Disappearance of spurious states in analog associative memories

Yasser Roudi* and Alessandro Treves
SISSA, Programme in Neuroscience, via Beirut 4, 34014 Trieste, Italy
(Received 20 August 2002; published 17 April 2003)

We show that symmetric *n*-mixture states, when they exist, are almost never stable in autoassociative networks with threshold-linear units. Only with a binary coding scheme, we could find a limited region of the parameter space in which either 2-mixture or 3-mixture states are stable attractors of the dynamics.

DOI: 10.1103/PhysRevE.67.041906 PACS number(s): 87.19.La, 87.18.Sn

I. INTRODUCTION

Autoassociative networks are useful models of one of the basic operations of cortical networks [1]. ''Hebbian'' plasticity on recurrent connections, e.g., in the higher level areas of sensory cortex and in the hippocampus, is the crucial ingredient for autoassociation to work, with real neurons [2]. Neural network models, although very simplified and abstract, allow a comprehensive analysis, indicating whether associative memory retrieval can proceed safely, or whether it must face dynamical hurdles, such as ''spurious'' local minima in a free-energy landscape. The dynamics of such networks, in the simplest models, is governed by a number p of dynamical attractors, each of which corresponds to a distribution of neural activity, i.e., a pattern, which represents a long term memory. Memory is stored by superimposed synaptic weight changes, and the basic operation proceeds by supplying the network with an external signal that acts as a cue, correlated, perhaps only weakly, with a pattern, and which leads through attractor dynamics to the retrieval of the full pattern.

How smoothly can such an operation proceed, and how wide are the basins of attraction of the p memory states? Clearly, these issues depend critically on whether other attractors exist, that could hinder or obstruct retrieval. As a crude example, if the cue is correlated with the image of a mule, the net may be able to retrieve either a horse or a donkey, if no ''mixed'' attractor exist. If instead the encoding procedure has, unintentionally, created a spurious attractor for the mule itself, the network will likely be stuck in such a mixed memory state. In a slightly more complicated model endowed with some topographic mapping of visual space, a horse cue and a donkey cue might be presented simultaneously in neighboring positions. If they are too close in visual space and spurious attractors exist, this topographic map might retrieve two mules next to each other. Returning to nets without spatial structure and considering for simplicity only symmetric mixtures of patterns embedded with equal strengths, there are obviously $p(p-1)/2$ 2-mixture, $p(p-1)(p-2)/6$ 3-mixture states, and so on. Do they correspond to stable attractors, and as such do they influence the network dynamics?

In addition, connectionist modelers have proposed to describe in terms of spurious states certain psychiatric dysfunctions [3]. Speech disorders in schizophrenic patients, for instance, might arise from the existence of a large number of spurious states, that obstruct the retrieval of correct patterns [4].

In their seminal investigation of the Hopfield model [5], Amit, Gutfreund, and Sompolinsky found that while symmetric mixtures of an even number of patterns are unstable, odd mixtures and the spin-glass phase can be stable, in a certain region of phase space [6]. In the Hopfield model, though, neurons are modeled as binary units, and correspondingly each distribution of activity, in particular each memory pattern, is a binary vector. Either or both of these aspects might be essential in producing the additional minima in the free-energy landscape. Real neurons behave very differently from binary units in many respects, a basic one being that their spiking activity, once filtered with a short time kernel [7], is better approximated by an analog variable. Threshold-linear units reproduce this graded nature of neural response, yet still allow for a simple and complete statistical mechanics analysis of autoassociative network models [8]. With threshold-linear units, the memory patterns encoded in the synaptic weights can still be taken to be binary vectors, but can also be taken to be drawn from a distribution with several discrete activity values, or from a continuous distribution [2]. Exponential distributions, in particular, can be argued to be not far from experimentally observed spike count distributions [9].

The question of mixture states in analog nets was first addressed in Ref. [10], arguing that the multiple local minima of the spin glass phase are fewer in number in an associative net of units with more continuous (sigmoid) transfer function. Later it was found, considering threshold-linear units, that are both realistic and amenable to analytical treatment [2], that the region of stability of the spin-glass phase is severely restricted with such units [11], again indicative of a general smoothing of the free-energy landscape with analog variables. Although these analyses provide a good starting point, they are not complete in the sense that they did not show what will happen to n-mixture states with n small (the ones relevant to models of schizophrenia), and what is the effect of different coding schemes, that is pattern distributions. Here, we consider instead symmetric n-mixtures, with $n=2,3,\ldots$, and we consider nonbinary memory vectors. Also, from the biological point of view, it is important to study nets with diluted (incomplete) connectivity, which are much more realistic descriptions of cortical [12] and hippocampal networks [13], where the probabiliy of

*Email address: yasser@sissa.it

a recurrent connection between any two units may be of the order of a few percent.

