Thresholds in layered neural networks with variable activity

2000 J. Phys. A: Math. Gen. 33 2597

(http://iopscience.iop.org/0305-4470/33/14/301)

Thresholds in layered neural networks with variable activity

D Bollé†§ and G Massolo†‡

† Instituut voor Theoretische Fysica, K U Leuven, B-3001 Leuven, Belgium
‡ Università degli Studi dell'Insubria, Facoltà di Scienze, I-22100 Como, Italy

E-mail: `desire.bolle@fys.kuleuven.ac.be`

Received 1 October 1999, in final form 10 February 2000

Abstract. The inclusion of a threshold in the dynamics of layered neural networks with variable activity is studied at arbitrary temperature. In particular, the effects on the retrieval quality of a self-controlled threshold obtained by forcing the neural activity to stay equal to the activity of the stored patterns during the whole retrieval process, are compared with those of a threshold chosen externally for every loading and every temperature through optimization of the mutual information content of the network. Numerical results, mostly concerning low-activity networks are discussed.

1. Introduction

Recently, the introduction of a threshold in the dynamics of neural networks with low activity was discussed again by several authors [1–3] (and references therein). Extremely diluted models [1,3] and models for sequential patterns [2] have been looked at. In all cases it is found that the retrieval quality—overlap, basin of attraction, critical storage capacity, information content—depends on the methods of activity control employed. New insights in the dynamical properties of these models have been obtained and new suggestions have been put forward for the choice of threshold functions in order to get enhanced retrieval.

For these extremely diluted asymmetric networks the dynamics is relatively simple. This is due to the fact that there are no feedback correlations and no common ancestors in this system such that the neurons are completely uncorrelated in the course of the time evolution ([4,5] and references therein). The evolution equations can then be written down in closed form and the structure of these equations does not even change anymore after the first time step. Therefore, the question is relevant whether the introduction of such a self-control threshold function in models having more complicated dynamics still leads to enhanced retrieval.

In this context it is interesting to study layered networks. First, as is common knowledge by now, exactly these models are used in many applications in several areas of research [6–8]. Secondly, these networks contain correlations among the neurons because of the presence of common ancestors. Nevertheless, these correlations can be handled exactly giving rise to layer-to-layer evolution equations in closed form [9–14]. We remark that such a closed form is no longer possible for fully connected networks [5, 15].

In view of the fact that in practical applications of pattern recognition, information is generally encoded by a small fraction of bits and that in neurophysiological studies the activity levels of real neurons is found to be low, the models discussed in what follows are allowed

§ Also at Interdisciplinair Centrum voor Neurale Netwerken, K U Leuven, Belgium.

0305-4470/00/142597+13$30.00

to have a variable pattern activity [16]. The limit of low activity (or in other words sparse coding) is especially interesting. Sparsely coded models have indeed a very large storage capacity behaving as $1/(a \ln a)$ in the limit a going to 0, where a is the pattern activity (see, e.g., [17–21] 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. For the models mentioned above these drawbacks can be avoided and an optimal retrieval performance can be reached by introducing an appropriate threshold in the dynamics [1–3, 21]. In this paper we study whether this can also be done for layered models.

In the layered models discussed in what follows we take two different approaches. The first one consists in forcing the neural activity to be the same as the activity of the stored patterns during the whole retrieval process. In order to guarantee this we introduce a time-dependent threshold in the dynamics chosen as a function of the noise and the pattern activity in the network and adapting itself autonomously in the course of the time evolution. This is the self-control method proposed in [3].

The second approach chooses a threshold by optimizing the information content of the network since for very small pattern activities the number of active neurons and the information represented by a single pattern decreases. The relevant quantity we use here is the mutual information function [3,22,23] and the threshold will be called the optimal threshold. Here the threshold is time-independent and externally chosen for every loading and every temperature. Both methods are compared for zero and non-zero temperatures for networks with various activities.

The rest of this paper is organized as follows. In section 2 we introduce the layered network model and define the relevant order parameters. Section 3 presents the dynamical evolution equations for these order parameters obtained by the probabilistic signal-to-noise ratio analysis. In section 4 we discuss the different thresholds mostly in the context of low activity. In section 5 we present numerical results at zero and non-zero temperatures. Finally we end with some concluding remarks in section 6.

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 $t = 1, \ldots, L$ is the layer index and $i = 1, \ldots, N$ labels the site. Each neuron in layer t is unidirectionally connected to all neurons on layer $t + 1$. We want to store $p = \alpha N$ patterns $\{\xi_i^\mu(t)\}$, $i = 1, \ldots, N$, $\mu = 1, \ldots, p$ on each layer t, taking the values $\{0, 1\}$. They are assumed to be independent identically distributed random variables (IIDRV) with respect to i, μ 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: $\mathrm{Cov}(\xi_i^\mu(t), \xi_i^\nu(t)) \equiv E[\xi_i^\mu(t)\xi_i^\nu(t)] - E[\xi_i^\mu(t)]E[\xi_i^\nu(t)] = 0$. In what follows it will be convenient to make the change of variables $\eta_i^\mu(t) = \xi_i^\mu(t) - a$ such that the interesting expectation values are $E[\eta_i^\mu(t)] = 0$ and $E[\eta_i^\mu(t)^2] = a(1 - a) \equiv \tilde{a}$.

