Journal of Statistical Physics, Vol. 49, Nos. 5/6, 1987

Learning and Forgetting on Asymmetric, Diluted Neural Networks

B. Derrida[1] and J. P. Nadal[2]

Received May 18, 1987

It is possible to construct diluted asymmetric models of neural networks for which the dynamics can be calculated exactly. We test several learning schemes, in particular, models for which the values of the synapses remain bounded and depend on the history. Our analytical results on the relative efficiencies of the various learning schemes are qualitatively similar to the corresponding ones obtained numerically on fully connected symmetric networks.

KEY WORDS: Dynamics; neural networks.

1. INTRODUCTION

For the last few years a great deal of work, numerical and analytical, has been done on the Little–Hopfield model[1,2] of neural networks and on generalizations of it. Analytical work on the equilibrium properties were made possible using two basic simplifications: the synaptic connections were taken as symmetric, and each neuron was connected to every other neuron. The thermal properties were then computed using the replica method.[3] For this model several attempts have been done to analyze the effect of dilution and/or asymmetry.[4-9]

Recently a diluted, asymmetric version of the Little–Hopfield model has been introduced.[10] For this model, the dynamics can be solved exactly. Hence, it is tempting to look at the properties of the Little–Hopfield model and its generalizations for this diluted, asymmetric architecture in order to determine in an exactly soluble case the effect of the relevant parameters. It turns out that models that are difficult to study analytically for the fully connected network, with symmetric interactions,

[1] SPT, CEN-Saclay, 91191 Gif sur Yvette, France.
[2] GPS, ENS, 75231 Paris, France.

0022-4715/87/1200-0993$05.00/0

can be solved exactly on this diluted, asymmetric network. This is the motivation of the present paper, which deals with learning schemes where forgetting occurs: we solve here for the dynamical properties of the diluted, asymmetric network with learning schemes leading to short-term memory effects[11–13] or long-term memory effects.[14]

In Section 2 we give the general framework of dilute, asymmetric models and the equations that give the storage capacity of the network for any given learning scheme. In Section 3 we recall the definition of the learning schemes in which we are interested and introduce the quantities that measure the number of stored and memorized patterns. In Section 4 we give the solution for the diluted, asymmetric network with these schemes. We always find a phase diagram very similar to the ones obtained by numerical or analytical calculations for the fully connected symmetric network. In all cases, the problem can be reduced to the study of a one-dimensional random walk with constraints depending on the learning rule.

2. THE DILUTED, ASYMMETRIC NETWORK

We consider the general framework introduced in Ref. 10, that is, we work with a system of N Ising spins $\sigma_i = \pm 1$ with a very low connectivity. The interactions J_{ij} are given by

$$J_{ij} = C_{ij} T_{ij} \tag{1}$$

where C_{ij} is chosen (independently of C_{ji}) at random according to the distribution

$$\rho(C_{ij}) = \frac{C}{N} \delta(C_{ij} - 1) + \left(1 - \frac{C}{N}\right) \delta(C_{ij}) \tag{2}$$

and T_{ij} is a matrix that depends on the stored patterns: a given learning scheme is characterized by a given prescription for fixing the T_{ij}. We will limit ourselves to the large-C limit, with $N \to \infty$ first (the solution is valid if $C \ll \log N$).[10]

The dynamics is defined by the following updating rule. If spin i is updated at time t, this means that

$$\begin{aligned} \sigma_i(t) &= +1 \qquad \text{with probability } (1 + e^{-2\beta h_i})^{-1} \\ \sigma_i(t) &= -1 \qquad \text{with probability } (1 + e^{2\beta h_i})^{-1} \end{aligned} \tag{3}$$

where

$$h_i = \sum_j J_{ij} \sigma_j(t) \tag{4}$$

Since the interactions J_{ij} are not symmetric, there is no Hamiltonian, no partition function, and the temperature β^{-1} is only defined through the updating rule.

The retrieval quality $m_\mu(t)$ of a learned pattern ξ_i^μ is defined by

$$m_\mu(t) = 1/N \sum_i \xi_i^\mu \sigma_i(t) \tag{5}$$

As shown in Ref. 10, the evolution of $m_\mu(t)$ is given, for parallel dynamics (all sites are updated at each time step), by

$$m_\mu(t+1) = f_\mu(m_\mu(t)) \tag{6}$$

and for random sequential updating (a randomly chosen site is updated during the time interval $dt = 1/N$) by

$$dm_\mu(t)/dt = f_\mu(m_\mu(t)) - m_\mu(t) \tag{7}$$

where $f_\mu(m)$, in the limit $C \to \infty$, is given by

$$f_\mu(m) = \int_{-\infty}^{\infty} dz/(2\pi)^{1/2}\, e^{-z^2/2} \tanh[\beta C A_\mu (m + z\Delta_\mu)] \tag{8}$$

with

$$\Delta_\mu = [(D_\mu - A_\mu{}^2)/C A_\mu{}^2]^{1/2} \tag{9}$$

where A_μ and D_μ are the following averages, on the randomly stored patterns:

$$A_\mu = \overline{\xi^\mu{}_i \xi^\mu{}_j T_{ij}} \tag{10}$$

$$D_\mu = \overline{[\xi^\mu{}_i \xi^\mu{}_j T_{ij}]^2} = \overline{T_{ij}^2} \tag{11}$$

One can notice that by making a gauge transformation such that $\xi^\mu{}_i = +1$ for all i, A_μ is given by $A_\mu = \overline{T_{ij}}$.

