Tome 46 N° 14 15 JUILLET 1985

LE JOURNAL DE PHYSIQUE - LETTRES

J. Physique Lett. **46** (1985) L-623 - L-630 15 JUILLET **1985,** PAGE L-623

Classification
Physics Abstracts
05.50 — 75.40 — 87.10

Scaling laws for the attractors of Hopfield networks

G. Weisbuch

Laboratoire de Physique de l'Ecole Normale Supérieure, 24, rue Lhomond, 75231 Paris, France

and F. Fogelman-Soulié

Laboratoire de Dynamique des Réseaux, 1, rue Descartes, 75005 Paris
and Paris V University

(*Reçu le 28 mars 1985, accepté sous forme définitive le 29 mai 1985*)

Résumé. — Les réseaux d'automates à seuil sont des systèmes dynamiques à structure aléatoire semblables aux verres de spins dont J. Hopfield a proposé l'application comme mémoires associatives. Nous établissons les lois d'échelles reliant le nombre maximum d'attracteurs utiles et la distance d'attraction, au nombre des automates du réseau. Notre approche permet aussi un meilleur choix des seuils, ce qui double les performances du réseau en nombre d'attracteurs.

Abstract. — Networks of threshold automata are random dynamical systems with a large number of attractors, which J. Hopfield proposed to use as associative memories. We establish the scaling laws relating the maximum number of « useful » attractors and the radius of the attraction basin to the number of automata. A by-product of our analysis is a better choice for thresholds which doubles the performances in terms of the maximum number of « useful » attractors.

J. Hopfield [1] has recently renewed the interest for networks of threshold automata (further abbreviated as NTA) by proposing their use as content addressable memories. The basis of the device is to encode the information to be retrieved as a sequence of binary signals, 0 or 1, which are taken as states of threshold automata. A sequence of bits is considered as the state of a NTA and Hopfield's device, explained further, is based on a formula allowing to build the network such that a set of predefined sequences, referred to as the reference sequences, are among the attractors of the dynamics of the network. Furthermore, if sequences which differ by a few bits from a reference sequence are taken as initial states, the NTA converges with a high probability towards the reference.

This property of convergence exists only if the number of reference sequences is not too large. J. Hopfield gives some data based on simulations for networks of 30 and 100 automata and

proposes as a limit of m, the number of reference sequences, a ratio of 0.15 to the number of automata n. Simulations done by Peretto [2] for larger n (500) give less optimistic results.

As we further explain, NTA are formally equivalent to spin glasses at zero temperature. Such systems are known to have a large number of attractors, which also includes « spurious attractors » as opposed to the reference attractors. These attractors and the relations between NTA and spin glasses are fully discussed in Peretto [2] and Amit [3] where a more comprehensive bibliography can be found.

However, the purpose of this paper is limited to the « useful attractors » : we shall derive the scaling laws relating m to n and test them against numerical simulations. Furthermore we give the expression for the probability that a sequence differing by d states from a reference sequence converges towards the reference. Finally the discussion of these scaling laws gives indications about the best choice for thresholds.

1. Definitions.

A threshold automaton i is a device which receives inputs j and computes its state at each discrete time step from the binary input signals x_j according to the following expression :

$$x_i(t) = Y\left(\sum_j a_{ij}\, x_j(t-1) - b_i\right) \tag{1.1}$$

where x_i and x_j are binary variables 0 or 1,

$Y(x)$ is the Heaviside function, taking value 0 if x is negative and 1 otherwise.

The a_{ij} are called the synaptic weights of the i-j link, b_i is the threshold for automaton i.

A network of threshold automata is built by connecting a set of automata, the state of each automaton being an input signal for the connected automata.

We are interested in particular NTA's such that :

— all inputs are coming from other automata, $j \neq i$;

— the graph of connections is complete : every automaton is connected with all the others;

— the connections are symmetrical $a_{ij} = a_{ji}$.

In order to completely define the dynamics of the network, an iteration mode has to be chosen : either parallel, when all automata are simultaneously changed, or sequential, when the new state of only one automaton is computed at each time step. Dynamical properties generally depend upon this choice. In fact, we shall see that this is not the case for the results of sections 2 and 3.

