Hopfield Network as Associative Memory with Multiple Reference Points

Domingo López-Rodríguez, Enrique Mérida-Casermeiro, and Juan M. Ortiz-de-Lazcano-Lobato

***Abstract*—**Hopfield model of associative memory is studied in this work. In particular, two main problems that it possesses: the apparition of spurious patterns in the learning phase, implying the well-known effect of storing the opposite pattern, and the problem of its reduced capacity, meaning that it is not possible to store a great amount of patterns without increasing the error probability in the retrieving phase. In this paper, a method to avoid spurious patterns is presented and studied, and an explanation of the previously mentioned effect is given. Another technique to increase the capacity of a network is proposed here, based on the idea of using several reference points when storing patterns. It is studied in depth, and an explicit formula for the capacity of the network with this technique is provided.

***Keywords*—**Associative memory, Hopfield network, Network capacity, Spurious patterns.

I. Introduction

Associative memory has received much attention for the last two decades. Though numerous models have been developed and investigated, the most influential is Hopfield Associative Memory [1], based on his studies of collective computation in neural networks.

Hopfield's model consists in a fully-interconnected series of bi-valued neurons (outputs are either -1 or +1). Neural connection strength is determined in terms of weight matrix W, $w_{i,j}$ representing the synaptic connection between neurons i and j. This matrix is fixed, that is, once the learning phase (an application of Hebb's postulate of learning [2]) has finished, no further synaptic modification is considered.

Two main problems are found in this model: the apparition of spurious patterns and its low capacity.

Spurious patterns are local minima of the corresponding energy function and not associated to any stored pattern.

The capacity parameter α is usually defined as the quotient between the maximum number of patterns to load into the network and the number of used neurons that achieve an acceptable error probability in the retrieving phase. It has been shown that this constant is approximately $\alpha = 0.15$ for Hopfield's model. This value means that if the net is formed by N neurons, a maximum of $K \leq \alpha N$ patterns can be stored and retrieved with very little error probability.

McEliece [3] showed that the asymptotic capacity of the network is at most $\frac{N}{2\log N}$, if most of the prototype patterns are to remain as fixed points. This capacity decreases to $\frac{N}{4\log N}$ if every pattern must be a fixed point.

In this work, a technique to avoid the apparition of spurious states in Hopfield's model is explained in terms of the decrease of the energy function associated to state vectors.

The main contribution of this paper consists in an extension of this model as associative memory to overcome the problem of its reduced capacity.

The organization of the paper is as follows: in Sec. II, a description of Hopfield model is given, putting special emphasis on its application as content-addressable memory. In Sec. III, the method to avoid the apparition of spurious patterns is presented. In Sec. IV, the associative memory model is extended by the use of multiple reference points, and in Sec. V, a study of the capacity of this new model is presented, similar to that presented in [4], followed by several consequences of importance. Finally, in Sec. VI some final remarks and conclusions are given, as well as possible future research lines.

II. Hopfield's Model

A. The Network

Hopfield's model consists in a net formed by N neurons, whose outputs (states) are either -1 or +1. Thus, the state of the net at time t is completely defined by a N-dimensional state vector $V(t) = (V_1(t), V_2(t), \ldots, V_N(t)) \in \{-1,1\}^N$.

Associated to every state vector there is an energy function that determines the behavior of the net:

$$E(V) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}V_iV_j + \sum_{i=1}^{N}\theta_i V_i \qquad (1)$$

where $w_{i,j}$ is the connection weight between neurons i and j, and θ_i is the threshold corresponding to i-th neuron. Since

Manuscript received July 15, 2005.

D. López-Rodríguez is with the Department of Applied Mathematics, at the University of Málaga, Spain. (Corresponding author. Phone: +34 952 132 866. e-mail: dlopez@ ctima.uma.es).

E. Mérida-Casermeiro is also with the Department of Applied Mathematics, at the University of Málaga, Spain. (e-mail: merida@ ctima.uma.es).

J. M. Ortiz-de-Lazcano-Lobato is with the Department of Computer Science and Artificial Intelligence, at the University of Málaga, Spain. (e-mail: jmortiz@lcc.uma.es).

thresholds are not used in the case of associative memory, all of them are considered to be 0.

