Retrieval dynamics of neural networks constructed from local and nonlocal learning rules

J. Krätzschmar, G.A. Kohring

J. Phys. France **51** (1990) 223-229 1er FÉVRIER **1990,** PAGE 223

Classification
Physics Abstracts
87.30G — 06.50

Retrieval dynamics of neural networks constructed from local and nonlocal learning rules

J. Krätzschmar and G. A. Kohring

Universität Bonn, Physikalisches Institut, Nussallee 12, D-5300 Bonn I, F.R.G.

(*Reçu le 7 avril 1989, révisé le 18 septembre 1989, accepté le 20 octobre 1989*)

Résumé. — Les dynamiques de rappel de réseaux neuronaux construits à partir de règles d'apprentissage locales et non locales sont comparées à l'aide de simulations numériques et apparaissent très similaires. Nos simulations indiquent qu'il ne semble pas possible de déterminer le comportement des systèmes aux temps longs en fonction de la dynamique aux premiers instants, comme cela avait été récemment suggéré par Kepler et Abbott.

Abstract. — The retrieval dynamics of neural networks constructed from local and nonlocal learning rules are compared *via* computer simulations and shown to be very similar. Furthermore, the simulations show no hope for determining the long time behavior of the system in terms of the first step dynamics as has been advocated by Kepler and Abbott.

Introduction.

The thermodynamic properties of the original Little-Hopfield model have been studied in detail by Amit, Gutfreund, Sompolinsky and their co-workers [1]. They showed that a strongly connected network having N nodes, or Ising-spins, behaves as an associative memory device provided the number of patterns, P, scales with N as : $\alpha \equiv \lim_{N \to \infty} P/N \leqslant 0.14$. The relatively low storage capacity of the model prompted Personnaz *et al.* [2] to invent the nonlocal, or pseudo-inverse, learning rule for determining the couplings, J_{ij}, between the spins. Such a model has a maximum storage capacity of $\alpha = 1$. Vekentash [3], however, showed that the maximum storage capacity for any model with only two-spin interactions is $\alpha = 2$. This prompted Gardner [4] to find a local learning rule which could reach this maximum storage capacity and in doing so she developed the analytical tools needed for completely determining the static properties of networks having asymmetric couplings.

Static properties alone are not sufficient for understanding neural networks because their ultimate use is as dynamical attractors. In that respect it has been argued [5] that networks constructed from nonlocal learning rules should have better dynamical properties than those constructed from local learning rules. However, extensive numerical simulations by Forrest [6] and Kanter and Sompolinsky [2] at two values of α do not bear that out. An approximate analytic treatment [7] also indicates that at least in the neighborhood of $\alpha = 1$, the two learning rules should have similar dynamical behavior. Furthermore, it has been shown by

Diederich and Opper [8] that some local learning rules can in fact reproduce exactly the coupling matrix of the pseudo-inverse.

In this paper then, we make explicit comparison between these two types of networks using as a measure of the dynamical response, the radius of attraction, R. R is defined as one minus the minimum overlap, $m_{\min}$, with a stored state such that as $N \to \infty$ *almost all* of the states having $m > m_{\min}$ ($m \equiv (1/N) \sum_j \xi_j S_j$, where ξ represents the stored state and S an arbitrary state) will evolve towards the stored state. Previously Kepler and Abbott [9] gave a phenomenological rule for calculating this quantity, but the simulations to be presented here will show that care is required when using their rule. In the next section the numerical procedures used will be explained and an interpretation is given in the last section.

Numerical methods.

Local learning rules are designed to produce a coupling matrix, J_{ij}, such that the following stability condition is fulfilled for every site of every pattern :

$$\xi_i^\mu \sum_{j \neq i} J_{ij} \xi_j^\mu \geqslant \kappa \sqrt{\sum_{\ell \neq i} (J_{i\ell})^2} \,. \tag{2.1}$$

The learning rules proposed by Krauth and Mèzard [10], Forrest [6] and Gardner [4] can be shown to produce equivalent networks at saturation, i.e., when the network has the maximum storage capacity allowed for a given κ. In the case of random, uncorrelated patterns, α is related to κ at saturation by the Gardner formula [4] :

$$\alpha = \frac{1}{\int_{-\kappa}^{\infty} \frac{dt}{\sqrt{2\pi}} e^{-t^2/2} (t + \kappa)^2} \,. \tag{2.2}$$

To construct a network at saturation we used the approach advocated by Gardner [4]. Starting with pattern 1 and going sequentially through all of the patterns to be stored, the coupling matrix is changed by :

$$\Delta J_{ij} = \Theta\left(\xi_i^\mu \sum_{j \neq i} J_{ij} \xi_j^\mu - \kappa \sqrt{\sum_{\ell \neq i} (J_{i\ell})^2} \right) \xi_i^\mu \xi_j^\mu \,. \tag{2.3}$$

