Optimally adapted multistate neural networks trained with noise

R. Erichsen, Jr.* and W. K. Theumann†
*Instituto de Fı́sica, Universidade Federal do Rio Grande do Sul, Caixa Postal 15051, 91501-970 Porto Alegre,
Rio Grande do Sul, Brazil*

(Received 29 April 1998; revised manuscript received 21 July 1998)

The principle of adaptation in a noisy retrieval environment is extended here to a diluted attractor neural network of Q-state neurons trained with noisy data. The network is adapted to an appropriate noisy training overlap and training activity, which are determined self-consistently by the optimized retrieval attractor overlap and activity. The optimized storage capacity and the corresponding *retriever* overlap are considerably enhanced by an adequate threshold in the states. Explicit results for improved optimal performance and new retriever phase diagrams are obtained for $Q=3$ and $Q=4$, with coexisting phases over a wide range of thresholds. Most of the interesting results are stable to replica-symmetry-breaking fluctuations.
[S1063-651X(99)04201-4]

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

I. INTRODUCTION

Since the pioneering work of Hopfield [1], there has been much interest in both the training and performance of attractor neural networks. Training consists in encoding an appropriate synaptic matrix that enables the network to store a macroscopic number of patterns, while the performance of a network refers to the ability to retrieve one or a specific set of stored patterns [2]. Training and performance are usually thought of as separate stages in the operation of a network.

The retrieval performance of an attractor network can be studied in two different scenarios [3]. One is characterized by a *fixed* synaptic prescription, as in the case of the Hopfield model [1] or the maximally stable network (MSN) [4–6], while in the other one, the entire space of synaptic interactions is searched for optimal performance whenever there is a change in the retrieval environment. The synapsis in the first scenario are determined in an ordinary learning stage and the performance of the network is optimized separately in a given training environment. In the second scenario one resorts to a continuous adaptive training process in which the network performance is optimized in an adiabatically evolving retrieving environment [3]. For each value of the noise parameter T (temperature of the retrieval dynamics), and storage ratio α, the network has a unique interaction configuration, the so-called *retriever*. This is in distinction to the retrieval performance that yields the phase diagrams for the Hopfield model or the MSN, in which the interaction configuration determined in the separate learning stage is the same for all T and α [7].

Adaptive training processes seem to be biologically appealing as a means to learn from the environment. The adaptive process in the second scenario requires training the network with noisy patterns [8,9] and it is a procedure that does not separate the training process as a distinct step from the operating stage of the network. The principle of adaptation in a network of binary units consists of the search of the interaction space for the optimized network performance adjusting the training noise to be the same as the retrieval noise in each step of the adiabatically evolving retrieving environment. Both noises refer to the Hamming distances between the actual states of the network and the encoded patterns.

Training noises have been introduced in feedforward networks [10,11] in order to avoid overfitting to training examples and in attractor networks with the purpose of enlarging their basins of attraction [8,9]. A slightly distorted set of random patterns is presented to the network in the process of encoding the synaptic matrix by means of a stepwise updating procedure following the perceptron learning rule [9]. The MSN is generated by an infinitesimal amount of training noise and, except for low retrieval noise T and low load α, the performance of the optimally adapted network is clearly superior to that of the MSN [3]. In particular, for low to moderate T and higher load α, a second optimal solution in interaction space appears for each value of the training noise in the optimally adapted network. This solution is a weaker retriever that can be interpreted as an attractor of self-adaptation.

The point is that the second retriever constitutes a further solution to the optimization process, with its own interaction, in a neighborhood of interaction space where there is no solution for the MSN. This second, optimal solution appears as a low performance solution in the absence of or for low to moderate retrieval noise, with improving performance, up to a certain point, as the retrieval noise is raised. Thus, there has to be already a certain level of retrieval noise for the weak retriever to have an interesting performance. Moreover, whenever the solution exists it is only within a narrow range of α.

The principle of adaptation has been worked out, so far, only for a network of binary neurons and the purpose of the present paper is to explore the merits of an extension of the principle to a multistate attractor network in which both the neurons and the noisy training versions of the encoded patterns can be in $Q(>2)$ states. This adds two new dimensions to the study of the performance of the network. First, the randomly distributed noisy patterns presented to the network in the training process introduce a training activity a_t. Sec-

*Electronic address: rubem@if.ufrgs.br

†Electronic address: theumann@if.ufrgs.br

ond, the firing rate of the neurons is determined by one or more thresholds, or a growth parameter in the dynamical output function. Thus, in the extension considered in this work, an evolving dynamical overlap $m(\tau)$ and a dynamical activity $a(\tau)$ are generated at each time step τ of the neuron updating procedure. The search for the optimized network performance by means of the extended adaptation principle consists now of the adjustment of the training overlap m_t and the training activity a_t to be $m_t = m(\tau)$ and $a_t = a(\tau)$, respectively, i.e., the same as the retrieval overlap and dynamical activity in each step of the adiabatically evolving retrieval environment. Adaptive performance in this wider sense is a self-consistent procedure in which the retrieval environment continuously optimizes the attractor performance of the network.

Networks of multistate neurons have interesting features and applications. Feedforward networks of such units can be used to study multiclass classification problems [12], while multistate attractor networks, which are useful for the recognition of various gray-toned patterns, are networks that have interesting inferential properties, by means of which the storage capacity and the retrieval ability can be enhanced when they are trained with patterns of low activity [13–16]. Also, the categorization ability can be improved in a multistate network with hierarchical patterns. There has been lately considerable interest in such networks [17–19].

We consider an extremely diluted network and, for simplicity, restrict ourselves to binary unbiased encoded patterns. The main emphasis of the paper is on the storage capacity, the quality of the performance of the strong and the second retrievers, and on the characterization of the various phases that can appear. With that purpose we produce explicit results for a network with $Q=3$ or $Q=4$ states. It will be shown that, within a finite range of a threshold parameter, there is considerable improvement of the storage capacity and in the high performance of the second retriever solution, in the absence of or for low retrieval noise, when compared with the optimally adapted network of binary neurons. In particular, we show that the second retriever may attain a fairly high retrieval overlap for small training noise in a regime where there is no solution for the optimally adapted network of binary neurons. These are important results in the search for improvement of the behavior of attractor neural networks. We restrict ourselves to finite-Q state networks, in place of addressing the general (large-Q) case.

The outline of the paper is the following. In Sec. II we extend the training with noise procedure in the space of synaptic interactions to a Q-state Ising network by means of a quenched optimization approach [3,20], within the replica-symmetry ansatz, introducing a smooth cost function given by an average squared Hamming distance. The equations for the adaptation process in a noisy retrieval environment are formulated in that section. The explicit results for the fixed-point behavior, the storage capacity and the corresponding phase diagrams for self-adaptation for the three and the four-state models are discussed in Sec. III, and compared with the MSN. The domain of validity of the replica symmetric results is determined by the de Almeida–Thouless lines [21] in terms of the retrieval noise and the threshold in the dynamical updating procedure. A summary and concluding remarks are presented in Sec. IV.