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

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

The parameter $\beta = 1/T$ controls the stochasticity of the network dynamics, it measures the noise level. Given the configuration $\{\sigma_i(t)\}; i = 1, \ldots, N$ on layer t, the local field $h_i(t)$ in site i on the next layer $t+1$ is given by

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

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

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

These couplings then permit one to store sets of patterns to be retrieved by the layered network. We remark that in the limit $T \to 0$ the updating rule (1) reduces to the deterministic form

$$\sigma_i(t+1) = \Theta(h_i(t)) \tag{4}$$

where $\Theta(x)$ is the standard step function taking the value $\{0, 1\}$.

We take parallel updating. The dynamics of this network is defined as follows (see [9,10] and references therein). 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. Layered feed-forward networks allow an exact analytic treatment of their parallel dynamics stemming from the independent choice of the representations of the patterns on different layers. By exact analytic treatment we mean that, given the configuration of the first layer as the initial state, the configuration on layer t that results from the dynamics is predicted by recursion formulae for the relevant order parameters. This configuration is known through the calculation of macroscopic quantities obtained by averaging over the thermal noise associated with the dynamics, as well as over the random choice of the stored patterns.

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

$$M_N^{\mu}(t) = \frac{1}{N\tilde{a}} \sum_{i=1}^{N} \eta_i^{\mu}(t)(\sigma_i(t) - a) \qquad q_N(t) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(t). \tag{5}$$

These order parameters determine the Hamming distance between the state of the network and the pattern $\{\xi_i^{\mu}(t)\}$

$$d_H(\xi^{\mu}(t), \sigma(t)) = \frac{1}{N} \sum_{i=1}^{N} [\xi_i^{\mu}(t) - \sigma_i(t)]^2. \tag{6}$$

It is known that the Hamming distance is a good measure for the performance of a network when the neural activity $a \sim \frac{1}{2}$. For low-activity networks, however, it does not give a complete description of the information content [3]. Therefore, the mutual information function $I(\sigma_i(t); \xi_i^{\mu}(t))$ has been introduced [3,23]:

$$I(\sigma_i(t); \xi_i^{\mu}(t)) = S(\sigma_i(t)) - \langle S(\sigma_i(t)|\xi_i^{\mu}(t))\rangle_{\xi^{\mu}(t)} \tag{7}$$

where $\xi_i^{\mu}(t)$ is considered as the input and $\sigma_i(t)$ as the output with $S(\sigma_i(t))$ its entropy and $S(\sigma_i(t)|\xi_i^{\mu}(t))$ its conditional entropy, namely,

$$S(\sigma_i(t)) = -\sum_{\sigma} p(\sigma_i(t)) \ln[p(\sigma_i(t))] \tag{8}$$

$$S(\sigma_i(t)|\xi_i^{\mu}(t)) = -\sum_{\sigma} p(\sigma_i(t)|\xi_i^{\mu}(t)) \ln[p(\sigma_i(t)|\xi_i^{\mu}(t))]. \tag{9}$$

Here $p(\sigma_i(t))$ denotes the probability distribution for the neurons at time t and $p(\sigma_i(t)|\xi_i^{\mu}(t))$ indicates the conditional probability that the ith neuron is in a state $\sigma_i(t)$ at time t given that the ith site of the stored pattern to be retrieved is $\xi_i^{\mu}(t)$.

3. Dynamics at arbitrary temperature

We suppose that the initial configuration $\{\sigma_i(1)\}$ is a collection of IIDRV with average and variance given by $E[\sigma_i(1)] = E[(\sigma_i(1))^2] = q_0$. We furthermore assume that this configuration is correlated with only one stored pattern, say pattern $\mu = 1$, such that

$$\mathrm{cov}(\xi_i^\mu(1), \sigma_i(1)) = \delta_{\mu,1} M_0^1 \tilde{a}. \tag{10}$$

We then obtain the order parameters (5) at the initial time step $t = 1$ in the thermodynamic limit by the law of large numbers (LLN). For the main overlap we have

$$M^\mu(1) \equiv \lim_{N\to\infty} M_N^\mu(1) \overset{LLN}{=} \frac{1}{\tilde{a}} E[\eta_i^\mu(1)(\sigma_i(1) - a)] = \frac{1}{\tilde{a}} \mathrm{cov}(\xi_i^\mu(1), \sigma_i(1)) = \delta_{\mu,1} M_0^1 \tag{11}$$

and for the neural activity

$$q(1) \equiv \lim_{N\to\infty} q_N(1) \overset{LLN}{=} E[\sigma_i(1)] = q_0. \tag{12}$$

