected subcluster.

[3] Chunking denotes the notion of grouping together elementary units of information for memory formation

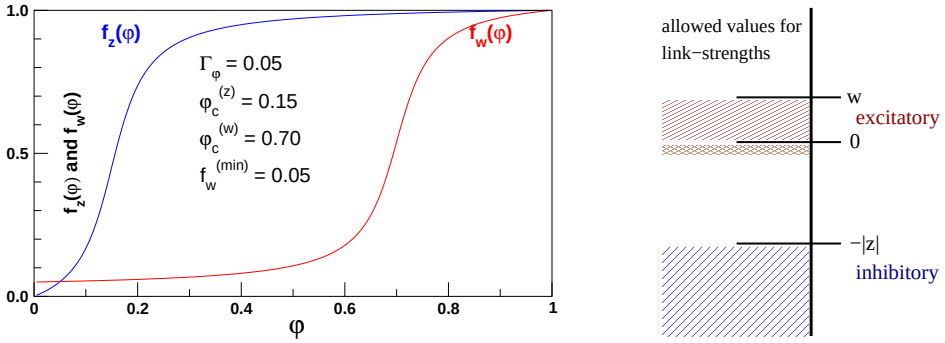

FIGURE 2. Left: Illustration of the reservoir functions $f_{z/w}(\varphi)$, see Eq. 2, of sigmoidal form with respective turning points $\varphi_c^{(f/z)}$, a width Γ_φ and a minimal value $f_z^{(min)} = 0$.
Right: Distribution of the synapsing strength leading to clique encoding. Weak inhibitory synapsing strengths do not occur.

for any component of an autonomous cognitive system. The network we examine is suitable to store patterns occurring in the sensory data input stream via unsupervised learning.

ASSOCIATIVE THOUGHT PROCESSES

We now present an implementation, in terms of a set of appropriate coupled differential equations, of the notion of associative thought processes as a time series of transient attractors. We start by a network with a large number of stable attractors (the cliques) and then turn these attractors into transient attractors by coupling to a second variable with a long time scale. To be concrete, we denote with $x_i \in [0,1]$ the activities of the local computational units constituting the network and with $\varphi_i \in [0,1]$ a second variable which we call *'reservoir'*. The differential equations [13]

$$\dot{x}_i = (1-x_i)\,\Theta(r_i)\,r_i + x_i\,\Theta(-r_i)\,r_i \tag{1}$$

$$r_i = \sum_{j=1}^{N}\Big[f_w(\varphi_i)\Theta(w_{ij})w_{i,j} + z_{i,j}\,f_z(\varphi_j)\Big]x_j \tag{2}$$

$$\dot{\varphi}_i = \Gamma_\varphi^+\,(1-\varphi_i)(1-x_i/x_c)\Theta(x_c-x_i) - \Gamma_\varphi^-\,\varphi_i\,\Theta(x_i-x_c) \tag{3}$$

$$z_{ij} = -|z|\,\Theta(-w_{ij}) \tag{4}$$

generate associative thought processes. We now discuss some properties of (1-4).

- Normalization
  Eqs. (1-3) respect the normalization $x_i, \varphi_i \in [0,1]$, due to the prefactors x_i,$(1-x_i)$, φ_i and $(1-\varphi_i)$ in Eqs. (1) and (3), for the respective growth and depletion processes. $\Theta(r)$ is the Heaviside-step function: $\Theta(r<0)=0$ and $\Theta(r>0)=1$.
- Synapsing strength
  The synapsing strength is split into an excitatory contribution $\propto w_{i,j}$ and an inhibitory contribution $\propto z_{i,j}$, with $w_{i,j}$ being the primary variable: The inhibition $z_{i,j}$

is present only when the link is not excitatory (4). With $z \equiv -1$ one sets the inverse unit of time.