II. TRAINING WITH NOISE AND ADAPTATION

Consider a network of N nodes with a dynamical variable $S_i(\tau)$, at time step τ on node i, that indicates the extent to which the unit on node i fires. Each unit can be in any one of Q Ising states

$$\sigma_k = -1 + \frac{2(k-1)}{Q-1} \tag{1}$$

in the interval $[-1,+1]$, for $k=1,\dots,Q$. A macroscopic set of p binary patterns $\{\xi_i^\mu = \pm 1;\ \mu=1,\dots,p;\ i=1,\dots,N\}$, with $p=\alpha C$, is encoded in the network in the learning process, where C is the connectivity of a node. The patterns constitute a set of independent identically distributed random variables. Training consists in presenting to the network a noisy version $\{R_i^\mu(\tau)\}$ of the patterns, at time τ, and in the optimization of the network output after one time step. This involves a dynamical process in the space of state configurations of the network and, to keep the dynamics simple, we restrict ourselves to an extremely diluted network. Each $R_i^\mu(\tau)$ is assumed to be in one of Q states, σ_k, and can be thought of as an example of the pattern ξ_i^μ. Assuming that every noisy pattern has the same overlap m_t with the corresponding pattern ξ_i^μ, and that the activity a_t is the same for all patterns in the training set, we define

$$m_t = \frac{1}{N} \sum_i \xi_i^\mu \langle R_i^\mu(\tau) \rangle_R \tag{2}$$

and

$$a_t = \frac{1}{N} \sum_i \langle [R_i^\mu(\tau)]^2 \rangle_R , \tag{3}$$

where the brackets $\langle \dots \rangle_R$ denote averages over the probability distribution of R_i^μ. Thus, the noisy training inputs are constrained to satisfy the mean $\langle R_i^\mu(\tau)\rangle_R = m_t \xi_i^\mu$ and variance $\langle (R_i^\mu(\tau))^2\rangle_R - \langle R_i^\mu(\tau)\rangle_R^2 = a_t - m_t^2$.

The normalized local field at node i, due to the activity at the other nodes, is given by

$$h_i(\tau) = \frac{1}{\sqrt{C}} \sum_{j=i_1}^{i_c} J_{ij} S_j(\tau), \tag{4}$$

where J_{ij} is the synaptic connection between nodes i and j, independently in what state the dynamical variable S_j is, while $i_1,\dots,i_c$ denote the nodes feeding node i. The connections follow the spherical constraint $\Sigma_j J_{ij}^2 = C$, and we consider the extremely diluted network in the limit of large connectivity in which $1 \ll C \ll \ln N$. The one time-step dynamics is exact in this limit.

We deal in this paper with the asymptotic, equilibrium configuration $\{J_{ij}\}$, for the synaptic matrix elements of the learning process that follows from a Langevin dynamics with a noise term. This involves an annealing temperature T_a that takes care that the network does not get trapped in local minima of the free energy. The distribution of equilibrium states of the J_{ij} can then be described by a canonical ensemble with temperature T_a. Thus, there are two time scales in this approach: a short-time scale for the dynamical evolu-

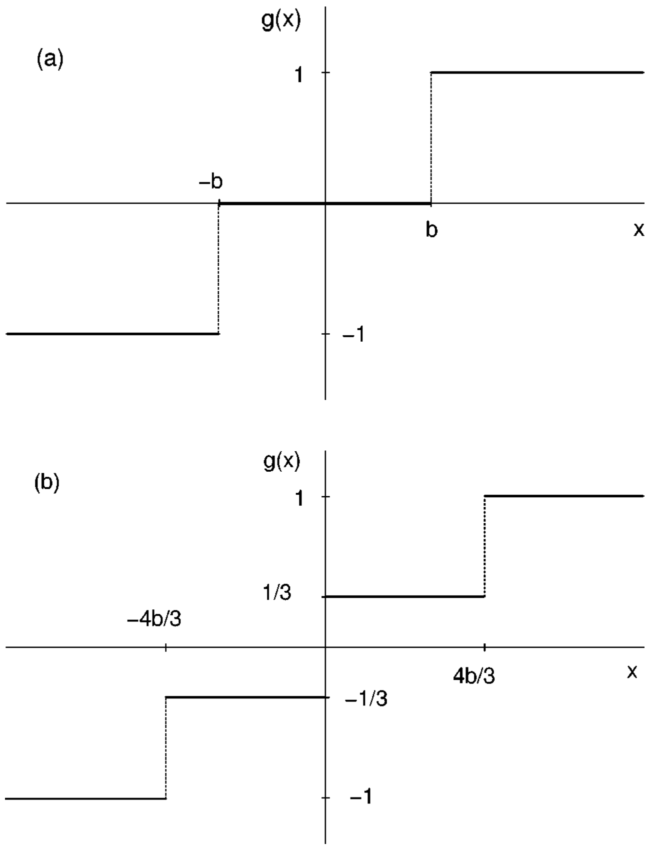

FIG. 1. The nondecreasing step function $g(x)$ for $Q=3$ (a) and $Q=4$ (b).

tion of the synaptic matrix $\{J_{ij}\}$ and a long-time scale for the dynamical evolution of the training and of the retrieval parameters.

The dynamical variables are updated according to the rule

$$S_i(\tau+1)=g(h_i(\tau)), \tag{5}$$

where $g(h_i(\tau))$ is the nondecreasing step function

$$g(x)=\sum_{k=1}^{Q}\sigma_k[\theta(b(\sigma_{k+1}+\sigma_k)-x)-\theta(b(\sigma_k+\sigma_{k-1})-x)] \tag{6}$$

shown in Fig. 1 for $Q=3$ and $Q=4$, in which $\theta(x)$ is the unitary step function, $\sigma_0=-\infty$, $\sigma_{Q+1}=\infty$, $b\geqslant 0$ is the threshold parameter and σ_k are the uniformly spaced Ising states of Eq. (1). According to Eq. (6), there is a zero activity state whenever Q is odd and none if Q is even.

