Memory maintenance in neural networks

1987 J. Phys. A: Math. Gen. 20 L1305

(http://iopscience.iop.org/0305-4470/20/18/015)

J. Phys. A: Math. Gen. **20** (1987) L1305-L1309. Printed in the UK

LETTER TO THE EDITOR

Memory maintenance in neural networks

S Shinomoto

Department of Physics, Kyoto University, Kyoto 606, Japan

Received 14 October 1987

Abstract. A rule of synaptic modification in neural networks is proposed under a principle which minimises 'free energy' with synaptic strength roughly bounded. The rule does not allow the overloading of memories which can be a cause of failure in memory process and it stabilises the synaptic connection, provided that memories are properly stored.

Interest in network models of memory functions in the brain has concentrated upon suitabale forms of information storage and information retrieval. The forms of information storage may be classified into two groups, localised and extended, in which each bit of information either corresponds to a firing of a specific neuron or to a firing pattern of a set of neurons (see Hopfield 1982). There have also been several proposals for methods of recording information for each of the localised or extended storage forms. We thus have several candidates for memory functions among neural network models. A unit qualified as a memory should be able to *record*, *maintain* and *retrieve* environmental information. The recording of information in the above-mentioned forms is due to the plasticity of synaptic couplings. On the other hand, each synaptic coupling should be fixed before the system overloads the memories and some catastrophic deterioration (Hopfield 1982, Amit *et al* 1985, 1987) occurs. For this purpose, one might introduce an external fixation unit for the synaptic couplings. But how does the external unit know the stage of storage in the memory process? The task would be complex. There might be some intrinsic mechanism such that a network spontaneously ceases to alter its synaptic strength. I have tried to find a plasticity rule in line with the latter hypothesis for the extended form of information storage. In the present letter, I will present an *ad hoc* principle which meets the hypothesis.

Both neuronic states and synaptic couplings are assumed to be variable in time. It is physiologically plausible, however, to suppose that a timescale of synaptic modification is sufficiently large compared to that of neuronic modification. This separation of timescales allows us to employ the adiabatic approximation (Caianiello 1961, Takeuchi and Amari 1979) in which synaptic connection is regarded as permanent during short-term dynamics of neurons. Against the hierarchy in which synaptic connection dominates neuronic dynamics, I will introduce a feedback, where the resulting neuronic states adiabatically modify the synaptic connection which rule themselves. The self-organisation of systems composed of two different species with comparative difference in control is also of current interest.

Although some dynamical characteristics of networks of permanent asymmetric connection have been clarified (Amari 1971, 1974, Shinomoto 1986, 1987, Sompolinsky and Kanter 1986, Amit 1987) measure-theoretic knowledge is as yet insufficient for the present purpose. I will restrict the present investigation within symmetric connections

0305-4470/87/181305+05$02.50

to rest on a knowledge of their equilibrium properties (Amit *et al* 1985, 1987). The advantage of this restriction is the existence of a unique characteristic function which is called the energy or

$$E = -\tfrac{1}{2}\sum_i \sum_j K_{ij} s_i s_j \tag{1}$$

where K_{ij} is the synaptic connection from j to i. The connection is assumed to be symmetric, $\{K_{ij} = K_{ji}\}$, with $\{K_{ii} = 0\}$. Each neuron is supposed to be a binary element whose representation is symmetric: $s_i = +1$ (firing) or -1 (resting). I will introduce the stochastic threshold rule (Little 1974) which is to readjust each state iteratively by a probability $p(s_i) = 1/[1+\exp(\Delta E_i / T)]$, where $\Delta E_i = s_i(\Sigma_j K_{ij} s_j)$. Here, T is 'temperature', indicating the degree of noise against downhill motion in energy surfaces. An invariant measure or an equilibrium distribution function of the state $\boldsymbol{s} = (s_1, \ldots, s_n)$ for this rule is $\rho(\boldsymbol{s}) \propto \exp[-E(\boldsymbol{s})/T]$. A standard scheme of the extended form of information storage is the autocorrelation matrix memory or the Hopfield model (Hopfield 1982) whose connection is

