Self-Control in Sparsely Coded Networks

D. R. C. Dominguez and D. Bollé

Instituut voor Theoretische Fysica, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium

(Received 4 November 1997)

A *complete self-control* mechanism is proposed in the dynamics of neural networks through the introduction of a time-dependent threshold, determined in function of both the noise and the pattern activity in the network. Especially for sparsely coded models this mechanism is shown to considerably improve the storage capacity, the basins of attraction, and the mutual information content. [S0031-9007(98)05698-1]

PACS numbers: 87.10.+e, 02.50.–r, 64.60.Cn, 89.70.+c

In the past years, neural network models which can execute complex computations have been proposed using, e.g., statistical mechanics and probabilistic methods. On the one hand, it is suggested that these models might mimic computations by the cerebral cortex and, on the other hand, it is believed that they have a power similar to big assemblies of simple processors or devices [1]. In this development, both from the device oriented and biologically oriented point of view, sparsely coded models [2,3] are of particular interest since it is well known that they have a large storage capacity, which behaves as $1/(a\ln a)$ for a small where a is the pattern activity. However, it is clear that the basins of attraction, e.g., should not become too small because then sparse coding is, in fact, useless.

In this context the necessity of an activity control system has been emphasized, which tries to keep the activity of the network in the retrieval process the same as the one for the memorized patterns [4–6]. This has led to several discussions imposing external constraints on the dynamics (see the references in [3]). Clearly, the enforcement of such a constraint at every time step destroys part of the autonomous functioning of the network.

An important question is then whether the capacity of storage and retrieval with non-negligible basins of attraction can be improved and even be optimized without imposing these external constraints, keeping at the same time the simplicity of the architecture of the network.

In this Letter we answer this question by proposing a *complete self-control* mechanism in the dynamics of neural networks. This is done through the introduction of a time-dependent threshold in the transfer function. This threshold is chosen as a function of the noise in the system and the pattern activity, and adapts itself in the course of the time evolution. The difference with existing results in the literature ([3–6]) precisely lies in this adaptivity property. This immediately solves, e.g., the difficult problem of finding the mostly narrow interval for an optimal threshold such that the basins of attraction of the memorized patterns do not shrink to zero.

We have worked out the case of sparsely coded, diluted models. We find that the storage capacity, the basins of attraction, as well as the mutual information content are improved. These results are shown to be valid also for not so sparse models. Indeed, a similar self-control mechanism should even work in more complicated architectures, e.g., layered and fully connected ones. Furthermore, this idea of self-control might be relevant for dynamical systems in general, when trying to improve the basins of attraction and the convergence times.

Consider a network of N binary neurons. At time t and zero temperature the neurons $\{\sigma_{i,t}\} \in \{0,1\}$, $i = 1,\ldots,N$ are updated in parallel according to the rule

$$\sigma_{i,t+1} = F_{\theta_t}(h_{i,t}), \qquad h_{i,t} = \sum_{j(\neq i)}^{N} J_{ij}(\sigma_{j,t} - a). \quad (1)$$

In general, the input-output relation F_{θ_t} can be a monotonic function with θ_t a time-dependent threshold. In the sequel we restrict ourselves to the step function $F_{\theta_t}(x) = \Theta(x - \theta_t)$. The quantity $h_{i,t}$ is the local field of neuron i at time t and a is the activity of the stored patterns, $\xi_i^\mu \in \{0,1\}$ $\mu = 1,\ldots,p$. The latter are independent identically distributed random variables (IIDRV) with respect to i and μ determined by the probability distribution $p(\xi_i^\mu) = a\delta(\xi_i^\mu - 1) + (1-a)\delta(\xi_i^\mu)$. At this point we remark that the activity can be written as $a = (1-b)/2$ with $-1 < b < 1$ the bias of the patterns as defined, e.g., in [4]. In fact, $\langle \xi_i^\mu \rangle = a$ but no correlations occur, i.e., $\langle \xi_i^\mu \xi_i^\nu \rangle - \langle \xi_i^\mu \rangle \langle \xi_i^\nu \rangle = 0$. We now consider an extremely diluted asymmetric version of this model in which each neuron is connected, on average, with C other neurons. In that case the synaptic couplings J_{ij} are determined by the covariance rule

$$J_{ij} = \frac{C_{ij}}{C\tilde{a}} \sum_{\mu=1}^{p} (\xi_i^\mu - a)(\xi_j^\mu - a), \qquad \tilde{a} \equiv a(1-a). \quad (2)$$

