Generalization in a diluted neural network

1995 J. Phys. A: Math. Gen. 28 1593

(http://iopscience.iop.org/0305-4470/28/6/014)

J. Phys. A: Math. Gen. **28** (1995) 1593–1602. Printed in the UK

Generalization in a diluted neural network

Crisógono R da Silva†‖, Francisco A Tamarit†, Ney Lemke‡, Jeferson J Arenzon‡ and Evaldo M F Curado§

† Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290 Rio de Janeiro, RJ, Brazil

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

§ International Centre of Condensed Matter Physics, CP 04667, 70919-900, Brasília, DF, Brazil and Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290 Rio de Janeiro, RJ, Brazil

Received 17 August 1994, in final form 3 January 1995

Abstract. We study the *generalization capability* of an extreme and asymmetrically diluted version of the Hopfield model through analytical and simulation techniques. Generalization is the ability of the system for grouping a given set of correlated patterns (the examples) in distinct classes (concepts), in such a way that each concept represents the common features of a set of examples and are attractors of the network dynamics. The dynamics of the system can be solved exactly and the generalization error for the long-time regime can be evaluated. As occur for the storage capacity, here it is shown that dilution improves the performance of the network as a categorization device when confronted with the fully connected Hopfield model.

1. Introduction

The modelling of the generalization ability of the brain has been one of the most active subjects in the field of neural networks and for a long time it was concentrated mainly in the context of feed-forward networks, where the task consists in training the net to infer an optimal rule that maps a set of inputs to a set of outputs. After having presented a given set of examples of the mapping to the net, it is desirable that the system would be able to associate any new input with the correct output. In the last few years, an increasing amount of work has also been devoted to extending the study of generalization abilities to attractor neural models, particularly, the Hopfield one [1–5].

The Hopfield model [6] was originally introduced in the context of pattern recognition and in spite of its importance as a precursor work, it is seriously limited as a recognition device. One of its most troublesome restrictions is due to the necessity of statistical independence among the stored configurations: any correlation generates a noise which drastically damages its recognition ability. After the pioneering work of Hopfield [6], a huge effort has been made to overcome this limitation, usually by introducing suitable, although more complex, modifications in the synaptic prescription [7]. Approaching this seemingly undesirable property from a new point of view, Fontanari [1] was able to show that, when correlated patterns are stored in the fully connected Hopfield model, in spite of loosing its retrieval capacity, it displays an ability to work as a categorizing machine.

‖ Permanent address: Departamento de Física, Universidade Federal de Alagoas.

0305-4470/95/061593+10$19.50

This means that, although the system may not recognize the stored patterns, it can decide to which category the input belongs.

Since the synaptic couplings in the Hopfield model are symmetric, the dynamics is ruled by an energy-like function from which the long-term behaviour of the system can be obtained [8]. This approach allows us to study both the recognition and the generalization abilities. Nevertheless, real neurons behave in a very different way: on average, each one is connected with 10^4 other neurons and these synaptic couplings are known to be asymmetric. When these two ingredients, *dilution* and *asymmetry*, are taken into account the dynamics of the system can easily be evaluated.

In this paper we analyse the generalization ability of an asymmetric diluted version of the Hopfield model introduced by Derrida *et al* [9] in the context of pattern recognition and for which the dynamics can be solved exactly. They proved that in this case the system is not only able to recognize the patterns, but also improves its storage capacity when compared with the fully connected version. Following [9] we extend the study of this model and show that, as occurs with the storage capacity, dilution and asymmetry improve the generalization ability.

In section 2 we introduce the model and describe the generalization task in the context of attractor neural networks. In section 3 we obtain a recurrent equation for the generalization error and in section 4 we analyse the limit of infinite connectivity. Numerical simulations are presented in section 5. Finally, in section 6 we discuss the main results.