The critical temperature $1/\beta^*$ is given by $(df_\mu/dm)(0) = 1$ and therefore is a solution of

$$\beta^* C A_\mu \int_{-\infty}^{\infty} dz/(2\pi)^{1/2}\, e^{-z^2/2} \cosh^{-2}(\beta^* C A_\mu \Delta_\mu z) = 1 \tag{12}$$

and the transition is of second order.

At zero temperature, the stationary value m_μ, i.e., the value of $m_\mu(t)$ in the limit $t \to \infty$, is a solution of

$$m_\mu = (2/\pi)^{1/2} \int_0^{m_\mu/\Delta_\mu} dz\, e^{-z^2/2} \tag{13}$$

and m_μ is nonzero if

$$\Delta_\mu < (2/\pi)^{1/2} \tag{14}$$

For a general learning scheme, the averages A_μ and D_μ depend on μ, and criterion (14) gives which patterns are memorized. If one wants the patterns retrieved with a quality at least equal to a given value M, one has the criterion

$$\Delta_\mu < M/X \tag{15}$$

where X is defined by

$$M = (2/\pi)^{1/2} \int_0^X dz\, e^{-z^2/2} \tag{16}$$

Hence, for any given learning scheme, the storage properties of the system are obtained by the computation of the mean A_μ and mean square D_μ of the synaptic efficacies (this means also that two different learning schemes, leading to the same values of A_μ and D_μ have the same storage properties). This would not be true for the fully connected network, where the result depends also on the correlations between the J_{ij} (such as $J_{ij}J_{jk}J_{kl}$). One can note, however, that the qualitative results obtained by a signal-to-noise analysis provides a criterion similar to (14) [but with an unknown parameter instead of $(2/\pi)^{1/2}$]. Such a criterion leads to a continuous transition, whereas in the fully connected network, one has a discontinuous transition. Apart from this crucial difference, the scaling properties with the *connectivity* are (qualitatively) the same for the diluted and nondiluted networks.

3. LEARNING AND FORGETTING SCHEMES

Simple modifications of the Hopfield scheme, keeping a Hebbian type learning rule, but leading to forgetting effects (instead of a complete deterioration of the memory due to overloading), have been studied either numerically or analytically.[1,11–13] These models are characterized by an iterative learning rule: once some configurations have been stored, the acquisition of a new configuration is obtained via a modification of the synaptic efficacies that depends only on the new configuration. Furthermore, the rule is local at the synaptic level, that is this modification depends only on the activities of the two neurons involved (presynaptic and postsynaptic neurons). We will consider here three of these models. Let us first recall the definition of these models, and then their main properties.

Although in the fully connected network the connectivity C is identical to the number of neurons N, having in mind the diluted network we study here, we will write C instead of N whenever appropriate.

Model A: Hopfield scheme. The learning of the pth pattern $(\xi^p{}_i)$ is obtained via

$$T_{ij}(p) = T_{ij}(p-1) + (1/C)\,\xi^p{}_i \xi^p{}_j \tag{17}$$

Model B: Marginalist scheme and weighted schemes.[11,12]

B1. Marginalist scheme. The learning of the pth pattern is

$$T_{ij}(p) = \lambda[T_{ij}(p-1) + (\varepsilon/C)\,\xi^p{}_i \xi^p{}_j] \tag{18}$$

with $\lambda = \exp(-\varepsilon^2/2C)$. If p_s is the total number of stored patterns, (18) leads to the formula

$$T_{ij}(p) = (\varepsilon/C) \sum_{\mu=1}^{p_s} e^{-\mu\varepsilon^2/2C} \xi_i^{p_s-\mu+1} \xi_j^{p_s-\mu+1} \tag{19}$$

B2. Weighted schemes. A straightforward generalization[12] of this model is the prescription

$$T_{ij}(p) = (1/C) \sum_{\mu=1}^{p_s} \Lambda(\mu/C)\,\xi_i^{p_s-\mu+1} \xi_j^{p_s-\mu+1} \tag{20}$$

where Λ is any given positive function with the appropriate normalization

$$(1/C) \sum_{\mu=1}^{p_s} \Lambda^2(\mu/C) = K \tag{21}$$

where K is a constant and does not depend on C. The thermodynamics of model B for the fully connected and symmetric network have been solved[12] in the same way as those of model A.[3]

Model C: Learning within bounds.[1,11,13] The synaptic efficacies are constrained between a lower and an upper bound, $-L \leqslant T_{ij} \leqslant L$:

$$T_{ij}(0) = 0$$

$$T_{ij}(p) = \begin{cases} T_{ij}(p-1) + (\varepsilon L/\sqrt{C})\,\xi^p{}_i \xi^p{}_j & \text{if no bound is reached} \\ +L \text{ or } -L & \text{otherwise} \end{cases} \tag{22}$$

Hence, each T_{ij} makes a random walk between two nonabsorbing walls. This model C has been studied numerically[1,11,13] for the fully connected symmetric network and has the very same qualitative behavior as the

marginalist scheme. Analytical solution of this model is difficult because the T_{ij} are correlated.

