Information capacity of a hierarchical neural network

David Renato Carreta Dominguez*
Instituut voor Theoretische Fysica, Katholieke Universeteit Leuven, Celestijnenlaan, B-3001 Leuven, Belgium
(Received 8 June 1998)

The information conveyed by a hierarchical attractor neural network is examined. The network ''learns'' sets of correlated patterns (the examples) in the lowest level of the hierarchical tree and can categorize them at the upper levels. A way to measure the nonextensive information content of the examples is formulated. Curves showing the transition from a large retrieval information to a large categorization information behavior, when the number of examples increase, are displayed. The conditions for the maximal information are given as functions of the correlation between examples and the load of concepts. Numerical simulations support the analytical results. [S1063-651X(98)06110-8]

PACS number(s): 87.10.+e, 64.60.Cn, 02.50.−r

I. INTRODUCTION

In the context of learning rules by perceptrons, generalization by a neural network is the capability of correctly classifying patterns after some examples are ''taught'' to the network (see, e.g., Ref. [1]). For attractor neural networks, another type of generalization was suggested, the categorization, that emerges from an encoding stage where a hierarchical tree of patterns is stored [2]. The ability of the network to classify the patterns on a lower level of the tree (i.e., the *examples*) into categories defined by their ancestors (i.e., the *concepts*), arises from the Hopfield model if the examples are correlated with their concepts [3].

A minimal number S of examples for each concept is necessary to start the categorization. An extensive number of concepts is then ''learned'' by memorizing finite sets of examples. This was shown for networks of binary neurons with fully connected [4,5], diluted [6], or layered [7] architectures, and for analog [8], ternary [9], and nonmonotonic [10] neurons, using Hebbian synapses. A similar behavior was found for pseudoinverse synapses [11]. Categorization is achieved through the appearance of symmetric spurious states. This ability to categorize starts just when the capacity of the network recovering the original examples is lost, because of the interference generated by their correlations.

As in most models for pattern recognition, an adequate analysis of the memory capacity of this network requires the tools of information theory. In the case of nonbiased independent patterns, one can avoid it and measure the performance through the Hamming distance D between the neuron and the retrieved pattern, and the load capacity α. One scenario, where D and α are not enough to characterize the system, is that of sparse coded patterns [12]. Another is that of dependent patterns. This is the case for categorization models, since the information conveyed by the examples is not extensive in them.

Our goal in this work is to establish a reliable measure for the capacity of retrieving examples, and their categorization, based in the information theory. In Sec. II, we define the model and its parameters. After obtaining expressions for the information capacity in Sec. III, in Sec. IV we study some special cases which present the transition from a retrieval phase to a categorization phase. Finally we conclude with some remarks in Sec. V.

II. MODEL

Consider a network of N binary neurons, with states $\{\sigma_{i,t}=\pm 1\}_{i=1}^{N}$ at time t. The neurons states are updated in parallel according to the deterministic rule

$$\sigma_{i,t+1}=\mathrm{sgn}(h_{i,t}), \quad h_{i,t}=\sum_{i(\neq j)}^{N} J_{ij}\sigma_{jt}, \tag{1}$$

where $h_{i,t}$ is the local field of neuron i at time t. The elements of the Hebbian-like synaptic matrix between neurons i and j are given by

$$J_{ij}=\frac{1}{N}\sum_{\mu}^{p}\sum_{\rho}^{S}\eta_i^{\mu\rho}\eta_j^{\mu\rho}, \tag{2}$$

where $\{\eta_i^{\mu\rho}\}_{\rho=1}^{S}$ are the *examples* of the *concept* ξ_i^{μ}. The concepts are independent identically distributed random variables (IIDRV's), $\{\xi_i^{\mu}=\pm 1\}_{i=1}^{p}$, with equal probability.

