J. Phys. A: Math. Gen. **22** (1989) 2069-2080. Printed in the UK

Distribution of the activities in a diluted neural network

B Derrida

The Institute for Advanced Studies, The Hebrew University of Jerusalem, Jerusalem 91904, Israel and Service de Physique Théorique de Saclay†, F-91191 Gif-sur-Yvette Cedex, France

Received 13 December 1988

Abstract. The dynamics of asymmetrically diluted neural networks can be solved exactly. In the present work, the distribution of the neural activities is calculated analytically for zero-temperature parallel dynamics. This distribution depends on the number of stored patterns and is a continuous function in the good retrieval phase. The continuous part of the distribution of activities is due to the asymmetry of the synapses since it is known that networks with symmetric interactions always have a distribution of activities which is a sum of a few delta functions. The expression for the distribution of activities is also given for a mixture of two patterns which have a non-zero overlap.

1. Introduction

Neural network models with symmetric synapses ($J_{ij} = J_{ji}$) have been extensively studied in the last few years, (Little 1974, Hopfield 1982, 1984, Amit *et al* 1985a, b, 1987). The main approach to these systems was to investigate their properties at thermal equilibrium, trying to relate the appearance of phase transitions or of metastable states in the mean-field equations to the structure of the attractors for Glauber-like dynamics. At present, the properties of these networks at thermal equilibrium are considered to be well understood, at least if one believes in the validity of the replica approach, whereas less is known of the dynamics when the initial condition is far from equilibrium. At zero temperature, the dynamics of these models becomes simpler and one can show that starting with any initial condition, the system converges to a fixed point in phase space (for sequential dynamics) implying that the activity of each neuron does not change with time. Once the attractor has been reached, a given neuron is either quiet for ever or firing for ever. This is a direct consequence of the assumption that the synapses are symmetric. Of course, these fixed activities as well as the symmetry of the synapses have no justification from a neurobiological point of view.

For non-symmetric synapses ($J_{ij} \neq J_{ji}$) the equilibrium approach is no longer possible (with one exception (Coolen and Ruijgrok 1988)) and the properties of the network have to be understood by directly studying its dynamics. This usually makes the problem much harder since the dynamics is always more difficult to understand than equilibrium properties. There exists, however, a class of models for which the asymmetry of interactions makes the problem easier, allowing for a full analytic solution of the dynamics (Derrida *et al* 1987). This class of diluted models has the following two properties.

† Laboratoire de l'Institut de Recherche Fondamentale du Commissariat à L'Énergie Atomique.

0305-4470/89/122069+12$02.50

(i) When the number, N, of neurons becomes infinite, each neuron i receives non-zero synaptic inputs from only a finite number of other neurons j chosen at random among the N neurons.

(ii) The events (neuron j has an input to neutron i) and (neuron i has an input to neuron j) are not correlated, implying that almost all the synapses are unidirectional.

For several models belonging to this class (Derrida *et al* 1987, Derrida and Nadal 1987, Kree and Zippelius 1987, Gutfreund and Mézard 1988, Noest 1988), properties like overlaps, projections on stored patterns and storage capacities have already been calculated analytically. The purpose of the present work is to extend these results by calculating the full probability distribution of the neuronal activities. As a consequence of the non-symmetry of the synapses ($J_{ij} \neq J_{ji}$), we will see that even for zero-temperature dynamics, the states of almost all neurons change with time.

In this work the calculation of the distribution of activities will be described for the simplest diluted neural network (Derrida *et al* 1987). There should be no difficulty in extending these calculations to more complicated models (Derrida and Nadal 1987, Gutfreund and Mézard 1988, Noest 1988) which belong to the same class of diluted neural networks.

The model is defined as a system of N neurons $S_i = \pm 1$ ($S_i(t) = +1$ if neuron i is firing at time t and -1 if it is quiet). The synaptic strengths J_{ij} are given by the expression

$$J_{ij} = C_{ij} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \tag{1}$$

where p is the number of stored patterns, $\xi_i^{(\mu)} = \pm 1$ is the value of the μth pattern at site i and C_{ij} represents the dilution and has a statistical distribution given by

$$\rho(C_{ij}) = (C/N)\delta(C_{ij} - 1) + [1 - C/N]\delta(C_{ij}). \tag{2}$$

By definition of the model, C_{ij} and C_{ji} are independent random variables. Thus if $J_{ij} \neq 0$, one has $J_{ij} = 0$ with probability 1.

For this neural network, only zero-temperature parallel dynamics will be discussed here:

$$S_i(t+1) = \operatorname{sgn}(h_i(t)) \tag{3a}$$

with

$$h_i(t) = \sum_j J_{ij} S_j(t). \tag{3b}$$

Zero-temperature parallel dynamics means that at each time step, all the neurons are updated simultaneously according to (3). (If $h_i(t) = 0$, one can choose, for example, $S_i(t+1) = \pm 1$ at random but the results presented below will not depend on this choice because only the limit of large C will be considered and in that limit $h_i(t)$ is non-zero with probability 1.)

