Generalization in a Hopfield network

J.F. Fontanari

J. Phys. France **51** (1990) 2421-2430 1er NOVEMBRE **1990,** PAGE 2421

Classification
Physics Abstracts
87.10 — 64.60C

Generalization in a Hopfield network

J. F. Fontanari

Instituto de Física e Química de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560 São Carlos SP, Brasil

(*Received 23 April 1990, accepted in final form 12 July 1990*)

Abstract. — The performance of a Hopfield network in learning an extensive number of concepts having access only to a finite supply of typical data which exemplify the concepts is studied. The minimal number of examples which must be taught to the network in order it starts to create representations for the concepts is calculated analitically. It is shown that the mixture states play a crucial role in the creation of these representations.

1. Introduction

Learning and generalization in neural networks has been the subject of intensive research in the past few years [1-7]. The most recent studies have been carried out in the context of supervised learning in single-layer feedforward neural networks [5-7], following the theoretical framework presented in Gardner's seminal paper [8]. Comparatively, little attention has been devoted to the ability of simple feedback neural networks, e.g. Hopfield's model [9], to perform computational tasks beyond the simple storage of a set of activity patterns.

The learning process in Hopfield's model consists of setting the value of the coupling J_{ij} between neurons i and j for all pairs of neurons such that a given set of activity patterns is memorized by the network. In this model the states of the neurons are represented by Ising spins, $S_i = +1$ (firing) or $S_i = -1$ (rest). Storage of an activity pattern $\{\xi_i^\mu = \pm 1, i = 1, ..., N\}$ into the memory of the network is achieved by modifying the couplings according to the generalized Hebb rule

$$\Delta J_{ij} = \frac{1}{N} \xi_i^\mu \xi_j^\mu \qquad i \neq j \,. \tag{1.1}$$

After being exposed to p activity patterns the couplings are set to

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \qquad i \neq j \tag{1.2}$$

where we have assumed $J_{ij} = 0$ initially (tabula rasa).

Once the couplings are fixed, the dynamical retrieval process is governed by the Hamiltonian [9]

$$H = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} J_{ij} S_i S_j . \tag{1.3}$$

The network can potentially retrieve a given activity pattern if it is a minimum or if it is very near a minimum of H. The natural parameters for measuring the performance of the network in retrieving the stored patterns are the overlaps

$$m^{\mu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} S_i \qquad \mu = 1, ..., p \tag{1.4}$$

where the state $\{S_i, i = 1, ..., N\}$ is a minimum of H. The properties of these minima have been fully studied by Amit *et al.* [10, 11] using statistical mechanics tools developed in the analysis of infinite range spin-glasses [12]. It has been shown that besides the retrieval states which have macroscopic overlaps, $O(1)$, with only one stored pattern there exist mixture states which have macroscopic overlaps with several stored patterns [10]. Since the interest in Hopfield's model is mainly due to its prospective use as an associative memory, the attention has been focused on the retrieval states, the mixtures states being regarded as a nuisance which can be eliminated either by adding external noise to the system [10] or by modifying the learning rule [13]. On the other hand, these spurious states have been seen as proof of the ability of the network to create new representations to handle the information contained in the stored patterns [14]. In this paper we show that the mixture states play a crucial role when the task posed to the network is to extract meaningful information from the activity patterns it is exposed to during the learning stage.

We consider the following problem. Let us suppose that during the learning stage the network is exposed to s examples of a given concept. The examples are embedded in the memory of the network by the Hebbian learning process, equation (1.2). The question we address is whether the network can create a representation for the concept to which it had been exposed only through examples. We say that the network has a representation for a concept if the concept is a minimum or if it is very near a minimum of H. More specifically, we consider $p = \alpha N$ concepts represented by the activity patterns $\{\xi_i^{\mu}\}$, $\mu = 1, ..., p$. For each concept, a finite number of examples $\{\xi_i^{\mu\nu}\}$, $\nu = 1, ..., s$ is generated. Their components are statistically independent random variables drawn from the distribution

$$P(\xi_i^{\mu\nu}) = \frac{1}{2}(1 + \xi_i^{\mu} b)\, \delta(\xi_i^{\mu\nu} - 1) + \frac{1}{2}(1 - \xi_i^{\mu} b)\, \delta(\xi_i^{\mu\nu} + 1) \tag{1.5}$$

with $0 \leqslant b \leqslant 1$. The examples can be thought of as noisy versions of the concepts they exemplify. The parameter b measures the difficulty of the task posed to the network : Small values of b result in low correlations between examples and concepts, making the grasping of the concepts more difficult. For simplicity, the components of the concepts are randomly chosen as ± 1 with equal probability.

