IJMSOR Vol. 1, Issue. 1
January - June, 2016

Information, Stability and Learning Complexity in Associative Memories

Información, Estabilidad y Complejidad de Aprendizaje en Memorias Asociativas

dx.doi.org/10.17981/ijmsor.01.01.07

Research Article - Reception Date: November 3, 2015 - Acceptance Date: February 17, 2016

Enrique Carlos Segura
Universidad de Buenos Aires
esegura@dc.uba.ar, Buenos Aires, Argentina

To reference this paper:
E. C. Segura "Information, Stability and Learning Complexity in Associative Memories", *IJMSOR*, vol. 1, no. 1, pp. 49-53, 2016. DOI: dx.doi.org/10.17981/ijmsor.01.01.07

Abstract-- The relationship between amount of information and learning complexity is studied for the discrete Hopfield model of associative memory. More precisely, we analize, from a statistical point of view, the relation between the probability of a state of the system to be a stable equilibrium (i.e. a memory) and a value of entropy or uncertainty associated to it. Computer experiments are made to confirm these results.

Key Words-- associative memory - Hopfield network - information - entropy - learning complexity.

Resumen-- En este trabajo se estudia la relación entre la cantidad de información y la complejidad de aprendizaje en el modelo discreto de Hopfield de memoria asociativa. Más específicamente, se analiza, desde un punto de vista estadístico, la relación entre la probabilidad de que un estado del sistema sea un equilibrio estable (memoria) y un valor de entropía o incertidumbre asociado a él. Se realizan experimentos computacionales para corroborar estos resultados.

Palabras claves-- memoria asociativa – red de Hopfield – información – entropía – complejidad de aprendizaje.

Address correspondence to Enrique Carlos Segura, Department of Computer Science, Faculty of Exact and Natural Sciences, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria 1428, Buenos Aires, Argentina. E-mail: esegura@dc.uba.ar

IJMSOR vol. 1 no. 1, pp. 49-53. January - December, 2016
Barranquilla. ISSN 2539-5416 Online

I. Introduction

Up to now, the capacity of neural networks (and, especially, of the Hopfield associative memory [1]) as dynamical systems to memorize, i.e. to have attractor stable states, has been studied in depth [6],[7],[8]. Our main interest here is the question concerning the complexity of learning in relation with the information to be learned. In this work we investigate whether (and, if so, how) the probability that a state, after taught to the network, become a stable equilibrium, depends on some property inherent to the structure of the state. We focus on the amount of information (or of uncertainty) contained in the state.

In the first experiments we tried to find a relation between the complexity of learning several parameters of the problem, such as dimension (i.e. number of neurons in the network) and the number of states taught to the system.

The first fact observed was the independence of the complexity with respect to the number of neurons. This was previsible, since what the network learns is a matrix of synaptic weights, and the convergence of each entry of this matrix to its correct value is "in parallel" (i.e. simultaneous) with the others. If we consider each memory to be learned as a concept (a relationship between a sequence of attributes) and each unit (bit) of that memory as an attribute of the concept, then to learn an attribute consists of learning N connections. But that learning is accomplished in parallel; hence the complexity, in terms of number of examples, does not increase.

II. Basic Definitions

The Hopfield model of associative memory [1] is a simple model in its theoretic formulation and easy to implement in a computer. It consists of a system of interconnected units ("neurons"), having each one two possible states, 1 and 0. The state of neuron i at time t will be 1 ("activated") if the sum of the input due to the other units, properly weighted, is greater than a critical value or "action threshold". That is:

$$x_i(t+1) = sgn\{\sum_{j=1}^{N} T_{ij} x_j(t) - d_i\}$$

being N the number of neurons that form the net, $\vec{x} = (x_1,..., x_N)^T$ the vector that describes the status of the net, d_i the action threshold associated to neuron i and T the matrix of weights ("synaptic connections"), which is computed according to Hebb's rule:

$$T_{ij} = \begin{cases} \sum_{k=1}^{M} (2x_i^k - 1)(2x_j^k - 1) & if\; i \neq j \\ 0 & if\quad i = j \end{cases}$$

where $\vec{x}^k = (x_1^k,..., x_N^k)$ are the configurations intended to be stable (i.e. memorized) for the network. We make the assumption that no neuron interacts with itself, hence the main diagonal of T is zero.