If one defines the projection $m_\mu(t)$ of the configuration $\{S_i(t)\}$ of the system at time t on the μth pattern by

$$m_\mu(t) = \frac{1}{N} \sum_{i=1}^{N} S_i(t) \xi_i^{(\mu)} \tag{4}$$

it was shown that in the thermodynamic limit ($N \to \infty$) the evolution of $m_1(t)$ is given by equation (9) of Derrida *et al* (1987):

$$m_1(t+1) = f(m_1(t)) \tag{5}$$

with

$$f(m) = \sum_{K=0}^{\infty} \frac{C^K \exp(-C)}{K!} \sum_{n=0}^{K} \sum_{s=0}^{K(p-1)} \frac{(1+m)^{K-n}(1-m)^n}{2^{Kp}} \binom{K}{n} \times \binom{K(p-1)}{s} \operatorname{sgn}(Kp-2n-2s). \tag{6}$$

It is assumed here that the patterns $\{\xi_i^{(\mu)}\}$ are chosen at random (with probability $\frac{1}{2}$ for the two orientations $\xi_i^{\mu} = \pm 1$) and that the initial configuration $\{S_i(0)\}$ has a finite projection on a single pattern only:

$$m_1(0) = O(1) \qquad m_\mu(0) \sim N^{-1/2}. \tag{7}$$

For large C and p, (6) simplifies to

$$f(m) = \sqrt{\frac{2}{\pi\alpha}} \int_0^m dy \exp\left(-\frac{y^2}{2\alpha}\right) = \operatorname{erf}\left(\frac{m}{\sqrt{2\alpha}}\right) \tag{8}$$

where α is defined by

$$\alpha = p/C \tag{9}$$

and where

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-z^2)\, dz. \tag{10}$$

There is a critical value which is the storage capacity of the network

$$\alpha_c = 2/\pi \simeq 0.6366\ldots. \tag{11}$$

For $\alpha > \alpha_c$, the only fixed point of (5) is $m_1 = 0$ and $m_1(t)$ always converges to 0. For $\alpha < \alpha_c$, $m_1 = 0$ is an unstable fixed point and there appear two attractive fixed points m_1^* and $-m_1^*$ where m_1^* is solution of $m_1^* = f(m_1^*)$.

2. The probability distribution of activities

The projection $m_1(t)$ represents the activity averaged over all the neurons (see (4)). One can define a more microscopic quantity $a_i(t)$, the activity of neuron i at time t:

$$a_i(t) = \overline{S_i(t)\xi_i^{(1)}} \tag{12}$$

where the overbar means an average over an ensemble of initial conditions. If this ensemble of initial conditions consists of the set of $\{S_i(0)\}$ which have a given non-zero projection $m_1(0)$ on a single pattern $\{\xi_i^{(1)}\}$ and zero projections on the other patterns, one has

$$m_1(t) = \frac{1}{N} \sum_{i=1}^{N} a_i(t). \tag{13}$$

The probability distribution $Q_t(a)$ of local activities which will be calculated in the present work is defined by

$$Q_t(a) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta(a - a_i(t)). \tag{14}$$

Let us assume that the neurons are uncorrelated in the initial condition, i.e. that the probability distribution $\mathcal{P}(\{S_i(0)\})$ has the form

$$\mathcal{P}(\{S_i(0)\}) = \prod_{i=1}^{N} \frac{1+b_i(0)S_i(0)}{2}. \tag{15}$$

If one defines $b_i(t)$ to be the value of neuron i at time t averaged over the initial conditions (15):

$$b_i(t) = \overline{S_i(t)} \tag{16}$$

then one can write a relation between $b_i(t+1)$ of neuron i at time $t+1$ and the $b_{j_1}(t), \ldots, b_{j_K}(t)$ of its K inputs $j_1, \ldots, j_K$ at time t (asuming that for site i, only $J_{ij_1}, \ldots, J_{ij_K}$ are non-zero):

$$b_i(t+1) = \sum_{\sigma_{j_1}=\pm 1} \cdots \sum_{\sigma_{j_K}=\pm 1} \left(\prod_{r=1}^{K} \frac{1+\sigma_{j_r} b_{j_r}(t)}{2} \right) \operatorname{sgn}\left(\sum_{r=1}^{K} J_{ij_r}\sigma_{j_r} \right). \tag{17}$$

Expression (17) is valid for almost all sites i. It expresses the fact that, at time t, the states of the neurons $S_{j_1}(t), \ldots, S_{j_K}(t)$ are uncorrelated. This is true for almost all sites i because (Derrida *et al* 1987, Derrida and Flyvbjerg 1987, Flyvbjerg 1988) the tree of all the ancestors of a site i from time t to time 0, has no loop and the K immediate ancestors of this site are uncorrelated. (Equation (17) for the activities of the neurons is the analogue of equation (7) of Derrida and Flyvbjerg (1987) for the Kauffman model.)

