

# Adaptive Thresholds for Layered Neural Networks with Synaptic Noise

D. Bollé and R. Heylen

Institute for Theoretical Physics  
Katholieke Universiteit Leuven, Celestijnenlaan 200 D  
B-3001, Leuven, Belgium

**Abstract.** The inclusion of a macroscopic adaptive threshold is studied for the retrieval dynamics of layered feedforward neural network models with synaptic noise. It is shown that if the threshold is chosen appropriately as a function of the cross-talk noise and of the activity of the stored patterns, adapting itself automatically in the course of the recall process, an autonomous functioning of the network is guaranteed. This self-control mechanism considerably improves the quality of retrieval, in particular the storage capacity, the basins of attraction and the mutual information content.

## 1 Introduction

As is common knowledge by now, layered feedforward neural network models are the workhorses in many practical applications in several areas of research and, therefore, any new insight in their capabilities and limitations should thus be welcome. In view of the fact that in many of these applications, e.g., pattern recognition in general, information is mostly encoded by a small fraction of bits and that also in neurophysiological studies the activity level of real neurons is found to be low, any reasonable network model has to allow variable activity of the neurons. The limit of low activity, i.e., sparse coding is then especially interesting. Indeed, sparsely coded models have a very large storage capacity behaving as  $1/(a \ln a)$  for small  $a$ , where  $a$  is the activity (see, e.g., [1,2,3,4] and references therein). However, for low activity the basins of attraction might become very small and the information content in a single pattern is reduced [4]. Therefore, the necessity of a control of the activity of the neurons has been emphasized such that the latter stays the same as the activity of the stored patterns during the recall process. This has led to several discussions imposing external constraints on the dynamics of the network. However, the enforcement of such a constraint at every time step destroys part of the autonomous functioning of the network, i.e., a functioning that has to be independent precisely from such external constraints or control mechanisms. To solve this problem, quite recently a self-control mechanism has been introduced in the dynamics of networks for so-called diluted architectures [5]. This self-control mechanism introduces a time-dependent threshold in the transfer function [5,6]. It is determined as a function of both the cross-talk noise and the activity of the stored patterns in the network,

and adapts itself in the course of the recall process. It furthermore allows to reach optimal retrieval performance both in the absence and in the presence of synaptic noise [5,6,7,8]. These diluted architectures contain no common ancestors nodes, in contrast with feedforward architectures. It has then been shown that a similar mechanism can be introduced successfully for layered feedforward architectures but, without synaptic noise [9].

The purpose of the present contribution is to generalise this self-control mechanism for layered architectures when synaptic noise is allowed, and to show that it leads to a substantial improvement of the quality of retrieval, in particular the storage capacity, the basins of attraction and the mutual information content.

## 2 The Model

Consider a neural network composed of binary neurons arranged in layers, each layer containing  $N$  neurons. A neuron can take values  $\sigma_i(t) \in \{0, 1\}$  where  $t = 1, \dots, L$  is the layer index and  $i = 1, \dots, N$  labels the neurons. Each neuron on layer  $t$  is unidirectionally connected to all neurons on layer  $t + 1$ . We want to memorize  $p$  patterns  $\{\xi_i^\mu(t)\}, i = 1, \dots, N, \mu = 1, \dots, p$  on each layer  $t$ , taking the values  $\{0, 1\}$ . They are assumed to be independent identically distributed random variables (i.i.d.r.v.) with respect to  $i, \mu$  and  $t$ , determined by the probability distribution:  $p(\xi_i^\mu(t)) = a\delta(\xi_i^\mu(t) - 1) + (1 - a)\delta(\xi_i^\mu(t))$ . From this form we find that the expectation value and the variance of the patterns are given by  $E[\xi_i^\mu(t)] = E[\xi_i^\mu(t)^2] = a$ . Moreover, no statistical correlations occur, in fact for  $\mu \neq \nu$  the covariance vanishes.

