IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Generalization and chaos in a layered neural network

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1997 J. Phys. A: Math. Gen. 30 1403

(http://iopscience.iop.org/0305-4470/30/5/011)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.112
The article was downloaded on 02/06/2010 at 06:12

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **30** (1997) 1403–1414. Printed in the UK PII: S0305-4470(97)76426-6

Generalization and chaos in a layered neural network

David R C Dominguez† and W K Theumann‡

† Theoretical Physics Department C-XI, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain

‡ Instituto de Física, Universidade Federal do Rio Grande do Sul, Caixa Postal 15051, 91501-970, Porto Alegre, RS, Brazil

Received 12 July 1996, in final form 22 October 1996

Abstract. The generalization performance of a multi-state and a graded response layered attractor neural network trained with examples of low activity is established exactly for monotonic and non-monotonic input/output functions. Complex behaviour is found which goes from fixed-point attractors to chaos through a cascade of bifurcations, depending on an appropriate threshold or cut-off parameter. The effect of the irregular behaviour on the generalization curves is explicitly demonstrated and phase diagrams for the recognition ratio of concepts α in terms of the threshold/cut-off exhibit ordered (generalization), disordered (paramagnetic or self-sustained activity) and chaotic phases.

1. Introduction

The generalization ability of perceptrons has been the subject of much interest in recent works in the context of learning an unknown rule [1–3]. The main point is to establish how well the network recognizes correct outputs when it is exposed to *new* examples if it has been trained with a closed pair of examples-correct outputs. In contrast, generalization in an attractor neural network is a different problem that deals with the ability of a network to create a representation for concepts through the extraction of common features from a *fixed* set of examples to which the network has been exposed in the learning stage. The concepts may be thought of as the ancestors of a hierarchically correlated set of patterns [4], which are recognized when they become stable states of the attractor dynamics of the network and this is possible through the presence of symmetric mixture states with the stored patterns [5–7]. These are states that have a sizeable common overlap with a finite number of training examples and they appear already in networks with binary neurons [8]. Such states are responsible for the strong inferential properties in networks with discrete multi-state neurons trained with low-activity patterns in which there is a finite fraction of inactive neurons. Indeed, patterns of full activity are generated in such networks through the merging of low-activity prototype patterns [9, 10].

In a recent work we showed that the generalization ability of an extremely dilute three-state feedback neural network, with a non-decreasing gain function, can be considerably improved either by training the network with low-activity examples or by moderately increasing the threshold beyond which the neurons are active [11]. The purpose of the present paper is to extend that work in two important aspects. One is the use of a network with a more complex architecture than the extremely dilute one and the other is the introduction of non-monotonic gain functions. We consider a layered feedforward

0305-4470/97/051403+12$19.50 © 1997 IOP Publishing Ltd

network with no loops between neurons which is known to have interesting features and has the property that a deterministic dynamics can be solved exactly [12, 13].

It has recently been found that non-monotonic gain functions are appealing from both computational and biological points of view. Indeed, the critical storage capacity α_c was found to increase [14] beyond Gardner's limiting value for binary neurons $\alpha_c = 2$ [15]. It has been argued that a non-sigmoidal analogue transfer function can implement a local-stability criterion in a network with binary patterns [16]. On the other hand, it has been suggested that characteristic features of firing in cortical neurons could be described by a slanted sigmoidal gain function [17].

The zero-temperature stationary regime of the retrieval dynamics in an extremely dilute network of analogue neurons, with the Hebbian learning rule and piecewise linear non-sigmoidal gain functions, has been analysed in a recent work [18]. In addition to fixed-point attractors, all the steps up to a chaotic behaviour appear for a low loading level of the network [19]. Our purpose here is to deal with a different task, namely the generalization dynamics which aims at recognizing the concepts when the network is trained only with examples of the latter. For the network to be successful in its goal it is important for the states to escape from the attractors of the stored patterns. Otherwise one would just have the retrieval behaviour found before.

The generalization task in an attractor neural network can be realized in a manifold of contexts. First, categorization in its simplest form can be implemented by the use of an alternative Hebbian learning algorithm which stores s examples $\{\eta_i^{\mu\nu}\}_{\nu=1}^{s}$ of the hierarchical ancestor (the concept) $\{\xi_i^\mu\}$ on neuron i $(i = 1, \ldots, N)$, for each $\mu = 1, \ldots, p$, having correlation $\langle \eta_i^{\mu\nu} \xi_j^\mu \rangle = b\delta_{ij}$. The states of the network where the neurons come close to the recognition of a concept, such that, $\{\sigma_i\}_{i=1}^N \sim \{\xi_i^\mu\}_{i=1}^N$, are called the *generalization* states. The phase transition from a disordered to a generalization phase was found to be discontinuous with b for a fully connected network [6], or smooth for a diluted network [20]. For sufficiently large s or b, and a not too large ratio of generated concepts, $\alpha \equiv p/N$, the generalization error may become small enough to consider that the task has been achieved successfully, and the network can even turn out to be robust against synaptic noise [7].

