CNPq

CBPF - CENTRO BRASILEIRO DE PESQUISAS FÍSICAS

Notas de Física

CBPF-NF-038/92

OPTIMAL HEBBIAN LEARNING RULES AND THE ROLE OF ASYMMETRY

by

D.A. STARIOLO and C. TSALLIS

Abstract

We study the storage properties associated with generalized Hebbian learning rules which present four free parameters that allow for asymmetry. We also introduce two extra parameters in the post-synaptic potentials in order to further improve the critical capacity. Using signal-to-noise analysis, as well as computer simulations on an analog network, we discuss the performance of the rules for arbitrarily biased patterns and find that the critical storage capacity α_c becomes maximal for a particular symmetric rule (α_c diverges in the sparce coding limit). Departures from symmetry decrease α_c but can increase the robustness of the model.

Key-words: Hebbian learning; Asymmetry; Biased patterns.

PACS: 87.10+e - 64.60Ht - 75.10Nr

Since the now classical Hopfield's proposal[1],work in neural networks has progressed continuously towards the description of more realistic models. In spite of the very interesting and rich behaviour of the original model, the Hopfield learning rule is biologically unplausible from several points of view. One of the main problems is the symmetry of the synapses. It is well known that a realistic learning rule must be asymmetric. Asymmetry has been introduced mainly through dilution of synapses[2,3]. In the limit of extreme dilution, the dynamics of several models could be solved exactly[4-6].

Another restriction of the original model was the 50% level of activity of the neurons. In cortex areas where associative functions are detected, representative levels of activity would be 4-5% [7]. Work on networks with low activity levels was motivated also by the found by Gardner[8] that, independently of any particular learning rule, optimal storage capacity can be obtained in the limit of very low activities (sparce coding). Consistently, straightforward modifications of the original Hopfield rule were introduced which took into account the possibility of storing biased patterns[9-12] (though improperly, quite often referred to as "correlated"); they presented, in the limit of sparce coding, storage capabilities which approach the ideal bound found by Gardner.

In the present paper we consider a class of Hebbian learning rules that incorporate several biologically appealing features

like possible structural asymmetry and biased patterns, and that recover, as particular cases, several well known models. These generalized learning rules operate in networks that contain N neurons $\{x_i\}_{i=1,\ldots,N}$ and store p binary independent patterns $\{\xi_i^\mu = \pm 1\}_{i=1,\ldots,N}^{\mu=1,\ldots,p}$ with mean activity $(1+a)/2$, chosen according to the probability distribution

$$P(\xi_i^\mu) = \frac{1+a}{2}\,\delta(\xi_i^\mu - 1) + \frac{1-a}{2}\,\delta(\xi_i^\mu + 1) \qquad (1)$$

where $a \in [-1,1]$ is the "bias" between the patterns ($a=0$ corresponds to the unbiased case, i.e., 50% of activity level).

If we assume that a new memorized pattern modifies a given synapse only through the activities of the pre-synaptic neuron j and post-synaptic neuron i, then the most general analytic form (without mixing of the patterns) for the variation of the synaptic strength is given by

$$\Delta J_{ij}(\xi_i^\mu\ ,\ \xi_j^\mu) = A + B\xi_i^\mu + C\xi_j^\mu + D\xi_i^\mu\xi_j^\mu$$

where A,B,C and D are arbitrary constants. If we further assume, as in the Hopfield model, that the learnt patterns add linearly, we obtain the following class of learning rules[13]:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}\left(A + B\xi_i^\mu + C\xi_j^\mu + D\xi_i^\mu\xi_j^\mu\right) \qquad (2)$$

The factor 1/N has been introduced in order to properly normalize the signal-to-noise analysis.

These rules are in general asymmetric provided B ≠ C, and the asymmetry is structural in the sense that it is not produced by a dilution process, but rather is an intrinsic property of the learning rule. By adjusting the values of the parameters A,B,C and D, we can recover various well known models. For instance, the case A=B=C=0 and D=1 corresponds to the Hopfield rule. The case A=B=C=D=1/4 corresponds to the original Hebb proposal: only those neurons that are simultaneously active on a pattern will contribute to the modification (reinforcement) of the corresponding synapse. If we consider patterns with bias a, the choice $A=a^2$, $B=C=-a$ and D=1 corresponds to the rule introduced by Amit,Gutfreund and Sompolinsky[9] for the storing of biased patterns.