By using the function $f(x,y)=2\delta_{x,y}-1$, which takes value 1 if $x=y$ and -1 otherwise, (1) can be rewritten as:

$$E(V)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}w_{i,j}\,f(V_i,V_j) \tag{2}$$

In this work, we have also considered discrete time and semi-parallel dynamics, where only one neuron is updated at a time, achieving the greatest possible descent of the energy function.

B. The Associative Memory

Let us consider $\{X^{(k)}:k=1,...,K\}$, a set of bipolar patterns to be loaded into the network. In order to store these patterns, weight matrix *W* must be determined. This is achieved by applying Hebb's classical rule for learning. So, the increment of the weights, when pattern $X=(X_i)$ is introduced into the network, is given by $\Delta w_{i,j}=X_iX_j=f(X_i,X_j)$ [5]. Thus, the final expression for the weights is:

$$w_{i,j}=\sum_{k=1}^{K}X_i^{(k)}X_j^{(k)}=\sum_{k=1}^{K}f(X_i^{(k)},X_j^{(k)})\,. \tag{3}$$

In this case, the energy function that is minimized by the network can be expressed in the following terms:

$$E(V)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K}f(X_i^{(k)},X_j^{(k)})f(V_i,V_j) \tag{4}$$

In order to retrieve a pattern, once the learning phase has finished, the net is initialized with the known part of the pattern. Then, the dynamics makes the network converge to a stable state (due to the decrease of the energy function), corresponding to a local minimum. Usually this stable state is close to the initial one.

III. How to Avoid Spurious Patterns

Definitions. Given a state *V*, its *associated matrix* is defined as $G_V=(g_{i,j})$ such that $g_{i,j}=f(V_i,V_j)$.

Its *associated vector* is $A_V=(a_k)$ with $a_{j+N(i-1)}=g_{i,j}$, that is, it is built by expanding the associated matrix as a vector of N^2 components.

When a pattern *X* is loaded into the network, by modifying weight matrix *W*, not only the energy corresponding to state $V=X$ is decreased. This fact can be explained in terms of the associated vectors.

With this notation, the energy function can be expressed as:

$$E(V)=-\frac{1}{2}\sum_{k=1}^{K}<A_{X^{(k)}},A_V> \tag{5}$$

where $<\cdot,\cdot>$ denotes the usual inner product.

Lemma. Give a state *V*, we have $A_V=A_{-V}$.

If two states have the same associated vector (as is the case of *V* and *–V*), they will have the same energy value. More concretely, the increment of energy of a state *V* when the pattern *X* is loaded into the network, by using (3), is given by

$$\Delta E(V)=-\frac{1}{2}<A_X,A_V> \tag{6}$$

Since A_V, A_X are vectors of N^2 components taking value in $\{-1,1\}$, their norms are the same $\|A_V\|_E=N$ for all *V*. This fact implies that what the network actually stores is the orientation of the vectors associated to loaded patterns.

So, as *X* and *–X* have the same associated vector, the increase of energy will be the same for both of them when a pattern is introduced into the network. This fact explains the well-known problem of loading the opposite pattern of Hopfield's associative memory.

From (6), and using that terms in A_V, A_X are either -1 or 1, the following expression for the decrease of energy when a pattern is loaded is obtained [5]:

$$-\Delta E(V)=\frac{1}{2}(N-2d_H(V,X))^2 \tag{7}$$

where $d_H(V,X)$ is the Hamming distance between vectors *V* and *X*.

Definition. The *augmented pattern* $\hat{X}$, associated to *X*, is defined by appending to *X* the possible values of its components. In the case of bipolar outputs, we have $\hat{X}=(X_1,...,X_N,-1,1)$.

By making use of augmented vectors, the problem of spurious patterns is solved, as stated in the next result:

Theorem 1. The function ψ, that associates an augmented pattern to its corresponding associated vector, is injective.

It can be shown that if augmented patterns are used, the state *V* whose energy decreases most when pattern *X* is introduced in the net, is $V=X$. By using (7), if $V\neq X$, then we have $1\leq d_H(V,X)\leq N$, and this inequality holds: $2-N=N+2-2N\leq N+2-2d_H(V,X)\leq N+2-2=N$.