As an alternative viewpoint, one may consider the concepts as defining p classes each one containing s individuals (examples). Thus, the task of the network would be to group the examples in their respective classes, i.e. the network should categorize the examples. However, in this paper we follow the point of view expressed in Denker *et al.* [1] that categorization is a particular case of rule extraction (generalization) in which the rule may be roughly described by « nearby inputs should produce nearby outputs ». We return to this issue in the conclusion of the paper.

Finished the learning stage, the couplings are set to

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \sum_{\nu=1}^{s} \xi_i^{\mu\nu} \xi_j^{\mu\nu} \qquad i \neq j \,. \tag{1.6}$$

The quantities we focus on are the generalization errors ε^μ defined by

$$\varepsilon^\mu = \frac{1 - m^\mu}{2} \qquad \mu = 1, ..., p \tag{1.7}$$

where m^μ, the generalization overlaps, are the overlaps between the concepts and a certain minimum of H which will be specified later.

The aim of this paper is to calculate the dependence of the generalization error (ε^μ) on the number of examples taught to the network (s) for a given task characterized by the parameter b. To achieve this we study the thermodynamics of Hopfield's Hamiltonian, equation (1.3), with the couplings set as in equation (1.6). Our analysis is restricted to the noiseless (zero temperature) limit. A first attempt to tackle this problem has been published recently [15]. This paper presents a simpler and more general approach.

The paper is organized as follows. In section 2 we study the thermodynamics of the model in the limit $\alpha = p/N \to 0$. The simplicity of this limit, which dispenses the use of the replica trick in the computation of the averaged free energy density, allows us to find analytical expressions relating ε^μ with s and b. The analysis of the nonzero α limit is performed within the replica symmetric framework in section 3. We summarize our results and present some concluding remarks in section 4.

2. Finite number of concepts.

The Hamiltonian of Hopfield's model with the couplings given by equation (1.6) can be written as

$$H = -\frac{1}{2N} \sum_{\mu\nu} \left(\sum_i \xi_i^{\mu\nu} S_i \right)^2 + \sum_{i\mu} h^\mu \xi_i^\mu S_i \tag{2.1}$$

where we have omitted a constant term and included an additional term in order to compute m^μ. In fact, writing the averaged free energy density as

$$f = -\lim_{N \to \infty} \frac{1}{N\beta} \langle \ln Z \rangle \tag{2.2}$$

where $Z = \mathrm{Tr}_s \, \mathrm{e}^{-\beta H}$ one has

$$m^\mu = \frac{\partial f}{\partial h^\mu} \bigg|_{h^\mu = 0} . \tag{2.3}$$

The parameter $\beta \equiv T^{-1}$ in the expression for the partition function Z is a measure of the amount of noise acting on the system. The noiseless limit is obtained by taking $\beta \to \infty$. The notation $\langle ... \rangle$ stands for the averages over the examples and over the concepts taken in this order. The calculation of f is straightforward in the limit $\alpha = 0$ [10] so we present only the final result

$$f = \frac{1}{2} \sum_{\mu\nu} (m^{\mu\nu})^2 - \beta^{-1} \left\langle \ln 2 \cosh \left[\beta \left(\sum_{\mu\nu} m^{\mu\nu} \xi^{\mu\nu} - \sum_\mu h^\mu \xi^\mu \right) \right] \right\rangle . \tag{2.4}$$

The order parameters,

$$m^{\mu\nu} = \langle \xi^{\mu\nu} \langle S \rangle_T \rangle \tag{2.5}$$

with $\langle \ldots \rangle_T$ standing for the thermal average, are given by the saddle-point equations

$$m^{\mu\nu} = \left\langle \xi^{\mu\nu} \tanh \left(\beta \sum_{\rho\sigma} m^{\rho\sigma} \xi^{\rho\sigma} \right) \right\rangle \tag{2.6}$$

with $h^\mu = 0$. Using equation (2.3) we can write the generalization overlaps m^μ in terms of the retrieval overlaps $m^{\mu\nu}$

$$m^\mu = \left\langle \xi^\mu \tanh \left(\beta \sum_{\rho\sigma} m^{\rho\sigma} \xi^{\rho\sigma} \right) \right\rangle . \tag{2.7}$$