All these particular rules are symmetric, and a priori one could expect that more general rules (with B ≠ C) would perform better than the symmetric ones. To test this hypothesis, we performed a signal-to-noise analysis searching for an optimal rule in the whole (A,B,C,D) parameter space, much in the spirit of the work of Dayan and Willshaw[14]. The analytical results were checked through computer simulations on a network of analog neurons.

The local field on neuron i will be assumed as follows:

$$h_i = \sum_{j=1}^{N} J_{ij} \, (x_j - b) - U \qquad (3)$$

The parameters b and U modify the neuronal thresholds and, as we shall see later on, they enable us to further optimize the storage capabilities of the network. The parameter U (an uniform external input on every neuron) permits the cancelation of perturbing terms on the signal part of the local field. Furthermore, as we will require a vanishing mean value for the noise, b will also become a suitable parameter. The introduction of these parameters has already been considered by Perez Vicente and Amit[10] in a symmetric learning rule.

Let us suppose that the network is in pattern $\xi^1 \equiv \{\xi_i^1\}$. Then

$$h_i = \frac{1}{N} \sum_{j=1}^{N} \sum_{\mu=1}^{P} (A + B\xi_i^\mu + C\xi_j^\mu + D\xi_i^\mu \xi_j^\mu)(\xi_j^1 - b) - U \qquad (4)$$

In analyzing the stability of the pattern we can decompose the local field into two parts:

$$h_i = S + R$$

The signal S will be produced by the selected pattern and will be given by

$$S = \frac{1}{N} \sum_{j=1}^{N} (A + B\xi_i^1 + C\xi_j^1 + D\xi_i^1 \xi_j^1)(\xi_j^1 - b) - U \qquad (5)$$

The noise R will be produced by the remaining (p-1) patterns and will be given by

$$R = \frac{1}{N} \sum_{j=1}^{N} \sum_{\mu=2}^{p} (A + B\xi_i^\mu + C\xi_j^\mu + D\xi_i^\mu\xi_j^\mu)(\xi_j^1 - b) \qquad (6)$$

Without loss of generality we can take D=1 (provided D≠0), and our task is now to optimize-the-signal to noise ratio ρ≡S/R in the (A,B,C,b,U,a) parameter space.

With a probability distribution for the patterns given by (1) we have, for the mean value of the signal,

$$\langle S \rangle = A(a-b) + C(1-ab) + [B(a-b) + D(1-ab)]\xi_i^1 - U \qquad (7)$$

where ξ_i^1 has been kept fixed as it represents the activity of the neuron in the pattern whose stability we are analyzing . The value of U which optimizes the signal clearly is:

$$U = A(a-b) + C(1-ab) \qquad (8)$$

The mean noise is:

$$\langle R \rangle = (a-b)[A(p-1) + aC(p-1) + (B+a) \sum_{\mu=2}^{p} \xi_i^\mu] \qquad (9)$$

It seems natural to impose $\langle R \rangle$=0 in order to maximize the stability of pattern ξ^1, so the optimum value for b is b=a. As mentioned before, the other (p-1) patterns generate a noise in the retrieval of ξ^1 and the parameter b enables us to cancel out its mean value. So we only have to consider now the effect of the variance of the noise for the case b=a, i.e.,

$$\langle R^2\rangle = \alpha(1-a^2)\{\, p\,[(A + a^2) + a(B+C)]^2 + B^2 + C^2 + 2a(aBC+B+C) + 1\} \qquad (10)$$

where $\alpha=p/N$ (with $p \to \infty$ and $N \to \infty$) is the storage capacity parameter.To assure, for arbitrary p, a small value for $\langle R^2\rangle$ the term in brackets proportional to p must vanish, hence

$$A + a^2 + a(B+C) = 0 \qquad (11)$$

The symmetric rule of ref.(10) cancels this term automatically. As we are a priori interested in more general (asymmetric) rules we will try to optimize $\langle R^2\rangle$ with

$$B = \gamma C \qquad (12)$$

for general values of γ (γ=1 recovers the symmetric case). From (11) we get

$$C = -\frac{a^2 + A}{a(\gamma + 1)} \qquad (13)$$

Replacing these values in (10) yields

$$\langle R^2\rangle = \alpha(1-a^2)[\, A^2(2+q/a^2) + 2A(3a^2-1+q) + a^2q + 2a^4 + (a^2-1)^2] \qquad (14)$$

