IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

On the unlearning procedure yielding a high-performance associative memory neural network

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1998 J. Phys. A: Math. Gen. 31 L463

(http://iopscience.iop.org/0305-4470/31/25/001)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.122
The article was downloaded on 02/06/2010 at 06:55

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **31** (1998) L463–L471. Printed in the UK PII: S0305-4470(98)90393-6

LETTER TO THE EDITOR

On the unlearning procedure yielding a high-performance associative memory neural network

Jorge A Horas and P Marcelo Pasinetti

Universidad Nacional de San Luis, Facultad de Ciencias Físico-Matemáticas y Naturales, Instituto de Matemática Aplicada San Luis (IMASL), Departamento de Física, Ejército de los Andes 950, (5700) San Luis, Argentina

Received 7 January 1998, in final form 30 March 1998

Abstract. We consider a fully connected Hopfield-like neural network as a set of N independent perceptrons. We trained these perceptrons using the so-called *inverse perceptron* rule, obtaining a matrix **J** of synaptic couplings, that make a number of spurious states *unstable*. We numerically determine the optimum number of spurious states, obtained by random shooting, that must be destabilized in order to obtain an improvement in performance.

The *unlearning* procedure generated, is shown to be able to give a high-performance associative memory characterized by: (1) an enhancement in storaging capacity; (2) an enlargement in the size of attraction basins; (3) a reduction in the number of spurious attractors and (4) a reliable and fast retrieval.

One of the most impressive capacities of the human brain is its behaviour as a content addressable memory (CAM). Artificial neural networks (ANN) intend to mimic this capacity and much work has been made in this direction [1–6]. For example, the seminal work of Hopfield [7] must be mentioned, who introduced an energy-like Liapounov function and the corresponding statistical physics methodology used to describe the relaxation of symmetric networks. However, the Hopfield network suffers a main drawback, consisting of the appearance of the blackout catastrophe or overloading. This occurs if too much information is stored, and as a consequence the neural network ceases to function as an associative memory. The performance is seriously affected because the number of spurious or parasitic states grows exponentially with respect to the number of intentionally stored memories.

Thus, a reduction of these spurious states can be expected to notably improve the network efficiency. Numerous contributions have been made to reach this aim [6, 8–10], of which we chose *an unlearning procedure* so far not fully understood. It was first implemented by Hopfield *et al* [11] and later by other authors [12–15]. Basically it consists of starting a relaxation process from a random initial state and iteratively adding the result to each synaptic coupling, with a small negative coefficient, which means 'unlearning' the effect of random retrievals.

The effect of unlearning is to erase the most strongly attracting spurious memories while doing no harm to the true ones. This procedure is inspired by the suggestion of Crick and Mitchison [16], who hypothesized that the purpose of dream (REM) sleep is to weaken certain undesirable modes in the network cells on the cerebral cortex.

Previous authors [11–15] implemented the unlearning mechanism sharing similar ideas, focusing on the number of times that the mechanism must be applied [14]. From another

0305-4470/98/250463+09$19.50 © 1998 IOP Publishing Ltd

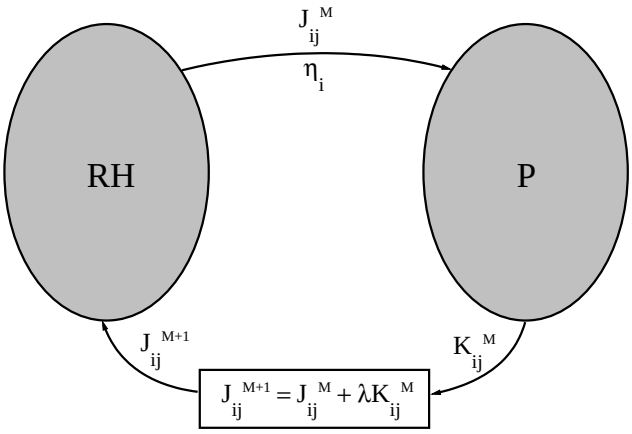

Figure 1. Architecture used. J_{ij}^{M-1} are the weights resulting from the last iteration, K_{ij}^{M} are the weights destabilizing a spurious $\eta_i (i = 1 \ldots N)$ and J_{ij}^{M} are the weights resulting from the present iteration.

point of view, in this letter we shall show that the unlearning procedure actually involves the *destabilization* of spurious states that shall be unlearned, focusing now on determining the number of spurious states that must be destabilized.