The state  $\sigma_i(t + 1)$  of neuron  $i$  on layer  $t + 1$  is determined by the state of the neurons on the previous layer  $t$  according to the stochastic rule

$$P(\sigma_i(t + 1) | \sigma_1(t), \dots, \sigma_N(t)) = \{1 + \exp[2(2\sigma_i(t + 1) - 1)\beta h_i(t)]\}^{-1}. \tag{1}$$

The right hand side is the logistic function. The “temperature”  $T = 1/\beta$  controls the stochasticity of the network dynamics, it measures the synaptic noise level [10]. Given the network state  $\{\sigma_i(t)\}; i = 1, \dots, N$  on layer  $t$ , the so-called “local field”  $h_i(t)$  of neuron  $i$  on the next layer  $t + 1$  is given by

$$h_i(t) = \sum_{j=1}^N J_{ij}(t)(\sigma_j(t) - a) - \theta(t) \tag{2}$$

with  $\theta(t)$  the threshold to be specified later. The couplings  $J_{ij}(t)$  are the synaptic strengths of the interaction between neuron  $j$  on layer  $t$  and neuron  $i$  on layer  $t + 1$ . They depend on the stored patterns at different layers according to the covariance rule

$$J_{ij}(t) = \frac{1}{Na(1 - a)} \sum_{\mu=1}^p (\xi_i^\mu(t + 1) - a)(\xi_j^\mu(t) - a). \tag{3}$$

These couplings then permit to store sets of patterns to be retrieved by the layered network.

The dynamics of this network is defined as follows (see [11]). Initially the first layer (the input) is externally set in some fixed state. In response to that, all neurons of the second layer update synchronously at the next time step, according to the stochastic rule (1), and so on.

At this point we remark that the couplings (3) are of infinite range (each neuron interacts with infinitely many others) such that our model allows a so-called mean-field theory approximation. This essentially means that we focus on the dynamics of a single neuron while replacing all the other neurons by an average background local field. In other words, no fluctuations of the other neurons are taken into account. In our case this approximation becomes exact because, crudely speaking,  $h_i(t)$  is the sum of very many terms and a central limit theorem can be applied [10].

It is standard knowledge by now that mean-field theory dynamics can be solved exactly for these layered architectures (e.g., [11,12]). By exact analytic treatment we mean that, given the state of the first layer as initial state, the state on layer  $t$  that results from the dynamics is predicted by recursion formulas. This is essentially due to the fact that the representations of the patterns on different layers are chosen independently. Hence, the big advantage is that this will allow us to determine the effects from self-control in an exact way.

The relevant parameters describing the solution of this dynamics are the *main overlap* of the state of the network and the  $\mu$ -th pattern, and the *neural activity* of the neurons

$$M^\mu(t) = \frac{1}{Na(1-a)} \sum_{i=1}^N (\xi_i^\mu(t) - a)(\sigma_i(t) - a), \quad q(t) = \frac{1}{N} \sum_{i=1}^N \sigma_i(t). \quad (4)$$

In order to measure the retrieval quality of the recall process, we use the mutual information function [5,6,13,14]. In general, it measures the average amount of information that can be received by the user by observing the signal at the output of a channel [15,16]. For the recall process of stored patterns that we are discussing here, at each layer the process can be regarded as a channel with input  $\xi_i^\mu(t)$  and output  $\sigma_i(t)$  such that this mutual information function can be defined as [5,15]

$$I(\sigma_i(t); \xi_i^\mu(t)) = S(\sigma_i(t)) - \langle S(\sigma_i(t)|\xi_i^\mu(t)) \rangle_{\xi^\mu(t)} \quad (5)$$

where  $S(\sigma_i(t))$  and  $S(\sigma_i(t)|\xi_i^\mu(t))$  are the entropy and the conditional entropy of the output, respectively

$$S(\sigma_i(t)) = - \sum_{\sigma_i} p(\sigma_i(t)) \ln[p(\sigma_i(t))] \quad (6)$$

$$S(\sigma_i(t)|\xi_i^\mu(t)) = - \sum_{\sigma_i} p(\sigma_i(t)|\xi_i^\mu(t)) \ln[p(\sigma_i(t)|\xi_i^\mu(t))] . \quad (7)$$

These information entropies are peculiar to the probability distributions of the output. The quantity  $p(\sigma_i(t))$  denotes the probability distribution for the neurons at layer  $t$  and  $p(\sigma_i(t)|\xi_i^\mu(t))$  indicates the conditional probability that the

$i$ -th neuron is in a state  $\sigma_i(t)$  at layer  $t$  given that the  $i$ -th site of the pattern to be retrieved is  $\xi_i^\mu(t)$ . Hereby, we have assumed that the conditional probability of all the neurons factorizes, i.e.,  $p(\{\sigma_i(t)\}|\{\xi_i(t)\}) = \prod_j p(\sigma_j(t)|\xi_j(t))$ , which is a consequence of the mean-field theory character of our model explained above. We remark that a similar factorization has also been used in Schwenker et al. [17].