For the adapted optimization a temperature T is introduced as a noise parameter, not to be confused with the annealing temperature T_a, to characterize the noisy retrieval environment. We assume a Gaussian thermal noise term added to the local field to write the one-step output as

$$S_i(\tau+1)=g(h_i(\tau)+Tz), \tag{7}$$

where z has mean zero and unit variance. The optimization, in the extremely diluted limit, consists in penalizing deviations from the minimal output error in one time step on any node which is independent of the optimization on all the other nodes. Thus, it is sufficient to consider the cost function for a single node. We choose this to be

$$\sum_{\mu} d_i^{\mu}(\tau+1)=\sum_{\mu}\langle\langle[1-2\xi_i^{\mu}S_i(\tau+1)+S_i^2(\tau+1)]\rangle_z\rangle_R, \tag{8}$$

where $d_i^{\mu}(\tau+1)$ is the average squared Hamming distance to a stored pattern in which $\langle\cdots\rangle_z$ denotes the average over the Gaussian thermal noise. The training noise enters only through the local fields, via Eqs. (4) and (7). In the case of binary patterns, the local field is a Gaussian random variable with mean $\langle h_i(\tau)\rangle_R=m_t\Lambda_i^{\mu}$ and variance $\langle h_i^2(\tau)\rangle_R-\langle h_i(\tau)\rangle_R^2=a_t-m_t^2$, in which

$$\Lambda_i^{\mu}=\frac{1}{\sqrt{C}}\sum_{j=i_1}^{i_c}J_{ij}\xi_j^{\mu} \tag{9}$$

is the local field on node i due to the pattern μ.

The optimization of the Hamming distance between the one-step output of the network in the noisy *training* environment and a given pattern in a network of binary neurons is equivalent to finding the optimal output overlap after one time step. In the case of a network of multistate neurons, the Hamming distance also depends on the activity through the local field, and our first goal is to find the optimal output Hamming distance $d(m_t,a_t)$, after one time step, for a given training overlap *and* a given training activity.

For that purpose, and for later use, we need the averages

$$S_{m_t,a_t}(\Lambda_i^{\mu})\equiv\langle\langle S_i(\tau+1)\rangle_z\rangle_R=\frac{1}{2}\sum_{k=1}^{Q}\sigma_k\mathrm{Erf}(u_k,l_k;\Lambda_i^{\mu}) \tag{10}$$

and

$$S^2_{m_t,a_t}(\Lambda_i^{\mu})\equiv\langle\langle S_i^2(\tau+1)\rangle_z\rangle_R=\frac{1}{2}\sum_{k=1}^{Q}\sigma_k^2\mathrm{Erf}(u_k,l_k;\Lambda_i^{\mu}), \tag{11}$$

which follow from Eq. (7), where

$$\mathrm{Erf}(u,l,\Lambda)=\mathrm{erf}\left(\frac{u-m_t\Lambda}{\sqrt{2(a_t-m_t^2+T^2)}}\right)-\mathrm{erf}\left(\frac{l-m_t\Lambda}{\sqrt{2(a_t-m_t^2+T^2)}}\right), \tag{12}$$

with

$$u_k/2b=\sigma_k+1/(Q-1),\quad u_Q=\infty \tag{13}$$

and

$$l_k/2b=\sigma_k-1/(Q-1),\quad l_1=-\infty. \tag{14}$$

The quenched optimization approach [3,20] requires the introduction of the partition function

$$\mathcal{Z}(\beta)=\int\prod_j dJ_{ij}\,\delta\left(\sum_j J_{ij}^2-C\right)\exp\left[-\beta\sum_\mu d(\Lambda_i^\mu)\right] \tag{15}$$

to obtain first an annealed average over the space of synaptic connections J_{ij}, in which $\beta=T_a^{-1}$ is the inverse annealing temperature, and $d(\Lambda_i^\mu)$ is the squared Hamming distance, for a given configuration $\{\xi_i^\mu\}$ of encoded patterns, averaged over thermal and training noises. Its dependence on the noise parameters m_t, a_t, and T is left implicit. The quenched average free energy is then obtained making use of the replica method to write

$$\langle\ln\mathcal{Z}\rangle_\xi=\lim_{n\to 0}\frac{1}{n}(\langle\mathcal{Z}^n\rangle_\xi-1), \tag{16}$$

where $\langle\cdots\rangle_\xi$ denotes the average over the set of stored patterns $\{\xi_i^\mu\}$. Using the standard technique in the space of synaptic interactions, with the assumption of replica symmetry [22,23], we obtain the optimal one-step output Hamming distance for training

$$\begin{aligned} d(m_t,a_t)&=-\lim_{\beta\to\infty}\frac{1}{\alpha\beta C}\langle\ln\mathcal{Z}\rangle_\xi\\ &=\mathrm{extr}_x\left\{\int Dy\,\min_\lambda F(\lambda,x,y)-\frac{1}{2\alpha x}\right\} \end{aligned} \tag{17}$$

as a function of the overlap m_t and activity a_t of the noisy input patterns, in which $Dy=e^{-y^2/2}dy/\sqrt{2\pi}$ is a Gaussian measure and

$$F(\lambda,x,y)=d(\lambda)+\frac{(\lambda-y)^2}{2x}. \tag{18}$$

Here, $d(\lambda)$ is the squared Hamming distance averaged over ξ_i^μ, while $x=\beta(1-q)$ and

$$q=\frac{1}{C}\sum_j J_{ij}^\rho J_{ij}^\sigma \tag{19}$$

for all $\rho\neq\sigma$ is the spin-glass order parameter for the problem.

The optimization in the training process amounts to take the limits $\beta\to\infty$ and $q\to 1$ keeping x finite. A single solution in the space of interactions is thus obtained out of the full multiplicity of solutions when $q\to 1$ [23]. The minimization with respect to λ yields

$$y(\lambda)=\lambda+xd'(\lambda), \tag{20}$$

where $\lambda=\lambda(y)$ is the inverse function of $y(\lambda)$. On the other hand, the extremum in x gives the saddle-point equation

$$\alpha^{-1}=\int Dy[\lambda(y)-y]^2, \tag{21}$$

which determines the storage capacity α for a given training environment.

In cases where $\lambda(y)$ is a multivalued function of y, which is the case for $Q\geqslant 2$, there may be one or more transitions, each with a fixed y_0 between an upper and a lower value $\lambda_>$ and $\lambda_<$, respectively, ruled by a Maxwell construction

$$\int_{\lambda_<}^{\lambda_>}d\lambda\,y(\lambda)=y_0(\lambda_>-\lambda_<), \tag{22}$$

where $y_0=y(\lambda_<)=y(\lambda_>)$. It turns out that the function $F(\Lambda,x,y)$ is the same on both sides of the ‘‘first-order’’ transition.