Therefore

$$\begin{aligned}&-2\Delta E(V)=(N+2-2d_H(V,X))^2\leq\\&\leq\max\left\{(2-N)^2,N^2\right\}=N^2<(N+2)^2=-2\Delta E(X)\end{aligned} \tag{8}$$

which demonstrates our statement.

Then, in order to load a pattern *X*, it will suffice to load its augmented pattern, which will be the unique state maximizing the decrease of energy.

It must be noted that it will only be necessary *N* neurons, since the last 2 are fixed.

IV. Associative Memory with Many Reference Points

In Hopfield's classical model, the unique reference point is the origin in $\mathbf{R}^N$. As the network stores the orientations of the associated vectors, it could be useful to shift patterns by different amounts in order to be capable of distinguish them more accurately.

In this work, to load the set $\{X^{(k)}:k=1,...,K\}$, we use as reference points the set $\{O^{(1)},\ldots,O^{(Q)}\}\subset\{-1,1\}^N$.

For each k and q, $X_i^{(k)} - O_i^{(q)} \in \{-2,0,2\}$ for every i. Then, the augmented pattern (associated to) $X^{(k)} - O^{(q)}$ will be (using the same notation for simplicity) $X^{(k)} - O^{(q)} = (X_1^{(k)} - O_1^{(q)}, ..., X_N^{(k)} - O_N^{(q)}, -2, 0, 2)$ for each k and q.

By extending what was exposed in Sec. II, the weights will now be defined by:

$$w_{i,j} = \sum_{q=1}^{Q} w_{i,j}^{(q)} = \sum_{q=1}^{Q}\sum_{k=1}^{K} f((X^{(k)} - O^{(q)})_i, (X^{(k)} - O^{(q)})_j) \quad (9)$$

and a function F is introduced as:

$$F(V_i, V_j) = \sum_{q=1}^{Q} f((V - O^{(q)})_i, (V - O^{(q)})_j) \quad (10)$$

such that the new energy function is

$$\begin{aligned} E(V) &= -\frac{1}{2}\sum_{i=1}^{N+3}\sum_{j=1}^{N+3} w_{i,j} F(V_i, V_j) = \\ &= -\frac{1}{2}\sum_{q_1=1}^{Q}\sum_{q_2=1}^{Q}\sum_{i=1}^{N+3}\sum_{j=1}^{N+3} w_{i,j}^{(q_1)} f((V - O^{(q_2)})_i, (V - O^{(q_2)})_j) \end{aligned} \quad (11)$$

The above expression can be rewritten in the following terms:

$$E(V) = \sum_{q=1}^{Q} E_q^{=}(V) + \sum_{q_1=1}^{Q}\sum_{q_2 \neq q_1} E_{q_1,q_2}^{\neq}(V), \quad (12)$$

where $E_q^{=}$ is the corresponding to the terms with $q_1 = q_2$ and $E_{q_1,q_2}^{\neq}$ is the rest.

V. Capacity of the Network

The capacity of the network is a measure of the amount of patterns that can be introduced into the network such that at the retrieving phase the probability of error does not exceed a threshold, p_e.

Let us suppose that K patterns $\{X^{(k)} : k = 1, ..., K\}$ have been loaded into the network, and that state vector V matchs a stored pattern, $X^{(k_0)}$. Suppose that state V' coincides with V except in one component. Without loss of generality, this component can be assumed to be the first one, that is $V_i = V_i'$ if $i > 1$ and $V_1 \neq V_1'$.

By denoting $D = \Delta E = E(V') - E(V)$ the energy increment between these two states V and V', the pattern $X^{(k_0)}$ is correctly retrieved when pattern V' is introduced into the net if $D > 0$. So, in order to calculate the error probability in the retrieval phase, the probability $P(D < 0)$ must be computed.