$$K_{ij} = C_{ij} \equiv (1/\sqrt{M}) \sum_m s_i^m s_j^m \tag{2}$$

where $\boldsymbol{s}^m = (s_1^m, \ldots, s_N^m)$ is a firing pattern of the mth memory and M is the number of memories ($m = 1, \ldots, M$). I have chosen here a numerical factor $1/\sqrt{M}$ in order to make the coupling strength $|C_{ij}|$ independent of the number of memories, which are given independent of each other. There are global attractor basins for each of the memory patterns, provided that the number of memories M is sufficiently small compared to the number of elements N and the temperature T is sufficiently small compared to $\sqrt{M}/N$. Detailed characteristics were clarified by the statistical mechanical study by Amit *et al* (1985, 1987) for the system of an infinite number of elements ($N \to \infty$). A network of a large but finite number of elements ($N \gg 1$) keeps the characteristics of the infinite system.

The most elementary rule of synaptic modification needed to construct a connection similar to the Hopfield model is the generalised Hebb rule:

$$\Delta K_{ij} = s_i^m s_j^m \tag{3}$$

where $\boldsymbol{s}^m$ is a pattern to be acquired by the network. The complete autocorrelation matrix is obtained if the acquisition begins from a *tabula rasa* $\{K_{ij} = 0\}$ and it ceases at a reasonable stage, $1 \leqslant m \leqslant M \ll N$. In the absence of an external fixation unit, however, the system overloads the memories, $M \sim N$, and then catastrophic deterioration of memories (Amit *et al* 1985, 1987) occurs. In order to avoid the deterioration and to keep each synaptic strength bounded, palimpsest schemes (Nadal *et al* 1986, 1987, Parisi 1986, Mézard *et al* 1986) were proposed. A representative form is

$$\Delta K_{ij} = -\gamma K_{ij} + s_i^m s_j^m. \tag{4}$$

The system has a kind of steady state in the continual presentation of patterns. The statistical mechanics of the system in an asymptotic regime was solved by Mézard *et al* (1986) to reveal the existence of the threshold value γ_c such that the system acquires a stationary capacity for $\gamma > \gamma_c$. The palimpsest models are discussed by Nadal *et al* (1987) in relation to the behaviour of human short-term memory.

Rules such as (3) and (4) by themselves are not able to maintain the acquired memories for a long time. The mechanism of information storage in human long-term

memory, which may even last throughout a lifetime, is of current interest. I introduce here a minimisation procedure of an *ad hoc* cost function. For a choice of the cost function, I take the following plausible assumptions into account:

(i) the strength of synaptic coupling $|K_{ij}|$ is roughly bounded by a physiological structural constraint;

(ii) the rule of synaptic modification does not depend on the number of temporarily stored memories; and

(iii) information for modification of each of the synaptic couplings is only available locally.

I will choose cost functions for each of assumptions (i) and (ii). The choice will be checked *a posteriori* by examining the plasticity rule deduced from the sum of the cost functions in the light of assumption (iii).

A possible form of a cost function in relation to (i) may be $U = \frac{1}{2}\Sigma_i \Sigma_j |K_{ij}|^2$. The one in relation to (ii) may be a general characteristic function of networks of symmetric connections. Any functional of $E(s)$ can be a candidate for the latter cost function. Almost all functionals, however, do not satisfy assumption (iii), i.e. the resultant plasticity rules are not local. An exceptional functional which will satisfy criterion (iii) is the free energy $F = -T \ln(\mathrm{Tr} \cdot \exp(-E/T))$, where $\mathrm{Tr}\cdot$ represents the summation over all patterns of $\{s_i\}$. The total cost function is the sum of them, $\Gamma = F + \gamma U$, where γ is the parameter indicating the relative weight between F and U. The variation of the total cost function with regard to the coupling is

$$\Delta\Gamma = -\Delta K_{ij}[\langle s_i s_j \rangle_K - \gamma K_{ij}] \tag{5}$$

where $\langle \ldots \rangle$ represents a thermal average in the system characterised by temporal values of $\{K_{ij}/T\}$. I will take a unit $T = 1$, without loss of generality. A minimisation procedure of the cost function Γ provides a rule of synaptic modification:

$$\Delta K_{ij} = -\gamma K_{ij} + \langle s_i s_j \rangle K. \tag{6}$$

The rule (6) keeps a local form which meets assumption (iii). The above rule, seemingly similar to (4), however, has a completely different meaning because the second term in (6) is not simply a presented pattern, but an average over 'autonomous' states organised by the stochastic rule implemented by the temporal connection itself. I will assume that the network is also subjected to *infrequent renewal* in which neuronic states are clamped at a pattern of environmental information. The relaxation process after the clamping corresponds to the information processing. It would be possible to choose the frequency of clamping sufficiently small to keep the ergodicity in which the ensemble average can be replaced by the time average. Furthermore, the renewals would play a part in avoiding the non-ergodicity which will be mentioned later.