The calculation of the different terms in the expression (5) proceeds as follows. Because of the mean-field character of our model the following formula hold for every neuron  $i$  on each layer  $t$ . Formally writing (forgetting about the pattern index  $\mu$ )  $\langle O \rangle \equiv \langle \langle O \rangle_{\sigma|\xi} \rangle_{\xi} = \sum_{\xi} p(\xi) \sum_{\sigma} p(\sigma|\xi) O$  for an arbitrary quantity  $O$  the conditional probability can be obtained in a rather straightforward way by using the complete knowledge about the system:  $\langle \xi \rangle = a$ ,  $\langle \sigma \rangle = q$ ,  $\langle (\sigma - a)(\xi - a) \rangle = M$ ,  $\langle 1 \rangle = 1$ .

The result reads

$$p(\sigma|\xi) = [\gamma_0 \xi + (\gamma_1 - \gamma_0) \xi] \delta(\sigma - 1) + [1 - \gamma_0 - (\gamma_1 - \gamma_0) \xi] \delta(\sigma) \tag{8}$$

where  $\gamma_0 = q - aM$  and  $\gamma_1 = (1 - a)M + q$ , and where the  $M$  and  $q$  are precisely the relevant parameters (4) for large  $N$ . Using the probability distribution of the patterns we obtain

$$p(\sigma) = q \delta(\sigma - 1) + (1 - q) \delta(\sigma) . \tag{9}$$

Hence the entropy (6) and the conditional entropy (7) become

$$S(\sigma) = -q \ln q - (1 - q) \ln(1 - q) \tag{10}$$

$$S(\sigma|\xi) = -[\gamma_0 + (\gamma_1 - \gamma_0) \xi] \ln[\gamma_0 + (\gamma_1 - \gamma_0) \xi] - [1 - \gamma_0 - (\gamma_1 - \gamma_0) \xi] \ln[1 - \gamma_0 - (\gamma_1 - \gamma_0) \xi] . \tag{11}$$

By averaging the conditional entropy over the pattern  $\xi$  we finally get for the mutual information function (5) for the layered model

$$I(\sigma; \xi) = -q \ln q - (1 - q) \ln(1 - q) + a[\gamma_1 \ln \gamma_1 + (1 - \gamma_1) \ln(1 - \gamma_1)] + (1 - a)[\gamma_0 \ln \gamma_0 + (1 - \gamma_0) \ln(1 - \gamma_0)] . \tag{12}$$

### 3 Adaptive Thresholds

It is standard knowledge (e.g., [11]) that the synchronous dynamics for layered architectures can be solved exactly following the method based upon a signal-to-noise analysis of the local field (2) (e.g., [4,12,18,19] and references therein). Without loss of generality we focus on the recall of one pattern, say  $\mu = 1$ , meaning that only  $M^1(t)$  is macroscopic, i.e., of order 1 and the rest of the patterns causes a cross-talk noise at each step of the dynamics.

We suppose that the initial state of the network model  $\{\sigma_i(1)\}$  is a collection of i.i.d.r.v. with average and variance given by  $E[\sigma_i(1)] = E[(\sigma_i(1))^2] = q_0$ . We furthermore assume that this state is correlated with only one stored pattern, say pattern  $\mu = 1$ , such that  $\text{Cov}(\xi_i^\mu(1), \sigma_i(1)) = \delta_{\mu,1} M_0^1 a(1 - a)$ .

