Hierarchically correlated patterns in Potts neural networks

D. Bollé*,† and J. Huyghebaert*,‡
Instituut voor Theoretische Fysica, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium
(Received 19 December 1994)

The Q-state Potts neural network is extended to allow for storage and retrieval of hierarchically correlated patterns. A Markovian scheme is used for generating the patterns and their ancestors. Two learning rules are considered. The first one is a modified Hebbian learning rule involving a ferromagnetic term. The second one is derived from the pseudoinverse learning rule. Using replica-symmetric mean-field theory, the free energy and the fixed-point equations for the order parameters are derived for general Q and arbitrary temperature T. To compare the performance of both learning rules, the storage capacity and the retrieval quality are calculated for a $Q=3$ network at $T=0$ and different hierarchies of two generations.

PACS number(s): 87.10.+e, 64.60.Cn, 75.10.Hk

I. INTRODUCTION

In many cases of data classification and analysis hierarchical organization is a natural feature. Objects belonging to the same group are strongly correlated while objects sitting in distinct groups are only weakly correlated. The letters of the alphabet, e.g., serve as an example, because they can be organized in classes of letters such as $\{E,F,P,R\}$ and $\{C,G,O,Q\}$.

In the context of multistate Potts neural networks, minimal correlations induced by a so-called bias [1] have been considered before [2,3]. Hierarchical correlations in these multistate models have not yet been treated. This is the purpose of the present contribution.

Several approaches to memorize hierarchically correlated patterns in binary networks have been examined in the past. One procedure is to represent the hierarchical organization explicitly into the spatial structure of the network [4–6]. The couplings between the neurons in different spatial blocks are defined separately from the couplings between the neurons inside the blocks. Another procedure is to take the network spatially homogeneous. The synaptic couplings between all neurons are of the same type, but the patterns are still hierarchically organized in the sense described in the beginning. Such models have been introduced by Parga and Virasoro [7]. They prosed a Markovian scheme to generate the patterns, presented an appropriately generalized Hebb rule and studied the retrieval behavior in the limit of low loading. Related models memorizing an extensive number of patterns have been examined in [8–11] by using the replica approach. In this work, these results are extended to the retrieval of hierarchically correlated patterns in a spatially homogeneous Q-state Potts network. Hereby, two different learning rules are proposed. The first one is a modification of the Hebb rule, the second one is derived from the pseudoinverse rule. Their performance is compared for $Q \leq 3$ models at zero temperature.

The rest of this paper is organized as follows. In Sec. II, the model is described. The two learning rules are introduced in Sec. III and a naive signal-to-noise ratio analysis is performed. Section IV studies the network with the modified Hebb rule involving a ferromagnetic term. Using replica-symmetric mean-field theory, the free energy is written down for general Q and arbitrary T. Relevant order parameters are defined. The retrieval solutions of the fixed-point equations for $Q \leq 3$ models are studied in detail. In particular, for finite loading, $\alpha=0$, the critical temperature for retrieval is calculated and for extensive loading, $\alpha \neq 0$, the storage capacity and the retrieval quality are discussed at $T=0$. In Sec. V, an analogous treatment is given for networks equipped with the learning rule derived from the pseudoinverse rule. Section VI presents some conclusions about the performance of both types of networks. Finally, in the Appendix some details are given on the derivation of the learning rule used in Sec. V.

II. MODEL

Consider a system of N neurons. Each neuron can be described by a Potts spin $\sigma_i \in \{1,2,\ldots,Q\}$, $i=1,2,\ldots,N$. The neurons are interconnected with all the others by a synaptic matrix of strength J_{ij}^{kl} which determines the contributions of a signal fired by the jth presynaptic neuron in state l to the postsynaptic potential that acts on the ith neuron in state k. The energy potential h_{i,σ_i} of neuron i, which is in a state σ_i, is given by

$$h_{i,\sigma_i} = -\sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{k,l=1}^{Q} J_{ij}^{kl} u_{\sigma_i,k} u_{\sigma_j,l} \ , \tag{1}$$

with u the Potts spin operator defined as $u_{\sigma_i,k} = Q\delta_{\sigma_i,k} - 1$. We assume that the synaptic couplings are sym-

*Also at Interdisciplinair Centrum voor Neurale Netwerken, K.U. Leuven, Leuven, Belgium.
†Electronic address: Desire.Bolle@fys.kuleuven.ac.be
‡Electronic address: Jacky.Huyghebaert@fys.kuleuven.ac.be

metric, i.e., $J_{ij}^{kl}=J_{ji}^{lk}$. The dynamics of the Q-state Potts model is defined as in [3] and [12]. At zero temperature the state of the neuron in the next time step is fixed to be the state that minimizes the induced local energy (1). The stable states of the system are those configurations $\{\sigma_i\}$ where every neuron is in a state that gives a minimum value to $\{h_{i,\sigma_i}\}$. For symmetric couplings this stability is equivalent to the requirement that the configurations $\{\sigma_i\}$ are the local minima of the Potts Hamiltonian

$$H=-\frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{k,l=1}^{Q}J_{ij}^{kl}u_{\sigma_i,k}u_{\sigma_j,l}\ . \tag{2}$$

In the presence of noise there is a finite probability of having configurations other than the local minima. This can be taken into account by introducing an effective temperature $T=1/\beta$.

To build in the capacity for learning and memory in this network, its stationary configurations representing the retrieved patterns must be correlated with the stored patterns $\{\xi^{\mu}\}$, $\mu=1,2,\ldots,p$ fixed by the learning process. The latter are allowed to be hierarchically correlated. To generate this set of patterns we follow a procedure analogous to that proposed by Gutfreund [11]. Suppose that the p patterns are organized in a hierarchical structure of L levels with $p_1\times p_2\times\cdots\times p_l$, $1\le l\le L$ patterns at the lth level. Here p_k denotes the number of descendants of a pattern of generation $k-1$. Hereby it is assumed that all patterns of the same generation have an equal number of descendants. At the lth generation level, the patterns are grouped in $p_1\times\cdots\times p_{l-1}$ classes, each containing p_l members. To identify a pattern of generation l we use l numbers $\mu_1\cdots\mu_l\equiv\bar{\mu}_l$ indicating how to go down in the hierarchy. An example of such a structure with two levels is shown in Fig. 1.

The hierarchical structure is generated by a Markov process in the flowing way. The oldest ancestor is generated first. It determines the probability distribution of the p_1 patterns at the first generation level. The latter in turn define the probability distribution of the $p_1\times p_2$ patterns at the next generation and so on. So, each generation is determined by its ancestors only.