Models B and C lead qualitatively to the same short-term memory effects. The parameter ε characterizes the amplitude with which a pattern is stored. Two other parameters, g and α, are of interest to study the large-C limit:

$$g = p_s/C \tag{23}$$

where p_s is the total number of stored patterns since the beginning of the learning process, and the capacity α

$$\alpha = p_m/C \tag{24}$$

where p_m is the number of patterns that are effectively memorized.

For ε smaller than a critical value ε_c, these models present three regimes (Fig. 1): $g < g^*(\varepsilon)$; $g^*(\varepsilon) < g < g_c(\varepsilon)$; and $g > g_c(\varepsilon)$:

1. A good learning regime, $g < g^*(\varepsilon)$, where every stored pattern is well retrieved and therefore $p_m = p_s$ (i.e., $g = \alpha$).
2. When one increases the number of stored patterns g, one then finds a forgetting regime $g^* < g < g_c$ where only the most recent patterns are memorized and therefore $\alpha < g$.
3. Finally, above a critical value $g_c(\varepsilon)$, one reaches a complete deterioration regime where no stored pattern is memorized, $\alpha = 0$.

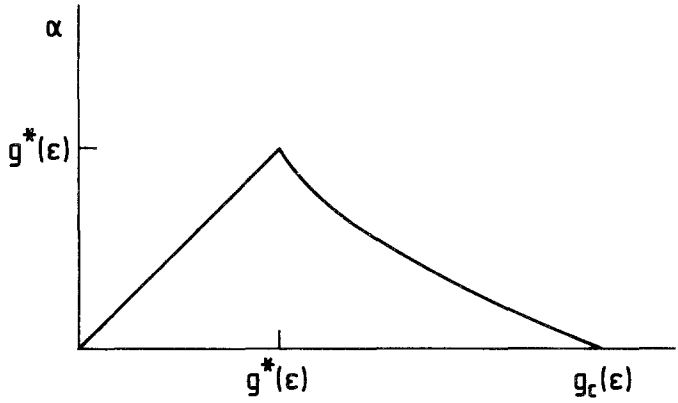

Fig. 1. For $\varepsilon < \varepsilon_c$, capacity α as a function of $g = p_s/C$, where p_s is the total number of stored patterns. For $g < g^*$, all patterns are memorized ($\alpha = g$). For $g^* < g < g_c$, only the strongest ones are memorized. For $g > g_c$, no pattern is memorized.

The Little–Hopfield model is recovered in the $\varepsilon \to 0$ limit. In that limit only the two extreme regimes can be observed and $g^* = g_c$ (≈ 0.14 for the fully connected, symmetric network, and $2/\pi$ for the diluted, asymmetric network).

For $\varepsilon > \varepsilon_c$, the complete deterioration regime never occurs ($g_c = \infty$) and for $g \to \infty$ the network reaches a stationary regime: the capacity α has a limit $\alpha_c(\varepsilon)$, which is the stationary number of memorized patterns (Fig. 2). When ε increases from ε_c, this stationary capacity $\alpha_c(\varepsilon)$ has a maximum α_{opt} at a certain value ε_{opt} (Fig. 3). All these functions $g^*(\varepsilon)$, $g_c(\varepsilon)$, $\alpha_c(\varepsilon)$ and ε_c, α_{opt}, ε_{opt} depend on the learning scheme and will be computed in the next section. Note that α_{opt} is always much smaller than the maximal capacity $g^*(0)$ of the Little–Hopfield model.

Model D: Learning within absorbing bounds. This model has been introduced and studied numerically Peretto.[14] The learning rule is identical to that of model C, except that now once a synaptic efficacy has reached a bound $+L$ or $-L$, it remains fixed at this value forever. This leads to long-term memory effects: it is now the oldest learned patterns that are memorized; clearly, when the number of stored patterns increases, the number of frozen synaptic efficacies increases, hence there is less and less plasticity to store the relevant information on the new patterns. This indicates, as confirmed numerically,[14] that one has similar diagrams (those of Figs. 1–3), where now α gives the number of first learned patterns that are still memorized.

In the next section we consider these models on the diluted, asymmetric network.