In the context of multi-state patterns, generalization takes place through inference. The coherence between the learned patterns, $\{\eta_i^\mu\}$ say, with low activity $a \equiv \frac{1}{N}\sum_i (\eta_i^\mu)^2 \ll 1$, allows the simultaneous retrieval of patterns on many neurons [9]. Thus, by learning small patterns, one can infer the existence of a whole pattern, with activity $a \sim 1$ and extract more information than that available in the original patterns.

Here, as in [11], we consider generalization in both contexts. The outline of the paper is the following. The layered network model for the generalization problem is described in section 2 for various non-monotonic transfer functions, and recursion relations for the symmetric mixture states are discussed in section 3. The equations for the dynamics are obtained in section 4, and the results for the attractors and irregular behaviour are discussed in section 5. We end with concluding remarks in section 6.

2. The model

We consider a layered feedforward network of neurons $i = 1, \ldots, N$, with either discrete or continuous states, σ_{it} in each layer t, which may also be viewed as a discrete-time index, that are updated according to a zero-temperature parallel deterministic dynamics given by

$$\sigma_{i,t+1} = F_\theta(h_{it}) \qquad i = 1, \ldots, N. \tag{1}$$

Here, $F_\theta(x)$ is anyone of the odd transfer functions defined below and θ is the threshold/cut-off parameter that eventually specifies a deviation from the sign function. The local field at site i on layer t is given by

$$h_{it} = \sum_{j\neq i}^{N} J_{ij}^{t}\sigma_{jt} \tag{2}$$

J_{ij}^{t} being the elements of the synaptic matrix between the neurons i and j in two consecutive layers. We assume here the modified Hebbian rule

$$J_{ij}^{t} = \frac{1}{sb^2N}\sum_{\mu}^{p}\sum_{\nu}^{s}\eta_{i,t+1}^{\mu\nu}\eta_{jt}^{\mu\nu} \tag{3}$$

where $\mu = 1,\ldots,p$ and $\nu = 1,\ldots,s$ label the concepts and the examples, respectively. The examples $\eta_{it}^{\mu\nu}$ are independent identically distributed random variables (IIDRV) built from the *concepts* ξ_{it}^{μ} through the following stochastic process

$$\eta_{it}^{\mu\nu} = \xi_{it}^{\mu}\lambda_{it}^{\mu\nu} \tag{4}$$

$$\langle\lambda_{it}^{\mu\nu}\rangle = \langle\xi_{it}^{\mu}\eta_{it}^{\mu\nu}\rangle = b \geqslant 0 \tag{5}$$

$$\langle(\lambda_{it}^{\mu\nu})^2\rangle = a \tag{6}$$

via the random variables $\lambda_{it}^{\mu\nu}$, where the concepts are assumed to be IIDRV, $\xi_{it}^{\mu} =_{-}^{+} 1$, with equal probability. Thus, a is the activity of an example, defined above, and b is the correlation between an example and the concept to which it belongs, while $\langle\eta_{it}^{\mu\nu}\eta_{it}^{\mu\rho}\rangle = b^2$, for $\nu\neq\rho$, is the correlation between two examples of the same concept. There is no correlation between examples of different concepts as well as between a given concept and the examples of another one. Higher moments of $\lambda_{it}^{\mu\nu}$ are not needed for our purpose.

The pure generalization model is recovered by setting $\lambda_{i}^{\mu\nu} =_{-}^{+} 1$, which amounts to activity $a = 1$, with a positive bias b and a binary transfer function $F_\theta(x)$. On the other hand, the pure multi-state model is obtained by taking the number of examples $s = 1$ and correlation $b = 1$. A low activity $a \ll 1$ indicates that in many sites the patterns are not active, that is $|\eta_{i}^{\mu\nu}| \neq 1$, with the effective size of the learned patterns being $N_e = aN$. Thus, when the activity a is not close to 1, one refers to *small* patterns [10]. In our model the new point of view is that the small examples are samples of the full activity concepts that are to be inferred.

The generalization task (inference) is considered successful if the Hamming distance between the state of the neurons and the concepts ξ_{it}^{μ}, at time t, defined as

$$E_{Nt}^{\mu} = \frac{1}{N}\sum_{i}|\xi_{it}^{\mu} - \sigma_{it}| \tag{7}$$

becomes small for large enough t. For binary concepts, as we deal here, the Hamming distance can be directly related to the overlap of the state of the neurons with a concept,

$$M_{Nt}^{\mu} = \frac{1}{N}\sum_{j}\xi_{jt}^{\mu}\sigma_{jt} \tag{8}$$

through $E_{Nt}^{\mu} = 1 - M_{Nt}^{\mu}$. Thus, E^{μ} will be called the *generalization error*. Note that it is twice the usual error. A particular solution for the overlap of the state of the network $\{\sigma_{it}; i = 1,\ldots,N\}$ with the examples, given by

$$m_{Nt}^{\mu\nu} = \frac{1}{N}\sum_{j}\eta_{jt}^{\mu\nu}\sigma_{jt} \tag{9}$$

determines the generalization phase characterized by the symmetric solution of s components of equal overlap. In terms of the overlaps, the local field becomes

$$h_{it} = \frac{1}{sb^2} \sum_{\mu}^{p} \sum_{\nu}^{s} \eta_{it+1}^{\mu\nu} m_{Nt}^{\mu\nu}. \tag{10}$$

and the dynamics of the network requires the study of the evolution of the overlaps from one layer to the next.