The conditional probability to generate the patterns of level l, given the patterns at level $l-1$, is defined by

$$P_l(k,k')\equiv P(\xi_i^{\bar{\mu}_l}=k\,|\,\xi_i^{\bar{\mu}_{l-1}}=k')=\frac{1+B_{k,k'}^{(l)}}{Q}\ , \tag{3}$$

where $1\le l\le L$, $k,k'\in\{1,2,\ldots,Q\}$, and $\xi_i^0=1$, for all $i\in\{1,2,\ldots,N\}$. The $B_{k,k'}^{(l)}$ are considered as bias parameters. In contrast with [2,3], the bias parameters are now elements of a $Q\times Q$ matrix $[B^{(l)}]$. Since the $P_l(k,k')$ are probabilities, the matrix elements $B_{k,k'}^{(l)}$ have to satisfy

$$-1\le B_{k,k'}^{(l)}\le Q-1$$

and

$$\sum_{k=1}^{Q}B_{k,k'}^{(l)}=\sum_{k'=1}^{Q}B_{k,k'}^{(l)}=0\ . \tag{4}$$

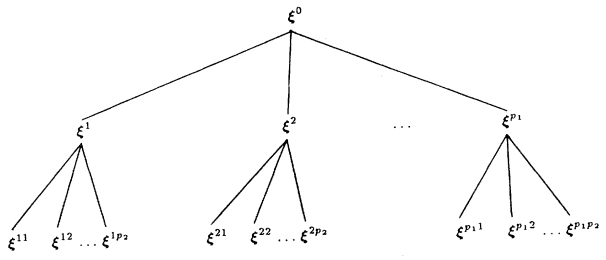

FIG. 1. Hierarchical structure consisting of two levels ($L=2$). The patterns $\xi^{\mu_1\mu_2},\mu_1=1,\ldots,p_1;\mu_2=1,\ldots,p_2$ belong to the second level, the patterns $\xi^{\mu_1},\mu_1=1,\ldots,p_1$ belong to the first one. The pattern ξ^0 is the oldest ancestor.

To ensure that a component $\xi_i^{\bar{\mu}_l}$ of a pattern at generation l has a higher probability to be equal to its nearest ancestor $\xi_i^{\bar{\mu}_{l-1}}$ than to be equal to any other state, it is assumed that

$$B_{k,k}^{(l)}\ge B_{k,k'}^{(l)}\quad\text{with } k\neq k'\ . \tag{5}$$

It is convenient to rewrite the matrix $[B^{(l)}]$ in the following form:

$$[B^{(l)}]=a_l[b^{(l)}]\ , \tag{6}$$

where $0\le a_l\le 1$. We call a_l the bias amplitude and the $Q\times Q$ matrix $[b^{(l)}]$ the bias structure. We recall that the elements of the bias matrix $[b^{(l)}]$ have been chosen in accordance with the conditions (4) and (5).

Generating the patterns according to the probability distribution (3) permits us to classify them in a hierarchical structure. Indeed, we first define the correlation matrix $[C]$ with elements

$$C_{\bar{\mu}_m,\bar{\nu}_n}=\frac{1}{(Q-1)N}\sum_{i=1}^{N}u_{\xi_i^{\bar{\mu}_m},\xi_i^{\bar{\nu}_n}}\ . \tag{7}$$

We remark that for equal indices $m=n$, correlations within the same level are measured whereas for $m\neq n$, correlations between different levels are indicated. For $m=n=1$, the correlations are taken to be zero in the corresponding binary networks of [7–10] but not in [11]. Here we also allow those to be nonzero.

Then, using the conditions (4) and (5), it is easy to prove the following inequalities:

$$C_{\bar{\mu}_m\mu_{m+1}\cdots\mu_l,\bar{\mu}_m\mu_{m+1}\cdots\mu_{l-1}} \ge C_{\bar{\mu}_m\mu_{m+1}\cdots\mu_l,\bar{\mu}_m\nu_{m+1}\cdots\nu_{l-1}} \tag{8}$$

$$C_{\bar{\mu}_{l-1}\mu_l,\bar{\mu}_{l-1}\nu_l}\ge\cdots\ge C_{\bar{\mu}_m\mu_{m+1}\cdots\mu_l,\bar{\mu}_m\nu_{m+1}\cdots\nu_l} \ge\cdots\ge C_{\bar{\mu}_l,\bar{\nu}_l}\ . \tag{9}$$

The first inequality (8) expresses the fact that a pattern is more strongly correlated with its own ancestor than with the other ancestors. The other inequalities (9) indicate that within the same level patterns of the same class are more strongly correlated than patterns from distinct classes. In particular, the larger the distance from two

patterns to their nearest common ancestor, the smaller their correlations. We notice that a level l the patterns are classified in $l-1$ classes, leading to l different values for the correlation coefficients, in contrast with the model for biased patterns studied in [3].

In the sequel, we consider a hierarchical structure of two generations. The first generation consists of a finite number p_1 of ancestor patterns. The second generation has a finite number p_1 of classes, each containing an extensive number p_2 of patterns. Only the patterns of the second level are interpreted as patterns that have to be retrieved. The other patterns are just artificial constructions permitting an efficient classification of the patterns of the second level.

III. LEARNING RULES AND SIGNAL-TO-NOISE ANALYSIS

To store the $p=p_1+p_1p_2$ patterns defined above, we propose two different learning rules. First, we consider the learning rule

$$J_{ij}^{kl}=\frac{1}{Q^2N}\left\{\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}(u_{\xi_i^{\mu_1\mu_2},k}-B_{\xi_i^{\mu_1},k})\times(u_{\xi_j^{\mu_1\mu_2},l}-B_{\xi_j^{\mu_1},l})+\epsilon_1\sum_{\mu_1=1}^{p_1}(u_{\xi_i^{\mu_1},k}-B_k)(u_{\xi_j^{\mu_1},l}-B_l)+\gamma u_{k,l}\right\}. \tag{10}$$

The patterns of the first generation are denoted by $\xi^{\mu_1},\mu_1=1,\ldots,p_1$, the patterns of the second generation by $\xi^{\mu_1\mu_2},\mu_1=1,\ldots,p_1;\ \mu_2=1,\ldots,p_2$ where $\mu_1\mu_2\equiv\bar{\mu}_2$. Furthermore, $B_{k,k'}\equiv B_{k,k'}^{(2)}$ and $B_k\equiv B_{k,1}^{(1)}$ specify the matrices $[B^{(2)}]$ and $[B^{(1)}]$. The latter is in fact a Q-dimensional vector.

The coefficients in (10) can be chosen freely. The following choices are made: $\epsilon_2=1$, $\epsilon_1=\gamma=0$, indicating that only the patterns of the second generation are memorized, $\epsilon_2=\epsilon_1=1$, $\gamma=0$ telling that patterns of both generations are stored and

$$\epsilon_2=(1-K_2)^{-1}\quad \epsilon_1=\left[1-\frac{1}{Q(Q-1)}\sum_{k=1}^{Q}B_k^2\right]^{-1}$$

and

$$\gamma=Q-1\ , \tag{11}$$

with

$$K_2=\frac{1}{Q(Q-1)}\sum_{k=1}^{Q}P(k)\sum_{k'=1}^{Q}B_{k,k'}^2 \tag{12}$$

and $P(k)=(1+B_k)/Q$. The last choice explicitly includes a ferromagnetic term and is motivated by the naive signal-to-noise analysis studied below.

In choosing these storage prescriptions we were inspired by the work of Gutfreund [11] and Buhmann, Divko, and Schulten [13] on the $Q=2$ model. Gutfreund has shown that an appropriate bias term has to be subtracted to ensure the storage of an extensive number of patterns at a certain generation of the hierarchy. In the work of Buhmann, Divko, and Schulten it is shown that the retrieval properties of neural networks are improved by introducing a ferromagnetic term and by choosing appropriate values for the prefactors of the other terms.

