Generalization in a multi-state neural network

1996 J. Phys. A: Math. Gen. 29 749

(http://iopscience.iop.org/0305-4470/29/4/006)

J. Phys. A: Math. Gen. **29** (1996) 749–761. Printed in the UK

Generalization in a multi-state neural network

D R C Dominguez† and W K Theumann‡

† Departamento de Física Teórica, Universidad Autónoma de Madrid, Canto Blanco, 28049 Madrid, Spain

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

Received 5 June 1995, in final form 2 November 1995

Abstract. The generalization ability of an extremely dilute feedback neural network with multi-state neurons is studied by means of a deterministic noiseless parallel dynamics. The overlap with any one of a macroscopic number of binary, full activity, concepts is determined when the network is trained with examples of variable activity according to a Hebbian learning algorithm that favours stable symmetric mixture states. Explicit results about the phase diagram and the generalization error are obtained for a network with three-state neurons which remain inactive below a threshold θ. It is shown that the generalization ability can be considerably enhanced either by training the network with low-activity examples or by means of a moderate increase in θ.

1. Introduction

The spontaneous emergence of features that were not originally built into a neural network during its learning stage, known as the generalization (or rule extraction) ability, has been a subject of much interest in recent years [1]. The categorization problem, in which individuals, or examples, are grouped into classes is a particular kind of generalization [2–4]. The process of creating a representation for concepts involves the extraction of common information from the activity patterns to which the network has been exposed during the learning stage and it is now known that, in attractor neural networks, this can be achieved through the presence of stable symmetric mixture states with the stored patterns [5–7].

Most of the works on generalization in attractor neural networks deal with the Hopfield model with two-state (firing or not firing) neurons. On the other hand, interesting features concerning the retrieval performance appear in networks with multi-state neurons that are active beyond a threshold [8–14]. In the case of training with full-activity patterns, a finite threshold below a certain limit tends to turn off those neurons which cause errors, allowing for a moderate increase in the storage capacity. When the network is trained with low-activity patterns, instead, in which there is a finite fraction of zero neurons, a careful selection of the threshold can lead to a drastic increase in the storage capacity [9]. It has also been found that such networks have strong inferential properties. Indeed, patterns of full activity, so-called *large* patterns, can implicitly be stored through the merging of low-activity (that is, *small*) prototype patterns, via the action of mixture states [8, 9]. On the other hand, since symmetric mixture states can be used to characterize the generalization phase, one may ask if the inferential properties of a network could not be used to enlarge this phase

0305-4470/96/040749+13$19.50

and to enhance the generalization ability. In its simplest form, this ability is determined by the generalization error associated with a given concept. This error is defined as the Hamming distance between the concept and the asymptotic state of the network.

Multi-state neurons are interesting from various points of view. One is the recognition of pictures from different levels of grey-toned patterns. Another is the biologically motivated analogous neural network with a continuous gain function [15–17]. On the other hand, discrete multi-state neurons could be of interest for hardware implementations.

It should be of considerable interest, therefore, to study the generalization ability of an attractor neural network with multi-state neurons. An attempt in that direction has been made recently for an analogous network that creates a representation for a single concept from a finite set of full activity patterns that act as examples [18]. Here we consider the more interesting situation in which a representation for a macroscopic number of concepts of full size is created in a network that learns from a set of examples of low activity [9]. Since we have to deal with a large parameter space, we restrict ourselves to a network in the extremely dilute limit which yields an exact dynamical description for all times or, alternatively, an approximate one-step dynamics for the fully connected network [19].

It will be shown that, in the case of the three-state network, the generalization ability can be considerably improved by training the network with examples of low activity in place of full activity. A slight gain in the performance can then be obtained through a small increase in the threshold. If the network is trained, instead, with examples of moderate-to-large activity, the generalization ability is considerably reduced, although an increase in the threshold can then lead to an improvement of the result.

The outline of the paper is the following. In section 2 we introduce the model and the relevant parameters; in section 3 we discuss briefly the dynamics for the generalization problem to make clear the distinction with the retrieval dynamics. We present there formal results for any transfer function and discuss the explicit results for the three-state case with either low or full activity examples in section 4. We end with concluding remarks in section 5.

2. The model

Consider a network of N neurons in the extremely dilute limit in which each neuron is connected, on average, with $C \ll \log N$ ($C \gg 1$) randomly chosen other neurons through the synaptic couplings J_{ij}, between neurons i and j. The parallel dynamics for this network can be solved exactly, in that the first time step describes the full evolution of the network for all later times [19].

