Learning within bounds and dream sleep

1987 J. Phys. A: Math. Gen. 20 L1299

(http://iopscience.iop.org/0305-4470/20/18/014)

J. Phys. A: Math. Gen. **20** (1987) L1299-L1303. Printed in the UK

LETTER TO THE EDITOR

Learning within bounds and dream sleep

T Geszti and F Pázmándi

Department of Atomic Physics, Eötvös University, H-1088 Budapest, Hungary

Received 23 October 1987

Abstract. In a bounded-synapses version of Hopfield's model for neural networks the quasienergy of a given memory, which is approximately equal to the depth of the corresponding energy well, is calculated exactly by treating the change of a synaptic strength on learning as a random walk within bounds. Attractors corresponding to stored memories are found to be considerably flattened before serious retrieval errors arise. This allows dream sleep to be interpreted as random recall and relearning of fresh strong memories, in order to stack them on top of weak incidental memory imprints of a day.

The Little-Hopfield model (Little 1974, Hopfield 1982) initiated an explosive development of physically motivated neural network models, aimed mainly at simulating the associative memory of the human brain and providing a means to transplant it into artificial intelligence. The Little-Hopfield model is based on a quasispin representation of the state of a neural network (McCulloch and Pitts 1943): memories to be stored are encoded into binary sequences (patterns) of quasispin variables $S_i = \pm 1$, physically realised as firing or non-firing neurons in a network ($i = 1, \ldots, N$ for N neurons). Such a sequence can be regarded as an N-component vector $\{S\}$. Patterns represented by such vectors are stored if they are made attractors of the spin-flip dynamics, which is governed by the signs of the sums $\Sigma_j J_{ij} S_j$, where the coupling constants J_{ij} are called synaptic strengths. The latter are of both signs, which endows the system with spin-glass-like properties; in particular, the possibility of having many attractors. According to Hebb's mechanism (Hebb 1949), turning given patterns into attractors can be achieved by appropriate modifications of the synaptic strengths J_{ij} ('learning algorithms').

The learning algorithm used by Hopfield has a major drawback (Hopfield 1982, Amit *et al* 1987): if the number p of stored patterns passes above a sharp bound $p_c = \alpha_c N$ where N is the number of fully interconnected neurons in the network and $\alpha_c \approx 0.14$, then the memory abruptly collapses and no stored information can be retrieved any more. This can be avoided if one makes the synaptic strength bounded ('learning within bounds'): $|J_{ij}| \leq A$ (Parisi 1986, Nadal *et al* 1986, also hinted at by Hopfield (1982)). Then freshly learned memories gradually erase the older ones, and the whole memory can be visualised as a stack with fresh patterns above, and with increasingly deteriorated older ones as we go downwards. A memory organised in this way is called a palimpsest (Nadal *et al* 1986).

The present letter gives some new insight into the change of attractors of the older memories. In particular, their flattening and its role in making them unstable is emphasised. The results of our analysis are used to offer a modified and perhaps more realistic version of a proposed explanation of dream sleep by Crick and Mitchison

0305-4470/87/181299+05$02.50

(1983) and Hopfield *et al* (1983) who suggested that dream sleep may serve to eliminate spurious (parasitic) memories, by means of (i) retrieving them by generating random initial states and letting the system relax to the closest attractor, and (ii) weakening them by *unlearning*, i.e. opposite-to-Hebbian synaptic changes. Although generation of highly chaotic and possibly random states in dream sleep seems to be real (Babloyantz *et al* 1985, Dvorak and Siska 1986, Mayer-Kress and Layne 1986), no evidence has been found for unlearning since 1983. Besides, spurious memories seem to be easy to escape by including thermal noise or smooth threshold conditions in the relaxation process.

We start with blank synapses of strengths 0 (actually, in view of ergodicity of the learning process, any symmetric matrix of initial strengths within the bounds would give the same results); then teach a new pattern $\{\xi\}$ ($\xi_i = \pm 1$, $i = 1, \ldots, N$) by the bounded synaptic modification algorithm

$$z = J_{ij}^{\text{old}} + N^{-1/2}\xi_i\xi_j$$
$$J_{ij}^{\text{new}} = \begin{cases} z & \text{if } |z| \leq A \\ J_{ij}^{\text{old}} & \text{otherwise.} \end{cases} \tag{1}$$

This is roughly (if A is a multiple of $N^{-1/2}$ then exactly) equivalent to Parisi's (1986) algorithm. Like the one used by Hopfield (1982), it gives rise to a symmetric matrix of synaptic strengths: $J_{ij} = J_{ji}$. This symmetry of the coupling constants, which is a biologically non-realistic feature of this class of models, allows one to define an energy-like function

$$E\{S\} = -\tfrac{1}{2}\sum_{i,j(i\neq j)} J_{ij}S_iS_j \tag{2}$$

of the configuration. Learning is achieved since the patterns taught approximately minimise $E\{S\}$ with respect to nearby configurations, i.e. become approximate centres of wells on the energy surface in the space of configurations. Memories stored in this way can be retrieved since under single-spin-flip dynamics at temperature $T = 0$ the system evolves towards such energy minima. With thermal noise ($T \neq 0$) spurious memories, represented by high-lying local minima, are escaped and only deep energy wells act as efficient attractors.