Equation (17) gives the activity $b_i(t+1)$ as a function of the activities $b_{j_r}(t)$ and of the patterns $\{\xi_i^{(\mu)}\}$ through the J_{ij}. Let us define $P_t(b, \eta^{(1)}, \ldots, \eta^{(p)})$ as the probability that a site i has an activity $b_i(t)$ at time t knowing that the values of the patterns on this site i are $\eta^{(1)}, \ldots, \eta^{(p)}$ (i.e. knowing that $\eta^{(\mu)} = \xi_i^{(\mu)}$ for $1 \leqslant \mu \leqslant p$).

From (17) one can write:

$$P_{t+1}(b, \eta^{(1)}, \ldots, \eta^{(p)}) = \sum_{K=0}^{\infty} \frac{C^K \exp(-C)}{K!} \left\langle \int db_1 P_t(b_1, \xi_1^{(1)}, \ldots, \xi_1^{(p)}) \ldots \int db_K P_t(b_K, \xi_K^{(1)}, \ldots, \xi_K^{(p)}) \times \delta\left\{ b - \sum_{\sigma_1=\pm 1} \cdots \sum_{\sigma_K=\pm 1} \left[\prod_{r=1}^{K} \left(\frac{1+\sigma_r b_r}{2} \right) \right] \operatorname{sgn}\left(\sum_{r=1}^{K} \sum_{\mu=1}^{p} \eta^{(\mu)} \xi_r^{(\mu)} \sigma_r \right) \right\} \right\rangle \tag{18}$$

where $\langle\ \rangle$ means the average over the patterns $\{\xi_r^{(\mu)}\}$.

This equation gives the time evolution of the 2^p distributions $P_t(b, \eta^{(1)}, \ldots, \eta^{(p)})$ for arbitrarily correlated patterns $\{\xi_i^{(\mu)}\}$. It plays the same role as equation (8) of Derrida and Flyvbjerg (1987) for the Kauffman model.

Equation (18) is valid provided that the statistical properties of the patterns are homogeneous in space (i.e. $\langle \xi_i^{(\mu_1)} \ldots \xi_i^{(\mu_l)} \rangle$ is independent of i) and that the spins $S_i(t)$ are not correlated in the initial condition.

Equation (18) is general (arbitrary correlations between the patterns) but complicated: b is a continuous variable and $\eta^{(1)}, \ldots, \eta^{(p)}$ can take 2^p possible values. So one has to iterate 2^p functions of one variable. This difficulty can be greatly simplified if one considers simple cases for which, in the initial condition, the spin configuration has non-zero projections on a few patterns only.

3. Finite projection on a single pattern

Let us first consider the case of a finite projection on a single pattern and no correlation between the patterns. We choose the initial condition at $t=0$ to be

$$P_0(b, \eta^{(1)}, \ldots, \eta^{(p)}) = Q_0(b\eta^{(1)}). \tag{19}$$

So P_0 depends only on the product $b\eta^{(1)}$. One can easily check by looking at (18) (and by using the fact that the patterns are not correlated) that if P_0 has the shape (19), then P_t has exactly the same shape:

$$P_t(b, \eta^{(1)}, \ldots, \eta^{(p)}) = Q_t(b\eta^{(1)}). \tag{20}$$

Therefore the recursion (18) becomes a recursion for a single function of one variable:

$$Q_{t+1}(a) = \sum_{K=0}^{\infty} \frac{C^K \exp(-C)}{K!} \int da_1 Q_t(a_1) \ldots \int da_K Q_t(a_K) \times \left\langle \delta\left[a - \sum_{\tau_1=\pm 1} \ldots \sum_{\tau_K=\pm 1} \left(\prod_{r=1}^{K} \frac{1+\tau_r a_r}{2} \right) \operatorname{sgn}\left(\sum_{r=1}^{K} \tau_r + \sum_{\mu=2}^{p} \sum_{r=1}^{K} \psi_r(\mu)\tau_r \right) \right] \right\rangle \tag{21}$$

where $\tau_r = \sigma_r \xi_r^{(1)} = \pm 1$ and $\psi_r^{(\mu)} = \eta^{(1)}\eta^{(\mu)}\xi_r^{(1)}\xi_r^{(\mu)} = \pm 1$. Equation (21) means that a is a random variable given by

$$a = \overline{\operatorname{sgn}\left(\sum_{r=1}^{K} \tau_r + \sum_{\mu=2}^{p} \sum_{r=1}^{K} \psi_r^{(\mu)} \tau_r \right)} \tag{22}$$

where the overbar in (22) is an average over the τ_r only ($\tau_r = +1$ with probability $\frac{1}{2}(1+a_r)$ and -1 with probability $\frac{1}{2}(1-\alpha_r)$). So a is a random variable which depends on the $\psi_r^{(\mu)} = \pm 1$ and on the a_r. One can rewrite (22) in the form

$$a = \overline{\operatorname{sgn}(A+B)} \tag{23}$$

with

$$A = \frac{1}{K} \sum_{r=1}^{K} (\tau_r - a_r) + \frac{1}{K} \sum_{\mu=2}^{p} \sum_{r=1}^{K} \psi_r^{(\mu)} (\tau_r - a_r) \tag{24}$$

and

$$B = \frac{1}{K} \sum_{r=1}^{K} a_r + \frac{1}{K} \sum_{\mu=2}^{p} \sum_{r=1}^{K} \psi_r^{(\mu)} a_r. \tag{25}$$

