PHYSICAL REVIEW E VOLUME 49, NUMBER 3 MARCH 1994

Response functions improving performance in analog attractor neural networks

Nicolas Brunel*
Istituto Nazionale di Fisica Nucleare, Dipartimento di Fisica, Piazzale Aldo Moro 2, 00185 Roma, Italy

Riccardo Zecchina†
Dipartimento di Fisica Teorica e Istituto Nazionale di Fisica Nucleare, Università di Torino, Via P. Giuria 1, 10125 Torino, Italy

(Received 10 December 1993)

In the context of attractor neural networks, we study how the equilibrium analog neural activities, reached by the network dynamics during memory retrieval, may improve storage performance by reducing the interferences between the recalled pattern and the other stored ones. We determine a simple dynamics that stabilizes network states which are highly correlated with the retrieved pattern, for a number of stored memories that does not exceed $\alpha_* N$, where $\alpha_* \in [0, 0.41]$ depends on the global activity level in the network and N is the number of neurons.

PACS number(s): 87.15.−v, 05.20.−y

Attractor neural networks (ANN's) have been the subject of intensive study among physicists since the original paper of Hopfield [1,2]. The analogy between the thermodynamics of ANN's and of spin glasses has been used to interpret the associative processes taking place in neural networks in terms of collective nonergodic phenomena. The identification of attractors with the internal representation of the memorized patterns, though still an object of open discussion, has received some basic experimental confirmations [3] and represents one of the basic issues for the biologically motivated models currently under study. Several authors have studied ANN's composed of analog neurons instead of the discrete spinlike neurons of the original model [4–7] and have shown that such more realistic networks may perform as associative memories. In the present paper we will be concerned with the following issue: assuming the interaction (synaptic) matrix to be the simple Hebb-Hopfield correlation matrix, we discuss how the storage performance of an ANN may depend on the equilibrium analog neural activities reached by the dynamics during memory retrieval.

In both discrete and analog Hopfield-like attractor neural networks, the phase transition of the system from associative memory to spin glass is due to temporal correlations arising from the static noise produced by the interference between the retrieved pattern and the other stored memories. The introduction of a suitable cost function in the space of neural activities allows us to study how such a static noise may be reduced and to derive a class of simple response functions for which the dynamics stabilizes the "ground-state" neural activities, i.e., the ones that minimize the cost function, for a macroscopic number of patterns.

In what follows, we first give some basic definitions and successively do the following.

(i) We study the ground states of a cost function E defined in the phase space $\mathcal{E}$ of the neural activities and proportional to the sum of the squared overlaps of the network state with the stored memories except the retrieved one.

(ii) We derive the associated gradient flow in $\mathcal{E}$ and show that it converges to the ground state of E.

(iii) We show that when the minimum of the cost function is zero there is a linear relation between the afferent current and the activity at each site of the network.

(iv) We determine a simple dynamics (which turns out to be characterized by a nonmonotonic response function) stabilizing the ground state of the system when the number of stored memories is smaller than $\alpha_* N$, where $\alpha_* \in [0, 0.41]$ depends on the global activity level.

The neural network model is assumed to be fully connected and composed of N neurons whose activities $\{V_i\}_{i=1,N}$ belong to the interval $[-1, 1]$. The global activity of the network is defined by

$$\gamma = \frac{1}{N}\sum_i \epsilon_i \quad , \tag{1}$$

where $\epsilon_i = |V_i|$ and we denote by $\mathcal{E} = [0,1]^N$ the space of all ϵ_i. A macroscopic set of $P = \alpha N$ random binary patterns $\Xi \equiv \{\{\xi_i^\mu = \pm 1\}_{i=1,N}, \mu = 1, P\}$, characterized by the probability distribution $P(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1)$, is stored in the network by means of the Hebb-Hopfield learning rule $J_{ij} = \frac{1}{N}\sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu$ for $i \neq j$ and $J_{ii} = 0$.

At site i, the afferent current (or local field) h_i is given by the weighted sum of the activities of the other neurons $h_i = \sum_j J_{ij} V_j$. We consider a continuous dynamics for the depolarization I_i at each site i [$\tau \dot{I}_i(t) = -I_i(t) + h_i(t)$], in which the activity of neuron i at time t is given by $V_i(t) = f(I_i(t))$, where f is the neuronal response function. No assumptions are made on f, except that it is such as to align the neural activity with its afferent current [$xf(x) \geq 0$ for all x].