After all the patterns have been checked, the procedure is repeated until $\Delta J_{ij} = 0$ for every pattern, μ and site, i. For a finite size network, the system saturates much sooner than would be expected by equation (2.2). In order to correct for this, we note that the learning time on an infinite system should diverge as $\kappa \to \kappa_s(\alpha)$, where κ_s is the saturation value of κ [6, 11]. Hence, we looked for a similar finite size, κ_{fs}, at which the learning time tends to infinity and chose for the simulations a value of κ slightly smaller so as to produce reasonable learning times. This procedure produces a network with properties very similar to that of a fully saturated, infinite volume network. In fact, a calculation of the average local alignment, $\left\langle \xi_i^\mu \sum_{j \neq i} J_{ij} \xi_j^\mu \right\rangle_\xi$, gives a value on a 100 node network within 0.5 % of that expected for an infinite size system at saturation [9].

Systems constructed from these local learning rules will be contrasted with those constructed from nonlocal learning rules. One reason similar behavior is expected from these two networks is the proof by Diederich and Opper [8] which showed that by changing the

stability condition in (2.1) to, $\xi_i^\mu \sum_{j \neq i} J_{ij} \xi_j^\mu = 1$, could reproduce the coupling matrix structure of the pseudo-inverse method of Personnaz *et al.*, as $N \to \infty$ and in the limit of an infinite number of learning steps. The pseudo-inverse method defines the synaptic couplings as :

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu (C^{-1})_{\mu\nu} \xi_j^\nu \tag{2.4}$$

where the correlation matrix is defined as : $C_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$. This method has the advantage that no refinements are necessary for finite size systems.

After preparing a network using one of these prescriptions, the system was initialized to some state having an overlap, m_0, with one of the stored patterns. Then the network was updated in parallel according to the zero-temperature, deterministic rule (Glauber Dynamics) :

$$S_i(t+1) = \text{sgn}\left(S_i(t) \sum_{j \neq i} J_{ij} S_j(t)\right) . \tag{2.5}$$

Biologically, this rule is not correct ; however, a naive comparison of time scales indicates that real systems should update more in parallel than in serial, hence, there has been a great amount of attention paid to this updating scheme [9, 12, 7, 13]. Other updating rules will give different quantitative results for the radius of attraction, but the comparative results should be the same.

The updating rule is then applied until the system reaches a fixed point or a cycle. Thus, the percentage of times the pattern μ is recalled as a function of m_0^μ can be calculated. Repeating this procedure for many different values of m_0 and using the definition given above, enables one to calculate the radius of attraction. This radius, however, must be renormalized, since, as Kantor and Sompolinsky [2] have pointed out, for finite size systems the radius is reduced due to the $O(1/\sqrt{N})$ overlap of every stored state with every other stored state. Hence, the actual definition in use is : $R = \left\langle\!\!\left\langle \frac{1 - m_{\min}}{1 - m_{\text{av}}} \right\rangle\!\!\right\rangle$, where m_{av} is the average overlap with the other stored states and the brackets indicate an average over all stored states and all possible sets of patterns.

We have also checked in the course of these simulations the proposed phenomenological rule of Kepler and Abbott [9] :

$$P(S_0 \to \xi) = \Theta\left(\frac{m_1 - m_0}{1 - m_0} - \chi\right) , \tag{2.6}$$

where, m_1 is the overlap of the system at the first time step and $P(S_0 \to \xi)$ is the probability that initial system is mapped onto the stored state. Kepler and Abbott suggested that χ has the value 1/2.

For the simulation results to be presented here, an $N = 100$ node network was used and typically 75 different sets of patterns were randomly chosen to stabilize for each distinct value of α. For each set of patterns, the radius of attraction was measured for each pattern as indicated above using typically 100 different initial states.

Results and discussion.