Since the τ_r variables appear only in A, this means that in (23) the overbar average is an average on A only. When $K \to \infty$, A is a sum of a large number of random independent variables. So A becomes a Gaussian variable whose distribution $\rho_1(A)$ is given by

$$\rho_1(A) = \frac{1}{\sqrt{2\pi\alpha(1-\langle a^2 \rangle_t)}} \exp\left(-\frac{A^2}{2\alpha(1-\langle a^2 \rangle_t)} \right). \tag{26}$$

Since

$$\bar{A} = 0 \tag{27}$$

$$\overline{A^2} = \frac{1}{K^2} \sum_{r=1}^{K} (1-a_r^2) \left(1 + \sum_{\mu=2}^{p} \psi_r^{(\mu)} \right)^2 \to \alpha(1-\langle a^2 \rangle_t) \tag{28}$$

when $K \to \infty$. So (23) becomes

$$a = \int \mathrm{d}A\, \rho_1(A)\, \mathrm{sgn}(A+B) = \mathrm{erf}\left(\frac{B}{\sqrt{2\alpha(1-\langle a^2\rangle_t)}}\right). \tag{29}$$

The variable a as given by (29) is still a random variable since it depends on the a_r and the $\psi_r^{(\mu)}$ through B. When $K \to \infty$, B is again a sum of a large number of random variables and so becomes Gaussian.

For $K \to \infty$

$$\begin{aligned} &\langle B\rangle \to \langle a\rangle_t = \int a\, \mathrm{d}a\, Q_t(a) \\ &\langle B^2\rangle - \langle B\rangle^2 \to \alpha\langle a^2\rangle_t = \int a^2 \mathrm{d}a\, Q_t(a) \end{aligned} \tag{30}$$

and the distribution of B becomes

$$\rho_2(B) = \frac{1}{\sqrt{2\pi\alpha\langle a^2\rangle_t}} \exp\left(-\frac{(B-\langle a\rangle_t)^2}{2\alpha\langle a^2\rangle_t}\right). \tag{31}$$

So we see that for large K, (21) becomes

$$\begin{aligned} Q_{t+1}(a) &= \int \mathrm{d}B\, \rho_2(B)\delta\left(a - \int \mathrm{d}A\, \rho_1(A)\mathrm{sgn}(A+B)\right) \\ &= \int \mathrm{d}B\, \rho_2(B)\delta\left[a - \mathrm{erf}\left(\frac{B}{\sqrt{2\alpha(1-\langle a^2\rangle_t)}}\right)\right]. \end{aligned} \tag{32}$$

It is possible to find a parametrisation which allows one to draw $Q_{t+1}(a)$:

$$Q_{t+1}(a) = \frac{1}{2}\left(\frac{1-\langle a^2\rangle_t}{\langle a^2\rangle_t}\right)^{1/2} \exp\left(z^2 - \frac{(z\sqrt{1-\langle a^2\rangle_t} - \langle a\rangle_t/\sqrt{2\alpha})^2}{\langle a^2\rangle_t}\right) \tag{33}$$

where $a = \mathrm{erf}\,(z)$.

Clearly $Q_{t+1}(a)$ is fully known once the first two moments $\langle a\rangle_t$ and $\langle a^2\rangle_t$ are known. This is not surprising because in the limit $K \to \infty$, one has to deal with sums of a large number of random variables. If K or C had been finite, more information than $\langle a\rangle_t$ or $\langle a^2\rangle_t$ would be needed and one would have to solve (21).

From (32), one can show that

$$\langle a\rangle_{t+1} = \int a\, \mathrm{d}a\, Q_{t+1}(a) = \mathrm{erf}\left(\frac{\langle a\rangle_t}{\sqrt{2\alpha}}\right) \tag{34}$$

$$\langle a^2\rangle_{t+1} = -1 + \frac{2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp(-y^2)\, \mathrm{d}y\, \mathrm{erf}\left(\left|\frac{y\sqrt{1+\langle a^2\rangle_t} + \langle a\rangle_t/\sqrt{\alpha}}{\sqrt{1-\langle a^2\rangle_t}}\right|\right). \tag{35}$$

Starting with an arbitrary distribution $Q_0(b)$, one can iterate $\langle a\rangle_t$ and $\langle a^2\rangle_t$ using (34) and (35) and then calculate the distribution $Q_t(a)$. Of course the expression (34) is identical to (8) since, by definition, $\langle a\rangle_t = m_1(t)$.

When $t \to \infty$, $\langle a\rangle_t$ and $\langle a^2\rangle_t$ converge in general to fixed-point values $\langle a\rangle^*$ and $\langle a^2\rangle^*$ and $Q_t(a)$ converges to a fixed distribution $Q_\infty(a)$. For all α, the fixed point $\langle a\rangle^* = \langle a^2\rangle^* = 0$ always exists. It corresponds to

$$Q_\infty(a) = \delta(a). \tag{36}$$

This fixed point describes the configurations which are attracted by no pattern.

