Infinite Dimensional Associative Memories: Foundations (with Theorems)

Enrique Carlos Segura
Department of Computer Science, University of Buenos Aires
Ciudad Universitaria, Pab.I, (1428) Buenos Aires, Argentina
e-mail: esegura@dc.uba.ar
(Paper received on July 06, 2007, accepted on September 1, 2007)

Abstract

We present a formal theoretical background, including theorems and their proofs, for our neural network model with associative memory and continuous topology, i.e. its processing units are elements of a continuous metric space and the state space is Euclidean and in¯nite dimensional. This approach is intended as a generalization of the previous ones due to Little and Hop¯eld. Thus we integrate two levels of continuity: continuous response units and continuous topology of the neural system, obtaining a more biologically plausible model of associative memory. A theoretical framework is provided so as to make this integration consistent. We ¯rst present some general results concerning attractors and stationary solutions. Then we focus on the case of orthogonal memories, proving theorems on their stability, size of attraction basins and spurious states. Finally, we get back to discrete models, i.e. we discuss new viewpoints arising from the present continuous approach and examine which of the new results are also valid for the discrete models.

Keywords: associative memory, continuous topology, dynamical systems, Hop¯eld model, in¯nite dimensional state space, stability.

1 Introduction

In seminal papers, Little [8],[9] and Hop¯eld [6] constructed a content-addressable memory as a dense network of arti¯cial neurons that are represented as elementary bistable processors. Addressability is guaranteed by the dissipative dynamics of the system. It consists of switching each processor from one of its stable con¯gurations to the other as a consequence of the intensity of the local ¯eld acting on it. The memories, corresponding to ¯xed points of the dynamics, are stored in the system in a distributed manner through the matrix of two-body interactions (synaptic e±cacies) between the neurons. If this matrix is properly de¯ned, the above dynamics is enough so as to ensure a monotonic decrease of an \energy" function. Thus, starting from an arbitrary con¯guration the system is led to a local minimum that corresponds to the nearest stored memory.

Special Issue in Neural Networks and Associative Memories
Research in Computing Science 28, 2007, pp. 15-28

In a later paper, Hopfield [7] aimed at a more realistic model by replacing bistable neurons by graded response devices. In fact, a classical objection to the former model [6],[8],[9] was that a two-state representation of the neural output is, from a biological point of view, an oversimplification and that it is necessary to describe relevant neural activity by firing rates, rather than merely by the presence or the absence of an individual spike[1]. In either case the retrieval process is again guaranteed by the nature of the matrix of synaptic efficacies. However, in [7] the space of states describing the patterns of activity remained discrete, in the sense that the number of units was, at most, countable. This was an open gap in the plausibility of the model. In fact, since the Little model was formulated to describe the computational ability of an ensemble of simple processing units, it was necessary to reconcile the biological evidence of a true continuum of the neural tissue with the descriptions provided by discrete models inspired in an Ising system. While the empirical evidence always shows patterns of activity or quiescence involving patches with finite sizes, the ferromagnetic approach suggests systems with discrete processing units with no finite dimensions. In spite of this simplification all the discrete models have been remarkably successful in describing emergent processing abilities that correspond to stylized facts concerning basic elementary cognitive processes.

In this paper we introduce a solid theoretical background, including theorems and their proofs, for our neural network model with associative memory and processing units defined as elements of a continuous metric space (some subsidiary results are omitted by length limitations). This model [12],[13] is intended as a generalization of the previous ones due to Little and Hopfield. Our main purpose is to provide proofs in the sense that it is actually possible to formulate a system of associative memory (AM) with continuous response units and a continuous topological structure on the set of such units. We conceive the network so as to preserve the salient features that made attractive all the discrete models, especially the levels of continuity that the Hopfield model with graded response [7] added to the discrete one [6]: continuous-valued units and continuous scale of time, via the transition from discrete to continuous, differential equation dynamics. In spite of the fact that the corresponding space of configurations is an infinite dimensional functional space, we can define a basic simple dynamics having asymptotic, stationary solutions which can be associated to minima of an energy functional of Lyapunov type and can be taken to represent the memories stored in the system.

We have already introduced in a previous paper [12] several of the results included here, but without any rigorous proof. The present article is devoted to provide theoretical foundations for that sketch. We place emphasis on a detailed analysis of orthogonal memories, a relevant particular case of the general theory (deep comprehension of orthogonal memories is essential to understand the general pseudo-orthogonal case). However, we also present some other more general results.

Some approaches related to ours have appeared in recent years [10],[11]. The concept of *field computation*, introduced by MacLennan, has a similar inspiration since many neural phenomena can be described as a field, i.e. the distribution of some physical quantity over a continuous space, with a topology associated to it. On the other hand, the big number of neurons per sq. milimeter that can be found

[1]However, there is at present an increasing agreement that spiking neurons have some properties for describing certain aspects of neural dynamics not completely covered by rate-coding models.

throughout most of the brain cortex, justifies treatment of neural activity as a field.

All these arguments are related to our approach. However, this is aimed to a different purpose, which is that of formulating an extended model of AM and theoretically founding it, including justification of its potential as a tool for modelling cognitive processes of memory and learning.

Moreover, in another previous paper [13], we have already proposed a generalization of the nondeterministic, finite temperature Glauber dynamics [3] to the case of a finite number of graded response neurons (Hopfield'84). We did this by casting the retrieval process of a Hopfield model with continuous-valued units, into the framework of a diffusive process governed by the Fokker-Plank equation. We thus provided a description of the transitional regime that rules the retrieval process, which is currently disregarded. In other words, we unified the graded response units model [7] and the stochastic approach, obtaining a description of the retrieval process at both the microscopic, individual neuron level and the macroscopic level of time evolution of the probability density function over the space of activation patterns, i.e. an equation describing, for each possible pattern, how, given an initial probability for the system being in it, this probability changes upon time.