Then the full recall proces is described by [11,12]

$$M^1(t + 1) = \frac{1}{2} \left\{ \int \mathcal{D}x \tanh \left[ \beta((1 - a)M^1(t) - \theta(t) + \sqrt{\alpha D(t)} x) \right] + \int \mathcal{D}x \tanh \left[ \beta(-aM^1(t) - \theta(t) + \sqrt{\alpha D(t)} x) \right] \right\} \tag{13}$$

$$q(t + 1) = aM^1(t + 1) + \frac{1}{2} \left\{ 1 + \int \mathcal{D}x \tanh \left[ \beta(-aM^1(t) - \theta(t) + \sqrt{\alpha D(t)} x) \right] \right\} \tag{14}$$

$$D(t + 1) = Q(t + 1) + \frac{\beta}{2} \left\{ 1 - a \int \mathcal{D}x \tanh^2 \beta \left[ (1 - a)M^1(t) - \theta(t) + \sqrt{\alpha D(t)} x \right] - (1 - a) \int \mathcal{D}x \tanh^2 \beta \left[ -aM^1(t) - \theta(t) + \sqrt{\alpha D(t)} x \right] \right\}^2 D(t) \tag{15}$$

where  $\alpha = p/N$ ,  $\mathcal{D}x$  is the Gaussian measure  $\mathcal{D}x = dx(2\pi)^{-1/2} \exp(-x^2/2)$ , where  $Q(t) = [(1 - 2a)q(t) + a^2]$  and where  $D(t)$  contains the influence of the cross-talk noise caused by the patterns  $\mu > 1$ . As mentioned before,  $\theta(t)$  is an adaptive threshold that has to be chosen.

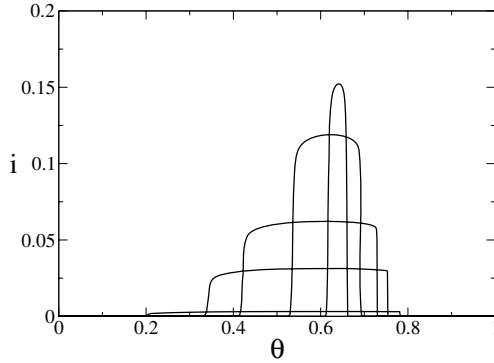
In the sequel we discuss two different choices and both will be compared for networks with synaptic noise and various activities. Of course, it is known that the quality of the recall process is influenced by the cross-talk noise. An idea is then to introduce a threshold that adapts itself autonomously in the course of the recall process and that counters, at each layer, the cross-talk noise. This is the self-control method proposed in [5]. This has been studied for layered neural network models without synaptic noise, i.e., at  $T = 0$ , where the rule (1) reduces to the deterministic form  $\sigma_i(t + 1) = \Theta(h_i(t))$  with  $\Theta(x)$  the Heaviside function taking the value  $\{0, 1\}$ . For sparsely coded models, meaning that the pattern activity  $a$  is very small and tends to zero for  $N$  large, it has been found [9] that

$$\theta(t)_{sc} = c(a)\sqrt{\alpha D(t)}, \quad c(a) = \sqrt{-2 \ln a} \tag{16}$$

makes the second term on the r.h.s of Eq.(14) at  $T = 0$ , asymptotically vanish faster than  $a$  such that  $q \sim a$ . It turns out that the inclusion of this self-control threshold considerably improves the quality of retrieval, in particular the storage capacity, the basins of attraction and the information content.

The second approach chooses a threshold by maximizing the information content,  $i = \alpha I$  of the network (recall Eq. (12)). This function depends on  $M^1(t)$ ,  $q(t)$ ,  $a$ ,  $\alpha$  and  $\beta$ . The evolution of  $M^1(t)$  and of  $q(t)$  (13), (14) depends on the specific choice of the threshold through the local field (2). We consider a layer independent threshold  $\theta(t) = \theta$  and calculate the value of (12) for fixed  $a$ ,  $\alpha$ ,  $M_0^1$ ,  $q_0$  and  $\beta$ . The optimal threshold,  $\theta = \theta_{opt}$ , is then the one for which the mutual information function is maximal. The latter is non-trivial because it is even rather difficult, especially in the limit of sparse coding, to choose a threshold interval by hand such that  $i$  is non-zero. The computational cost will thus be

larger compared to the one of the self-control approach. To illustrate this we plot in Figure 1 the information content  $i$  as a function of  $\theta$  without self-control or a priori optimization, for  $a = 0.005$  and different values of  $\alpha$ . For every value of