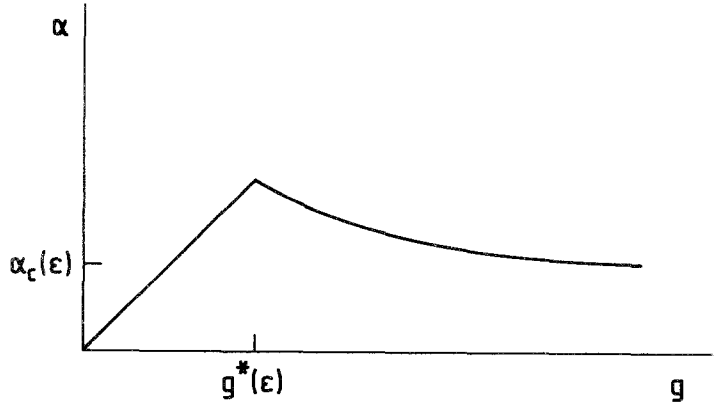

Fig. 2. The same as Fig. 1, for $\varepsilon > \varepsilon_c$: g_c is infinite, and α has a nonzero limit $\alpha_c(\varepsilon)$ for $g \to \infty$.

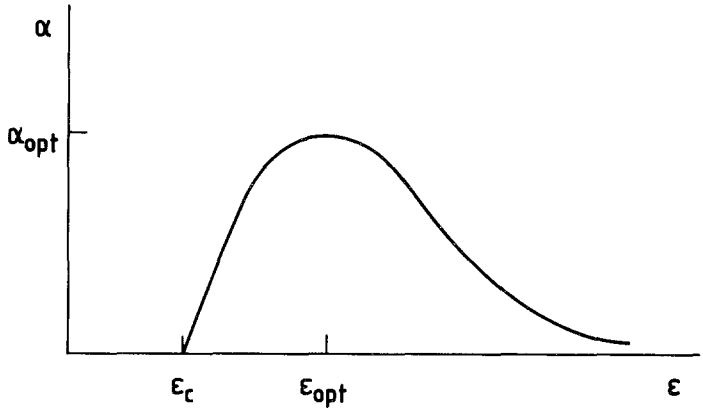

Fig. 3. Asymptotic capacity $\alpha_c(\varepsilon)$ as a function of ε.

4. RANDOM WALK SOLUTIONS

4.1. General Solution

We now consider a diluted, asymmetric network as defined in Section 2. The synaptic efficacies are given by (1), and we take the T_{ij} as given by the learning schemes of Section 2. We restrict ourselves to the equilibrium (i.e., the infinite-time limit) properties, that is, we consider the storing capacities as given by Eqs. (13)–(16). For each model, we have to compute the mean values A_μ and D_μ.

Recall that $p_s = gC$ is the total number of patterns that have been learned, and $p_m = \alpha C$ is the total number of memorized patterns that are effectively memorized (these are the most recently learned patterns in models B1 and C, the oldest ones in model D). Whenever condition (14) is true for all μ, $1 < \mu < p_s$, then $\alpha = g$: all of what has been learned is memorized.

When g reaches g^*, (14) becomes an equality:

$$\max_\mu \Delta_\mu = (2/\pi)^{1/2} \tag{25}$$

For g larger than g^*, the capacity α decreases, and α is given by the number of μ such that (14) is true. When g increases, α might reach zero at a value g_c or an asymptotic value $\alpha_c > 0$ for $g \to \infty$ (see Section 2).

In all cases that we consider, Δ_μ is a monotonic function of μ (increasing with ancestry for models B1 and C, and with recency for model D). Equation (25) can be written, in the large-C limit,

$$\Delta(\alpha, g) = (2/\pi)^{1/2} \tag{26}$$

with

$$\Delta(\alpha, g) = [D(\alpha, g)/A^2(\alpha, g)]^{1/2} \tag{27}$$

$$A(\alpha, g) = \lim_{C \to \infty} A_\mu \tag{28}$$

$$D(\alpha, g) = \lim_{C \to \infty} (D_\mu - A_\mu{}^2)/C \tag{29}$$

where $\mu = p_s - \alpha C$ for models B1 and C, and $\mu = \alpha C$ for model D. Once the function $\Delta(\alpha, g)$ has been determined, the thresholds $g^*(\varepsilon)$ and $g_c(\varepsilon)$ are respectively given by

$$\Delta(g^*, g^*) = (2/\pi)^{1/2} \tag{30}$$

$$\Delta(0, g_c) = (2/\pi)^{1/2} \tag{31}$$

For $g^* < g < g_c$, the capacity $\alpha(\varepsilon, g)$ is given by (26). In particular, for $\varepsilon > \varepsilon_c$, there is an asymptotic capacity $\alpha_c(\varepsilon)$ given by

$$\Delta(\alpha_c, \infty) = (2/\pi)^{1/2} \tag{32}$$

and the optimal capacity is obtained by

$$[d\alpha/d\varepsilon](\varepsilon_{\text{opt}}) = 0, \qquad \alpha_{\text{opt}} = \alpha_c(\varepsilon_{\text{opt}}) \tag{33}$$