We take the J_{ij} to be given by a generalized Hebbian rule:

$$J_{ij} = \frac{C_{ij}}{C} \sum_{\mu}^{p} \sum_{\rho}^{s} \xi_i^{\mu\rho} \xi_j^{\mu\rho} \tag{1}$$

where $\{C_{ij}\}$ is a set of random independent parameters that take the value 1 with probability C/N and zero with probability $1 - C/N$, in which C_{ij} is independent of C_{ji} and, thus, the J_{ij} are asymmetric. The network learns from a set $\{\xi_i^{\mu\rho}; \mu = 1, \ldots, p; \rho = 1, \ldots, s\}$ of random independent examples of each concept ξ_i^{μ} that takes the values ± 1, with equal probability, and we assume, for simplicity, that the concepts are uncorrelated. Specifically, we set

$$\xi_i^{\mu\rho} = \xi_i^{\mu} \lambda_i^{\mu\rho} \tag{2}$$

where $\lambda_i^{\mu\rho} = \pm 1$ or 0, on an active or passive site, respectively, belongs to a set of independent random microscopic activities with probability distribution

$$p(\lambda_i^{\mu\rho}) = \frac{a-b}{2}\delta(\lambda_i^{\mu\rho}+1) + (1-a)\delta(\lambda_i^{\mu\rho}) + \frac{a+b}{2}\delta(\lambda_i^{\mu\rho}-1) \tag{3}$$

with $a \geqslant b$, where δ is the Kronecker delta.

The first moment of $\lambda^{\mu\rho}$ gives the correlation

$$b \equiv \langle \xi_i^{\mu\rho}\xi_i^{\mu}\rangle \tag{4}$$

between an example and the concept to which it belongs, while the *activity* of an example,

$$a \equiv \frac{1}{N}\sum_i (\xi_i^{\mu\rho})^2 \tag{5}$$

is given by the second moment of $\lambda^{\mu\rho}$, in the $N \to \infty$ limit. Note that $N_e = aN$ may be regarded as the effective size of the learned examples, and we refer to these as *small* patterns whenever $a < 1$. It follows from equations (2) and (3) that the correlation between two examples of the same concept is given by $\langle \xi^{\mu\rho}\xi^{\mu\zeta}\rangle = b^2$, if $\rho \neq \zeta$, while examples of a given concept will not be useful in creating other concepts since $\langle \xi^{\nu}\xi^{\mu\rho}\rangle = 0$, if $\nu \neq \mu$.

Thus, the examples that are used in the training stage are biased with the concepts ξ^{μ} and the learning rule given by equation (1) is the most suitable and simple Hebbian-type rule that does not suppress the symmetric mixture states that characterize the generalization phase in our problem. This is in contrast with the retrieval problem, where the unwanted symmetric mixture states can be suppressed by the choice of an appropriately modified learning rule [20, 21].

The relevant parameters describing the performance of the network are the following [5, 22]. First, the overlap of the state $\sigma_i(t)$ with example $\xi_i^{\mu\rho}$ is defined as

$$m_N^{\mu\rho}(t) = \frac{1}{N}\sum_j \xi_j^{\mu\rho}\sigma_j(t) \tag{6}$$

which remains bounded between a and $-a$ as $\xi_j^{\mu\rho} = \pm\sigma_j$. Next, the generalization overlap with concept ξ^{μ} is given by

$$M_N^{\mu}(t) = \frac{1}{N}\sum_i \xi_i^{\mu}\sigma_i(t) \tag{7}$$

which yields the generalization error defined as

$$\epsilon^{\mu} = \frac{1}{2N}\sum_i |\xi_i^{\mu} - \sigma_i(t)| = (1 - M_N^{\mu})/2. \tag{8}$$

Finally, the *dynamical activity* [9]

$$Q(t) = \frac{1}{N}\sum_i [\sigma_i(t)]^2 \tag{9}$$

plays a crucial role in determining the generalization and chaotic phases, the latter being analogous to a spin-glass phase in a fully connected network. We allow for a finite, non-zero, capacity $\alpha = p/C$ of generated concepts, in the large-p and large-C limits, where C/N is the fraction of uncut synapses of the model and we restrict ourselves, for simplicity, to a noiseless (zero-temperature) parallel dynamics.