In order to reach this aim we consider that a fully connected Hopfield-like network may be viewed as a set of N independent perceptrons which are trained in order to make a number of spurious states unstable [17]. In this system, the synaptic coupling J_{ij} is independent of the connection J_{ji} †. A given system state $\zeta_i (i = 1 \ldots N)$, is both the output of the perceptron i and the input for all the other perceptrons. If a given state is (un)stable for all (or some) perceptrons, it is (un)stable for the whole system, which means that this state is (is not) a fixed point of the neural dynamics.

We address the following issues.

(i) To apply this reverse learning procedure, shedding light on the underlying mechanism.

(ii) To implement the procedure, describing it and discussing its main results.

(iii) To show that the proposed unlearning mechanism, implies an important improvement on each one of the four points that characterize a high-performance CAM, i.e. *high capacity:* the ANN must be able to store the maximum possible number of patterns; *size of attraction basins:* the network should be tolerant to noisy or partial inputs; *the existence of* only relatively *few spurious memories, and few or no limit cycles* with a negligible size of basins of attraction; and finally *fast and reliable memory retrievals*.

The basic architecture used, shown in figure 1, is briefly described as follows.

RH. A Hopfield-like network, fully connected, operated asynchronously following the Monte Carlo or Glauber dynamics at zero temperature

$$h_i(t) \equiv \sum_{j=1}^{N} J_{ij} s_j(t) \tag{1}$$

$$s_i(t+1) = \mathrm{sign}(h_i(t)). \tag{2}$$

† The asymmetry that appears considering N independent perceptrons is negligible, as was determined in all our simulations. We only find fixed points and no limit cycles, i.e. an energy function yet drives the relaxation process and is only a little perturbation of a Hopfield-like one.

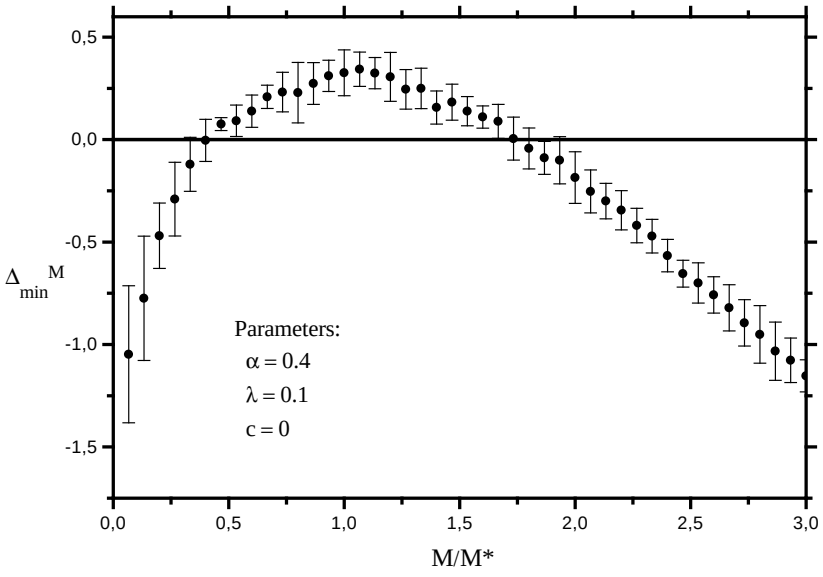

Figure 2. Pattern stability (see text) versus destabilization steps. The maximum is found in $M* = 500$. Each point is an average on at least 10 networks, the bars show the standard deviation of each one.

P. The same RH network considered now as a collection of N independent perceptrons. These perceptrons are trained with the inverse-perceptron rule, whose operation is defined asas follows.

(1) Start with the connection weights existents in RH, $K_{ij}^{0} = J_{ij}$.