2. The model

The model consists of N binary neurons, each one modelled by an Ising-like variable S_i which can take the values $\{-1, +1\}$, representing the passive and active states, respectively. The state of the network at time t is then given by an N-bit word $\{S_i\}$ whose evolution is ruled by synchronous heat bath dynamics

$$S_i(t+1) = \begin{cases} +1 & \text{with probability} \quad (1+\exp[-2\beta_0 h_i(t)])^{-1} \\ -1 & \text{with probability} \quad (1+\exp[+2\beta_0 h_i(t)])^{-1}. \end{cases} \tag{1}$$

The parameter β_0 (the inverse of the temperature T_0) measures the noise level of the net, $h_i(t)$ is the post-synaptic potential on neuron i at time t, defined by

$$h_i(t) = \sum_{j \neq i}^{N} J_{ij} S_j(t) \tag{2}$$

where J_{ij} is the synaptic matrix connecting the pre and post-synaptic neurons j and i and its elements have the following form:

$$J_{ij} = C_{ij} T_{ij} \tag{3}$$

where the C_{ij}s are random independent variables that control the asymmetric dilution of the synapses and are chosen according to the following distribution:

$$\rho(C_{ij}) = \frac{C}{N}\delta(C_{ij}-1) + \left(1-\frac{C}{N}\right)\delta(C_{ij}). \tag{4}$$

We assume that the synaptic efficacies of those connections that survive after dilution are given by the following Hebbian rule [1]:

$$T_{ij} = \sum_{\mu=1}^{p}\sum_{\nu=1}^{s} \xi_i^{\mu\nu}\xi_j^{\mu\nu} \tag{5}$$

where the $\xi_i^{\mu\nu} = \pm 1$ are random independent variables and the N-bit words $\{\xi_1^{\mu\nu}, \xi_2^{\mu\nu}, \ldots, \xi_N^{\mu\nu}\}$ stand for the sp stored configurations ($\nu = 1, 2, \ldots, s$ and $\mu = 1, 2, \ldots, p$). These memories are grouped in p categories (indicated by the letter μ), each one containing s configurations (indicated by the letter ν) that represent different elements or examples of the category. The probability distribution of these variables is given by

$$P(\xi_i^{\mu\nu}) = \tfrac{1}{2}(1 + \xi_i^{\mu} b)\delta(\xi_i^{\mu\nu} - 1) + \tfrac{1}{2}(1 - \xi_i^{\mu} b)\delta(\xi_i^{\mu\nu} + 1) \tag{6}$$

where the ξ_i^{μ} are random independent variables which can take the values ± 1 with the same probability and represent the p *concepts*. The parameter b measures the overlap (the similarity) between a concept and an example

$$\langle\langle \xi_i^{\mu\nu} \xi_j^{\gamma} \rangle\rangle = b\delta_{\mu\gamma}\delta_{ij} \tag{7}$$

and also between two examples:

$$\langle\langle \xi_i^{\mu\nu} \xi_j^{\gamma\rho} \rangle\rangle = \delta_{ij}\,\delta_{\mu\gamma}\left[b^2 + \delta_{\nu\rho}\,(1 - b^2)\right]. \tag{8}$$

The stored memories are then organized in a hierarchical tree.

Due to the asymmetry in the couplings the dynamics given by (1) does not allow the definition of an energy-like function, preventing a thermodynamical approach. We must then restrict ourselves to consider the time evolution of the network and, since it is not possible to describe this evolution on a microscopic level, we look for a recurrent relation for the relevant macroscopic parameters. Due to the effect of temporal correlation between different neurons, the time evolution was only solved for the fully connected Hopfield model in some limiting cases (a finite number of stored patterns or few initial time steps) until very recently, when a solution valid in all time-scales was proposed by Coolen and Sherrington [10]. Following a different approach, Derrida *et al* [9] have proved that in the limit $C \ll \ln N$ (extreme dilution) these temporal correlations are negligible and the quenched disorder can be treated as an annealed one, yielding the exact solution for the network dynamics. In the next section, following this approach, we solve the dynamics in the context of generalization.