3. The dynamics

The zero temperature, parallel, dynamics of the network yields the state $\sigma_i(t+1)$ of neuron i at time $t+1$ through

$$\sigma_i(t+1) = F_\theta[h_i(t)] \tag{10}$$

where $F_\theta(x)$ is an odd transfer function of the local field

$$h_i(t) = \sum_{i \neq j} J_{ij}\sigma_j(t) \tag{11}$$

at time t. Here, θ could represent a series of thresholds for a discrete function or a gain parameter in the continuum case. For instance, $F_\theta(x)$ could be one of the multi-state functions considered recently [12–14] or a graded response function such as $\tanh(x/\theta)$ [15, 18]. Explicit results will be obtained for the three-state case, $\sigma_i = 1, 0, -1$, where [8–11]

$$F_\theta(x) = \begin{cases} \operatorname{sign}(x) & |x| > \theta \\ 0 & |x| \leqslant \theta. \end{cases} \tag{12}$$

Since the concepts are uncorrelated, we concentrate on the properties of concept $\mu = 1$. In the case of binary concepts $\xi_i^1 = \pm 1$, it is convenient to introduce the new state $\tau_i(t)$ and the new field $\Lambda_i(t)$,

$$\tau_i(t) = \xi_i^1\sigma_i(t) \qquad \Lambda_i(t) = \xi_i^1 h_i(t) \tag{13}$$

in terms of which the evolution of the states, equation (10), becomes $\tau(t+1) = F_\theta[\Lambda(t)]$ with the site-dependence being implicit here and in what follows.

In the large-N and large-C limits we may write

$$\Lambda(t) = \Omega(t) + \omega(t) \tag{14}$$

where

$$\Omega(t) = \sum_\rho \lambda^{1\rho} m^\rho(t) \tag{15}$$

is the part of the local field that favours ordering with the concept $\mu = 1$, in which $m^\rho(t) = \lim m_N^{1\rho}(t)$ as $N \to \infty$, is the overlap that characterizes the phase of interest. For the retrieval of a particular example, say $\rho = 1$, $m^\rho = m\delta_{\rho,1}$, whereas the overlap that characterizes the generalization phase is the symmetric one of s components defined below. The second term,

$$\omega(t) = \lim_{C,N\to\infty} \xi_i^1 \sum_{\nu>1,\rho,j\neq i} \frac{C_{ij}}{C} \xi_i^{\nu\rho}\xi_j^{\nu\rho}\sigma_j(t) \tag{16}$$

is the noise in the local field due to the presence of the other $p-1$ concepts. This noise will be finite whenever $\alpha = p/C \neq 0$.

The sum over sites in equation (6), for $\mu = 1$, becomes then in the limit $N \to \infty$, due to the law of large numbers,

$$m^\rho(t) = \langle\langle \lambda^{1\rho}\tau(t)\rangle_\Omega\rangle_\omega \tag{17}$$

where the brackets denote averages over the probability distributions for the components of the local field. It should be noted that, through its dependence on $\Lambda(t)$, the new state variable $\tau(t)$ is a function of the full set $\{\lambda^{1\delta}\}$, each member of which is correlated to $\lambda^{1\rho}$,

so that the average over the latter cannot be done separately from that over $\tau(t)$. Similarly, the generalization overlap becomes

$$M(t) = \lim_{N\to\infty} M_N^1(t) = \langle\langle \tau(t)\rangle_\Omega\rangle_\omega. \tag{18}$$

The fully symmetric overlap of the state of the network with the examples, given by

$$m^\rho(t)/b = m(t) \qquad \rho = 1, \ldots, s \tag{19}$$

characterizes the generalization phase and enables the network to extract the common features of the training examples in the generalization process. From equations (15) and (19) one finds that

$$\Omega(t) = \gamma m(t) x_s \tag{20}$$

in which $\gamma = sb^2$ and where x_s is (for $s \geqslant 10$) approximately a Gaussian random variable with mean $\langle x_s\rangle = 1$ and variance $\mathrm{Var}(x_s) = (a - b^2)/\gamma$, while

$$m(t) = \langle\langle x_s \tau(t)\rangle_{x_s}\rangle_\omega. \tag{21}$$

On the other hand, the noise term turns out to be given by

$$\omega(t) = z_p\sqrt{\alpha r Q(t)} \tag{22}$$

where $z_p \doteq N(0, 1)$ is distributed ($\doteq$) according to a Gaussian random variable with mean zero and unit variance. Here, $\alpha = p/C$ is the capacity of generated concepts, while

$$r = s[a^2 + (s-1)b^4] \tag{23}$$

and $Q(t)$ is the dynamical activity, given by

$$Q(t) = \langle\langle \tau^2(t)\rangle_{x_s}\rangle_\omega \tag{24}$$

in the large-N limit. Higher moments of ω are of order N^{-2} and, hence, vanishingly small in this limit.