Second, the following learning rule is proposed:

$$J_{ij}^{kl}=\frac{1}{Q^2N}\left\{\frac{1}{1-K_2}\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}(u_{\xi_i^{\mu_1\mu_2},k}-B_{\xi_i^{\mu_1},k})(u_{\xi_j^{\mu_1\mu_2},l}-B_{\xi_j^{\mu_1},l})+\frac{1}{K_2-K_1}\sum_{\mu_1=1}^{p_1}\left[B_{\xi_i^{\mu_1},k}-\sum_{\rho=1}^{Q}P(\rho)B_{\rho,k}\right]\left[B_{\xi_j^{\mu_1},l}-\sum_{\rho=1}^{Q}P(\rho)B_{\rho,l}\right]+\frac{1}{K_1}\sum_{\rho=1}^{Q}P(\rho)B_{\rho,k}\sum_{\rho=1}^{Q}P(\rho)B_{\rho,l}\right\}, \tag{13}$$

where K_2 is given by (12) and

$$K_1=\frac{1}{Q(Q-1)}\sum_{k=1}^{Q}\left[\sum_{k'=1}^{Q}P(k')B_{k,k'}\right]^2. \tag{14}$$

Extending the method in [14], this rule is derived from the multistate generalization of the pseudoinverse learning rule [15,16] by expanding the inverse correlation matrix $[C^{-1}]$ in powers of $1/p_2$ and keeping the terms up to order $1/p_2^2$. Consequently, we name this rule the truncated pseudoinverse learning rule. A sketch of this calculation is given in the Appendix. In comparison with the pseudoinverse learning rule itself, (13) has the advantage that it is not required to invert the correlation matrix $[C]$. Inverting $[C]$ is a huge numerical problem, because it is a $p\times p$ matrix with p growing linearly with the system size N. Similarly as in (10), the patterns of both the first and the second generation are memorized. We remark that here the prefactors of each term are fixed.

At this point it is interesting to note the following. For a two-state model, the synaptic couplings (10) reduce to

$$J_{ij}=\frac{1}{N}\left\{\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}(\xi_i^{\mu_1\mu_2}-\xi_i^{\mu_1}a_2)(\xi_j^{\mu_1\mu_2}-\xi_j^{\mu_1}a_2)+\epsilon_1\sum_{\mu_1=1}^{p_1}(\xi_i^{\mu_1}-a_1)(\xi_j^{\mu_1}-a_1)+\gamma\right\}, \tag{15}$$

where $\xi_i^{\mu_1\mu_2},\xi_i^{\mu_1}=\pm 1$. For $\epsilon_2=1$ and $\epsilon_1=\gamma=0$, the learning rule reduce to the one in [11]. Taking $\epsilon_2=1/[1-a_2^2]$, $\epsilon_1=1$, $\gamma=a_1=0$, and making the substitution $\xi_i^{\mu_1}a_2\to\xi_i^{\mu_1}$ on the ancestor patterns, the model studied in [8,9] is recovered. Choosing $\epsilon_2=1/(1-a_2^2)$, $\epsilon_1=1/(1-a_1^2)$,and $\gamma=1$, the learning rule (15) completely coincides with the $Q=2$ limit of the rule (13).

The difference in storage properties between both learning rules is already suggested by a naive signal-to-noise analysis. For an arbitrary pattern of the second generation, say $\xi^{\lambda\lambda'}$, stored with the learning rule (10), the local energy at neuron i in the limit $N\to\infty$ is given by

$$\begin{aligned}h_i(\sigma_i)=&-\epsilon_2(u_{\xi_i^{\lambda\lambda'},\sigma_i}-B_{\xi_i^{\lambda},\sigma_i})[Q-1-\langle\langle B_{\xi^{\lambda},\xi^{\lambda\lambda'}}\rangle\rangle]\\&-\epsilon_1(u_{\xi_i^{\lambda},\sigma_i}-B_{\sigma_i})\langle\langle u_{\xi^{\lambda},\xi^{\lambda\lambda'}}-B_{\xi^{\lambda\lambda'}}\rangle\rangle\\&-\gamma\langle\langle u_{\sigma_i,\xi^{\lambda\lambda'}}\rangle\rangle\end{aligned} \tag{16}$$

where $\langle\langle\ \rangle\rangle$ denotes an average over the patterns ξ^{λ} and $\xi^{\lambda\lambda'}$. From this equation it follows that only for appropriate values of the coefficients ϵ_1, ϵ_2, and γ and a specific choice of the bias the pure signal $-(Q-1)u_{\xi_i^{\lambda\lambda'},\sigma_i}$ is restored. Indeed, precisely for the choice (11) we get

$$h_i(\sigma_i)=-(Q-1)\left\{(u_{\xi_i^{\lambda\lambda'},\sigma_i}-B_{\xi_i^{\lambda},\sigma_i})+(u_{\xi_i^{\lambda},\sigma_i}-B_{\sigma_i})\times\frac{\langle\langle u_{\xi^{\lambda},\xi^{\lambda\lambda'}}-B_{\xi^{\lambda\lambda'}}\rangle\rangle}{Q-1-\frac{1}{Q}\sum_{k=1}^{Q}B_k^2}+\langle\langle u_{\sigma_i,\xi^{\lambda\lambda'}}\rangle\rangle\right\}. \tag{17}$$

such that the local energy $h_i(\sigma_i)$ equals the pure signal term for a second generation bias matrix of the form $B_{k,k'}=a_2u_{k,k'}$, $0\le a_2\le 1$, whatever the choice for the bias B_k in the first generation.

In the case of the truncated pseudoinverse learning rule (13) the local energy at neuron i for the pattern $\xi^{\lambda\lambda'}$ is

$$\begin{aligned}h_i(\sigma_i)=&-\frac{1}{1-K_2}(u_{\xi_i^{\lambda\lambda'},\sigma_i}-B_{\xi_i^{\lambda},\sigma_i})[Q-1-\langle\langle B_{\xi^{\lambda},\xi^{\lambda\lambda'}}\rangle\rangle]\\&-\frac{1}{K_2-K_1}\left[B_{\xi_i^{\lambda},\sigma_i}-\sum_{\rho=1}^{Q}P(\rho)B_{\rho,\sigma_i}\right]\\&\times\left\langle\!\!\left\langle B_{\xi^{\lambda},\xi^{\lambda\lambda'}}-\sum_{\rho=1}^{Q}P(\rho)B_{\rho,\xi^{\lambda\lambda'}}\right\rangle\!\!\right\rangle\\&-\frac{1}{K_1}\sum_{\rho=1}^{Q}P(\rho)B_{\rho,\sigma_i}\sum_{\rho=1}^{Q}P(\rho)\langle\langle B_{\rho,\xi^{\lambda\lambda'}}\rangle\rangle .\end{aligned} \tag{18}$$

Working out the average $\langle\langle\ \rangle\rangle$ it becomes clear that the pure signal term is restored for all possible bias types, in contrast with the foregoing rule (10).

In the case of extensive loading of patterns in the $Q=2$ model with any of these learning rules the variance of the noise caused by the other patterns equals $\sqrt{\alpha}$, exactly as in the unbiased case. So we expect a storage capacity $\alpha_c=0.1379$ independent of the bias. Since for $Q\ge 3$ the variance of that noise depends explicitly on the bias, the storage capacity will also depend on it. These assertions are verified by the numerical study of the replica-symmetric fixed-point equations presented in the next section.