Here the $C_{ij} \in \{0,1\}$ are IIDRV with probability $\Pr\{C_{ij} = 1\} = C/N \ll 1, C > 0$. For $a = 1/2$ and $\theta_t = 0$ we recover the diluted Hopfield model.

The relevant order parameters measuring the quality of retrieval are the overlap of the microscopic state of the network and the μth pattern, and the neural activity

$$m_{N,t}^\mu \equiv \frac{1}{Na} \sum_i \xi_i^\mu \sigma_{i,t}, \qquad q_{N,t} \equiv \frac{1}{N} \sum_i \sigma_{i,t}. \quad (3)$$

The $m_{N,t}^{\mu}$ are normalized within the interval $[0,1]$. They have to be considered over the diluted structure such that the loading α is defined by $p = \alpha C$. The Hamming distance between the state of the neuron and the pattern ξ^{μ} can be written as $d_t^{\mu} = a - 2am_{N,t}^{\mu} + q_{N,t}$.

To fix the ideas and without loss of generality, we take an initial network configuration correlated with only one pattern, say $\mu = 1$, meaning that only $m_{N,t}^{1}$ is macroscopic, i.e., of order $\mathcal{O}(1)$ in the thermodynamic limit $C, N \to \infty$. The rest of the patterns cause a residual noise at each time step of the dynamics. Depending on the architecture of the network this noise might be extremely difficult to treat [7]. A novel idea is then to let the network itself autonomously counter this residual noise at each step of the dynamics, by introducing an adaptive, hence time-dependent, threshold. We propose the general form $\theta_t(x) = c(a)[\mathrm{Var}(\omega_t)]^{1/2}$ with ω_t this residual noise. This self-control mechanism of the network is complete if we find a way to determine $c(a)$.

In order to do so we first write down the evolution equations governing the dynamics. We recall that for the particular model we are considering the parallel dynamics can be solved exactly following the methods involving a signal-to-noise analysis (see, e.g, [8,9]). Such an approach leads to the following equations for the order parameters in the thermodynamic limit $C, N \to \infty$:

$$m_{t+1}^{1} = \langle F_{\theta_t}[(1 - a)M_t^1 + \omega_t]\rangle_{\omega}, \tag{4}$$

$$q_{t+1} = am_{t+1}^{1} + (1 - a)\langle F_{\theta_t}(-aM_t^1 + \omega_t)\rangle_{\omega}, \tag{5}$$

with $M_t^1 = (m_t^1 - q_t)/(1 - a)$, where we have averaged over ξ^1 and where the angular brackets indicate that we still have to average over ω_t which can be written as $\omega_t = [\alpha Q_t]^{1/2}\mathcal{N}(0,1)$ with $Q_t = (1 - 2a)q_t + a^2$ and $\mathcal{N}(0,1)$ a Gaussian random variable with mean zero and variance unity. The m_t^1 and q_t are the thermodynamic limits of (3). The quantity M_t^1 reduces to the overlap of the Hopfield model, again when taking $a = 1/2$ and $\theta_t = 0$. From now on we forget about the superscript 1.

Equations (4) and (5) give a self-controlled dynamics if we can completely specify, *a priori*, the threshold θ_t proposed before. For the present model, θ_t is a macroscopic parameter, thus no average must be done over the microscopic random variables at each time step t. This is different from some existing models, e.g., [10] where a local threshold $\theta_{i,t}$ is taken (to study periodic behavior of patterns). Therefore, we have a mapping with a threshold which changes each time step but no statistical history effects the evolution process. What is left then is to find an optimal form for $c(a)$.

