Generalization in an analog neural network

D. A. Stariolo and F. A. Tamarit*
Centro Brasileiro de Pesquisas Fisicas, Rua Xavier Sigaud 150, 22290 Rio de Janiero, Brazil
(Received 30 July 1991)

We analyze the generalization ability of an iterated-map neural network when an extensive number of patterns is stored through a Hebbian learning mechanism. We show that the model is able to create a concept representative of a set of correlated patterns if a critical minimum number of patterns is presented. This critical number depends on the correlation among the patterns, the slope of the transfer function at the origin, and the ratio between the number of memories and the total number of neurons.

PACS number(s): 87.10.+e, 64.60.Ht, 75.10.Nr

I. INTRODUCTION

One of the most interesting issues of neural-network models is their generalization capacity, i.e., the ability of the system to create a concept that represents a set of patterns. In other words, irrespective of fine details, the network will be able to generalize if it succeeds in capturing the common features shared by the patterns. This problem has been studied mainly in the context of feedforward networks [1–4], where the system is trained to infer an optimal rule which maps a particular set of inputs to a given set of outputs. When the system learns how to solve a certain set of examples, it eventually reaches a configuration in which, for any subsequent input, the correct output is inferred.

In a recent work [5], Fontanari analyzed the generalization capacity of the Hopfield model with a two-level hierarchical organization of its memories. He found that the system is able to generalize when the sets at the lower level are composed of a minimum number of patterns. In this kind of model, generalization means that, as the number of patterns in each branch increases, the system loses the capability for recognizing single patterns, but instead it creates a representation that contains the common features of each set. In the context of feedback networks, it has been found that models with continuous-state variables and deterministic dynamics perform better than discrete-stochastic systems. From a biological point of view, a continuous (generally sigmoidal) input-output relation is more realistic than the simpler step transfer function.

In a series of interesting papers, Marcus, Waugh, and Westervelt have studied the dynamical properties of analog feedback neural networks in the context of associative memory [6–8]. In particular, they proved that under general conditions, these models present two important properties for the implementation of fast computing devices, namely, (i) the stability problems inherent to the parallel dynamics of two-state networks can be suppressed [6], and (ii) the number of spurious attractors can be greatly reduced [8]. Trying to provide more insight into the behavior of this kind of system, we study in this work the dynamical behavior of a neural network modeled by real-state variables, and whose dynamics is defined by a set of coupled nonlinear maps with parallel updating. Taking into account the properties found by Marcus, Waugh, and Westervelt, we study whether the use of analog neurons also improves the performance of these models as categorizing devices.

In Sec. II we define the model: its dynamics, its architecture, how it stores information, its relation with the generalization task, and the magnitudes of interest. In Sec. III we first perform a statistical analysis of the model, and find a set of equations for the relevant magnitudes. We then solve the equations and present our results for two important cases: (i) when the input-output relation approaches a step function, and (ii) for general values of the slope of the transfer function at the origin. Finally, in Sec. IV we present a discussion of our results.

II. THE MODEL

The system is composed of N neurons whose states at instant t are represented by the real-valued variables $X_i(t)$, $i=1,\ldots,N$. The whole network is updated in parallel according to the deterministic dynamics defined by the following system of coupled maps:

$$X_i(t+1)=\tanh[gh_i(t)], \quad i=1,\ldots,N \ , \tag{1}$$

where g is the gain parameter of the transfer function and

$$h_i(t)=\sum_{j=1}^{N} J_{ij}X_j(t) \tag{2}$$

is the input to neuron i at time t. J_{ij} is the symmetric sinaptic matrix that fully connects the neurons, and is defined by the following Hebbian rule:

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{\mu}\xi_j^{\mu}, \quad J_{ii}=0 \ , \tag{3}$$

where $\{\xi_i^{\mu}=\pm 1\}$, with $\mu=1,\ldots,p$, are the stored patterns. In order to study the generalization properties of the model, we first create a random configuration $\{\eta_i\}$, where $\eta_i\pm 1$ with equal probability. This configuration will have the common features of a set of s patterns. The components of these s patterns are statistically independent random variables chosen according to the following distribution:

$$P(\xi_i^\mu)=\tfrac{1}{2}(1+\eta_i b)\delta(\xi_i^\mu-1)+\tfrac{1}{2}(1-\eta_i b)\delta(\xi_i^\mu+1)\ , \tag{4}$$

with $0\le b\le 1$. We will consider the case in which s is finite and $p=\alpha N$. The components of the remaining p-s patterns are also statistically independent random variables, taking the values ± 1 with equal probability. With this choice, the model has the simplest possible hierarchical organization that allows the storage of an extensive number of memories. The configuration $\{\eta\}$ can be thought as a *concept* not explicitly present in the learning rule but introduced through s noisy examples. The parameter b measures the proximity between each example and the concept.

Much work has been done to overcome the limitations of the Hopfield model for storing correlated patterns [9,10]. In these papers the retrieval properties of neural networks with hierarchically correlated memories are studied by introducing new learning rules. In the present work we do not focus on the retrieval problem, but we consider instead the learning process as a training strategy, i.e., we look for the minimum number of examples required by the system defined by Eqs. (1)–(4) in order to have the concept as an attractor. Although all these systems share a hierarchical structure, the nature of the problems is different, and the results cannot be easily compared.

We are now interested in the asymptotic behavior of the overlap between the state of the system and the concept, given by

$$m_c(t)=\frac{1}{N}\sum_{i=1}^{N}\eta_i X_i(t)\ . \tag{5}$$

Let us define the *generalization error* ϵ_g as the Hamming distance between the concept and the asymptotic state of the network. Then ϵ_g is related to the value of m_c at the fixed points of the maps given by Eqs. (1) through

$$\epsilon_g=\frac{1-m_c}{2}\ . \tag{6}$$

III. STATISTICAL ANALYSIS AND FIXED-POINT EQUATIONS

We start by looking for the overlaps between each memory and the fixed-point state $\{X_i^*\}$ of Eqs. (1):

$$m^\mu=\frac{1}{N}\sum_i \xi_i^\mu X_i^*,\quad \mu=1,\ldots,p\ . \tag{7}$$

Following the ideas introduced by Marcus, Waugh, and Westervelt [8], we search for a set of self-consistent equations for these overlaps. Taking into account that

$$X_i^*=\tanh\left[g\sum_\mu \xi_i^\mu m^\mu\right]\ , \tag{8}$$

and using the self-averaging property, we can rewrite Eqs. (7) as

$$m^\mu=\left\langle \xi^\mu \tanh\left[g\sum_\mu \xi^\mu m^\mu\right]\right\rangle\ , \tag{9}$$

where $\langle\ \rangle$ denotes an average over the random variables $\{\xi_i^\mu\}$ and $\{\eta_i\}$, in this order. Since we want to recognize the concept (and not the memories) and it has the same overlap with all the examples, we only consider solutions having the same macroscopic overlap with the s examples, and an overlap of order $O(1/N^{1/2})$ with the remaining p-s patterns. In other words, we are interested in symmetric solutions of the form

$$sm_s=\left\langle z_s\tanh\left[g\left[m_s z_s+\sum_{\mu(>s)}\xi^\mu m^\mu\right]\right]\right\rangle\ , \tag{10}$$

where

$$z_s=\sum_{\mu=1}^{s}\xi^\mu\ .$$

We consider the approximation in which, in order to average over the s first ξ variables, z_s is treated as a Gaussian variable, with mean value $sb\eta$ and variance $\sigma_1^2=s(1-b^2)$. Note that, at this stage, the average over η has not yet been performed. For finite values of α, the second term in the argument of the transfer function is a Gaussian variable with mean value zero and variance $\sigma_2^2=\alpha(g-C)/g(1-C)^2$. Here, we have used the previous results of Marcus, Waugh, and Westervelt [8]. C is defined by

$$C=\left\langle g\,\mathrm{sech}^2\left[g\left[m_s z_s+\sum_{\mu(>s)}\xi^\mu m^\mu\right]\right]\right\rangle\ . \tag{11}$$