IV. THE MODEL WITH A FERROMAGNETIC TERM

A. Replica-symmetric mean-field theory

To study in detail the neural network model defined by (2), which has learned a two-level hierarchical set of patterns, we use the standard replica-symmetric mean-field theory. This means that we have to calculate the free energy of the model and derive the corresponding fixed-point equations for the order parameters.

Since both (10) and (13) memorize the ancestors patterns and their descendants, we have to look for solutions representing a macroscopic overlap with patterns of both generations. Consequently, it is assumed that all the patterns of the first generation and a finite number s_2 of the p_2 patterns in each class of the second generation ($s_2<p_2$) are condensed.

After some algebra we arrive at the following expression for the free energy density for the model with the ferromagnetic term (10):

$$\begin{aligned}f=&\frac{1}{2}\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{s_2}m_{\mu_1\mu_2}^2+\frac{1}{2}\epsilon_1\sum_{\mu_1=1}^{p_1}m_{\mu_1}^2+\frac{\gamma}{2Q}\sum_{l=1}^{Q}M_l^2\\&+\frac{1}{2}\alpha\epsilon_2^2\beta(\bar{r}\bar{q}-rq)+\frac{1}{2}\alpha\epsilon_2\bar{q}-\frac{\alpha\epsilon_2 q}{2[1-\beta\epsilon_2(\bar{q}-q)]}\\&+\frac{\alpha}{2\beta}\ln[1-\beta\epsilon_2(\bar{q}-q)]\\&-\frac{1}{\beta}\left\langle\!\!\left\langle\int_{\mathbb{R}^{Q\times Q}}Dz\ln\left[\sum_{\sigma=1}^{Q}\exp[\beta\mathcal{H}_\sigma(z,\xi)]\right]\right\rangle\!\!\right\rangle ,\end{aligned} \tag{19}$$

where $\alpha=p_1p_2/N$ is the loading of the second generation and with $\langle\langle\ \rangle\rangle$ denoting the average over the p_1 ancestor patterns and the p_1s_2 condensed patterns of the second generation. The integral is taken over a $Q\times Q$ dimensional space and the Gaussian measure $D\mathbf{z}$ is given by

$$D\mathbf{z}=\prod_{k,k'=1}^{Q}(dz_{kk'}/\sqrt{2\pi})\exp(-z_{kk'}^2/2) \tag{20}$$

and $\mathcal{H}_\sigma(\mathbf{z},\xi)$ reads

$$\mathcal{H}_\sigma(\mathbf{z},\xi)=\epsilon_2\sum_{k,k'=1}^{Q}\sqrt{\alpha rP(k)P(k,k')}(u_{k',\sigma}-B_{k,\sigma})z_{kk'}+\frac{1}{2}\alpha\beta\epsilon_2(\bar{r}-r)\sum_{k,k'=1}^{Q}P(k)P(k,k')(u_{k',\sigma}-B_{k,\sigma})^2$$
$$+\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{s_2}(u_{\xi^{\mu_1\mu_2},\sigma}-B_{\xi^{\mu_1},\sigma})m_{\mu_1\mu_2}+\epsilon_1\sum_{\mu_1=1}^{p_1}(u_{\xi^{\mu_1},\sigma}-B_\sigma)m_{\mu_1}+\frac{\gamma}{Q}\sum_{l=1}^{Q}u_{l,\sigma}M_l\ . \tag{21}$$

This free energy (19) depends on the following order parameters:

$$m_{\mu_1\mu_2}=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle u_{\xi_i^{\mu_1\mu_2},\sigma_i}-B_{\xi_i^{\mu_1},\sigma_i}\rangle\rangle\rangle\ , \tag{22}$$

$$m_{\mu_1}=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle u_{\xi_i^{\mu_1},\sigma_i}-B_{\sigma_i}\rangle\rangle\rangle\ , \tag{23}$$

$$M_l=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle u_{l,\sigma_i}\rangle\rangle\rangle\ , \tag{24}$$

$$q=\frac{1}{N}\sum_{i=1}^{N}\sum_{k,k'=1}^{Q}P(k)P(k,k')\langle\langle\langle u_{k',\sigma_i}-B_{k,\sigma_i}\rangle^2\rangle\rangle\ , \tag{25}$$

$$\tilde{q}=\frac{1}{N}\sum_{i=1}^{N}\sum_{k,k'=1}^{Q}P(k)P(k,k')\langle\langle\langle(u_{k',\sigma_i}-B_{k,\sigma_i})^2\rangle\rangle\rangle\ , \tag{26}$$

$$r=\frac{1}{\alpha}\sum_{\mu_1=1}^{p_1}\sum_{\nu_2=s_2+1}^{p_2}\left\langle\!\!\left\langle\left[\frac{1}{N}\sum_{i=1}^{N}\langle u_{\xi_i^{\mu_1\nu_2},\sigma_i}-B_{\xi_i^{\mu_1},\sigma_i}\rangle\right]^2\right\rangle\!\!\right\rangle\ , \tag{27}$$

$$\bar{r}=\frac{1}{\alpha}\sum_{\mu_1=1}^{p_1}\sum_{\nu_2=s_2+1}^{p_2}\left\langle\!\!\left\langle\!\!\left\langle\left[\frac{1}{N}\sum_{i=1}^{N}(u_{\xi_i^{\mu_1\nu_2},\sigma_i}-B_{\xi_i^{\mu_1},\sigma_i})\right]^2\right\rangle\!\!\right\rangle\!\!\right\rangle-\frac{1}{\beta\epsilon_2}\ , \tag{28}$$

where $\langle\ \rangle$ stands for the thermal average.

The order parameter $m_{\mu_1\mu_2}$, which is a component of the p_1s_2-dimensional vector $\mathbf{m}^{(2)}$, measures the overlap of the network configuration with a pattern $\xi^{\mu_1\mu_2}$ of the second generation. The overlap with a pattern ξ^{μ_1} of the first generation is presented by m_{μ_1}. It is a component of the p_1-dimensional vector $\mathbf{m}^{(1)}$. In both definitions the bias has been subtracted to account for the random overlap caused by the nonuniform probability distribution (3).

The order parameter M_l gives information about the number of neurons in the state l. Its value varies between -1 and $Q-1$.

The order parameter q is the extended Edwards-Anderson order parameter, which measures the correlations between the neurons. In contrast with the Hopfield model with hierarchically organized patterns [8–11] it explicitly contains the bias. For $Q=2$, using $u_{k,\sigma_i}=k\sigma_i$, $B_{k,\sigma_i}=k\sigma_i a_2$, and $B_{\sigma_i}=\sigma_i a_1$, where the k and σ_i are now Ising spins ±1, Eq. (25) reduces to

$$q=(1-a_2^2)\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle\sigma_i^2\rangle\rangle\rangle\ . \tag{29}$$

Hence the multiplicative bias factor $(1-a_2^2)$ can be taken out of the definition [3].

The order parameter $\tilde{q}$ measures the autocorrelation of the neurons. For the $Q=2$ model, $\tilde{q}$ plays no role as an order parameter since it takes the constant value $\tilde{q}=1-a_2^2$.

The order parameters r and $\bar{r}$ give information about the overlap with the noncondensed patterns, i.e., r represents the total mean-square random overlap with these patterns and $\bar{r}$ measures their total autocorrelation.