(2) Relax from a random configuration and test, for the obtained spurious state $\eta_i (i = 1 \ldots N)$, whether the stability condition for the neurode i is satisfied:

$$\eta_i \sum_j K_{ij} \eta_j > c \tag{3}$$

where $c(\geqslant 0)$ is the stability. While condition (3) is true, we change each weight according to

$$K_{ij} \rightarrow K_{ij} - \varepsilon \eta_i \eta_j \qquad (i, j = 1 \ldots N; j \neq i) \tag{4}$$

where ϵ adjusts the change in the connection weights made in each perceptron step (note the minus sign).

(3) Repeat (2) for each neurode.

A description of the whole sequence to be followed for this iterative and local procedure is as follows.

(a) Storing the patterns (of components $+1$ and -1 with equal probability) through the Hebb rule, in RH.

(b) Attainment of metastable spurious states in RH, resulting from: (i) random shooting: the network is initially started in a random configuration (note that this fits the neurophysiological picture given in [16]), (ii) relaxation towards an attractor: the network evolves to a stationary configuration, $\eta_i (i = 1 \ldots N)$.

(c) Unlearning spurious states by destabilizing them (see (4)). The new connections obtained in such a way, K_{ij}^{M}, are used to correct the synaptic couplings in RH (see [18]),

according to

$$J_{ij}^{M+1} = J_{ij}^{M} + \lambda K_{ij}^{M}. \tag{5}$$

(d) Iterate (b) and (c).

Then, there are two learning phases; the first (point (a)) consisting of storing the pattern to memorize in the Hopfield network, using the Hebb rule. The second (points (b)–(d)) consists of the modification of the synaptic couplings in the original Hopfield network.

We aim to obtain the matrix of connections **J** which determine the optimum number ($M*$) of metastable spurious states to destabilize. Such an optimum can be obtained in many ways. We use the method which gives the maximum stability to the patterns $\xi_i^{\mu}(i = 1 \ldots N)$. This is justified because a strong indication exists [18] that by proceeding in such a way, the maximization of the attraction basins is achieved and, the network performance is also optimized.

The pattern's stability is given by (see figure 2)

$$\Delta_{\min}^{M} = \min\left\{\xi_i^{\mu}\sum_j \xi_j^{\mu} J_{ij}^{M} \text{ with } i = 1 \ldots N \text{ and } \mu = 1 \ldots p\right\} \tag{6}$$

and its maximum

$$\Delta_{\min}^{M*} = \max\{\Delta_{\min}^{M} \text{ with } M = 0 \ldots M_{\max}\} \tag{7}$$

For the optimum number $M*$ we propose

$$M*(\lambda, p, N, c) = f^{\lambda}(\lambda) \cdot f^{p}(p) \cdot f^{N}(N) \cdot f^{c}(c) \tag{8}$$

where λ (see (5)) adjusts the change made over the couplings of RH with the weights K_{ij}^{M} while p is the number of patterns, N is the network size (the number of neurodes) and c is the stability in the inverse-perceptron algorithm. The assumption of a multiplicative dependence was verified numerically.

A working zone is defined taking five values of each parameter λ, p, N and c, which includes the typical cases, being $p \propto N$ the really interesting one. Applying these values and collecting the results we have obtained

$$M*(\lambda, p, N, c) = \frac{1}{\lambda} \cdot p \cdot \left(\frac{\kappa_1}{N} + \kappa_2\right) \exp(\kappa_3 \cdot c) \tag{9}$$

where $\kappa_1 = 16.684 \pm 2.22$, $\kappa_2 = 1.5 \pm 0.047$, $\kappa_3 = 0.91 \pm 0.02$.

This equation gives the optimum number of spurious states that need to be destabilized in order to obtain an optimal performance. This deserves some comments about the specific scaling of the parameters. At this point one must realize that (9) has a validity range given by the data collected.

Thus, within the limits of our numerical results, for increasing correction strengths λ (see (5)), $M*$ correspondingly decreases, and the observed linear dependence of $\alpha(= p/N)$ and p on $M*$, gives the desired performance.

In the range studied, we have observed that the spurious stability shows a distribution in which the parameter c operates like a cut-off value. Thus, an exponential number of steps are needed to destabilize the remaining spurious states.