3. The recurrent equations

Since the interest is in the generalization ability, let us suppose that the initial configuration represents an example of the first category and we look for a solution with macroscopic overlap only with the first concept and with the same overlap $m^{\mu\nu}$ with all the stored examples, that is $m^{1\nu}(0) = m_s$ for any ν and $m^{\mu\nu}(0) \sim \mathcal{O}(N^{-1/2})$ for $\mu > 1$, where

$$m^{\mu\nu}(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^{\mu\nu}\langle S_i(t)\rangle\,. \tag{9}$$

Here $\langle\cdots\rangle$ means both a thermal average at temperature T and over an ensemble of initial conditions. The categorization ability of the network is measured by the generalization error

$$\varepsilon = \frac{1 - m^1}{2} \tag{10}$$

where m^1 is the overlap between the first concept and the state of the network

$$m^1(t) = \frac{1}{N}\sum_{i}^{N} \xi_i^1\langle S_i(t)\rangle\,. \tag{11}$$

Note that ε is the fraction of bits in which $\{S_i(t)\}$ and $\{\xi_i^1\}$ differ. Here it is important to stress that this definition of the generalization error is not related to the one used in the context of learning theory, where it measures the ability of the system for extracting a rule from a given set of examples. Actually it would be better, in this paper, to use the term *categorization* instead of *generalization*, since our device associates a given example with a pre-existing category. Nevertheless, the word generalization has been widely used in all the related papers, so we prefer to mantain this terminology.

We then look for two functions f and g such that

$$m_s(t+1) = f(m_s(t)) \tag{12}$$

$$m^1(t+1) = g(m_s(t)) \tag{13}$$

from which we will extract the long-time behaviour of ε.

From the definitions (9) and (11) and after performing the thermal average we obtain

$$m_s(t+1) = \left\langle\!\!\left\langle \frac{1}{s}\sum_{\nu=1}^{s} \xi_i^{1\nu} \tanh\left[\beta_0 h_i(t)\right] \right\rangle\!\!\right\rangle \tag{14}$$

$$m^1(t+1) = \left\langle\!\!\left\langle \xi_i^1 \tanh[\beta_0 h_i(t)] \right\rangle\!\!\right\rangle \tag{15}$$

where $\langle\!\langle \ \ \rangle\!\rangle$ is an average over the patterns ξ_i^μ and $\xi_i^{\mu\nu}$. If after dilution the neuron i is connected to k_i other neurons, denoted by $\{j_1, j_2, \ldots, j_{k_i}\}$, the local field $h_i(t)$ can be written as:

$$\begin{aligned} h_i(t) &= \sum_{q=1}^{k_i}\sum_{\mu=1}^{p}\sum_{\nu=1}^{s} \xi_i^{\mu\nu}\xi_{j_q}^{\mu\nu} S_{j_q}(t) \\ &= \sum_{q=1}^{k_i}\sum_{\nu=1}^{s} \xi_i^{1\nu}\xi_{j_q}^{1\nu} S_{j_q}(t) + \sum_{\mu>1}^{p}\sum_{\nu=1}^{s} \xi_i^{\mu\nu}\sum_{q=1}^{k_i} \xi_{j_q}^{\mu\nu} S_{j_q}(t)\,. \end{aligned} \tag{16}$$

After averaging over the disorder, we obtain the following recurrent equations:

$$\begin{aligned} m_s(t+1) = \frac{1}{s\,2^{s+1}} &\sum_{k=0}^{\infty}\sum_{l=0}^{s}\sum_{n=0}^{k}\sum_{m=0}^{ks(p-1)} (s-2l)\,F^{+}_{kmn}(s,b) \\ &\times \tanh[\beta_0((s-2l)(k-2n)+ks(p-1)-2m)] \end{aligned} \tag{17}$$

$$\begin{aligned} m^1(t+1) = \frac{1}{2^{s+1}} &\sum_{k=0}^{\infty}\sum_{l=0}^{s}\sum_{n=0}^{k}\sum_{m=0}^{ks(p-1)} F^{-}_{kmn}(s,b) \\ &\times \tanh[\beta_0((s-2l)(k-2n)+ks(p-1)-2m)] \end{aligned} \tag{18}$$

where

$$\begin{aligned} F^{\pm}_{kmn} = \frac{C^k\,\mathrm{e}^{-C}}{k!\,2^{k(sp-s+1)}} &\binom{s}{l}\binom{k}{n}\binom{ks(p-1)}{m}\left[(1+m_s(t))^{k-n}\,(1-m_s(t))^{n}\right] \\ &\times\left[(1+b)^l\,(1-b)^{s-l} \pm (1-b)^l\,(1+b)^{s-l}\right]. \end{aligned} \tag{19}$$