For $\alpha < \alpha_c$, this fixed point becomes unstable ($\langle a \rangle^* = 0$ is an unstable fixed point of (34)) and a new solution $Q_\infty(a)$ appears corresponding to the attractive fixed point $\langle a \rangle^* \neq 0$ of (34). This fixed distribution $Q_\infty(a)$ is shown in figure 1(a) for $\alpha = 0.4$ and figure 1(b) for $\alpha = 0.6$. One can see that for small α the distribution is very concentrated near $a = 1$ and $a = -1$ with a stronger weight near $a = 1$. As a increases, the divergences at $a = \pm 1$ disappear (they disappear when $\langle a^2 \rangle^* = \frac{1}{2}$ (see (33)) and the distribution of activities becomes a broad distribution with a maximum at a positive value of a.

It is interesting to note that the fixed distribution $Q_\infty(a)$ has another interpretation. One can replace the averages over the initial conditions by time averages. Until now, we have defined $a_i(t)$ as an average over initial conditions (see (12)).

One can also define $\tilde{a}_i$ as the activity averaged over time for a given initial condition $\{S_i(0)\}$

$$\tilde{a}_i = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} S_i(t) \xi_i^{(1)}. \tag{37}$$

As long as the thermodynamic limit ($N \to \infty$) is taken before the limit $T \to \infty$ in (37) or, more precisely, as long as $T \ll \log N$ (Derrida *et al* 1987) the argument leading to equation (17) about the absence of correlation between the inputs $j_1(i), \ldots, j_K(i)$ of almost all sites i remains valid and one can write for the $\tilde{a}_i$ an expression similar to (27):

$$\tilde{a}_i = \sum_{\sigma_{j_1} = \pm 1} \cdots \sum_{\sigma_{j_K} = \pm 1} \left(\prod_{r=1}^{K} \frac{1 + \sigma_{j_r} \tilde{a}_{j_r}}{2} \right) \operatorname{sgn}\left(\sum_{r=1}^{K} \xi_i^{(1)} \xi_{j_r}^{(1)} J_{ij_r} \sigma_{j_r} \right). \tag{38}$$

One can then define the probability distribution $\tilde{Q}(\tilde{a})$ of these activities

$$\tilde{Q}(\tilde{a}) = \frac{1}{N} \sum_{i=1}^{N} \delta(\tilde{a} - \tilde{a}_i). \tag{39}$$

The calculation is identical to the calculation of $Q_t(a)$ and one finds that

$$\tilde{Q}(\tilde{a}) = Q_\infty(\tilde{a}). \tag{40}$$

So $Q_\infty(a)$ is the distribution of activities when these are defined as averages over initial conditions. It is also the distribution of activities when the activities are defined as averages over time for one fixed initial condition.

If we come back to the possible distributions $Q_\infty(a)$, we see that for $Q_\infty(a) = \delta(a)$, all the neurons have an average over time $\tilde{a} = 0$. This means that all neurons flip all the time, probably in a random manner, spending half of their time $+1$ and half of their time -1.

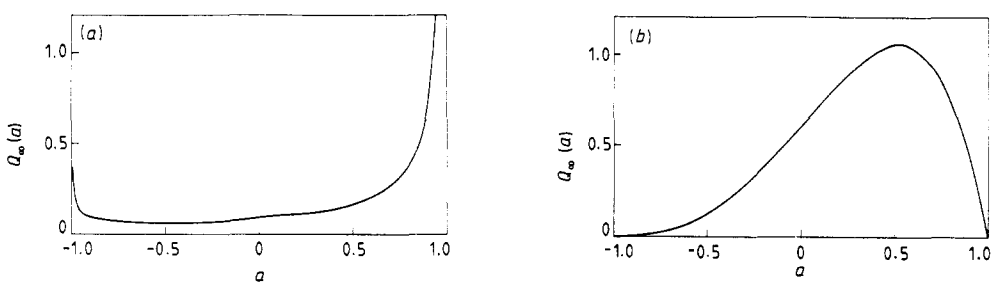

Figure 1. The distribution of the neural activities $Q_\infty(a)$ for (a) $\alpha = 0.4$ and (b) $\alpha = 0.6$.

For $\alpha < \alpha_c$, the attractive distribution (figure 1) $Q_\infty(a)$ is continuous with no delta function. This means that all neurons keep moving for ever. This is very different from what would happen for networks with symmetric synapses ($J_{ij} = J_{ji}$): at zero temperature, systems with symmetric interactions are always attracted by metastable states with all spins fixed (for sequential dynamics) or cycles 2 (for parallel dynamics). This means that $\tilde{Q}(\tilde{a})$ for a system with symmetric interactions is always the sum of delta functions at $+1$ and -1 (and 0 for parallel dynamics). So it is in the shape of the distribution $\tilde{Q}(\tilde{a})$ that the non-symmetry of the interactions is the most apparent.