On the other hand, at least for some values outside the working zone, the α and p scaling appears to be exponential and the λ and c dependence shows a breakdown for $\lambda > 1.5$ and $c > 2$ respectively.

It is particularly instructive to compare a given model with the Hopfield model [7]. Consequently, our results using (9) will be contrasted with those obtained by the standard

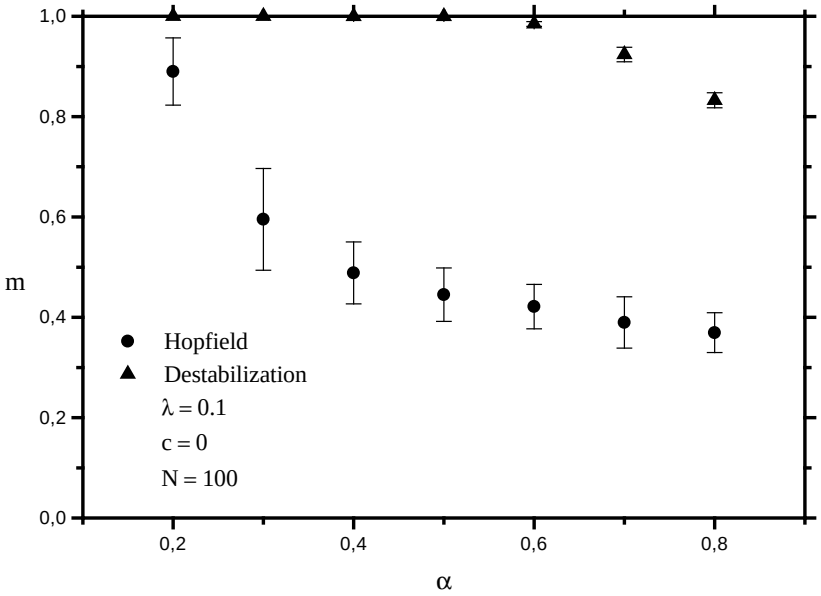

Figure 3. Overlap m versus $\alpha = p/N$. Each point or triangle is the average of at least 10 networks, and the bars indicate the standard deviation.

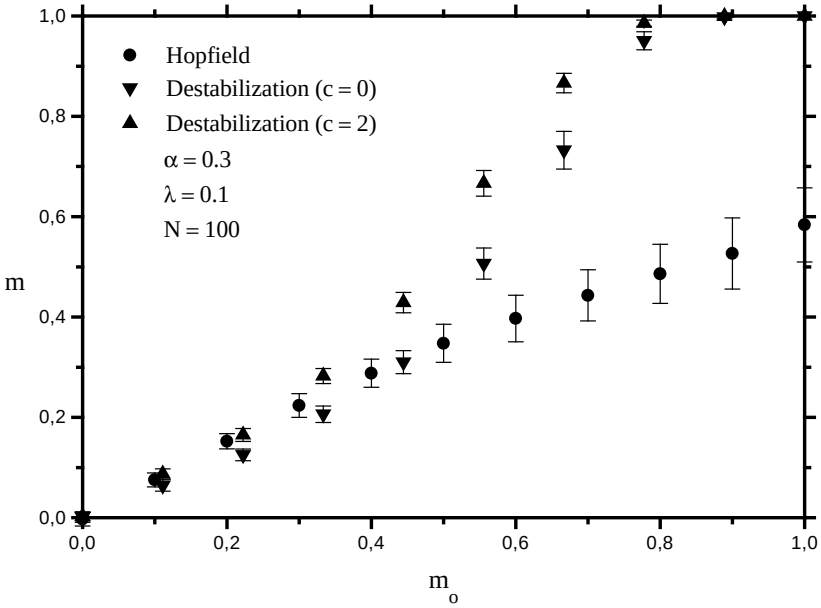

Figure 4. Final overlap m versus initial overlap m_0 for a reference pattern. Each point or triangle is the average of at least 10 networks, and the bars indicate the standard deviation.

Hopfield model. For this purpose we follow the following four points characterizing a high-performance CAM.