A further relevant quantity is the *dynamical activity*, defined as

$$Q_t = \frac{1}{N} \sum_{i} (\sigma_{it})^2 \tag{11}$$

and, to complete the definition of the model, we work here with the following monotonic and non-monotonic transfer functions. First,

$$F_\theta^{M3}(x) = \operatorname{sgn}(x) \qquad |x| > \theta \tag{12}$$

and zero otherwise, is the usual monotonic three-state transfer function that leads to the firing of the neurons if the local field is larger than the threshold θ. The zero state is responsible for the low dynamical activity of the network. Next, consider the *non*-monotonic three-state function

$$F_\theta^{N3}(x) = \operatorname{sgn}(x) \qquad |x| < \theta \tag{13}$$

and zero otherwise. It is assumed here that the neurons stop firing, which may be the case due to fatigue, when the local field reaches a cut-off θ. Alternatively, one may consider the *non*-monotonic two-state function

$$\begin{aligned} F_\theta^{N2}(x) &= \operatorname{sgn}(x) \qquad |x| < \theta \\ &= -\operatorname{sgn}(x) \qquad |x| \geqslant \theta. \end{aligned} \tag{14}$$

In addition to the multi-state functions introduced so far, it is of interest to check if eventual irregular behaviour also appears for graded response neurons and for that purpose we also consider the non-monotonic analogue function

$$F_\theta^{NA}(h) = \sin\left(\frac{1}{\theta} h\right) \qquad \left|\frac{1}{\theta} h\right| < \pi \tag{15}$$

and zero otherwise, in which θ is the gain parameter and at the same time a cut-off.

3. The symmetric solution

Since we are interested in the generalization ability of the network, we take a configuration in which the overlaps of the state of the network with the examples stored in the learning stage are the symmetric overlaps of s components, of macroscopic size of $O(b)$ for the first concept say, and of $O(\frac{1}{\sqrt{N}} b)$ for the remaining ones. Noting that the concepts are uncorrelated random variables, the generalization process becomes a recognition of p independent variables ξ_i^μ, $\mu = 1, \ldots, p$, so that we may concentrate on a single one. The overlaps with the examples of the remaining $p - 1$ concepts contribute to the noise term in the local field. Thus, we may write for any time t,

$$\begin{aligned} m_{Nt}^{\mu\nu} &= b m_{Nt}^{s} \qquad \mu = 1 \\ &= \frac{b}{N} R_{Nt}^{\mu} \qquad \mu > 1 \end{aligned} \tag{16}$$

for $\nu = 1, \ldots, s$, where the main and residual overlaps m^s_{Nt} and R^μ_{Nt}, respectively, are the same for all ν. For simplicity, we assume the initial configuration to have the same symmetric form. The interesting question of the dependence on a different initial configuration will be discussed elsewhere [21].

We follow here the signal-to-noise approach to the layered network [13]. In the thermodynamic limit, equation (9) yields in the first time step, for $\mu = 1$,

$$m^s_{t=1} \equiv \lim_{N\to\infty} m^s_{Nt=1} = \langle\langle y^{1s}_{t=1}\xi^1_{t=1} F_\theta(h_{t=0})\rangle_s\rangle_\omega \tag{17}$$

according to the law of large numbers (LLN) and it is, thus, i-site independent. Here, $y^{\mu s}_t = \frac{1}{sb}\sum_\nu^s \lambda^{\mu\nu}_t$, and $\langle\langle \ldots \rangle\rangle$ denotes first the average over y^{1s}_t and then over the noise term in the local field,

$$\begin{aligned} h_{t=0} &= \xi^1_{t=1} y^{1s}_{t=1} m^s_{t=0} + \omega_0 \\ \omega_0 &= \frac{1}{\sqrt{N}} \sum_{\mu>1}^{p} \xi^\mu_{t=1} y^{\mu s}_{t=1} R^\mu_{Nt=0} \end{aligned} \tag{18}$$

where $m^s_{t=0}$ is the initial symmetric overlap, and ω_0 is the initial noise produced by the $p-1$ residual symmetric overlaps in equation (16). The first term in the local field favours an alignment of the states with the first concept. Because of the feedforward architecture of the network, these equations are reproduced at all time steps.