The shapes of the $Q_\infty(a)$ in figure 1 corresponding to an attractor near a pattern are very different from the delta function corresponding to an attractor far from all patterns. So the activity of the neurons is very different when the system remembers a stored pattern and when it does not. This again is a major difference from systems with symmetric interactions for which these activities are the same (Parisi 1986).

Lastly one can relate the distribution $Q_t(a)$ defined in (20) to another interesting quantity (Derrida *et al* 1987) which is the overlap between two configurations. If one starts at $t = 0$ with two initial conditions $\{S_i(0)\}$ and $\{\tilde{S}_i(0)\}$, one can show (Derrida *et al* 1987) that the time evolution of their overlap

$$q(t) = \frac{1}{N} \sum_{i=1}^{N} S_i(t) \tilde{S}_i(t) \tag{41}$$

is given by

$$q(t+1) = -1 + \frac{2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp(-y^2)\, dy\, \operatorname{erf}\left(\left| \frac{y\sqrt{1+q(t)} + m_1(t)/\sqrt{\alpha}}{\sqrt{1-q(t)}} \right| \right) \tag{42}$$

when the two configurations have the same projection $m_1(t)$ on pattern 1. We see that since $m_1(t) = \langle a \rangle_t$, the time evolution of $q(t)$ is exactly the same as the time evolution of $\langle a^2 \rangle_t$ given in (35). This equality between $q(t)$ and $\langle a^2 \rangle_t$

$$q(t) = \int a^2\, da\, Q_t(a) = \langle a^2 \rangle_t \tag{43}$$

is due to the fact that if two initial configurations $\{S_i(0)\}$ and $\{\tilde{S}_i(0)\}$ are chosen at random according to (15), the probability that $S_i(t) = \tilde{S}_i(t) = \xi_i^{(1)}$ is $\frac{1}{4}(1 + a_i(t))^2$, the probability that $S_i(t) = -\tilde{S}_i(t) = \xi_i^{(1)}$ is $\frac{1}{4}(1 - a_i^2(t))$ and so on. Therefore one has:

$$\overline{S_i(t)\tilde{S}_i(t)} = a_i^2(t) \tag{44}$$

for almost all sites i, and this implies (43).

4. Finite projections on two patterns

The calculations carried out in § 3 can be repeated to study situations where the initial configuration $\{S_i(0)\}$ has finite projections on more than one pattern. One can then describe the activity of the neurons in mixed patterns and try to see whether the distribution of activities allows one to distinguish between single-pattern attractors and attractors corresponding to mixed patterns. In this section, an example of mixed patterns will be discussed and we will see that the distribution of activities is rather different to that corresponding to a single-pattern attractor.

Let us consider an initial configuration $\{S_i(0)\}$ with finite projections on two patterns $\{\xi_i^{(1)}\}$ and $\{\xi_i^{(2)}\}$ and zero projections on the $p-2$ other patterns. We assume that the two patterns $\{\xi_i^{(1)}\}$ and $\{\xi_i^{(2)}\}$ have a finite overlap Q

$$Q=\frac{1}{N}\sum_{i=1}^{N}\xi_i^{(1)}\xi_i^{(2)} \tag{45}$$

whereas the $p-2$ other patterns ($\{\xi_i^{(\mu)}\}$ for $\mu \geqslant 3$) are random and uncorrelated. If one defines the projections

$$m_1(t)=\frac{1}{N}\sum_{i=1}^{N}\xi_i^{(1)}S_i(t) \qquad m_2(t)=\frac{1}{N}\sum_{i=1}^{N}\xi_i^{(2)}S_i(t) \tag{46}$$

it was shown in equation (30) of Derrida *et al* (1987) that the time evolution of m_1 and m_2 is given for zero-temperature dynamics by

$$m_1(t+1)=\frac{1+Q}{2}\operatorname{erf}\left(\frac{m_1(t)+m_2(t)}{\sqrt{2\alpha}}\right)+\frac{1-Q}{2}\operatorname{erf}\left(\frac{m_1(t)-m_2(t)}{\sqrt{2\alpha}}\right) \tag{47}$$

$$m_2(t+1)=\frac{1+Q}{2}\operatorname{erf}\left(\frac{m_1(t)+m_2(t)}{\sqrt{2\alpha}}\right)-\frac{1-Q}{2}\operatorname{erf}\left(\frac{m_1(t)-m_2(t)}{\sqrt{2\alpha}}\right). \tag{48}$$

For $Q\neq 0$, there exist two thresholds $\alpha_c^{(1)}$ and $\alpha_c^{(2)}$, given by

$$\alpha_c^{(1)}=\frac{2}{\pi}(1+Q)^2 \qquad \alpha_c^{(2)}=\frac{2}{\pi}(1-Q)^2. \tag{49}$$

for $\alpha<\alpha_c^{(2)}$, each pattern has its own attractor (there is an attractive fixed point $0\neq m_1^*\neq m_2^*\neq 0$) and the mixed state is unstable. For $\alpha_c^{(2)}<\alpha<\alpha_c^{(1)}$, only the mixed state ($m_1^*=m_2^*\neq 0$) is stable. For $\alpha>\alpha_c^{(2)}$, the only attractive fixed point of (47) and (48) is $m_1^*=m_2^*=0$.