The paper is organized as follows. In Sect. 2 we give basic concepts and definitions. Sect. 3 provides general results on attractors and stationary solutions. In Sect. 4, we focus on the orthogonal case, proving theorems on stability of the memories and of the origin. Sect. 5 presents a result on the size of basins of attraction and Sect. 6 deals with spurious states. Finally, in Sect. 7 we get back to discrete models, i.e. discuss new viewpoints arising from the present continuous approach and examine which of the new results are also valid for the discrete models.

2 Basic Definitions and First Results

Assume that $\mathbf{v}(x,t)$ describes the activity of a point-like neuron located in x at time t. This pattern of activity evolves according to:

$$\frac{\partial \mathbf{v}(x,t)}{\partial t} = -\mathbf{v}(x,t) + g_\sigma \left(\int_K \mathbf{T}(x,y)\mathbf{v}(y,t)dy \right) \tag{1}$$

with $\mathbf{v}(x,t) : K \times R_{\geq 0} \rightarrow R$, $K \subset X$. X is a metric space, K a compact domain, g_σ a *sigmoid* function, i.e. $g_\sigma \in C^1(R)$, non decreasing, odd and satisfying $\lim_{x \rightarrow \pm\infty} g_\sigma(x) = \pm \mathrm{V_M}$, $\lim_{\sigma \rightarrow \infty} g_\sigma(x) = sgn(x)\ \forall x \neq 0$, $\mid g_\sigma(x) \mid < \min\{\mathrm{V_M}, \sigma x\}$ and $g'_\sigma(0) = \sigma$.

Let S be the set of all possible states $\mathbf{v}(x)$ (patterns of activity) of the system. Then a solution $\mathbf{v}(x,t)$ fulfilling (1) is a trajectory in S.

We can assume that $\mathrm{V_M} = 1$. As for $\mathbf{T}$:$K \times K \rightarrow R$, we assume it is continuous almost everywhere (a.e.) in order to ensure that the integral is well defined. As a natural extension of the discrete case we introduce the *local field* on (or net input to) the neuron located in x when the state of the system is $\mathbf{v}(y,t)$:

$$h_t^{\mathbf{v}}(x) = \int_K \mathbf{T}(x,y)\mathbf{v}(y,t)dy$$

For $t = 0$ we write $h^{\mathbf{v}}(x) = h_0^{\mathbf{v}}(x)$. Note that $h^{\mathbf{v}}$ is linear in $\mathbf{v}$.

Let $\mathbf{v}_0^{\mu}(x) = \mathbf{v}^{\mu}(x, 0)$ be an initial condition (i.c.) and $\mathbf{v}(x, t)$ the solution of (1) associated to it. We say that $\mathbf{v}^{\mu}(x)$ is a *memory* or an *attractor* if and only if:

1) $\mathbf{v}^{\mu}$ is an equilibrium point, i.e. $\mathbf{v}^{\mu}(x) = g_{\sigma}(h_t^{\mathbf{v}^{\mu}}(x))$ *a.e.*

2) For every $t_0 \geq 0$ and $\mathbf{v}_0$ a different i.c. corresponding to $\mathbf{v}$, there exists $\delta(t_0) > 0$ such that if $\| \mathbf{v}^{\mu} - \mathbf{v}_0 \| < \delta$ then $\| \mathbf{v}^{\mu}(\cdot, t) - \mathbf{v}(\cdot, t) \| \rightarrow 0$ when $t \rightarrow \infty$.

Hence, attractors are stationary solutions of (1). Assume that $\mathbf{S}=L^2(K)$ and that $| K | < \infty$ (K has finite Lebesgue measure).

We define the *energy* of the system at time t_0 as the functional:

$$\mathbf{H}[\mathbf{v}(\cdot, t_0)] = -\frac{1}{2} \int_K \int_K \mathbf{T}(x, y)\mathbf{v}(x, t_0)\mathbf{v}(y, t_0)dxdy + \int_K \int_0^{\mathbf{v}(x,t_0)} g_{\sigma}^{-1}(s)dsdx \qquad (2)$$

where $\mathbf{H}[\mathbf{v}(\cdot, t_0)]$ means that $\mathbf{v}$ is viewed as a function of x. Thus, each $\mathbf{v}$ in $\mathbf{S}$ has an energy $\mathbf{H}(\mathbf{v})$. This is an extension of the energy as defined in [7] for the (discrete) model with graded response functions.

2.1 Attractors and Stationary Solutions

From now on we assume that $\mathbf{T}$ is symmetric.

Theorem 2.1: $\mathbf{H}$ is monotonically decreasing with t and reaches its minima at $\mathbf{v}_{t_e}(x) = \mathbf{v}(x, t_e)$ such that

$$\left[\frac{\partial \mathbf{v}}{\partial t}(x, t)\right]_{t_e} = 0 \qquad (3)$$

a.e. in K (in words, given a solution $\mathbf{v}(x, t)$ corresponding to some i.c., the minima of $\mathbf{H}$ are equilibrium points of the system). This theorem generalizes the classical result for the discrete Hopfield model with graded responses [7] (see e.g. [1],[5]).

Proof: omitted by length limitation.