We consider the case in which one of the stored patterns, $\mu = 1$ for example, is presented to the network via

*Electronic address: brunel@roma1.infn.it

†Also at Physics Department, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy; electronic address: zecchina@to.infn.it

an external current which forces the initial configuration of the network, thus we take $V_i = \epsilon_i \xi_i^1$. The current arriving at neuron i becomes

$$h_i = \gamma \xi_i^1 + \frac{1}{N} \sum_{\mu \neq 1} \xi_i^\mu \sum_{j \neq i} \xi_j^\mu \xi_j^1 \epsilon_j \tag{2}$$

which, in terms of the overlaps of the network state with the stored patterns $m_\mu \equiv \frac{1}{N} \sum_j \xi_j^\mu \xi_j^1 \epsilon_j$, reads

$$h_i = m_1 \xi_i^1 + \left(\sum_{\mu > 1} \xi_i^\mu m_\mu - \alpha \epsilon_i \xi_i^1 \right) . \tag{3}$$

Notice that the global activity is equal to the overlap of the network configuration with the retrieved pattern $(m_1 = \gamma)$.

The first term in the right-hand side (rhs) of Eq. (3) is the signal part, whereas the second term, when the ϵ_i are fixed and for N large, is a Gaussian random variable with zero mean and variance $\sigma^2 = \sum_{\mu > 1} m_\mu^2$ and represents the static noise part, or cross-talk, due to the interference between the stored patterns and the recalled one. In the following we will be interested in minimizing the overlaps of the network state with the stored memories $\mu \neq 1$ which are not recalled. If one succeeds in finding such states ($m_\mu = 0$ for all $\mu \neq 1$), then the second term in the rhs of Eq. (3) reduces to $-\alpha \epsilon_i \xi_i^1$, and the interference effect vanishes. We will see that this is indeed the case in a finite range of the parameter α and that it is also possible to derive a class of effective response functions realizing such a minimization of the interference and thus leading to an improvement of the network storage capacity.

(i) In order to study how the interference may be reduced, we define a cost function $E(\Xi, \boldsymbol{\epsilon})$ depending on the neural activities and the set of stored patterns, proportional to σ^2, i.e., to the sum of the squared overlaps of the network state with all the stored memories except the retrieved one

$$E(\Xi, \boldsymbol{\epsilon}) = \frac{1}{N} \sum_{\mu \neq 1} \left(\sum_j \xi_j^\mu \xi_j^1 \epsilon_j \right)^2 \tag{4}$$

and we study its ground states. If no constraints are present on $\boldsymbol{\epsilon}$, the minimum is always $E = 0$ for $\boldsymbol{\epsilon} = \mathbf{0}$; obviously, in this case no information is obtained when one presents the pattern and therefore we impose the constraint $\gamma = K$ on the average activity, where $K \in [0, 1]$. The geometrical picture of the problem is indeed very simple. In the space $\mathcal{E}$ we have to find the vectors $\boldsymbol{\epsilon}$ as orthogonal as possible to the $P - 1$ vectors $\boldsymbol{\eta}^\mu = \{\eta_i^\mu = \xi_i^1 \xi_i^\mu\}$, with the constraint $\gamma = K$. If there exists (with probability 1) at least one $\boldsymbol{\epsilon}$ orthogonal to the $P - 1$ vectors satisfying the constraint, then we have $E = 0$. The subspace corresponding to the condition $E = 0$ is connected since it is the intersection of $P - 1$ hyperplanes $m_\mu = 0$ and one hyperplane $\gamma = K$.

In order to determine the typical "free energy" we compute $\langle \ln Z \rangle_\Xi$, where $\langle \, \rangle_\Xi$ stands for the average over the quenched random variables Ξ and Z is the partition function at temperature $T = 1/\beta$ given by $Z = \mathrm{Tr}_{\boldsymbol{\epsilon}} \{ \exp[-\beta E(\Xi, \boldsymbol{\epsilon})] \delta(\gamma(\boldsymbol{\epsilon}) - K) \}$, using the standard replica method. Starting from the typical partition function of n replicas of the system $\langle Z^n \rangle_\Xi$, for n integer, we perform an analytic continuation for noninteger values of n, thus obtaining $\langle \ln Z \rangle_\Xi = \lim_{n \to 0} \frac{\langle Z^n \rangle_\Xi - 1}{n}$.