The main results of this paper are presented in figure 1, where it can clearly be seen that local and nonlocal learning rules have similar dynamical behavior as manifested by the similarity of their radius of attraction. At $\alpha = 1$, R_{nonlocal} must tend to zero [2], however, it is also clear that R_{local} is becomming very small. Hence, even though the local learning rule has a storage capacity of $\alpha = 2$, the systems ability to function as an associative memory device is already seriously impaired at $\alpha = 1$. These results confirm the predictions of Krauth *et al.* [13] who argued on the basis of a one-pattern-neural-network that the stability of the patterns is the most important variable for determining the radius of attraction. Indeed, for the nonlocal learning the stability, Δ, is [2] : $\Delta = \sqrt{(1 - \alpha)/\alpha}$, while for the local learning rule we have

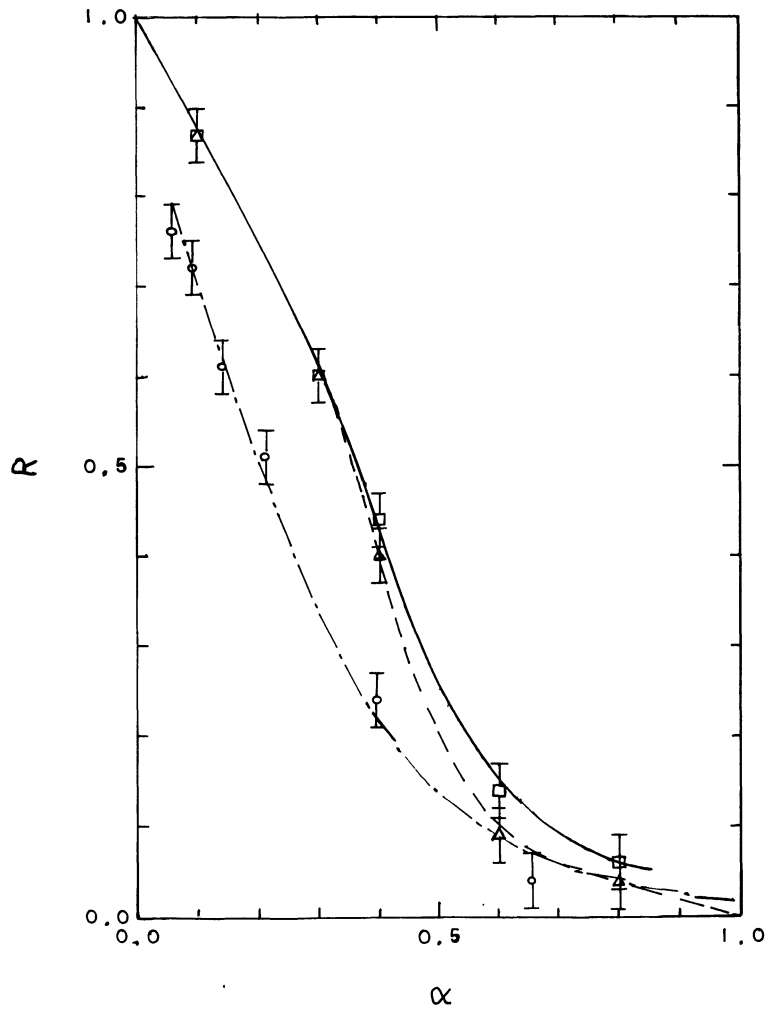

Fig. 1. — The radius of attraction, R *vs.* α. (□) For the local learning rule, (△) for the nonlocal learning rule, (○) using the Kepler and Abbott phenomenological rule. The broken line is the analytical realization of the Kepler-Abbott rule. The solid and dotted lines are guides to the eye.

[4]: $\alpha = \left[\left(\frac{\kappa^2+1}{2} \right) \text{erfc}\,(-\kappa/\sqrt{2}) + (\kappa/\sqrt{2})\,e^{-\kappa^2/2} \right]^{-1}$, and as $\kappa \to \infty$, $\alpha \to 1/(\kappa^2+1)$, $\Rightarrow \kappa = \sqrt{(1-\alpha)/\alpha}$. Hence, these results confirm Krauth *et al.*'s extropolation from a one-pattern-network to a full network. Furthermore, since the local learning rule produces couplings which are symmetric only in the mean (i.e., the quantity, $y = T_{ij} - T_{ji}$ is Gaussian distributed with zero mean), while the nonlocal learning produces completely symmetric couplings ; these results indicate the lesser importance of the asymmetry of the couplings (at least when such asymmetry is not too large).

Figure 1 also shows the radius of attraction as predicted by the phenomenological rule of Kepler and Abbott along with the data from their paper [9]. The disagreement between their results and the present results can be understood by looking at figure 2a. Here the probability that $P(S_0 \to \xi)$ as a function of $\frac{m_1 - m_0}{1 - m_0}$ is plotted for $\alpha = 0.10$, $\kappa = 2.5$ and two different values of m_0. Although simulations at $m_0 = 0.15$ give results in agreement with their proposed rule, results for $m_0 = 0.50$ do not. Clearly, a universal value for the adjustable parameter χ cannot be determined from this data because its dependence upon m_0 is not clear at this finite value of N. Furthermore, the results in figure 2b for $\alpha = 0.40$ and $\kappa = 0.92$ indicate that the value of χ may also have a κ dependence. In figure 3 additional evidence of this is shown at $\alpha = 0.4$, $\kappa = 0.92$, where $P(S_0 \to \xi)$ is plotted as a function of m_0 using the normal procedure and the Kepler and Abbott rule. The former predicts the correct radius of