We restrict our analysis to a particular class of solutions for $m^{\mu\nu}$, i.e. to a particular class of minima of f (or H for $T = 0$). Since there are no macroscopic correlations between different concepts we consider solutions of the form $m^{\mu\nu} = m^{1\nu} \delta_{\mu 1}$. Moreover, we choose $m^{1\nu}$ to be of the form

$$m^{1\nu} = \delta_{1\nu}(m^{11} - m_{s-1}) + m_{s-1} . \tag{2.8}$$

The motivation for choosing this solution is that it gives a bias to the network to behave as an associative memory : each example is singled out (the solution is s degenerate since we can select any of the examples to be m^{11}) and treated as an independent piece of information to be stored. If any other behaviour emerges, it will be a spontaneous property of the network and not an artifice due to the particular choice expressed by equation (2.8). At $T = 0$, m^{11} and m_{s-1} satisfy the equations

$$m^{11} = \langle \xi^{11} \operatorname{sign} (m^{11} \xi^{11} + m_{s-1} x_{s-1}) \rangle \tag{2.9a}$$

$$(s-1)\, m_{s-1} = \langle x_{s-1} \operatorname{sign} (m^{11} \xi^{11} + m_{s-1} x_{s-1}) \rangle \tag{2.9b}$$

where $x_{s-1} = \sum_{\nu > 1} \xi^{1\nu}$. For $s \geqslant 10$ the binomial distribution of x_{s-1} can be replaced by a Gaussian with mean $(s-1)\, b\xi^1$ and variance $\Delta_0^2 = (s-1)(1-b^2)$. Performing the averages over ξ^{11}, x_{s-1} and ξ^1 in this order, we find

$$m^{11} = \frac{1+b}{2} \operatorname{erf} (\Xi_+) + \frac{1-b}{2} \operatorname{erf} (\Xi_-) . \tag{2.10a}$$

$$m_{s-1} = b \left(\frac{1+b}{2} \operatorname{erf} (\Xi_+) - \frac{1-b}{2} \operatorname{erf} (\Xi_-) \right) + (1-b^2)\, C \tag{2.10b}$$

where

$$C = \frac{2}{\sqrt{\pi}} (2\Delta_0^2)^{-1/2} \left(\frac{1+b}{2} \exp(-\Xi_+^2) + \frac{1-b}{2} \exp(-\Xi_-^2) \right) \tag{2.11}$$

$$\Xi_\pm = \frac{m^{11} \pm (s-1)\, b m_{s-1}}{(2\Delta_0^2 m_{s-1}^2)^{1/2}} . \tag{2.12}$$

Following the same procedure we can compute m^1 from equation (2.7). The result is

$$m^1 = \frac{1+b}{2} \operatorname{erf} (\Xi_+) - \frac{1-b}{2} \operatorname{erf} (\Xi_-) . \tag{2.13}$$

Next we discuss the solutions of equations (2.10). For small s the network retrieves the examples almost perfectly, i.e. $m^{11} \approx 1$ and $m_{s-1} \approx b^2$. Hence the generalization error $\varepsilon \equiv \varepsilon^1$, equation (1.7), is

$$\varepsilon = \frac{1 - m^1}{2} \approx \frac{1 - b}{2} . \tag{2.14}$$

Strictly, the retrieval is perfect for $(s-1) \leqslant b^{-2}$ as can be easily seen from equations (2.9) since $|x_{s-1}| \leqslant s - 1$. This result is not recovered by equations (2.10) because the replacement of the discrete distribution of x_{s-1} by a Gaussian makes $|x_{s-1}|$ unbounded. As s increases, m^{11} decreases slightly until s reaches a critical value s_c, above which m^{11} jumps to a much smaller value, almost equating with m_{s-1}. This behaviour is reflected in the generalization error which we show in figure 1 as a function of s for several values of b. For $s < s_c$ the network just memorizes the patterns it is exposed to. This behaviour is referred to as the retrieval phase (R). For $s > s_c$ the network no longer treats the examples as independent pieces of information and starts to mix them, creating the representation for the concept. This is the generalization phase (G). The phase diagram in the (s, b) plane indicating the regions where each regime occurs is shown in figure 2.