Here we have taken the thermodynamic limit and averaged over all possible connectivities k_i of neuron i.

4. The limit of infinity connectivity

In the limit of infinity connectivity ($C \to \infty$ but $C \ll \ln N$) and $p \to \infty$, with constant $\alpha \equiv p/C$ and finite s, after some standard calculations [9] we obtain

$$m_s(t+1) = \frac{1}{s} \sum_{l=0}^{s} G_l^+(s,b)\,(s-2l) \int \mathrm{D}y \tanh\left[\beta((s-2l)m_s(t) + \sqrt{\alpha' s}\,y)\right] \tag{20}$$

and

$$m^1(t+1) = \sum_{l=0}^{s} G_l^-(s,b) \int \mathrm{D}y \tanh\left[\beta((s-2l)m_s(t) + \sqrt{\alpha' s}\,y)\right] \tag{21}$$

where $T \equiv \beta^{-1} = T_0/C$ is a reduced temperature, $\mathrm{D}y$ is given by

$$\mathrm{D}y = \frac{\mathrm{d}y}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2} \tag{22}$$

and α' is defined as

$$\alpha' = \alpha\left[1 + (s-1)b^4\right] \tag{23}$$

and

$$G_l^{\pm}(s,b) = \frac{1}{2^{s+1}} \binom{s}{l} \left[(1+b)^l (1-b)^{s-l} \pm (1-b)^l (1+b)^{s-l}\right]. \tag{24}$$

Note that for $s = 1$ and $b = 1$ we recover the expression obtained by Derrida *et al* [9] in the problem of pattern recognition.

It the limit $\alpha \to 0$, we recover the equations obtained for the fully connected Hopfield model studied in [1] for the case $T = 0$ and in [5] for the case $T \neq 0$. Notice that although their equations look different, the difference stems from the chosen ansatz: the fully symmetric one (in this paper) and the asymmetric one in their case.

We focus on the problem of having an *extensive* number of concepts ($\alpha > 0$) while the system is presented with a finite number of examples and starts analysing the noiseless limit $T = 0$. Figure 1 shows ε as a function of s for $b = 0.2$ and several values of α/α_0, where $\alpha_0 = 2/\pi$ is the critical value above which the system does not recognize uncorrelated patterns [9]. For small values of α the generalization error is a monotonically decreasing function of s. As α increases this behaviour changes and ε presents a maximum for $s > 1$. For $\alpha > \alpha_p$ a plateau emerges with $\varepsilon = 0.5$ for $s^- < s < s^+$ and at s^+ the system undergoes a smooth transition to a generalization phase. Assuming that $s^+ \gg 1$, the binomial distribution of z_s can be approximated by a gaussian distribution, and after some simple calculations we obtain

$$m_s(t+1) = b\,\mathrm{erf}\left(\frac{\Delta}{\sqrt{2}}\right) + \sqrt{\frac{2}{\pi}}\,\frac{(1-b^2)\Delta}{bs} \exp\left(-\frac{\Delta^2}{2}\right) \tag{25}$$

$$m^1(t+1) = \mathrm{erf}\left(\frac{\Delta}{\sqrt{2}}\right) \tag{26}$$

where

$$\Delta = \frac{b\sqrt{s}\,m_s(t)}{[\alpha' + m_s^2(t)(1-b^2)]^{1/2}}. \tag{27}$$

