Nuclear Physics B (Proc. Suppl.) 4 (1988) 673-677
North-Holland, Amsterdam

INTRODUCTION TO NEURAL NETWORK MODELS

B. DERRIDA

Service de Physique Théorique, CEN Saclay, 91191 Gif-sur-Yvette cedex, France

This talk is an introduction to neural network models which have been studied recently using methods of statistical mechanics.

1. INTRODUCTION

It would be a formidable task to review all the recent and interesting works on neural networks and this would not be very useful because several excellent reviews[1-6] have already been published. Therefore, in these few pages, I will only introduce the neural network models which are considered by most of the physicists of the field and mention a few recent results. My hope is that this will encourage the reader to go deeper into the subject[1-6].

One of the reasons why physicists (from statistical mechanics) are interested by the brain is rather obvious. The brain is composed by a large number N ($\simeq 10^{12}$) of neurons which interact through synapses (10^2-10^4 per neuron). It is therefore tempting to try to describe the properties of such a system by the technics of statistical mechanics.

The simplest neural network models which have been considered consist in assuming that the state of each neuron i at time t is represented by an Ising variable $S_i(t)$

$$S_i(t) = +1 \quad \text{if the neuron } i \text{ is firing} \qquad (1)$$

$$S_i(t) = -1 \quad \text{if the neuron } i \text{ is quiescent}$$

and that the synapsis J_{ij} between neuron j and neuron i is a real number ($J_{ij} > 0$ if the synapsis is excitatory and $J_{ij} < 0$ if it is inhibitory). In general the matrix J_{ij} is nonsymmetric ($J_{ij} \neq J_{ji}$) because the synapses are non symmetric. One then has to choose a dynamical rule to make the system evolve in time. A simple way consists in saying that at time t the neuron i receives a potential $V_i(t)$ given by

$$V_i(t) = \sum_j J_{ij} S_j(t) \qquad (2)$$

and that the state of neuron i at time $t+1$ depends on $V_i(t)$ in a probabilistic way

$$S_i(t+1)=+1 \quad \text{with probability} \quad f(V_i(t)) \qquad (3)$$

$$S_i(t+1)=-1 \quad \text{with probability} \quad 1-f(V_i(t))$$

$f(x)$ is an increasing function such that $f(x) \to 0$ if $x \to -\infty$ and $f(x) \to 1$ if $x \to +\infty$. For example

$$f(x) = \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{x}{T}\right) \qquad (4)$$

where T plays the role of a temperature.

Dynamics like (2-4) are commonly used in Monte Carlo simulations. The main difference with more usual spin models of Statistical Mechanics is that here the matrix J_{ij} is non symmetric. Therefore there is no hamiltonian, no partition function.

The first property of such neural network models is that they can memorize patterns by choosing properly the synapses J_{ij}. Assume that we have N neurons $Si = \pm 1$ and we want to store p patterns (ξ_i^μ) of N bits each

0920-5632/88/$03.50

$$\begin{array}{ll} 1^{st} \text{ pattern} & \xi_i^{(1)} = +1 \text{ or } -1 \quad 1 \leqslant i \leqslant N \\ \bullet & \\ \bullet & \qquad (5) \\ \bullet & \\ p^{th} \text{ pattern} & \xi_i^{(p)} = +1 \text{ or } -1 \quad 1 \leqslant i \leqslant N \end{array}$$

We will say that pattern $\{\xi_i^{(\mu)}\}$ is memorized if for the dynamics (2 -4) there is an attractor near this pattern. So the problem is to choose the J_{ij} in order to make the attractors as close as possible to the stored patterns. A simple way of measuring the distance between a spin configuration $\{S_i(t)\}$ and a pattern is to calculate their overlap

$$m_\mu(t) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{(\mu)} S_i(t) \qquad (6)$$

There exist several choices of the J_{ij} which give attractors in the neighbourhood of the stored patterns $\xi_i^{(\mu)}$. Some of these choices lead to interesting effects like short on long term memory, forgetting[12-13]. For the moment I will limit the discussion to one of the simplest rules (the Hebb rule[4]) which give an expression of the J_{ij} in terms of the patterns

$$J_{ij} = \frac{1}{C} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \qquad (7)$$

where C is the number of synapses of each neuron (for simplicity one can assume that C does not depend on i).