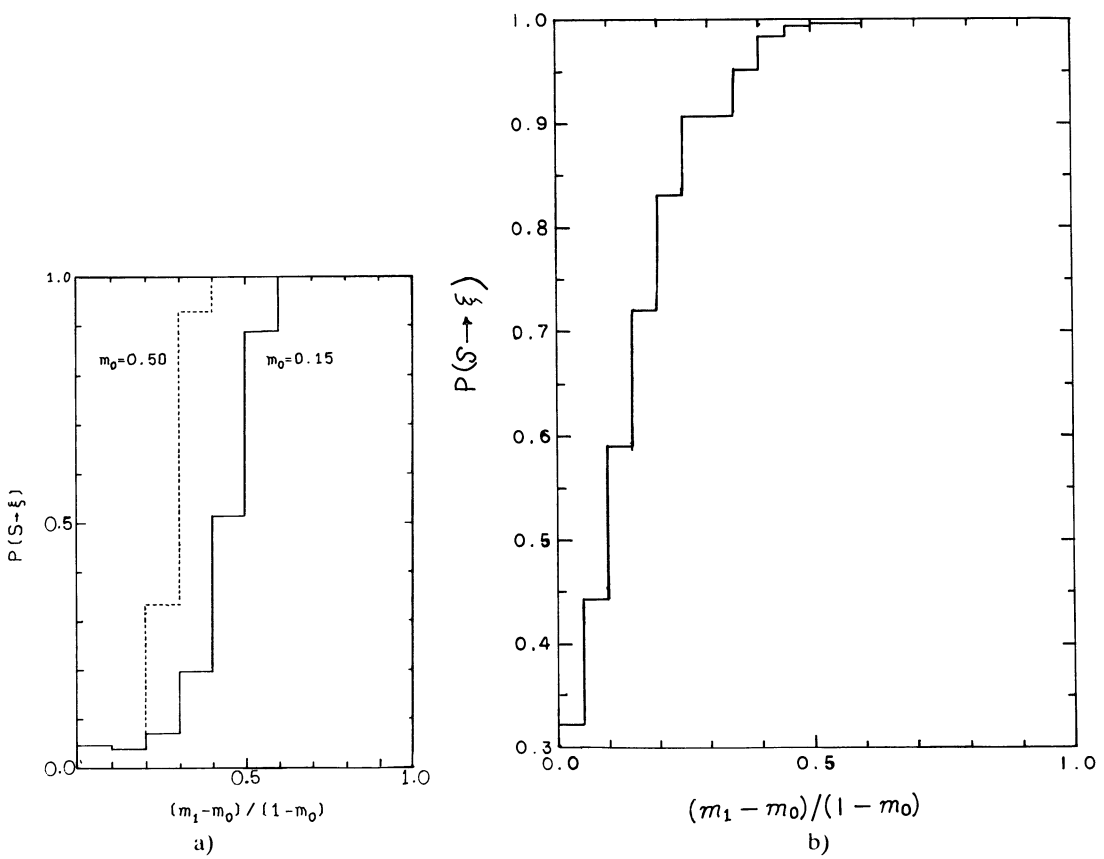

Fig. 2. — $P(S \to \xi)$ as a function of $(m_1 - m_0)/(1 - m_0)$ for : a) $\alpha = 0.10$, $\kappa = 2.50$ and $m_0 = 0.15$, and 0.5, and b) $\alpha = 0.40$, $\kappa = 0.92$ and $m_0 = 0.66$.