**Fig. 1.** The information  $i = \alpha I$  as a function of  $\theta$  for  $a = 0.005$ ,  $T = 0.1$  and several values of the load parameter  $\alpha = 0.1, 1, 2, 4, 6$  (bottom to top)

$\alpha$ , below its critical value, there is a range for the threshold where the information content is different from zero and hence, retrieval is possible. This retrieval range becomes very small when the storage capacity approaches its critical value  $\alpha_c = 6.4$ .

Concerning then the self-control approach, the next problem to be posed in analogy with the case without synaptic noise is the following one. Can one determine a form for the threshold  $\theta(t)$  such that the integral in the second term on the r.h.s of Eq.(14) at  $T \neq 0$  vanishes asymptotically faster than  $a$ ?

In contrast with the case at zero temperature where due to the simple form of the transfer function, this threshold could be determined analytically (recall Eq. (16)), a detailed study of the asymptotics of the integral in Eq. (14) gives no satisfactory analytic solution. Therefore, we have designed a systematic numerical procedure through the following steps:

- Choose a small value for the activity  $a'$ .
- Determine through numerical integration the threshold  $\theta'$  such that

$$\int_{-\infty}^{\infty} \frac{dx e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \Theta(x - \theta) \leq a' \quad \text{for } \theta > \theta' \tag{17}$$

for different values of the variance  $\sigma^2 = \alpha D(t)$ .

- Determine as a function of  $T = 1/\beta$ , the value for  $\theta'_T$  such that

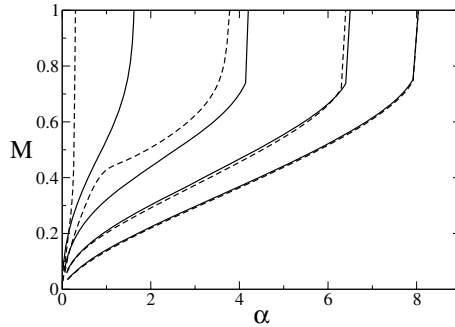
$$\int_{-\infty}^{\infty} \frac{dx e^{-y^2/\sigma^2}}{2\sigma\sqrt{2\pi}} [1 + \tanh[\beta(x - \theta)]] \leq a' \quad \text{for } \theta > \theta' + \theta'_T. \tag{18}$$

The second step leads precisely to a threshold having the form of Eq. (16). The third step determining the temperature-dependent part  $\theta'_T$  leads to the final proposal

$$\theta_t(a, T) = \sqrt{-2 \ln(a) \alpha D(t)} - \frac{1}{2} \ln(a) T^2. \tag{19}$$

This dynamical threshold is again a macroscopic parameter, thus no average must be taken over the microscopic random variables at each step  $t$  of the recall process.

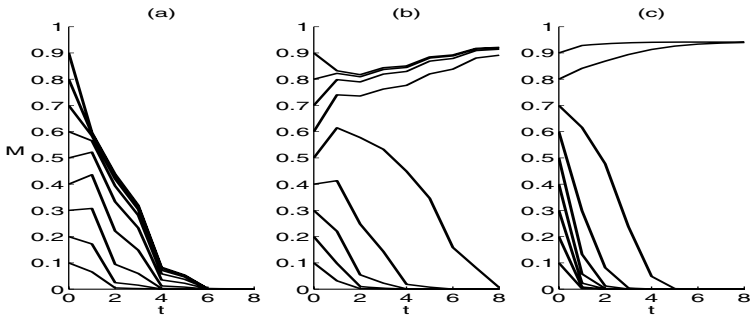
We have solved these self-controlled dynamics, Eqs.(13)-(15) and (19), for our model with synaptic noise, in the limit of sparse coding, numerically. In particular, we have studied in detail the influence of the  $T$ -dependent part of the threshold. Of course, we are only interested in the retrieval solutions with  $M > 0$  (we forget about the index 1) and carrying a non-zero information  $i = \alpha I$ . The important features of the solution are illustrated, for a typical value of  $a$  in Figures 2-4. In Figure 2 we show the basin of attraction for the whole retrieval