Computer simulations show that after a rather short transient time, starting from any initial distribution of states, the configuration of states evolves towards attractors which are fixed points, at least for sequential iterations.

(The condition $a_{ij} = a_{ji}$ ensures that the attractors are of period 1.)

The use of NTA as associative memories is based on the possibility to build the connection matrix A such that given configurations are attractors. If these configurations are vectors $x^1, \ldots, x^s, \ldots, x^m$, the a_{ij} are computed by

$$a_{ij} = \sum_s (2\, x_i^s - 1)(2\, x_j^s - 1) \tag{1.2}$$

which is written with a simplifying notation

$$a_{ij} = \sum_s X_i^s\, X_j^s \tag{1.3}$$

where

$$X_i^s = (2\, x_i^s - 1)\,.$$

The x_i being 0 or 1, the X_i are -1 or $+1$.

At this stage the formal equivalence between NTA and spin glasses can be explicited : the a_{ij} can be interpreted as exchange terms between spins X_i and X_j, and the thresholds as random local magnetic fields. The equivalence is complete if one chooses random sequential iteration mode, when automata are randomly selected for the application of the transition rule. A similar model proposed by Little [4] to model neural networks has some features of spins glasses at finite temperature, but uses parallel iteration.

2. Invariance of the reference sequences.

Let us recall the conditions imposed by the invariance of the reference sequences on the threshold. If in sequence s automaton i is in state 1, it remains in 1 if :

$$b_i \leqslant \sum_{j \neq i} a_{ij} x_j^s . \tag{2.1}$$

if x_i^s is in state 0, stability requires :

$$b_i > \sum_{j \neq i} a_{ij} x_i^s . \tag{2.2}$$

Since such conditions have to be verified by threshold b_i for all the different sequences, b_i must then belong to an interval defined by :

$$\text{Max} \left\{ \sum_j a_{ij} x_j^s \right\} < b_i \leqslant \text{Min} \left\{ \sum_j a_{ij} x_j^s \right\} \tag{2.3}$$

where the maximum is taken for all s such that $x_i^s = 0$, and the minimum for all s such that $x_i^s = 1$.

When the number of sequences is increased, it may happen that the minimum on the right is smaller than the maximum on the left and the interval to choose b_i does not exist anymore [5].

Replacing the a_{ij} by their expressions as functions of the X_i^s one obtains for the right hand side inequality :

$$b_i \leqslant \sum_{s'} \sum_j X_i^{s'} X_j^{s'} x_j^s \tag{2.4}$$

which can be decomposed into :

$$b_i \leqslant \sum_{j \neq i} x_j^s + \sum_{s' \neq s} \sum_{j \neq i} X_i^{s'} X_j^{s'} x_j^s \tag{2.5}$$

by taking out the term $s' = s$.

In a similar way for $x_i^s = 0$ one obtains :

$$b_i > - \sum_{j \neq i} x_j^s + \sum_{s' \neq s} \sum_{j \neq i} X_i^{s'} X_i^{s'} x_i^s . \tag{2.6}$$

We now introduce two fundamental assumptions about the reference sequences :

First they are random collections of zeroes and ones, which implies that the $\sum x_j^s$ term is large and close to $(n - 1)/2$. We call it further the polarized term.

Second, the sequences are not correlated with each other, and the double sums (called further the random sums) are random collections of 0, + 1 and − 1 terms. The result of the summation is then the same as that of a random walk with $(n - 1)(m - 1)/2$ steps.

The factors 1/2 come from the fact that half x_j^s are zeroes.

For symmetry reasons $b_i = 0$ is a good first guess. We can then evaluate the probability that one condition of invariance for one automaton is satisfied : it is the probability that the random sum is inferior to the polarized term. (The probability is the same for symmetry reasons for the two cases $x_i^s = 0$ or 1.)

For large n and m we approximate the binomial distribution by a Gaussian. The probability that after p steps the abscissa of the end of a random walk is smaller than x_0 is given by :

$$P(x < x_0) = \text{erf}\,(x_0/\sqrt{p}) \tag{2.7}$$

where erf, the error function is defined by :

$$\text{erf}\,(x_0/\sqrt{p}) = 1/\sqrt{2\,\pi p} \int_{-\infty}^{x_0} \exp(-\,x^2/2\,p)\,\mathrm{d}x\,.$$

Here $x_0 = (n - 1)/2$ and $p = (n - 1)(m - 1)/2$.