In this manuscript, we show that symmetric mixture states give rise to dynamical attractors only in very restricted circumstances, in associative networks of threshold-linear units, both with full and diluted connectivity. We have analyzed the validity of this statement in different coding schemes, and did not find any stable mixture state at all, when memory patterns are not binary. Essentially, we conclude that this type of spurious states are a pathological feature of the simplified binary models considered in the initial studies.

II. THRESHOLD-LINEAR MODEL

We use a model very similar to that analyzed in Ref. [8]. We consider a fully connected network of N units, taken to model excitatory neurons. The level of activity of unit i is a dynamical variable $V_i \geqslant 0$, which corresponds to the short time averaged firing rate of the neuron. Units are connected to each other through symmetric weights. The specific covariance Hebbian learning rule we consider prescribes that the synaptic weight between units i and j be given as

$$J_{ij}=\frac{1}{Na^2}\sum_{\mu=1}^{p}(\eta_i^\mu-a)(\eta_i^\mu-a), \tag{1}$$

where η_i^μ represents the activity of unit i in pattern μ. Each η_i^μ is taken to be a quenched variable, drawn independently from a distribution $p(\eta)$, with the constraints $\eta \geqslant 0$, $\langle \eta \rangle_\eta = \langle \eta^2 \rangle_\eta = a$. As in one of the first extensions of the Hopfield model [14], we thus allow for the mean activity a of the patterns to differ from the value $a=1/2$ of the original model [8].

The model further assumes that the input to unit i takes the form

$$h_i=\sum_{j\neq i} J_{ij}^c V_i+\sum_\nu s^\nu \frac{(\eta_i^\nu-a)}{a}+b\left(\frac{1}{N}\sum_j V_j\right), \tag{2}$$

where the first term enables the memories encoded in the weights to determine the dynamics, the second term allows for external signals s^ν to cue the retrieval of one or several patterns, and the third term is unrelated to the memory patterns, but is designed to regulate the activity of the newtork, so that at any moment in time $(1/N)\Sigma_i V_i=(1/N)\Sigma_i V_i^2=a$. The activity of each unit is determined by its input through a threshold-linear function

$$V_i=g(h_i-T_{thr})\Theta(h_i-T_{thr}), \tag{3}$$

where T_{thr} is a threshold below which the input elicits no output, g is a gain parameter, and $\Theta(\cdots)$ is the Heaviside step function. Units are updated, for example, sequentially in random order, possibly subject to fast noise. The exact details of the updating rule and of the noise are not specified further, here, because they do not affect the steady states of the dynamics, and we take the noise level T to be vanishingly small, $T\to 0$. Discussions about the biological plausibility of this model for networks of pyramidal cells can be found in Refs. [2,15], and will not be repeated here.

Subject to the above dynamics, the network evolves towards one of a set of attractor states. In a given attractor, the network may still wander among a variety of configurations, but it reaches a stationary probability distribution of being in any particular configuration. The average of any quantity over such ''annealed'' probability distribution is denoted by $\langle\ \rangle$ [whereas $\langle\ \rangle_\eta$ denotes the average over the quenched distribution $p(\eta)$]. To analyze such a model one can introduce, as in Ref. [8] the order parameters are

$$x=\sum_{i=1}^{N} v_i, \tag{4}$$

$$x^\sigma=\frac{1}{Na}\sum_{i=1}^{N}\eta_i^\sigma v_i-x, \tag{5}$$

$$y_0=\frac{1}{N}\sum_{i=1}^{N}\langle V_i^2\rangle, \tag{6}$$

$$y_1=\frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle^2, \tag{7}$$

where x is simply the mean activity of the network, and x^σ, is the subtracted, or specific, overlap of the current state of the network with each of the stored patterns. Two further parameters,

$$\psi=(y_0-y_1)T_0/T, \tag{8}$$

$$\rho=\frac{py_1}{[N(1-\psi)^2]}, \tag{9}$$

can be defined as a function of y_0 and y_1, and play a particularly useful role in the analysis in the limit we consider, $T\to 0$, when one configuration dominates the annealed average, and $y_1\simeq y_0+O(T)$. The characteristic noise scale of the system is $T_0\equiv(1-a)/a$ [8], and we define the storage load $\alpha\equiv p/N$. In the limit $N\to 0, T\to 0$, the system is thus characterized by the parameters a (mean pattern activity, which also parametrizes the coding sparseness [8] in the sense that decreasing a makes the code sparser), α (storage load), g (gain), and T_{thr} (threshold).