with

$$q = \frac{(\gamma^2+1)(1-3a^2) - 4\gamma a^2}{(\gamma-1)^2} \qquad (15)$$

For general values of the bias *a* we have to maximize the signal-to-noise ratio as a function of two variables: A and the asymmetry γ. As the optimized signal depends neither on A nor on γ, we only have to minimize the mean square of the noise. It presents an absolute minimum at $\gamma=1$ and $A=a^2$. So, from the great number of rules defined by the parameters A,B,C and D, the optimal one (in the sense that it maximizes the signal-to-noise ratio) is the symmetric rule $A=a^2$, $B=C=-a$, $D=1$ studied in ref.(10).

When the magnitude of the noise becomes comparable with that of the signal ($\rho\cong1$), the patterns become destabilized and we can consequently estimate the critical storage capacity α_c by means of eqs.(7) and (14) by setting $\rho=1$. In the case $\gamma=1$ and $a=0$ (Hopfield rule) we obtain $\rho=\alpha^{-1/2}$ hence $\alpha_c = 1$.

In figure 1 we show, for several values of the bias *a*, α_c as a function of the asymmetry parameter γ, obtained from the signal-to-noise analysis by setting $\rho=1$ with the optimal condition $A=a^2$. In the sparce coding limit ($a\to1$), α_c diverges at the optimal value $\gamma=1$. This solution is definitively not robust against asymmetry; indeed, a weak assymetry is sufficient for the collapse of the retrieval capabilities of the system as can be seen in fig.1(a). We can see in figs.1 (b)-(d) that, as *a* decreases, asymmetric rules perform, in what concerns robustness, gradually better except in the vicinity of $\gamma=-1$ where a minimum $\alpha_c=0$ appears. In the limit $a\to0$, the critical storage capacity $\alpha_c=1$ independently from γ. We can see, through the figures, that the

minimum at $\gamma=-1$ is always present and causes a non-uniform convergence to the value $\alpha_c=1$ in the limit $a\to 0$.

As a signal-to-noise analysis is useful only as a first approximation to the behaviour of a model, we have checked our results through computer simulations in a network composed of N=400 analog neurons whose deterministic dynamics is defined by the following set of coupled maps:

$$x_i(t+1) = \tanh\{g(\sum_{j=1}^{N} J_{ij}(x_j(t)-b) - U)\}, \qquad (i=1,\ldots,N) \qquad (16)$$

where $x_i(t)$ represents the firing rate of neuron i at time t and can take continuous values between -1 and 1, both extremes representing a neuron at rest and firing at maximum rate respectively. The gain parameter g of the transfer function measures its degree of nonlinearity. In the limit $g \to \infty$ we recover a system with two discrete states neurons; in the limit $g \to 0$ we can expand the tanh to first order in g and the only stable state will be that with all neurons in the $x_i=0$ state (the system will consequently loose all its memory properties). J_{ij} is the synaptic matrix defined by Eqs.(1) and (2). The parameters b and U were fixed to the optimum values as discussed in the text above. At each time step all neurons were updated sincronously. Although this type of updating may present limit cycles in certain regions of the parameter space (a problem that is absent in sequential updating), we made this choice because sincronous dynamics is more

suited for performing simulations in massively parallel computing devices. In any case, neither the parallel dynamics nor the sequential one are fair approximations to the dynamics of a biological neural network. In order to test the storage capacities of these models we began the simulations from one of the (p) stored patterns and let it evolve according to the dynamics defined by eq.(16). When the system reached a fixed point, we compute and store the final overlap with the pattern. We repeat this process for half of the stored patterns in each of 40 different realizations of them (a total of 20p runs). We then compute the final "mean" overlap from all the runs that had reached a fixed point. We considered the system in a retrieval state (for a particular value of α) when the final overlap was greater than 95% (conventional value) and the number of cycles was less than 5% of the total number of runs; α_c was defined as the value of α at which any one of these conditions was not fulfilled. For g=100 we perturbed around the optimal rule found above, looking for a direction in the B-C plane for which some rule could have a critical storage capacity greater than that of the symmetric case for typical values of the bias a. We did not find such rules, thus confirming the predictions of the signal-to-noise analysis. We also tested the effect of the asymmetry parameter γ (keeping all other parameters at their optimal values) for the case a=0.2 . The results can be seen in fig.(2) for two values of g, namely g=100 and g=2.5. We note that for g=100 where the model becomes almost a discrete one the