As observed by Parisi (1986) and Nadal *et al* (1986), the so-called retrieval probability, i.e. the probability of faithfully associating a stored pattern to the memory imprint it leaves, starts to drop as some $0.02N$ new patterns are taught subsequently to the network and becomes negligible after $0.07N$ new patterns. We want to demonstrate that the sooner this drop of retrieval probability becomes apparent, the probability of retrieving the same pattern from a *random* initial state already drops to practically nothing, due to a considerable loss of depth of the corresponding potential well. This feature is the basis of the explanation of dream sleep proposed below.

To demonstrate this property, we observe that as far as the retrieval probability of a given pattern $\{\xi^{(0)}\}$ is still close to unity after having taught t more patterns to the network, the corresponding potential well is essentially not displaced from the pattern, and its depth can be approximated by the mean value of the energy in the pattern configuration, $E_0(t) = \frac{1}{2}N(N-1)\bar{B}(t)$, where $\bar{B}(t)$ is obtained from the projection of any synaptic strength onto $\{\xi^{(0)}\}$,

$$B_{ij} = J_{ij}\xi_i^{(0)}\xi_j^{(0)} \tag{3}$$

by averaging over possible sequences of the stored patterns up to $\{\xi^{(0)}\}$ and t more, which determine J_{ij} through algorithm (1). This can be done exactly as follows. B_{ij},

according to algorithm (1), assumes values $N^{-1/2}s (s = -M, \ldots, +M$ where $M =$ integer part of $N^{1/2}A)$, with probabilities p_s over which the averaging has to be taken. Let the stored patterns $\{\xi^{(t)}\}$ be statistically independent random binary sequences, one being taught at each integer time. By algorithm (1) this implies a bounded random walk of B_{ij} on its $2M+1$ possible values, with the simple law of evolution

$$\begin{aligned} &p_s(t+1) = \tfrac{1}{2}(p_{s-1}(t) + p_{s+1}(t)) \qquad \text{for } |s| < M \\ &p_{\pm M}(t+1) = \tfrac{1}{2}(p_{\pm M}(t) + p_{\pm(M-1)}(t)). \end{aligned} \tag{4}$$

If teaching of patterns is started at some large negative initial time, then at, say, $t = -1$ we have a uniform distribution $p_s = (2M+1)^{-1}$ for all s. Then teaching the distinguished pattern $\{\xi^{(0)}\}$ at $t = 0$ turns this into

$$\begin{aligned} &p_s(0) = (2M+1)^{-1} \qquad \text{for } |s| < M \\ &p_{-M} = 0 \qquad p_M = 2(2M+1)^{-1}. \end{aligned} \tag{5}$$

Now it is straightforward to solve equation (4) by numerical iteration to $t > 0$ and calculate the mean value $\bar{B}(t)$ of B_{ij}, from which $E_0(t)$ can be obtained (see the full curve in figure 1).

We have repeated Parisi's (1986) simulations for $N = 200$ neurons to evaluate the final energy to which the system relaxes at $T = 0$ from the initial configuration $\{\xi^{(0)}\}$, subsequent to which t more patterns had been memorised. As seen from figure 1, a 30% reduction of the energy depth, i.e. a considerable flattening of the corresponding