- Winners-take-all network
  Eqs. (1) and (2) describe, in the absence of a coupling to the reservoir via $f_{z/w}(\varphi)$, a competitive winners-take-all neural network with clique encoding. The system relaxes towards the next attractor made up of a clique of Z sites $(p_1, \ldots, p_Z)$ connected via excitatory $w_{p_i,p_j} > 0$ $(i, j = 1, .., Z)$.
- Reservoir functions
  The reservoir functions $f_{z/w}(\varphi) \in [0,1]$ govern the interaction in between the activity levels x_i and the reservoir levels φ_i. They may be chosen as washed out step functions of sigmoidal form with a suitable width Γ_φ and inflection points $\varphi_c^{(w/z)}$, see Fig. 2.
- Reservoir dynamics
  The reservoir levels of the winning clique depletes slowly, see Eq. (3) and Fig. 3, and recovers only once the activity level x_i of a given site has dropped below x_c. The factor $(1 - x_i/x_c)$ occurring in the reservoir growth process, see the r.h.s. of (3), serves for a stabilization of the transition between subsequent memory states [13]
- Separation of time scales
  A separation of time scales is obtained when the $\Gamma_\varphi^\pm$ are much smaller than the average strength of an excitatory link, $\bar{w}$, leading to transient-state dynamics. Once the reservoir of a winning clique is depleted, it looses, via $f_z(\varphi)$, its ability to suppress other sites and the mutual intra-clique excitation is suppressed via $f_w(\varphi)$.

In Fig. 3 the transient-state dynamics resulting from Eqs. (1-4), in the absence of any sensory signal, is illustrated. When the growth/depletion rates $\Gamma_\varphi^\pm \to 0$ are very small, the individual cliques turn into stable attractors. The possibility to regulate the 'speed' of the associative thought process arbitrarily by setting the $\Gamma_\varphi^\pm$ is important for applications. For a working cognitive system it is enough if the transient states are just stable for a certain minimal period, anything longer just would be a 'waste of time'.

Cycles. The system in Fig. 3 is very small and the associative thought process soon settles into a cycle, since there are no incoming sensory signals in the simulation of Fig. 3. For networks containing a somewhat larger number of sites, the number of attractors can be however very large and such the resulting cycle length. We performed simulations for a 100-site network, to give an example, containing 713 clique-encoded memories. We found no cyclic behavior even for thought processes with up to 4400 transient states. For a working cognitive system prolonged periods without sensory signals will be anyhow rare events and it will be unlikely that the system will settle into a stable cycle of memories.

Dual functionalities for memories. The network discussed here is a dense and homogeneous associative network (dHAN). It is homogeneous since memories have dual functionalities:

- Memories are the transient states of the associative thought process.

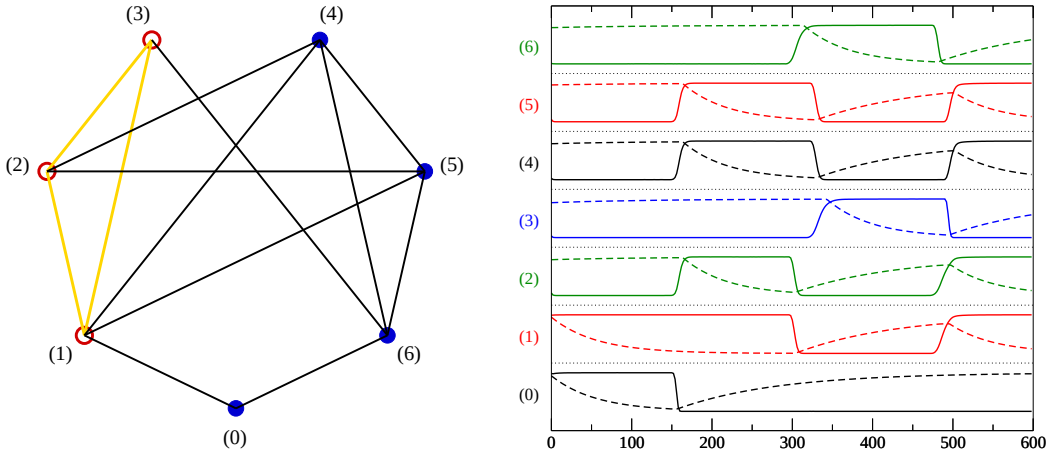

FIGURE 3. Left: A 7-site network, shown are links with $w_{i,j} > 0$, containing six cliques, (0,1), (0,6), (3,6), (1,2,3) (which is highlighted), (4,5,6) and (1,2,4,5).
Right: The activities $x_i(t)$ (solid lines) and the respective reservoirs $\varphi_i(t)$ (dashed lines) for the transient-state dynamics $(0,1) \to (1,2,4,5) \to (3,6) \to (1,2,4,5)$.

- Memories define the associative overlaps, between two subsequent transient states.

Recognition. Any sensory stimulus arriving to the dHAN needs to compete with the ongoing intrinsic dynamics to make an impact. If the sensory signal is not strong enough, it cannot deviate the autonomous thought process. This feature results in an intrinsic recognition property of the dHAN: A background of noise will not influence the transient state dynamics.