In the next time step one needs the limit of the residual symmetric overlaps $R^\mu_{t=1} \equiv \lim_{N\to\infty} R^\mu_{N,t=1}$ which does not satisfy the LLN because their dispersions are of the same magnitude as their mean values. Using the central limit theorem (CLT), however, to evaluate the probability distribution at the first time step yields

$$\lim_{N\to\infty} \frac{R^\mu_{t=1} - \langle R^\mu_{N,t=1}\rangle}{\sqrt{\mathrm{Var}(R^\mu_{N,t=1})}} \doteq Z_{t=1} \tag{19}$$

where the brackets denote the average with respect to the examples at the previous layer (time) and $\doteq$ means the convergence in the distribution to the Gaussian random variable $Z_{t=1} \doteq N(0,1)$, of mean zero and unit variance. The average and variance can be calculated using the LLN giving respectively,

$$\begin{aligned} \lim_{N\to\infty} \langle R^\mu_{N,t=1}\rangle &= A R^\mu_{t=0} C_{t=1} \\ \lim_{N\to\infty} \mathrm{Var}(R^\mu_{N,t=1}) &= A Q_{t=1} \end{aligned} \tag{20}$$

where

$$A \equiv \langle (y^s)^2\rangle = 1 + \frac{a-b^2}{sb^2} \tag{21}$$

and the dynamical variables are

$$\begin{aligned} C_{t+1} &\equiv \langle\langle F'_\theta(h_t)\rangle_s\rangle_\omega \\ Q_{t+1} &\equiv \langle\langle [F_\theta(h_t)]^2\rangle_s\rangle_\omega. \end{aligned} \tag{22}$$

The prime denotes derivative with respect to the argument and Q_t is the dynamical activity, equation (11). It accounts for the active neurons, and plays a similar role as the spin-glass order parameter in the thermodynamic equilibrium approach for binary neurons [22].

4. The macro-dynamics

Since ω_0 is a sum over random uncorrelated variables we may apply the CLT to the sum in equation (18), to get that the noise in the local field is a Gaussian random variable

$$\omega_0/\sqrt{\alpha r_{t=0} A} \doteq z_p = N(0,1) \tag{23}$$

as one expects for a feedforward layered network [13], where we have defined the variance of the residual overlaps, in the $\lim N \to \infty$,

$$r_t \equiv \mathrm{Var}(R_t^\mu). \tag{24}$$

Then, taking $\langle R_t^\mu = 0 \rangle$, where the brackets here denote the average over the (unknown) distribution of the limiting R_t^μ, and the corresponding equation for its variance, we obtain

$$\begin{aligned} m_{t+1}^s &= \langle\langle y^s F_\theta(\Lambda_t)\rangle_s\rangle_p \\ r_{t+1} &= (AC_{t+1})^2 r_t + AQ_{t+1} \\ \Lambda_t &= y^s m_t^s + z_p\sqrt{\alpha r_t A} \end{aligned} \tag{25}$$

with the functions C_t, Q_t given in equations (22). We have used the odd-property of F_θ to introduce the new field $\Lambda_t \equiv \xi^1 h_t$, recalling that ξ^1 is a binary variable. The averages are now over the two random contributions to the field, y^s and z_p. Assuming a large number of examples, say $s \geqslant 10$, $y^s \doteq 1 + z_s\sqrt{A-1}$ with the Gaussian variable $z_s \doteq N(0,1)$, independent of z_p. Then one finds

$$m_{t+1}^s = M_{t+1} + m_t^s \frac{a-b^2}{sb^2} C_{t+1} \tag{26}$$

where

$$M_{t+1} = \langle\langle F_\theta(\Lambda_t)\rangle_{z_s}\rangle_{z_p} \tag{27}$$

is the generalization overlap with the first concept defined in equation (8). These are the dynamical equations that have to be solved. We make no restrictions about the values which the parameters a and b can assume within the $(0, 1]$ interval, except that they must satisfy $a \geqslant b^2$, the equality corresponding to constant microscopic activities $\lambda \equiv b$.

5. Fixed-point and irregular attractors

The phases associated with fixed-points that can appear are those already present in the extremely dilute network [11]. Indeed, there is a paramagnetic phase where $M = 0 = Q$, also called the zero phase (Z), a generalization phase (G) with $M > 0$ and $Q > 0$ as well as a self-sustained activity phase (S) with $M = 0$ and $Q > 0$ [10, 23]. More interesting, however, is the presence of irregular attractors corresponding to non-steady macroscopic phases, that follow from the large-time behaviour of the dynamic equations (25), as will be discussed below.