The evolution equations governing the dynamics are then obtained following the methods based upon a *signal-to-noise* analysis of the local field (see, e.g., [9–14] for the case without threshold and without bias, i.e., $a = \frac{1}{2}$). The local field is split as the sum of a signal (from the condensed pattern $\mu = 1$) and a noise (from the non-condensed patterns $\mu > 1$). For a recent overview comparing various architectures we refer the reader to [5]. Since the method is standard by now we only write down the final results. For a general time step at zero temperature we obtain

$$M^1(t+1) = 1 - \frac{1}{2}\left\{ \mathrm{erfc}\left(\frac{(1-a)M^1(t) - \theta(t)}{\sqrt{2\alpha D(t)}}\right) + \mathrm{erfc}\left(\frac{aM^1(t) + \theta(t)}{\sqrt{2\alpha D(t)}}\right)\right\} \tag{13}$$

$$q(t+1) = aM^1(t+1) + \frac{1}{2}\mathrm{erfc}\left(\frac{aM^1(t) + \theta(t)}{\sqrt{2\alpha D(t)}}\right) \tag{14}$$

$$\begin{aligned} D(t+1) = Q(t+1) + \frac{1}{2\pi\alpha}\Bigg\{ & a \exp\left(-\frac{(\theta(t) - (1-a)M^1)^2}{2D(t)\alpha}\right) \\ & + (1-a)\exp\left(-\frac{(\theta(t) + aM^1)^2}{2D(t)\alpha}\right)\Bigg\}^2 \end{aligned} \tag{15}$$

where $Q(t) = (1-2a)q(t) + a^2$ and $D(t)$ is the variance of the residual overlap containing the influence of the non-condensed patterns $\mu > 1$. The residual overlap is defined as

$$r_N^\mu(t) = \frac{1}{\sqrt{N\tilde{a}}} \sum_{i=1}^N \eta_i^\mu(t)(\sigma_i(t) - a) \qquad \mu > 1 \tag{16}$$

and causes the intrinsic noise in the dynamics of the main overlap $M^1(t)$. Finally, $\mathrm{erf}(x) = (2/\sqrt{\pi}) \int_0^x \mathrm{d}y \exp(-y^2)$.

For non-zero temperatures thermal averages denoted by $\langle \cdots \rangle$ have to be taken in agreement with the distribution (1) such that

$$\langle \sigma_i(t+1) \rangle = \tfrac{1}{2}[1 + \tanh(\beta \langle h_i(t) \rangle)] \tag{17}$$

and

$$M^\mu(t) \equiv \lim_{N\to\infty} \langle M_N^\mu(t) \rangle \qquad q(t) \equiv \lim_{N\to\infty} \langle q_N(t) \rangle. \tag{18}$$

The stochastic dynamics can then be described through the following equations for the order parameters:

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

$$+\int \mathcal{D}x \tanh\left[\beta\left(-aM^1(t)-\theta(t)+\sqrt{\alpha D(t)}x\right)\right]\Bigg\} \tag{19}$$

$$q(t+1) = aM^1(t+1)+\tfrac{1}{2}\left\{1+\int \mathcal{D}x \tanh\left[\beta\left(-aM^1(t)-\theta(t)+\sqrt{\alpha D(t)}x\right)\right]\right\} \tag{20}$$

$$\begin{aligned} D(t+1) = Q(t+1)+\frac{\beta}{2}\Bigg\{&1-a\int \mathcal{D}x \tanh^2\beta\left[(1-a)M^1(t)-\theta(t)+\sqrt{\alpha D(t)}x\right] \\ &-(1-a)\int \mathcal{D}x \tanh^2\beta\left[-aM^1(t)-\theta(t)+\sqrt{\alpha D(t)}x\right]\Bigg\} \end{aligned} \tag{21}$$

where $\mathcal{D}x$ is the Gaussian measure $\mathcal{D}x = x(2\pi)^{-1/2}\exp(-x^2/2)$. One can show that in the special case of $\theta(t) = 0$ and $a = \frac{1}{2}$ these equations become equivalent to those derived in [9, 10].

4. Thresholds

4.1. Low activity and self-control

In the limit of low activity it has been emphasized already in the study of extremely diluted and fully connected architectures that, in general, one should try to keep the pattern activity of the network during the retrieval process the same as the one for the memorized patterns [1, 3, 16, 21, 24–26]. In addition, for the layered model considered here one easily finds for fixed α and zero threshold by using equations (13) and (14) that in the limit $a \to 0$ the neural activity behaves as $q(t) \sim \frac{1}{2}+aM^1(t)$ and always tends to $\frac{1}{2}$. The way to avoid this is to choose, given a, the capacity α such that $aM^1(t) \sim \sqrt{2\alpha D(t)}$, however, this means that when a decreases the critical capacity α_c is going to decrease too. In fact, numerical experiments on the layered model show that for $\theta = 0$ and $a \approx 10^{-3}$, $\alpha_c \approx 10^{-4}$. Similar considerations stay valid for non-zero temperatures.

Therefore, in the retrieval process we need to control the neural activity and keep it, at each layer, the same as the one for the stored patterns: $q(t) = a$. For a network with low activity this requires the introduction of a threshold $\theta(t)$ in the definition of the local field (2). For the extremely diluted model a time-dependent threshold has been chosen in [3] as a function of the noise in the system and the pattern activity, adapting itself in the course of the time evolution. The novel idea was to let the network itself autonomously counter the residual noise at each time step of the dynamics without having to impose any external constraints. Here, we want to pursue this idea for the layered model.