AUTONOMOUS ONLINE LEARNING

An external stimulus, $\{b_i^{(ext)}(t)\}$, influences the activities $x_i(t)$ of the respective neural centers. This corresponds to a change of the respective growth rates r_i,

$$r_i \;\to\; r_i + f_w(\varphi_i)\, b_i^{(ext)}(t)\;, \tag{5}$$

compare Eq. (2), where $f_w(\varphi_i)$ is an appropriate coupling function, depending on the local reservoir level φ_i. The task is then to formulate principles which let the dHAN learn and store on-thy-fly patterns found in the stimuli $b_i(t)$.

Short- and long-term synaptic plasticities

There are two fundamental considerations for the choice of synaptic plasticities adequate for the dHAN.

- Learning is a very slow process without a short-term memory. Training patterns need to be presented to the network over and over again until substantial synaptic

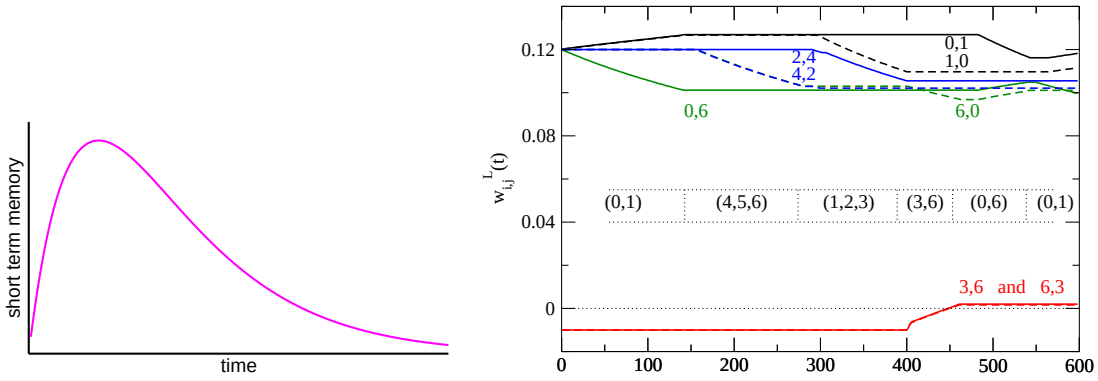

FIGURE 4. Left: Typical activation pattern of the short-term plasticities of an excitatory link (short-term memory).
Right: The time evolution of the long-term memory, for some selected links $w^L_{i,j}$ and the network illustrated in Fig. 3, without the link (3,6). The transient states are $(0,1) \to (4,5,6) \to (1,2,3) \to (3,6) \to (0,6) \to (0,1)$. An external stimulus at sites (3) and (6) acts for $t \in [400,410]$ with strength $b^{(stim)} = 3.6$. The stimulus pattern (3,6) has been learned by the system, as the $w_{3,6}$ and $w_{6,3}$ turned positive during the learning-interval $\approx [400,460]$. The learning interval is substantially longer than the bare stimulus length due to the activation of the short-term memory.

changes are induced [10]. A short-term memory can speed-up the learning process substantially as it stabilizes external patterns and hence gives the system time to consolidate long-term synaptic plasticity.

- Systems using sparse coding are based on a strong inhibitory background, the average inhibitory link-strength $|z|$ is substantially larger than the average excitatory link strength $\bar{w}$,

$$|z| \gg \bar{w} .$$

It is then clear that gradual learning affects dominantly the excitatory links: Small changes of large parameters do not lead to new transient attractors, nor do they influence the cognitive dynamics substantially.

We then have

$$w_{ij} = w_{ij}(t) = w^S_{ij}(t) + w^L_{ij}(t) , \tag{6}$$

where $w^{S/L}_{ij}$ correspond to the short/long-term synaptic plasticities.

Negative baseline. Eq. (4), $z_{ij} = -|z|\,\Theta(-w_{ij})$, states that the inhibitory link-strength is either zero or $-|z|$, but is not changed directly during learning, in accordance to (6). When a $w_{i,j}$ is slightly negative, as default (compare Fig. 2), the corresponding total link strength is inhibitory. When $w_{i,j}$ acquires, during learning, a positive value, the corresponding total link strength becomes excitatory.