A very intuitive reasoning based on the detailed behavior of Eqs. (4),(5) goes as follows. To have $m \sim 1 - \mathrm{erfc}(n)$ and $q \sim a + \mathrm{erfc}(n)$ with $n > 0$ at a given t such that good retrieval properties, i.e., $\sigma_i = \xi_i$ for most i are realized, we want the following inequalities to be satisfied: $(1 - a)M_t - n[\alpha Q_t]^{1/2} \geq \theta_t$ and $-aM_t + n[\alpha Q_t]^{1/2} \leq \theta_t$. Using the general form for θ_t we obtain that $2n \sim M_t[\alpha Q_t]^{-1/2}$. This leads to $c(a) \sim n(1 - 2a)$. Here we remark that n itself depends on α in the sense that for increasing α it gets more difficult to have good retrieval such that n decreases. But it can still be chosen *a priori*.

In the limit of sparse coding meaning that a is very small and tends to zero for $N \to \infty$, we can present a more refined result for $c(a)$ by rewriting the second term on the right-hand side of Eq. (5) asymptotically as

$$\langle [F_{\theta}[-aM + \omega]\rangle_{\omega} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{aM}{\sqrt{2\alpha Q}} + \frac{c(a)}{\sqrt{2}}\right)\right] \to \frac{e^{-c^2/2}}{c\sqrt{2\pi}}.$$

This term must vanish faster than a so that we obtain $c = [-2\ln(a)]^{1/2}$. Using this and the first inequality written down above we can evaluate the maximal capacity for which some small errors in the retrieval are allowed. The result is $\alpha = \mathcal{O}(|a\ln(a)|^{-1})$, which is of the same order as the critical capacity found for non-self-controlled sparsely coded neural networks [3,6,11–13].

Next, it is known that while the Hamming distance is a good measure for the performance of a uniform network ($a \sim 1/2$), it does not give a complete description of the information content for sparsely coded networks. It cannot distinguish between a situation where most of the wrong neurons ($\sigma_i \neq \xi_i$) are turned off and a situation where these wrong neurons are turned on. This distinction is extremely critical because the inactive neurons carry less information than the active ones. For example, when $\sigma_i = 0$ for all i, $d = a$ and hence vanishes in the sparsely coded limit, while for $\sigma_i = 1$ for all i, $d = 1 - a$ and hence goes to 1. However, in both cases no information is transmitted. To solve this problem we introduce the mutual information content of the network.

The mutual information function ([14]) is a concept in information theory which measures the average amount of information that can be received by the user by observing the signal at the output of a channel. For the problem at hand, i.e., retrieval dynamics of the pattern $\mu = 1$, where each time step is regarded as a channel, it can be defined as (we forget about the time index t)

$$I(\sigma_i;\xi_i) = S(\sigma_i) - \langle S(\sigma_i|\xi_i)\rangle_{\xi_i}, \tag{6}$$

$$S(\sigma_i) \equiv -\sum_{\sigma_i} p(\sigma_i)\ln[p(\sigma_i)], \tag{7}$$

$$S(\sigma_i|\xi_i) \equiv -\sum_{\sigma_i} p(\sigma_i|\xi_i)\ln[p(\sigma_i|\xi_i)]. \tag{8}$$

Here $S(\sigma_i)$ and $S(\sigma_i|\xi_i)$ are the entropy and the conditional entropy of the output, respectively. We note that they are not used, as in thermodynamics, to determine the number of possible realizations of the network giving good retrieval. They have a dynamical content and are peculiar to the probability distribution of the output.