**Fig. 2.** The basin of attraction as a function of  $\alpha$  for  $a = 0.005$  and  $T = 0.2, 0.15, 0.1, 0.05$  (from left to right) with (full lines) and without (dashed lines) the  $T$ -dependent part in the threshold (19)

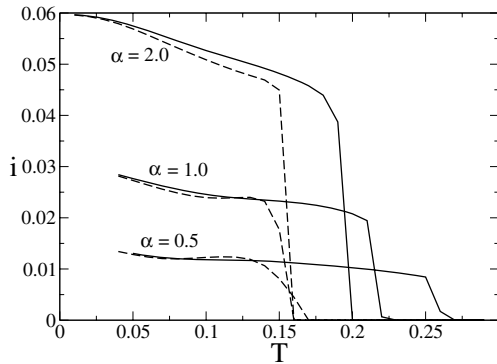
phase for the model with threshold (16) (dashed curves) compared to the model with the noise-dependent threshold (19) (full curves). We see that there is no clear improvement for low  $T$  but there is a substantial one for higher  $T$ . Even near the border of critical storage the results are still improved such that also the storage capacity itself is larger.

This is further illustrated in Figure 3 where we compare the evolution of the retrieval overlap  $M(t)$  starting from several initial values,  $M_0$ , for the model with (Figure 3 (a)) and without (Figure 3 (b)) the  $T$ -correction in the threshold and for the optimal threshold model (Figure 3 (c)). Here this temperature correction is absolutely crucial to guarantee retrieval, i.e.,  $M \approx 1$ . It really makes the difference between retrieval and non-retrieval in the model. Furthermore, the model with the self-control threshold with noise-correction has even a wider basin of attraction than the model with optimal threshold.



**Fig. 3.** The evolution of the main overlap  $M(t)$  for several initial values  $M_0$  with  $T = 0.2$ ,  $q_0 = a = 0.005$ ,  $\alpha = 1$  for the self-control model (19) without (a) and with  $T$ -dependent part (b) and for the optimal threshold model (c)

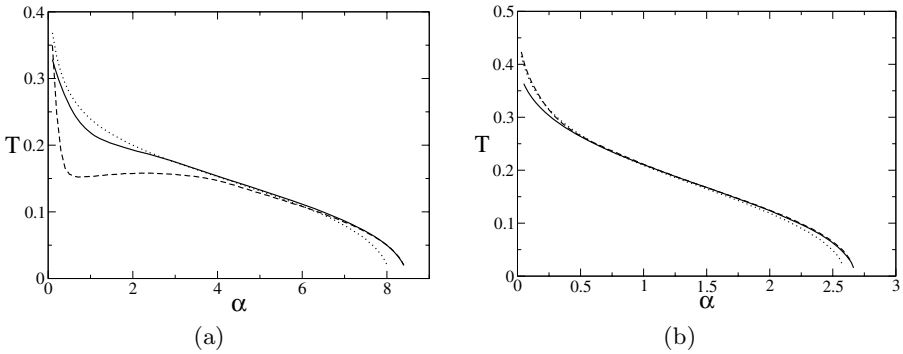
In Figure 4 we plot the information content  $i$  as a function of the temperature for the self-control dynamics with the threshold (19) (full curves), respectively (16) (dashed curves). We see that a substantial improvement of the information content is obtained.



**Fig. 4.** The information content  $i = \alpha I$  as a function of  $T$  for several values of the loading  $\alpha$  and  $a = 0.005$  with (full lines) and without (dashed lines) the  $T$ -correction in the threshold

Finally we show in Figure 5 a  $T - \alpha$  plot for  $a = 0.005$  (a) and  $a = 0.02$  (b) with (full line) and without (dashed line) noise-correction in the self-control threshold and with optimal threshold (dotted line). These lines indicate two phases of the layered model: below the lines our model allows recall, above the lines it does not. For  $a = 0.005$  we see that the  $T$ -dependent term in the self-control threshold leads to a big improvement in the region for large noise and small loading and in the region of critical loading. For  $a = 0.02$  the results for the self-control threshold with and without noise-correction and those for the optimal





**Fig. 5.** Phases in the  $T - \alpha$  plane for  $a = 0.005$  (a) and  $a = 0.02$  (b) with (full line) and without (dashed line) the temperature correction in the self-control threshold and with optimal threshold (dotted line)

thresholds almost coincide, but we recall that the calculation with self-control is autonomously done by the network and less demanding computationally.