In the following we start from an analogous general form for this self-control threshold

$$\theta(t)_{\mathrm{sc}} = c(a)\sqrt{\alpha D(t)} \tag{22}$$

where we recall that $D(t)$ is the variance of the noise contribution in the local field. For the determination of $c(a)$ we consider the form (14) for the layered architecture and require that the term

$$\operatorname{erfc}\left(\frac{aM^1(t)+\theta(t)}{\sqrt{2\alpha D(t)}}\right) = \operatorname{erfc}\left(\frac{aM^1(t)}{\sqrt{2\alpha D(t)}}+\frac{c(a)}{\sqrt{2}}\right) \sim \frac{\mathrm{e}^{-c(a)^2/2}}{c(a)\sqrt{2\pi}} \tag{23}$$

must vanish faster than a. This can be realized by choosing $c(a) = \sqrt{-2\ln a}$. Furthermore, we remark that in the low-activity limit the recursion relation (15) for $D(t+1)$ leads to $D(t+1) \sim Q(t+1)$. This shows explicitly that in this limit the result for the layered model is similar to the one for the extremely diluted model [3]. Indeed, we intuitively expect that in the limit of very low activity all models roughly behave in the same way.

The line of arguments above is also valid at arbitrary temperature. In the limit of low activity it is straightforward to show that the second term on the rhs of equation (20) vanishes faster than the activity a.

We recall that this self-control threshold (22) is a macroscopic parameter, thus no average must be taken over the microscopic random variables at each time step t. We have in fact a mapping with a threshold changing each time step, but no statistical history intervenes in this process.

In the next section we study explicitly the influence of this threshold on the retrieval quality of the network dynamics. For the extremely diluted model [1, 3] and in the case of sparsely coded sequential patterns [2] it has been shown that this retrieval quality is considerably improved for low activity. In the case of the extremely diluted model this improvement also works for not so low activity [3]. Furthermore, although the form of the threshold has been derived at zero temperature, we also want to find out whether it works at finite temperatures.

4.2. Optimizing the mutual information

We have argued that the mutual information function (7) is a better concept than the Hamming distance in order to measure the retrieval quality especially in the limit of low activity. So, a second type of threshold we introduce is obtained by optimizing this mutual information.

We start by calculating the mutual information for the case at hand using the equations (7)–(9). In what follows we drop the index t. Because of the mean-field character of our model the following formula hold for every site index i on each layer t. After some algebra we find for the conditional probability

$$p(\sigma|\xi) = [\gamma_0\xi + (\gamma_1 - \gamma_0)\xi]\delta(\sigma - 1) + [1 - \gamma_0 - (\gamma_1 - \gamma_0)\xi]\delta(\sigma) \qquad (24)$$

where $\gamma_0 = q - aM^1$ and $\gamma_1 = (1 - a)M^1 + q$, and where the M^1 and q are precisely the order parameters (5) for $N \to \infty$. Using the probability distribution of the patterns we obtain

$$p(\sigma) = q\delta(\sigma - 1) + (1 - q)\delta(\sigma). \qquad (25)$$

Hence the entropy (8) and the conditional entropy (9) become

$$S(\sigma) = -q \ln q - (1 - q)\ln(1 - q) \qquad (26)$$

$$\begin{aligned} S(\sigma|\xi) = &-[\gamma_0 + (\gamma_1 - \gamma_0)\xi]\ln[\gamma_0 + (\gamma_1 - \gamma_0)\xi] \\ &-[1 - \gamma_0 - (\gamma_1 - \gamma_0)\xi]\ln[1 - \gamma_0 - (\gamma_1 - \gamma_0)\xi]. \end{aligned} \qquad (27)$$

By averaging the conditional entropy over the pattern ξ we get

$$\begin{aligned} \langle S(\sigma|\xi)\rangle_\xi = &-a[\gamma_1 \ln\gamma_1 + (1 - \gamma_1)\ln(1 - \gamma_1)] \\ &-(1 - a)[\gamma_0 \ln\gamma_0 + (1 - \gamma_0)\ln(1 - \gamma_0)] \end{aligned} \qquad (28)$$

such that the mutual information function (7) for the layered model is given by

$$\begin{aligned} I(\sigma;\xi) = &-q\ln q - (1 - q)\ln(1 - q) + a[\gamma_1 \ln\gamma_1 + (1 - \gamma_1)\ln(1 - \gamma_1)] \\ &+(1 - a)[\gamma_0 \ln\gamma_0 + (1 - \gamma_0)\ln(1 - \gamma_0)]. \end{aligned} \qquad (29)$$

At time t the mutual information function depends on the main overlap $M^1(t)$, the neural activity $q(t)$, the pattern activity a, the load parameter α and the inverse temperature β. The evolution of the main overlap and of the neural activity (equations (13), (14) for zero temperature and (19), (20) for arbitrary temperature) depends on the specific choice of the threshold in the definition of the local field (2). We consider a time-independent threshold $\theta(t) = \theta$ and calculate the value of (29) at equilibrium for fixed a, α, M_0, q_0 and β. The optimal choice for this threshold chosen at equilibrium, $\theta = \theta_{\rm opt}$, is then the one for which the mutual information function is maximal.