Now summing both Gaussians in equations (20) and (22), the local field becomes

$$\Lambda(t) = \gamma m(t) + z V(t) \tag{25}$$

where $z \doteq N(0, 1)$ is also a Gaussian random variable with zero mean and unit variance, and

$$V(t) = [(a - b^2)\gamma m^2(t) + \alpha r Q(t)]^{1/2}. \tag{26}$$

The one-step recursion relations for the dynamics, for any odd transfer function $F_\theta(x)$, are then

$$m(t+1) = M(t+1) + m(t)(a - b^2)C(t+1) \tag{27}$$

where

$$M(t+1) = \langle F_\theta[\Lambda(t)]\rangle_z \tag{28}$$

and $C(t+1) = \langle F'_\theta[\Lambda(t)]\rangle_z$ is related to the spin-glass order, in which $F'_\theta(x) = \mathrm{d}F_\theta(x)/\mathrm{d}x$. The dynamical activity is given by

$$Q(t+1) = \langle\{F_\theta[\Lambda(t)]\}^2\rangle_z. \tag{29}$$

The fixed-point solutions of these equations solve exactly the dynamics of the network in the extremely dilute limit for all times after an initial step. Since the search for the fixed-point solutions becomes quite complex in the general case, in terms of a, b, s, θ and α as well as the dependence on the shape of $F_\theta(x)$, we specialize in the following to the three-state case.

4. Results: the three-state neuron

The fixed-point equations for $m = m(t)$, $M = M(t)$, $Q = Q(t)$ and $C(t) = C$ are, together with equation (27),

$$M = \tfrac{1}{2}[\mathrm{erf}(A_+/\sqrt{2}) + \mathrm{erf}(A_-/\sqrt{2})] \tag{30}$$

$$Q = 1 - \tfrac{1}{2}[\mathrm{erf}(A_+/\sqrt{2}) - \mathrm{erf}(A_-/\sqrt{2})] \tag{31}$$

and

$$C = \frac{1}{V}[\varphi(A_+) + \varphi(A_-)] \tag{32}$$

with $\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$,

$$\mathrm{erf}(x/\sqrt{2}) = 2\int_0^x \mathrm{d}z\, \varphi(z) \tag{33}$$

and

$$A_-^+ = \frac{1}{V}(m\gamma \pm \theta). \tag{34}$$

The dependence on the parameters a, b and s is implicit in V, the fixed-point value of $V(t)$.

At this point it is interesting to note the correspondence between our equations for the generalization problem, with the equations derived by Yedidia [9] for the retrieval problem in the extremely dilute network. This correspondence helps to understand the phase diagram for α as a function of θ that we obtain below. Indeed, as a and $b \to 0$ for $s \to \infty$, such that sa^2 and $\gamma = sb^2$ are finite, the dominant contribution to $V(t)$, equation (26), comes from the last term. Also, $m(t+1)$ in equation (27) reduces to $M(t+1)$ in this limit and we find that

$$A_-^+ = \frac{M \pm \hat{\theta}}{\sqrt{\hat{\alpha} Q}} \tag{35}$$

where $\hat{\theta} = \theta/\gamma$ and $\hat{\alpha} = \alpha[1 + sa^2/\gamma^2]$ can be viewed as the effective threshold and ratio of generated concepts, respectively. Our equations for the generalization overlap M with *binary* concepts and for Q become then the equations of Yedidia for the retrieval overlap m and the dynamical activity, when $a = 1$. For general values of the activity a, the correlation parameter b and the number of examples s, our equations are, however, more complex than those of the retrieval problem. The important point of this correspondence is that it illustrates that there should be generalization even if the activity and the correlation between examples are very small.

In order to obtain extensive results, we first reduce the parameter space, taking $a = b$. This is a compromise choice since an independently increasing activity a leads, in general, as will be seen later on, to a poorer generalization while the increase in the correlation parameter b provides an improvement. Noting that $\xi^{\mu\rho}$ still takes the values ± 1 *and* 0, we have the simplest possibility of exploring the generalization ability of the network when the training is with small examples, i.e. with low-activity patterns.