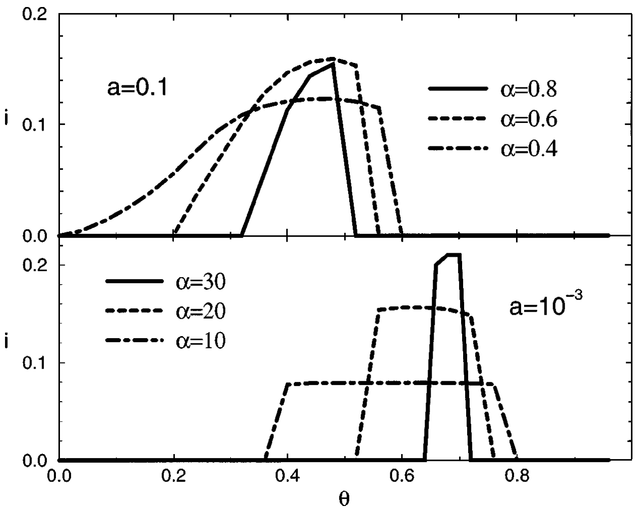

FIG. 1. The information i as a function of θ without self-control for $a = 0.1$ (top) and $a = 0.001$ (bottom) for several values of α.

The term $S(\sigma_i|\xi_i)$ is the equivocation term in the retrieval process. In (8) $p(\sigma_i|\xi_i)$ is the conditional probability that the ith neuron is in a state σ_i at time t, given that the ith site of the pattern being retrieved is ξ_i:

$$p(\sigma|\xi) = [\gamma_0 + (m - \gamma_0)\xi]\delta_{\sigma-1} + [1 - \gamma_0 - (m - \gamma_0)\xi]\delta_\sigma$$
$$\gamma_0 = (q - am)/(1 - a), \qquad (9)$$

where we have assumed that this formula holds for every site index i, and where the m and q are precisely the order parameters (3) for $N \to \infty$. We have also used the normalizations $\sum_\sigma p(\sigma|1) = \sum_\sigma p(\sigma|0) = 1$. Using the probability distribution of the patterns, we obtain

$$p(\sigma) \equiv \sum_\xi p(\xi)p(\sigma|\xi) = q\delta(\sigma - 1) + (1 - q)\delta(\sigma). \qquad (10)$$

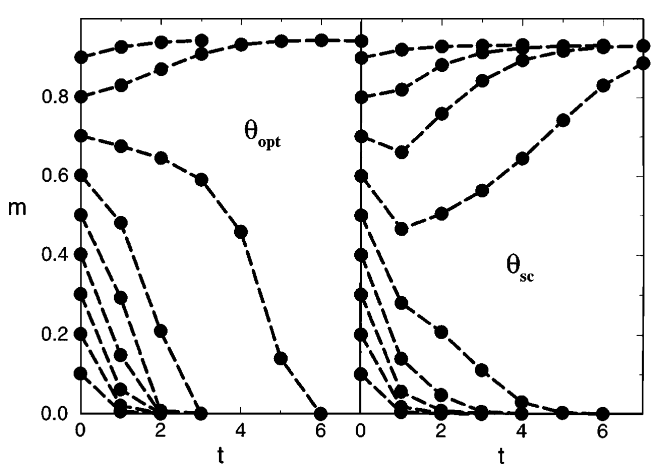

FIG. 2. The evolution of the overlap m_t for several initial values m_0, with $q_0 = 0.01 = a$ and $\alpha = 4$ for the self-control model (right) and the optimal threshold model (left).

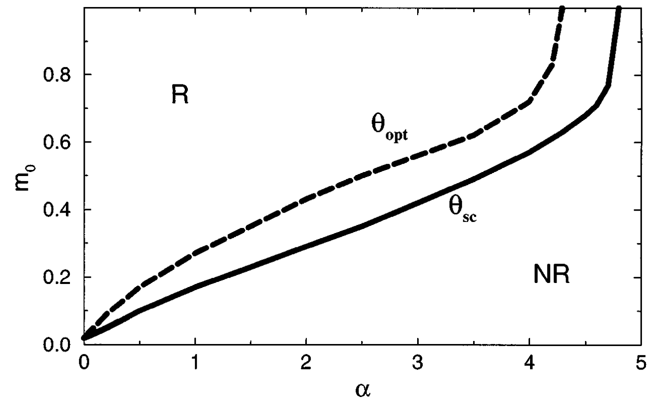

FIG. 3. The basin of attraction as a function of α for $a = 0.01$ and initial $q_0 = a$ for the self-control model (full line) and the optimal threshold model (dashed line).