In the encoding stage, the examples are built from the concepts, according to the stochastic process

$$p(\eta_i^{\mu\rho}|\xi_i^{\mu})=b\,\delta(\eta_i^{\mu\rho}-\xi_i^{\mu})+(1-b)\,\delta(|\eta_i^{\mu\rho}|^2-1), \tag{3}$$

where $b=\langle\eta_i^{\mu\rho}\xi_i^{\mu}\rangle$ gives the correlation between the ancestors (the concepts) and the descendants (the examples) of this tree of patterns. The second δ of this conditional distribution gives the component of the examples which is independent on the concepts. This process can equivalently be formulated as $\eta_i^{\mu\rho}=\xi_i^{\mu}\lambda_i^{\mu\rho}$, where the *biased* IIDRV's $\lambda_i^{\mu\rho}$ are distributed according to

$$p_B(\lambda_i^{\mu\rho})=B_+\,\delta(\lambda_i^{\mu\rho}-1)+B_-\,\delta(\lambda_i^{\mu\rho}+1), \tag{4}$$

with $B_{\pm}=(1\pm b)/2$.

The macroscopic parameters which describe the state of the network are the *retrieval* and *categorization overlaps*, respectively:

*Electronic address: david@tfdec1.fys.kuleuven.ac.be

$$m_{Nt}^{\mu\rho} \equiv \frac{1}{N} \sum_j \eta_j^{\mu\rho} \sigma_{jt}, M_{Nt}^{\mu} \equiv \frac{1}{N} \sum_j \xi_j^{\mu} \sigma_{jt}. \quad (5)$$

In the thermodynamic limit, the qualities of the retrieval and of the categorization can be measured by taking the lim $N \to \infty$ of the overlaps for a single concept, say $\mu = 1$, which give

$$m_t^{1\rho} \equiv \langle \eta^{1\rho} \mathrm{sgn}[h_{t-1}] \rangle, \quad M_t^1 \equiv \langle \xi^1 \mathrm{sgn}[h_{t-1}] \rangle, \quad (6)$$

where the brackets mean averages over the set of examples $\eta^{1\rho}$ and the local field h_{t-1} for a single neuron.

The *generalization error* [1,3] can be defined as $E_t^1 \equiv \langle |\sigma_t - \xi^1|^2 \rangle = 1 - M_t^1$, as a function of the categorization overlap. The stationary states are given by macroscopic overlaps with examples of a given concept, say $m_\infty^{1\rho} \equiv m^{1\rho}$, and microscopic remaining overlaps $\nu > 1$ and $m^{\nu\rho} \sim 1/\sqrt{N}$. The general solution is represented by a retrieval overlap with a single example, say $m^{11} \equiv m$, and the *quasisymmetric* overlaps with the other examples, $m^{1\rho} \equiv m^S$ and $\rho > 1$. In the retrieval phase one has $m \sim 1$ and $m^S \sim b^2$, while in the categorization phase the stable state is $m = m^S \sim b$, which may lead to a large categorization overlap $M_\infty^1 \equiv M \sim 1$. In the following we will consider a situation where the network relaxes to equilibrium states, so we can drop the time t on the parameters.

III. INFORMATION CAPACITIES

In this section we describe a way to measure the storage of information by the network in the retrieving and categorizing regimes. There are two types of information to be extracted from the patterns in these networks: retrieval information and categorization information. The former is that which can be conveyed from the examples to the neurons, while the latter is that which can be conveyed from the concepts. In each case one must calculate the information entropy of the pattern distributions, $H[\{\xi_i^\mu\}_{i,\mu}^{N,p}] = -\Sigma_{\{\xi_i^\mu\}} p(\{\xi_i^\mu\}) \ln[p(\{\xi_i^\mu\})]$, and $H[\{\eta_i^{\mu\rho}\}_{i,\mu,\rho}^{N,p,S}] = -\Sigma_{\{\eta_i^{\mu\rho}\}} p(\{\eta_i^{\mu\rho}\}) \ln[p(\{\eta_i^{\mu\rho}\})]$, where $p(\{\xi_i^\mu\})$ and $p(\{\eta_i^{\mu\rho}\})$ are the concepts and examples joint probability distributions, respectively.