2. THE HOPFIELD MODEL[7,8,4]

As long as the matrix J_{ij} is non symmetric, it is not easy to use the methods of Statistical Mechanics (Partition function etc ...). The idea of Hopfield was to consider a simpler situation where

$$C = N - 1 \qquad (8)$$

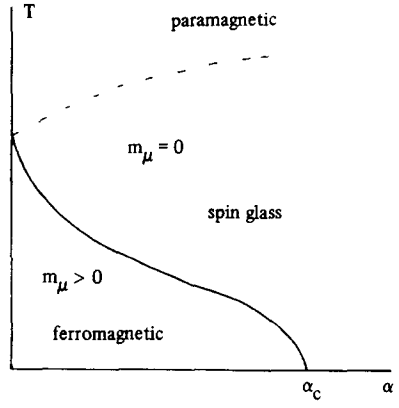
i.e. each neuron interacts with each other neuron and the J_{ij} are given by (7). Then one knows that with the dynamics (2-4) the system will evolve to an equilibrium described by an Hamiltonian $\mathcal{H}$ at temperature T

$$\mathcal{H}(\{S_i\}) = -\sum_{ij} J_{ij} S_i S_j \qquad (9)$$

i.e. each configuration $\{S_i\}$ is visited in the long time limit with a probability $P_{eq}(\{S_i\}) = \exp[-\mathcal{H}(\{S_i\})/T]$.

When one considers the J_{ij} given by the Hebb rule (7), the J_{ij} take both positive and negative values and phase space is composed of many valleys like in spin glass problems[9,10].

Amit, Gutfreund and Sompolinsky[4,11] have studied the equilibrium properties (the thermal equilibrium) in great detail using replica technics which had been developed previously in the study of spin glasses. They found a phase diagram with the following shape :



when the p patterns are chosen at random (i.e. : $\xi_i^{(\mu)} = +1$ or -1 with equal probability). The parameter α which appears in figure 1 is defined as the ratio of the number of stored patterns p divided by the number of synapses C per neuron (in the case of the Hopfield model C = N - 1 but this will not be the case for the diluted model discussed in section 3)

$$\alpha = p/C \qquad (10)$$

The important line is the phase boundary between the ferromagnetic phase and the spin glass phase. In the ferromagnetic phase, there exists a minimum in the free energy landscape with $m_\mu > 0$, i.e. one expects a valley near each pattern $\xi^{(\mu)}$. In the spin glass and the paramagnetic phases $m_\mu = 0$ and therefore this local minimum disappears. The phase dia-

gram obtained by Amit at al[4,11] has more structure than what is shown on figure 1 (transition line where the symmetry of replica is broken, spin glass phase, etc ...) but this will not be discussed here.

We see that with the Hebb rule (7) the system is able to memorize the patterns as long as the number of stored patterns $p = C\alpha$ does not exceed a certain value $\alpha_c(T)$. For $\alpha > \alpha_c(T)$, there is a complete deterioration and no pattern is memorized. It turns out[4,11] that the transition from the ferromagnetic phase to the spin glass phase is a first order transition and that m_μ has a jump. At T = 0, one finds that

$$\alpha_c \simeq .14$$

and m_μ jumps from a value $\simeq .95$ to 0. Since the fraction of wrong bits is given by $\frac{1-m_\mu}{2}$, we see that up to α_c the patterns are memorized with very few mistakes.

The calculations done on the Hopfield model can be generalized to various situations (see the reviews 4-5) by modifying or by replacing the Hebb rule (7) by other rules[12,13] (this can be used in particular to describe short or long term memory effects).

There are however several difficulties in the Hopfield model

(1) The calculations are done at equilibrium (using replica) but one does not know how to describe analytically dynamics.

(2)The symmetry of the synapses ($J_{ij} = J_{ji}$) is essential in this approach although the synapses are known to be non symmetric in the brain.

(3) All the neurons are connected (C = N -1)and that too is not realistic.

(4) The forgetting catastrophe : if $\alpha > \alpha_c$, i.e. the number of input patterns exceeds a maximal value, all the patterns are forgotten at once.