But

$$\begin{aligned} D = \Delta E &= \sum_{q=1}^{Q} E_q^{=}(V') + \sum_{q_1=1}^{Q}\sum_{q_2 \neq q_1} E_{q_1,q_2}^{\neq}(V') - \\ &- \left(\sum_{q=1}^{Q} E_q^{=}(V) + \sum_{q_1=1}^{Q}\sum_{q_2 \neq q_1} E_{q_1,q_2}^{\neq}(V) \right) = \\ &= \sum_{q=1}^{Q} \left(E_q^{=}(V') - E_q^{=}(V) \right) + \sum_{q_1=1}^{Q}\sum_{q_2 \neq q_1} \left(E_{q_1,q_2}^{\neq}(V') - E_{q_1,q_2}^{\neq}(V) \right) = \\ &= \sum_{q=1}^{Q} \Delta E_q^{=} + \sum_{q_1=1}^{Q}\sum_{q_2 \neq q_1} \Delta E_{q_1,q_2}^{\neq} \end{aligned} \quad (13)$$

and so we have to compute $D_q = \Delta E_q^{=}$ and $D'_{q_1,q_2} = \Delta E_{q_1,q_2}^{\neq}$.

To this end, we present some technical results which will guide us to the main result of this section, the capacity of the network with multiple reference points. Proofs for these lemmas are given in the Appendix.

Lemma 1. We have

$$D_q = N + 3 - \sum_{i=2}^{N} \phi_i + \sum_{k \neq k_0}\sum_{i=2}^{N+3} \xi_i, \quad (14)$$

where

1) ϕ_i is a random variable with mean $E(\phi_i) = -\frac{1}{2}$ and variance $V(\phi_i) = \frac{3}{4}$, for all i.
2) ξ_i is a random variable with mean $E(\xi_i) = 0$ and variance $V(\xi_i) = 3$, for every i.

Lemma 2. It is

$$D'_{q_1,q_2} = \sum_{i=2}^{N+3} \psi_i - \sum_{i=2}^{N+3} \phi_i^* - \sum_{k \neq k_0}\sum_{i=2}^{N+3} \xi_i^*, \quad (15)$$

where:

1) For $2 \le i \le N$, ψ_i is a random variable with mean $E(\psi_i) = \frac{1}{16}$ and variance $V(\psi_i) = \frac{255}{256}$.
2) For $i > N$, ψ_i is a random variable with $E(\psi_i) = \frac{1}{4}$ and $V(\psi_i) = \frac{15}{16}$.
3) ϕ_i^* is a random variable with mean $E(\phi_i^*) = \frac{1}{16}$ and variance $V(\phi_i^*) = \frac{255}{256}$, for all i.
4) ξ_i^* is a random variable with mean $E(\xi_i^*) = 0$ and variance $V(\xi_i^*) = 3$, for every i.

The exact formula for D is given in the following lemma, since last two results have provided us of the precise expressions for $D_q = \Delta E_q^{=}$ and $D'_{q_1,q_2} = \Delta E_{q_1,q_2}^{\neq}$.

Lemma 3. For $N \ge 30$ and $Q \ge 4$,

$$D = Q(N+3) + \Omega, \quad (16)$$

where Ω is a Gaussian random variable with mean

$$\mu = \frac{Q(9Q+8N-17)}{16} \quad (17)$$

and variance given by

$$\sigma^2 = \frac{3Q[Q(256K(N+2)-86N-187)-106N-389]}{256}. \quad (18)$$

This result allows us to calculate

$$P(D<0) = P(Q(N+3)+\Omega<0) =$$
$$= P(\Omega < -Q(N+3)) = P\left(\frac{\Omega-\mu}{\sigma} < \frac{-Q(N+3)-\mu}{\sigma}\right) \quad (19)$$

Since $Z = \frac{\Omega-\mu}{\sigma}$ is a Gaussian with mean 0, and variance 1, there is a z_α such that $P(Z<z_\alpha)=p_e$. For example, for $p_e = 0.05$, it is $z_\alpha = -1.645$ and for $p_e = 0.01$, it is $z_\alpha = -2.326$. By comparing the last expression to (19), and taking into account that our objective is $P(D<0)=p_e$, we arrive at

$$\frac{-Q(N+3)-\mu}{\sigma} = z_\alpha. \quad (20)$$

The next step is to use the proper definition of the parameter of capacity α. It is the quotient between the number of patterns and the number of neurons which achieve an error probability lower than p_e. So, $\alpha = \frac{K}{N}$, that is, $K = \alpha N$.