4.2. Weighted Schemes

We consider first the simplest case, model B2, which contains as a particular case the Little–Hopfield model (model A) for $\Lambda = \Lambda_H$:

$$\begin{aligned} \Lambda_H(u) &= 1/\sqrt{g}, & u \leqslant g = p_s/C \\ &= 0, & u > g \end{aligned} \tag{34}$$

and the marginalist scheme (model B1) for $\Lambda = \Lambda_m$:

$$\Lambda_m(u) = \begin{cases} \varepsilon \exp(-\varepsilon^2 u/2), & u \leqslant g \\ 0, & u > g \end{cases} \tag{35}$$

For any decreasing function Λ, one finds

$$A(\alpha, g) = \Lambda(\alpha) \tag{36}$$

$$D(\alpha, g) = \int_0^g \Lambda^2(u)\, du \tag{37}$$

Resolution of Eqs. (30)–(33) gives for the marginalist scheme (35)

$$\varepsilon_c = (\pi/2)^{1/2} \tag{38}$$

and the functions

$$g^*(\varepsilon) = (1/\varepsilon^2)\log(1+\varepsilon^2/\varepsilon_c{}^2) \tag{39}$$

$$g_c(\varepsilon) = -(1/\varepsilon^2)\log(1-\varepsilon^2/\varepsilon_c{}^2) \qquad \text{for} \quad \varepsilon < \varepsilon_c \tag{40}$$

For $g^* < g \leqslant g_c$, the capacity α is

$$\alpha(\varepsilon, g) = (1/\varepsilon^2)\log\{\varepsilon^2/\varepsilon_c{}^2[1-\exp(-\varepsilon^2 g)]\} \tag{41}$$

and the asymptotic capacity $\alpha_c(\varepsilon)$ for $\varepsilon \geqslant \varepsilon_c$ is

$$\alpha_c(\varepsilon) = (2/\varepsilon^2)\log(\varepsilon/\varepsilon_c) \tag{42}$$

The optimal capacity α_{opt} is reached at $\varepsilon = \varepsilon_{\text{opt}}$:

$$\varepsilon_{\text{opt}} = \varepsilon_c\sqrt{e} = 2.066 \tag{43}$$

$$\alpha_{\text{opt}} = 1/\varepsilon_{\text{opt}}^2 = 0.234 \tag{44}$$

One should notice that the ratio $1/e = 0.368$, between the optimal capacity and the maximal capacity of the Little–Hopfield scheme, is very close to the corresponding ratio obtained for the nondiluted, symmetric case,[12] namely $0.0489/0.138 = 0.354$.

If one requires a retrieval quality at least equal to M, one has to replace $(2/\pi)^{1/2}$ by M/X [see (15), (16)] in Eqs. (25), (26), and (30)–(32). This gives the same solutions (39)–(44), with now $\varepsilon_c = X/M$. For example, if one choses $M = 0.97$, for the marginalist scheme one has $\varepsilon_c = 2.23$, $\alpha_{\text{opt}} = 1/(\varepsilon_c{}^2 e) = 0.074$, $\varepsilon_{\text{opt}} = \varepsilon_c\sqrt{e} = 3.68$. In the limit of very good retrieval (M very close to one), Eqs. (39)–(44) become identical to those obtained in the same limit for the nondiluted, symmetric case.[12]

4.3. Learning within Bounds (Model C)

In this scheme each synaptic efficacy T_{ij} is bounded above and below:

$$-L \leqslant T_{ij} \leqslant L \tag{45}$$

and each pattern is stored with a uniform acquisition amplitude, provided a bound is not reached. That is, with the appropriate scaling, one has

$$\begin{aligned} T_{ij}(p) &= T_{ij}(p-1) + (\varepsilon L/\sqrt{C})\, \xi^p{}_i \xi^p{}_j \qquad \text{if no bound is reached} \\ &= \text{value of the bound } (+L \text{ or } -L) \qquad \text{otherwise} \end{aligned} \tag{46}$$

Let us fix the scale by

$$L = C^{1/2}/\varepsilon \tag{47}$$

Thus, the probability distribution of a given T_{ij} as a function of the number of stored patterns, chosen at random, is the probability distribution of a one-dimensional random walker making unit steps ± 1 with equal probability within two nonabsorbing walls at $\pm L$. Solving this problem for the fully connected network ($C = N$) is difficult because one has to deal with the correlated random walks of different T_{ij}. Here, in the dilute case, one needs only to consider the average properties of *one* synaptic efficacy: we just have to solve for the problem of one random walk between walls.

We will only consider the simplest case where L is an integer. There is no difficulty *a priori* to generalize the calculation for a noninteger L, but the calculations are a little more complicated for finite L, and should become identical in the infinite-L limit. One could also generalize to synaptic efficacies having a given sign (that is, $0 \leqslant T_{ij} \leqslant L$ or $-L \leqslant T_{ij} \leqslant 0$ for each T_{ij}).