These order parameters satisfy fixed-point equations obtained in the standard way by taking the relevant derivatives of the free energy. Their retrieval solutions correspond to overlap vectors of the form $\mathbf{m}^{(1)}=(m_1,0,\ldots,0)$ and $\mathbf{m}^{(2)}=(m_2,0,\ldots,0)$ with $m_2\gg m_1>0$. Since each pattern of the second generation has an ancestor at the first generation, the network configuration has a macroscopic overlap with both patterns. Due to the condition $m_2\gg m_1>0$, the above mentioned solution represents retrieval of a pattern of the second generation. The maximal value of m_2, representing perfect overlap with a pattern of the second generation, reads

$$m_2=Q-1-\frac{1}{Q}\sum_{k,k'=1}^{Q}\frac{1+B_k}{Q}B_{k,k'}^2\ . \tag{30}$$

The corresponding overlap with the ancestor pattern is then given by

$$m_1=\sum_{k=1}^{Q}\frac{1+B_k}{Q}B_{k,k}-\frac{1}{Q^2}\sum_{k,k'=1}^{Q}B_kB_{k,k'}B_{k'} \tag{31}$$

which can be shown [by using Eqs. (4) and (5)] to be smaller than the overlap m_2. We remark that solutions of the form $\mathbf{m}^{(1)}=\mathbf{0}$ and $\mathbf{m}^{(2)}=(m_2,0,\ldots,0)$ with $m_2\neq0$ do not appear.

B. Discussion of the retrieval properties

To illustrate the retrieval properties of these networks we have solved numerically the fixed-point equations for the retrieval solutions of $Q \leq 3$ models at $T=0$. In particular, we have calculated the storage capacity as a function of the bias amplitude and the retrieval quality as a function of the loading $\alpha = p_1 p_2 / N$ of the second generation.

For the $Q=3$ model we choose two representative bias matrices, i.e.,

$$[B_1] = a \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}, \qquad [B_2] = a \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{32}$$

with $0 \leq a \leq 1$. For simplicity we have taken $a \equiv a_1 = a_2$. Since the diagonal elements of both matrices are larger than the nondiagonal ones, $\xi_i^{\mu_1 \mu_2}$ has a higher probability to be equal to its ancestor $\xi_i^{\mu_1}$ than to be equal to any other state. For $[B_1]$ all the other states have equal probability, for $[B_2]$ they have different probability. At this point we notice that $[B_1]$ is of the specific form $B_{k,k'} = a u_{k,k'}$ restoring the pure signal term in the signal-to-noise analysis of the local field corresponding with learning rule (10) [see Eq. (17)].

Due to the conditions (4) and (5) the bias matrix of the $Q=2$ model is fixed and given by

$$[B] = a \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{33}$$

Let us start with a finite number of patterns in both the first and second generation. If there are only a finite number of patterns memorized, the free energy reduces to the $\alpha \to 0$ limit of (19), which depends on $m_{\mu_1 \mu_2}$, m_{μ_1}, and M_l. Solving numerically the appropriate fixed-point equations teaches one that the retrieval solutions exist up to a critical temperature T_c. At T_c the system has a second order ($Q=2$) or first order ($Q \geq 3$) phase transition to the paramagnetic phase. We find that the behavior of T_c as a function of the bias amplitude a is qualitatively the same for the $Q=2$ and $Q \geq 3$ networks.

In the following, we focus attention on the $Q=3$ model. At $a=0$ we find that $T_c = 2.185$. For the models without the ferromagnetic term, i.e., $\gamma = 0$, T_c decreases rapidly with increasing a, similarly as in [2]. For the bias type $[B_1]$, T_c becomes even zero at $a=1$.

Introducing the ferromagnetic term and choosing the coefficients ϵ_2, ϵ_1, and γ according to Eq. (11) increases T_c substantially. For the choice $[B_1]$, $T_c = 2.185$ whatever the value of the bias amplitude a. This result can be proved analytically. Indeed the fixed-point equations of the retrieval solutions of the network with $[B_1]$ type patterns are equivalent with those of the Potts model with unbiased patterns [17]. However, for $[B_2]$ type patterns, T_c slightly depends on a. This behavior is also in agreement with the signal-to-noise analysis since for all possible choices of the bias matrices, except $[B_1]$, the signal term (17) depends on the bias amplitude.

Next we allow an extensive number of patterns in each class of the second generation. First, we investigate the effect of storing, in addition, the (finite number of) ancestor patterns. Therefore, we turn to Fig. 2 indicating the maximal value of the loading of the second generation, α_c, for which retrieval solutions exist. The dotted lines correspond with the network that only stores patterns of the second generation ($\epsilon_2 = 1$ and $\epsilon_1 = \gamma = 0$). The solid curves denote the model that memorizes the patterns of both generations ($\epsilon_2 = \epsilon_1 = 1$ and $\gamma = 0$). Comparing these two situations, we conclude that storing also the ancestor patterns increases the storage capacity. Clearly, if also the ancestors patterns ar memorized, it is easier for the network to discriminate between the different classes.

Comparing the two different choices of bias matrices we see that in all cases the $[B_2]$ model has the largest storage capacity. This seems to be consistent with the fact that this model has the weakest correlations (7). In particular in the neighborhood of $a=1$, the enhancement is very small for the $[B_1]$ model, whereas it is still significant for the $[B_2]$ model. The corresponding curves for the $Q=2$ model show a similar behavior as for the $Q=3$ network with $[B_1]$ type patterns.

Second, we study the consequences of introducing the ferromagnetic term and choosing appropriate values for the coefficients ϵ_2, ϵ_1, and γ [see (11)]. The dashed lines in Fig. 2 denote the model with the ferromagnetic term and the appropriate prefactors (11). We see that α_c is increased by introducing the ferromagnetic term. In particular, for the most extreme choice of the bias matrix, i.e., $[B_1]$, the storage capacity is still substantially different from zero at $a=1$. However, its maximal value $\alpha_c = 0.4144$ at $a=0$ is never reached. We remark that the storage capacity of the corresponding $Q=2$ model, however, is independent of the bias amplitude: $\alpha_c(a \geq 0) = 0.1379$. This different behavior is in agreement with the signal-to-noise analysis. We further notice that the network loaded with $[B_2]$ type patterns has a higher storage capacity than the network loaded with $[B_1]$ patterns. This is in agreement with the fact that the former model corresponds with the weakest correlations (7).