- *High capacity*. In figure 3 we show a typical capacity plot of overlap m versus

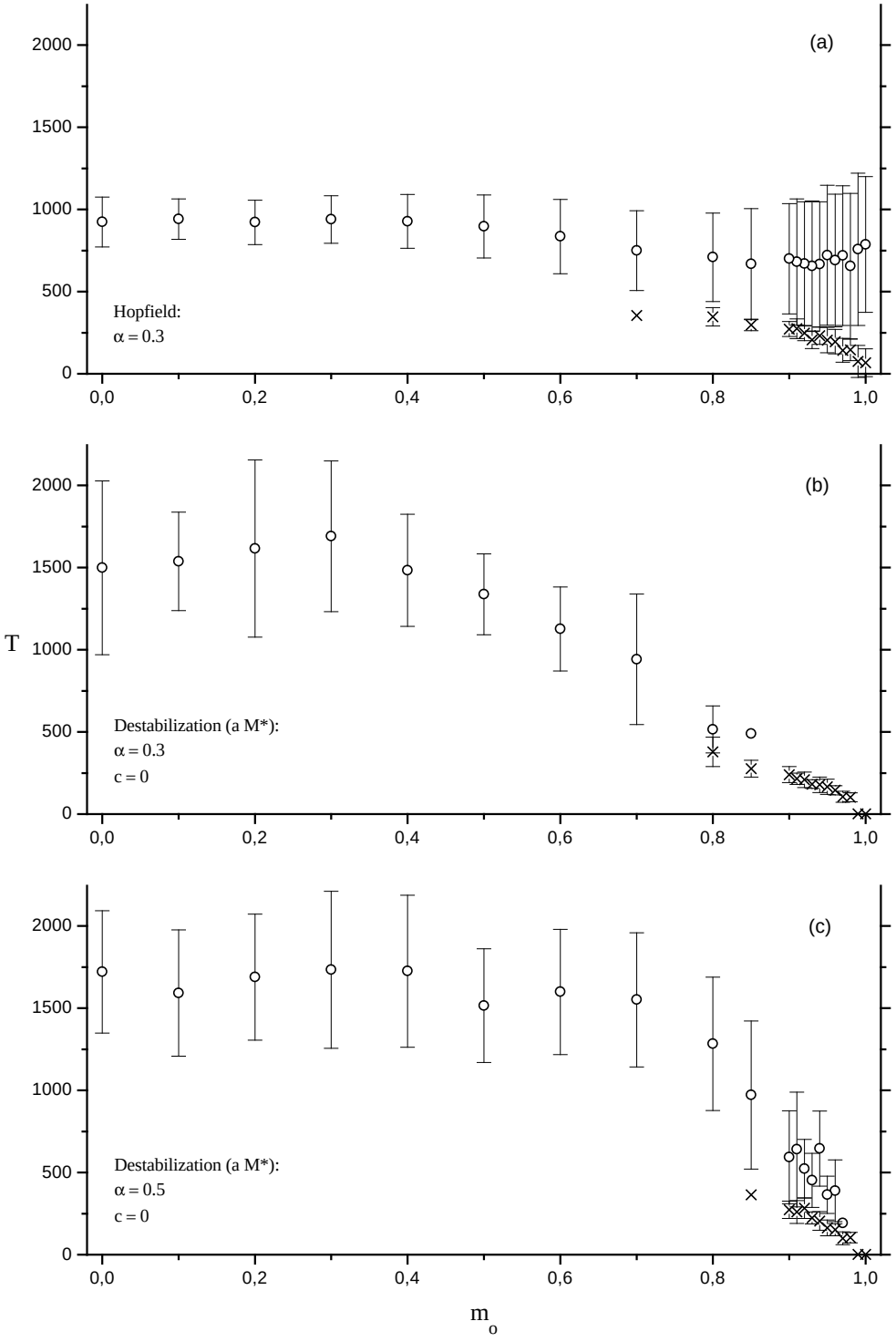

Figure 5. Time spent to relax (T) to reach a fixed point starting from an initial overlap m_0 (see text).