The expressions for the entropies defined above become

$$S(\sigma) = -q\ln q - (1 - q)\ln(1 - q), \qquad (11)$$

$$\langle S(\sigma|\xi)\rangle_\xi = -a[m\ln(m) + (1 - m)\ln(1 - m)] - (1 - a) \times [\gamma_0\ln\gamma_0 + (1 - \gamma_0)\ln(1 - \gamma_0)]. \qquad (12)$$

Recalling Eq. (6) this completes the calculation of the mutual information content of the present model.

We have solved this self-controlled dynamics for the sparsely coded network numerically and compared its retrieval properties with non-self-controlled models. We are interested only in the retrieval solutions leading to $M > 0$ and carrying a nonzero information I.

In Fig. 1 we have plotted the information content $i \equiv pNI/\#J = \alpha I$ as a function of θ for $a = 0.1$ and $a = 0.001$ and different values of α, *without* self-control. This illustrates that it is difficult, especially for sparse coding, to choose a threshold interval such that i is nonzero.

In Fig. 2 we compare the time evolution of the retrieval overlap, m_t, starting from several initial values, m_0, for

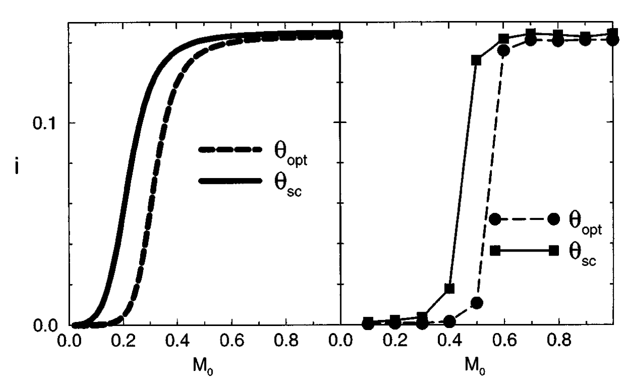

FIG. 4. The information i as a function of M_0 for $a = 0.3, \alpha = 0.4, t = 8$ with $(c = 1)$ and without self-control. Left: analytic results. Right: simulations for $N = 47.000, C = 50, p = 20$ averaged over ten samples.

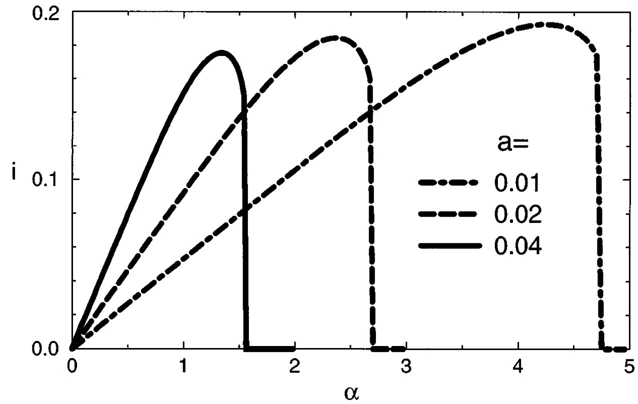

FIG. 5. The information i as a function of α for the self-control model with several values of a.

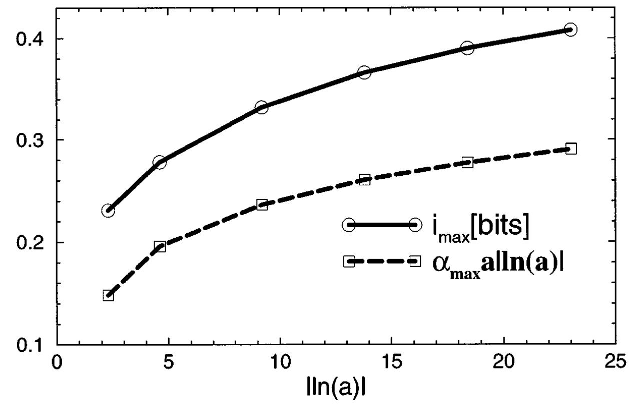

FIG. 6. The maximal information $i_{\max}/\ln(2)$ and $\alpha_{\max}a|\ln(a)|$ as a function of $|\ln(a)|$.