The influence on the retrieval quality becomes clear from Figs. 3 and 4. They show the overlap m_1 and m_2 as a function of the loading α of the second generation. For $a=0$, the corresponding overlap diagram for the Potts model with unbiased patterns [3] is recovered. We see that the overlap m_2 is nearly maximal over a long interval in α, indicating almost perfect retrieval. Increasing a, the overlap diagram retains a similar form, but the maximal overlap decreases. We notice that the maximal values of m_2 and m_1 are exactly given by expressions (30) and (31), respectively. The overlap for the $[B_1]$ model decreases more quickly as a function of a. For each value of α, m_2 is substantially larger than m_1 indicating retrieval of a pattern of the second generation. Both figures show that the model with the ferromagnetic term

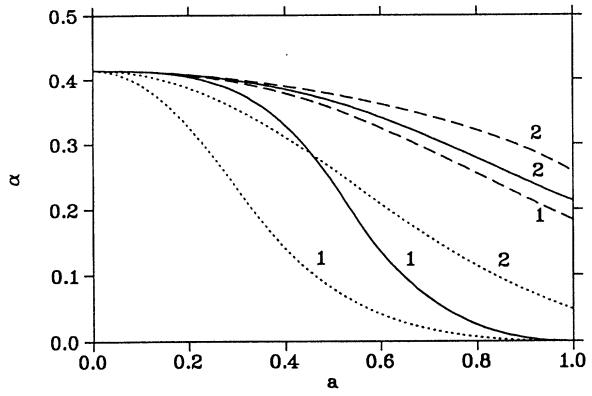

FIG. 2. The storage capacity at $T=0$ as a function of the bias amplitude a for the $Q=3$ network with the rule (10) and with $[B_1]$ (1) and $[B_2]$ (2) type patterns [see Eq. (32)]. The dotted (solid and dashed, respectively) curves correspond with $\epsilon_2=1$ and $\epsilon_1=\gamma=0$ [$\epsilon_2=\epsilon_1=1$ and $\gamma=0$, respectively, a ferromagnetic term according to (11)]. It is seen that storing also the ancestor patterns and introducing a ferromagnetic term improves the storage capacity.

has the largest overlap, indicating that its retrieval quality is the highest.

In this section, we have studied the ability of Potts neural networks to retrieve hierarchically correlated patterns, which are memorized with the learning rule (10). We have seen that memorizing the ancestor patterns of the first generation enhances the storage capacity of the patterns of the second generation. Furthermore, we have shown that the presence of a ferromagnetic term and the choice of appropriate prefactors [see Eq. (11)] even further increases this storage capacity. It also gives rise to the best retrieval quality.

V. THE TRUNCATED PSEUDOINVERSE MODEL

A. Replica-symmetric mean-field theory

The study of the neural network model defined by (2) and the truncated pseudoinverse rule (13) proceeds via the replica-symmetric mean-field approximation of the free energy density which reads

$$f=\frac{1}{2}\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{s_2}m_{\mu_1\mu_2}^2+\frac{1}{2}\epsilon_1\sum_{\mu_1=1}^{p_1}m_{\mu_1}^2+\frac{\gamma}{2}M^2+\tfrac{1}{2}\alpha\epsilon_2^2\beta(\bar{r}\bar{q}-rq)+\tfrac{1}{2}\alpha\epsilon_2\tilde{q}-\frac{\alpha\epsilon_2 q}{2[1-\beta\epsilon_2(\tilde{q}-q)]}+\frac{\alpha}{2\beta}\ln[1-\beta\epsilon_2(\tilde{q}-q)]$$
$$-\frac{1}{\beta}\left\langle\!\!\left\langle\int_{\mathbf{R}^{Q\times Q}}D\mathbf{z}\ln\left[\sum_{\sigma=1}^{Q}\exp[\beta\mathcal{H}_\sigma(\mathbf{z},\boldsymbol{\xi})]\right]\right\rangle\!\!\right\rangle \tag{34}$$

where

$$\epsilon_2=\frac{1}{1-K_2}\ ,\quad \epsilon_1=\frac{1}{K_2-K_1}\ ,\quad \gamma=\frac{1}{K_1}\ , \tag{35}$$

and $\mathcal{H}_\sigma(\mathbf{z},\boldsymbol{\xi})$ is given by

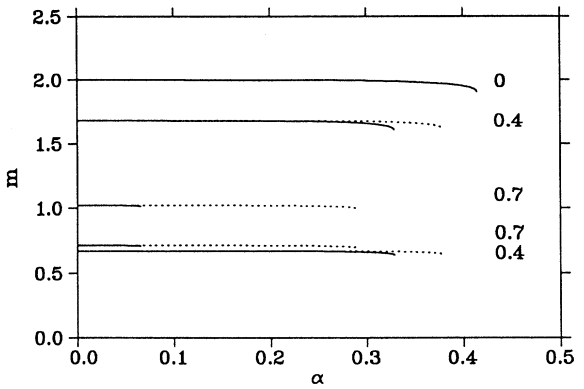

FIG. 3. The overlap m_2 (upper five curves) and m_1 (lower four curves) at $T=0$ as a function of α for the $Q=3$ $[B_1]$ model with the rule (10) for different values of the bias amplitude a. The dotted (solid) lines represent the model with (without) the ferromagnetic term and appropriate coefficients (11). A ferromagnetic term is seen to enhance the retrieval quality.

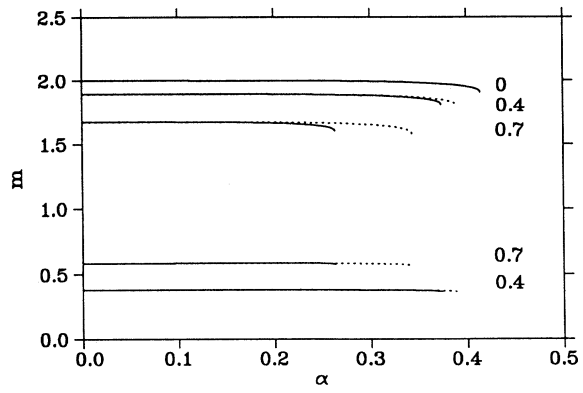

FIG. 4. The overlap m_2 (upper five curves) and m_1 (lower four curves) at $T=0$ as a function of α for the $Q=3$ $[B_2]$ model with the rule (10) for different values of the bias amplitude a. The dotted (solid) lines represent the model with (without) the ferromagnetic term and appropriate coefficients (11). A ferromagnetic term is seen to enhance the retrieval quality.

$$\mathcal{H}_\sigma(\mathbf{z},\boldsymbol{\xi})=\epsilon_2\sum_{k,k'=1}^{Q}\sqrt{\alpha r P(k)P(k,k')}(u_{k',\sigma}-B_{k,\sigma})z_{kk'}+\frac{1}{2}\alpha\beta\epsilon_2(\tilde{r}-r)\sum_{k,k'=1}^{Q}P(k)P(k,k')(u_{k',\sigma}-B_{k,\sigma})^2$$
$$+\epsilon_2\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{s_2}(u_{\xi^{\mu_1\mu_2},\sigma}-B_{\xi^{\mu_1},\sigma})m_{\mu_1\mu_2}+\gamma\sum_{k=1}^{Q}P(k)B_{k,\sigma}M+\epsilon_1\sum_{\mu_1=1}^{p_1}\left[B_{\xi^{\mu_1},\sigma}-\sum_{k=1}^{Q}P(k)B_{k,\sigma}\right]m_{\mu_1}\,. \quad (36)$$

In this case, the following order parameters appear:

$$m_{\mu_1\mu_2}=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle u_{\xi_i^{\mu_1\mu_2},\sigma_i}-B_{\xi_i^{\mu_1},\sigma_i}\rangle\rangle\rangle \quad (37)$$

$$m_{\mu_1}=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle B_{\xi_i^{\mu_1},\sigma_i}-\sum_{k=1}^{Q}P(k)B_{k,\sigma_i}\rangle\rangle\rangle \quad (38)$$

$$M=\sum_{k,k'=1}^{Q}P(k)P(k,k')M_{k'}$$

with

$$M_{k'}=\frac{1}{N}\sum_{i=1}^{N}\langle\langle\langle u_{k',\sigma_i}\rangle\rangle\rangle \quad (39)$$

together with q, $\tilde{q}$, r, and $\tilde{r}$ given by Eqs. (25), (26), (27), and (28), respectively.