The solution of the fixed-point equations yields mainly three phases: a generalization (G) phase where $M > 0$ and $Q > 0$, a chaotic (C) phase with self-sustained activity in which $M = 0$ and $Q > 0$ and a zero (Z) or paramagnetic phase with $M = 0 = Q$. The chaotic phase is the analogue of a spin-glass phase in a fully connected network. There is also a retrieval phase for sufficiently small values of α, in which $M = O(b)$ while the

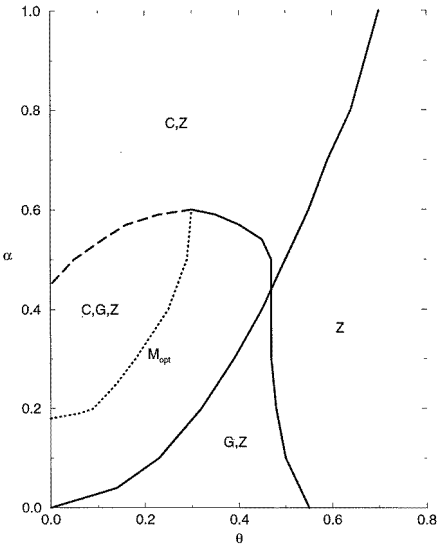

Figure 1. Zero-temperature low-activity phase diagram in the concept ratio α against threshold θ plane, for a three-state network with $s = 20$ examples and the activity $a = b = 0.2$, where b is the correlation between examples and a given concept. G, C, Z denote the generalization, the chaotic and the zero phases, respectively, defined in the text. Full curves represent first-order transitions, the broken curve represents a second-order transition and the dotted curve is the locus of the optimal generalization overlap $M_{\rm opt}$.

overlap with a single example, $m = O(1)$. This appears only within a small region of the phase diagram and, for simplicity, we shall not be concerned with it in what follows.

The three phases we are interested in are shown in figure 1, for a typical $s = 20$ and $b = 0.2$ when $a = b$. Each region is labelled with the possible phases that can appear depending on the initial state of the dynamical variables. The full curves represent first-order transitions, while the broken curve represents a second-order transition. The dotted curve, $M_{\rm opt}$, gives the threshold that optimizes the generalization overlap for a given α. There is a generalization phase within a finite range of threshold values and a limiting critical concept ratio $\alpha_{\rm c} = \alpha(\theta)$, beyond which generalization is not possible for a given number of examples and correlation b^2.

The shape of the phase diagram and the existence of the various phase-transition lines are similar to those in the retrieval problem [9], where a line for optimal retrieval has also been found. The main new feature of our phase diagram is the existence of a stable generalization phase in place of a retrieval phase over a considerable part of the phase diagram. The reason for this is the dynamical outcome of stable symmetric mixture states due to the training of the network with a sufficiently large number of correlated examples. The extraction of the common features of the examples through these mixture states yields now a finite overlap between the state of the network and each concept, if the random noise produced by the other concepts is not too large.

As one would expect, even when there is a generalization phase, there could be a competition for stability between this phase and the other ones, in particular with the chaotic phase, unless $\alpha(\theta)$ is small enough. This is the case below the lower first-order transition line where only the generalization phase is stable, besides the always possible Z phase. Thus, the generalization dynamics of the network trained only with the lowest level of a

set of correlated hierarchical patterns differs basically from the retrieval dynamics of the network trained with uncorrelated patterns, which yields a retrieval phase over a large part of the phase diagram.

To understand the role of the threshold, a distinction has to be made between moderately high and low values of α within the generalization phase. The line for $M_{\rm opt}$ in the phase diagram provides the boundary for this distinction. Consider first the case where $\alpha = 0.4$, say, which is close to the maximum ratio of concepts for the presence of the generalization phase when the threshold is set to zero. For such a high value of α, the neurons that have the lowest local fields are the most important ones in producing an error in the overlap with a given concept, due to the random noise produced by the other concepts. The effect of a finite threshold is to turn off those neurons and, eventually, an overlap on the optimal generalization curve, $M_{\rm opt}$, in figure 1 may be reached, with an appropriate value of the threshold. Increasingly higher values of the threshold should start to deteriorate the generalization ability of the network.

On the other hand, for $\alpha = 0.1$ say, i.e. well below the value where the line for $M_{\rm opt}$ starts, the local fields should be less sensitive to the random noise produced by the other concepts and the recognition of a concept may already be harmed by a small threshold.

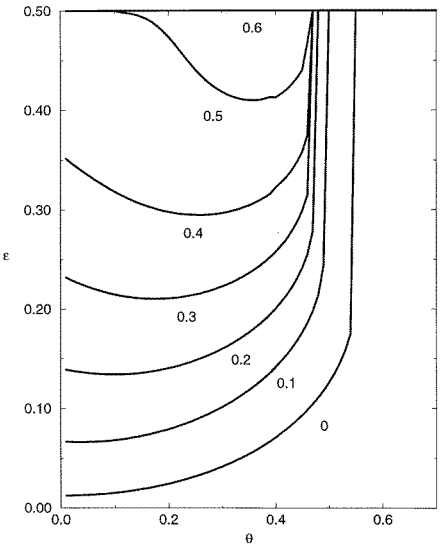

Figure 2. Generalization error ϵ as a function of θ for various values of α, with the same s, a and b as in figure 1. The minima in ϵ corespond to points on $M_{\rm opt}$.