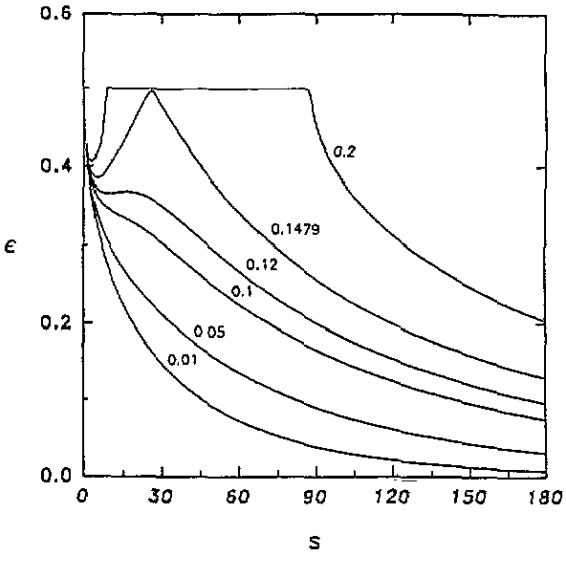

Figure 1. The generalization error ε as function of the number of examples s for $T = 0$, $b = 0.2$ and $\alpha/\alpha_0 = 0.01$, 0.05, 0.1, 0.12, 0.1479 and 0.2.

Since the transition is continuous, for values of s close to but greater than s^+ we know that $m_s \ll 1$. Expanding the former equations we obtain the following relation among the parameters α, b and s at the critical point:

$$\alpha = \frac{2[1 + b^2(s-1)]^2}{\pi s[1 + (s-1)b^4]}. \tag{28}$$

Here it can be seen how the performance of the system as a generalization device strongly depends on the correlation b. Observe that for $s = 1$ we recover the critical value $\alpha_0 = 2/\pi$ obtained in [9]. From equation (28) it is possible to get the following expression for s^- and s^+:

$$s^{\pm} = -\frac{\alpha_0(1-b^2)}{b^2(\alpha_0 - \alpha)} + \frac{\alpha(1-b^4)}{2b^4(\alpha_0 - \alpha)}\left[1 \pm \left(1 - \frac{4\alpha_0 b^2}{\alpha(1+b^2)^2}\right)^{1/2}\right]. \tag{29}$$

The value α_p at which a plateau emerges is then given by

$$\alpha_p = \frac{8b^2}{\pi(1+b^2)^2}. \tag{30}$$

At $\alpha = \alpha_p$ the generalization error reaches the value 0.5 at a unique point

$$s_p = \frac{1+b^2}{b^2} \tag{31}$$

and for $\alpha < \alpha_p$ it never does attain the plateau. In figure 1, with $b = 0.2$ we obtain $s_p = 26$ and $\alpha_p/\alpha_0 \approx 0.148$.

For $s \gg 1$, $m^1 \to 1$ and $m_s \to b$ and ε decreases exponentially as

$$\varepsilon = \frac{1}{2(\pi\Delta)^{1/2}} \exp\left(-\frac{\Delta^2}{2}\right) \tag{32}$$

where Δ is given by (27). Observe that this equation implies that for all values of α there is always a value of s great enough above which the system starts generalizing.

When the above results are confronted with those obtained for the fully connected Hopfield model at $T = 0$ [1], one observes some novel properties. The transition to the generalization phase is always continuous for the diluted version but it is discontinuous in

Table 1. The generalization error ε for $b = 0.4$, $\alpha = 0.028$ and different values of s, both for the fully connected and for the extremely diluted versions of the Hopfield model.

s	Fully connected	Diluted
10	0.500	0.1046
20	0.500	0.0321
30	0.032	0.0123
40	0.021	4×10^{-3}
50	0.017	9×10^{-4}