The categorization information can be easily measured by computing the categorization overlap of a single concept, M, and its entropy. Since the concepts $\{\xi_i^\mu\}_{i,\mu}^{N,p}$ are IIDRV's, their probability distribution is factorial, $p(\{\xi_i^\mu\}_{i,\mu}^{N,p}) = \Pi_{i,\mu}^{N,p} p(\xi_i^\mu)$. Thus the entropy of the concepts is extensive, $H[\{\xi_i^\mu\}_{i,\mu}^{N,p}] = \Sigma_{\mu,i}^{p,N} H[\xi_i^\mu] = pNH[\xi]$, where the entropy of a single concept on a single neuron is $H[\xi] = \log_2(2)$. As we study binary patterns, we shall use base-2 logarithm in order to count information in bits, then we have $H[\xi] = 1$. The equivocation in the categorization can be evaluated by the square of the overlap, in such a way that no information is transmitted by the concepts if $M = 0$ and the information is maximal if $M = \pm 1$, showing that the information is symmetric in this overlap, because an inverted concept $\sigma_i = -\xi_i$ carries the same information than $\sigma_i = \xi_i$. Therefore, the total categorization information is $I_C = pNM^2 H[\xi]$, and the categorization information (per synapse) is

$$i_C = \alpha M^2. \quad (7)$$

The retrieval information can be similarly measured, by computing the retrieval overlap and the entropy of the examples, since this entropy can also be factorized as $p(\{\eta_i^{\mu\rho}\}_{i,\mu,\rho}^{N,p,S}) = \Pi_{i,\mu}^{N,p} p(\{\eta_i^{\mu\rho}\}_\rho^S)$, such that the entropy is extensive in the concepts and in the neurons, $H[\{\eta_i^{\mu\rho}\}_{i,\mu,\rho}^{N,p,S}] = \Sigma_{\mu,i}^{p,N} H[\{\eta_i^{\mu\rho}\}_\rho^S] = pNH[\{\eta^\rho\}_\rho^S]$. Thus it is enough to calculate the entropy of a set of examples of a single concept, $\{\eta^\rho\}_\rho^S \equiv \{\eta_i^{\mu\rho}\}_\rho^S$, on a single neuron, to obtain the entropy of the whole set $\{\eta_i^{\mu\rho}\}_{i,\mu,\rho}^{N,p,S}$.

On the other hand, $\{\eta^\rho\}_\rho^S$ is *not* a set of IIDRV's, so $p(\{\eta^\rho\}_\rho^S)$ is *not* factorizable in example probabilities, and the entropy is *not* extensive in the examples, $H[\{\eta^\rho\}_\rho^S] \neq \Sigma_\rho^S H[\eta^\rho]$. So the retrieval information is not the naive one, $i_R \neq \alpha S$.

Let $\{\eta^\rho\} \equiv \{\eta^\rho\}_\rho^S$ be a set of examples of a given concept on a given neuron. In calculating $p(\{\eta^\rho\})$ we proceed as follows: we take the conditional probability of the examples given the concept, $p(\{\eta^\rho\}|\xi)$, from Eq. (3), and average it on the distribution of ξ,

$$p(\{\eta^\rho\}) = \langle p(\{\eta^\rho\}|\xi) \rangle_\xi = \prod_{\rho=1}^{S} \frac{p_B(\eta^\rho) + p_B(-\eta^\rho)}{2}, \quad (8)$$

where p_B is the probability distribution in Eq. (4). After expanding this product, we calculated the entropy of this distribution, obtaining

$$H[\{\eta^\rho\}] = -\sum_{k=0}^{S} C_k^S A_k \ln(A_k),$$
$$A_k = [B_+^k B_-^{S-k} + B_-^k B_+^{S-k}]/2, \quad (9)$$

where C_k^S are the combinatorial numbers.