the self-control model with an initial neural activity $q_0 = 0.01 = a$ and $\theta_{sc} = [-2(\ln a)\alpha Q_t]^{1/2}$, with the model where the threshold is chosen by hand in an optimal way in the sense that we took the one with the greatest information content i, by looking at the corresponding results of Fig. 1 for $a = 0.01$. We see that the self-control forces more of the overlap trajectories to go to the retrieval attractor. It does improve substantially the basin of attraction. This is further illustrated in Fig. 3 where the basin of attraction for the whole retrieval phase R is shown for the model with a θ_{opt} selected for every loading α and the model with self-control θ_{sc}. We remark that even near the border of critical storage the results are still improved. Hence the storage capacity itself is also larger. These results are not strongly dependent upon the initial value of q_0 as long as $q_0 = \mathcal{O}(a)$.

Furthermore, we find that self-control gives a comparable improvement for not so sparse models, e.g., $a \sim 0.3$. This is illustrated in Fig. 4 where we show some analytic results together with a first set of simulations for the basins of attraction. This type of simulation for extremely diluted models is known to be difficult because of the theoretical limits $C, N \to \infty$ and $\ln C \ll \ln N$. Nevertheless, it is clear that concerning the self-control aspect, qualitative agreement with the analytic results is obtained. For these values of a, M_0 is the relevant quantity. The quantitative difference is mostly due to the fact that $M_0^{\text{anal}} \sim M_0^{\text{simul}} + \mathcal{O}(1/\sqrt{Ca})$.

Figure 5 displays the information i as a function of α for the self-controlled model with several values of a. We observe that $i_{\max} \equiv i(\alpha_{\max})$ is reached somewhat before the critical capacity and that it slowly increases with increasing α.

Finally, in Fig. 6 we have plotted $i_{\max}/\ln(2)$ and $\alpha_{\max}a|\ln(a)|$ as a function of the activity on a logarithmic scale. It shows that $i_{\max}$ increases with $|\ln(a)|$ until it starts to saturate. The saturation is rather slow, in agreement with results found in the literature [12,13].

In conclusion, we have found a novel way to let a diluted network autonomously control its dynamics such that the basins of attraction and the mutual information content are maximal.

We thank S. Amari and G. Jongen for useful discussions. This work has been supported by the Research Fund of the K. U. Leuven (Grant No. OT/94/9). One of us (D. B.) is indebted to the Fund for Scientific Research–Flanders (Belgium) for financial support.

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, 1991).
[2] J. Nadal and G. Toulouse, Netw., Comput. Neural Syst. **1**, 61 (1990).
[3] M. Okada, Neural Netw. **9**, 1429 (1996).
[4] D. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **35**, 2293 (1987).
[5] S. Amari, Neural Netw. **2**, 451 (1989).
[6] J. Buhmann, R. Divko, and K. Schulten, Phys. Rev. A **39**, 2689 (1989).
[7] E. Barkai, I. Kanter, and H. Sompolinsky, Phys. Rev. A **41**, 590 (1990).
[8] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[9] D. Bollé, G. M. Shim, B. Vinck, and V. A. Zagrebnov, J. Stat. Phys. **74**, 565 (1994).
[10] D. Horn and M. Usher, Phys. Rev. A **40**, 1036 (1989).
[11] M. V. Tsodyks, Europhys. Lett. **7**, 203 (1988).
[12] C. J. Perez-Vicente, Europhys. Lett. **10**, 621 (1989).
[13] H. Horner, Z. Phys. B **75**, 133 (1989).
[14] C. E. Shannon, Bell Syst. Techn. J. **27**, 379 (1948).