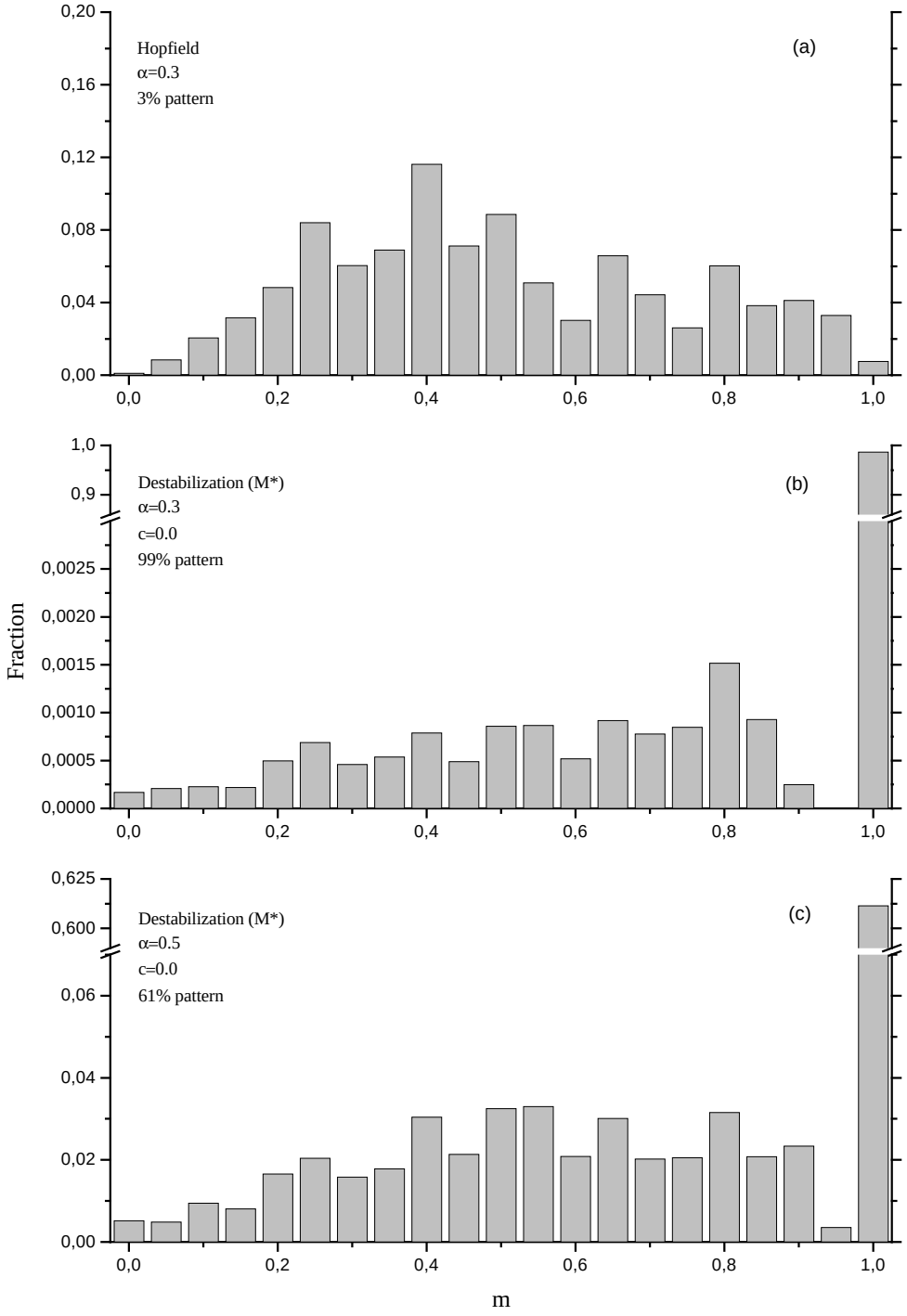

Figure 6. Histograms of retrieval overlaps (m). The histograms are averaged over at least 50 networks and 1000 trials have been performed on each one.

loading $\alpha = p/N$, the enhancement of the capacity being evident. The overlap m between the reached final stationary state $\zeta_i (i = 1 \ldots N)$ and the starting pattern $\xi_i^\mu (i = 1 \ldots N, \mu = 1 \ldots p)$ is defined as $\frac{1}{N} \sum_{i=1}^{N} \zeta_i \xi_i^\mu$.