Each replica a $(a = 1, \ldots, n)$ is characterized by its neural activities $\boldsymbol{\epsilon}^a$ and by the order parameter $Q^a = \frac{1}{N} \sum_i (\epsilon_i^a)^2$, whereas the overlap between neural activities in two different replicas defines the other order parameters (for $a < b$) $q^{ab} = \frac{1}{N} \sum_i \epsilon_i^a \epsilon_i^b$. We indicate with R^a and r^{ab} the conjugate parameters of Q^a and q^{ab}, respectively. The typical free energy F per site is then given, in the thermodynamical limit, by $F(\beta) = -G(\beta)/\beta$ where $G(\beta) = \lim_{N \to \infty} \langle \ln Z(\beta) \rangle_\Xi / N$. The free energy at zero temperature, F_0, gives the ground state of the system. G can be calculated using a saddle-point method that, once a replica symmetric (RS) ansatz $q^{ab} = q$, $r^{ab} = r$ (for all $a < b$), and $Q^a = Q$, $R^a = R$ (for all a) has been done, leads to

$$G = \min_{\mathcal{M}} \left[\frac{1}{2}(rq + RQ + Ku) - \frac{\alpha}{2} \frac{\beta q}{1 + \beta(Q - q)} - \frac{\alpha}{2} \ln[1 + \beta(Q - q)] + \int D\zeta \ln \Gamma_{u,r,R}(\zeta) \right] , \tag{5}$$

where $\mathcal{M} = \{q, r, Q, R, u\}$,

$$\Gamma_{u,r,R}(\zeta) = \int_0^1 dz \exp \left[-\frac{R + r}{2} z^2 - \left(\frac{u}{2} + \sqrt{r} \zeta \right) z \right] , \tag{6}$$

and $D\zeta$ is the Gaussian measure $d\zeta g(\zeta)$ with $g(\zeta) = \exp(-\zeta^2/2)/\sqrt{2\pi}$. Depending on the storage level α one obtains the following results.

(a) For $\alpha < \alpha_0(K)$ we have $F_0 = 0$. The parameter q, which is the typical overlap between the activity configurations in two replicas such that the free energy vanishes, increases from $q = K^2$ at $\alpha = 0$ to $q = Q$ at $\alpha = \alpha_0$. At $\alpha = \alpha_0$, the space of neural activities such that $F_0 = 0$ shrinks to zero, and F_0 becomes positive.

(b) For $\alpha > \alpha_0(K)$, we have $q = Q$ and thus, in the limit $\beta \to \infty$, we introduce the new scaled variables $r = \rho \beta^2$, $R + r = \sigma \beta$, $u = 2\omega \beta$, and $x = \beta(Q - q)$. The order parameters are given by the following saddle-point equations:

$$Q = \int_{\zeta_0}^{\infty} D\zeta + \int_{\zeta_1}^{\zeta_0} D\zeta \left(\frac{\zeta - \zeta_1}{\zeta_0 - \zeta_1} \right)^2 , \tag{7}$$

$$K = \int_{\zeta_0}^{\infty} D\zeta + \int_{\zeta_1}^{\zeta_0} D\zeta \left(\frac{\zeta - \zeta_1}{\zeta_0 - \zeta_1} \right) , \tag{8}$$

$$x\sqrt{\rho} = \int_{\zeta_0}^{\infty} \zeta D\zeta + \int_{\zeta_1}^{\zeta_0} \zeta D\zeta \left(\frac{\zeta - \zeta_1}{\zeta_0 - \zeta_1} \right) , \tag{9}$$

$$\rho = \frac{\alpha Q}{(1 + x)^2} , \quad \sigma = \frac{\alpha}{1 + x} , \tag{10}$$

with $\zeta_0 = \sigma/\sqrt{\rho} + \zeta_1$, $\zeta_1 = \omega/\sqrt{\rho}$. $\alpha_0(K)$ (Fig. 1) is obtained in the limit $x \to \infty$ and the minimum F_0 is given by $F_0 = \alpha Q(1+x)^{-2}$ where Q and x are fixed by their saddle-point values.

The condition of local stability of the RS solution with respect to small fluctuations in replica space has been calculated. It turns out to be verified at $T = 0$ for all α, which is not surprising since, as previously noticed, the space of neural activities that minimize the cost function is connected.