After averaging over the Gaussian noise, the random variables z_s and η, in this order, we arrive at the following set of coupled equations:

$$m_s=\tfrac{1}{2}\int Dz\int Dh\left[(b+\sigma_1 z/s)\tanh(\Theta^+)-(b-\sigma_1 z/s)\tanh(\Theta^-)\right]\ , \tag{12a}$$

$$C=\frac{g}{2}\int Dz\int Dh\left[\mathrm{sech}^2(\Theta^+)+\mathrm{sech}^2(\Theta^-)\right]\ , \tag{12b}$$

with $\Theta^\pm=m_s(z\sigma_1\pm sb)+\sigma_2 h$ and

$$Dz=\frac{dz}{\sqrt{2\pi}}\exp(-z^2/2),\quad Dh=\frac{dh}{\sqrt{2\pi}}\exp(-h^2/2)\ .$$

A similar analysis for the overlap m_c leads to the following expression:

$$m_c=\tfrac{1}{2}\int Dz\int Dh\left[\tanh(\Theta^+)-\tanh(\Theta^-)\right]\ . \tag{12c}$$

In the next section, we study the numerical solutions of Eqs. (12).

A. The limit of infinite gain

In the limit $g\to\infty$, the equations for m_s, C, and m_c take the form

$$m_s=\sqrt{2/\pi}\,\frac{m_s(1-b^2)}{\Delta}\exp\left[-\frac{(m_s sb)^2}{2\Delta^2}\right]+b\,\mathrm{erf}\left[\frac{m_s sb}{\sqrt{2}\Delta}\right]\ , \tag{13a}$$

$$C=\left[\frac{2}{\pi\Delta^2}\right]^{1/2}\exp\left[-\frac{(m_s sb)^2}{2\Delta^2}\right]\ , \tag{13b}$$

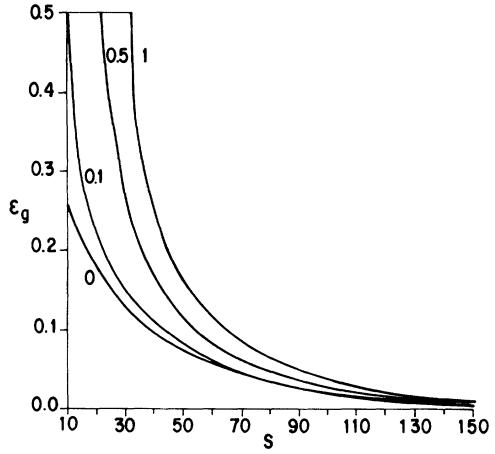

FIG. 1. Generalization error ϵ_g vs number of examples s in the limit $g \to \infty$ for $b=0.2$ and $\alpha=0$, 0.1, 0.5, and 1.

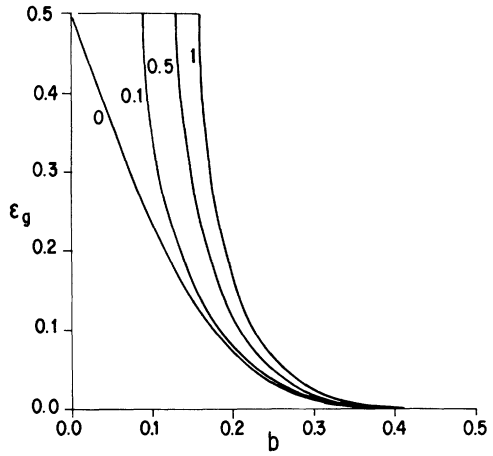

FIG. 3. Generalization error ϵ_g vs b in the limit $g \to \infty$ for $s=50$ and for $\alpha=0$, 0.1, 0.5, and 1.

$$m_c = \operatorname{erf}\left[\frac{m_s s b}{\sqrt{2}\Delta}\right], \tag{13c}$$

where $\Delta^2 = s(1-b^2)m_s^2 + \alpha/(1-C)^2$.