First we consider, for reference, the case of the monotonic three-state function, equation (12), for which there is always a fixed-point behaviour. In the lower part of figure 1 we show the dependence of the generalization error $E = \lim_{N\to\infty} E_{Nt}^\mu$, for $\mu = 1$, on the threshold in the long-time limit, which is actually reached after a finite number of steps, as a function of the activity a, for $\alpha = 0.01$, $b = 0.2$ and $s = 20$ examples per concept. As in the case of the extremely dilute network [11], learning with examples of a small size yields a better generalization performance, if the threshold is not too large, than in the case of learning with large examples. The reason for this is that by lowering

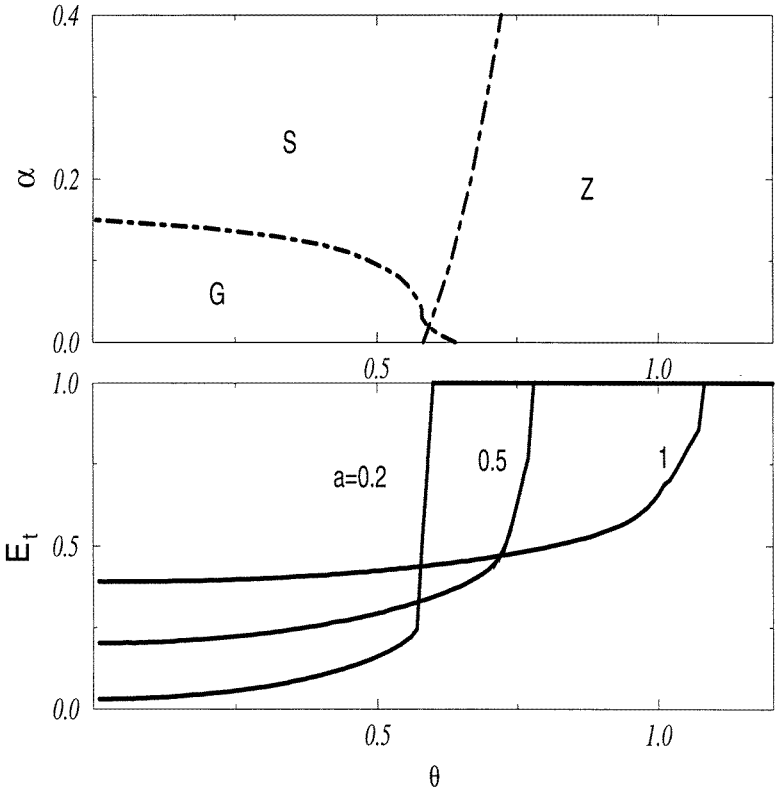

Figure 1. Generalization error E for $\alpha = 0.01$, correlation $b = 0.2$, $s = 20$ examples and various activities (below) and phase diagram for $a = 0.2 = b$ and the same s (above) as functions of the threshold θ for monotonic three-state neurons. G denotes the generalization phase, S the phase of self-sustained activity and Z the fully disordered phase.

the activity the noise term in the local field is more effectively suppressed than the main part favouring a recognition of the concepts by the state of the network. If the threshold is too large, however, the inactive part of the neurons becomes more important and the recognition of the concepts is harder. In the upper part of the figure we exhibit the phase diagram for the recognition ratio α as function of θ, for the same monotonic gain function with the values of $a = 0.2 = b$ and $s = 20$. Thus, the network manages to recognize the concepts only if α is below a critical α_c, beyond which the network remains active in the self-sustained phase S with dynamic activity, but fails to recognize the concepts. As one would expect, when the threshold is too large even the dynamical activity is suppressed and only the zero state, Z, remains. There is no chaotic phase in this case and the transitions between the various phases are discontinuous.

Next, we consider various forms of non-monotonic neurons with a transfer function that has some sort of a cut-off but no threshold. All of these yield irregular, in addition to fixed-point behaviour with a time-dependent generalization error E_t, even for long t. For the purpose of comparison, the results are presented in the following figures for $a = 0.2 = b$, $s = 20$ and the generalization curves $E_t(\theta)$ are shown for $\alpha = 0$, that is, for a finite number of concepts. This is to emphasize that, for the present model, the irregular behaviour is not due to a macroscopic number of concepts that the network attempts to recognize.