In evaluating the equivocation in the retrieval, here we have to multiply this entropy by the square of the retrieval overlap of a single example. Since we have to subtract the information due to the categorization, and the overlaps between examples and their concepts are $b = \langle \eta^\rho \xi \rangle$, we estimate the total retrieval information as $I_R = pN(m - bM)^2 H[\{\eta^\rho\}]$. Therefore the retrieval information (per synapse) is

$$i_R = \alpha(m - bM)^2 H[\{\eta^\rho\}]. \quad (10)$$

Although other measures for the information could be used, they must be monotonous functions of those we consider in the Eqs. (7) and (10). Nevertheless, these have the advantage that both are equivalently scaled, and they can be directly compared to each other.

IV. RESULTS

We now present the equilibrium states for the networks which are used to obtain the retrieval and categorization information. These states are studied for two systems: an asymptotic network ($N \to \infty$), for which analytical stationary equations were derived [3], and finite-sized systems, for which simulations of the dynamics in Eq. (1) are carried on.

While the information measures obtained in Sec. III are functions of asymptotic parameters M and m, the results from simulation use the overlaps in Eqs. (5).

A. Asymptotic network

First we study the stationary states of the overlaps in Eqs. (5), in the thermodynamic limit $N\to\infty$. Using the Hebbian synapses in Eq. (2) in the dynamics in Eq. (1), taking the local field at the fixed point, and averaging over the distribution of a single example, one obtains

$$m=\sum_{k=0}^{S-1} p_S(k)\int_{-\infty}^{\infty} Dz[B_+G_+-B_-G_-],$$

$$M=\sum_{k=0}^{S-1} p_S(k)\int_{-\infty}^{\infty} Dz[B_+G_++B_-G_-], \qquad (11)$$

$$m^S=\sum_{k=0}^{S-1} p_S(k)\frac{x_S}{S-1}\int_{-\infty}^{\infty} Dz[B_+G_++B_-G_-],$$

for the retrieval, categorization, and quasisymmetric overlap, respectively. Here

$$G_\pm=\mathrm{sgn}[x_S m^S\pm m+z\sqrt{\alpha r}], \qquad (12)$$

with $x_S\equiv\Sigma_{\rho=2}^{S}\lambda_\rho\equiv 2k-(S-1)$. and the averages are over the remaining $S-1$ examples from the first concept, and the remaining $p-1$ concepts. The first is the binomial variable $x_S=2k-(S-1)$, distributed according to

$$p_S(k)=C_k^{S-1}B_+^kB_-^{S-1-k}; \qquad (13)$$

the last is a Gaussian noise, distributed according to

$$Dz=\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}. \qquad (14)$$

In the present case of a fully connected network, there is a strong feedback in the dynamics, but an expression for the variance of the noise can be obtained using a replica symmetric approach [3,5],

$$r=s\frac{[1-C(1-b^2)(1-b^2+sb^2)]^2+(s-1)b^4}{[1-C(1-b^2)]^2[1-C(1-b^2+sb^2)]^2}, \qquad (15)$$

with

$$C=\frac{1}{\sqrt{\alpha r}}\sum_{k=0}^{S-1} p_S(k)\int_{-\infty}^{\infty} Dz\, z[B_+G_++B_-G_-]. \qquad (16)$$

We have to solve Eqs. (11)–(16), then introduce the overlaps in the expressions for the information [Eqs. (7) and (10)]. These analytical results for the information are then presented in comparison with the results from simulations.

B. Simulation

The simulations we have performed are for networks of $N=5000$ and 10^4 neurons, which are updated in parallel according to the dynamics in Eq. (1), up to $t=10$ time steps, or when the overlaps converge. Thus we have *almost* stationary states in most cases, except when a state of noninformation is obtained, for which the times of convergence are typically much larger.