This simple model, widely studied, exhibits interesting properties related to its information capacity. In [1] it is proven that, if the states that define the matrix T (the memories $\vec{x}^k$, k = 1,..., M) are pseudoorthogonal, i.e. they behave as random variables following a probability distribution such that

$$E\left(\sum_{\substack{k,\, l=1 \\ k =/ 1}}^{M} \sum_{i=1}^{N} x_i^k x_i^l \right) = 0$$

where E stands for the expectation and is computed over the sum of all inner products between crossed pairs of vectors to be stored, then the network can memorize (have as attractors in its evolution) a certain amount of states, in proportion to the number of neurons. This capacity has been rigorously analysed and quantified [9],[10].

The purpose of this paper is to study an experimental fact: in the "classic" Hopfield model, the size of the basin of attraction for a stable state (subset of initial states that evolve into that state) depends strongly on the considered state. This implies that not all the memories would be equally complex to be learned. This dependence of the learning complexity with regard to the information to be learned, opens the question which we are here interested in: the relation between the complexity of learning and some measure of the amount of information contained in the state to be learned.

We used a simplified formulation of the law of evolution between states, as follows:

$$x_i(t+1) = sgn\{\sum_{j=1}^{N} T_{ij} x_j(t)\}$$

i.e. considering all thresholds equal to zero.

III. Information and Complexity

As announced previously, from the experiencies made up to now in connection with the Hopfield model, it follows that the size of each basin of attraction depends strongly on the structure of the corresponding attracting state. We intend to sketch and test a hypothesis about the relation between the size of these basins and the structure of the corresponding attracting state.

Suppose that our concept is represented by the vector $\vec{x} = (x_1,..., x_N)$ with x_i taking values 1 or 0. We define the entropy or uncertainty associated to $\vec{x}$ as

$$I(\vec{x}) = -[f_{\vec{x}}(1) \log f_{\vec{x}}(1) + f_{\vec{x}}(0) \log f_{\vec{x}}(0)]$$

where

$$f_{\vec{x}}\,(k) = \#\,\{i/x_i = k\}\,/\,\mathrm{N} \qquad k = 1 \; or \; 0$$

i.e. frequence of the value k in $\vec{x}$ [4].

According to this definition, there is a direct relation between the amount of information that a vector contains and the similarity between the frequencies with which their components take each of the possible values (in this case, only two). States with all their components equal (1 or 0), have the less possible entropy. The latter fact agrees exactly with what happens in the Hopfield model with a state and its complement (the vector having 1's instead of 0's and conversely), in the sense that, if one of them is stable, also is the other: from the point of view of the information that they contain, they are also equivalent. Moreover, our experiments confirmed that the states the most easy to learn are those which minimize I and, at the same time, those which, once learned, have the largest attracting basins. In the Hopfield model the complexity of learning a concept seems to be directly proportional to the uncertainty associated with it and inversely proportional to the size of the basin that the concept generates after learned. An intuitive but reasonable interpretation of this fact is that a state with a large uncertainty can be recognized only from another very similar state (that is what means, ultimately, that its attracting basin is small).

We want to study from a statistical point of view the relation in Hopfield's model between the probability that a state of the system be an equilibrium state (i.e. a "memory") and the value I associated to that state and defined as above.

Suppose that $j1, j2,..., jL$ with $1 \leq j \leq N$ is the sequence of natural values such that

$$x_k = \begin{cases} 1 \; if \; k = ji \; for \; some \; i \\ 0 \; if \; not \end{cases}$$

that is, exactly L components of $\vec{x}$ are 1's and N-L are 0's.