By combining (20) and the above expression for *K*, we get the following result:

Theorem 2. The capacity of Hopfield's associative memory with multiple reference points is given by

$$\alpha = \frac{1}{V}\left(\frac{T^2}{z_\alpha^2}+U\right) \quad (21)$$

where:

$$T = -Q(N+3)-\mu$$
$$U = \frac{3Q}{256}[Q(86N+187)+106N+389] \quad (22)$$
$$V = 3Q^2N(N+2)$$

From this theorem, two important corollaries can be stated:

Corollary 1. The capacity of the network increases with the number of reference points.

This results from the fact that (21) can be rewritten in the following terms:

$$\alpha = aQ^2 + bQ + c + \frac{d}{Q}, \quad (23)$$

with $a = \frac{27}{256N(N+2)z_\alpha^2} > 0$. So, for a fixed number of neurons *N*. This implies that, by increasing the number of reference points, capacity greater than 1 may be achieved, as can be verified in Fig. 1.

Corollary 2. There exists a positive constant $\alpha_{\min}$ such that, for fixed *Q*, $\alpha \ge \alpha_{\min}$.

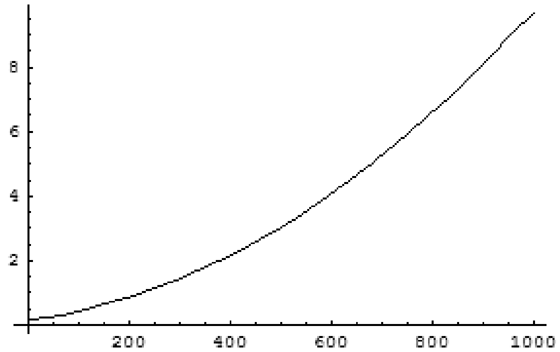

Fig. 1 Capacity of a network with *N*=50 neurons as a function of *Q*, which varies from 1 to 1000

This fact can be proved just by noting that the capacity is a decreasing function of *N*, so we can fix *Q* and calculate the value $\alpha_{\min} = \lim_{N\to\infty} \alpha = \frac{3}{4z_\alpha^2}$ (it does not depend on *Q*). If we consider $p_e = 0.01$, a value of $\alpha_{\min} = 0.1386$ is obtained, and for $p_e = 0.05$, a value of $\alpha_{\min} = 0.2771$.

VI. Conclusions and Future Work

In this paper, Hopfield associative memory has been studied to overcome some of the most important problems or lacks it possesses: spurious patterns and low capacity.

A method to avoid the apparition of spurious patterns has been presented. This method also explains the well-known (and undesirable) phenomenon of storing the opposite of a pattern.

A new technique to increase the network capacity as a content-addressable memory has also been proposed, based on the use of multiple reference points, which contributes many new possibilities of study and research.

Our future work covers several aspects of these methods:

1) Apply the technique of multiple reference points to a multivalued network, as the one described in [5].
2) Find the optimum configuration of $O^{(q)}$ for a given set of patterns (randomly distributed or with a specific distribution), that is, the value of $O^{(q)}$ which discriminates most the patterns.
3) Consider a mix of fixed and random reference points.

Appendix

Proof of Lemma 1.

Let us denote $O^{(q)} = O$.

As the difference between vectors V and V' is only in the first component, in the expression for D_q all terms with $i \ne 1$ and $j \ne 1$ vanish, resulting:

$$D_q = \sum_{i=2}^{N+3}\left(\sum_{k=1}^{K} f((X^{(k)}-O)_1,(X^{(k)}-O)_i)\right)\cdot$$
$$\cdot\left(f((V-O)_1,(V-O)_i) - f((V'-O)_1,(V'-O)_i)\right) \quad (24)$$

Equation (24) can be expanded in the following terms, since $V = X^{(k_0)}$:

$$D_q = \sum_{i=2}^{N+3} f((V-O)_1,(V-O)_i)(f((V-O)_1,(V-O)_i) - \\ - f((V'-O)_1,(V'-O)_i) + \sum_{i=2}^{N+3}\sum_{k\neq k_0} f((X^{(k)}-O)_1,(X^{(k)}-O)_i)\cdot \\ \cdot\left(f((V-O)_1,(V-O)_i) - f((V'-O)_1,(V'-O)_i)\right) \quad (25)$$

Let us compute the value of each one of its terms.

The term $\sum_{i=2}^{N+3} f((V-O)_1,(V-O)_i)^2 = \sum_{i=2}^{N+3} 1 = N+2$.