(5) Various difficulties when one extends the calculations to the case of correlated patterns.

(6) With symmetric interactions, one can expect minima in the free energy landscape but there is no way to memorize temporal sequences of patterns.

3. NON SYMMETRIC - DILUTED NETWORKS

It turns out that one can construct a neural network model with non symmetric synapses for which the dynamics can be solved exactly[14]. The model consists of N neurons $S_i = \pm 1$ and the synapses J_{ij} are given by

$$J_{ij} = C_{ij} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \qquad (11)$$

where the $\{\xi_i^\mu\}$ is the μ^{th} pattern and C_{ij} is a random number which represents the dilution

$$\begin{aligned} C_{ij} &= 1 \quad \text{with probability } C/N \\ C_{ij} &= 0 \quad \text{with probability } 1 - C/N \end{aligned} \qquad (12)$$

The C_{ij} and C_{ji} are independent random variables and therefore the matrix J_{ij} is no longer symmetric. The spin still evolve according to the following rules

$$\begin{aligned} S_i(t+1) &= +1 \quad \text{with prob} \quad p_i(t) \\ S_i(t+1) &= -1 \quad \text{with prob} \quad 1 - p_i(t) \end{aligned} \qquad (13)$$

where

$$p_i(t) = \frac{1}{2} + \frac{1}{2} \tanh\left(\sum_j J_{ij} S_j(t)/T\right) \qquad (14)$$

which gives in the low temperature limit

$$S_i(t+1) = \mathrm{sgn}\left[\sum_j J_{ij} S_j(t)\right] \qquad (15)$$

The situation for which this model can be solved is the limit $N \to \infty$, C being finite or infinite with the constraint that

$$C \ll \log N \qquad (16)$$

The reason why this condition makes the model soluble would be too long to explain here in detail[10,14]. Let me just say that because the system is very diluted (see eq. 12), the structure of the

network is locally a tree. At zero temperature, and for large C, one gets a simple expression for the time evolution of $m_\mu(t)$

$$m_\mu(t+1) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} dy\ e^{-y^2}\ \mathrm{sign}\left[m_\mu(t) - y\sqrt{2\alpha}\right] \tag{17}$$

where $\alpha = p/C$ (p is the number of stored patterns). The dynamics are fully described by the map (17). One sees from (17) that there is a critical value α_c of α

$$\alpha_c = 2/\pi \tag{18}$$

If $\alpha > \alpha_c$, the number of patterns memorized is too large and the only attractive fixed point of (17) is $m^*_\mu = 0$. The system does not remember anything.

If $\alpha < \alpha_c$, there appears an attractive fixed point $m^*_\mu \neq 0$ of (17) corresponding to the attractor near the pattern μ. One should notice that the retrieval is not perfect since $m^*_\mu \neq 1$ (the fraction of wrong bits is given by $(1 - m^*_\mu)/2$).

In the above calculation, the typical projection of one pattern μ on another pattern ν was $N^{-1/2}$:

$$\frac{1}{N} \sum_i \xi_i^{(\mu)}\ \xi_i^{(\nu)} \sim N^{-1/2} \tag{19}$$

for all pairs μ and ν. On can generalize it to describe other situations. If one considers that p patterns are random but that two of them (patterns 1 and 2) are correlated

$$\frac{1}{N} \sum_i \xi_i^{(1)}\ \xi_i^{(2)} = q \tag{20}$$

one can write[14] equations similar to (17) to describe the time evolution of $m_1(t)$and $m_2(t)$. Because of (20), the time evolution of $m_1(t)$ and $m_2(t)$ are coupled and one finds that there are two critical values of α :

$$\alpha_1 = \frac{2}{\pi}(1-q)^2 \qquad \alpha_2 = \frac{2}{\pi}(1+q)^2 \tag{21}$$

For $\alpha > \alpha_2$, the only fixed point is $m^*_1 = m^*_2 = 0$. Too many patterns have been stored. The system does not remember anything.

For $\alpha_1 < \alpha < \alpha_2$, there is an attractive fixed point $m^*_1 = m^*_2 \neq 0$. The system remembers patterns 1 and 2 but cannot distinguish them.