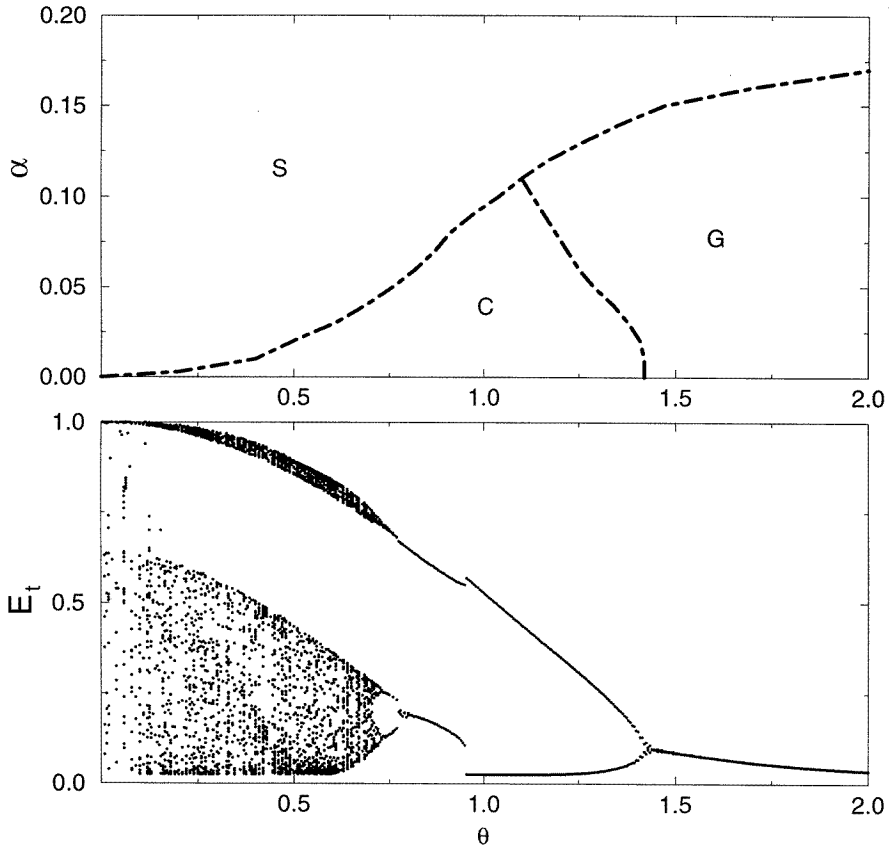

Figure 2. Generalization error E_t for $\alpha = 0$ (below) and phase diagram (above) as functions of θ, for the same values of the other parameters as in figure 1 for non-monotonic three-state neurons. C denotes the chaotic phase.

First, we deal with the three-state gain function equation (13). The drastic effect of the fatigue on the generalization ability is exhibited in the lower part of figure 2. When the threshold is large enough one expects to have regular behaviour with a time-independent generalization error, E, in which there is good generalization because the local fields are almost everywhere within the sigmoidal regime of the neurons. This region of fixed-point behaviour goes over into a periodic cycle-two attractor when θ is about 1.4, and the probability of the local field being lower than the threshold becomes important. When θ is about 0.9 a pair of discontinuities appear with a further period doubling and chaotic behaviour somewhat below. The interpretation for the first period doubling is that the network is in a waiting mode [18] in which the collective state of the neurons hesitates in a state between a clear and a poor recognition of the concepts. This state is characterized by a pair of finite overlaps with the concepts. Such hesitating behaviour is more visible for the other irregular states.

The phase diagram in the upper part of the figure, in which the transitions are discontinuous, exhibits the 'chaotic' phase (C) in which the irregular behaviour takes place. It is interesting to note that if the threshold is such that a network trained with the examples of a finite number of concepts ($\alpha = 0$) is in the waiting mode, training with the same number

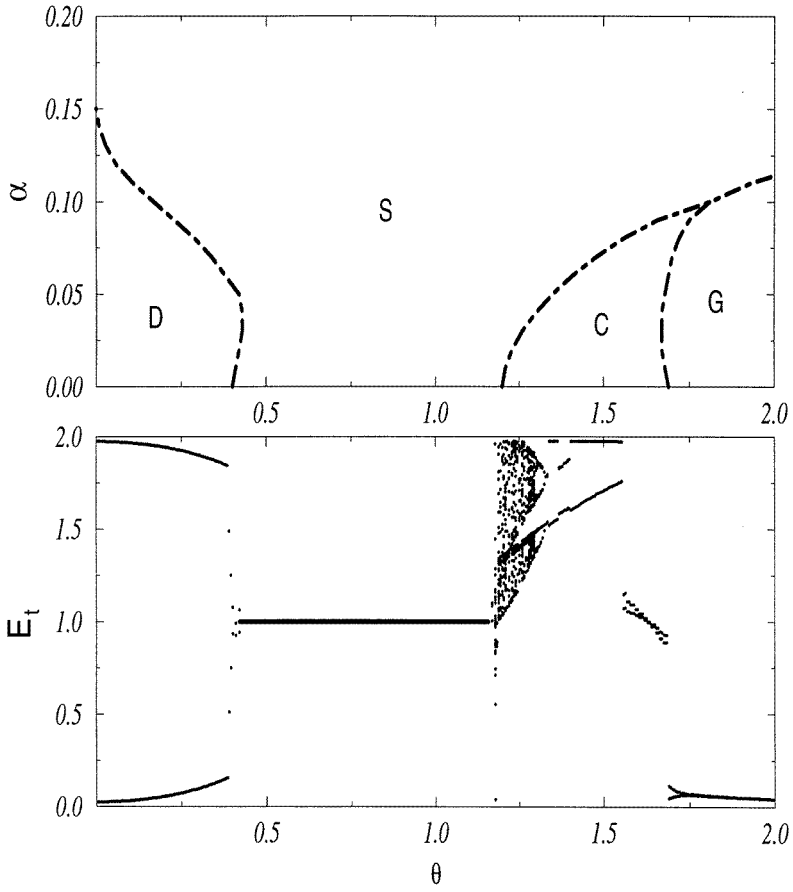

Figure 3. Generalization error (below) and phase diagram (above) for non-monotonic binary neurons with the same parameters as in figure 2. *D* is the period doubling phase.

of examples of a macroscopic number of the concepts, for which α is finite, may lead to a recognition of concepts in the generalization phase. This somewhat unexpected behaviour is illustrated in the figure by the re-entrance of the generalization into the *C* phase.