To solve our problem we need the averages $\langle z \rangle$ and $\langle z^2 \rangle / C$ in the large-C limit, after t patterns have been stored, z being the position of the random walker, knowing that at time τ the walker makes a $+1$ step. Let $p_z(t)$ be the probability distribution of the position z at time t. At any time $t \neq \tau$ the evolution equations for p_z are

$$p_z(t) = \tfrac{1}{2}[p_{z-1}(t-1) + p_{z+1}(t-1)] \qquad \text{for} \quad |z| < L \tag{48}$$

$$p_{\pm L}(t) = \tfrac{1}{2}[p_{\pm(L-1)}(t-1) + p_{\pm L}(t-1)] \tag{49}$$

with the initial condition

$$p_z(0) = \delta_{z,0} \tag{50}$$

and at time τ

$$\begin{aligned} p_z(\tau) &= p_{z-1}(\tau-1) \qquad \text{for} \quad |z| < L \\ p_L(\tau) &= p_{L-1}(\tau-1) + p_L(\tau-1) \\ p_{-L}(\tau) &= 0 \end{aligned} \tag{51}$$

One can solve these equations, and one obtains

$$p_z(t) = \frac{1}{2L+1} + \frac{2}{2L+1} \sum_{K=1}^{L} \cos\left(\frac{2K\pi z}{2L+1}\right)\left[\cos\left(\frac{2K\pi}{2L+1}\right)\right]^t$$
$$+ \frac{4}{(2L+1)^2} \sum_{K=0}^{L} \sin\left[\frac{(2K+1)\pi z}{2L+1}\right]\left\{\cos\left[\frac{(2K+1)\pi}{2L+1}\right]\right\}^{t-\tau} (-)^K$$
$$\times \cos\left[\frac{(2K+1)\pi}{2(2L+1)}\right]\left\{1 + \sum_{Q=1}^{L} (-)^Q \cos\left(\frac{Q\pi}{2L+1}\right)\left[\cos\left(\frac{2Q\pi}{2L+1}\right)\right]^{\tau-1}\right.$$
$$\left. \times \frac{1-\cos[(2K+1)\pi/(2L+1)]}{\cos[2Q\pi/(2L+1)] - \cos[(2K+1)\pi/(2L+1)]}\right\} \tag{52}$$

The two first terms contain the symmetric (with respect to $z \to -z$) part of p_z, which will give $\langle z^2 \rangle$. The remainder is the antisymmetric part, which will give $\langle z \rangle$. As explained in Section 2, we are interested in the large-C limit, hence in the large-L limit, with the scaling

$$t = gC \tag{53}$$

$$t - \tau = \alpha C \tag{54}$$

Since $L = C^{1/2}/\varepsilon$, where ε is a measure of the acquisition amplitude of the patterns, one finds

$$A(\alpha, g) = \lim_{C \to \infty} \langle z \rangle$$
$$= \sum_{K \geqslant 0} [8/\pi^2(2K+1)^2] \exp[-\alpha\varepsilon^2\pi^2(2K+1)^2/8]$$
$$\times \left(1 + \sum_{Q \geqslant 1} 2(-)^Q \{(2K+1)^2/[(2K+1)^2 - 4Q^2]\}\right.$$
$$\left. \times \exp[-(g-\alpha)\varepsilon^2\pi^2Q^2/2]\right) \tag{55}$$

$$D(\alpha, g) = \lim_{C \to \infty} \langle z^2 \rangle / C$$
$$= (1/3\varepsilon^2)\left[1 + \sum_{Q \geqslant 1} (-)^Q (12/\pi^2Q^2) \exp(-g\varepsilon^2\pi^2Q^2/2)\right] \tag{56}$$

Let us comment on these formulas.

1. Note first that $\langle z^2 \rangle$ is only a function of the total time, and not of the time τ: this is expected, since for each walk with a $+1$ step at time τ

one can consider the symmetric walk with a -1 step, showing that the same result for $\langle z^2 \rangle$ should be obtained whatever the direction of the step at time τ. Note also that at time 0, that is, $g=0$, one must have $\langle z^2 \rangle = 0$, which is indeed the case, since $\sum_Q (-)^Q 12/\pi^2 Q^2 = -1$.

2. The infinite-time limit (α and g going to infinity) is obvious: far from a specific learning event, $\langle z \rangle$ has to be zero; the asymptotic distribution of p_z is easily found from (48), (49) to be $p_z = 1/(2L+1)$, which gives $\langle z^2 \rangle = L(L+1)/3$, in agreement with $D(\infty, \infty) = 1/3\varepsilon^2$.

3. For $\alpha = 0$, that is, looking exactly at the time when the walker makes the $+1$ step, one must find $\langle z \rangle = 1$. This can be checked on the formula (55) for $\langle z \rangle$, noting that

$$\sum_{K \geqslant 0} 8/\pi^2 (2K+1)^2 = 1$$

and for any integer $Q \geqslant 1$,

$$\sum_{K \geqslant 0} [(2K+1)^2 - 4Q^2]^{-1} = 0$$

The critical value ε_c is

$$\varepsilon_c = (\pi/6)^{1/2} = 0.7236... \tag{57}$$

Numerical solutions of Eqs. (30)–(33) give the curves $g^*(\varepsilon)$, $g_c(\varepsilon)$, and $\alpha_c(\varepsilon)$, which are displayed on Figs. 4 and 5. The optimal stationary capacity is found at