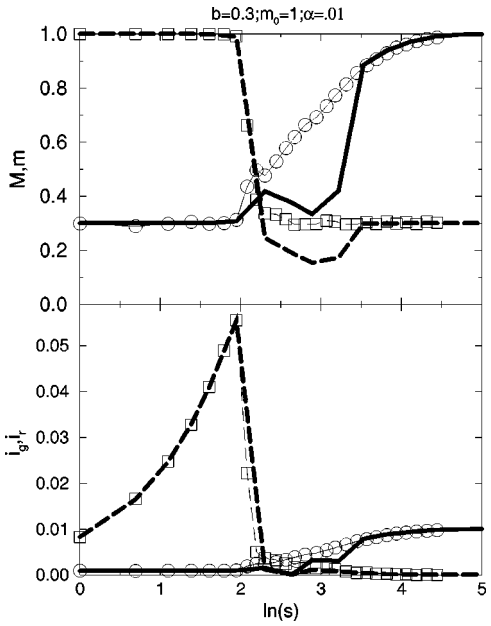

FIG. 1. The overlaps (top) and information (bottom) as functions of ln(S), for $b=0.3$ and $\alpha=0.01$. The squares (circles) are the simulation results for retrieval (categorization) for $N=10^4$ and $t=10$; the dashed (solid) curves are the asymptotic results.

The capacity is analyzed as a function of the two parameters of loading of the network: the rate of loading of concepts, $\alpha=p/N$, and the number of examples per concept, S. The sample averages are taken over an interval in ln(S) or in α. When simulating the information as a function of S, we first generate the concepts and then consecutively store the examples of each concept. When simulating the information as a function of α, we generate the S examples of the concept generated at each step of the learning.

The network is trained then storing examples, while the retrieval and categorization overlaps are monotorized. For a fixed α, it is expected that on increasing S the network passes from a regime where the retrieving information is large to another where the categorizing information increases up to saturation in a upper bound. This behavior is seen in Fig. 1, where the overlaps, as well as the information, are plotted as a function of ln(S), with a correlation $b=0.3$, for a loading of concepts $\alpha=0.01$. When more and more examples are learned, the retrieval information increases until a maximum at $S_R=7$; then it falls down. After a while, when no information is transmitted, the network reaches, at $S_C\sim 33$, the categorization phase, where the categorization information jumps to a higher value. It continues to increase until it saturates at $i_C=0.01$, when the network reaches $M\sim 1$ after $S\sim 90$. The retrieval information capacity of the network is $i_R\sim 0.06$. The asymptotic theory for $N\to\infty$ fits quite well the simulation for $N=10^4$, except in the region of no information. This is due to the finite number of steps used in the dynamical simulation, $t=10$, while the convergence to the fixed point there is very slow.

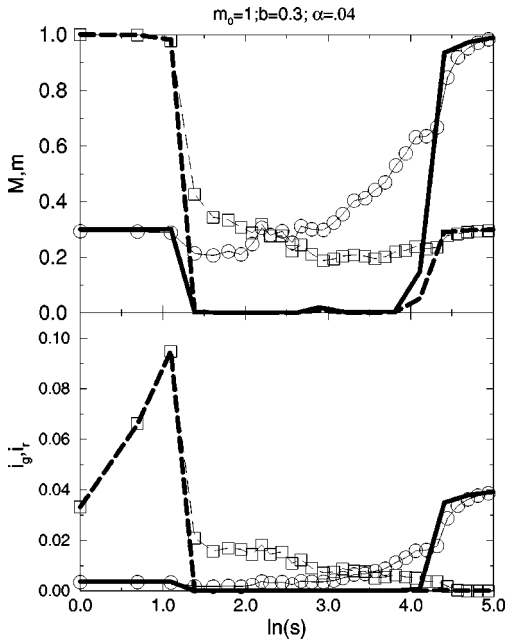

FIG. 2. Same as Fig. 1, for $b=0.3$ and $\alpha=0.04$.