The case of the two-state non-monotonic transfer function, equation (14), is particularly interesting since θ inverts the sign of the neurons without an inversion of the sign in the local field, and the results are shown in figure 3. Consider first the lower part of the figure obtained with the initial conditions $m^s_{t=0} = q_{t=0} = r_{t=0} = 1$ for the dynamic relations. The strange behaviour can be qualitatively understood in the following manner. When the threshold for inversion is large enough, the network is expected to behave in a similar way to that for a sigmoidal gain function with good generalization. As this threshold is decreased, the local field may exceed θ and the network falls in a region of poor generalization, when $1.55 < \theta < 1.70$. The crossing at $E_t = 1.0$ ($M = 0$) when $\theta \sim 1.60$ indicates that the number of wrong neurons surpasses the correct ones and the sign inversion effect starts to take place. When $1.20 < \theta < 1.55$, the fluctuation of the local field around θ becomes important and a chaotic behaviour appears, which ends abruptly when the threshold goes below a value for which the state of the neurons starts to fluctuate randomly giving a vanishing average for the overlap with the concepts. Below $\theta \sim 0.40$, the neurons are

almost everywhere in the inverted regime and the response is each time the opposite of that at the previous time. Thus, a concept and its inverse are successively recognized. Note that the generalization is asymmetric until the number of wrong neurons matches the correct ones.

The phase diagram in the upper part of the figure shows the D phase where period doubling occurs, while the other phases are those described before. There is a slight re-entrance of the generalization phase into the chaotic phase which should persist for other values of a, b and s. Here again, the transitions are discontinuous.

The behaviour of the network with the usual $\tanh(h/\theta)$ analogue transfer function is qualitatively the same as that for three-state neurons, including the phase diagram of α in terms of θ, except that the phase transitions are continuous, so this case does not need to be discussed any further.

More interesting is the case of a non-monotonic analogue transfer function. The generalization performance for the function given by equation (15) is shown in figure 4. The reason for the behaviour that appears is the following. When θ is large, the small growth rate of the state of a neuron requires the build up of a large local field for an appropriate firing to recognize the concepts. This is unlikely and thus, the network fails to generalize. On the other hand, if θ is small, the growth rate becomes large and firing could take place for small values of the local field, but the cut-off also decreases. Thus, there can be at most an intermediate region of values of θ where one can have acceptable generalization, but this is spoiled again by the onset of a period doubling regime followed by a chaotic region for smaller values of θ. For large and small values of θ one expects the S phase to appear, and the G and C phases for intermediate θ. Note that there is again a re-entrance of the G into the C phase, as in the case of non-monotonic three-state neurons, figure 2, showing that training of the network with the examples of a sufficiently large number of concepts may be advantageous in restoring the generalization ability lost due to the presence of irregular behaviour. Thus, one can see that although the onset of the chaotic regime may differ in details between the case of discrete against continuous neurons, it is present in both of them.

6. Conclusions

We studied the generalization problem in a feedforward layered neural network in which a set of concepts is to be recognized when the network is only presented with examples in the training stage. A modified Hebbian learning rule has been used and we found that for non-monotonic neurons, in addition to fixed-point behaviour, irregular behaviour may appear with period doubling cascades ending in a chaotic regime. As the comparison of our results with the case of monotonic neurons shows, the irregular behaviour in this model is not due to the fact that the synaptic matrix is not symmetric and that the model has no Hamiltonian. It is expected, in general, that such models should exhibit a rich dynamical behaviour. The interest of our work for the understanding of neural network behaviour is not in the different structure of irregular attractors that may appear, which seems to depend on the form of the specific non-monotonic transfer function, but rather to point out the existence of this behaviour which seems to be there whenever the neurons are not monotonic. The interesting feature is the possibility of avoiding irregular behaviour for either multi-state or analogue non-monotonic transfer functions, by training the network with a fixed number of examples of a sufficiently large set of concepts, as shown in figures 2 and 4.

The work presented here generalizes previous results showing the presence of irregular attractors for the retrieval problem in a diluted network [18]. It has been proposed there that behaviour associated with such attractors means that the neurons are in a waiting mode in