Although $m^{11} = m_{s-1}$ is solution of equations (2.10) only in the limit $s \to \infty$, in the generalization regime one has $m^{11} \approx m_{s-1} \approx m_s$ where m_s is the symmetric solution of equation (2.6),

$$m^{\mu\nu} = \delta_{\mu 1} m_s \qquad \forall \nu . \tag{2.15}$$

In fact, m^{11} never equals m_{s-1} because when averaging over the examples we have explicitly ruled out this possibility by retaining the discrete nature of ξ^{11} while making the Gaussian approximation for the distribution of x_{s-1}. Nevertheless, since the differences between m^{11}, m_{s-1} and m_s are negligible in the G phase we can approximately characterize this regime by the symmetric solution m_s. Following similar steps to the ones leading to equations (2.10) one finds

$$m_s = \left(\frac{2(1 - b^2)}{s\pi} \right)^{1/2} \exp(-\Xi_0^2) + b \operatorname{erf}(\Xi_0) \tag{2.16}$$

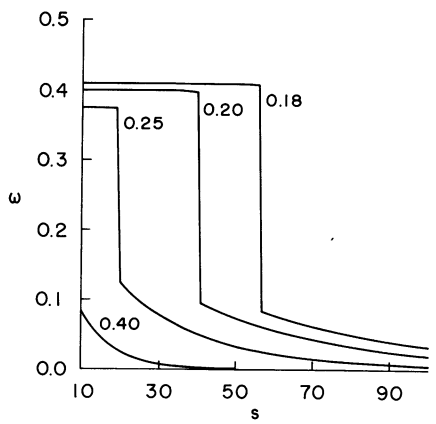

Fig. 1. — The generalization error as function of the number of examples for $b = 0.18$, 0.20, 0.25, 0.40 and $\alpha = 0$.

where

$$\Xi_0 = \left(\frac{sb^2}{2(1-b^2)} \right)^{1/2}. \tag{2.17}$$

Thus, the generalization error in the G phase can be approximate to

$$\varepsilon = \frac{1}{2}\,(1 - \mathrm{erf}\,(\Xi_0)) \qquad s > s_c\,. \tag{2.18}$$

The values of ε computed through this equation are indistinguishable from the exact generalization error, computed through equations (1.7) and (2.13), in the scale of figure 1.

3. Infinite number of concepts.

In this section we consider the case where the network has to create representations for an extensive number of concepts, $\alpha = p/N > 0$, being exposed to a finite number of examples s of each concept. In this case we have to take into account the fact that the combined overlap of a concept with all the other concepts is of $\mathcal{O}(\sqrt{\alpha})$. To handle this situation we follow Amit *et al.* [11] and assume that only the overlaps $m^{1\,\nu}$ condense, i.e. are of $\mathcal{O}(1)$ while the others are of $\mathcal{O}(N^{-1/2})$. The averaged free energy density is calculated through the replica trick

$$f = -\lim_{N\to\infty}\lim_{n\to 0} \frac{\langle Z^n\rangle - 1}{nN\beta} \tag{3.1}$$

where Z^n is the partition function replicated n times,

$$Z^n = \mathrm{Tr}_{s^\rho} \exp\left(\frac{\beta}{2N} \sum_{\mu\nu}^{ps} \sum_{\rho=1}^{n} \left(\sum_i^N \xi_i^{\mu\nu}\, S_i^\rho \right)^2 - \beta h^1 \sum_i^N \sum_{\rho=1}^{n} \xi_i^{1}\, S_i^\rho \right). \tag{3.2}$$