Following the calculations done in §§ 2 and 3 we have to introduce two functions $Q_t^{(1)}$ and $Q_t^{(2)}$ defined by

$$Q_t^{(1)}(b\eta)=P_t(b,\eta,\eta,\eta^{(3)},\ldots,\eta^{(p)}) \tag{50}$$

and

$$Q_t^{(2)}(b\eta)=P_t(b,\eta,-\eta,\eta^{(3)},\ldots,\eta^{(p)}). \tag{51}$$

$Q_t^{(1)}(b)$ represents the distribution of activity of the $N(1+Q)/2$ neurons i such that $\xi_i^{(1)}=\xi_i^{(2)}$ whereas $Q_t^{(2)}$ is the distribution of the activities of the $N(1-Q)/2$ neurons i such that $\xi_i^{(1)}=-\xi_i^{(2)}$. The calculation is a direct generalisation of what was done in § 3. Let us just discuss here the final result, which is analogous to (33):

$$Q_{t+1}^{(1)}(a)=\frac{1}{2}\left(\frac{1-\langle a^2\rangle_t}{\langle a^2\rangle_t}\right)^{1/2}\exp\left(z^2-\frac{[z\sqrt{1-\langle a^2\rangle_t}-(m_1(t)+m_2(t))/\sqrt{2\alpha}]^2}{\langle a^2\rangle_t}\right) \tag{52}$$

$$Q_{t+1}^{(2)}(a)=\frac{1}{2}\left(\frac{1-\langle a^2\rangle_t}{\langle a^2\rangle_t}\right)^{1/2}\exp\left(z^2-\frac{[z\sqrt{1-\langle a^2\rangle_t}-(m_1(t)-m_2(t))/\sqrt{2\alpha}]^2}{\langle a^2\rangle_t}\right) \tag{53}$$

where $a=\operatorname{erf}(z)$.

The time evolutions of $m_1(t)$ and $m_2(t)$ are given in (47) and (48) whereas $\langle a^2\rangle_t$ is defined by

$$\langle a^2\rangle_t=\frac{1+Q}{2}\int a^2\,\mathrm{d}a\,Q_t^{(1)}(a)+\frac{1-Q}{2}\int a^2\,\mathrm{d}a\,Q_t^{(2)}(a) \tag{54}$$

and evolves according to

$$\langle a^2\rangle_{t+1} = -1 + \frac{(1+Q)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp(-y^2)\, dy\, \mathrm{erf}\left(\left| \frac{y\sqrt{1+\langle a^2\rangle_t} + (m_1(t)+m_2(t))/\sqrt{\alpha}}{\sqrt{1-\langle a^2\rangle_t}} \right| \right)$$
$$+ \frac{(1-Q)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp(-y^2)\, dy\, \mathrm{erf}\left(\left| \frac{y\sqrt{1+\langle a^2\rangle_t} + (m_1(t)-m_2(t))/\sqrt{\alpha}}{\sqrt{1-\langle a^2\rangle_t}} \right| \right). \tag{55}$$

So, as in § 3, in the limit $C \to \infty$, one needs to iterate only three parameters $m_1(t)$, $m_2(t)$ and $\langle a^2\rangle_t$, to describe the distributions $Q_t^{(1)}$ and $Q_t^{(2)}$.

In the long-time limit, these distributions converge to fixed distributions $Q_\infty^{(1)}(a)$ and $Q_\infty^{(2)}(a)$ and again these distributions are identical to $\tilde{Q}^{(1)}(\tilde{a})$ and $\tilde{Q}^{(2)}(\tilde{a})$ where $\tilde{a}_i$ is the time average of $S_i(t)$.

In figure 2, the distributions $Q_\infty^{(1)}(a)$ and $Q_\infty^{(2)}(a)$ are shown for an attractor corresponding to mixed patterns ($Q = 0.2$, $\alpha = 0.7$ so that $\alpha_c^{(2)} < \alpha < \alpha_c^{(1)}$). We see that the distribution of activities of the neurons such that $\xi_i^{(1)} = \xi_i^{(2)}$ (figure 2(a)) looks rather similar to what it was for single-pattern attractors whereas the activities of the neurons such that $\xi_i^{(1)} = -\xi_i^{(2)}$ (figure 2(b)) look rather different. So again we see that the distribution of activities has rather different shapes for single-pattern attractors and mixed-patterns attractors.

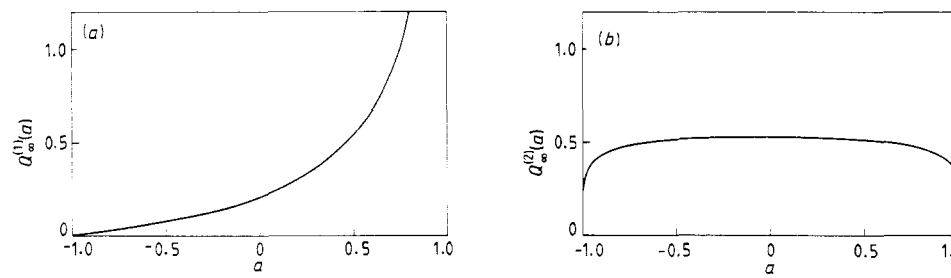

Figure 2. The distribution of the neural activities (a) $Q_\infty^{(1)}(a)$ and (b) $Q_\infty^{(2)}(a)$ for a mixture of two patterns which have an overlap $Q = 0.2$ for $\alpha = 0.7$. $Q_\infty^{(1)}(a)$ represents the neurons for which the two patterns are the same and $Q_\infty^{(2)}(a)$ represents the neurons for which the two patterns are opposite.