By a suitable choice of the parameter γ, the cost function Γ may have a number of local minima in a $(N(N-1)/2)$-dimensional space of the variables $\{K_{ij}\}$. To get a glimpse of this, I will show that couplings similar to the Hopfield model $\{C_{ij}\}$ for a given set of memories may actually be stable solutions of (6), provided that $\alpha = M/N \ll 1$.

An introduction of a new set of variables $\{B_{ij}\}$ by $K_{ij} = B_{ij}C_{ij}$ leads equation (6) to

$$\Delta B_{ij} = -\gamma B_{ij} + \langle s_i s_j \rangle_{BC} / C_{ij} \tag{7}$$

where each C_{ij} defined by equation (2) is supposed non-vanishing. Let us investigate a case where temporal values of multipliers $\{B_{ij}\}$ are uniform, or $\{B_{ij} = B\}$. Then the multiplier B is regarded as an inverse of the temperature for a system of the connection $\{C_{ij}\}$ and one can estimate the correlation $\langle s_i s_j \rangle$ by making use of the statistical

mechanical knowledge from Amit *et al* (1985, 1987). The system has three characteristic phases in parameter space spanned by α and B. The first is a paramagnetic phase characterised by ergodicity and the absence of order. In this regime bounded by $B < B_g$, the correlation $\langle s_i s_j \rangle$ vanishes. The threshold B_g as a function of α is $B_g = \sqrt{\alpha}[(1+\sqrt{\alpha})\sqrt{N}]^{-1}$ in the present unit of $\{C_{ij}\}$. The second is a spin-glass phase in $B > B_g$. In this phase the system is non-ergodic and may have an exponential number of local equilibria. The third, which I will call a retrieval phase, is characterised by global attractor basins for each of the memory patterns. The retrieval phase is localised in the parameter region $\alpha \ll 1$ and $B > B_c \sim O(\sqrt{(\alpha/N)})$, and the phase transition from the glass phase is first order. Ergodicity missed in these glass and retrieval phases can be recovered by the previously mentioned infrequent renewals of the neuronic states by external inputs. If there is no specific correlation in the input patterns, $\langle s_i s_j \rangle$ is also expected to be exponentially small in the glass phase because there may be large amounts of uncorrelated local equilibria. In the retrieval phase, the correlation $\langle s_i s_j \rangle$ is proportional to the average of $s_i s_j$ of all the memory patterns. Especially in a limit $B \gg B_c$, we may approximate the correlation by $(1/M)\,\Sigma_m\, s_i^m s_j^m = (1/\sqrt{M})C_{ij}$. The following are discussions on equation (7) in two limiting cases of α values.

In the case $\alpha \ll 1$, the second term in (7) as a function of B is vanishing for $B < B_c$ and is of $O(1/\sqrt{M})$ for $B > B_c$, where B_c is of $O(\sqrt{M}/N)$. The rule (7) provides a non-trivial stable fixed point $B = B_f \neq 0$, if the parameter γ is chosen as $\gamma \ll N/M = \alpha^{-1}$ (see figure 1). Note that the inequality of γ for the proper storage is opposite to the one for the rule (4). Irregularity of the second term of (7) for each (i, j) due to some inaccuracy of the memories would be present but would not cause a drastic change of the stability. The existence of a non-trivial fixed point implies a tremendous number of local minima in the cost function Γ, because memory patterns (2) can be chosen almost arbitrarily under rough orthogonality. Rule (7) also has a trivial stable fixed point, $B = 0$. Separation of two attractor domains implies that the desired coupling can be made up from a sufficiently large synaptic strength but *not* from a *tabula rasa*, $\{K_{ij} = 0\}$.