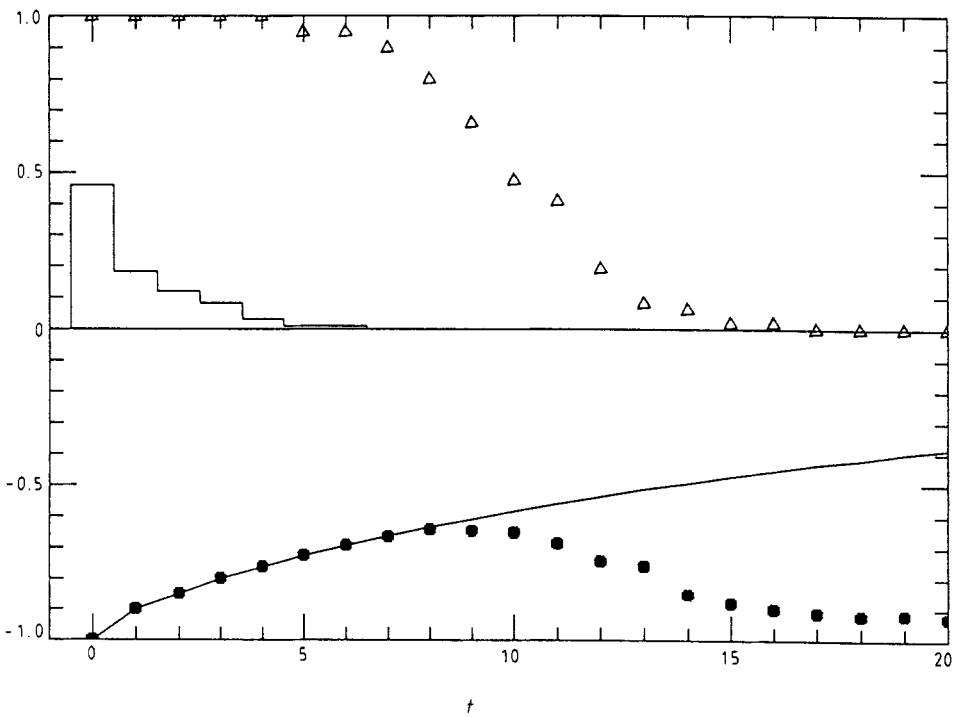

Figure 1. Forgetting for 200 neurons. Upper part: reduction of the probability of retrieval from the pattern $\{\xi^{(0)}\}$ (△, simulations by Parisi (1986)) and from a random initial configuration (histogram). Lower part: reduction of $|E_0(t)$ (full curve) and mean final energy after relaxation from a pattern after which t more had been taught (■).

potential well is observed before the final state (full squares) begins to deviate from the initial one (full curve) to any considerable extent.

The loss of depth of the ageing memory is accompanied by a drastic reduction of the probability of being retrieved from a random initial configuration. Our corresponding simulation results are shown in the histogram of the upper half of figure 1. For comparison we include Parisi's data (open triangles) on the 'retrieval probability' defined as the probability of 98% agreement between pattern and minimum energy configurations.

Let us now discuss briefly what happens to older memories for which the retrieval probability has been considerably reduced. As seen in the lower-right part of figure 1, starting from such long-ago-learned patterns the system relaxes to remote final configurations of deeper energy values. In some 86% of the cases these final configurations can be identified as some of the fresh memories, against which the older ones become unstable, letting the flow leak into the deep energy wells before the old patterns might get noisy. In this picture the fact that for large N the retrieval probability is a scaling function of the variable $x = t/N$ is explained by the diffusion constant $D = 1/N$ of the random walk of the synaptic strength. More details about this will be published subsequently. It should be mentioned that a random walk treatment similar to ours has been applied to a solvable model with asymmetric diluted synapses by Derrida and Nadal (1987).

The distinction between associative recall of the stored patterns, controlled by leakage from wells that just keep the flow or not, and retrieval from random initial states, in which nothing but the freshest and deepest energy wells are competitive, is the basis of our modified interpretation for dream sleep. We suggest that a possible aim of sleeping dreams is to eliminate unintentionally stored *weak* memory imprints of whatever you see and hear during the day. These have shallow energy minima but can be retrieved associatively (starting sufficiently close to them), and they occupy much of the retrievable capacity of the memory. On the contrary, what you *wanted* to learn leaves deep energy wells.

The elimination of weak incidental memories can then happen by (i) generating random initial states from each of which the system relaxes to one of the strong fresh memories, and (ii) a memory retrieved in this way is *relearned* (not unlearned!) by the usual Hebbian mechanism and thereby put on top of the stack above those more recent but much weaker accidental memories which are not recalled nor relearned during this sequence of events. Therefore such weak memories are subject to forced 'ageing' and are eliminated from the memory sooner than would happen without this (conjectured) action of dream sleep.

The proper context for this mechanism may be a filtering of what would be transferred from a medium-term (about one day) memory to long-term storage. A prediction perhaps testable in real-life experiments (Gardner-Medwin 1987) is the weakening of weak memories during the sleeping period.

It is a pleasure to thank J Kertész, I Kondor and T Tél for helpful comments and suggestions on a preliminary version.

References

Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
Babloyantz A, Salazar J M and Nicolis C 1985 *Phys. Lett.* **111A** 152

Crick F and Mitchison G 1983 *Nature* **304** 111
Derrida B and Nadal J P 1987 *J. Stat. Phys.* in press
Dvorak I and Siska J 1986 *Phys. Lett.* **118A** 63
Gardner-Medwin A R 1987 private communication
Hebb D O 1949 *The Organization of Behaviour* (New York: Wiley)
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
—— 1984 *Proc. Natl Acad. Sci. USA* **81** 3088
Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158
Little W A 1974 *Math. Biosci.* **19** 101
Mayer-Kress G and Layne S 1986 *Proc. Conf. on Perspectives in Biological Dynamics and Theoretical Medicine, Bethesda, MA* (New York: New York Academy of Sciences) to be published
McCulloch W S and Pitts W 1943 *Bull. Math. Biophys.* **5** 115
Nadal J P, Toulouse G, Changeux J P and Dehaene S 1986 *Europhys. Lett.* **1** 535
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617