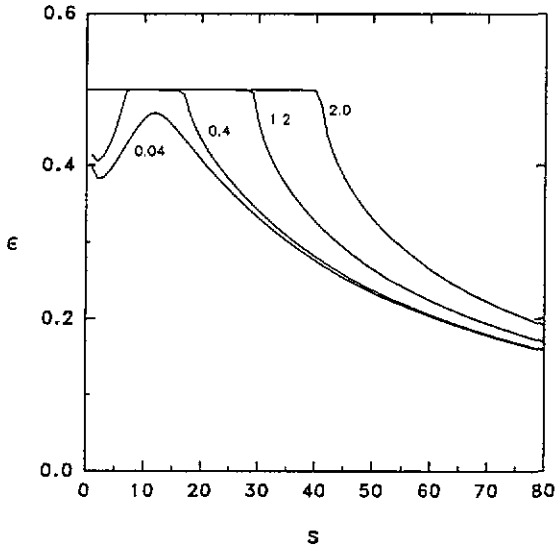

Figure 2. The generalization error ε as function of the number of examples s for $b = 0.3$, $\alpha/\alpha_0 = 0.3$ and $T = 0.04$, 0.4, 1.2 and 2.0.

the fully connected case. For small values of s (few examples) the behaviour of the error ε as a function of s is completely different in the two limits. Depending on the value of b and α, ε can start increasing until it attains a maximum or even a plateau regime.

In comparing the values of α in these papers, it is important to emphasize that, although their definitions are different ($\alpha = p/C$ in our case and $\alpha = p/N$ in [1]), they measure the same quantity, actually, the number of bits of stored concepts (pN) in units of the number of calculated couplings (CN and N^2, respectively). Note also that the results are all presented in different units of α ($\alpha_0 = 2/\pi \sim 0.64$ for the diluted version and 0.138 for the fully connected case). In order to clarify this point, in table 1 we present the values of ϵ for the two models, obtained for $\alpha = 0.028$, $b = 0.4$ and for different values of s. For these values, the fully connected model generalizes at $s_c = 14$ with a generalization error $\varepsilon \sim 0.09$, while the diluted version attains the same value at $s_c = 9$. We observe that dilution improves the performance of the model as a generalization device.

Next we present the results for α and $T \neq 0$. Figure 2 displays ε versus s for $\alpha/\alpha_0 = 0.3$, $b = 0.3$ and several values of T. The temperature T seems to play the same role as the parameter α in the noiseless case. As T increases, ε goes from a monotonically decreasing to a plateau-like behaviour. In all cases analysed in this paper we found that temperature damages the generalization ability of the system: as T increases, a greater number of examples must be presented to the net to start generalization, similarly to fully connected Hopfield networks [5]. Here again we find that dilution improves the capacity of the

network. Finally, in figure 3 we show the critical lines T versus α/α_0 below which the system is able to generalize with only 50 examples and with an error ε less than 0.05, for $b = 0.3$ and $b = 0.4$. As expected, as b increases the generalization phase also increases.

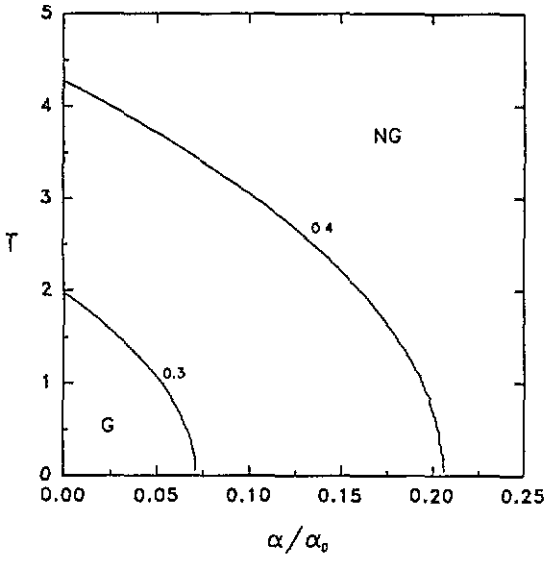

Figure 3. The critical lines T versus α/α_0 separating the generalization phase (G) and the non-generalization phase (NG) for 50 examples, $b = 0.3$ and $b = 0.4$. The lines correspond to an error $\varepsilon = 0.05$.