The optimal output Hamming distance for training becomes then

$$d(m_t,a_t)=\int Dy\,d(\lambda(y)). \tag{23}$$

It is convenient to introduce the distribution of the local fields due to the encoded patterns, defined as [3–5]

$$\rho(\Lambda)=\left\langle\left\langle\delta\left(\Lambda-\frac{1}{\sqrt{C}}\sum_j J_{ij}\xi_j\right)\right\rangle_J\right\rangle_\xi, \tag{24}$$

where the ensemble average $\langle\cdots\rangle_J$ is performed with the partition function $\mathcal{Z}(\beta)$, Eq. (15). It turns out that this distribution becomes

$$\rho(\Lambda)=\int Dy\,\delta(\Lambda-\lambda(y)), \tag{25}$$

and the transition between the lower and upper bonds, $\lambda_<$ and $\lambda_>$, respectively, implies a gap in the distribution of local fields $\rho(\lambda)$ whenever $\lambda(y)$ is a multivalued function of y.

The optimal one-step output Hamming distance for training with noise may now be written as

$$d(\tau+1)=1-2f_{m_t,a_t}(m_t,a_t)+g_{m_t,a_t}(m_t,a_t), \tag{26}$$

where

$$f_{m_t,a_t}(m,a)=\int d\Lambda\,\rho_{m_t,a_t}(\Lambda)S_{m,a}(\Lambda) \tag{27}$$

is the optimized overlap between the encoded patterns and their noisy versions and

$$g_{m_t,a_t}(m,a)=\int d\Lambda\,\rho_{m_t,a_t}(\Lambda)S_{m,a}^2(\Lambda) \tag{28}$$

is their optimized activity. The distribution of the local fields, $\rho_{m_t,a_t}(\Lambda)$, is a characteristic property of the training set and, as such, it depends on m_t and a_t.

The formal results presented so far assume that replica symmetry holds in the space of interactions. The condition for local stability of the replica symmetric saddle point can be writen as [20,24]

$$\alpha^{-1}>\int Dy\,[\lambda'(y)-1]^2, \tag{29}$$

in which $\lambda'=d\lambda/dy$, and this is to be solved together with Eq. (21). When the distribution of the local fields has a gap, λ' diverges and the condition cannot be satisfied. Then, the network becomes unstable to replica-symmetry-breaking fluctuations. The limiting load for which Eq. (29) is still satisfied yields the de Almeida–Thouless (AT) line, $\alpha_{AT}(T)$ [21]. The dependence on the retrieval noise T comes from λ. Note that the AT line must lie within the one-band region or, at most, on the band-merging surface where the gap in $\rho(\lambda)$ disappears [24]. This completes the formal description of the training process in itself. In order to become optimally adapted, we consider now the retriever process.

The calculation of the one-step output Hamming distance between *any* input state $\{S_i(\tau)\}$ and a given encoded pattern in a noisy retrieval environment, with temperature T, is now obtained as follows. First, the training parameters m_t and a_t in Eqs. (10)–(12) are replaced by the overlap $m(\tau)$ and the dynamical activity $a(\tau)$ of the noisy *retrieval* state $\{S_i(\tau)\}$, expressed, respectively, as Eqs. (2) and (3) with $\{S_i(\tau)\}$ in place of the the noisy pattern $\{R_i^\mu(\tau)\}$. The one-step output Hamming distance in the *retrieval* environment is now given by an expression similar to Eq. (26), depending on the pair (m_t, a_t) through the distribution of local fields *and* on the pair (m,a) through the present state of the network as given, literally, by Eqs. (27) and (28).

Now, the training overlap m_t and the training activity a_t, which give the optimal performance for retrieval at a fixed temperature T, storage level α, and threshold parameter b, are given by the adaptation principle. The optimal adaptation consists of a search in the space of interactions $\{J_{ij}\}$ simultaneously with a search in the space of state configurations $\{S_i(\tau)\}$. The best adapted performance of the network is attained by adjusting the training noise and activity to the same level as the retrieval noise and activity. For the parallel dynamics in the extremely diluted network we are dealing with, the stable fixed point of the set of equations

$$f_{m,a}(m,a) = m \tag{30}$$

and

$$g_{m,a}(m,a) = a \tag{31}$$

gives at the same time the optimal training condition and the optimized performance. The stable fixed point for each value of the synaptic noise parameter T, the storage ratio α, *and* the threshold parameter b is a *retriever*, for which the network has a unique interaction configuration. In other words, in distinction to the usual phase diagrams for retrieval, every point of the phase diagrams that will be discussed next represents a different network.

III. RESULTS AND DISCUSSION

We present next the results for the optimally adapted retrievers. The rich structure of locally stable states and the corresponding phase diagrams for self-adaptation that arise as the threshold parameter b is increased will be discussed now, separately for $Q=3$ and $Q=4$.

A. Three-state network

To illustrate the role of the threshold parameter b, we discuss first the fixed-point solutions for m and a and the corresponding phase diagram for α versus b, in the absence of retrieval noise shown in Fig. 2. For fixed b within the range $0 \leqslant b \leqslant 0.57$ and $0 \leqslant \alpha \leqslant \alpha_1(b)$, there is a perfect retriever with $m=1=a$, which is the only stable fixed point, and a solution with $m=0$ and either $a \neq 0$ or $a=0$, which is an unstable fixed point. This suggests that one can conceive a network capable of perfect retrieval operating with a limited threshold, as long as the training is with infinitesimal noise $m_t = 1^-$ and almost full activity $a_t = 1^-$. The corresponding retriever is that of the MSN.

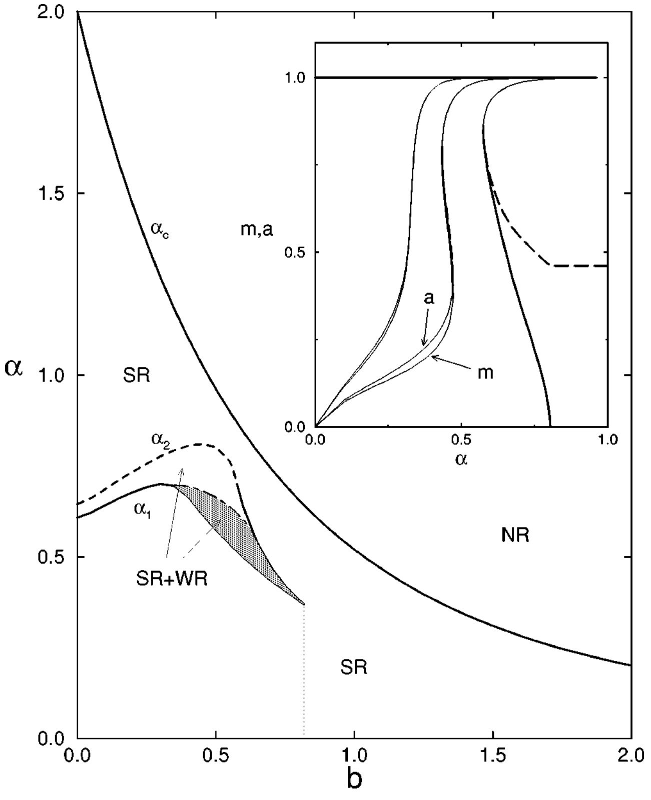

FIG. 2. Phase diagram for the load α as a function of the threshold b for $Q=3$, at $T=0$, and the corresponding optimal overlap m (solid lines) and activity a (dashed lines) for $b=0.5$ (right), $b=0.7$ (center), and $b=0.9$ (left), in the inset. Unstable fixed-point solutions are shown in light lines. SR and WR are strong and weak retrievers, respectively. The SR is a wide retriever at the left of the light dotted line and below $\alpha_1(b)$. Solid lines in the phase diagram indicate discontinuous transitions and a dashed line a continuous transition. The dash-dotted line is the de Almeida-Thouless line (cf. the text). The WR is unstable to replica-symmetry-breaking in the shaded area.