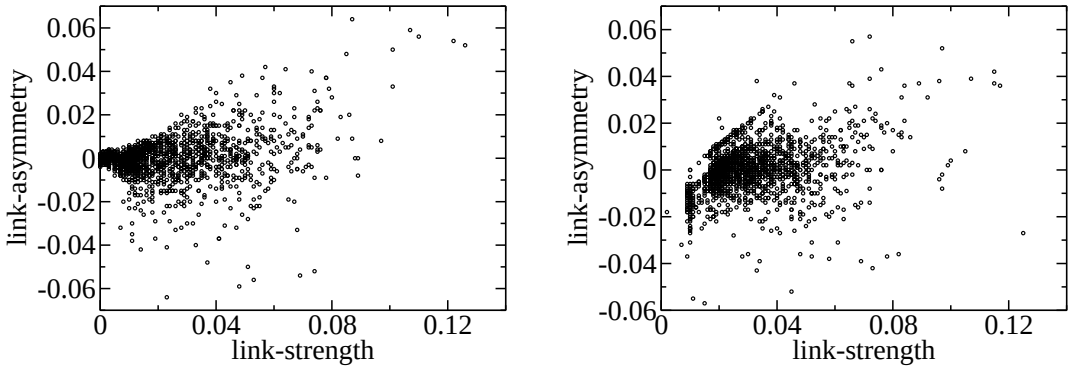

FIGURE 5. The link-asymmetry $w^L_{ij} - w^L_{ji}$ for the positive w^L_{ij} for a 100-site network with 713 cliques at time $t = 500\,000$, corresponding to circa 4500 transient states.
Left: After learning from scratch, learning was finished at $t \approx 50\,000$.
Right: Starting with $w_{i,j} \to 0.12$ for all links belonging to one or more cliques.

Short-term memory dynamics

It is reasonable to have a maximal possible value $W_S^{(max)}$ for the short-term synaptic plasticities. The appropriate Hebbian-type autonomous learning rule is then

$$\begin{aligned} \dot{w}^S_{ij}(t) &= \Gamma^+_S \left(W_S^{(max)} - w^S_{ij} \right) f_z(\varphi_i)\, f_z(\varphi_j)\, \Theta(x_i - x_c)\Theta(x_j - x_c) \\ &- \Gamma^-_S\, w^S_{ij}\ . \end{aligned} \tag{7}$$

It increases rapidly when both the pre- and the post-synaptic centers are active, it decays to zero otherwise, see Fig. 4. The coupling functions $f_z(\varphi)$ preempt prolonged self-activation of the short-term memory. When the pre- and the post-synaptic centers are active long enough to deplete their respective reservoir levels, the short-term memory is shut-off via $f_z(\varphi)$, compare Fig. 2.

Long-term memory dynamics

Dynamical systems retain normally their functionalities only when they keep their dynamical properties in certain regimes. They need to regulate their own working point. This is a long-term affair, it involves time-averaged quantities, and it is therefore a job for the long-term synaptic plasticities, w^L_{ij}.

Effective incoming synaptic strength. The average magnitude of the growth rates r_i, see Eq. (2), determine the time scales of the autonomous dynamics and such the working point. The $r_i(t)$ are however quite strongly time dependent. The effective incoming synaptic signal

$$\tilde{r}_i = \sum_j \Big[w_{i,j} x_j + z_{i,j} x_j\, f_z(\varphi_j) \Big]\ ,$$

which is independent of the post-synaptic reservoir, φ_i, is a more convenient control parameter. The working point of the cognitive system is optimal when the effective incoming signal is, on the average, of comparable magnitude $t^{(opt)}$ for all sites,

$$\tilde{r}_i \;\rightarrow\; r^{(opt)} \;. \tag{8}$$

Eq. (8) is an implementation of the principle of homeostatic self-regulation.

The long-term memory has two tasks: To encode the stimulus patterns and to keep the working point of the dynamical system in its desired range. Both tasks can be achieved by a single local learning rule,

$$\begin{aligned} \dot{w}^L_{ij}(t) \;&=\; \Gamma^{(opt)}_L \Delta\tilde{r}_i \Big[\Big(w^L_{ij} - W^{(min)}_L \Big)\, \Theta(-\Delta\tilde{r}_i) + \Theta(\Delta\tilde{r}_i) \Big] \\ &\quad \cdot \Theta(x_i - x_c)\, \Theta(x_j - x_c), \qquad\qquad \Delta\tilde{r}_i \;=\; r^{(opt)} - \tilde{r}_i \;. \end{aligned} \tag{9}$$