Memories or attractor states, as defined in this section, satisfy the above conditions. However, the reciprocal implication is not necessarily true: from the previous theorem it does not follow that if a solution $\mathbf{v}(x, t)$ of (1) satisfies condition (3) for some t^*, then $\mathbf{v}(x, t^*)$ is an attractor. For example, the trivial solution $\mathbf{v} \equiv 0$ satisfies it for all t but, as we soon will see, its stability or instability depends on the slope σ of g_{σ} at the origin. In general, the possibility to construct nontrivial memories strongly depends on such parameter.

The sigmoid function g_{σ} plays an important role in determining in which cases the system has nontrivial stationary solutions. A necessary condition is given by:

Theorem 2.2: (existence and uniqueness of the solution) If $\sigma < \frac{1}{\mathrm{M}|K|}$, being M such that $| \mathbf{T}(x, y) | \leq \mathrm{M}$, then the unique stationary solution of (1) is $\mathbf{v} \equiv 0$.

Proof: by definition of g_σ, σ is a Lipschitz constant for it. Then, assuming that $\mathbf{v}^1$ and $\mathbf{v}^2$ are two fixed points of the operator A defined as

$$A\mathbf{v} = g_\sigma \left[\int_K \mathbf{T}(x,y)\mathbf{v}(y)dy \right]$$

we get (using the L^2 norm):

$$\mid A\mathbf{v}^1(x) - A\mathbf{v}^2(x) \mid = \mid g_\sigma(\int_K \mathbf{T}(x,y)\mathbf{v}^1(y)dy) - g_\sigma(\int_K \mathbf{T}(x,y)\mathbf{v}^2(y)dy) \mid$$

$$\leq \sigma \mid \int_K \mathbf{T}(x,y)\mathbf{v}^1(y)dy - \int_K \mathbf{T}(x,y)\mathbf{v}^2(y)dy \mid \leq \sigma M \mid K \mid^{\frac{1}{2}} \parallel \mathbf{v}^1 - \mathbf{v}^2 \parallel$$

being M an upper bound for $\mid \mathbf{T}(x,y) \mid$ (which exists since $\mathbf{T}$ is continuous and K is compact). Finally:

$$\parallel A\mathbf{v}^1 - A\mathbf{v}^2 \parallel \leq \mid K \mid^{\frac{1}{2}} \sup_{x \in K} \mid A\mathbf{v}^1(x) - A\mathbf{v}^2(x) \mid < \sigma M \mid K \mid \parallel \mathbf{v}^1 - \mathbf{v}^2 \parallel$$

Then A is a contraction and has a unique fixed point provided $\sigma < \frac{1}{M \mid K \mid}$.

Besides the condition $\sigma \mathrm{M} \mid K \mid \geq 1$, other ones (see next section) have to be fulfilled in order to ensure the actual existence of nontrivial solutions.

3 Orthogonal Memories, Hebbian Synapses

The case we are specially interested in is the storage of orthogonal memories when the matrix of synaptic weights is Hebbian. This can be achieved through a straightforward generalization of the Hebb rule [4]. Let $\{\mathbf{v}^\mu\}$ be an orthogonal set of functions in some space $S(K)$, that is to say $(\mathbf{v}^\mu, \mathbf{v}^\nu) = 0$ if $\mu \neq \nu$. In principle, $S(K)$ may be noncountable and hence we can define in general:

$$\mathbf{T}(x,y) = \frac{1}{\mid K \mid} \int_P \mathbf{v}^\rho(x)\mathbf{v}^\rho(y)d\rho$$

for $\rho \in P$ some index set. In the case $\{\mathbf{v}^\mu\}$ is an orthogonal set in $L^2(K)$, it is at most countable (provided the separability of $L^2(K)$). Therefore it is natural to restrict to the case in which P is countable:

$$\mathbf{T}(x,y) = \frac{1}{\mid K \mid} \sum_{\mu=1}^{p} \mathbf{v}^\mu(x)\mathbf{v}^\mu(y) \tag{4}$$

Then the following theorem holds:

Theorem 3.1: The system (1), with $\mathbf{T}(x,y)$ defined as in (4), may have any finite number p of orthogonal memories taking values in $\{\mathrm{V}_*, -\mathrm{V}_*\}$, with $g_\sigma(\pm \mathrm{V}_*^3) = \pm \mathrm{V}_*$.

Proof: let p be a positive integer and $\{\mathbf{v}^\mu\}_{\mu=1}^{p} \subset L^2(K)$, $(\mathbf{v}^\mu, \mathbf{v}^\nu) = 0$ if $\mu \neq \nu$, $\mathbf{v}^\mu(x) \in \{-V_*, V_*\}$, $1 \leq \mu \leq p$, $x \in K$, with V_* such that $g_\sigma(\pm V_*^3) = \pm V_*$ (which exists and depends on σ). Defining $\mathbf{T}$ by (4) it holds that, for any μ:

$$g_\sigma(h^\mu(x)) = g_\sigma(\int_K \mathbf{T}(x,y)\mathbf{v}^\mu(y)dy) = g_\sigma(\int_K \frac{1}{|K|}\sum_{\nu=1}^{p} \mathbf{v}^\nu(x)\mathbf{v}^\nu(y)\mathbf{v}^\mu(y)dy)$$

Since the integrals are finite, we can interchange the sum and the integration:

$$g_\sigma(\frac{1}{|K|}\sum_{\nu=1}^{p} \mathbf{v}^\nu(x)\delta_{\nu\mu} \parallel \mathbf{v}^\mu \parallel^2) = g_\sigma(\frac{\mathbf{v}^\mu(x)\, V_*^2\, |K|}{|K|}) = \mathbf{v}^\mu(x)$$

Then, $\mathbf{v}^\mu(x,t) = \mathbf{v}^\mu(x)$ $\forall\, t > 0$ is a fixed point of equation (1).