5. Conclusion

In this work, we have seen that the probability distribution of activities can be calculated analytically for asymmetrically diluted neural networks. These distributions are continuous and not delta functions as in the case of symmetric synapses. It should be possible to extend the calculations of the present work to other diluted models (Derrida and Nadal 1987, Gutfreund and Mézard 1988, Noest 1988), to layered networks (Meir and Domany 1987a, b, 1988, Meir 1988, Derrida and Meir 1988), to finite-temperature or to sequential dynamics (Derrida *et al* 1987) (for sequential dynamics it is probably sufficient to replace everywhere recursions in time $x_{t+1} = f(x_t)$ by differential equations $dx/dt = f(x_t) - x_t$). In particular it would be interesting to study systems with low activites (i.e. for which the distribution of activities is concentrated around -1) because they should be more relevant from a biological point of view.

From the point of view of the statistical mechanics of disordered systems, finding the distribution of activities is a problem very similar to the problem of finding the distribution of local magnetisation or local fields of spin glasses on Bethe lattices (Bowman and Levin 1982, Thouless 1986, Mottishaw 1987, de Oliviera 1988a, b), of diluted spin glasses (Viana and Bray 1985, Orland 1985, De Dominicis and Mottishaw 1987, Mézard and Parisi 1987, Kanter and Sompolinsky 1987, Katsura 1987) or other automata models (Derrida and Flyvbjerg 1987, Kanter 1988). It is interesting to dispose of a case like the one described here for which the distribution of activities is known exactly and has a rather simple analytic expression (33). It would be interesting to see how this distribution is changed when the connectivity decreases and to know whether the singularities observed by Derrida and Flyvbjerg (1987) in the case of automata with finite connectivity are also present in asymmetrically diluted neural networks. It would also be interesting to study how a non-zero fraction of symmetric bonds would modify the distribution of activities.

Acknowledgments

I am very grateful to D Amit, H Gutfreund and H Sompolinsky for their hospitality at the Hebrew University of Jerusalem where most of this work was done. I gratefully acknowledge also the Max Planck Institute für Mathematik at Bonn, West Germany, and the IBM Bergen Research Centre, Norway, where this work was finished.

I would like to thank Professor H Gutfreund for his comments on the manuscript.

References

Amit D, Gutfreund H and Sompolinsky H 1985a *Phys. Rev.* A **32** 1007
—— 1985b *Phys. Rev. Lett.* **55** 1530
—— 1987 *Ann. Phys., NY* **173** 30
Bowman D R and Levin K 1982 *Phys. Rev. B* **25** 3438
Coolen A C C and Ruijgrok Th W 1988 *Phys. Rev.* A **38** 4253
De Dominicis C and Mottishaw P 1987 *Europhys. Lett.* **3** 87
Derrida B and Flyvbjerg H 1987 *J. Phys. A: Math. Gen.* **20** L1107
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Derrida B and Meir R 1988 *Phys. Rev.* A **38** 3116
Derrida B and Nadal J-P 1987 *J. Stat. Phys.* **49** 993
Flyvbjerg H 1988 *J. Phys. A: Math. Gen.* **21** L955
Gutfreund H and Mézard M 1988 *Phys. Rev. Lett.* **61** 2351
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
—— 1984 *Proc. Natl Acad. Sci. USA* **81** 3088
Kanter I 1988 *Phys. Rev. Lett.* **60** 1891
Kanter I and Sompolinsky H 1987 *Phys. Rev. Lett.* **58** 164
Katsura S 1987 *Physica* **141A** 556
Kree R and Zippelius A 1987 *Phys. Rev.* A **36** 4421
Little W A 1974 *Math. Biosci.* **19** 101
Meir R 1988 *J. Physique* **49** 201
Meir R and Domany E 1987a *Phys. Rev. Lett.* **59** 359
—— 1987b *Europhys. Lett.* **4** 645
—— 1988 *Phys. Rev.* A **37** 608
Mézard M and Parisi G 1987 *Europhys. Lett.* **3** 1067
Mottishaw P 1987 *Europhys. Lett.* **4** 333
Noest A J 1988 *Europhys. Lett.* **6** 469

de Oliviera M J 1988a *Physica* **148A** 567
—— 1988b *Physica* **150A** 614
Orland H 1985 *J. Physique Lett.* **46** L763
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L675
Thouless D J 1986 *Phys. Rev. Lett.* **56** 1082
Viana L and Bray A J 1985 *J. Phys. C: Solid State Phys.* **18** 3037