5. Numerical simulations

In this section we compare the analytical and simulation results. Although it seems that the limit of extreme dilution discussed in this paper would be hard to implement in a computer (because the conditions $C \ll \ln N$ and $C \gg 1$ require a huge number of neurons, greater than any previously presented in the literature), in a recent work, Arenzon and Lemke [11] reproduced numerically the results obtained by Derrida *et al* [9] working with large networks (up to $N = 16\,000$) and finite, fixed connectivity ($C = 20$). Note that the extremely diluted version requires a significatively smaller numerical effort than the fully connected version: instead of storing or calculating a complete $N \times N$ coupling matrix, one need only store two $N \times C$ matrices, one to store the neighbours of a given neuron and the other for the values of each non-zero synaptic connection.

Following these ideas, we also studied numerically the generalization ability of the model for $T = 0$. The simulation was performed for $\alpha/\alpha_0 = 0.12$ and 0.20, and the sizes used were $N = 10\,000$ and $20\,000$. The number of concepts was always $p = 2$ and the connectivity was $C = 26$ for $\alpha/\alpha_0 = 0.12$ and $C = 16$ for $\alpha/\alpha_0 = 0.20$.

The algorithm used is the following: initially a set of p uncorrelated concepts is created. For each concept, one example ($s = 1$) is generated and stored in the network. The initial state is chosen as the first concept and the network is updated synchronously during a fixed number of time units, until the fluctuations in the overlap m^{11} are negligible (the system may not achieve a fixed point). Then, a temporal average of ε is calculated over 50 time units. Next, another example is created ($s \to s + 1$) and the same steps are repeated. In order to make a configurational average, this procedure is repeated for 20 different samples, using different concepts, examples and initial configurations.

In figures 4 and 5 we present the results for $\alpha/\alpha_0 = 0.12$ and 0.20, respectively.

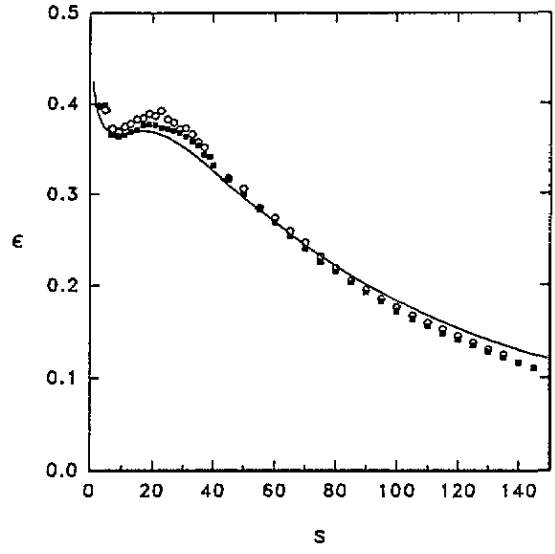

Figure 4. The generalization error ε versus the number of examples s for $b = 0.2$ and $\alpha/\alpha_0 = 0.12$. The full curves correspond to the analytical solutions. For the simulations we used $p = 2$, $C = 26$ and $N = 10000$ (open circles) and 20000 (full squares).

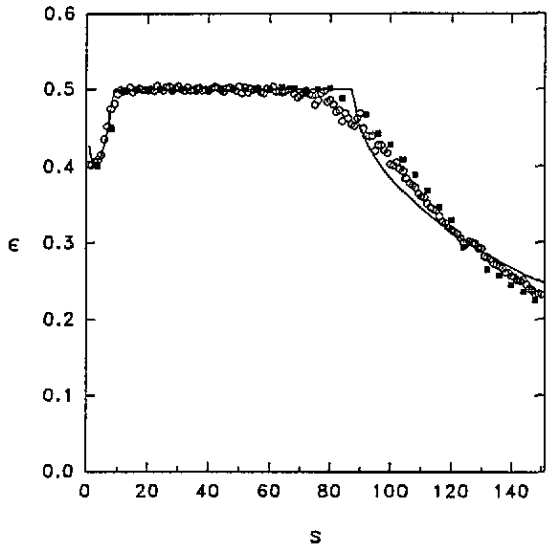

Figure 5. The generalization error ε versus the number of examples s for $b = 0.2$ and $\alpha/\alpha_0 = 0.2$. The full curves correspond to the analytical solutions. For the simulations we used $p = 2$, $C = 16$ and $N = 10000$ (open circles) and 20000 (full squares).