In Fig. 1 we show the generalization error ϵ_g versus the number of examples s for different values of α and for $b=0.2$. As α grows so does the noise due to the uncorrelated patterns, and consequently the system must be exposed to a greater number of examples in order to generalize. Given a fixed value of α, there exists a critical value s_c above which the system starts recognizing the class in a continuous way, in contrast with results found for the Hopfield model, where the transition is discontinuous [5]. For $s \gg 1$, $m_c \to 1$, $m_s \to b$, and ϵ_g falls off exponentially as

$$\epsilon_g \simeq \left[\frac{\gamma_1}{2\pi^{1/2}}\right] \exp\left[-\frac{1}{2\gamma_1}\right], \tag{14}$$

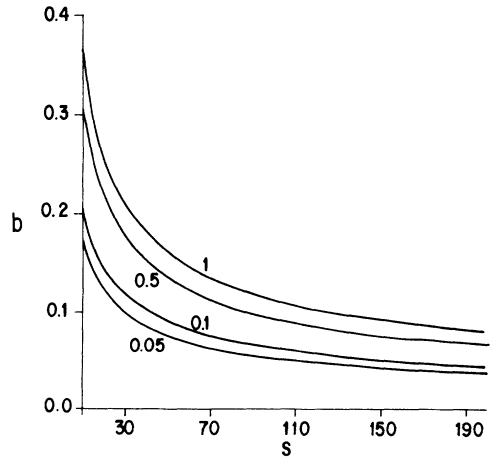

FIG. 2. The critical lines $b(s)$ in the limit $g \to \infty$ for $\alpha=0.05$, 0.1, 0.5, and 1. The system begins to generalize above the corresponding line.

with

$$\gamma_1 = \frac{1-b^2}{sb^2}.$$

Since close to the transition $m_s \ll 1$, we expand Eqs. (13a) and (13b), obtaining the following relation for the critical values α_c, b_c, and s_c:

$$\alpha_c = \frac{2}{\pi} b_c^4 (s_c - 1)^2, \tag{15}$$

which is in agreement with previous results [11,12]. In Fig. 2, we present the critical lines $b_c(s_c)$, above which nonzero solutions of Eq. (13a) appear for different values of α_c. It is important to note that α_c diverges with s_c and is not related to the critical storage capacity of the network. Below these lines, the system presents either spin-glass or retrieval solutions, depending on the value of α. The detailed study of these solutions will be presented in a future paper.

If we demand that the system creates a representation of the concept without error, when a fixed number of ex-

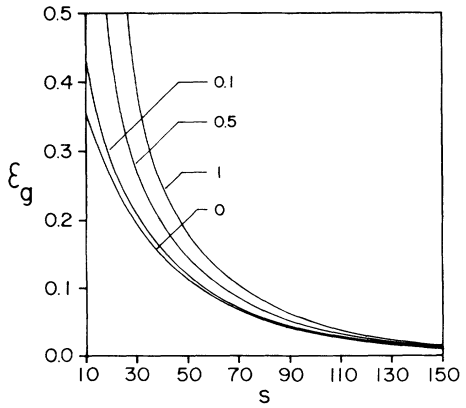

FIG. 4. Generalization error ϵ_g vs s for $g=1$, $b=0.2$ and for $\alpha=0$, 0.1, 0.5, and 1.

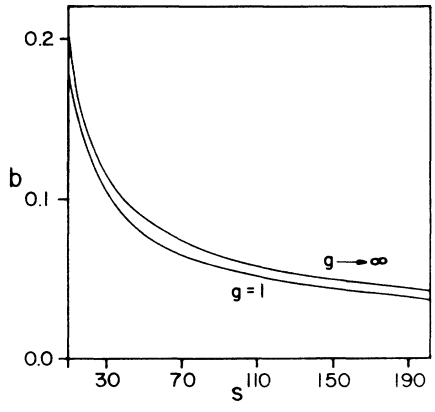

FIG. 5. The critical lines $b(s)$, above which the system generalizes, for $g=1$ and $g\to\infty$, with $\alpha=0.1$.