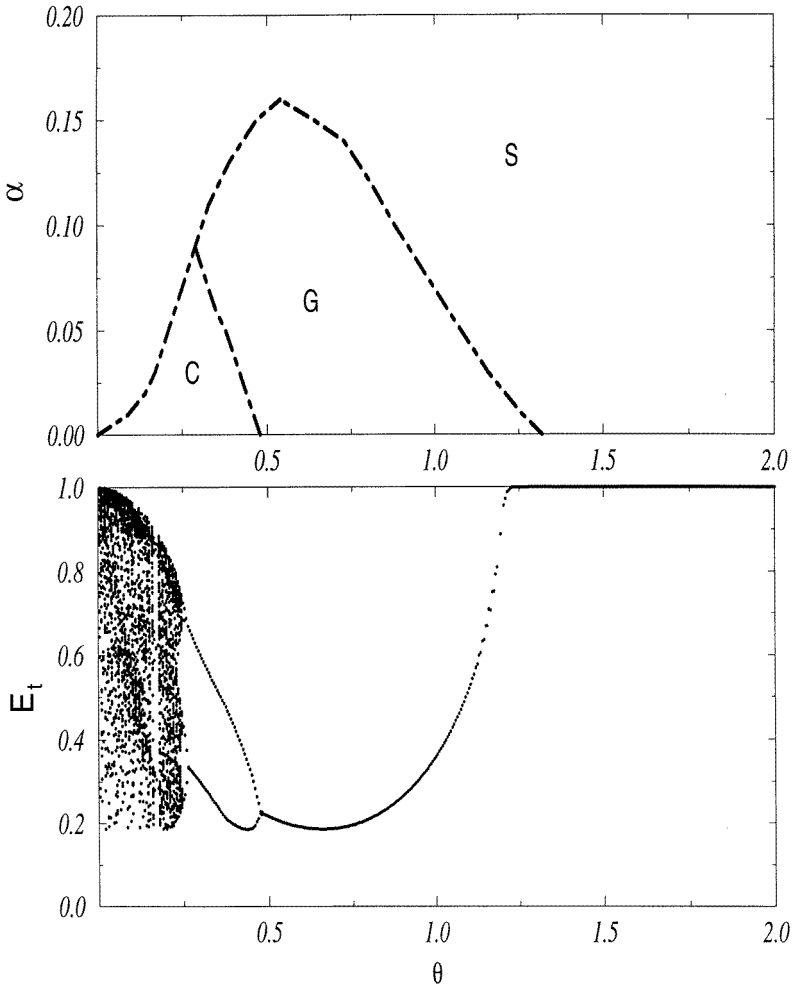

Figure 4. Generalization error (below) and phase diagram (above) for non-monotonic analogue neurons with the same parameters as in figure 2.

which the network fails to classify or disclassify the initial state as a condensed pattern taught to the network. In the context of our work, the presence of irregular attractors suggests the existence of a waiting mode in which the states of the network fail to recognize the concept used as a seed in the initial configuration. Some memory of this seed is preserved yet, because the waiting mode is characterized by a non-zero overlap with the concept. A waiting mode may be associated with big uncertainty, as illustrated by our results in figures 2 and 4 for the lower values of θ. There are other cases where this uncertainty may be greatly reduced as in the case of the linear analogue non-monotonic transfer function $F_\theta(h) = h/\theta$, for $|h| < \theta$ and zero otherwise. For this case we found a reduced dispersion of the overlap with the concepts in the regime of irregular behaviour and either generalization with a partial recognition of concepts is possible or else the S phase may be reached outside that regime.

Due to the architecture of the layered network considered here, the local fields are Gaussian random variables [13] and the dynamics can be solved exactly. It would be

interesting to investigate multi-state neural networks for the generalization problem in more complex architectures.

Acknowledgments

One of us (DRCD) is grateful for the hospitality of the Department of Theoretical Physics, of the Universidad Autónoma of Madrid, Spain, and thanks the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, for a postdoctoral fellowship. The research of WKT was supported in part by CNPq and FINEP (Financiadora de Estudos e Projetos), Brazil.

References

[1] György G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Koeberle (Singapore: World Scientific)
[2] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[3] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[4] Fontanari J F and Meir R 1989 *Phys. Rev.* A **40** 2806
[5] Fontanari J F 1990 *J. Physique* **51** 2421
[6] Miranda E 1991 *J. Physique* **1** 999
[7] Krebs P R and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** 398
[8] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007
[9] Meunier C, Hansel D and Verga A 1989 *J. Stat. Phys* **55** 859
[10] Yedidia J S 1989 *J. Phys. A: Math. Gen.* **22** 2265
[11] Dominguez D R C and Theumann W K 1996 *J. Phys. A: Math. Gen.* **29** 749
[12] Domany E and Meir R 1987 *Phys. Rev.* A **37** 608
[13] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[14] Kobayashi K 1991 *Network* **2** 237
[15] Gardner E 1987 *Europhys. Lett.* **4** 481
[16] Brunel N and Zecchina R 1994 *Phys. Rev.* E **49** R1823
[17] Meijilson I and Ruppin E 1994 *Network* **5** 277
[18] Bollé D and Vinck B 1996 *Physica* **223A** 293
[19] Eckmann J P and Ruelle D 1985 *Rev. Mod. Phys.* **57** 617
[20] Crisogono R, Tamarit F A, Lemke N, Arenzon J and Curado E 1995 *J. Phys. A: Math. Gen.* **28** 1593
[21] Martins J A and Theumann W K 1996 to be published
[22] Hertz J, Krogh A and Palmer R 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)
[23] Bouten M and Engel A 1993 *Phys. Rev.* E **47** 1397