As in the model studied in Sec. IV, the overlap with a pattern $\xi^{\mu_1\mu_2}$ of the second generation is again represented by $m_{\mu_1\mu_2}$. The order parameter m_{μ_1}, which is a component of the p_1-dimensional vector $\mathbf{m}^{(1)}$, measures the overlap of the network configuration with the pattern ξ^{μ_1} of the first generation. Indeed, due to the condition (5) on the $B_{k,k'}$ the expression (38) is maximized when $\sigma=\xi^{\mu_1}$ indicating perfect retrieval. The order parameter $M_{k'}$ measures the number of neurons in a state k'. So M corresponds to a weighted sum of the order parameters $M_{k'}$.

The meaning of the other order parameters is as before. We remark that for the specific bias choice $B_{k,k'}=a_2u_{k,k'}$, the expression (38) reduces to (23) but (39) does not reduce to (24).

B. Discussion of the retrieval properties

In this section, we study the retrieval properties of Potts neural networks that have memorized hierarchically correlated patterns by using the truncated pseudoinverse learning rule (13).

We start by considering the case of a finite number of patterns in both the first and the second generation. Then the free energy reduces to the $\alpha\to 0$ limit of (34), which only depends on $m_{\mu_1\mu_2}$, m_{μ_1}, and M. It turns out that the relevant fixed-point equations for the retrieval solutions are exactly equivalent to the fixed-point equation of the retrieval solutions of the Potts model with unbiased patterns [17], independent of the choice of the matrices $[B]$. Hence, in contrast with the networks studied in the former section, T_c is independent of a. Hence it equals the value of the unbiased Potts model, e.g., $T_c=1.000$ $(Q=2)$, $T_c=2.185$ $(Q=3)$,. . . .

Next we turn to the case where each class of the second generation contains an extensive number of patterns, while the number of ancestor patterns is still finite. Again for $Q=3$ models at $T=0$ with $a\equiv a_1=a_2$, we discuss the storage capacity as a function of a and the retrieval quality as a function of the loading $\alpha=p_1p_2/N$ of the second generation. To obtain these results, the relevant fixed-point equations for the retrieval solutions are solved numerically.

First, we compare in Fig. 5 the storage capacity of the truncated pseudoinverse network and the network with the ferromagnetic term for the two representative bias matrices indicated before. The dashed-dotted line represents the truncated pseudoinverse network, the dashed lines are taken from Fig. 2. They represent the network with the ferromagnetic term. The curve for $[B_1]$ type patterns coincides with that of the truncated pseudoinverse model. Hence, for this type of patterns there is no difference in storage capacity for both networks. This is in agreement with the signal-to-noise analysis since, for $[B_1]$ type patterns, both models lead to the same variance of the noise term. For $[B_2]$ type patterns, however, the truncated pseudoinverse model leads to a higher storage capacity than the model with the ferromagnetic term.

Second, the retrieval quality of the two models is compared. For $[B_1]$ type patterns, both models lead to identical overlap-loading diagrams. Hence there is no difference in retrieval quality. For $[B_2]$ type patterns, however, differences are found as is clear from Fig. 6. This figure shows the overlaps m_2 and m_1 as a function

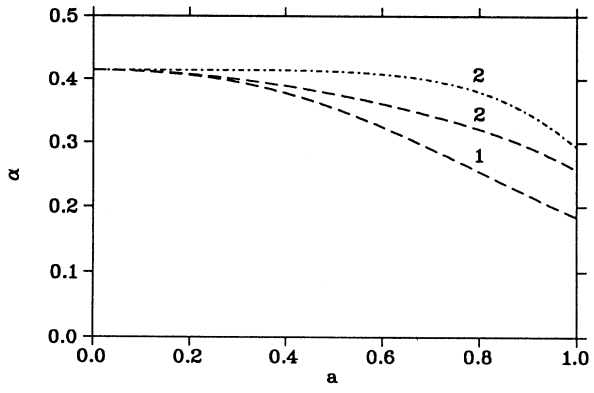

FIG. 5. The storage capacity at $T=0$ as a function of the bias amplitude a for $Q=3$ networks with $[B_1]$ (1) and $[B_2]$ (2) type patterns [see Eq. (32)]. The dashed (dashed-dotted) lines correspond with the model with the ferromagnetic term (11) [truncated pseudoinverse network (13)]. The truncated pseudoinverse network has the largest storage capacity.

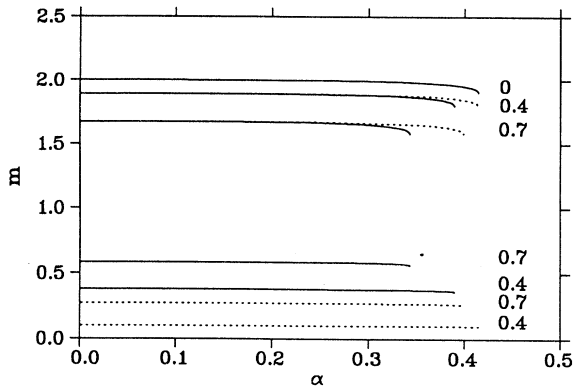

FIG. 6. The overlap m_2 (upper five curves) and m_1 (lower four curves) at $T=0$ as a function of α for the $Q=3$ networks with $[B_2]$ type patterns for different values of the bias amplitude a. The solid (dotted) lines represent the model with the ferromagnetic term (11) [truncated pseudoinverse model (13)]. The truncated pseudoinverse model leads to the best retrieval quality.

of the loading α of patterns of the second generation for different values of a. From this figure it follows that for each value of the loading α, the truncated pseudoinverse model has a higher overlap than the model with the ferromagnetic term. Consequently, for $[B_2]$ type patterns, the truncated pseudoinverse model leads to a better retrieval quality.

In this section, we have studied the ability of Potts neural networks to retrieve hierarchically correlated patterns, which are memorized with the learning rule (13). We have seen that for $Q=3$ models the truncated pseudoinverse model performs better than the model with the ferromagnetic term for certain bias types.

VI. CONCLUDING REMARKS

We have discussed Q-state Potts neural networks that are able to memorize hierarchically correlated patterns, generated by a Markovian scheme. Two different learning rules have been considered. The Hebbian-type learning rule (10) contains three relevant cases: only storing patterns of the second generation, storing both patterns of the first and second generation with and without the ferromagnetic term. The learning rule (13) is derived from the pseudoinverse rule by keeping the first terms in a series expansion. The free energy has been written down for both types of networks with general Q and arbitrary T in replica-symmetric mean-field theory.

For $Q=3$ models at $T=0$, where each class of the second generation consists of extensively many patterns, while the first generation still contains a finite number of patterns, we have calculated the storage capacity and the retrieval behavior (overlap).

First, for the models without the ferromagnetic term (Fig. 2), the storage capacity is enhanced by storing also the patterns of the first generation, and even further enhanced by introducing a ferromagnetic term [recall (10) and (11)]. However, for $Q>2$ models, the value of the unbiased case is never reached. The retrieval quality is the best for the model with the ferromagnetic term (Figs. 3 and 4).