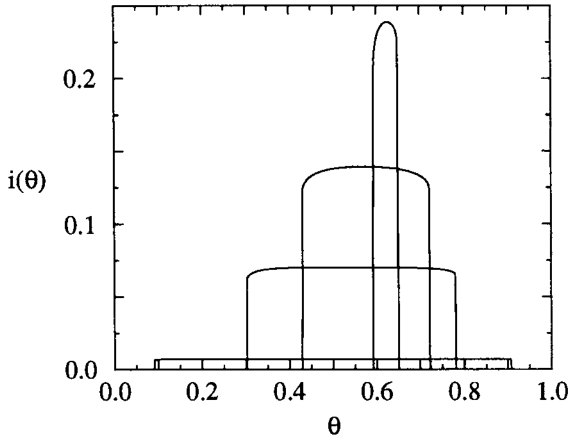

Figure 1. The information i as a function of θ for $a = 0.01$ and several values of the load parameter $\alpha = 0.1, 1, 2, 4$ (bottom to top).

5. Results

We have studied the retrieval properties for the layered model with θ_{sc} and θ_{opt} by numerically solving the recursion relations derived in section 3 with an activity ranging from $a = 0.001$ to 0.3 at various inverse temperatures $\beta = 3, 4, 5, 10, 100, \infty$. We are interested only in the retrieval solutions with $M^1 > 0$ (in what follows we drop the superscript 1) and carrying a non-zero information I. The results for zero and non-zero temperature have been analysed separately. Our main aim is to study how self-control introduced for extremely diluted networks also works for other models, in the case of a layered architecture at zero temperature, as claimed in [3], and to check whether such a threshold can still be useful at non-zero temperatures. Moreover, we compare this self-control method, which is mainly designed for low activity but also works for higher activities, with the optimization method. The latter works for all values of the activity although it has to be found externally for every loading and every temperature.

5.1. Zero temperature

In figure 1 we have plotted the information content $i = \alpha I$ as a function of θ without self-control or *a priori* optimization for pattern activity $a = 0.01$ and different values of the load parameter α. For every value of α, below its critical value, there is a range for the threshold where the information content is different from zero. For any choice of the threshold in this range retrieval is possible. This retrieval range becomes very small when the capacity approaches its critical value $\alpha_c = 4.72$.

Defining the basin of attraction as the range of initial values $M_0 \in [0, 1]$ which lead to the retrieval attractor $M(t) \sim 1$, we note at this point that the size of this basin strongly depends on the specific choice of the threshold in the retrieval range. Technically it turns out that the value to be chosen for the latter in order to have the largest basin is the minimal θ in the retrieval range. This, of course, has to be repeated for every α. This threshold optimizes the information content and is called, as specified before, θ_{opt}.

Figure 2 represents the dynamical evolution of the network. The retrieval overlap $M(t)$ is shown as a function of time for different initial values M_0, $q_0 = 0.001 = a$ and $\alpha = 25$. A self-control threshold $\theta_{\mathrm{sc}} = [-2(\ln a)\alpha Q(t)]^{-1/2}$ (figure 2(*a*)) is compared with an optimal

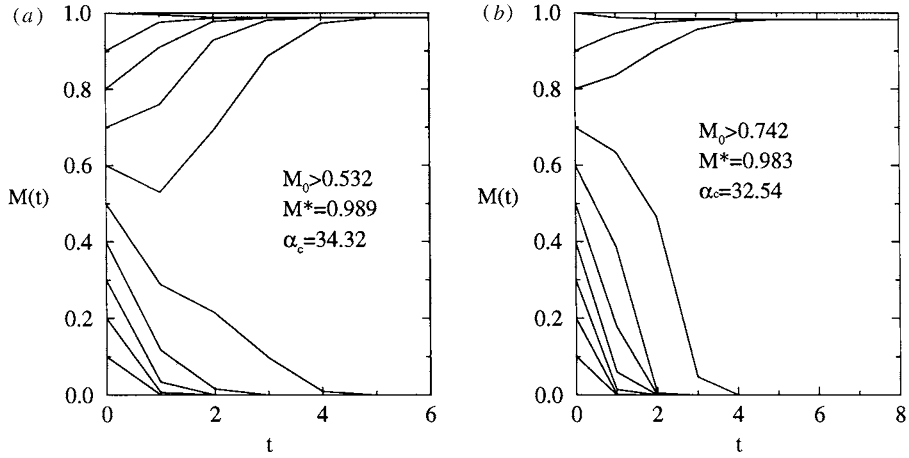

Figure 2. The evolution of the main overlap $M(t)$ for several initial values M_0 with $q_0 = a = 0.001$, $\alpha = 25$ for the self-control model (*a*) and the optimal threshold model (*b*).

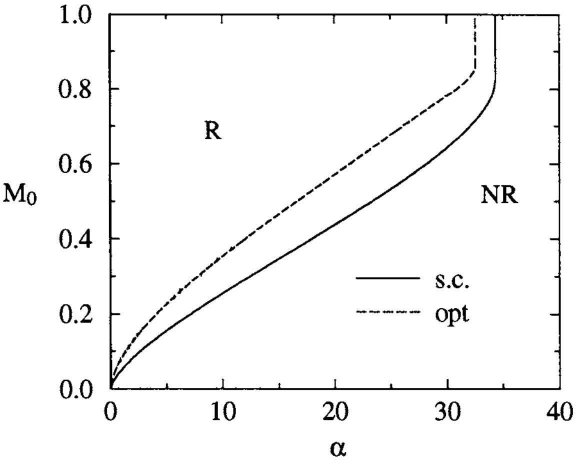

Figure 3. The basin of attraction as a function of α for $a = 0.001$ for the self-control model (full curve) and the optimal threshold model (dashed curve).

threshold θ_{opt} (figure 2(*b*)) concerning the values of the minimal M_0 for retrieval, the fixed-point M^* and the critical capacity α_c. It is seen that self-control works better than optimization and both much better than a zero threshold (where there is no retrieval at all since $\alpha_c = 5.3 \times 10^{-5}$ only). This can be interpreted as a result of the property of adaptivity in the course of the time evolution inherent in the self-control method.