saturation value of α_c for $\gamma \geq 1$ is close to that ($\alpha_c \cong 0.14$) of the Hopfield model. This is invariance with the prediction ($\alpha_c \cong 1$) of the signal-to-noise analysis. This numerical discrepancy is not surprising since, as discussed above, the signal-to-noise analysis gives only a qualitative description. The effect of lowering the gain is basically a reduction in the critical storage capacity. It is worthy to stress that, for arbitrary values for a and g, α_c vanishes for $\gamma=-1$ (extreme asymmetric case).

Returning to the general analysis we also note that the results found by Peretto[13] for a=0 can be greatly improved through the introduction of an optimizing field U=C. Nevertheless, for this case, the maximum α_c is still that of the Hopfield model, as found through our general analysis.

To conclude, let us point out that, although the Gardner analysis in the space of interactions admits networks with asymmetric synapses capable of excellent storage capacity, only symmetric rules were proposed up to now. We found that the asymmetric rules considered here, though not optimal, are more robust for not too low levels of activity and can accomodate much more patterns of simultaneous activities between pre-synaptic and post-synaptic neurons. As we restricted the present analysis mainly to the effect of asymmetry on the storage properties of the models, it would be interesting to investigate, for this type of rules, the influence of γ (and the rest of the present parameters) on other quantities of interest such as the information content,

the retrieval times, the size of the basins of attraction of the memories and the phase diagram.

The model can be further generalized by considering terms that mix the memories in the synaptic matrix (e.g. terms of the form $\xi_i^\mu \xi_j^\nu$ for arbitrary μ and ν). The particular cyclic case $\xi_i^\mu \xi_j^{\mu+1}$ has already been considered in relation to the retrieval of temporal sequences of patterns[7]. Finally, in the case A=B=C=U=b=0, the deterministic model with a gain g can be mapped into a stochastic one with a parameter β that plays the role of an inverse temperature[15]. It would be interesting to verify if this equivalence is preserved for the general model.

ACKNOWLEDGMENTS

We acknowledge useful remarks from E.M.F.Curado and F.A.Tamarit. We also thank the LAFEX for permiting us the use of their ACP II parallel processor, and very especially Carla Barros for computational assistance.

FIGURE CAPTIONS

Fig.1- Critical storage capacity α_c vs. asymmetry parameter γ from a signal-to-noise analysis (γ=1 corresponds to symmetric rules): (a) a=1; (b) a=0.6; (c) a=0.2; (d) a=0.05 .

Fig.2- Critical storage capacity α_c vs. asymmetry parameter γ from computer simulations on a network of 400 analog neurons for a=0.2. The empty squares (full triangles) correspond to a gain value g=100 (g=2.5) .