To judge the dependence of the generalization quality on the threshold we show in figure 2 the generalization error ϵ, equation (8), for a given concept, with the same s and b as in figure 1. The minima in ϵ for each α correspond to points on the line for $M_{\rm opt}$. The best improvement due to the threshold is seen to appear for the larger values of α, close to the critical $\alpha_{\rm c}$.

In order to check if there is a significant and stable generalization phase, corresponding to a finite basin of attraction of the G fixed point, we determined the basins of attraction of the three fixed points in the M–Q plane, corresponding to the G, C and Z phases for typical values of the parameters, and found that there is, usually, a large region of attraction of the G phase within the region where the dynamical activity $Q \geqslant M$. Consistent with figure 2,

the G fixed-point appears well below $M = 1$, indicating a not too good generalization quality, unless α is small. We also found that the Z phase has a small basin of attraction and that the C fixed-point is a saddle-point, as in the retrieval problem for the three-state network [11].

The generalization ability of a fully connected network with binary neurons is known to depend crucially on the number of examples used in the training stage and on the correlation between examples [5–7]. The symmetric mixture states are stabilized and the generalization phase should appear either when the network has been exposed to a sufficiently large number of examples or when the correlation between examples is large enough. If the threshold θ is small one would expect a continuous drop in the generalization error and for larger values of the threshold eventually a critical number s_c of examples should be needed in order to have a rapid decrease in the generalization error. Our results for the so-called generalization curves in the extremely dilute three-state network, for $\alpha = 0.3$ and $b = 0.2$, shown in figure 3 for various values of θ, show that this is precisely the case. As increasingly higher thresholds tend to turn off larger portions of the active part of each neuron, a larger number of examples is needed in order to generalize, except in the vicinity of the optimal generalization surface in the (α, θ, s) space. This is the surface generated by the optimal generalization line, M_{opt}, in figure 1 for varying s. Indeed, the generalization curves decrease slightly when θ increases from zero to $\theta \simeq 0.2$ but, for clarity, this is not shown in the figure.

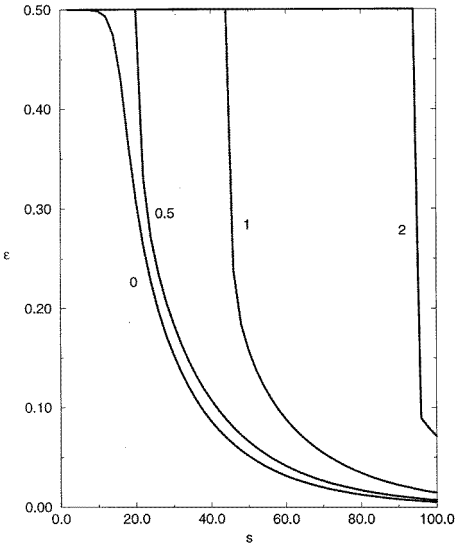

Figure 3. The generalization error ϵ as a function of the number of examples s, for various values of the threshold θ, when $\alpha = 0.3$ and $a = b = 0.2$.

A similar behaviour appears for the generalization error as a function of b (the activity or correlation parameter between examples and concepts) and, to emphasize this non-monotonic dependence on the threshold for the larger values of α, we show in figure 4 the results for $\alpha = 0.5$ and $s = 20$ for various values of θ. Although we have set $a = b$, we conclude on the basis of what will be discussed below that, unless the correlation between examples is large enough, there is no generalization due to the absence of stable symmetric mixture states. It will be shown that it is not necessary to have at the same time an increasingly large activity a. Note that the improvement with θ is not monotonic and that, within certain limits, it is possible to reduce the generalization error for a given value of b

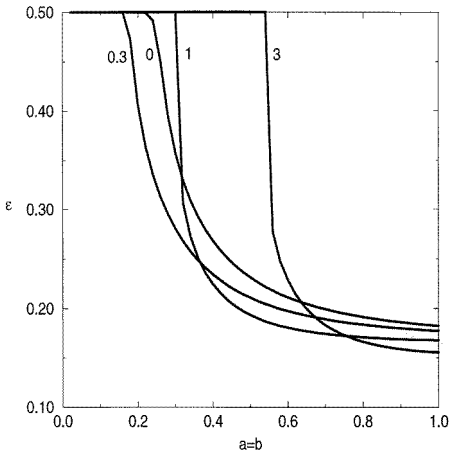

Figure 4. The generalization error ϵ as a function of $a = b$, for various values of the threshold θ, when $s = 20$ and $\alpha = 0.5$.

by increasing the threshold. Note also that as θ increases the transition to the generalization state becomes steeper and, for $\theta \geqslant 3$, it is practically a first-order transition.