These solutions $\mathbf{v}^\mu(x)$ look like the example depicted in Figure 1. Activation patterns of this kind agree with the intuitive generalization of the attractors of an Ising-type, spin glass discrete neural network in which patches of full activation alternate randomly with those of full quiescence. They can also be viewed as the vertices of an infinite (noncountable) dimensional hypercube.

A question arising is whether the set of orthogonal fixed points of (1) can be infinite. Note first that it is countable: the elements $\mathbf{v}^\mu$ as they were defined belong to $L^2(K)$, a separable space; hence every orthogonal set in it must be countable. However even an infinite countable number of orthogonal fixed points is not possible while preserving the integrability of $\mathbf{T}$. Observe that if there are k_μ changes of sign in $\mathbf{v}^\mu$ then each term of the form $\mathbf{v}^\mu(x)\mathbf{v}^\mu(y)$ divides the domain $K \times K$ in $(k_\mu + 1)^2$ square regions. Moreover, each region is separated from the next by discontinuity lines because such term takes the constant values $+\mathrm{V}_*^2$ or $-\mathrm{V}_*^2$. If the set of memories is infinite, the number of terms in $\mathbf{T}$ that are added is also infinite, therefore those discontinuity lines are dense at least in a neighborhood of some point, and $\mathbf{T}$ ceases to be piecewise continuous.

Note that theorem 3.1 implies a qualitative difference between discrete and continuous models. Since the memory capacity is now unbounded, there is nothing like a "phase diagram" in which, for a domain K and above some critical number p_c of memories, a transition to a "confusion phase" takes place, implying a rapid degradation of the retrieval ability. While in discrete models the size of the domain is determined by the dimension of the state space ($p_c = \alpha_c N$), in the present case this dimension is infinite and hence there is no p_c.

Discussions on discrete models are mostly in the thermodynamic limit in which either the number of neurons and the number of memories tend to infinity while their ratio is kept constant. This is not possible in the continuous limit, but it is certainly not a problem as far as the biological plausibility of the model is concerned.

We end this section with the following results that are easy to check.

Lemma 3.2: If the memories are orthogonal, the distance between any two of them is always the same.

Corollary 3.3: Orthogonal memories are never dense in $L^2(K)$.

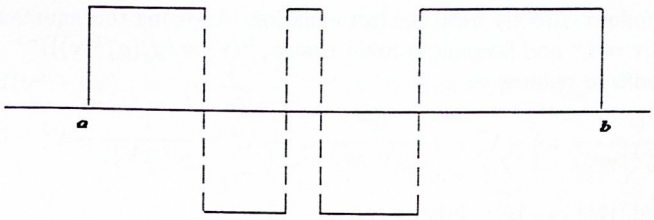

Figure 1: A memory in the space S=$L^2[a,b]$.

4 Stability of the Solutions

We will now derive conditions for the elements $\mathbf{v}^\mu$, as defined in Section 3, to be stable equilibria (i.e. memories) for equation (1).

Theorem 4.1: elements $\mathbf{v}^\mu$ are stable fixed points of (1) if and only if $g'_\sigma(\mathrm{V}_*^3) < \frac{1}{\mathrm{V}_*^2}$.

Theorem 4.2: (Stability of the origin) The solution $\mathbf{v} \equiv 0$ is stable if and only if $g'_\sigma(0) = \sigma < \frac{1}{\mathrm{V}_*^2}$.

Proof (both theorems): the computation of the directional derivatives of $\mathbf{H}(\mathbf{v})$ at an arbitrary point yields:

$$D_w\mathbf{H}(\mathbf{v}) = -\frac{1}{\mid K \mid}\sum_{\nu=1}^{p}(\mathbf{v}^\nu,\mathbf{v})(\mathbf{v}^\nu,w) + (g_\sigma^{-1}(\mathbf{v}),w)$$

with $w \in L^2(K)$ and $\parallel w \parallel = 1$. Now, if $\mathbf{v} = \mathbf{v}^\mu$, using the condition of orthogonality and noting that $\parallel \mathbf{v}^\mu \parallel^2 = \mathrm{V}_*^2 \mid K \mid$, it follows that $D_w\mathbf{H}(\mathbf{v}) = 0$. Similarly, it is easy to check that $D_w\mathbf{H}(\mathbf{v})$ vanishes for any element in $span\,\{\mathbf{v}^\mu\}_{\mu=1}^p$, i.e. linear combinations of the memories, when those combinations take values on $\{\mathrm{V}_*, -\mathrm{V}_*, 0\}$.

$$D_{w^2}^2\mathbf{H}(\mathbf{v}) \;=\; -\frac{1}{\mid K \mid}\sum_{\nu=1}^{p}(\mathbf{v}^\nu,w)^2 + (\frac{\partial}{\partial\mathbf{v}}g_\sigma^{-1}(\mathbf{v})w,w)$$