On the other hand, it is easy to see that the expression $\sum_{i=N+1}^{N+3} f((V-O)_1,(V-O)_i) f((V'-O)_1,(V'-O)_i) = -1$, because for $i > N$ we have $(V-O)_i = (V'-O)_i = m \in \{-2,0,2\}$, $(V-O)_1 = m_1$, $(V'-O)_1 = m_2$, $m_1 \neq m_2$. If $m = m_1$ or $m = m_2$, then the addend equals -1. Otherwise, it is 1.

Let us define

$$\phi_i = f((V-O)_1,(V-O)_i) f((V'-O)_1,(V'-O)_i) \quad (26)$$

for $2 \le i \le N$. Let us suppose that patterns, states as well as reference points are independent random variables distributed uniformly on $\{-1,1\}^N$. Then, ϕ_i is a random variable taking value -1 or +1.

It will take value -1 if:

1) $(V-O)_1 \neq (V-O)_i$ and $(V'-O)_1 = (V-O)_i$, and this happens with probability

$$P = P((V'-O)_1 = (V-O)_i)\cdot \\ \cdot P((V-O)_1 \neq (V-O)_i \mid (V'-O)_1 = (V-O)_i) = \\ = P((V'-O)_1 = (V-O)_i) = P(V_1' - V_i = O_1 - O_i) \quad (27)$$

since we know that $V_1 \neq V_1'$.

But then $P = \sum_{m\in\{-2,0,2\}} P(O_1 - O_i = m) P(V_1' - V_i = m)$.

As all V_i, V_i' and O_i are identically distributed, their pairwise differences will be identically distributed too. Let us calculate this new probability distribution.

It can be easily showed that $P(O_1 - O_i = 0) = \frac{1}{2}$ and that $P(O_1 - O_i = -2) = P(O_1 - O_i = 2) = \frac{1}{4}$.

Thus, we have $P = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{4}\cdot\frac{1}{4} + \frac{1}{4}\cdot\frac{1}{4} = \frac{3}{8}$.

2) $(V-O)_1 = (V-O)_i$ and $(V'-O)_1 \neq (V-O)_i$, and this happens with the same probability as in 1).

So, $P(\phi_i = -1) = 2\cdot\frac{3}{8} = \frac{3}{4}$, hence $P(\phi_i = 1) = 1 - \frac{3}{4} = \frac{1}{4}$.

From these probabilities, we can deduce that ϕ_i has mean $E(\phi_i) = -\frac{1}{2}$ and variance $V(\phi_i) = \frac{3}{4}$.

Now, to calculate the value of the second addend, we consider the random variable:

$$\xi_i = f((X^{(k)}-O)_1,(X^{(k)}-O)_i)\cdot \\ \cdot(f((V-O)_1,(V-O)_i) - f((V'-O)_1,(V'-O)_i)) \quad (28)$$

$\xi_i = -2$ if

1) $(X^{(k)}-O)_1 = (X^{(k)}-O)_i$, $(V-O)_1 = (V-O)_i$, and $(V'-O)_1 \neq (V'-O)_i$ (this term is redundant).

2) $(X^{(k)}-O)_1 \neq (X^{(k)}-O)_i$, $(V-O)_1 \neq (V-O)_i$ (this term is redundant), and $(V'-O)_1 = (V'-O)_i$.

In this case,

$$P(\xi_i = -2) = \\ = P((X^{(k)}-O)_1 = (X^{(k)}-O)_i) P((V-O)_1 = (V-O)_i) + \\ + P((X^{(k)}-O)_1 \neq (X^{(k)}-O)_i) P((V'-O)_1 = (V-O)_i) \quad (29)$$

But, as $P((V-O)_1 = (V-O)_i) = P((V'-O)_1 = (V-O)_i)$, and $P((X^{(k)}-O)_1 = (X^{(k)}-O)_i) + P((X^{(k)}-O)_1 \neq (X^{(k)}-O)_i) = 1$ we have $P(\xi_i = -2) = P((V-O)_1 = (V-O)_i) = \frac{3}{8}$.

An analogous explanation gives $P(\xi_i = 2) = \frac{3}{8}$ and therefore $P(\xi_i = 0) = \frac{1}{4}$. So $E(\xi_i) = 0$ and $V(\xi_i) = 8\cdot\frac{3}{8} = 3$.