The line $\alpha_1(b)$ deserves further attention. It is the upper bound of the region where the perfect retriever is the only attractor in the retriever dynamics with a wide basin of attraction for self-adaptation. Beyond that line, the basin of attraction of this retriever is greatly reduced in the three-state network, as will be seen next. Thus, for increasing b, in the small b regime, there is an enhancement of the associativity of the network, as long as α_1 is an increasing function of b.

A new pair of stable and unstable fixed points appears discontinuously at $\alpha_1(b)$. The stable fixed point represents a new retriever of weaker attractor overlap and reduced activity. Note, however, that for low to moderate b (illustrated in the inset by $b=0.5$), there is a considerably enhanced retrieval overlap when compared with the overlap for the optimally adapted network of binary units [3]. The second retriever has a rather wide basin of attraction for this larger retriever overlap. This higher performance can be attained through training with low-noise patterns with moderately high activity. For the threshold $b\simeq 0.3$ that maximizes $\alpha_1(b)$, the improvement in storage capacity with the *same* retrieval overlap as that of the network of binary neurons is about 20%. However, as one would expect, the performance deteriorates with a further increase in the threshold b.

The second stable fixed point means that there exists a second training condition, with higher noise, which results in a network with lower, but still optimal performance when compared with other three-state networks in its vicinity of the space of interactions, for this training condition. The unstable fixed points are repelors of the self-adaptation dynamics [3].

The overlap of this second retriever vanishes continuously as α increases approaching $\alpha_2(b)$. For $\alpha_2(b)\leqslant\alpha\leqslant\alpha_c(b)$, the perfect retriever and a nonretriever with $m=0$, and either finite or no activity, are the only stable fixed-point solutions. The nonretriever state with $a\neq 0$ appears as a self-sustained activity phase, which has been discussed first for a diluted network with a Hebbian learning rule [14]. When the activity is zero the network stops operating.

The presence of a nonretriever with finite activity follows from the fixed-point solution for (m,a) when $m=0$ is a stable fixed point. The expression for $S^2_{m,a}(\Lambda)$ becomes then independent of the local field Λ and, hence, of x and α. The fixed-point values for a are then given by the solutions of the equation $a=1-\mathrm{erf}[b/\sqrt{2(a+T^2)}]$. The solution $a=0$ is stable for all b, when $T=0$. There is a second stable fixed point that decreases monotonically from $a=1$, at $b=0$, and disappears discontinuously at $b\simeq 0.57$ when the value $a\simeq 0.23$ is reached. This is the origin of the ''tricritical'' point in the phase diagram for α versus b, where the line of continuous transitions for the overlap becomes discontinuous. We come back to this point below. It is important to point out that the term ''transition'' here only means that the network changes from one retriever state to another one. We remind the reader that it is not meant as an usual thermodynamic phase transition, since each point of the phase diagram corresponds to a different network.

Finally, when α reaches the critical storage capacity $\alpha_c(b)$, given by

$$\alpha_c^{-1}(b)=\int_{-\infty}^{b}Dy(b-y)^2, \qquad (32)$$

the perfect retriever is destabilized.

Consider next the case where $0.57<b\leqslant 0.82$. For $0\leqslant\alpha<\alpha_1(b)$, there is again a perfect retriever that is a stable fixed-point solution. In addition, a pair of stable and unstable fixed points appears. The stable fixed point is a nonretriever with $m=0$ and either $a\neq 0$ or $a=0$. A new pair of stable and unstable fixed points appears discontinuously at $\alpha_1(b)$. The stable fixed point is, again, a retriever of weaker attractor overlap and reduced activity. However, as α approaches $\alpha_2(b)$, this second retriever vanishes *discontinuously* and, thus, there is a changeover from the line of continuous transitions $\alpha_2(b)$ when b increases and reaches a tricritical point at $b\simeq 0.57$. When α increases beyond $\alpha_2(b)$, the perfect retriever and the nonretriever are, again, stable fixed-point solutions, and the perfect retriever, which has a narrow basin of attraction, is destabilized when the critical $\alpha_c(b)$ is reached. When b is increased, the retriever of weaker attractor overlap disappears at $b\simeq 0.82$, and beyond this point the perfect retriever is the only stable fixed point with finite overlap for $0\leqslant\alpha\leqslant\alpha_c(b)$.

Now we discuss the stability of the replica symmetric solution. First, the strong retriever state is always stable to replica-symmetry-breaking fluctuations below α_c. Thus, at most the weak retriever can become unstable. In view of this, we mapped out the region of the phase diagram where the stability condition, Eq. (29), is not satisfied for the weak retriever state, and this is shown as the shaded area in Fig. 2, the dash-dotted line being the AT line. Furthermore, we found that this line corresponds to the appearance of a gap in the distribution of local fields.

The phase diagram also yields the optimal basin boundary of the self-adaptation dynamics for a given α. Thus, as α increases for $b\leqslant 0.82$ the strong retriever is a ''wide'' retriever for $\alpha<\alpha_1(b)$, since it is the only attractor in the self-adaptation dynamics. For $\alpha>\alpha_1(b)$, the strong retriever becomes a ''narrow'' retriever that coexists with the weak retriever. Finally, for $b>0.82$, the strong retriever is a narrow retriever that coexists with the nonretriever state for all $\alpha\leqslant\alpha_c(b)$.

We consider next the results in the presence of retrieval noise T. In the case of a small to moderate threshold where the strong and weak retriever coexist, say, for $b=0.5$, the phase diagram for T versus α is not very different from the phase diagram for the network of binary units. The strong and the weak retriever coexist now over a wider range of α but the strong retriever disappears, as one would expect, for a lower T. More interesting are the results for the phase diagram and the underlying fixed-point solutions for the overlap and the activity when $b=1$, shown in Fig. 3. This threshold is typical of an optimally adapted network that has a perfect retriever as the only stable fixed point with nonzero overlap at $T=0$. For fixed and low $T\leqslant 0.4$, there is a strong retriever with rapidly decreasing m and a when α comes close to the line $\alpha_c(T)$, where both parameters vanish discontinuously. There is a second stable fixed point with $m=0$ and $a\sim 0$, for all $\alpha\geqslant 0$, and this nonretriever is the only stable solution for $\alpha>\alpha_c(T)$. There is also an unstable fixed point for m and a throughout the range $0\leqslant\alpha\leqslant\alpha_c(T)$ that separates the basin of attraction for self-adaptation of the two stable fixed points, and indicates that the strong retriever is a narrow retriever in this interval.