But the $\mathbf{v}^\nu$ are assumed orthogonal. Therefore, the use of Bessel's inequality yields:

$$\sum_{\nu=1}^{p}\frac{(\mathbf{v}^\nu,w)^2}{\parallel \mathbf{v}^\nu \parallel^2} \leq \parallel w \parallel^2 = 1 \Longleftrightarrow \sum_{\nu=1}^{p}(\mathbf{v}^\nu,w)^2 \leq \mathrm{V}_*^2 \mid K \mid$$

hence

$$D_{w^2}^2\mathbf{H}(\mathbf{v}) \geq (\frac{\partial}{\partial\mathbf{v}}g_\sigma^{-1}(\mathbf{v})w,w) - \mathrm{V}_*^2$$

for any w in S, $\parallel w \parallel = 1$. Then, a necessary and sufficient condition for an element $\mathbf{v}$ in S to be a minimum of $\mathbf{H}$ is:

$$(g_\sigma^{-1\prime}(\mathbf{v})w,w) - \mathrm{V}_*^2 > 0 \;\; \forall\, w \in \mathrm{S}, \parallel w \parallel = 1$$

Theorem 4.2 follows directly from the last equation. Applying this equation to the case in which $\mathbf{v} = \mathbf{v}^\mu$ and keeping in mind that $g_\sigma^{-1\prime}(\mathbf{v}) = (g'_\sigma(g_\sigma^{-1}(\mathbf{v})))^{-1}$ the above condition reduces to

$$(\frac{w}{g'_\sigma(g_\sigma^{-1}(\mathbf{v}^\mu))}, w) - \mathrm{V}_*^2 = (\frac{w}{g'_\sigma(\mathrm{V}_*^2\mathbf{v}^\mu)}, w) - \mathrm{V}_*^2 = \frac{1}{g'_\sigma(\pm\mathrm{V}_*^3)} - \mathrm{V}_*^2 > 0$$

Since g_σ is odd, this can be rewritten as:

$$g'_\sigma(\mathrm{V}_*^3) < \frac{1}{\mathrm{V}_*^2} \text{ or } g'_\sigma(\mathrm{V}_*^3)\mathrm{V}_*^2 < 1$$

Let us compare the necessary and sufficient condition given by theorem 4.2 for the stability of the origin with the uniqueness condition for the general case (theorem 2.2). When $\mathbf{T}$ is defined according to (4), the $\mathbf{v}^\mu$'s being stationary solutions of (1) and therefore $\mathbf{v}^\mu(x) \in \{\mathrm{V}_*, 0, -\mathrm{V}_*\}$, we have:

$$\mid \mathbf{T}(x,y) \mid \leq \frac{p\mathrm{V}_*^2}{\mid K \mid} = \mathrm{M}.$$

In this case the condition for the origin to be the only solution is that $\sigma < \frac{1}{\mathrm{M}\mid K\mid} = \frac{1}{p\mathrm{V}_*^2}$. This is more restrictive than what follows from theorem 4.2. Therefore, for the case of orthogonal memories there exists an intermediate range for the values of σ ($\sigma \in [\frac{1}{p\mathrm{V}_*^2}, \frac{1}{\mathrm{V}_*^2}]$, which degenerates into a point if $p = 1$) for which the trivial solution $\mathbf{v} \equiv 0$ is an attractor, but not necessarily the only solution of (1). Note, in addition, that the conditions derived in theorems 4.1 y 4.2 are independent of p (number of memories); this is a consequence of the orthogonality of the memories.

5 Basins of Attraction

Using the preceeding results, we can now estimate the size of the basins of attraction.

Theorem 5.1: for $p \geq 2$, the largest sphere contained in the basin of attraction of an orthogonal memory $\mathbf{v}^\mu$, $1 \leq \mu \leq p$, has a radius $k = \mathrm{V}_* \sqrt{\frac{\mid K \mid}{2}}$.

In other words, whenever $\| \mathbf{v}^\mu - \mathbf{v}_0 \| < k$, the distance $\| \mathbf{v}^\mu(\cdot,t) - \mathbf{v}(\cdot,t) \| \to 0$ when $t \to \infty$ (being $\mathbf{v}_0$ any i.c. for (1) and $\mathbf{v}$ the corresponding solution).

Proof: the radius of the basin will be the largest number $k > 0$ such that $D_w\mathbf{H}(\mathbf{v}^\mu + kw) > 0 \quad \forall w \in S, \| w \| = 1$. We know that

$$D_w\mathbf{H}(\mathbf{v}) = -\frac{1}{\mid K \mid}\sum_{\nu=1}^{p}(\mathbf{v}^\nu, \mathbf{v})(\mathbf{v}^\nu, w) + (g_\sigma^{-1}(\mathbf{v}), w)$$

Then:

$$\begin{aligned} D_w \mathbf{H}(\mathbf{v}^\mu + kw) &= -\frac{1}{\mid K \mid} \sum_{\nu=1}^{p} (\mathbf{v}^\nu, \mathbf{v}^\mu + kw)(\mathbf{v}^\nu, w) + (g_\sigma^{-1}(\mathbf{v}^\mu + kw), w) \\ &= \frac{1}{\mid K \mid} \left\{ V_*^2 \mid K \mid (\mathbf{v}^\mu, w) + k \sum_{\nu=1}^{p} (\mathbf{v}^\nu, w)^2 \right\} \\ &\quad + (g_\sigma^{-1}(\mathbf{v}^\mu + kw), w) \end{aligned}$$

which is positive if and only if

$$(g_\sigma^{-1}(\mathbf{v}^\mu + kw), w) > \frac{1}{\mid K \mid} \left\{ V_*^2 \mid K \mid (\mathbf{v}^\mu, w) + k \sum_{\nu=1}^{p} (\mathbf{v}^\nu, w)^2 \right\}$$

for every direction w. By virtue of Bessel's inequality:

$$\sum_{\nu=1}^{p} \frac{(\mathbf{v}^\nu, w)^2}{\| \mathbf{v}^\nu \|^2} \leq \| w \|^2 = 1$$

and, consequently (remembering that $\| \mathbf{v}^\nu \|^2 = V_*^2 \mid K \mid$), the condition is satisfied by imposing $(g_\sigma^{-1}(\mathbf{v}^\mu + kw), w) > (g_\sigma^{-1}(\mathbf{v}^\mu), w) + kV_*^2$ or equivalently

$$\int_K \frac{g_\sigma^{-1}(\mathbf{v}^\mu(x) + kw) - (g_\sigma^{-1}(\mathbf{v}^\mu(x))}{k} w(x)\, dx > V_*^2 \tag{5}$$