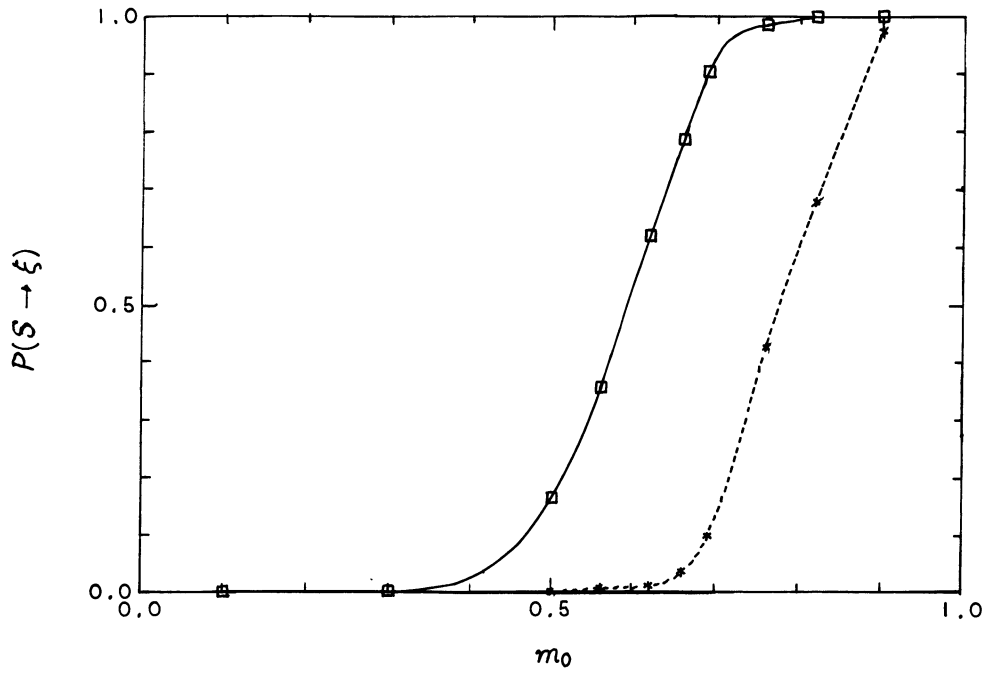

Fig. 3. — $P(S \rightarrow \xi)$ *vs.* m_0. The full line is for the normal procedure while the dotted line is for the Kepler and Abbott procedure. Both are for $\alpha = 0.40$ and $\kappa = 0.92$.

attraction, while the latter predicts a radius of attraction consistent with the results given by [9] and indicated in figure 1. The present results shown in figure 1 could in fact be accounted for by using $\chi = 0.3$ [14]. However, the results in figures 2 and 3 would still be difficult to interpret as supporting any such universal value of χ. Given this uncertainty for the correct value of χ, if any, prudence would dictate that any results should be based upon the more reliable method for computing the radius of attraction, namely the standard method as described in section 2 and used by other authors (see for example Kanter and Sompolinsky in [2]) which uses only the initial overlap m_0 as its dependent variable.

In summary then, the similarity of the dynamical behavior between networks constructed from local and nonlocal learning rules has been demonstrated and implies no advantage, from this viewpoint, of either network type in actual applications. Furthermore, it has been shown that attempting a description of the long time behavior of these systems in terms of only the first step dynamics may lead to spurious results.

References

[1] AMIT D., GUTFREUND H. and SOMPOLINSKY H., *Phys. Rev. Lett.* **55** (1985) 1530 ; *Ann. Phys.* **173** (1987) 30 ;
CRISANTI A., AMIT D. J. and GUTFREUND H., *Europhys. Lett.* **2** (1986) 337.

[2] PERSONNAZ L., GUYON I. and DREYFUS G., *Phys. Rev.* **A 34** (1986) 4217 ;
KANTER I. and SOMPOLINSKY H., *Phys. Rev.* **A 35** (1987) 380.

[3] VENKATESH S. S., *Neural Networks for Computing,* Ed. J. S. Denker (American Inst. of Phys., New York) 1986.
[4] GARDNER E., *Europhys. Lett.* **4** (1987) 481 ; *J. Phys.* **A 21** (1988) 257 ; GARDNER E. and DERRIDA B., *J. Phys.* **A 21** (1988) 271.
[5] AMIT D. J. (Racah Inst. of Phys., Preprint), RI/87/19 (to be published).
[6] FORREST B. M., *J. Phys.* **A 21** (1988) 245.
[7] KOHRING G. A., *Europhys. Lett.* **8** (1989) 697.
[8] DIEDERICH S. and OPPER M., *Phys. Rev. Lett.* **58** (1987) 949.
[9] KEPLER T. B. and ABBOTT L. F., *J. Phys. France* **49** (1988) 1657.
[10] KRAUTH W. and MÈZARD M., *J. Phys.* **A 20** (1987) L745.
[11] OPPER M., *Phys. Rev.* **A 38** (1988) 3824.
[12] GARDNER E., DERRIDA B. and MOTTISHAW M., *J. Phys. France* **48** (1987) 741 ; KRAUTH W., NADAL J.-P. and MÈZARD M., *Complex Syst.* **2** (1988) 387.
[13] KRAUTH W., NADAL J.-P. and MÈZARD M., *J. Phys.* **A 21** (1988) 2995.
[14] ABBOTT L. F. and KEPLER T. B., *private communication.*