The probability for one inequality to be true is then

$$\text{erf}\,(\sqrt{(n - 1)/2(m - 1)})\,. \tag{2.8}$$

A similar expression is found in Hopfield [1] and used to predict a linear dependance of maximum m with n.

As a matter of fact the invariance of one sequence requires n such conditions to be satisfied simultaneously for the n automata of the network. If the n samplings of the random terms were independent, the probability for the n random terms to be smaller than $(n - 1)/2$ would simply be the product of all the probabilities. We would then obtain for the probability of invariance of one sequence :

$$\text{erf}^n\,(((n - 1)/2(m - 1))^{1/2})\,. \tag{2.9}$$

An examination of the construction of the double sum shows that this simplifying hypothesis is not unreasonable.

Let us rewrite matrix X with the X_i^s on the first row :

$$\begin{matrix} X_1^s & \ldots & X_i^s & \ldots & X_n^s \\ X_1^{s'} & \ldots & X_i^{s'} & \ldots & X_n^{s'} \\ X_1^m & \ldots & X_i^m & \ldots & X_n^m\,. \end{matrix}$$

The evaluation of the double sum can be decomposed into several steps. First, half of the rows are decimated corresponding to the $x_j^s = 0$. Along rows s', $X_j^{s'}$ are added together.

These partial sums are then randomly flipped when multiplied by the $X_i^{s'}$. As long as the samplings and summation of the lines are in small number as compared to the total number of possible combinations ($n < 2^{(m-1)}$) they can be considered as independent. We then adopt this hypothesis at the present time. Its validity has been tested by computer simulations.

If we carry on with the same assumptions about the different sequences, based on similar examination about the construction of the double sums, we end up with the following probability that all automata for all sequences are stable

$$P = \text{erf}^{nm}\,(((n - 1)/2(m - 1))^{1/2})\,. \tag{2.10}$$

If we want this probability P to be equal to P^0, any number not too close to 0 or 1, the error functions should be taken for arguments very close to 1. The error function can then be replaced by its asymptotic approximation

$$\text{erf}\,(x) = 1 - \exp(-\,x^2/2)/\sqrt{2\,\pi}\,x\,.$$

This expression to power nm is approximated by its Taylor expansion :

$$\mathrm{erf}^{nm}(x) = 1 - nm \exp(-x^2/2)/\sqrt{2\pi}\, x \tag{2.11}$$

which gives the relationship between P^o and m

$$(n-1)/(4(m-1)) = \ln nm + 1/2 \ln((m-1)/(n-1)) - \ln(\sqrt{\pi}(1 - P^o)). \tag{2.12}$$

This expression allows the calculation of a critical m^o such that if $m < m^o$ the probability for all sequences to be invariant is larger than P^o. In fact computer simulations show that m^o does not vary much when P^o goes from 0.9 to 0.1 — for instance, for $n = 100$, $m^o(0.9) = 5$ and $m^o(0.1) = 7$.

By choosing $P^o = 1 - 1/\sqrt{\pi} = 0.44$ the last term is cancelled and by approximating $n-1$ (resp. $m - 1$) by n (resp. m) the following simple expression for $\beta = n/m$ is obtained :

$$\beta + 6 \ln \beta = 8 \ln n \tag{2.13}$$

β increases with n, but much slower.

For large n, β varies as $\ln n$ which implies that m scales as $n/\ln n$ instead of n.

Expression (2.10) is verified by numerical simulations for values of n varying from 50 to 500 (see Fig. 1). It is difficult to explore a larger range for n because the computer time to test nm conditions implying sums of $n - 1$ terms is roughly of order n^3. For $n = 50$ it takes a few minutes of VAX time and for $n = 500$ more than 4 hours. The independent samplings assumption explains for the fact that the theoretical curves lay slightly below the simulation data.