An increase in retrieval noise can be of use for the enhancement of the performance of the single, strong retriever, with a moderately large threshold, as in the present case of $b=1$. Indeed, for $0.4\leqslant T\leqslant 0.8$, the nonretriever becomes an *un*stable fixed point for α below the line $\alpha_1(T)$, leaving the strong retriever as a wide retriever. The overlap and the activity change discontinuously as α goes through $\alpha_1(T)$. For $T\geqslant 0.8$, the overlap of the wide retriever vanishes continuously as α approaches $\alpha_c(T)$. The results shown here confirm the general expectation that one cannot attain the best retriever overlap (as we have here for the narrow retriever) together with the best associativity, as for the wide retriever, in the same network except at the phase boundary.

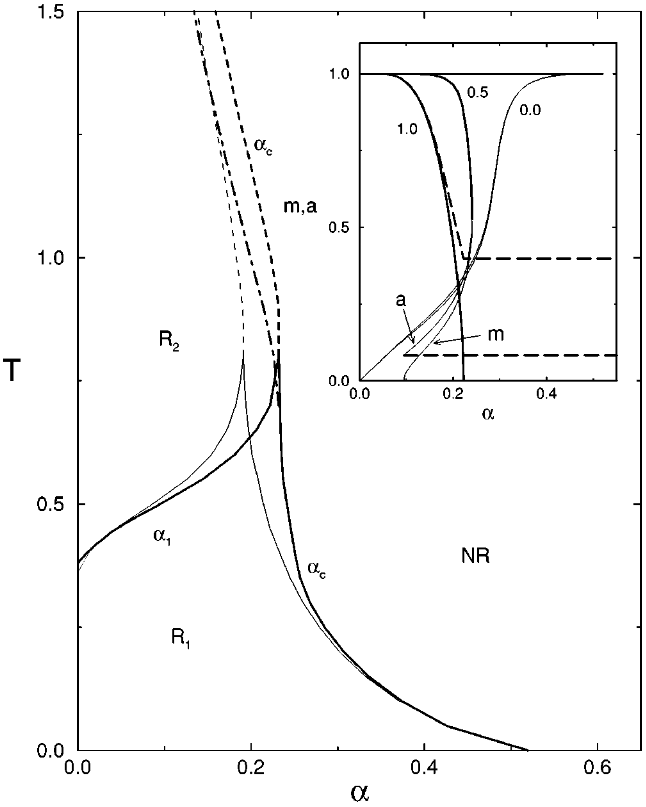

FIG. 3. Phase diagram, for T vs α, for $Q=3$ and $b=1$. In the inset are shown the optimal overlap (solid lines) and activity (dashed lines) for $T=0$, $T=0.5$, and $T=1$; the unstable solutions for m and a are in light lines. In $R_{1(2)}$ the strong retriever is a narrow (wide) retriever. NR is the nonretriever phase. The dash-dotted line is the de Almeida–Thouless line.

The AT line coinciding with the locus where the gap closes down is also shown in Fig. 3, and the region to the right of the line up to the α_c line is stable to replica-symmetry-breaking fluctuations. Thus, it seems that the part of the discontinuous transition line $\alpha_c(T)$ that is close to the tricritical point where the changeover to the line of continuous transitions takes place, is marginally stable. We also argue that for low T the line $\alpha_c(T)$ may be almost correct, since $\alpha_c(0)$ is the critical capacity of the MSN, which corresponds to a stable point. Note that the line $\alpha_c(T)$ of discontinuous transitions has an upper part of infinite slope which should also be correct since one would not expect a reentrant behavior for $\alpha_c(T)$. Finally, for comparison, we also show the phase boundaries for the MSN and conclude that the optimally adapted network with three-state neurons has an improved performance in the presence of retrieval noise.

B. Four-state network

To see now the effects of the threshold in the optimally adapted four-state network, we present first the results for the fixed-point solutions for the overlap and the activity in Fig. 4. Depending on the value of b there may be a domain in the values of α in which there are up to three stable fixed-point solutions with nonzero m, one for a perfect retriever and the other ones for weaker retrievers. The perfect retriever exists up to a critical $\alpha_c(b)$, given by

$$\alpha_c^{-1}(b)=\int_{-\infty}^{4b/3} Dy\left(\frac{4b}{3}-y\right)^2. \qquad (33)$$

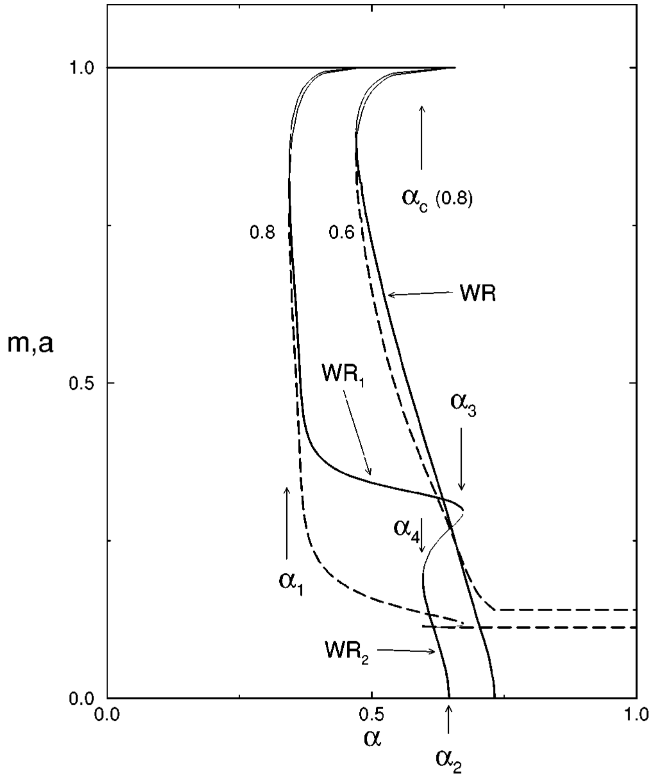

FIG. 4. Optimal overlap (solid lines) and activity (dashed lines) for $Q=4$, at $T=0$, for $b=0.6$, and $b=0.8$. The various α indicate the loads for which the optimal solutions appear or disappear, for $b=0.8$, and WR, WR_1, and WR_2 are weak retrievers.