A case with a larger load of concepts, $\alpha=0.04$, is plotted in Fig. 2. Although now the network can only retrieve well examples up to $S=3$ well, it has $i_R\sim 0.10$. Then there is a large waiting period where the information stays close to zero, up to $S_C\sim 74$, when the categorization information jumps to $i_C\sim 0.04$, which is much larger than in the case $\alpha=0.01$.

Comparing this with a network with a larger correlation, $b=0.4$, plotted in Fig. 3, we observe that the network can store only $S_R=2$ examples with a larger overlap, with a maximal retrieval information $i_R\sim 0.05$, which is somewhat smaller than the naive $S\alpha\sim 0.08$. However, the categorization information approaches its saturation value $i_C\sim 0.04$ much faster; only $S\sim 30$ examples must be learned. We have

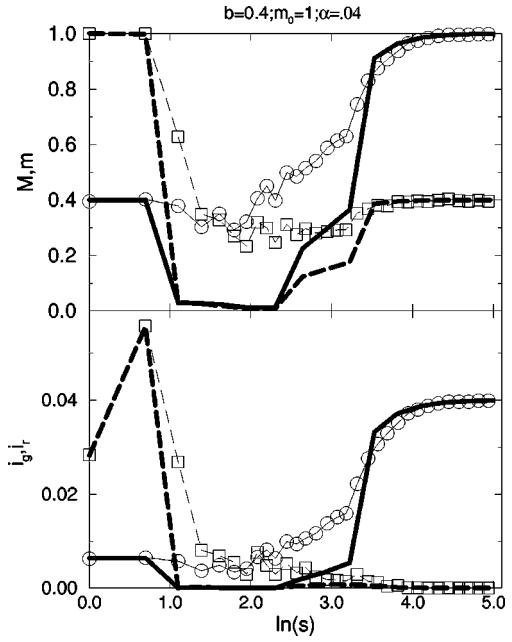

FIG. 3. Same as Fig. 1, for $b=0.4$ and $\alpha=0.04$.

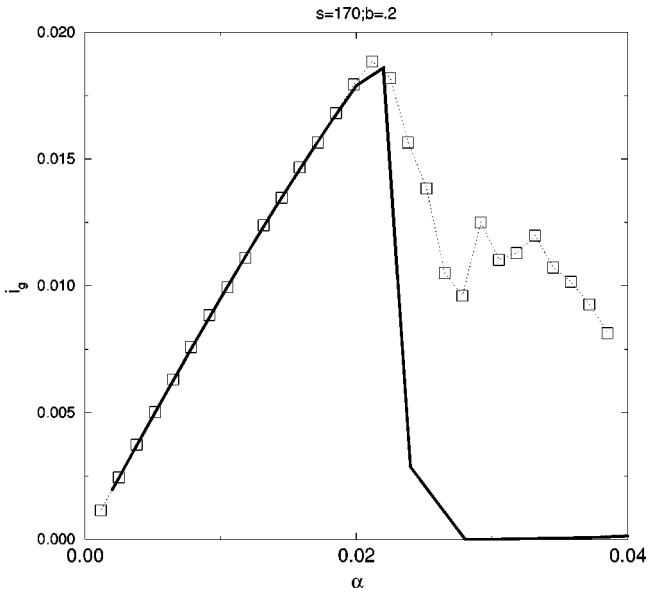

FIG. 4. The categorization information as a function of α, for $b=0.2$ and $S=170$. Asymptotitc (solid), and simulation for $N=5000$ (dashed).

checked that for a larger load of concepts ($\alpha\geqslant 0.06$) the categorization information is larger than the retrieval information. Also we verified that for higher correlations ($b\geqslant 0.6$) the categorization information can be the larger one, even for a small load $\alpha\sim 0.01$, while for smaller correlations ($b\leqslant 0.2$) the retrieval information is always the larger one.