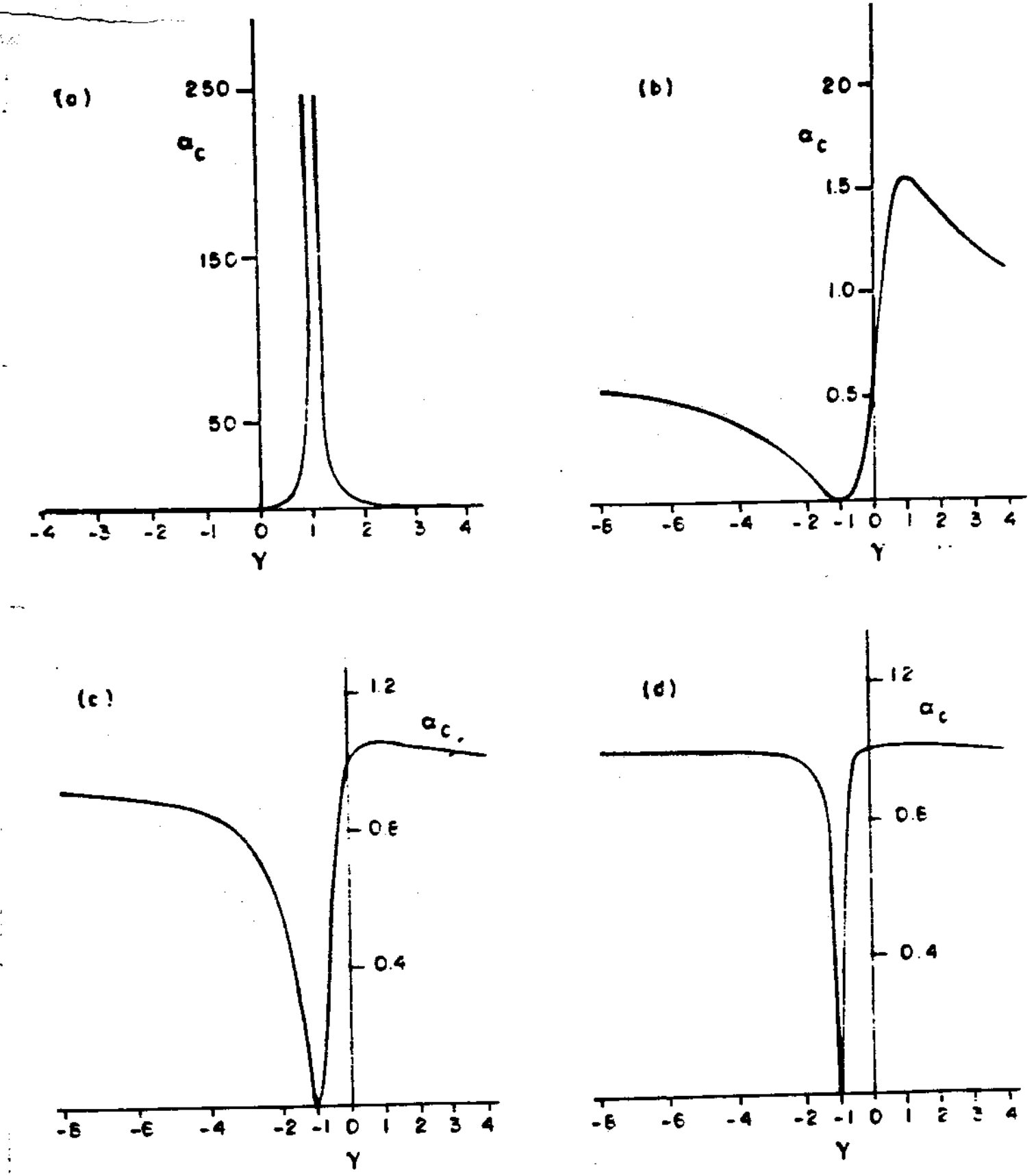
(a)
250
150
50
a_c
-4 -3 -2 -1 0 1 2 3 4
Y
(b)
20
1.5
1.0
0.5
a_c
-8 -6 -4 -2 -1 0 1 2 3 4
Y
(c)
1.2
0.8
0.4
a_c
-8 -6 -4 -2 -1 0 1 2 3 4
Y
(d)
1.2
0.8
0.4
a_c
-8 -6 -4 -2 -1 0 1 2 3 4
Y