$$\varepsilon_{\text{opt}} = 1.456... \tag{58}$$

with the value

$$\alpha_{\text{opt}} = 0.18788... \tag{59}$$

4.4. Learning within Absorbing Bounds (Model D)

A very similar computation can be done for model D. For this model, the probability distribution $p_z(t)$ obeys the evolution equations

$$\begin{aligned} p_z(t) &= \tfrac{1}{2}[p_{z-1}(t-1) + p_{z+1}(t-1)] \qquad \text{for} \quad |z| < L-1 \\ p_{\pm(L-1)}(t) &= \tfrac{1}{2} p_{\pm(L-2)}(t-1) \\ p_{\pm L}(t) &= p_{\pm L}(t-1) + \tfrac{1}{2} p_{\pm(L-1)}(t-1) \end{aligned} \tag{60}$$

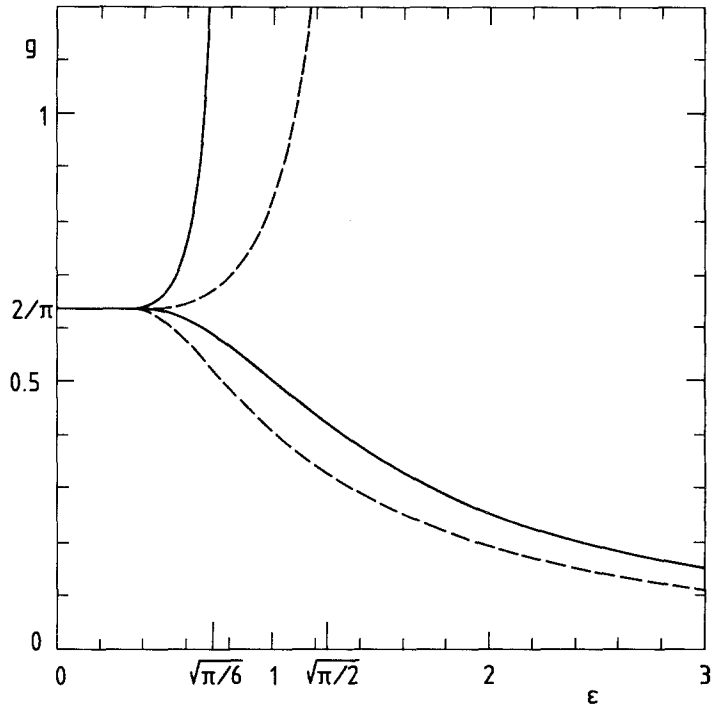

Fig. 4. Curves $g^*(\varepsilon)$ and $g_c(\varepsilon)$ versus ε for the (—) C (short-term memory model), and (- -) D (long-term memory model).

with the initial condition

$$p_z(0) = \delta_{z,0} \tag{61}$$

and at time τ the learning of the configuration ($\xi_i = +1$) implies

$$\begin{aligned} p_z(\tau) &= p_{z-1}(\tau-1) \qquad \text{for} \quad -(L-2) \leqslant z \leqslant L-1 \\ p_L(\tau) &= p_{(L-1)}(\tau-1) + p_L(\tau-1) \\ p_{-(L-1)}(\tau) &= 0 \\ p_{-L}(\tau) &= p_{-L}(\tau-1) \end{aligned} \tag{62}$$

With the appropriate scaling

$$t = gC, \qquad \tau = \alpha C \tag{63}$$

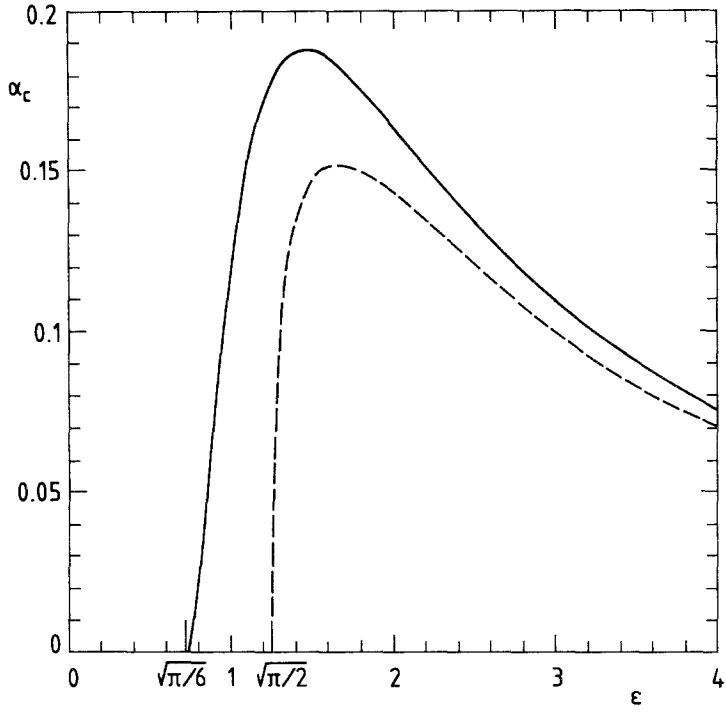

Fig. 5. Asymptotic capacity $\alpha_c(\varepsilon)$ for model (—) C and (- -) D.