The full curves are the solutions of (20) and (21) (valid for $C \to \infty$). Note that for small values of s the simulations reproduce the analytical results very well, even for small systems. As s increases, the behaviour displays a sensitive dependence on the size of the system. For $s \gg 1$ we observe a systematic but small deviation from the theoretical data, and this behaviour does not seem to be due to finite size effects. We believe that this discrepancy is due to the fact that in this region the simulations were done for $s \gg p$, while the analytical results are only valid for $p \gg s$. Anyway, note that the critical values of s^+ obtained in the simulations for α/α_0 tends to the exact value $s^+ = 88$ as N increases.

6. Conclusions

In this paper we studied the generalization ability of an extremely diluted version of the Hopfield model, by training it with examples of a given set of concepts and showed that after a certain number of examples have been taught, the system is able to generalize, that is, to create an attractor of the dynamics with a macroscopic overlap with the concept. The long-time behaviour of the generalization error was obtained through the exact equation for the time evolution of the overlap m^1 between the state of the system at time t and one of the concepts.

Summarizing our results, it can be said that, besides being more realistic from a biological point of view, the diluted model presents two main features as a generalization device. First, it generalizes better than the fully connected version of the model, i.e. it needs a smaller number of examples in order to correctly classify the inputs into categories. This behaviour is observed both for noisy and noiseless dynamics and for $\alpha = 0$ and $\alpha \neq 0$. Second, the results obtained in this work, although involving the limit $C \ll \ln N$, can be easily reproduced by numerical simulations, and with a considerably lower computational cost. We also found other interesting behaviours in the diluted Hopfield model: for $T = 0$ the system generalizes for *any* value of α and the inclusion of noise damages its performance as a generalization device (for a given α, b and s, $\varepsilon(T_1) \leqslant \varepsilon(T_2)$ if $T_1 \leqslant T_2$). It must be emphasized that this behaviour is a consequence of the chosen symmetric ansatz: if an asymmetric solution were considered, for $\alpha = 0$ and some range of T, the system would start generalizing earlier for higher temperatures. Also, for small values of s and depending on the value of α the curve $\varepsilon \times s$ displays a different behaviour: after a brief decreasing, it starts monotonically increasing as a function of s, until it attains a maximum or a plateau, from which falls down exponentially into a generalization phase.

Acknowledgments

We thank to D A Stariolo and R M C de Almeida for fruitful discussions and the Brazilian agencies CNPq, FINEP and FAPERGS for partial support.

References

[1] Fontanari J F 1990 *J. Physique I* **51** 2412
[2] Miranda E N 1991 *J. Physique I* **1** 999
[3] Stariolo D A and Tamarit F A 1992 *Phys. Rev.* A **43** 5249
[4] Branchtein M C and Arenzon J J 1992 *J. Physique I* **2** 2019
[5] Krebs P R and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** 3983
[6] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 91
[7] Parga N and Virasoro J 1986 *J. Phys.* **47** 1857
Gutfreund H 1988 *Phys. Rev.* A **37** 570
Personnaz L, Guyon I and Dreyfus G 1985 *J. Phys. Lett.* **46** L359
Feigelman M and Ioffe L 1987 *Int. J. Mod. Phys.* B **1** 51
Tamarit F A and Curado E M F 1991 *J. Stat. Phys.* **62** 473
[8] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007; 1987 *Ann. Phys., NY* **173** 30
[9] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[10] Coolen A C C and Sherrington D 1994 *Phys. Rev.* E **49** 1921
[11] Arenzon J J and Lemke N 1994 *J. Phys. A: Math. Gen.* at press