The calculation of the probability distribution of the activities in the ground state can be done with similar techniques and yields

$$\mathcal{P}(\epsilon) = \int_{-\infty}^{+\infty} \frac{D\zeta}{\Gamma_{u,r,R}(\zeta)} \exp\left[-\frac{R+r}{2}\epsilon^2 - \left(\frac{u}{2} + \sqrt{r}\zeta\right)\epsilon \right] , \tag{11}$$

where u, r, and R take their saddle-point values. Above the critical storage level $\alpha \geq \alpha_0$, the probability distribution reads, for $\epsilon \in [0,1]$,

$$\mathcal{P}(\epsilon) = H(\zeta_0)\delta(\epsilon - 1) + \Delta_\zeta g(\zeta_1 + \epsilon\Delta_\zeta) + H(-\zeta_1)\delta(\epsilon) \quad , \tag{12}$$

where $H(\zeta) \equiv \int_\zeta^\infty Dz$, $\Delta_\zeta = \zeta_0 - \zeta_1$, and ζ_0 and ζ_1 are given as a function of α and K by Eqs. (7)–(10). Notice the two δ functions in 0 and 1.

For brevity, we do not report here the results [10] concerning the case of discrete (three-states $\{\pm 1, 0\}$) neurons; we just mention that replica symmetry breaking is required.

(ii) The next step regards the gradient flow associated with a smooth version of the energy function (4), implementing a soft quadratic constraint for the global activity. The study of this gradient flow will allow us, on the one hand, to find the relation between activities and afferent currents in the ground state, and on the other, to check the outcome of the RS solution. We emphasize that this gradient flow does not correspond to the dynamics of the network – this point will be considered later in (iii) and (iv).

The new cost function E_λ can be written

$$E_\lambda(\Xi, \epsilon) = \frac{1}{N}\sum_{\mu \neq 1}\left(\sum_j \xi_j^\mu \xi_j^1 \epsilon_j\right)^2 + \frac{\lambda}{N}\left(\sum_i \epsilon_i - KN\right)^2 \quad , \tag{13}$$

where $\lambda > 0$ is a Lagrange multiplier. In the first term one recognizes the previous cost function (4) whereas the second term is introduced in order to favor configurations with activity K. The ground state F_λ of $E_\lambda(\Xi, \epsilon)$ can be calculated [10] with the same methods as the ground state of (4). For $\alpha < \alpha_0(K)$ we have $F_\lambda = 0$ and $\gamma = K$, while for $\alpha > \alpha_0(K)$ the ground state F_λ becomes positive, and we have $\gamma < K$. By computing the gradient of $E_\lambda(\Xi, \epsilon)$, it is now easy to derive the flow in the ε space: in $\mathbb{R}^N$ we have

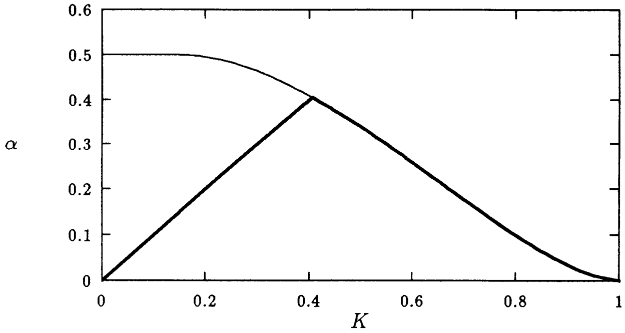

FIG. 1. Bold curve: α_* versus K. Light curve: α_0 versus K.

$$\dot{\boldsymbol{\epsilon}} = -\frac{\tau'}{2}\nabla E_\lambda = \tau'(A\boldsymbol{\epsilon} + \mathbf{b}) \quad , \quad \tau' > 0 \tag{14}$$

where $\{A_{ij} = -(\frac{1}{N}\sum_\mu \eta_i^\mu \eta_j^\mu - \frac{1-\lambda}{N})\}$ and $\{b_i = \lambda K\}$, $i, j = 1, N$. Discretizing time, constraining the ϵ_i to stay in the $[0,1]$ interval, and choosing $\tau' = \frac{1}{\alpha}$, we arrive at the following local equation:

$$\epsilon_i(t+1) = \phi\left(\frac{1}{\alpha}\left\{\lambda[K - \gamma(t)] + \gamma(t) - h_i^1(t)\xi_i^1\right\}\right) \quad , \tag{15}$$

where $\phi(x) = 0$ if $x < 0$, $\phi(x) = 1$ if $x > 1$ and $\phi(x) = x$ otherwise, $\gamma(t)$ is the global activity and $h_i(t)$ is the afferent current at site i. Since E_λ is a positive semidefinite quadratic form, every local minimum in $[0,1]^N$ is also an absolute minimum in the same interval and hence the gradient flow converges to the ground state of E_λ. In Fig. 2 we compare the ground-state energy F_λ computed analytically with that given by simulations of Eq. (15) for a network of $N = 1000$ neurons. It shows a remarkable agreement between the analytic solution and the numerical simulations.