Suppose we teach the network M memories $\vec{x}^m$ by means of the Hebb's rule, that is, the weight matrix Tij is

$$T_{ij} = \begin{cases} \sum\limits_{k=1}^{M} (2x_i^k - 1)(2x_j^k - 1) \quad if \; i \neq j \\ 0 \quad if \quad i = j \end{cases}$$

and that the system evolves according to the Cooper's law, i.e.:

$$x_i(t+1) = sgn\{\sum_{j=1}^{N} T_{ij}\, x_j(t)\}$$

Memories $\vec{x}^s$ are generated independently and pseudoorthogonally. Then, for a certain memory $\vec{x}^s$, with a sequence $\{jl\}_1^{L'}$ associated:

$$P(\vec{x}^s \; stable) = P(sgn T\vec{x}^s = sgn\; \vec{x}^s) =$$

$$P\left(\mathrm{sgn} \sum_{m=1}^{M} \sum_{l=1}^{L'} (2x_i^m - 1)(2x_{j_l}^m - 1) = \mathrm{sgn}\, x_i^s \; \forall i = 1 \ldots N\right) =$$

$$\prod_{i \in \{j_l\}} P\left(L' + \sum_{l=1}^{L'} \sum_{\substack{m=1 \\ m \neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) \geq 0 \right)$$

$$\prod_{i \notin \{j_l\}} P\left(-L' + \sum_{l=1}^{L'} \sum_{\substack{m=1 \\ m \neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) < 0 \right)$$

(note the product distributed between two lines due to space limitations)

Although the sums are not independent random variables, it can be seen that their correlation is positive. Then we can find the following less bound for the latter probability:

$$\prod_{i \in \{j_l\}} P\left(\sum_{l=1}^{L'} \sum_{m=1}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) \geq 0 \right) \prod_{i \notin \{j_l\}} P\left(\sum_{l=1}^{L'} \sum_{m=1}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) < 0 \right) =$$

$$\prod_{i \in \{j_l\}} P\left(\sum_{l=1}^{L'} (2x_i^s - 1)(2x_{j_l}^s - 1) + \sum_{l=1}^{L'} \sum_{\substack{m=1 \\ m \neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) \geq 0 \right)$$

$$\prod_{i \notin \{j_l\}} P\left(\sum_{l=1}^{L'} (2x_i^s - 1)(2x_{j_l}^s - 1) + \sum_{l=1}^{L'} \sum_{\substack{m=1 \\ m \neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) < 0 \right) =$$

$$\prod_{i\in\{j_l\}} P\left(L' + \sum_{l=1}^{L'} \sum_{\substack{m=1\\ m\neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) \geq 0\right) \prod_{i\notin\{j_l\}} P\left(-L' + \sum_{l=1}^{L'} \sum_{\substack{m=1\\ m\neq s}}^{M} (2x_i^m - 1)(2x_{j_l}^m - 1) < 0\right)$$

Double sums are sums of L'(M-1) independent binomial variables. Observing that each one takes value 1 or 0 with equal probability, it can be easily seen that their expectation equals 0 and their variance equals 1. Hence, if we consider a large number of neurons (N) and, consequently, large values for L' too, we can approximate these sums by normal variables with variance L'(M-1) and mean L' in the first case and -L' in

the second case (due to the constants which are summed in each case) [2],[3],[5]. Then the bound is obtained in the form of a normal N-valuated probability distribution:

$$\left(\frac{1}{2\pi L'(M-1)}\right)^{\frac{N}{2}} \int_0^\infty \dots \int_0^\infty \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \frac{(x_i - L')^2}{(L'(M-1))}\right) dx_1 \dots dx_N =$$

$$\left(\frac{1}{2\pi L'(M-1)}\right)^{\frac{N}{2}} \left[\int_0^\infty \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \frac{(x_i - L')^2}{(L'(M-1))}\right) dx\right]^N$$

Introducing the change of coordinates

$$y = \frac{(x- L')}{\sqrt{L'(M - 1)}}$$

we obtain finally:

$$P(\vec{x}^s\ stable) \geq [F(A(L', M)]N$$

where F is the normal distribution function with mean equal to zero and variance equal to 1 and

$$A(L',M) = \left(\frac{L'}{M - 1}\right)^{\frac{1}{2}}$$

According to this, for every fixed $\vec{x}^s$ and M, number of memories, it holds that

$$P(\vec{x}^s\ stable) \xrightarrow{N\to\infty} 1$$

but the limit case (the best, respecting rate of convergence) is

$$\vec{x}^s = (1,\dots,1)$$

IV. Experiments

In this section we present the results of computer simulations made to test the validity of the theoretic bound derived above.

We computed, for different values of N and M, the relation between the probability of a taught state to be stable and the value of I for that state. Figures 4.1 and 4.2 show some of those results.