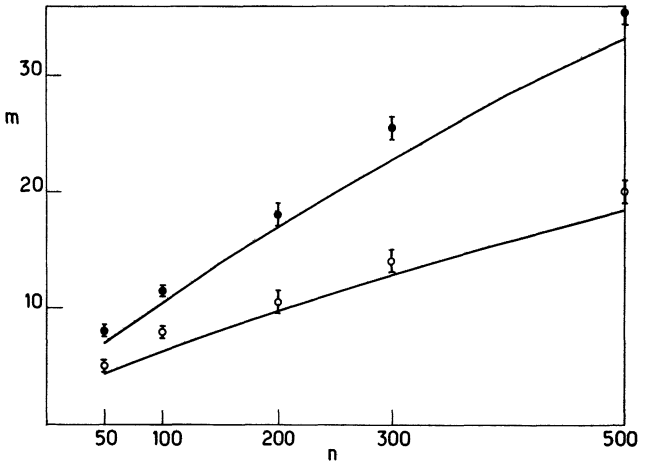

Fig. 1. — Variation of m, the maximum number of reference sequences as a function of n, the number of automata of the network. Solid lines correspond to theoretical calculations. Equation (2.10) allows to compute m for P, the probability of having at most one unstable reference, being equal to 0.5. The lower line is obtained from Hopfield's algorithm (zero thresholds) and the upper line from ours (thresholds as defined by expression (3.1)). The simulation results are obtained by testing nm inequalities (2.1) or (2.2) per network and by adjusting m so that the probability for all inequalities to be verified is 0.5. Circles correspond to zero thresholds, dots to non-zero thresholds.

3. A better choice for thresholds.

We can use expression (2.5) to select a threshold. For a given choice of sequences to retrieve, there always exist some residual correlations among sequences and usually the random sums are not centred around 0. The best choice for thresholds is therefore the mean value of the random term. Using the approximation $\langle x_j^s \rangle = 1/2$, we obtain for b_i

$$b_i = \left(\sum_s \sum_{j \neq i} X_i^s X_j^s \right) \Big/ 2 . \tag{3.1}$$

This choice is equivalent to subtract this expression from the sums while keeping a 0 threshold. The argument of the Heaviside function now becomes :

$$\sum_{j \neq i} a_{ij}(x_j - 1/2) = \sum_{j \neq i} a_{ij} X_j/2 . \tag{3.2}$$

Since the 1/2 does not change anything to the argument of a Heaviside function we can ignore it. If we now recalculate the two sums as in equation (2.5), we find that the polarized term is now exactly $n - 1$, and that the random sum contains $(n - 1)(m - 1)$ terms. This gives a twofold increase for the argument of the error function which implies a nearly equivalent increase for the maximum number of retrieved sequences as verified by the computer simulation on figure 1.

This simple and efficient improvement of the algorithm which consists in replacing the x_j by X_j in the evaluation of the argument of the Heaviside function is from now on adopted and the scaling law for β becomes :

$$\beta + 3 \ln \beta = 4 \ln n . \tag{3.3}$$

4. Memories as attractors.

In order for this system to be used as an associative memory, the reference sequences should be attractors for initial conditions which differ by a few bits from one of the reference sequences. The expressions which we started from in order to study invariance, can also be used to study stability. Let us suppose that we take as an initial state for the network a sequence which differs by d bits from a reference sequence s.

For d modified automata j, X_j^s has become $\overline{X}_j^s$, where $X_j^s = - \overline{X}_j^s$.

If none of the inequalities ensuring the invariance of sequence s has changed its direction when the dX_j's are inverted, the reference state is reached after one iteration per automaton. Let us compute the probability of this event, which is a sufficient condition for convergence toward the reference sequence. The sums $S_i^0 = \sum_j a_{ij} X_j^s$ are then changed into :

$$S_i = S_i^0 + \sum_j a_{ij}(\overline{X}_j^s - X_j^s) \tag{4.1}$$

$$S_i = S_i^0 - 2\, dX_i^s + \sum_{s' \neq s} \sum_j X_i^{s'} X_j^{s'}(\overline{X}_j^s - X_j^s) \tag{4.2}$$

where the sums are taken for the modified j only. The main effect of the modification is thus to decrease the amplitude of the polarized term by $2\,d$ (if automaton i has itself been modified this decrease is only $2(d - 1)$). On the other hand the random term is only randomly and partially modified : $d(m - 1)$ random terms are inverted.

Once more the modified inequalities remain satisfied as long as the random terms are kept inferior to the polarized terms.