FIG. 1

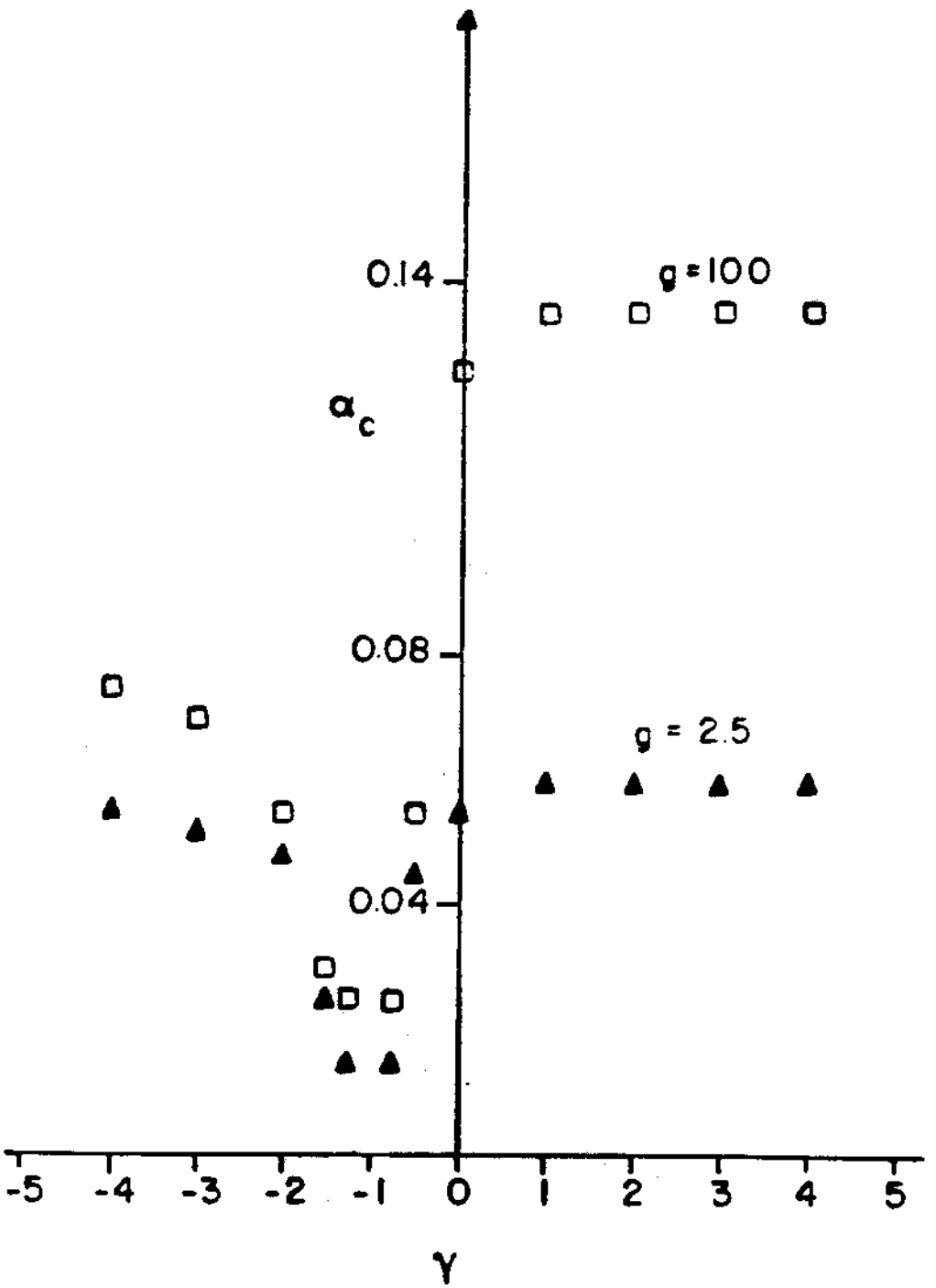

FIG. 2

REFERENCES

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

2) D.A.Stariolo, Phys.Lett. A152, 349(1991).

3) F.A.Tamarit, D.A.Stariolo and E.M.F.Curado, Phys.Rev.A43, 7083(1991).

4) B.Derrida, E.Gardner and A.Zippelius, Europhys.Lett. 4, 167(1987).

5) R.Kree and A.Zippeluis, Phys.Rev., A36, 4421(1987).

6) B.Tirozzi and M.V.Tsodyks, Europhys.Lett., 14, 727(1991).

7) D.Amit, *Modelling Brain Function*, (Cambridge University Press, 1989).

8) E.Gardner, J.Phys., A21, 257(1988).

9) D.J.Amit, H.Gutfreund and H.Sompolinsky, Phys.Rev., A35, 2293(1987).

10) C.J.Perez-Vicente and D.J.Amit, J.Phys., A22, 559(1989).

11) J.Buhmann, R.Divko and K.Schulten, Phys.Rev., A39, 2689(1989).

12) M.V.Tsodyks and M.V.Feigel'man, Europhyss.Lett., 6, 101(1988).

13) P.Peretto, J.Phys.France, 49, 711(1988).

14) P.Dayan and D.J.Willshaw, Biol.Cybern., 65, 253(1991).

15) C.M.Marcus, F.R.Waugh and R.M.Westervelt, Phys.Rev.A41, 3355(1990).