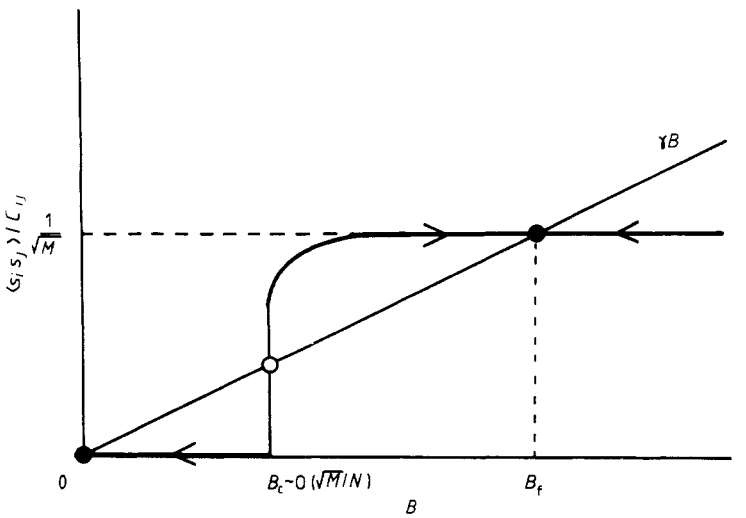

Figure 1. Schematic representation of the second term in (7) in the case $\{B_{ij} = B\}$ and $\alpha = M/N \ll 1$. The cross sections of the curve by a line γB are fixed points of (7).

In contrast to the above-mentioned case, the network with $\alpha \geqslant 1$ does not have a retrieval phase for any value of B. The system is subjected only to the glass transition at $B_g \sim O(1/\sqrt{N})$. We thus have only one fixed point, $B = 0$. Thus the Hopfield coupling with memories overloaded is unstable and each synaptic coupling is expected to decay monotonically.

The non-ergodicity and exponentially large number of local equilibria in the glass phase of C_{ij} with $\alpha \geqslant 1$ or the randomly connected network, however, may be utilised for a construction of a proper memory with regard to external inputs. In this phase, there may be a local equilibrium sufficiently close to an arbitrary input pattern. The state of the system is temporarily locked in a basin of the local equilibrium until tne next input arrives and while synaptic couplings are modified so as to deepen the basin. Thus, the system with strong and random couplings develops to record the input patterns. There may be a suitable frequency of presentation of the external information for this purpose. The problem in the learning stage is beyond the scope of the present study. The Boltzmann machine learning procedure (Ackley *et al* 1985, Sejnowski *et al* 1986) which is not 'autonomous' as for the present rule would be worth considering.

I have thus presented the rule of synaptic modification whose stable solutions are the Hopfield couplings. I recently noticed that an evolution equation similar to (6) had been proposed by Suzuki (1984) in another context, without any prediction of its solutions. In spite of the present findings of a set of solutions, the whole structure of the cost function is not yet revealed. There might be a set of stable solutions other than the ones presented in this letter. Findings from the other set of solutions would imply other forms of information storage.

I would like to express my sincere thanks to Y Kuramoto, H Sakaguchi, K Nemoto, T Ikegami, P C Davis, S Amari, K Fukushima and R Hecht-Nielsen for informative comments.

References

Ackley D H, Hinton G E and Sejnowski T J 1985 *Cognitive Sci.* **9** 147
Amari S 1971 *Proc. IEEE* **59** 35
—— 1974 *Kybern.* **14** 201
Amit D J 1987 *Preprint*
Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007
—— 1987 *Ann. Phys., NY* **173** 30
Caianiello E R 1961 *J. Theor. Biol.* **2** 204
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2550
Little W A 1974 *Mat. Biosci.* **19** 101
Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457
Nadal J P, Toulouse G, Changeux J P and Dehaene S 1986 *Europhys. Lett.* **1** 535
Nadal J P, Toulouse G, Mézard M, Changeux J P and Dehaene S 1987 *Computer Simulation in Brain Science* ed R R J Cotterill (Cambridge: Cambridge University Press)
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617, L675
Sejnowski T J, Kienker P K and Hinton G E 1986 *Physica* **22D** 260
Shinomoto S 1986 *Prog. Theor. Phys.* **75** 1313
—— 1987 *Biol. Cybern.* **57** 197
Sompolinsky H and Kanter I 1986 *Phys. Rev. Lett.* **57** 2861
Suzuki M 1984 *Prog. Theor. Phys. Suppl.* **79** 125
Takeuchi A and Amari S 1979 *Biol. Cybern.* **35** 63