In figure 3 the retrieval phase diagram is illustrated for $a = 0.001$ and $q_0 = a$. In the low-activity limit the basin of attraction is substantially improved by self-control even near the border of the critical storage. Hence, the storage capacity is also larger with self-control. Furthermore, we have compared these curves with the one for a model without threshold in the low-activity limit. Since we find a very small storage capacity (of order 10^{-4}) such a network without threshold has very little interest.

Plotting the retrieval fixed-point M^* as a function of α we have found a first-order transition

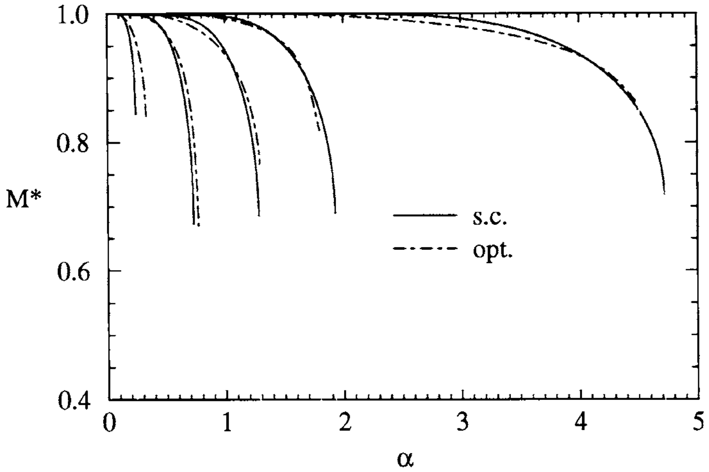

Figure 4. The retrieval fixed-points M^* as a function of α for the self-control model (full curve) and the optimal threshold model (dashed curve) with decreasing pattern activity: $a = 0.3, 0.1, 0.05, 0.03, 0.01$ (from left to right).

from the retrieval phase ($M^* > 0$) to the non-retrieval one ($M^* = 0$), see figure 4 for different values of a. We remark that the curves for $a = 0.001$ are out of the scale of this figure. In this case we find $\alpha_c = 34.32$ and $M^* \sim 1$ for $0 < \alpha < 20$. We compare the fixed-point behaviour found with self-control (solid curves) with the results obtained by choosing the threshold through the optimization of the mutual information function (dashed curves). Roughly speaking, self-control is the best choice for activities below 0.05. For a above this value, but still small compared with a homogeneous distribution $a = \frac{1}{2}$, e.g. $a = 0.3$, self-control continues to perform quite well, however it ceases to be better than optimization.

Finally, we have studied the leading behaviour of the critical capacity in the limit $a \to 0$. We have found that $\alpha_c(a) \sim (a|\ln a|)^{-1}$. This is consistent with former studies on other low-activity models (see [1,3] and references therein). Moreover, we remark that for a in the range $(10^{-4}, 10^{-3})$ the proportionality coefficient seems to be constant and given by 0.25.

5.2. Non-zero temperature

Since self-control is completely autonomous and since it also improves the retrieval quality for not so sparse networks it is worth checking how it performs for non-zero temperatures. In addition, in this case we compare it with the optimal threshold for which we recall that it has to be calculated by hand when the network has reached equilibrium for every loading α and every inverse temperature β.

In figure 5 we have studied the retrieval fixed-points of the main overlap as a function of the load parameter for different values of the temperature and of the pattern activity. The results are plotted for $a = 0.1$ (figure 5(a)) and $a = 0.001$ (figure 5(b)) and increasing β. The lines end at the critical capacity where a first-order transition to the non-retrieval phase occurs. At this point we also recall that for these non-zero temperatures in both cases the presence of a non-zero threshold is strictly necessary in order for the network to evolve toward the retrieval phase for these storage capacities. As in the zero-temperature case, without considering any threshold at all the storage capacity would be very small ($\alpha_c < 10^{-4}$ at low activity) and the dynamics of the network would be uninteresting.