III. MEAN-FIELD SOLUTIONS AND THEIR STABILITY

We calculate the free energy using the replica trick, for symmetric n-mixture states (where n overlaps take the same nonzero value, and the rest are zero) elicited by external signals $s^1=\cdots=s^n=s$. These signals can be purely transient, so that at steady state $s=0$, but we consider a nonzero steady value for the sake of generality. We look for symmetric states, characterized by nonzero $\hat{x}^1=\cdots=\hat{x}^n=\hat{x}$. The saddle point equations reduce to [8]

$$x = g' \left\langle \int_{h>T_t hr} Dz (h - T_{thr}) \right\rangle_\eta , \tag{10}$$

$$x^\sigma = g' \left\langle \left(\frac{\eta^\sigma}{a} - 1 \right) \int_{h>T_t hr} Dz (h - T_{thr}) \right\rangle_\eta , \tag{11}$$

$$\psi = T_0 g' \left\langle \int_{h>T_t hr} Dz \right\rangle_\eta , \tag{12}$$

$$y_0 = (g')^2 \left\langle \int_{h>T_t hr} Dz (h - T_{thr})^2 \right\rangle_\eta , \tag{13}$$

$$\rho^2 = \frac{\alpha y_0}{(1-\psi)^2}, \quad h_2 = \frac{\alpha T_0}{2(1-\psi)}, \quad g' = \frac{g}{1 - 2g h_2}, \tag{14}$$

where now the input to each unit can be expresed as

$$h = b(x) - \sum_\sigma x^\sigma + \sum_\sigma \frac{\eta^\sigma}{a} (x^\sigma + s^\sigma) - z T_0 \rho \tag{15}$$

and the free energy reads

$$f = -\frac{g'}{2} \left\langle \int_{h>T_t hr} Dz (h - T_{thr})^2 \right\rangle_\eta + \frac{1}{2} \sum_\sigma (x^\sigma)^2 + x b(x) - B(x) + \frac{T_0}{2} \psi \rho^2 .$$

If one defines new parameters $v = (\hat{x} + s)/(T_0 \rho)$ (the specific signal-to-noise ratio) and $w = [b(x) - n\hat{x} - T_{thr}]/(T_0 \rho)$ (a sort of uniform field-to-noise ratio), it is easy to show that the mean-field equations can be reduced to

$$E_1(w,v) = (A_1 + \delta A_2)^2 - \alpha A_3 = 0, \tag{16}$$

$$E_2(w,v) = (A_1 + \delta A_2) \left(\frac{1}{g T_0 (1+\delta)} - A_2 \right) - \alpha A_2 = 0, \tag{17}$$

where

$$A_1(w,v) = A_2(w,v) - \left\langle \int^+ Dz \right\rangle_\eta , \tag{18}$$

$$A_2(w,v) = \frac{1}{n v T_0} \left\langle \left(\frac{\Gamma}{a} - n \right) \int^+ Dz \left(w + v \frac{\Gamma}{a} - z \right) \right\rangle_\eta , \tag{19}$$

$$A_3(w,v) = \left\langle \int^+ Dz \left(w + v \frac{\Gamma}{a} - z \right)^2 \right\rangle_\eta , \tag{20}$$

with $\Gamma = \Sigma_{\sigma=1}^n \eta^\sigma$ and $\delta = s/\hat{x}$.

In the equations above, $Dz = (dz/\sqrt{2\pi}) e^{-z^2/2}$ and the subscript $+$ indicates that the z average has to be carried out only in the range where $w + v(\Gamma/a) - z > 0$. In the following, we take $\delta = 0$. Thus symmetric n-mixture attractors exist if we can find stable solutions of Eqs. (16) and (17).