Summarizing, we have

$$D_q = N + 3 - \sum_{i=2}^{N}\phi_i + \sum_{k\neq k_0}\sum_{i=2}^{N+3}\xi_i \quad (30)$$

Proof of Lemma 2.

Let us denote $O^{(q_1)} = O$ and $O^{(q_2)} = O'$.

Analogously to the case of D_q, we have

$$D'_{q_1,q_2} = \sum_{i=2}^{N+3} f((V-O)_1,(V-O)_i)(f((V-O')_1,(V-O')_i) - \\ - f((V'-O')_1,(V'-O')_i) + \sum_{i=2}^{N+3}\sum_{k\neq k_0} f((X^{(k)}-O)_1,(X^{(k)}-O)_i)\cdot \\ \cdot\left(f((V-O')_1,(V-O')_i) - f((V'-O')_1,(V'-O')_i)\right) \quad (31)$$

Let ψ_i be the random variable defined by

$$\psi_i = f((V-O)_1,(V-O)_i) f((V-O')_1,(V-O')_i) \quad (32)$$

For $2 \le i \le N$, $\psi_i = -1$ if

1) $(V-O')_1 = (V-O')_i$ and $(V-O)_1 \neq (V-O)_i$.

2) $(V-O')_1 \neq (V-O')_i$ and $(V-O)_1 = (V-O)_i$.

The probability of 2) is the same of 1), so

$$P(\psi_i = -1) = 2P((V-O')_1 = (V-O')_i)\cdot \\ \cdot P((V-O)_1 \neq (V-O)_i \mid (V-O')_1 = (V-O')_i) = \\ = 2\frac{3}{8}\cdot P(O_1' - O_1 \neq O_i' - O_i) = \frac{3}{4}(1 - \frac{3}{8}) = \frac{15}{32} \quad (33)$$

Thus, $P(\psi_i = 1) = 1 - P(\psi_i = -1) = \frac{17}{32}$.

Then, its mean is $E(\psi_i) = \frac{1}{16}$ and its variance is $V(\psi_i) = 1 - \frac{1}{256} = \frac{255}{256}$.

For $i > N$, since $(V-O)_i = (V-O')_i$, it is $\psi_i = -1$ if

1) $(V-O)_1 = (V-O)_i \neq (V-O')_1$.
2) $(V-O)_1 \neq (V-O)_i = (V-O')_1$.

Both 1) and 2) have the same probability. So,

$$P(\psi_i = -1) = 2P((V-O)_1 = (V-O)_i)\cdot$$
$$\cdot P((V-O')_1 \neq (V-O)_i \mid (V-O)_1 = (V-O)_i) =$$
$$= 2\frac{3}{8}\cdot P((V-O)_1 \neq (V-O')_1) = \frac{3}{4}P(O_1 - O_1' \neq 0) = \tag{34}$$
$$= \frac{3}{4}\cdot\frac{1}{2} = \frac{3}{8}$$

and $P(\psi_i = 1) = \frac{5}{8}$. Then, $E(\psi_i) = \frac{1}{4}$ and $V(\psi_i) = \frac{15}{16}$.

Let us consider the random variable

$$\phi_i^* = f((V-O)_1,(V-O)_i)\, f((V'-O')_1,(V'-O')_i). \tag{35}$$

$\phi_i^* = -1$ if:

1) $(V-O)_1 \neq (V-O)_i$ and $(V'-O')_1 = (V-O')_i$.
2) $(V-O)_1 = (V-O)_i$ and $(V'-O')_1 \neq (V-O')_i$.

The probability of 1) is the same as for 2), so

$$P(\phi_i^* = -1) = 2P((V-O)_1 = (V-O)_i)\cdot$$
$$\cdot P((V'-O')_1 \neq (V-O')_i) = 2\cdot\frac{3}{8}\cdot\frac{5}{8} = \frac{15}{32} \tag{36}$$

and so $P(\phi_i^* = 1) = \frac{17}{32}$. Therefore, $E(\phi_i^*) = \frac{1}{16}$ and $V(\phi_i^*) = \frac{255}{256}$.