(iii) When $F_\lambda = 0$ (which implies $\gamma = K$) the minimum of the cost function defined on $[0,1]^N$ is also an absolute minimum of the same function defined on $\mathbb{R}^N$ and therefore its gradient vanishes. This leads to a very simple relation between the activity and the stability $\Delta_i \equiv h_i \xi_i$ at each site [upon inserting in (15) $\gamma = K$]

$$K - \Delta_i = \alpha\epsilon_i \quad , \tag{16}$$

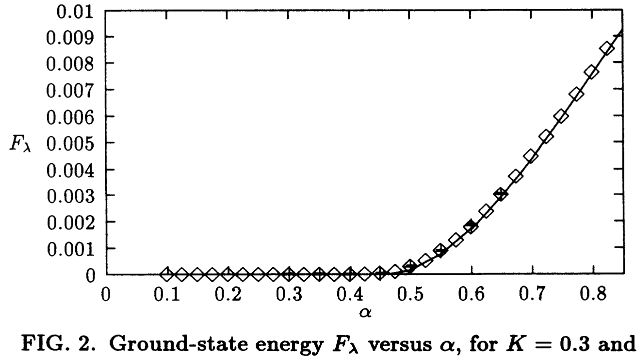

FIG. 2. Ground-state energy F_λ versus α, for $K = 0.3$ and $\lambda = 1.2$. Continuous curve: analytical calculation. Diamonds: simulations on a network of $N = 1000$ neurons.

which also coincides with the expression of the afferent currents, Eq. (3), in which $m_\mu = 0$ for $\mu > 1$. The above relation yields straightforwardly the probability distribution of the Δ's $\mathcal{P}(\Delta_i = \Delta) = \mathcal{P}[\epsilon_i = (K - \Delta)/\alpha]$, which implies that, for $\alpha \leq \alpha_0(K)$, this distribution is bounded between $\Delta = K$ and $\Delta = K - \alpha$. For $\alpha > \alpha_0(K)$, due to the nonlinear constraint on the bounds of the activities, the flow (14)–(15) reaches a fixed point which does not coincide with the minimum of E_λ in $\mathbb{R}^N$, and therefore the stabilities distribution is no longer given by Eq. (16).

(iv) Under the initial assumption on the neuronal transfer function $[xf(x) \geq 0]$, the condition for a ground-state configuration correlated with the presented pattern to be a fixed point of the dynamics is to have positive Δ's at all sites. This indeed happens if the storage level α satisfies $\alpha < \alpha_*(K)$ where $\alpha_*(K)$ is the critical line identified by $\alpha_*(K) = \min(K, \alpha_0(K))$ and shown in Fig. 1.

It follows from Eq. (16) that, when $\alpha < \alpha_*(K)$, the ground-state activities are fixed points of the network dynamics with the (nonmonotonic) transfer function f

$$f(h) = \begin{cases} \mathrm{sgn}(h) & \text{if } |h| \in [0, K-\alpha] \\ \mathrm{sgn}(h)(K - |h|)/\alpha & \text{if } |h| \in [K-\alpha, K] \\ 0 & \text{if } |h| > K \end{cases} \qquad (17)$$

shown in Fig. 3. In the same figure we also report, for $\alpha < \alpha_*(K)$, the equilibrium distribution of the local currents obtained by numerical simulations performed on a network with such a dynamics. All the Δ's belong to the interval $[K - \alpha, K]$, as expected. It is worth noticing that, at equilibrium, only the region $\mathcal{R} = [-K, -(K - \alpha)] \cup [K - \alpha, K]$ plays a role; as far as the equilibrium properties are concerned, outside this interval the form of the transfer function is arbitrary. In $\mathcal{R}$, the dynamical behavior induced by f corresponds to a regulation of the output activity at each site. The latter turns out to be proportional to the difference between the afferent current and a reference feedback signal equal to the global average activity.