To analyze the stability of the extrema of the free energy, one has to study the hessian matrix

$$H_{\mu\nu} = \delta_{\mu\nu} - \left\langle \left(\frac{\eta^\mu}{a} - 1 \right) \left(\frac{\eta^\nu}{a} - 1 \right) \int^+ Dz \right\rangle_\eta \tag{21}$$

around the saddle point.

In general, for n-mixture states, there are three types of eigenvalues as follows:

(1) A nondegenerate eigenvalue, which decides the stability against a uniform increase in the amplitude of the n patterns that contribute to the thermodynamic state (i.e., the ''condensed'' patterns), while the other overlaps remain zero. It is (for $\mu \neq \nu$)

$$\lambda_1 = 1 - \left\langle \left[\left(\frac{\eta^\mu}{a} - 1 \right)^2 + n \left(\frac{\eta^\mu}{a} - 1 \right) \left(\frac{\eta^\nu}{a} - 1 \right) \right] \int^+ Dz \right\rangle_\eta . \tag{22}$$

(2) An eigenvalue of degeneracy $n-1$, associated with any direction which tends to change the relative amplitude of the nonzero overlaps. It is (again for $\mu \neq \nu$)

$$\lambda_2 = 1 - \left\langle \left[\left(\frac{\eta^\mu}{a} - 1 \right)^2 - \left(\frac{\eta^\mu}{a} - 1 \right) \left(\frac{\eta^\nu}{a} - 1 \right) \right] \int^+ Dz \right\rangle_\eta . \tag{23}$$

(3) The third eigenvalue, with degeneracy $p-n$, measures the stability against the appearance of additional overlaps. It is

$$\lambda_3 = 1 - T_0 \left\langle \int^+ Dz \right\rangle_\eta . \tag{24}$$

IV. DIFFERENT CODING SCHEMES

In order to proceed further, we restrict the analysis to a number of specific coding schemes, i.e., to different choices for the distribution $p(\eta)$. We consider

$$p(\eta) = a\,\delta(\eta - 1) + (1-a)\,\delta(\eta), \quad \text{binary}$$

$$p(\eta) = \frac{a}{3} \delta\left(\eta - \frac{3}{2} \right) + a\,\delta\left(\eta - \frac{1}{2} \right) + \left(1 - \frac{4a}{3} \right) \delta(\eta), \quad \text{ternary}$$

$$p(\eta) = 4a e^{-2\eta} + (1-2a)\,\delta(\eta), \quad \text{exponential.} \tag{25}$$

For small values of the load α (and hence of the quenched noise ρ), Eq. (17) describes an hyperbole, whose center depends on the value of g. Eq. (16) instead, for small values of α, is a closed curve in the quartant $w < 0$, $v > 0$, so that with an appropriate choice of g the two curves intersect at two points. As α grows, the region $E_1(v,w) > 0$ shrinks in size,

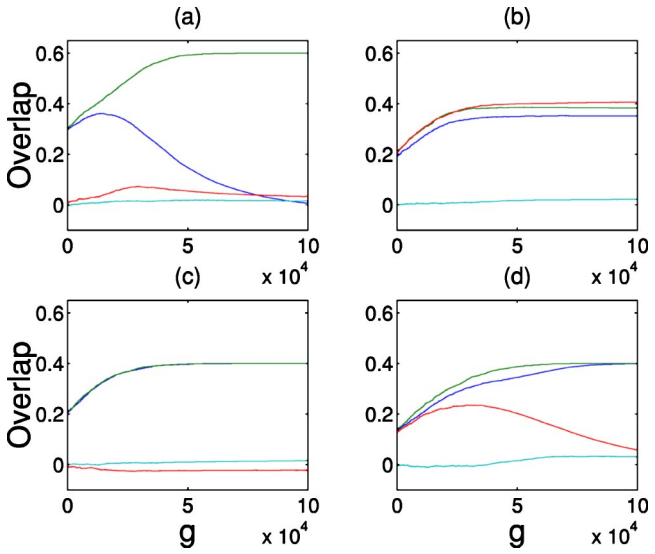

FIG. 1. Computer simulation result $N=10\,000$, $p=5$, and (a),(b) represent $g=1.2$, $a=0.4$; (c),(d) represent $g=3$, $a=0.6$. In (a),(c) the initial state was correlated equally with two patterns and in (b),(d) with 3.