Averaging over $\xi_i^{\mu\nu}(\mu > 1)$ explicitly and using the self-averaging property of $\xi_i^{1\,\nu}$ yields

$$\langle Z^n\rangle = \int \prod_\nu \mathrm{d}m^{1\,\nu} \prod_{\rho\neq\sigma} \mathrm{d}q_{\rho\sigma}\,\mathrm{d}r_{\rho\sigma} \exp\left\{ \beta N \left[-\frac{1}{2}\sum_{\rho\nu} (m_\rho^{1\,\nu})^2 - \frac{\alpha\beta}{2} \sum_{\rho\neq\sigma} r_{\rho\sigma}\, q_{\rho\sigma} + \right.\right.$$
$$\left.\left. + \frac{\alpha}{\beta} \ln G(\{q_{\rho\sigma}\}) + \beta^{-1} \left\langle \ln \mathrm{Tr}_{s^\rho}\, \mathrm{e}^{\beta H_\xi} \right\rangle \right] \right\} \tag{3.3}$$

where we have omitted multiplicative factors which vanish in the thermodynamic limit and

$$G(\{q_{\rho\sigma}\}) = \int_{-\infty}^{\infty} \prod_{\rho\nu}^{ns} \frac{\mathrm{d}y_{\rho\nu}}{(2\pi)^{1/2}} \exp\left\{ -\frac{1}{2}\sum_{\rho\nu} y_{\rho\nu}^2 + \frac{\beta}{2} \sum_{\rho\sigma} \sum_{\nu\lambda} y_{\rho\nu}\, y_{\sigma\lambda}\, B_{\nu\lambda}\, Q_{\rho\sigma} \right\} \tag{3.4}$$

$$H_\xi = \sum_{\rho\nu}^{ns} m_\rho^{1\,\nu}\, \xi^{1\,\nu}\, S^\rho + \frac{\alpha\beta}{2} \sum_{\rho\neq\sigma} r_{\rho\sigma}\, S^\rho\, S^\sigma - h^1\, \xi^1 \sum_\rho S^\rho \tag{3.5}$$

with $B_{\nu\lambda} = b^2 + (1-b^2)\,\delta_{\nu\lambda}$ and $Q_{\rho\sigma} = q_{\rho\sigma} + (1 - q_{\rho\sigma})\,\delta_{\rho\sigma}$. The integrals in equation (3.3) can be readily effected by saddle-point integration while the integrals in equation (3.4), the trace over S^ρ and the limit $n \to 0$ can be performed in the replica symmetric framework

$$m_\rho^{1\,\nu} = m^{1\,\nu} \tag{3.6a}$$

$$q_{\rho\sigma} = q \qquad \rho \neq \sigma \tag{3.6b}$$

$$r_{\rho\sigma} = r \qquad \rho \neq \sigma \tag{3.6c}$$

resulting in the following expression for the averaged free energy density

$$f = \frac{1}{2}\sum_{\nu} (m^{1\nu})^2 + \frac{\alpha r C}{2} + \frac{\alpha}{\beta} \ln G(q) - \beta^{-1} \int_{-\infty}^{\infty} Dz \langle \ln 2 \cosh (\beta \Xi) \rangle \tag{3.7}$$

where

$$\ln G(q) = -\frac{1}{2}\left\{(s-1)\ln(1 - C(1-b^2)) + \ln(1 - C(1-b^2+sb^2)) - \frac{\beta q s (1 - C(1-b^2)(1-b^2+sb^2))}{(1-C(1-b^2))(1-C(1-b^2+sb^2))}\right\} \tag{3.8}$$

$$\Xi = z\sqrt{\alpha r} + \sum_{\nu} m^{1\nu}\xi^{1\nu} - h^1\xi^1 \tag{3.9}$$

with $C \equiv \beta(1-q)$ and $Dz \equiv dz/\sqrt{2\pi}\, e^{-z^2/2}$. The order parameters are given by the saddle-point equations which, in the limits $\beta \to \infty$ and $h^1 \to 0$, are written as

$$m^{1\nu} = \left\langle \xi^{1\nu} \operatorname{erf}\left(\sum_{\nu} m^{1\nu}\xi^{1\nu}/\sqrt{2\alpha r}\right)\right\rangle \tag{3.10}$$

$$C = \sqrt{2/\pi\alpha r}\left\langle \exp\left(-\left(\sum_{\nu} m^{1\nu}\xi^{1\nu}\right)^2/2\alpha r\right)\right\rangle \tag{3.11}$$

$$r = s\frac{[1 - C(1-b^2)(1-b^2+sb^2)]^2 + (s-1)b^4}{[1-C(1-b^2)]^2\,[1-C(1-b^2+sb^2)]^2}. \tag{3.12}$$