The probability for one inequality to be satisfied is now :

$$P(d) = \text{erf}\left((n - 1 - 2d)/\sqrt{(m-1)(n-1)}\right). \tag{4.3}$$

An initial state will thus be attracted towards the reference if n inequalities are verified. At this point, some care has to be taken because the probability of convergence depends upon the mode of iteration, which was not the case for the results we have formerly derived. During a sequential iteration process, some of the d spins which were in the « wrong » direction are reset, thus increasing the probability for the polarized term to be larger than the random sum during the next iteration steps. For example, a process during which all modified automata would be first updated is more likely to converge towards the reference. In order to avoid easy to derive but cumbersome expressions, we can write the probability of convergence for parallel iteration, when all automata change their state at the same time :

$$P_p = P(d)^{n-d}\, P(d-1)^d. \tag{4.4}$$

The probabilities for sequential iteration are bounded below by, and close to this probability. As in section 2, further approximations are used to derive the scaling laws relating n, m and d :

$$\frac{(n-2d)^2}{nm} = 3\ln n + \ln m - 2\ln(n-2d) - \ln 2. \tag{4.5}$$

Figure 2 compares the plot of this equation in the m, d plane for several values of n to the simulation results obtained by direct tests of the modified inequalities. The theoretical predictions show the same behaviour slightly shifted towards the small values of m and d, probably because of residual correlation among simulated reference sequences.

Let us summarize our hypotheses and results.

— The basic assumptions concern the randomness and the non correlations of the sequences to be retrieved. Of course, the inequalities (2.3) are necessary and sufficient conditions to ensure invariance or attractivity whatever the sequences, but the derivation of the scaling laws is based on the above hypotheses. In contrast, in a preceding paper [7], Hopfield device was applied to pattern recognition of letters. In such cases, the patterns have spatial correlations, and the maximum number of retrieved patterns is smaller than predicted by the scaling laws given here.

— Performances are increased by 100 percent as compared to the original model by a proper choice of thresholds.

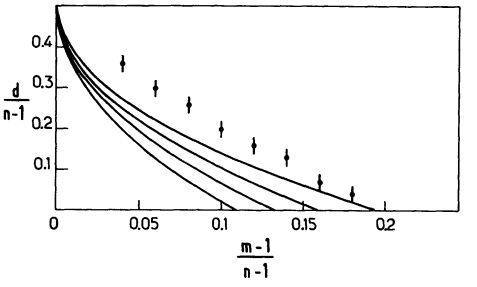

Fig. 2. — Relative width of the attraction basins. Theoretical curves relating $d/(n-1)$, the relative width of the attraction basin, to $(m-1)/(n-1)$, the relative number of automata, are plotted for several values of n, the number of automata (from right to left, n equals 50, 100, 200 and 500). The dots correspond to computer tests of the inequalities ensuring the return to the reference sequence for the case when $n = 50$.

— For a number of reference sequences definitely smaller than the limit for invariance, there exists a Hamming distance such that any initial state closer to the reference converges to this reference with a given probability. This property can be interpreted as a tesselation of the hypercube of the states, such that the reference sequences are well inside the frontiers. In fact the derivations concern sufficient conditions for convergence in one iteration step. The attraction basins are larger since they also contain states which converge in more than one iteration step.

Acknowledgments.

We thank Peretto for communicating his results prior to publication and D'Humières for many helpful comments. We used for the numerical simulations the computer facilities of GRECO 70 of CNRS.

References

[1] HOPFIELD, J., *Proc. Nat. Acad. Sci. USA*, **79** (1982) 2254.
[2] PERETTO, P., *Biological Cybernetics*, **50** (1984) 51, and private communications.
[3] AMIT, D., GUTFREUND, H., SOMPOLINSKY, H., Hebrew University of Jerusalem preprint (1985).
[4] LITTLE, W. A., *Math. Biosci.* **19** (1974) 101.
[5] FOGELMAN-SOULIE, F., WEISBUCH, G., submitted to *SIAM J.* on computing.
[6] Such a choice has also been proposed by LITTLE, W. A., *Math. Biosci.* **39** (1981) 281, for apparently different reasons.
[7] FOGELMAN-SOULIÉ, F., Proceedings of the International Meeting on Artificial Intelligence, Marseille (1984).