For $\alpha < \alpha_1$, there is an attractive fixed point $m^*_1 > m^*_2$. The system can distinguish the two patterns.

There are some limiting cases which can be easily understood.

If $q \to 0$, the patterns become uncorrelated and α_1 and $\alpha_2 \to \alpha_c$.

If $q \to 1$, $\alpha_1 \to 0$. If the 2 patterns become identical, it is impossible to distinguish them.

For diluted networks, one can extend the above calculation to describe more complex situations[15,17]. For example, one can produce short and long term memory effects by considering that the synapses J_{ij} are bounded ($|J_{ij}| < L$) and that adding a new pattern changes the synapsis only if the constraint $|J_{ij}| < L$ is satisfied. One can also choose the J_{ij} in order to produce temporal sequences of patterns[17].

4. CONCLUSION

The two models described in sections 2 and 3 have the following two simplifying features :

(1) - there is no architecture : all the neurons play similar roles

(2) - the synapses are given explicitly in terms of patterns

Recently it has been shown that none of these two symplifying assumptions is essential for the model to be soluble.

One can construct layered networks for which the dynamics can still be solved exactly[4,18]. The solution and the properties are similar to those of the diluted model.

One can also use J_{ij} which are no longer given explicitly in terms of the patterns (like in the Hebb rule (7)) but which are arbitrary with the condition that there are attractors near the stored patterns[19-20]. This might be a first

step in understanding the various learning rules which have been proposed and consist in using iterative procedures to modify the J_{ij} in order to create or to enlarge the basins of attraction near stored patterns[21-23].

References

1. E. Bienenstock, F. Fogelman Soulie and G. Weisbuch eds : Disordered Systems and Biological Organisation, Springer Verlag, Heidelberg 1986

2. P. Peretto and J.J. Niez, in ref.[1] and Biol. Cybern. 54, 53 (1986)

3. T. Kohonen, Self Organization and Associative Memory Springer Verlag Berlin 1984

4. D.J. Amit, H. Gutfreund, H. Sompolinsky, Ann. Phys. 173, 30 (1987)

5. E. Domany, 1987 to appear in J. Stat. Phys.

6. P. Rujan, Worshop on "Systems with learning and memory abilities" 1987

7. W.A. Little, Math. Bisoci. 19, 101 (1974))

8. J.J. Hopfield, Proc. Nat. Acad. Sci. USA 79, 2554 (1982)

9. see for example K. Binder and A.P. Young, Rev. Mod. Phys. 58, 801 (1986) for a review on spin glasses)

10. B. Derrida : "Dynamics of Automata, Spin Glasses and Neural Network models" lectures given at Noto (Sicily) 1987 and references therein

11. D.J. Amit, H. Gutfreund, H. Sompolinsky, Phys. Rev. Lett. 55, 1530 (1985), Phys. Rev. A32, 1007 (1985)

12. J.P. Nadal, G. Toulouse, J.P. Changeux and S. Dehaene, Europhys. Lett. 1, 535 (1986)

13. M. Mezard, J.P. Nadal, G. Toulouse, J. Phys. Paris 47, 1457 (1986)

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

15. B. Derrida, J.P. Nadal, J. Stat. Phys. 49, 993 (1987) in press

16. E. Gardner, J. Phys. A20, 453 (1987)

17. H. Gutfreund, M. Mezard, preprint 87, see also J. Buhman, K. Schulten Europhys. Lett. 4, 1205 (1987)

18. R. Meir, E. Domany, Phys. Rev. Lett. 59, 359 (1987), Europhys. Lett. 4, 465 (1987), preprints 87

19. E. Gardner, Europhys. Lett. 4, 1205 (1987)
E. Gardner, B. Derrida, J. Phys. A in press
E. Gardner, preprint

20. W. Krauth, M. Mezard, J. Phys. A 20, L745 (1987)

21. S. Diederich, M.. Opper, Phys. Rev. Lett. 58, 949 (1987)

22. E. Gardner, N. Stroud, D.J. Wallace, preprint 87

23. G. Pöppel, U. Krey, Europhys. Lett. 4, 979 (1987)