## 4 Conclusions

In this work we have studied the inclusion of an adaptive threshold in sparsely coded layered neural networks with synaptic noise. We have presented an analytic form for a self-control threshold, allowing an autonomous functioning of the network, and compared it with an optimal threshold obtained by maximizing the mutual information which has to be calculated externally each time one of the network parameters (activity, loading, temperature) is changed. The consequences of this self-control mechanism on the quality of the recall process have been studied.

We find that the basins of attraction of the retrieval solutions as well as the storage capacity are enlarged. For some activities the self-control threshold even sets the border between retrieval and non-retrieval. This confirms the considerable improvement of the quality of recall by self-control, also for layered network models with synaptic noise.

This allows us to conjecture that self-control might be relevant for other architectures in the presence of synaptic noise, and even for dynamical systems in general, when trying to improve, e.g., basins of attraction .

## Acknowledgment

This work has been supported by the Fund for Scientific Research- Flanders (Belgium).

## References

1. Willshaw D J, Buneman O P, and Longuet-Higgins H C, Nonholographic associative memory, *Nature* **222** (1969) 960.
2. Palm G, On the storage capacity of an associative memory with random distributed storage elements, *Biol. Cyber.* **39** (1981) 125.
3. Gardner E, The space of interactions in neural network models, *J. Phys. A: Math. Gen.* **21** (1988) 257.
4. Okada M, Notions of associative memory and sparse coding, *Neural Networks* **9** (1996) 1429.
5. Dominguez D R C and Bollé D, Self-control in sparsely coded networks, *Phys. Rev. Lett.* **80** (1998) 2961.
6. Bollé D, Dominguez D R C and Amari S, Mutual information of sparsely coded associative memory with self-control and ternary neurons, *Neural Networks* **13**(2000) 455.
7. Bollé D and Heylen R, Self-control dynamics for sparsely coded networks with synaptic noise, in *2004 Proceedings of the IEEE International Joint Conference on Neural Networks*, p.3195
8. Dominguez D R C, Korutcheva E, Theumann W K and Erichsen Jr. R, Flow diagrams of the quadratic neural network, *Lecture Notes in Computer Science*, **2415**, (2002) 129.
9. Bollé D and Massolo G, Thresholds in layered neural networks with variable activity, *J. Phys. A: Math. Gen.* **33** (2000) 2597.
10. Hertz J, Krogh A and Palmer R G, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City (1991).
11. Domany E, Kinzel W and Meir R, Layered Neural Networks, *J.Phys. A: Math. Gen.* **22** (1989) 2081.
12. Bollé D, Multi-state neural networks based upon spin-glasses: a biased overview, in *Advances in Condensed Matter and Statistical Mechanics* eds. Korutcheva E and Cuerno R., Nova Science Publishers, New-York,(2004 p. 321-349.
13. Nadal J-P, Brunel N and Parga N, Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction, *Network: Computation in Neural Systems* **9** (1998) 207.
14. Schultz S and Treves A, Stability of the replica-symmetric solution for the information conveyed by a neural network. *Phys. Rev. E* **57** (1998) 3302.
15. Blahut R E, *Principles and Practice of Information Theory*, Reading, MA: Addison-Wesley (1990).
16. Shannon C E, A mathematical theory for communication, *Bell Systems Technical Journal* **27** (1948) 379.
17. Schwenker F, Sommer F T and Palm G, Iterative retrieval of sparsely coded associative memory patterns, *Neural Networks* **9** (1996) 445.
18. Amari S, Neural theory and association of concept information, *Biol. Cyber.* **26** (1977) 175.
19. Amari S and Maginu K, Statistical neurodynamics of associative memory, *Neural Networks* **1** (1988) 63.