It turns out that there is a load $\alpha_1(b)$ for all b, where a weak retriever, which may or may not be the only one, appears discontinuously as α attains that point. For $b\leqslant 0.65$, it is the only weak retriever, as can be seen in the phase diagram for α versus b shown in Fig. 5. Note that, also for the four-state network, α_1 increases with b in the small b regime with a considerable enhancement of the strong retriever as a wide retriever. The perfect and the weak retriever coexist with increasing α until either the weak retriever disappears continuously at $\alpha_2(b)$, which is the case for $b\leqslant 0.44$, or the strong retriever ends at $\alpha_c(b)$ for $0.44\leqslant b\leqslant 0.65$. In the latter case, the weak retriever of nonzero overlap remains as the only attractor of self-adaptation up to $\alpha_2(b)>\alpha_c(b)$.

On the other hand, for b well above 0.65, a second weak retriever (WR_2) appears discontinuously as α attains the line $\alpha_4(b)$ while the first weak retriever (WR_1) extends up to a quite higher load $\alpha_3(b)$, where the state of the network changes discontinuously to the non-retriever state. The overlap of the WR_2 vanishes continuously as α approaches $\alpha_2(b)$. The two weak retrievers coexist for $\alpha_4(b)\leqslant\alpha\leqslant\alpha_2(b)$. Note that both the line where the first weak retriever disappears and the domain of α where the second weak retriever exists may lie well above the critical capacity $\alpha_c(b)$ for the existence of the perfect retriever.

The situation can become more involved for intermediate values of b, shown by the inset in Fig. 5. Around the end-

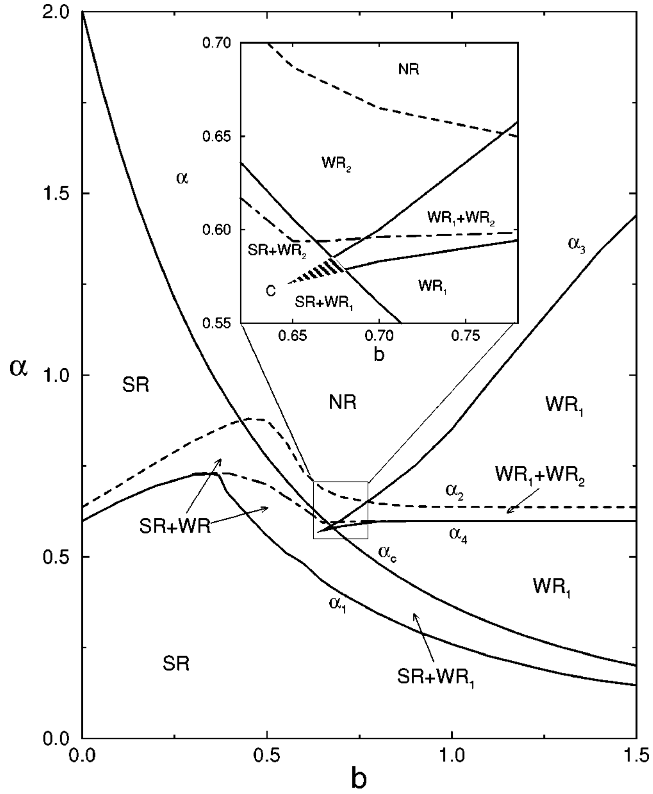

FIG. 5. Phase diagram for α as a function of b, for $Q=4$ at $T=0$, described in the text. The amplified central part is shown separately. The retrievers and the nature (continuous or discontinuous) of the phase boundaries are as in previous figures. The SR, WR_1, and WR_2 coexist in the shaded area of the inset. The de Almeida-Thouless line is the dash-dotted line.

point C of the wedge of discontinuous transitions lines, the second weak retriever can be reached continuously from the first one.

It is interesting to note that, for large b, the WR_1 state has an asymptotic overlap and activity $m\sim1/3$ and $a\sim1/9$, respectively. These correspond to the storage of binary patterns in a network with only the microscopic states $S_i=\pm1/3$ being activated. These are, practically, the only states favored in the high-b regime, since the states $S_i=\pm1$ can only become active by means of high local fields which are extremely unlikely in the absence of retrieval noise. Indeed, we found that the line $\alpha_3(b)$ goes to the critical value $\alpha_c=2$ for the optimal network of binary units with increasingly large b. Thus, as expected, the behavior of the network in the large-b limit should become that of the MSN with reduced overlap and activity.

The phase diagram in Fig. 5 also provides the optimal basin boundary of attraction, for a given α and b. For $b=1$, say, the strong retriever is a wide retriever for $\alpha<\alpha_1(b)$, and a narrow retriever when $\alpha_1(b)\leqslant\alpha\leqslant\alpha_c(b)$. On the other hand, in the interval $\alpha_c(b)\leqslant\alpha\leqslant\alpha_4(b)$, the weak attractor with higher overlap is a wide retriever, since it is the only attractor for the self-adapting dynamics in this interval. In distinction, in the interval $\alpha_4(b)\leqslant\alpha\leqslant\alpha_3(b)$ that weak retriever is a narrow retriever, that coexists with WR_2 if $\alpha\leqslant\alpha_2(b)$ and with the nonretriever state otherwise.

To discuss the validity of the replica symmetric results note that, whenever two weak retrievers coexist in the phase diagram, each one has to be analyzed separately since they refer to different levels of training noise, such that one may correspond to a gapless local field distribution and the other may not. The AT line is the dash-dotted line shown in Fig. 5, that starts on the boundary $\alpha_1(b)$ where the single weak retriever appears for small b and it merges with $\alpha_4(b)$ around $b=0.8$. That retriever is stable to replica-symmetry-breaking fluctuations above the AT line. The WR_2 is unstable around C and is stable in the strip $\alpha_4(b)\leqslant\alpha\leqslant\alpha_2(b)$, whereas the WR_1 is unstable everywhere below and at the boundary $\alpha_3(b)$. The left part of the boundary $\alpha_1(b)$ is marginally stable, as well as the boundary $\alpha_c(b)$ for the perfect retriever.

IV. SUMMARY AND CONCLUDING REMARKS

The principle of adaptation, formulated earlier for a network of binary neurons, has been extended in this work to study the training and performance of optimally adapted attractor neural networks of multistate neurons trained with noisy inputs in the presence of a noisy retrieval environment. Explicit results where obtained for the optimal attractor overlap and the optimal dynamical activity as functions of the retrieval noise T, the load α, and the threshold b, for a network with dilute connectivity. The maximum storage capacity was also obtained as a function of b and T and explicit retriever phase diagrams of performance and associativity of the retrievers are exhibited for a network of three- or four-state neurons. These are phase diagrams for *self-adaptation*, in distinction to phase diagrams for attraction, as pointed out in Ref. [3]. We remind the reader that, as pointed out by Wong and Sherrington, coexisting retrievers are solutions for different networks, which should correspond to distinct synaptic interactions.