• *Size of attraction basins.* In figure 4 we plot the final overlap reached (m) versus the starting one (m_0) in order to show a measure of the improvement obtained with the proposed procedure on the size of attraction basins. The network is initialized at configurations with an overlap m_0 relative to a target or reference pattern. For each initial state, the ordinate shows the overlap of the reference pattern with the fixed point reached after relaxation. Similar results have been obtained for greater network sizes ($N = 200$) and for other c values within the working zone.

• *The existence of* only relatively *few spurious memories and few or no limit cycles* with a negligible size of attraction basins. The results that have already been shown in figures 3 and 4 are a consequence of the effective reduction of spurious states. Limit cycles have never been observed. It is also interesting at this point to consider the retrieval time results.

• *Fast and reliable memory retrievals.* In order to have a measure of retrieval time, we determine the time spent to evolve to a fixed point (measured as the number of relaxation steps), starting from an initial overlap m_0, relative to a reference pattern. In figures 5(a)–(c) we represent with a cross the time taken to reach a pattern (final overlap $m > 0.95$) and with an open circle the time for the case moving away from a pattern (final overlap $m < 0.95$).

In order to complement and to make more evident the results shown in figures 5(a)–(c), we plot in figures 6(a)–(c) the corresponding histograms of retrieval overlaps (m) for the case of initial overlap $m_0 = 0.85$.

In the Hopfield case (figure 6(a)) only 4% of trials converged to the target pattern, and the remaining 96% of trials resulted in convergence to patterns other than the target pattern. Most of these are spurious attractors that have a retrieval overlap of about 0.5. We can observe that by applying the unlearning procedure (figure 6(b)), the inverse behaviour is obtained, i.e. 99% of the trials now converge to the target pattern.

Comparison of figures 5 and 6 shows the unlearning procedure gives a somewhat faster and notably more reliable memory retrieval than the Hopfield model, for networks which have the same relatively high loads, $\alpha = 0.3$. Moreover, an attenuated result is gained even for an extreme case ($\alpha = 0.5$) if the unlearning procedure if applied. It must be observed that at this load the breakdown catastrophe makes the Hopfield model useless as a CAM.

In conclusion we have given a new framework to obtain insight into the underlying mechanism of the unlearning procedure. We have also shown that destabilizing is an effective mechanism to reduce the number of spurious attractors. The application of the proposed procedure results then in achieving a high-performance CAM.

The authors acknowledge partial financial support from Secretaría de Ciencia y Técnica de la Universidad Nacional de San Luis. PMP is the recipient of a fellowship from CONICET.

References

[1] Domany E, van Hemmen J L and Shulten K 1991 *Models of Neural Networks* 1st edn (New York: Springer)
[2] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* vol 1 (London: Addison-Wesley)
[3] Amit D, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30–67
[4] Amit D, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293–303
[5] Geszti T 1990 *Physical Models of Neural Networks* (Singapore: World Scientific)
[6] Monoranjan P Singh, Zhang Chengxiang and Chandan Dasgupta 1995 *Phys. Rev.* E **52** 5261
[7] Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554–8

[8] Mori Y, Davis P and Nara S 1989 *J. Phys. A: Math. Gen.* **22** L525
[9] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167–73
[10] Gardner E 1987 *Europhys. Lett.* **4** 481–5
[11] Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158–9
[12] Keinfeld D and Pendergraft D B 1987 *Biophys. J.* **51** 47–53
[13] Poppel G and Krey U 1987 *Europhys. Lett.* **4** 979
[14] van Hemmen J L, Ioffe L B, Kuhn R and Vaas M 1990 *Physica* **163A** 386–92
[15] Christos G A 1996 *Neural Networks* **9** 427–34
[16] Crick F and Mitchison G 1983 *Nature* **304** 11
[17] Gardner E, Stroud N and Wallace D J 1987 *Preprint* Edinburgh 87/394
[18] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745–52