Let us define now the random variable

$$\xi_i^* = f((X^{(k)}-O)_1,(X^{(k)}-O)_i)\cdot$$
$$\cdot\left(f((V-O')_1,(V-O')_i) - f((V'-O')_1,(V'-O')_i)\right). \tag{37}$$

Its study is completely analogous to the already made for ξ_i. So, we have that $E(\xi_i^*) = 0$ and variance $V(\xi_i^*) = 3$.

Proof of Lemma 3.

By combining (13)-(15), we arrive at the expression:

$$D = Q(N+3) - \sum_{q=1}^{Q}\sum_{i=2}^{N}\phi_i + \sum_{q=1}^{Q}\sum_{i=2}^{N+3}\sum_{k\neq k_0}\xi_i +$$
$$+\sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N}\psi_i + \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=N+1}^{N+3}\psi_i - \tag{38}$$
$$-\sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N+3}\phi_i^* + \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N}\sum_{k\neq k_0}\xi_i^*$$

Let us define the following random variables. They will provide us of an easy way to handle the above expression for D.

$$\Xi = \sum_{q=1}^{Q}\sum_{i=2}^{N+3}\sum_{k\neq k_0}\xi_i + \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N}\sum_{k\neq k_0}\xi_i^*$$

$$\Phi = \sum_{q=1}^{Q}\sum_{i=2}^{N}\phi_i$$

$$\Phi^* = \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N+3}\phi_i^*$$

$$\Psi_1 = \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=2}^{N}\psi_i \tag{39}$$

$$\Psi_2 = \sum_{q_1=1}^{Q}\sum_{q_2\neq q_1}\sum_{i=N+1}^{N+3}\psi_i$$

$$\Omega = -\Phi + \Xi + \Psi_1 + \Psi_2 - \Phi^*$$

If N and Q are large enough to apply the Central Limit Theorem ($N \geq 30$ and $Q \geq 4$ is sufficient), we can affirm that:

1) Ξ is a Gaussian random variable with mean $\mu_1 = 0$ and variance $\sigma_1^2 = 3Q^2(N+2)(K-1)$.
2) Φ is a Gaussian random variable with mean $\mu_2 = \frac{-Q(N-1)}{2}$ and variance $\sigma_2^2 = \frac{3Q(N-1)}{4}$.
3) Φ^* is a Gaussian random variable with mean $\mu_3 = \frac{Q(Q-1)(N+2)}{16}$ and $\sigma_3^2 = \frac{255Q(Q-1)(N+2)}{256}$.
4) Ψ_1 is a Gaussian random variable with mean $\mu_4 = \frac{Q(Q-1)(N-1)}{16}$ and $\sigma_4^2 = \frac{255Q(Q-1)(N-1)}{256}$.
5) Ψ_2 is a Gaussian random variable with mean $\mu_5 = \frac{3Q(Q-1)}{4}$ and $\sigma_5^2 = \frac{45Q(Q-1)}{16}$.

Thus, Ω is a Gaussian random variable whose mean is $\mu = \mu_1 - \mu_2 - \mu_3 + \mu_4 + \mu_5 = \frac{Q(9Q+8N-17)}{16}$ and whose variance is

$$\sigma^2 = \frac{3Q}{256}[Q(256K(N+2)-86N-187)-106N-389].$$

Proof of Theorem 3.

It suffices to combine (20) with $K = \alpha N$, and solve for α.

References

[1] J. J. Hopfield, "*Neural networks and physical systems with emergent collective computational abilities*," Proc. Nat. Acad. Sci. USA, vol. 79, 1982, pp. 2554-2558.

[2] D. O. Hebb, *The Organization of Behavior*, New York. Wiley, 1949.

[3] R. J. McEliece, E. C. Posner, E. R. Rodemich, S. S. Venkatesh, *The Capacity of the Hopfield Associative Memory*, IEEE Transactions on Information Theory, vol. IT-33, no. 4, 1987, pp. 461-482.

[4] J. Hertz, A. Krogh and R. G. Palmer, *Introduction to the theory of neural computation*, Lecture Notes Volume I. Addison Wesley, 1991.

[5] E. Mérida-Casermeiro and J. Muñoz-Pérez, *MREM: An associative autonomous recurrent network*, Journal of Intelligent and Fuzzy Systems 12 (3-4), pp. 163-173, 2002.