The equations for the standard Hopfield model [11] are recovered in the cases $s = 1$ and $b = 1$ with an appropriate rescaling of C and r. For $b = 0$ the network is effectively storing sp uncorrelated patterns and Hopfield's equations are obtained by rescaling $\alpha' = s\alpha$. Once $m^{1\nu}$ and C are known we can compute m^1 through the equation

$$m^1 = \left\langle \xi^1 \operatorname{erf}\left(\sum_{\nu} m^{1\nu}\xi^{1\nu}/\sqrt{2\alpha r}\right)\right\rangle. \tag{3.13}$$

Next we discuss the solutions of the saddle-point equations (3.10)-(3.12). In addition to the solutions considered in the $\alpha = 0$ limit, there exist a spin-glass solution $m^{1\nu} = 0$ $\forall \nu$ which is stabilized by the Gaussian noise due to the overlaps of $O(N^{-1/2})$ between $\xi^{1\nu}$ and $\xi^{\mu\nu}$, $\mu > 1$. However, adding noise to the system has a destabilizing effect for the retrieval phase [13, 16, 17], reducing its domain to a region much smaller than the one shown in figure 2. Therefore the phase diagram in the (s, b) plane will be dominated by the generalization phase characterized by the symmetric solution, equation (2.15), and the spin-glass phase. In the following we will focus only on the interplay between these two phases.

For the symmetric solutions equations (3.10) and (3.11) reduce to

$$m_s = \langle x_s \operatorname{erf}(m_s x_s/\sqrt{2\alpha r})\rangle \tag{3.14}$$

and

$$C = \sqrt{2/\pi\alpha r}\langle \exp(-m_s^2 x_s^2/2\alpha r)\rangle \tag{3.15}$$

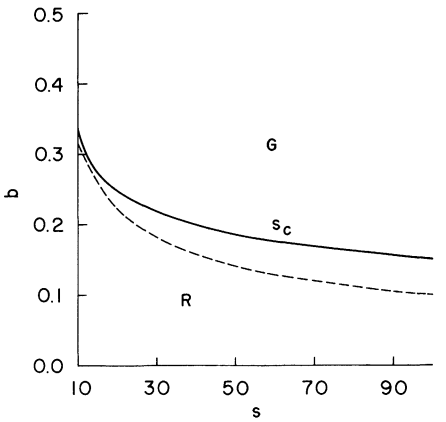

Fig. 2.

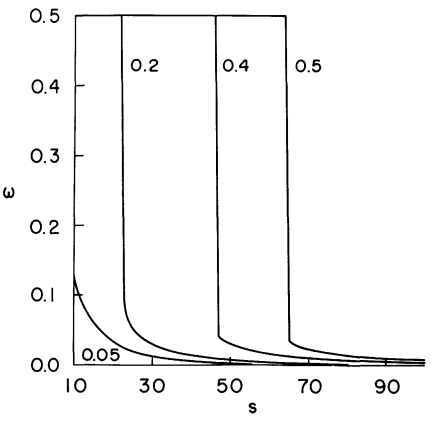

Fig. 3.

Fig. 2. — Phase diagram showing the generalization phase (G) and the retrieval phase (R) for $\alpha = 0$. The transition at $s = s_c$ is discontinuous. The retrieval of the examples is perfect below the dashed curve, $b = 1/\sqrt{s-1}$.

Fig. 3. — The generalization error as function of the number of examples for $\alpha/\alpha_0 = 0.05$, 0.2, 0.4, 0.5 and $b = 0.4$.

respectively, where $x_s = \sum_\nu \xi^{1\nu}$. For $s \geqslant 10$ one can replace x_s by a Gaussian variable of mean $sb\xi^1$ and variance $s(1-b^2)$. Performing the averages as in section 2 yields

$$m_s = \frac{b}{1 - C(1-b^2)} \operatorname{erf}\left(sbm_s/\sqrt{2\,\Delta_\alpha^2}\right) \tag{3.16}$$

$$C = \sqrt{2/\pi\,\Delta_\alpha^2}\exp\left(-(sbm_s)^2/2\,\Delta_\alpha^2\right) \tag{3.17}$$

where $\Delta_\alpha^2 = \alpha r + m_s^2 s(1-b^2)$. Equations (3.12), (3.16) and (3.17) must be solved numerically in order to compute the generalization overlap,

$$m^1 = \operatorname{erf}\left(sbm_s/\sqrt{2\,\Delta_\alpha^2}\right) \tag{3.18}$$

and, consequently, the generalization error $\varepsilon = (1 - m^1)/2$.