If the fixed points are stable, it means that for $\alpha < \alpha_*(K)$ a network with the discussed dynamics is capable of stabilizing a configuration with activity K highly correlated with the retrieved pattern: the optimal activity is $K_{\rm opt} = 0.41$ for which we have $\alpha_*(K_{\rm opt}) = 0.41$. The stability of the fixed points is difficult to prove, since the distribution of the local currents has peaks at points where the derivative of the transfer function is discontinuous. We have checked numerically their stability for $\alpha < \alpha_*(K)$: in order to correctly initialize the system, we have used Eq. (15) to find one set of initial ground-state activities $\{\epsilon_i(0)\}$ and then took $V_i(0) = \xi_i^1 \epsilon_i(0)$.

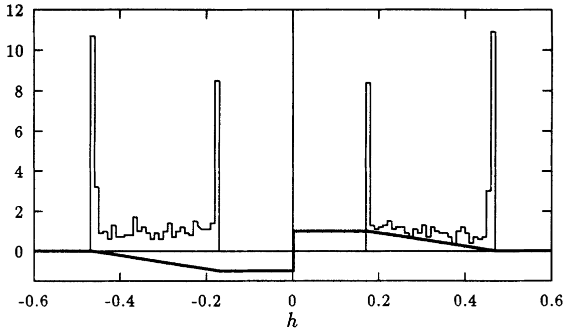

FIG. 3. Bold curve: neuronal transfer function for $\alpha = 0.3$ and $K = 0.47$. Light curve: distribution of the currents at equilibrium, for a network of size $N = 1000$ with a neuronal transfer function characterized by the same parameters and after presentation of one of the stored patterns ξ^ν $[V_i(t=0) = \xi_i^\nu$ for all $i]$.

Usually, the critical capacity is defined as an upper bound for the presence of retrieval states correlated with the presented pattern. In the case of a sigmoid transfer function, the critical capacity is obtained with a finite σ^2 and yields ~ 0.14 [7,8]. Here α_* is derived as an upper bound for the presence of retrieval states with $\sigma^2 = 0$. This means that α_* is a lower bound for the critical capacity, which is expected to be higher in the region where the Δ's are strictly positive at all sites, i.e., $K > \alpha_*$ or $K > 0.41$. Interestingly enough, our results on the maximal storage capacity $\alpha_*(K_{\rm opt})$ are very similar to an estimate obtained in [9] by a completely different method on a particular nonmonotonic transfer function.

The question of the size of basins of attraction in such a network remains open [11]. Obviously, the condition of local stability does not ensure that starting from an initial configuration highly correlated with a stored memory [as, for instance, $\{V_i(t=0) = \xi_i^\mu\}, i = 1, \ldots, N$], the network will converge to a ground-state configuration belonging to the same memory. Preliminary numerical simulations show that the basins of attraction can be considerably enlarged if one uses a dynamical threshold $\theta(t)$ instead of K in Eq. (17), determined at time t by the instantaneous global activity of the network $[\theta(t) = \gamma(t)]$, given by Eq. (1) at time t. Such dynamical nonmonotonic behavior might be seen as an effect of a regulatory mechanism of the global activity in the network, which in real cortical networks is supposed to be due to inhibitory interneurons.

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

[2] D.J. Amit, *Modeling Brain Function* (Cambridge University Press, New York, 1989).

[3] Y. Miyashita and H.S. Chang, Nature (London) **331**, 68 (1988); **335**, 817 (1988).

[4] J.J. Hopfield, Proc. Natl. Acad. U.S.A. **81**, 3088 (1984)

[5] D.J. Amit and M.V. Tsodyks, Network **2**, 259 (1991); **2**, 275 (1991).

[6] M. Shiino and T. Fukai, J. Phys. A: Math. Gen. **23**, L1009 (1990).

[7] R. Kuhn, in *Statistical Mechanics of Neural Networks*, edited by L. Garrido (Springer, Berlin, 1991); R. Kuhn and S. Bos, J. Phys A: Math. Gen. **26**, 831 (1993).

[8] D.J. Amit , H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).

[9] S. Yoshizawa , M. Morita, and S. Amari, Neural Networks **6**, 167 (1993).

[10] N. Brunel and R. Zecchina, Politecnico of Torino, Physics. Dept., Report No. POLFIS-TH-25, 1993 (unpublished).

[11] Work on an extremely diluted network is in progress.