So far, we discussed our results for a network in the extremely dilute limit trained with low-activity examples, in which $a = b \leqslant 1$. Two further aspects we consider next are, first, the comparison of the generalization ability with that of the network trained with full-activity examples, in which $a = 1$, and second, the dependence of ϵ on an independently varying activity unrelated to the correlation parameter b.

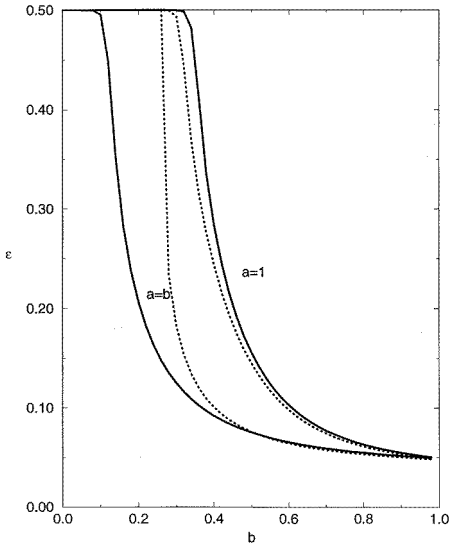

Figure 5. The generalization error ϵ as a function of the correlation b between examples and a given concept, for the 'low-activity' ($a = b$) and for the full-activity ($a = 1$) network, respectively, for two values of the threshold θ, when $s = 25$ and $\alpha = 0.3$. The full curves correspond to $\theta = 0$ (two-state neurons) and the dotted curves to $\theta = 1$.

The generalization error for $a = 1$ can be readily obtained and the comparison with the results for $a = b \leqslant 1$ is shown in figure 5, for $s = 25$, $\alpha = 0.3$ and two values of the threshold; $\theta = 0$ (i.e. two-state neurons), given by the full curves, and $\theta = 1$, given by the dotted curves. For the values of the parameters that have been chosen, the generalization error is already an increasing function of the threshold for the low-activity case, while it is still a decreasing function when $a = 1$. The behaviour of the latter reverts, however, already for $\theta \sim 2$ and the generalization error starts to increase thereafter. With these thresholds, and even for larger ones, we checked that for a given correlation between examples the generalization error for the low-activity case is always below that for $a = 1$.

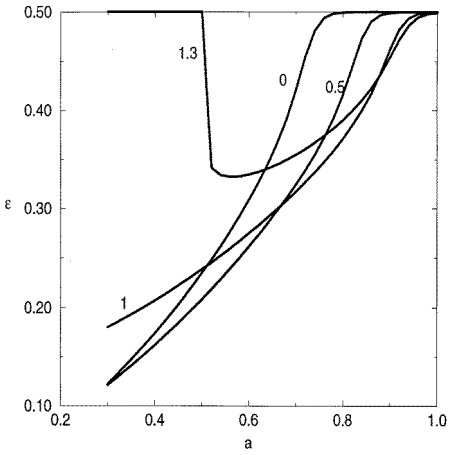

Figure 6. The generalization error ϵ as a function of the activity a of the training examples for fixed values of the threshold, as indicated, and $s = 25$, $b = 0.3$ and $\alpha = 0.3$.

To demonstrate that a much better performance can be obtained by training the network with low-activity examples, we show in figure 6 our results for $s = 25$, $b = 0.3$, $\alpha = 0.3$ and θ between zero and 1.3. The effect of a low activity is to decrease both the main part $\Omega(t)$ of the local field and the noise $\omega(t)$ term in equation (14). The low activity seems to be more efficient on the noise and the network has a quite good generalization ability, particularly for the lowest activities. For intermediate activities, the generalization ability becomes poorer although it is more robust to a moderate increase in the threshold, as one would expect. The generalization ability for the case of low activity is also sensitive to the number of examples presented to the network. If this is below a critical number, there is no generalization but it is quite good beyond that.