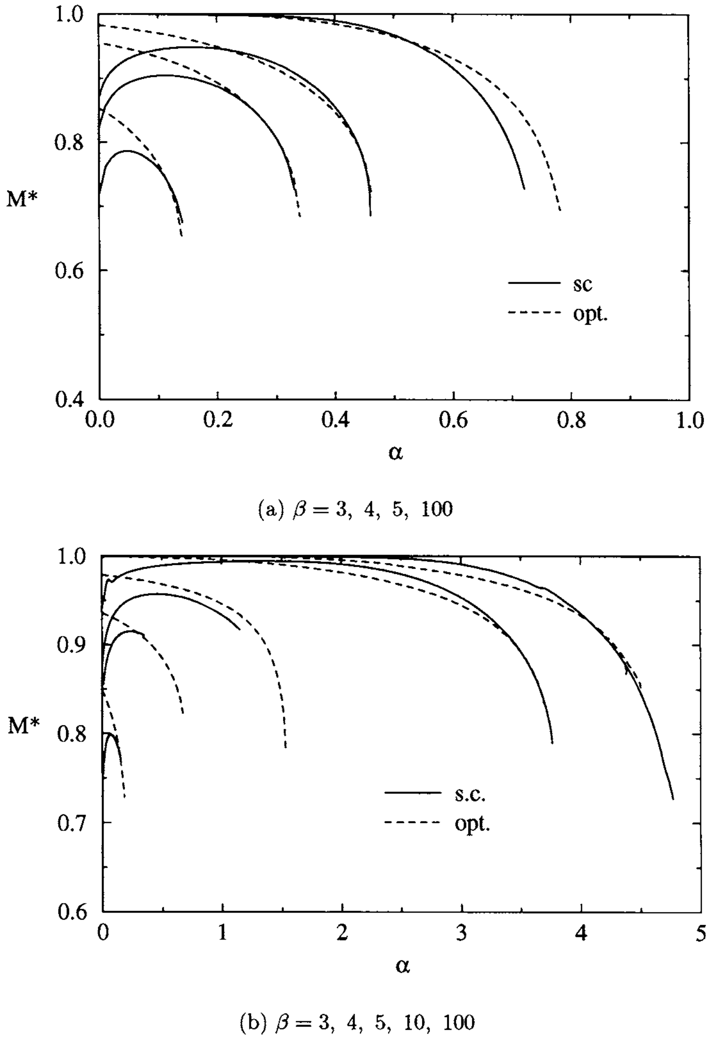

Figure 5. The retrieval fixed-points M^* as a function of α for several values of the inverse temperature for the self-control model (full curve) and the optimal threshold model (dashed curve) for $a = 0.1$ (*a*) and $a = 0.01$ (*b*).

For $\beta = 100$ the results of the deterministic network are found back. For $a = 0.1$ we already know from the previous analysis at zero temperature that optimization works better than self-control. For $a = 0.01$ the reverse situation is valid. For smaller β and smaller storage capacities self-control does not work as well. Optimization leads to a bigger value for the retrieval overlap than self-control does.

For the lowest pattern activity $a = 0.01$ (figure 5(*b*)) self-control works less well for increasing temperature. The critical capacity of the network with self-control is smaller than the critical capacity obtained by optimization. In fact, for $\beta = 3, 4$ it is about half. For pattern activity below $a = 0.01$ the critical capacity with self-control becomes still smaller and it is smaller than the critical capacity obtained by optimization.

We can then summarize the peculiar behaviour with self-control for small storage capacities as follows. We usually expect the retrieval fixed-points to have the greatest overlap values at zero storage capacity and then to slowly decrease until the critical capacity is reached, where there is a phase transition. This is, indeed, the behaviour at zero temperature with whatever choice of the threshold. At non-zero temperature this behaviour is found with the optimization approach, with the self-control method the retrieval fixed-points obtain their maximal retrieval overlap not at zero capacity, but at a higher value.

The analysis of the temperature-capacity phase diagram with self-control and optimization for different values of the pattern activity is summarized in figure 6. We discuss the results for decreasing a. For $a = 0.1$, figure 6(a), the two methods give similar results except near zero temperature where the critical capacity with optimization is slightly bigger than with self-control, like we expect from the analysis at zero temperature. Decreasing the value of the pattern activity to $a = 0.01$, figure 6(b), self-control starts to work less well for a bigger region of high temperatures but it is better at lower temperatures. The curves in figure 6(c) show that at high temperature the region of retrieval with self-control becomes rather small when the activity is further lowered to $a = 0.001$. We also remark that for any choice of the pattern activity below 0.05 there is a value of the temperature where the two curves intersect. This is consistent with the fact that at low activity in the limit of zero -temperature self-control works better than opimization for $a < 0.05$. We conclude that, compared with optimization, self-control gives quite good results for activities in the range $a \in [0.01, 0.05]$. When we want to consider lower activities ($a = 0.001$ and less) at arbitrary non-zero temperature self-control ceases to be a good method to control the noise during the dynamics of the network. In this case the temperature-dependent externally chosen threshold optimizing the mutual information function leads to better retrieval qualities than the temperature-independent self-control threshold.

6. Concluding remarks

In this paper we have studied the effects of a threshold in the gain function on the parallel dynamics in layered neural networks with variable activity. Such a threshold considerably enlarges the critical storage capacity of the network. Two different types of thresholds are compared. The first one forces the neural activity to be the same as the activity of the stored patterns at every step of the retrieval process and adapts itself for this purpose in the course of the time evolution. It provides a complete self-control mechanism. The second optimizes the mutual information function in equilibrium. It has to be given externally. Up to now such a systematic comparative study has only been performed for extremely diluted neural networks the dynamics of which is easier due to the fact that the neurons are completely uncorrelated beyond one time step.