In order to prove that inequality (5) holds no matter the direction w, let us take the worst case: w pointing to a different memory, say $\mathbf{v}^\nu$, i.e. $w = \frac{\mathbf{v}^\nu - \mathbf{v}^\mu}{\|\mathbf{v}^\nu - \mathbf{v}^\mu\|}$. It is easy to check that $\mathbf{v}^\nu - \mathbf{v}^\mu$ can take only values 0 and $\pm 2V_*$ and that, by virtue of the orthogonality, it is 0 exactly on one half of the domain K and $\pm 2V_*$ on the other half. Then w is 0 on a subdomain of size $\frac{|K|}{2}$ and $\sqrt{\frac{2}{|K|}}$ on the remaining subdomain of equal size. Thus, condition (5) can be rewritten as

$$\frac{\mid K \mid}{2} \frac{g_\sigma^{-1}(\mathbf{v}^\mu(x) + k\sqrt{\frac{2}{|K|}}) - (g_\sigma^{-1}(\mathbf{v}^\mu(x))}{k\sqrt{\frac{2}{|K|}}} \frac{2}{\mid K \mid} > V_*^2$$

(multiplying numerator and denominator by $\sqrt{\frac{2}{|K|}}$), which holds if

$$k\sqrt{\frac{2}{\mid K \mid}} < V_* \quad \Longleftrightarrow \quad k < V_* \sqrt{\frac{\mid K \mid}{2}}$$

Finally, the largest spherical basin of attraction has a radius equal to $V_* \sqrt{\frac{|K|}{2}}$, since otherwise the basins would not be disjoint, because the distance between any two $\mathbf{v}^\mu$ and $\mathbf{v}^\nu$ is twice that quantity.

Note that this result does not imply that the basins of attraction are spherical. It only limits the radius of *spherical* basins for memories $\mathbf{v}^\mu$ and, consequently, for $-\mathbf{v}^\mu$ as well. Figure 2 shows a simplified bidimensional sketch.

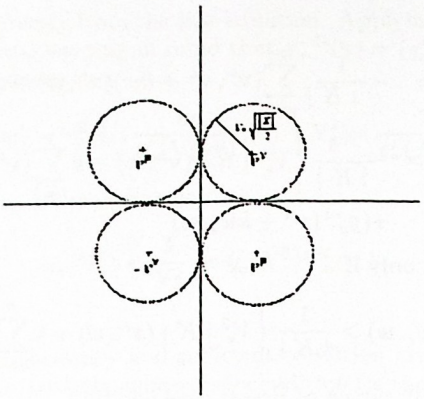

Figure 2: Two orthogonal memories and their inverses, each one with norm $V_* \sqrt{|K|}$ and a spherical basin of attraction of radius $V_* \sqrt{\frac{|K|}{2}}$.

6 Spurious Memories

As a consequence of the nonlinearity of the dynamics under consideration, undesired fixed points appear in addition to those purposely stored in the synaptic operator $\mathbf{T}$ with the Hebb prescription. These are called *spurious states* or *spurious memories*.

It is possible to distinguish two types of spurious states: *mixture* and *non-mixture* memories. $\mathbf{v}$ is said to be a mixture state if it can be expressed as a linear combination of the stored memories: $\mathbf{v} = \sum_{i=1}^{q} \alpha_{\mu_i} \mathbf{v}^{\mu_i}$ with $q \leq p$, $\mathbf{v}^{\mu_i} \in \{\mathbf{v}^{\mu}\}$ and α_{μ_i} real constants. If no such $\{\alpha_{\mu}\}$ exists, the spurious state is said non-mixture.

6.1 Mixture Spurious States

First note that, just like in the known discrete models, for every memory $\mathbf{v}^{\mu}$, $-\mathbf{v}^{\mu}$ is also a memory. In the simple case when $p = 1$, there exist only two spurious states: the origin ($\mathbf{v} \equiv 0$) and the inverse of the (unique) stored memory. Thus, there are no non-mixture states for $p = 1$. If $p \geq 2$, the analysis gets considerably harder.

We have already mentioned the fact that every mixture state is a fixed point if $\mathbf{v}(x) \in \{V_*, -V_*, 0\}$ $\forall x \in \omega$. This can be easily seen either by using the linearity of $h^{\mathbf{v}}$ or from the proof of theorems 4.1 and 4.2. It is also clear that only a small subset of $span\,\{\mathbf{v}^{\mu}\}_{\mu=1}^{p}$ contains spurious states. In particular, it follows that if $\mathbf{v}^{\mu}$ and $\mathbf{v}^{\nu}$ are memories, then $\pm\frac{1}{2}\mathbf{v}^{\mu} \pm \frac{1}{2}\mathbf{v}^{\nu}$ are fixed points of the dynamics. This implies that there exist at least $4\binom{p}{2}$ spurious (mixture) states. These are in general unstable, as stated by the following

Theorem 6.1: If $\mathbf{v}$ is a mixture spurious memory and there exists $x \in K$ such that $\mathbf{v}(x) = 0$, then $\mathbf{v}$ is a saddle point of the dynamics.