Second, there is no difference in the retrieval behavior of the network with the ferromagnetic term and the truncated pseudoinverse model when $[B_1]$ type patterns are stored. For $[B_2]$ type patterns, however, the pseudoinverse model leads to the best retrieval properties.

ACKNOWLEDGMENTS

This work has been supported in part by the Research Fund of the K. U. Leuven (Grant No. OT/91/13). We are indebted to A. Patrick, G. M. Shim, and K. Y. M. Wong for stimulating discussions. We also would like to thank the Belgian National Fund for Scientific Research and the Inter-University Institute for Nuclear Sciences for financial support.

APPENDIX

To derive the truncated pseudoinverse learning rule (13) we have followed the idea of Cortes, Krogh, and Hertz [14]. We start from the pseudoinverse learning rule [15,16]

$$J_{ij}^{kl}=\frac{1}{Q^2N}\sum_{\mu_1,\nu_1=1}^{p_1}\sum_{\mu_2,\nu_2=1}^{p_2} u_{\xi_i^{\mu_1\mu_2},k}([C]^{-1})_{\mu_1\mu_2,\nu_1\nu_2}u_{\xi_j^{\nu_1\nu_2},l} \tag{A1}$$

with

$$C_{\mu_1\mu_2,\nu_1\nu_2}=\frac{1}{N(Q-1)}\sum_{i=1}^{N} u_{\xi_i^{\mu_1\mu_2},\xi_i^{\nu_1\nu_2}} \tag{A2}$$

where $1\le\mu_1,\nu_1\le p_1$, and $1\le\mu_2,\nu_2\le p_2$.

First, we note that after dropping all terms of order $1/\sqrt{N}$, the correlation matrix $[C]$ can be written as

$$[C]=(1-K_2)\mathbb{1}+(K_2-K_1)\mathbb{1}_{p_2}+K_1\mathbb{1}\ , \tag{A3}$$

where $(\mathbb{1})_{\mu_1\mu_2,\nu_1\nu_2}=\delta_{\mu_1\nu_1}\delta_{\mu_2\nu_2}$, $(\mathbb{1}_{p_2})_{\mu_1\mu_2,\nu_1\nu_2}=\delta_{\mu_1\nu_1}$, and $(\mathbb{1})_{\mu_1\mu_2,\nu_1\nu_2}=1$ and K_1, K_2 are given by (14), (12), respectively. Using the specific form of $[C]$ its inverse is easily found

$$([C])^{-1}=R\mathbb{1}-S\mathbb{1}_{p_2}-T\mathbb{1}\ , \tag{A4}$$

where

$$R=\frac{1}{1-K_2}\ , \tag{A5}$$

$$S=\frac{K_2-K_1}{(1-K_2)[1-K_2+p_2(K_2-K_1)]}\ , \tag{A6}$$

$$T=\frac{K_1}{[1-K_2+p_2(K_2-K_1)+p_1p_2K_1][1-K_2+p_2(K_2-K_1)]} . \tag{A7}$$

Second, inserting (A4) into (A1) results in

$$J_{ij}^{kl}=\frac{1}{NQ^2}\left\{R\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}u_{\xi_i^{\mu_1\mu_2},ku}\xi_j^{\mu_1\mu_2},l-S\sum_{\mu_1=1}^{p_1}\sum_{\mu_2,\nu_2=1}^{p_2}u_{\xi_i^{\mu_1\mu_2},k}u_{\xi_j^{\mu_1\nu_2},l}-T\sum_{\mu_1,\nu_1=1}^{p_1}\sum_{\mu_2,\nu_2=1}^{p_2}u_{\xi_i^{\mu_1\mu_2},k}u_{\xi_j^{\nu_1\nu_2},l}\right\} . \tag{A8}$$

Third, applying the law of large numbers

$$\frac{1}{p_2}\sum_{\mu_2=1}^{p_2}u_{\xi_i^{\mu_1\mu_2},k}=B_{\xi_i^{\mu_1},k} , \tag{A9}$$

$$\frac{1}{p_1p_2}\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}u_{\xi_i^{\mu_1\mu_2},k}=\sum_{\rho=1}^{Q}P(\rho)B_{\rho,k} \tag{A10}$$

expanding S and T in powers of $1/p_2$ and truncating the expansion at terms of order $1/p_2^2$

$$S\simeq\frac{1}{p_2(1-K_2)}+\frac{1}{p_2^2(K_2-K_1)} , \tag{A11}$$

$$T\simeq\frac{1}{p_1p_2^2(K_1-K_2)}-\frac{1}{p_1^2p_2^2K_1} , \tag{A12}$$

results in the following expression for the synaptic couplings:

$$J_{ij}^{kl}=\frac{1}{NQ^2}\left\{\sum_{\mu_1=1}^{p_1}\sum_{\mu_2=1}^{p_2}\left[\frac{1}{1-K_2}u_{\xi_i^{\mu_1\mu_2},k}u_{\xi_j^{\mu_1\mu_2},l}-\left[\frac{1}{1-K_2}+\frac{1}{p_2(K_2-K_1)}\right]B_{\xi_i^{\mu_1},k}B_{\xi_j^{\mu_1},l}\right]\right.$$
$$\left.-\left[\frac{p_1}{K_1-K_2}-\frac{1}{K_1}\right]\sum_{\rho=1}^{Q}P(\rho)B_{\rho,k}\sum_{\rho=1}^{Q}P(\rho)B_{\rho,l}\right\} . \tag{A13}$$

Finally, after some algebra we end up with the learning rule (13).

[1] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **35**, 2293 (1987).
[2] D. Bollé, P. Dupont, and J. van Mourik, J. Phys. A **24**, 1065 (1991).
[3] D. Bollé, R. Cools, P. Dupont, and J. Huyghebaert, J. Phys. A **26**, 549 (1993).
[4] J. P. Sutton, J. S. Beis, and L. E. H. Trainor, J. Phys. A **21**, 4443 (1988).
[5] V. Dotsenko and B. Tirozzi, Int. J. Mod. Phys. B **3**, 1561 (1989).
[6] C. R. Willcox, J. Phys. A **22**, 4707 (1989).
[7] N. Parga and M. A. Virasoro, J. Phys. **47**, 1857 (1986).
[8] S. Bös, R. Kühn, and J. L. van Hemmen, Z. Phys. B **71**, 261 (1988).
[9] M. V. Feigel'man and L. B. Ioffe, Int. J. Mod. Phys. B **1**, 51 (1987).
[10] A. Krogh and J. A. Hertz, J. Phys. A **21**, 2211 (1988).
[11] H. Gutfreund, Phys. Rev. A **37**, 570 (1988).
[12] I. Kanter, Phys. Rev. A **37**, 2739 (1988).
[13] J. Buhmann, R. Divko, and K. Schulten, Phys. Rev. A **39**, 2689 (1989).
[14] C. Cortes, A. Krogh, and J. A. Hertz, J. Phys. A **20**, 4449 (1987).
[15] L. Personnaz, I. Guyon, and G. Dreyfus, J. Phys. Lett. **46**, L359 (1985).
[16] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).
[17] D. Bollé and F. Mallezie, J. Phys. A **22**, 4409 (1989).