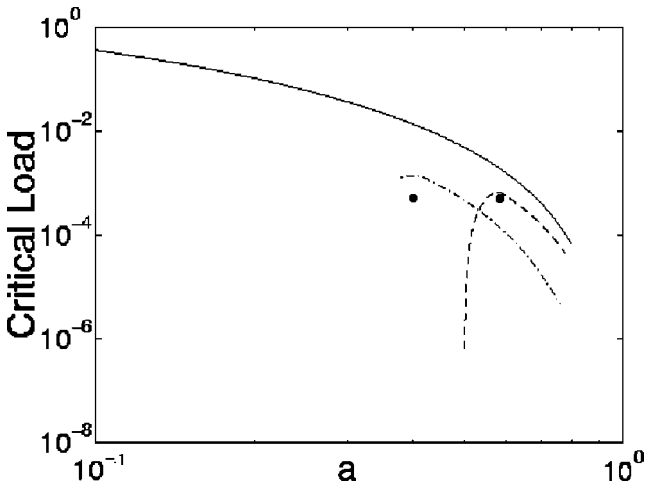

FIG. 2. Storage capacity as a function of sparseness for the single pattern states (full line), compared to the region of existence and stability of 2-mixture (dashed line) and 3-mixture states (dashed dotted line), for the binary coding scheme, in a fully connected network with threshold-linear units. Points denote the a,α values used in the simulations of Fig. 1.

until at a certain value of α, which depends only on a,n, and the coding scheme, it reduces to a point and then disappears.

We have investigated not just the existence but also the stability of solutions for symmetric 2- and 3-mixture states. The solutions behave exactly in the same manner in these two cases: for small values of a both intersections discussed above are unstable, in the sense that both λ_1 and λ_2 are negative. This finding is confirmed by computer simulation, in which one of the overlaps tends to grow, reaching the corresponding attractor, whereas the other one (or the other two in the case of 3-mixtures states) tends to zero. Increasing the value of the sparsity parameter, one finds different results with binary coding and with other types of coding.

Let us consider binary coding first. After a range of a values with only one unstable eigenvalues (λ_1 or λ_2), one finds a range where genuinely stable solutions can be found. Thus the retrieval of mixture patterns is possible for binary coding, as can be seen in the simulations shown in Fig. 1.

The exact stability region in the (α,a) plane differs for 2-mixture and 3-mixture states. In both cases, it is delimited to the right by the ''critical load'' $\alpha_c(a,n)$, i.e., the value at which the island with $E_1(v,w)>0$ shrinks to zero, and to the left by the load α beyond which no intersection with both $\lambda_1>0$ and $\lambda_2>0$ can be found. Fig. 2 illustrates these stability regions, compared with the critical load for the pure attractor states, as in Ref. [8].

For ternary and exponential coding, the solutions of the saddle point equations remain unstable even for very high values of the sparsity parameter a. Again, this was verified by computer simulations. Fig. 3 illustrates the different situation occurring with ternary and binary coding, by considering a very low load and a sparsity value for which stable solutions for 3-mixture states are easily found in the binary case. Note that the critical load for 3-mixture states would be considerably higher with ternary patterns (not shown); the fact is that at each position of the intersection, either λ_1 or λ_2 or both turn out to be negative. This complex behavior of eigenvalues will be discussed elsewhere in more detail.

V. DILUTED CASE

We have also extended the analysis to a highly diluted network [16]. In this case, the number of patterns that can be stored scales with the number C of connections each unit receives, rather than with the number of units N. One then redefines the load parameter as $\alpha\equiv p/C$. The essential difference introduced by the sparse (i.e., diluted) connectivity is that noise has less of an opportunity to reverberate along closed loops. In fact the signal, which during retrieval is simply contributed by the ''condensed'' patterns, propagates coherently and proportionally to C, independently of the density of feedback loops in the network. The fluctuations in the overlaps with the undecondensed patterns, which as $T\rightarrow 0$

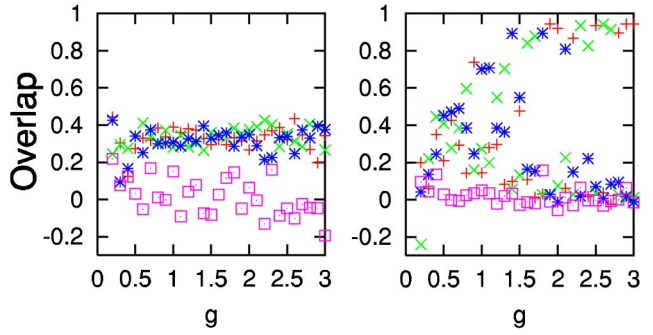

FIG. 3. The final overlaps with each of four stored patterns, averaged over the last 400 updates of the whole network, for left (binary) and right (ternary) coding, in both cases with $N=10000$, $a=0.5$, and $p=4$. Three patterns have initially a nonzero overlap with the activity of the network and retain it, for most g values, in the binary coding case, while a single pattern is always selected in the ternary case.