5. Concluding remarks

A neural-network model which is capable of inferring a representation for an extensive number of full-activity prototypes from a finite set of examples of small size that have been learnt with a generalized Hebbian rule, is studied in this paper in the extremely dilute limit. Recurrence relations for the overlaps with the examples and with the concepts are formally obtained for a general transfer function and worked out specifically for the case of three-state neurons. The simplest structure of a two-level hierarchy of patterns is used, of which only the lowest level is presented to the network in the training stage and the purpose is to recognize the patterns of the higher level. The choice of binary concepts in the higher

level is a natural one, in order to study the optimal ability of the network to build large patterns.

The possibility of using a threshold in a network with three-state neurons to stabilize the symmetric mixture states between a finite number of small patterns, in order to build up large patterns, discussed some time ago in the context of the retrieval problem, has been worked out here for the categorization problem. We have shown the interesting dependence of the generalization ability on the relevant parameters and found that it is advantageous to train a network with examples of low activity and to have a small-to-moderate threshold that cuts the lowest local fields which are responsible for the main error in the generalization process.

There are several interesting extensions of this work that may be considered. One is the role of graded response transfer functions $F_\theta(x)$ with a continuous gain parameter θ, which may be more suitable from a biological point of view. Graded response functions are, usually, monotonic but to appreciate their role it may be worth comparing the performance of such a network with that obtained using non-monotonic functions. Because of the relatively large number of parameters, we had to restrict the present work to a network in the extremely dilute limit and we considered the generalization process only in the absence of synaptic noise. Although we expect the latter to have only a damaging effect, it may still be worth exploring this explicitly.

More interesting perhaps, but more difficult, is the extension to a fully connected network which can have different behaviour from that of the network considered here [23]. An analytical treatment is expected to require replica symmetry breaking, in particular for low synaptic noise, since already the phase diagram for the generalization problem in the binary network presents a re-entrant behaviour that is believed to be related to the assumption of replica symmetry [7]. A recently worked out dynamics for a fully connected network can perhaps be extended to multi-state neurons [24].

Acknowledgments

One of us (DRCD) is grateful for the hospitality of the Department of Theoretical Physics, of the Universidad Autónoma of Madrid, where the work was completed, and would like to thank the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, for a postdoctoral fellowship first at the University of Rio Grande do Sul and then at the University of Madrid. The research of WKT was supported in part by CNPq and FINEP (Financiadora de Estudos e Projetos), Brazil.

References

[1] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
Seung H, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[2] Hertz J, Krogh A and Palmer R 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)
[3] Denker J, Schwartz D, Wittner B, Solla S, Howard R, Jackel L and Hopfield J J 1987 *Complex Syst.* **1** 877
[4] Patarnello S and Carnevali P 1987 *Europhys. Lett.* **4** 503
[5] Fontanari J F 1990 *J. Physique* **51** 2421
[6] Miranda E 1991 *J. Physique* I **1** 999
[7] Krebs P R and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** 3983
[8] Meunier C, Hansel D and Verga A 1989 *J. Stat. Phys.* **55** 859
[9] Yedidia J S 1989 *J. Phys. A: Math. Gen.* **22** 2265

[10] Rieger H 1990 *J. Phys. A: Math. Gen.* **23** L1273
[11] Bouten M and Engel A 1993 *Phys. Rev.* E **47** 1397
[12] Bollé D, Vinck B and Zagrebnov V A 1993 *J. Stat. Phys.* **70** 1099
[13] Bollé D, Shim G M, Vinck B and Zagrebnov V A 1994 *J. Stat. Phys.* **74** 565
[14] Bollé D, Rieger H and Shim G M 1994 *J. Phys. A: Math. Gen.* **27** 3411
[15] Marcus C M, Waugh F R and Westervelt R M 1990 *Phys. Rev.* A **41** 3355
[16] Kuhn R, Boss S and van Hemmen J L 1991 *Phys. Rev.* A **43** 2084
[17] Shiino M and Fukai T 1993 *J. Phys. A: Math. Gen.* **26** L831
[18] Stariolo D A and Tamarit F A 1992 *Phys. Rev.* A **46** 5249
[19] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[20] Fontanari J F and Theumann W K 1990 *J. Physique* **51** 375
[21] Bollé D and Huyghebaert J 1995 *Phys. Rev.* E **51** 732
[22] Fontanari J F and Meir R 1989 *Phys. Rev.* A **40** 2806
[23] Amit D J 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)
[24] Dominguez D R C and Theumann W K 1995 *J. Phys. A: Math. Gen.* **27** 63