Proof: let $\mathbf{v} = \sum_{\mu=1}^{q} \alpha_\mu \mathbf{v}^\mu$, $1 \leq q \leq p$ (renaming memories if necessary). $\mathbf{v}$ is piecewise constant (since so are the $\mathbf{v}^\mu$'s, and there is a finite number of them). Therefore, if $\mathbf{v}(x) = 0$ then it vanishes in a neighborhood U of x and it holds that $0 < \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i} = -\sum_{j=1}^{J} \alpha_{\mu_j} \mathbf{v}^{\mu_j}$ at every point in U, being $\{\mathbf{v}^{\mu_i}\}_{i=1}^{I} \cup \{\mathbf{v}^{\mu_i}\}_{j=1}^{J} = \{\mathbf{v}^{\mu_i}\}_{\mu=1}^{q}$

Let us choose $w = \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i} - \sum_{j=1}^{J} \alpha_{\mu_j} \mathbf{v}^{\mu_j}$ (we can neglect the normalizing constant).

Then $w = 2 \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}$ in U. Now compute

$$g_\sigma(h^{\mathbf{v}+\varepsilon w}(x)) = g_\sigma(h^{\mathbf{v}}(x) + \varepsilon h^w(x))$$

Keeping in mind that $h^{\mathbf{v}}(x) = V_*^2 \mathbf{v}(x)$ (by hypothesis and by virtue of the linearity of $h^{\mathbf{v}}$) and computing

$$h^w(x) = \sum_{\mu=1}^{q} \alpha_\mu h^{\mathbf{v}^\mu}(x) = V_*^2 \sum_{\mu=1}^{q} \alpha_\mu \mathbf{v}^\mu(x) = 2V_*^2 \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}(x)$$

(which holds for every $x \in U$), we obtain

$$g_\sigma(h^{\mathbf{v}+\varepsilon w}(x)) = g_\sigma(V_*^2 \mathbf{v}(x) + 2\varepsilon V_*^2 \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}(x))$$

in U. The dynamics at $\mathbf{v} + \varepsilon w$ is (always in $U \subset K$):

$$\frac{\partial(\mathbf{v} + \varepsilon w)}{\partial t} = -2\varepsilon \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i} + g_\sigma(2\varepsilon V_*^2 \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i})$$

No matter the sign of $\sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}$, the stability condition at $\mathbf{v}+\varepsilon w$ is $g_\sigma(2\varepsilon V_*^2 \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}) < 2\varepsilon \sum_{i=1}^{I} \alpha_{\mu_i} \mathbf{v}^{\mu_i}$ which is equivalent to $\sigma = g'_\sigma(0) < \frac{1}{V_*^2}$ which is false since, by hypothesis, $g_\sigma(\pm V_*^3) = g_{\sigma V_*^2}(\pm V_*) = \pm V_*$ (otherwise, g_σ would not have any fixed point apart from the origin). Then $\mathbf{v}$ is unstable in the direction w.

Now let us choose $w = \mathbf{v}$. With a similar reasoning, we get $g_\sigma(h^{\mathbf{v}+\varepsilon w}(x)) = g_\sigma(V_*^2 \mathbf{v}(1+\varepsilon))$ and the stability condition at $\mathbf{v} + \varepsilon w$ is $g_\sigma(V_*^2 \mathbf{v}(1+\varepsilon)) < \mathbf{v}(1+\varepsilon)$. Remembering that $\mathbf{v} = g_\sigma(V_*^2 \mathbf{v})$, the condition for the system to be stable at $\mathbf{v}+\varepsilon w$ is $g'_\sigma(V_*^3) < \frac{1}{V_*^2}$, which holds since the memories $\mathbf{v}^\mu$ are minima of $\mathbf{H}$ (cf. Sect. 2).

This leads to the useful

Corollary 6.2: The basins of attraction for mixture states have zero radius, in the sense of the $L^2(K)$ norm.

Example: for $q = 2$, all combinations of the form $\mathbf{v} = \pm\frac{1}{2}\mathbf{v}^\mu \pm \frac{1}{2}\mathbf{v}^\nu$ are spurious states (see Fig. 3). The direction of maximum unstability is given by $\mathbf{v}^\mu - \mathbf{v}^\nu$ and that of maximum stability is $\pm\mathbf{v}$ (directly towards or from the origin of coordinates).

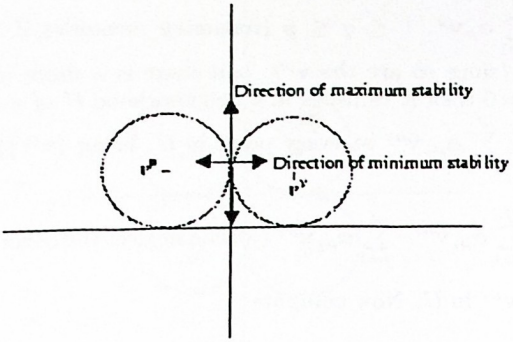

Figure 3: A mixture spurious state. Dotted lines indicate limits for the spherical basins of the two memories which compose the spurious state.

6.2 Non-mixture Spurious States

Unlike mixture spurious states, which can be calculated analytically, the non-mixture ones are difficult to find. Indeed, in the limit $p \to \infty$, the following property holds, no matter how $\mathbf{T}$ is defined.

Lemma 6.3: if $\{\mathbf{v}^\mu\}_{\mu=1}^{\infty}$ is complete[2] and $p \to \infty$, there are no non-mixture spurious memories.