represent the sole source of noise, propagate coherently along feedback loops, giving rise to the amplifying factor $1/(1-\psi)$ of the fully connected case. For a given load (fixed α), diluted connectivity reduces, therefore, the influence of this ''static'' noise, and performance is better than in the fully connected case with $N-1=C$. In particular, with the extreme dilution, that is, if the condition $(c/Ln(N))\rightarrow 0$ is satisfied, one can neglect correlations among the C inputs to a given unit [16], and the mean field equations become [17]

$$E_1(w,v)=(A_2+\delta A_2)^2-\alpha A_3=0, \tag{26}$$

$$E_2(w,v)=\left(\frac{1}{gT_0(1+\delta)}-A_2\right)=0. \tag{27}$$

Examining again the stability matrix, we find that the mixture solutions, that were present with binary coding and large values of a, still survive. By the token, the results for ternary and exponential coding are not affected, in the sense that no stable solutions can be found even in the highly diluted case.

VI. CONCLUSION

The conclusion is that the existence of stable mixture states in a restricted region of the parameter space should be regarded as almost a pathological feature, resulting from binary coding. If one considers mixture states as spurious states, to be avoided, then one notes that the introduction of analog variables, a more realistic description of neural activity, goes a long way towards disposing of spurious states, just as it almost eliminated the spin glass phase [11]. The remaining region of stability of spurious states is definitely eliminated by nonbinary coding schemes, that further contribute to smooth the free-energy landscape. This result casts doubts upon, e.g., models of schizophrenia that are based on the existence of spurious attractors.

These results may well have implications in domains outside computational neuroscience. The smoothness of the free-energy landscape is a crucial features of many interacting systems used to map optimization problems, such as the traveling salesman [18] or the graph matching problem [19]. Optimization generally fails if the dynamics gets stuck into local minima. Our result indicates that undesired local minima may be eliminated by a combination of analog variables and coding schemes, which may in some cases be manipulated while mapping the problem at hand onto a dynamical system.

[1] E.T. Rolls and A. Treves, *Neural Networks and Brain Function* (Oxford University Press, Oxford, 1998).
[2] A. Treves and E.T. Rolls, Network **2**, 371 (1991).
[3] D.J. Amit, *Modeling Brain Function*, (Cambridge University Press, Cambridge, 1989).
[4] R.E. Hoffman, Arch. Gen. Psychiatry **44**, 178 (1987).
[5] J.J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[6] D.J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); D.J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987).
[7] Such short time kernel may be thought to correspond to the conversion of presynaptic spikes into postsynaptic potentials.
[8] A. Treves, Phys. Rev. A **42**, 2418 (1990).
[9] W.B. Levy and R.A. Baxter, Neural Comput. **8**, 531 (1996); R.J. Baddeley *et al.*, Proc. R. Soc. London, Ser. B **264**, 1775 (1997); A. Treves *et al.*, Neural Comput. **11**, 611 (1999).
[10] F.R. Waugh *et al.*, Phys. Rev. Lett. **64**, 1986 (1990).
[11] A. Treves, J. Phys. A **24**, 2645 (1991).
[12] V. Braitenberg and A. Schütz, *Statistics of the Cortex: Anatomy and Geometry* (Springer, Berlin, 1991).
[13] A. Treves and E.T. Rolls, Hippocampus **2**, 199 (1992).
[14] M.V. Tsodyks and M.V. Feigel'man, Europhys. Lett. **6**, 101 (1988).
[15] D.J. Amit and M.V. Tsodyks, Network **2**, 259 (1991); **2**, 275 (1991).
[16] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[17] A. Treves, J. Phys. A **24**, 327 (1991).
[18] J.J. Hopfield and D.W. Tank, Science **233**, 625 (1986).
[19] Y. Fu and P.W. Anderson, J. Phys. A **19**, 1605 (1986).