For zero temperatures and low activity $a \leqslant 0.5$ it is found that self-control performs the best in considerably improving the storage capacity, the basin of attraction and the mutual information content, exactly as for extremely diluted models. And, in comparison with the optimization method, it even gives a comparable improvement for higher activities. Moreover, for non-zero temperatures self-control, although being designed at temperature zero, still gives quite good results for lower activities ($a < 0.5$) that are bigger than 0.01. Outside this region optimization done externally for every loading and every temperature leads to better overall retrieval qualities (except, obviously, near the critical capacity at zero temperature). It is worth studying whether self-control can still be improved by making it temperature dependent and/or whether optimization of the mutual information content can be done in a self-controlled way.

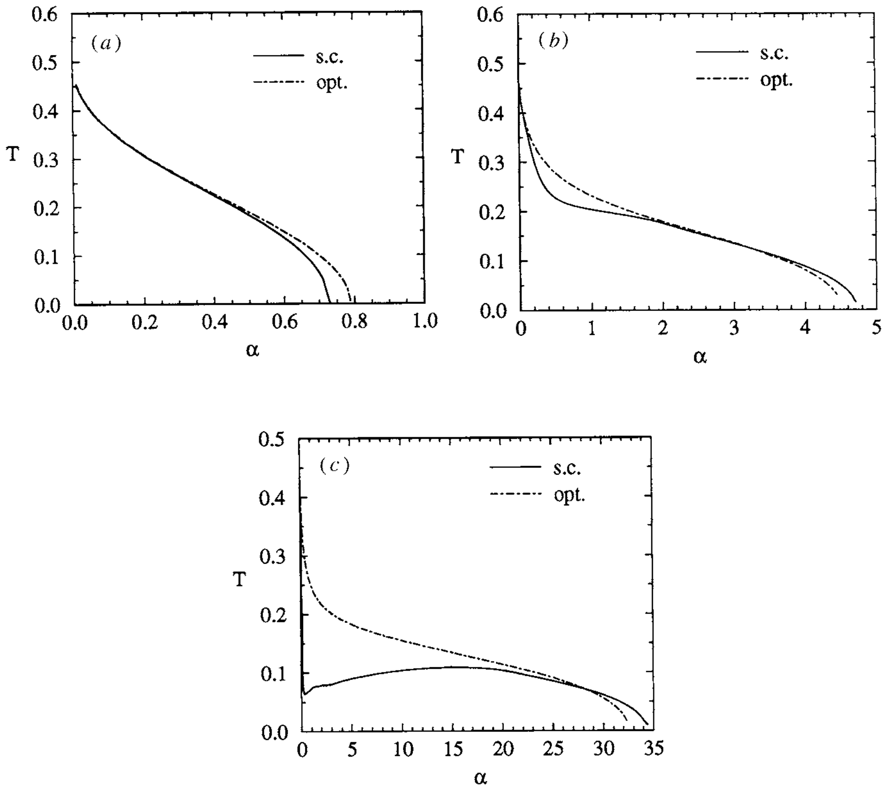

Figure 6. The temperature–capacity phase diagram for the self-control model (full curve) and the optimal threshold model (dashed-dotted curve) for $a = 0.1$ (*a*), $a = 0.01$ (*b*) and $a = 0.001$ (*c*).

Acknowledgments

This work has been supported in part by the Research Fund of the K U Leuven (Grant OT/94/9). The authors are indebted to G Jongen for constructive discussions. D B thanks the Fund for Scientific Research-Flanders (Belgium) for financial support.

References

[1] Grosskinsky S 1999 *J. Phys. A: Math. Gen.* **32** 2983

[2] Kitano K and Aoyagi T 1998 *J. Phys. A: Math. Gen.* **31** L613

[3] Dominguez D R C and Bollé D 1998 *Phys. Rev. Lett.* **80** 2961

[4] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

[5] Bollé D, Jongen G and Shim G M 1999 *Proc. Int. Conf. on Mathematical Physics and Stochastic Analysis (Lisbon, Oct. 1998)* ed S Albeverio *et al* (Singapore: World Scientific) at press
(Bollé D, Jongen G and Shim G M 1999 *Preprint* cond-mat/9907390)

[6] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)

[7] Müller B, Reinhardt J and Strickland M T 1995 *Neural Networks: An Introduction* (Heidelberg: Springer)

[8] Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press)

[9] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081

[10] Bollé D, Shim G M and Vinck B 1994 *J. Stat. Phys.* **74** 583

[11] Domany E, Meir R and Kinzel W 1986 *Europhys. Lett.* **2** 175

[12] Meir R and Domany E 1987 *Phys. Rev. Lett.* **59** 359

[13] Meir R and Domany E 1987 *Europhys. Lett.* **4** 645

[14] Meir R and Domany E 1988 *Phys. Rev.* A **37** 608

[15] Bollé D, Jongen G and Shim G M 1998 *J. Stat. Phys.* **91** 125

[16] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293

[17] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960

[18] Palm G 1981 *Biol. Cyber.* **39** 125

[19] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[20] Nadal J and Toulouse G 1990 *Network: Comput. Neural Syst.* **1** 61

[21] Okada M 1996 *Neural Networks* **9** 1429

[22] Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379

[23] Blahut R E 1997 *Principles and Practice of Information Theory* (Reading, MA: Addison-Wesley)

[24] Amari S 1989 *Neural Networks* **2** 451

[25] Buhmann J, Divko R and Schulten K 1989 *Phys. Rev.* A **39** 2689

[26] Horn D and Usher M 1989 *Phys. Rev.* A **40** 1036