For a fixed S, one expects that on increasing α the categorization information (if b or S are large enough) increases up to a maximum value, after which it decreases until it becomes zero at a critical α. This behavior can be seen in Fig. 4, where the case when $b=0.2$ and $S=170$ is plotted. We verified that the larger the values of b, the higher the maxima of i_C, and less examples are needed. We also observed that the retrieval information has a similar nonmonotonic behavior if b or S are small.

V. CONCLUSION

The information conveyed by the categorization model was studied. It was shown that the transition from the retrieval phase to the categorization phase causes a transition in the information: the retrieval information decreases when the network is oversaturated with examples, and, after a period of resting, the categorization information increases.

It is interesting to note that, although neither the retrieval nor the categorization information surpasses the usual Hopfield model ($S=1$, $b=1$), which is $i_R\sim 0.13$ at $\alpha=0.135$, the fact that the network can return to behave as an associative memory after a long period of *resting* between $S_R<S<S_C$ is an advantage with respect to Hopfield network. It is also worthy of note that the retrieval information can still be relatively large, as we see in Fig. 2, a quotation which to our knowledge has not been observed before in any work about the categorization model in the literature.

The simulation results fit very well with the theoretical results in both retrieval and categorization regimes, showing that almost no effect of finite size is present, but the time of convergence in the resting period must be much larger than

that used in this work. Both expressions for the information about the retrieval and the categorization in Eqs. (10)–(17) are not claimed to be exact. They are approximations for a more precise measure, the *mutual information* [13] between neuron and patterns, $\mathcal{I}[\sigma,\xi]=H[\xi]-\langle H[\sigma|\xi]\rangle_{\xi}$, where $H[\sigma|\xi]$ is the conditional entropy. Since we know that the conditional probability of the neuron, given the concept state, is $p(\sigma|\xi)=(1+M\sigma\xi)\,\delta(|\sigma|^2-1)$, we can replace the categorization information by

$$\mathcal{I}[\sigma,\xi]=\frac{1+M}{2}\,\ln(1+M)+\frac{1-M}{2}\,\ln(1-M). \quad (17)$$

This quantity gives the degree of information the neuron can ''catch'' from the concept. However we prefer to use the estimation in Eq. (7) to compare with the retrieval information with the same precision.

Finally, we hope that the present approach to the information content of a neural network of correlated patterns can be used in the context of more general architectures and learning rules. A more general distribution of the $\lambda_i^{\mu\rho}$ [14] may also deserve some attention.

ACKNOWLEDGMENT

This work was financially supported by the Research Fund of the K.U. Leuven (Grant No. OT/94/9).

[1] H. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[2] N. Parga and M. A. Virasoro, J. Phys. A **47**, 1857 (1986).
[3] J. F. Fontanari, J. Phys. **51**, 2421 (1990).
[4] E. Miranda, J. Phys. I **1**, 999 (1991).
[5] P. R. Krebs and W. K. Theumann, J. Phys. A **26**, 398 (1993).
[6] R. Crisogono, A. Tamarit, N. Lemke, J. Arenzon, and E. Curado, J. Phys. A **28**, 1593 (1995).
[7] D. R. C. Dominguez and W. K. Theumann, J. Phys. A **30**, 1403 (1997).
[8] D. A. Stariolo and F. A. Tamarit, Phys. Rev. A **46**, 5249 (1992).
[9] D. Dominguez and D. Bolle, Phys. Rev. E **56**, 7306 (1997).
[10] D. Dominguez, Phys. Rev. E **54**, 4066 (1996).
[11] C. Rodrigues Neto and J. F. Fontanari, J. Phys. A **31**, 531 (1998).
[12] D. Dominguez and D. Bolle, Phys. Rev. Lett. **80**, 2961 (1998).
[13] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[14] D. R. C. Dominguez and W. K. Theumann, J. Phys. A **29**, 749 (1996).