In the limit $s \to \infty$ the noise in the examples is averaged out, i.e. $x_s \to sb\xi^1$ and the standard Hopfield model with the concepts replacing the examples in equation (1.6) is recovered. This result can be easily obtained by rescaling $C' = sb^2C$ and $r' = r/s^2b^4$ which implies that $m_s = bm^1$ with m^1 satisfying the standard Hopfield's equations. Next we discuss the behaviour of the generalization error as a function of s, b and α. In figure 3 we show ε as a function of s for $b = 0.4$ and several values of α/α_0, where $\alpha_0 \approx 0.138$ gives the storage capacity of the standard Hopfield model. As in the $\alpha = 0$ case, there is a minimal number of examples which must be taught to the network in order it starts to generalize. The transition between the SG phase, where $\varepsilon = 0.5$ since the spin-glass states have no macroscopic correlations with the concepts, and the G phase is always discontinuous. The values of α/α_0 for which the transition occurs are shown in figure 4 for several values of b. As $b \to 1$ or $s \to \infty$, α_c tends to Hopfield's storage capacity α_0. Generalization occurs for $\alpha < \alpha_c$ and the

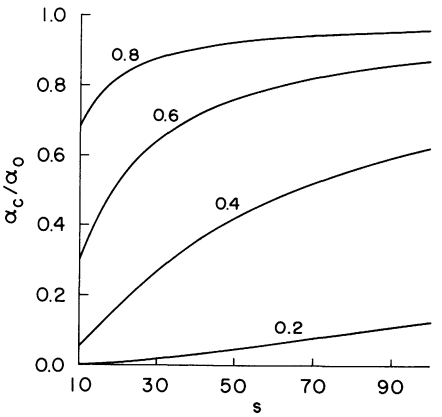

Fig. 4.

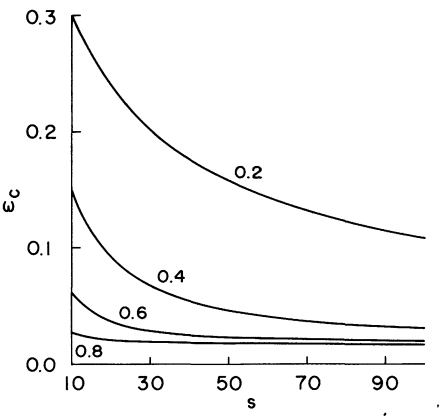

Fig. 5.

Fig. 4. — The critical values of α/α_0 below which the network starts to generalize as function of the number of examples for $b = 0.2,\ 0.4,\ 0.6,\ 0.8$.

Fig. 5. — The generalization error at criticality as function of the number of examples.

generalization error at the transition (ε_c) is shown in figure 5. As $b \to 1$ or $s \to \infty$, ε_c tends to 0.0165, the critical retrieval error for the standard Hopfield model [11].

4. Conclusion.

In this paper we have shown that a simple feedback neural network using a local Hebbian learning rule is able to learn a set of concepts having access only to a finite supply of typical data which exemplify the concepts. The network acomplishes that by creating representations, i.e. minima of the energy function governing the retrieval process, which capture the underlying statistical structure of the examples, allowing the extraction of meaningful information from the data supply. For the specific problem we have considered in this paper, the symmetric mixture states provide for the representations which allow the network to grasp the concepts. Since the research on Hopfield's model has focused mainly on the retrieval states, very little is known about the relation between the statistical structure of the patterns presented to the network during the learning stage and the structure of the representations created by the network [13, 17]. This seems to be a crucial issue if one intends to lead the study of Hopfield's model beyond its memorizing capabilities.