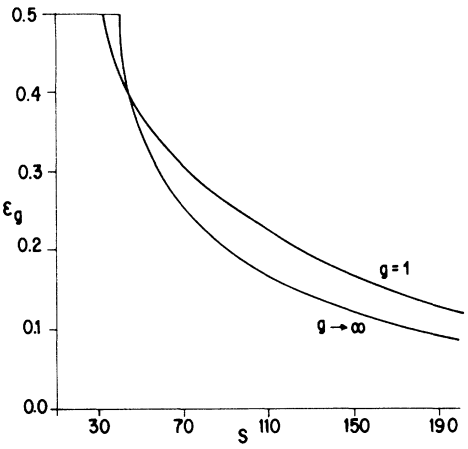

FIG. 6. Generalization error ϵ_g vs s for $g=1$ and $g\to\infty$ with $b=0.1$, $\alpha=0.1$.

amples is presented to it, then conditions $b\simeq 1$ and $\alpha\simeq 0$ must be fulfilled, and the error takes the same form as in Eq. (14):

$$\epsilon_g \simeq \left[\frac{\gamma_2}{2\pi}\right] \exp\left[-\frac{1}{2\gamma_2}\right], \tag{16}$$

where

$$\gamma_2=\frac{\alpha+sb^2(1-b^2)}{s^2b^4}.$$

Figure 3 shows ϵ_g versus b for $s=50$ and for several values of α.

B. The finite-gain case

Solving numerically Eqs. (12), we study the performance of the network for general values of the gain parameter g. In Fig. 4 we can see ϵ_g versus s for $g=1$, $b=0.2$ and for different values of α. Comparing these results with those obtained in the case $g\to\infty$, we observe an improvement of the network generalization capacity: the system needs fewer instances of the concept in order to begin to generalize. This improvement can be better seen in Fig. 5, where the critical lines $b_c(s_c)$ are shown for $\alpha=0.1$ and for two values of g. Finally, Fig. 6 compares the generalization error versus s for $g=1$ and $g\to\infty$, with $b=0.1$ and $\alpha=0.1$. Note that, although for $g=1$ the transition occurs for a smaller number of examples, the system with $g\to\infty$ works better when we require a small error ϵ_g ($s\gg 1$).

IV. DISCUSSION

In this work we showed that an analog neural network is able to classify a certain number of inputs according to their proximity, i.e., to generalize, provided a minimum number of examples of the class is presented to it. If compared with the two-state limit of the model ($g\to\infty$), finite values of the gain parameter introduce small improvements in the generalization capacity, that is, a smaller number of examples is needed in order that the network starts to generalize. In a future work we will present a detailed analysis of the different phases of this kind of model when correlation among the patterns is present.

Finally, we believe it would be interesting to study the generalization problem in feedback networks by working in the space of interactions, trying to find an algorithm for adjusting the couplings so that the generalization ability of the system is optimal.

ACKNOWLEDGMENTS

We want to thank E. M. F. Curado, W. K. Theumann, L. A. Florit, D. F. Rial, and C. Anteneodo for fruitful discussions.

*Also at Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Laprida 854, 5000 Córdoba, Argentina.

[1] P. del Giudice, S. Franz, and M. A. Virasoro, J. Phys. (Paris) Colloq. **50**, C2-121 (1989).

[2] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses,* edited by W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990), and references cited therein.

[3] D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990).

[4] H. Sompolinsky and N. Tishby, Phys. Rev. Lett. **65**, 1683 (1990).

[5] J. F. Fontanari, J. Phys. (Paris) **51**, 2421 (1990).

[6] C. M. Marcus and R. M. Westervelt, Phys. Rev. A **39**, 347 (1989); **40**, 501 (1989); **42**, 2410 (1990).

[7] F. R. Waugh, C. M. Marcus, and R. M. Westervelt, Phys. Rev. A **43**, 3131 (1991).

[8] C. M. Marcus, F. R. Waugh, and R. M. Westervelt, Phys. Rev. A **41**, 3355 (1990).

[9] N. Parga and M. A. Virasoro, J. Phys. (Paris) **47**, 1857 (1986).

[10] H. Gutfreund, Phys. Rev. A **37**, 570 (1988).

[11] J. F. Fontanari and R. Köberle, J. Phys. A **21**, 2477 (1988).

[12] R. Erichsen and W. K. Theumann (unpublished).