From the comparison between the resulting curves and the theoretical result, we infer that the bound proposed here is confirmed by experience. Nevertheless, it seems that for certain values (or ranges of values) of I, the actual behavior is closer to the theoretical bound than for others.

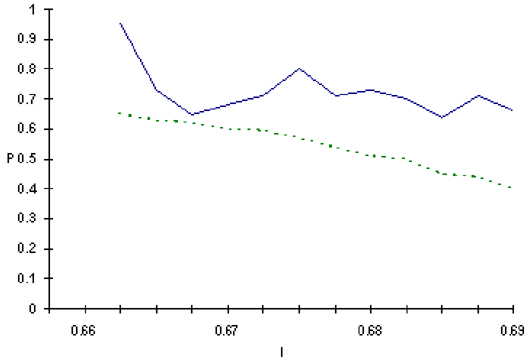

Fig. 4.1 Relation between the probability of stability of a state and the value of I for that state. Here N = 100 and M = 10. In dotted lines, the theoretic less bound.

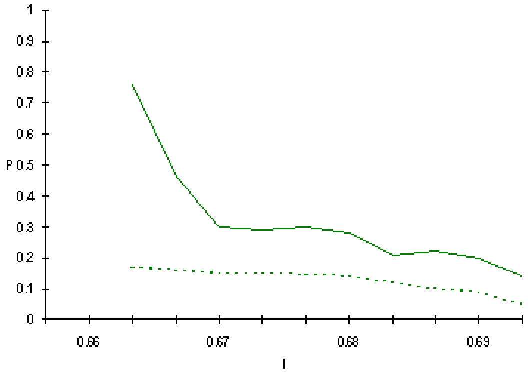

Fig. 4.2 Idem fig. 4.1 with N = 100 and M = 15. Note the decrease of both the actual probability of learning and the bound when increasing M.

V. Conclusions

In this paper we analyzed the relation between learning complexity and entropy of the information to be learned in the Hopfield model of associative memory. Specifically, a lower bound for the probability of equilibrium of a memory has been obtained. Computer experiments confirmed this theoretical result.

Several questions remain open. An interesting one concerns the search for some measure representing the structure of the information, and not only its probability distribution. For example, according to the definition given above, the vectors

(1 0 1 0 1 0 1 0 1 0 1 0)

and

(1 0 0 0 1 0 1 1 0 1 0 1)

have the same value I associated. However, the number of different structures in the second pattern is much greater than in the first. For future work we would like to find a measure (or measures) expressing the information contained in those topologic structures. As for the Hopfield model, it does not seem quite capable (at least in its original formulation) to quantify information according to finer measures than the used in this paper.

References

[1] J. J. Hopfield, “Neural networks and physical systems with emergent computational abilities”, Proc. Nat. Acad. Sc., USA, 79, pp. 2554-8, April 1982 (Biophysics).

[2] T. Hastie, R. Tibshirani and J. Friedman, The Ekements of Statistical Learning: Data Mining, Inference, and Prediction. Berlin, Germany: Springer, 2013.

[3] A. C. Rencher and W. F. Christensen, Methods of Multivariate Analysis. New Jersey, USA: Wiley, 2012.

[4] C. E. Shannon; “A mathematical theory of communication”, Bell System Technical Journal, vol. XXVII, no. 3, pp. 379-423, 1948.

[5] G. Grimmett and D. Stirzaker, Probability and random processes. Oxford, UK: Oxford University Press, 3rd. edition, 2001.

[6] D. J. Amit, H. Gutfreund and H. Sompolinsky, Statistical Mechanics of Neural Networks near Saturation,} Annals of Physics 173, pp. 30-67, 1987.

[7] D. J. Amit, Modeling Brain Function. Cambridge, USA: Cambridge University Press, 1989.

[8] S. Haykin, Neural Networks and Learning Machines Upper Saddle River, USA: Pearson–Prentice Hall, 2011.

[9] Y. S. Abu-Mustafa and J-M. St. Jacques, “Information capacity of the Hopfield model”, IEEE Trans. Inf. Th., vol. IT-31, no. 41, pp. 461-464, 1985.

[10] R. J. McEliece, E. C. Posner, E. R. Rodemich and S. S. Venkatesh, “The Capacity of the Hopfield Associative Memory”, California Institute of Technology, 1986.