An important issue of this work concerns the improvement in the associativity of multistate networks, when the width b of the intermediate states increases, in the small b regime. The enhanced performance of the second retrievers has also been emphasized. This is important because they are optimal *retriever* solutions on their own, rather than weaker *retrieval* solutions for the optimal network configuration, if such solutions exist [3]. We have shown that an improvement of the performance of the second retriever in the optimally adapted network with multistate units can be attained with relatively small training noise and large-activity input patterns. In practical terms, this may be a more accessible situation than training with an infinitesimal amount of noise and almost full activity. Furthermore, we have shown that the storage capacity of the second retriever is a nonmonotonic function of the threshold b with an increasing capacity for small b. With a moderately large threshold, as in the case of $b=1$ for the three-state network, an increase in retrieval noise T may help to enlarge the basin of attraction of the single, strong retriever. This can be understood noting that the increase in the noise should aid in overcoming the large gap in the local field in firing the units when the network has been trained with a moderate training noise. These are important results in the search for improvement of the behavior of attractor neural networks.

The work presented here is restricted, for simplicity, to binary encoded patterns. On the basis of results we obtained for three- or four-state patterns, we argue that this should not be a serious restriction. What is important is that the states of the noisy training set $\{R_i^{\mu}(\tau)\}$ have the same degrees of freedom as the arbitrary input set $\{S_i(\tau)\}$ for retrieval. This requires the introduction of a training activity a_t in the noisy inputs, in order to optimize both the training and the adaptation process in the Q-state network.

We have found, in accordance with earlier works, that networks are specialized [3,7]. Indeed, one cannot attain the best storage capacity for all T and b in a single network. Even if b is fixed the storage capacity of the strong retriever will be that of the MSN only at very low T and it will become that of the Hopfield model at high T.

All the results were obtained with the assumption of replica symmetry in the space of synaptic interactions and the limit of validity of this assumption has been established finding the de Almeida–Thouless lines $\alpha_{AT}(b)$ at $T=0$ and $\alpha_{AT}(T)$ for a given b. These lines coincide with the band-merging lines for the distribution of the local field. Due to the presence of optimal solutions for small-to-moderate training noise, there are gaps in the distribution of the local fields over sizeable domains of the phase diagram that are not stable to replica-symmetry-breaking fluctuations. Nevertheless, interesting phase boundaries and domains of the phase diagrams are stable or, at worst, marginally stable, confirming the validity of our results. Indeed, the enhancement of the line $\alpha_1(b)$, where the second retriever appears for small training noise and large activity, both for $Q=3$ and $Q=4$, lies on the replica-symmetric side of the AT line. Furthermore, the interesting weak retriever lies completely on this side. That is also the case for the tricritical point and the first-order transition line, $\alpha_2(b)$, for the three-state network, which at worst becomes marginally stable. Furthermore, the phase diagram for $T=T(\alpha)$ reveals that the line α_2 of continuous transitions is stable to replica-symmetry-breaking fluctuations, for both $Q=3$ and $Q=4$ and all b. In view of these results, it does not seem worthwhile to pursue a calculation beyond the replica-symmetry ansatz.

A closer look at our results reveals that although the critical capacity α_c, where the strong retriever terminates, decreases faster with increasing b for the four-state than for the three-state network, the trend is opposite for the lower and upper critical storage ratio α_1 and α_2, respectively, for the presence of a second retriever in the low-b regime. This suggests that the role of the threshold could become even more important in optimally adapted higher Q-state networks. The extended principle of adaptation of the present work assumes that both the training overlap and the training activity become continuously adapted to the noisy retrieval environment. In particular, the training activity follows the changes in the dynamical activity characteristic of the Q states of the units, and this makes difficult the study of the optimally adapted network for general Q. It may be possible to study a weaker version of the extended adaptation principle for the graded response network in which the training activity remains fixed. This, and other questions, will be considered in future work.

ACKNOWLEDGMENTS

We thank J. F. Fontanari and D. Bollé for critical comments, and one of us (W.K.T.) thanks the Institute for Theoretical Physics of the Catholic University of Leuven, Belgium, where part of the work was written, for their kind hospitality and financial support. The research of one of us (W.K.T.) was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil), and the work was supported in part by FINEP (Financiadora de Estudos e Projetos, Brazil). A grant from CNPq on a neural network project is gratefully acknowledged.

[1] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[2] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[3] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, 4659 (1990).

[4] T. B. Kepler and L. F. Abbott, J. Phys. (France) **49**, 1657 (1988).

[5] W. Krauth, J.-P. Nadal, and M. Mézard, J. Phys. A **21**, 2995 (1988).

[6] E. Gardner, J. Phys. A **22**, 1969 (1989).

[7] D. J. Amit, M. Evans, H. Horner, and K. Y. M. Wong, J. Phys. A **23**, 3361 (1990).

[8] E. Gardner, N. Stroud, and D. Wallace, J. Phys. A **22**, 2019 (1989).

[9] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, L175 (1990).

[10] G. Györgyi and N. Tishby, *Workshop on Neural Networks and Spin Glasses*, edited by W. K. Theumann and R. Koeberle (World Scientific, Singapore, 1990).

[11] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).

[12] T. L. H. Watkin, A. Rau, D. Bollé, and J. van Mourik, J. Phys. I **2**, 167 (1992).

[13] C. Meunier, D. Hansel, and A. Verga, J. Stat. Phys. **55**, 859 (1989).

[14] J. S. Yedidia, J. Phys. A **22**, 2265 (1989).

[15] D. R. C. Dominguez and W. K. Theumann, J. Phys. A **29**, 749 (1996); **30**, 1403 (1997).

[16] H. Rieger, J. Phys. A **23**, L1273 (1990).

[17] M. Bouten and A. Engel, Phys. Rev. E **47**, 1397 (1993).

[18] D. Bollé, G. M. Shim, B. Vinck, and V. A. Zagrebnov, J. Stat. Phys. **74**, 565 (1994).

[19] D. Bollé, H. Rieger, and G. M. Shim, J. Phys. A **27**, 3411 (1994).

[20] K. Y. M. Wong and D. Sherrington, Phys. Rev. E **47**, 4465 (1993).

[21] J. R. L. de Almeida and D. Thouless, J. Phys. A **11**, 983 (1978).

[22] E. Gardner, J. Phys. A **21**, 257 (1988).

[23] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).

[24] M. Bouten, J. Phys. A **27**, 6021 (1994).