Some comments:

- Hebbian learning
  The learning rule is local and of Hebbian type [10]. Learning occurs only when the pre- and the post-synaptic neuron are active. Weak forgetting, i.e. the decay of seldom used links is not present in (9), but could be added to it.
- Synaptic competition
  When the incoming signal is weak/strong, relative to the optimal value $r^{(opt)}$, the active links are reinforced/weakened, with $W^{(min)}_L$ being the minimal value for the w_{ij}. The baseline $W^{(min)}_L$ is slightly negative, compare Figs. 2 and 4.
  The Hebbian-type learning then takes place in the form of a competition between incoming synapses - frequently active incoming links will gain strength, on the average, on the expense of rarely used links.
- Fast learning of new patterns
  In Fig. 4 the time evolution of some selected w_{ij} from a simulation is presented. A simple input-pattern is learned by the network. In this simulation the learning parameter $\Gamma^{(opt)}_L$ has been set to a quite large value such that the learning occures in one step (fast learning).
- Suppression of runaway synaptic growth
  The link-dynamics (9) suppresses synaptic runaway-growth, a general problem common to adaptive, continuously active neural networks. It has been shown that similar rules for discrete neural networks optimize the overall storage capacity [14].
- Long-term dynamical stability
  In Fig. 5 the results for the long-term link matrices are presented for a 100-site network with 713 stored memories and for two simulations.
  - In the first simulation all excitatory links were set by hand right at the start to 0.12 [13]. The working-point optimization inherent in Eq. (9) then leads to a differentiation for the link-strengths during self-generated associative thought process, generating a total of about 4500 transient states.

– In the second simulation all excitatory links were learned on-the-fly, via Eqs. (7) and (9), from patterns presented to the network during $t \in [0, 50\,000]$. Afterwards the dynamics was 100% self-generated.

The resulting final link distributions are similar. This result indicates that self-sustained associative thought processes lead to stable long-term link distribution and such to stable cognitive dynamics. The system is self-adapting.

CONCLUSIONS

We have pointed out the importance of studying neural networks layouts compatible with the requirements for autonomously operating cognitive systems. We have formulated a set of basic requirements and discussed an implementation for a network capable to learn and store autonomously environmental data as they occur in the sensory stimuli.

We have pointed out (i) that fast online-learning is possible when a short term memory complements the usual long-term synaptic plasticities needed for pattern storage, (ii) that the working point of the self-sustained dynamics can be regulated homeostatically during the learning process and (iii) that clique-encoding allows at the same time for the generation of associative thought processes and for a very high storage capacity.

REFERENCES

1. J. Fiser, C. Chiu, and M. Weliky, *Nature*, **431**, 573–578 (2004).
2. G.M. Edelman, *Proc. Natl. Acad. Sci.* **100**, 5520–5524 (2003).
3. S. Dehaene, and L. Naccache, *Cognition* **79**, 1–37 (2003).
4. F.C. Crick, and C. Koch, *Nature Neurosci.* **6**, 119–126 (2003).
5. D.L. Ringach, *Nature* **425**, 912–913 (2003).
6. T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli, *Nature* **425**, 954–956 (2003).
7. M. Abeles *et al.*, *Proc. Natl. Acad. Sci. USA* **92**, 8616–8620 (1995).
8. D.L. Nelson, C.L. McEvoy, and T.A. Schreiber, "The University of South Florida word association, rhyme, and word fragment norms" (1998); URL http://www.usf.edu/FreeAssociation/.
9. H. Liu, and P. Singh, *BT Technology Journal* **22**, 211–226 (2004); URL http://web.media.mit.edu∼hugo/conceptnet/.
10. M.A. Arbib, *The Handbook of Brain Theory and Neural Networks*, MIT Press (2002).
11. F. Gobet, *et al.*, *Trends in Cogn. Sci.* **5**, 236–243 (2001).
12. L. Boucher, and Z. Dienes, *Cogn. Sci.* **27** 807–842 (2003).
13. C. Gros, "Self-Sustained Thought Processes in a Dense Associative Network", in *Proceedings of the 28th Annual German Conference on Artificial Intelligence (KI 2005)*, edited by U. Furbach, Springer Lecture Notes in Artificial Intelligence **3698**, 375-388 (2005); URL http://arxiv.org/abs/q-bio.NC/0508032/.
14. G. Chechik, I. Meilijson, and E. Ruppin, *Neural Computation* **13**, 817–840 (2001).