the results are

$$A(\alpha, g) = \sum_{K \geqslant 0} (-)^K [4/\pi(2K+1)] \exp[-\alpha\varepsilon^2\pi^2(2K+1)^2/8] \tag{64}$$

$$D(\alpha, g) = (1/\varepsilon^2) \left\{ 1 + \sum_{K \geqslant 0} (-)^{K+1} [32/\pi^3(2K+1)^3] \times \exp[-g\varepsilon^2\pi^2(2K+1)^2/8] \right\} \tag{65}$$

Note that the mean value A depends only on the number of learning events since the beginning. These results provide again a phase portrait similar to Fig. 1, with here α giving the number of earliest learned patterns still memorized.

Here the critical value ε_c is

$$\varepsilon_c = (\pi/2)^{1/2} = 1.2533... \tag{66}$$

Again numerical solution gives the curves $g^*(\varepsilon)$, $g_c(\varepsilon)$, and $\alpha_c(\varepsilon)$, which are also displayed on Figs. 4 and 5 for comparison with the previous model. The optimal values ε_{opt} and α_{opt} are here

$$\varepsilon_{\text{opt}} = 1.667... \tag{67}$$

$$\alpha_{\text{opt}} = 0.15216... \tag{68}$$

These results are rather similar to those obtained in the previous scheme of learning within bounds (Section 4.3). The performances are slightly diminished. For both models the Little–Hopfield model is recovered in the $\varepsilon \to 0$ limit.

For the symmetric, nondiluted case, a signal-to-noise analysis with a similar random walk approach has been made recently.[15]

5. CONCLUSION

In this work we have compared different learning schemes for diluted, asymmetric architectures that allow an exact solution. All these schemes where forgetting occurs have qualitatively similar properties, with slightly different performances.

Other questions, such as learning biased patterns,[16] hierarchies[17] of patterns, time-dependent patterns (sequences),[18] and schemes where active and inactive neurons, and presynaptic and postsynaptic neurons play nonsymmetric roles, could be studied analytically on this architecture.

ACKNOWLEDGMENTS

We have benefitted from enlightening discussions with Marc Mézard and Gérard Toulouse.

REFERENCES

1. J. J. Hopfield, *Proc. Natl. Acad. Sci USA* **79**:2554 (1982).
2. W. A. Little, *Math. Biosci.* **19**:101 (1974); W. A. Little and G. L. Shaw, *Math. Biosci.* **39**:281 (1978).
3. D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. Lett.* **55**:1530 (1985); *Ann. Phys. (NY)* **173**:30 (1987).
4. H. Sompolinsky, *Phys. Rev. A* **34**:2571 (1986).
5. H. Gutfreund and Y. Stein, to be published.
6. W. Kinzel, in *Proceedings of 1986 Heidelberg Colloquium on Glassy Dynamics and Optimization* (Springer, Lecture Notes in Physics).
7. G. Parisi, *J. Phys. A* **19**:L675 (1986).

8. J. A. Hertz, G. Grinstein, and S. A. Solla, preprint (1986); and in *Proceedings of 1986 Heidelberg Colloquium on Glassy Dynamics and Optimization* (Springer, Lecture Notes in Physics).
9. P. Peretto, communication at Journées de Physique Statistique, Paris (1987); and preprint (1987).
10. B. Derrida, E. Gardner, and A. Zippelius, *Europhys. Lett.* **4**:167 (1987).
11. J. P. Nadal, G. Toulouse, J. P. Changeux, and S. Dehaene, *Europhys. Lett.* **1**:535 (1986).
12. M. Mézard, J. P. Nadal, and G. Toulouse, *J. Phys. (Paris)* **47**:1457 (1986).
13. G. Parisi, *J. Phys. A* **29**:L617 (1986).
14. P. Peretto, private communication.
15. T. Geszti and F. Pázmándi, in preparation; M. Gordon, to appear.
16. D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Phys. Rev. A* (1987), in press.
17. N. Parga and M. A. Virasoro, *J. Phys. (Paris)* **47**:1857 (1986); M. V. Feigel'man and L. B. Ioffe, Landau Institute preprint (1986); H. Gutfreund, preprint (1987); V. S. Dotsenko, *J. Phys. C* **18**:L1017 (1985).
18. P. Peretto and J. J. Niez, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman, and G. Weibusch, eds. (Springer, Berlin, 1986), pp. 171–185; D. Kleinfield, *Proc. Natl. Acad. Sci USA* **83**:9469 (1986); H. Sompolinsky and I. Kanter, *Phys. Rev. Lett.* **57**:2861 (1986); D. W. Tank and J. J. Hopfield, preprint (1986); J. Buhmann and K. Schulten, *Europhys. Lett.* **4**:1205 (1987); S. Dehaene, J. P. Changeux, and J. P. Nadal, *Proc. Natl. Acad. Sci USA* **84**:2727 (1987).