Next we summarize our main results. For finite p we have found that there is a regime where the network simply memorizes the examples ignoring their statistical structure (retrieval phase). However, as the number of examples increases passing a certain critical number s_c the behaviour of the network undergoes an abrupt change entering a new regime where the representations of the concepts are created (generalization phase). For $p = \alpha N$ $(a > 0)$ the retrieval phase is confined to a very small region of the (s, b) plane. However, a spin-glass regime, where the network ignores both the examples and the concepts, appears to compete with the generalization regime. The interplay between these two regimes is qualitatively similar to the one discussed in the finite p limit with the spin-glass replacing the retrieval regime.

It is interesting to compare our results with the ones presented in the literature for feedforward neural networks [1, 3, 6]. As mentioned above, the most remarkable outcome of

the present work is the existence of a critical number of examples (s_c) beyond which the network generalizes well (Figs. 1 and 3). Whether a similar behaviour occurs in the case of feedforward networks is an unsettled issue. On the one hand, simulations of multilayer networks for the contiguity problem and some general theoretical arguments point out for the existence of a critical size of the training set above which the generalization error (ε) falls of exponentially fast [1, 3]. On the other hand, an analytical study of the performance of a single-layer perceptron in classifying examples according to their Hamming distance from a set of prototypes indicates that such a critical number does not exist [6]. However, since the behaviour of ε seems to depend strongly on the architecture of the network considered [3] there may be no simple answer for this issue. A similar controversy could very well arise in the context of feedback neural networks if we use other learning rules than the one considered in this paper.

Finally, we should mention the relevance of our results to the problem of categorization in neural networks. As pointed out in the Introduction, the generalization regime may be interpreted as a categorization of the examples into the classes defined by the concepts. We have found that categorization emerges spontaneously when a critical number of examples is presented to the network during the learning stage. This result corroborates the viewpoint that categorization is related with the limitation of an associative memory. This point was beautifully expressed by Virasoro : « we categorize not because we want to but because we cannot do otherwise » [18].

Acknowledgments

It is a pleasure to thank Ronny Meir for many helpful discussions.

References

[1] DENKER J., SCHWARTZ D., WITTNER B., SOLLA S., HOWARD R., JACKEL L. and HOPFIELD J. J., *Complex Systems* **1** (1987) 877.
[2] PATARNELLO S. and CARNEVALI P., *Europhys. Lett.* **4** (1987) 503 ; CARNEVALI P. and PATARNELLO S., *Europhys. Lett.* **4** (1987) 1199.
[3] TISHBY N., LEVIN E. and SOLLA S. A., *Proceedings of the International Joint Conference on Neural Networks* (1989).
[4] VALLET F., *Europhys. Lett.* **8** (1989) 747.
[5] DEL GIUDICE P., FRANZ S. and VIRASORO M. A., *J. Phys. France* **50** (1989) 121.
[6] HANSEL D. and SOMPOLINSKY H., *Europhys. Lett.* **11** (1990) 687.
[7] GYÖRGYI G. and TISHBY N., *Proceedings of the STATPHYS-17 Workshop on Neural Networks and Spin Glasses*, Eds. W. K. Theumann and R. Köberle (World Scientific, Singapore) 1990.
[8] GARDNER E., *J. Phys.* **A 21** (1988) 257.
[9] HOPFIELD J. J., *Proc. Natl. Acad. Sci. USA* **79** (1982) 2554.
[10] AMIT D. J., GUTFREUND H. and SOMPOLINSKY H., *Phys. Rev. A* **32** (1985) 1007.
[11] AMIT D. J., GUTFREUND H. and SOMPOLINSKY H., *Ann. Phys. N.Y.* **173** (1987) 30.
[12] MEZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.
[13] FONTANARI J. F. and THEUMANN W. K., *J. Phys. France* **51** (1990) 375.
[14] ANDERSON J. A., *IEEE Transactions on Systems, Man and Cybernetics* **13** (1983) 799.
[15] FONTANARI J. F. and MEIR R., *Phys. Rev. A* **40** (1989) 2806.
[16] FONTANARI J. F. and KÖBERLE R., *J. Phys. A* **21** (1988) 2477.
[17] ERICHSEN R. and THEUMANN W. K., Preprint IF-UFRGS (1990).
[18] VIRASORO M. A., Preprint 608, Università di Roma « La Sapienza ».