Remark: since S=$L^2(K)$ here, a question about the meaning of lemma 6.3 may arise, i.e. is there some basis of $L^2(K)$ whose elements take on only two values, say $\pm V_*$? The answer is yes. The relevant example for $K \subset R$ (that can be extended to R^n) are the *Haar wavelets*, which form an orthonormal and complete basis in $L^2(-\infty, +\infty)$. They are bi-valued, but since they are normalized, such values change from one function to another. If normality is relaxed, it is possible to force them to take values in $\{V_*, -V_*\}$. If restricted to a bounded interval $K \subset R$, they form an orthogonal (but not orthonormal) and complete set in $L^2(K)$. However, if we construct $\mathbf{T}$ according to (4), this completeness can be only asymptotical: as we saw, the number of orthogonal memories can be as large as desired, but not infinite.

7 Back to the Discrete Domain

In order for the results of Sect. 2 to be valid, the only condition on X is to be a metric space (continuous or discrete). Therefore, all theorems of Sect. 2 hold for the discrete Hopfield model with continuous activities [7], if we replace the $L^2(K)$ norm with the usual euclidean norm and $\mid K \mid$ with N (number of neurons).

Concerning the results of Section 3, the situation is different. Clearly, theorem 3.1 is no longer true and the same happens for lemmas and corollaries based on

[2]The set $\{\mathbf{v}^\mu\}$ is said to be *complete* in $L^2(K)$ if and only if the minimal subspace of $L^2(K)$ which contains it is the entire space $L^2(K)$.

the possibility of memories with an unbounded number of discontinuities, such as corollary 3.3 (of no meaning for the discrete case). Instead, results concerning stability (Sect. 4) and size of attraction basins (Sect. 5) remain valid, with slight changes. The same is true for Section 6 (spurious states), except for theorem 6.3 (non-mixture spurious memories), which has no meaning for the discrete model.

As for the Hopfield model with discrete activities [6], first remark that its metrics based on the Hamming distance is not euclidean. However, being this metrics the discrete version of the L^1 norm, which is equivalent[3] to the quadratic L^2 norm, some results remain qualitatively true, e.g. theorems 2.1 and 2.3; but the mathematical tools used here are of no help to obtain them. And, on the other hand, the results of Sections 3 to 6 have no meaning in general (when based on concepts of euclidean distance and directional derivatives, of no application in the discrete case).

8 Conclusions

We introduced a formal theoretical background, including theorems and their proofs, for our neural network model with AM in which processing units are elements of a continuous metric space. This approach is intended as a generalization of the previous ones due to Little and Hopfield. Our main purpose was to provide a mathematical foundation of the actual possibility to formulate a system that unifies graded response units and continuous topological structure on the set of such units, obtaining a more biologically plausible model of AM.

On the other hand, our approach preserves salient features that made attractive all discrete models, especially the levels of continuity that the second Hopfield model [7] added to the discrete one [6]: graded activation functions and continuous scale of time, via the transition from discrete to continuous, differential equation dynamics.

Firstly (Section 2) general results were proved assuming only a symmetric weight matrix $\mathbf{T}$ with non-negative diagonal elements. These results are generalizations of well known properties of discrete, Ising-type models.

Then (Sections 3 to 6) we analyzed the case of orthogonal memories and a synaptic operator constructed through the autocorrelation (Hebb) rule. We proved:
-*Hebb rule*: it can be naturally extended to the infinite dimensional case.
-*Capacity*: any finite set of orthogonal memories can be stored and retrieved. However there are some differences in capacity with regard to discrete approaches.
-*Stability*: necessary and sufficient conditions for the memories and the origin to be stable, in terms of the relation between parameters of the transfer function g_σ.
-*Basins of attraction*: with the same radius for all memories, positive (L^2 norm).
-*Spurious memories*: they exist. If a spurious state vanishes at some point, then its basin of attraction has zero radius ($\| \ L^2 \ \|$) and it is a saddle point of the dynamics.

We also discussed the validity of these results when applied to the discrete models of AM [6],[7]. Such application looks more natural for the model with graded response [7], as in this case the concepts of euclidean distance and directional derivatives remain valid, while in the discrete case [6] only some general properties (concerning stability and convergence to attractors) are preserved, maintaining anyhow

[3]Two norms $\| \cdot \|$ and $\| \cdot \|'$ on a vector space V are said to be *equivalent* norms if there exist positive real numbers c and d such that $c \| x \| \leq \| x \|' \leq d \| x \| \ \forall x \in V$.

qualitative similarity with the infinite dimensional system.

This approach can be useful for modelling in biology and neurophysiology. It retains all the stylized facts that made attractive the Hopfield neural network and its modifications, yet giving the possibility of modelling the brain cortex as a continuous space. In other words, it integrates two levels of continuity:

-Continuous response units, which was already present in [7] and permits description of relevant neural activity by firing rates, rather than merely by the presence or the absence of an individual spike.

-Continuous topology of the neural system, obtaining a model of AM that reconciles biological evidence of a continuum of the neural tissue with descriptions provided by discrete models inspired in Ising systems.

In addition, the results proved here can be useful, with their limitations (Sect. 7), when performing the reverse track of what we have done, i.e. reconsidering the discrete case through the knowledge of what happens if the state space is continuous.

References

[1] Dreyfus, G.: Neural Networks. Methodology and Applications. Springer (2005)

[2] Fodor, J. A.: The Modularity of Mind. Cambridge, MIT Press (1983)

[3] Glauber R. J.: Time-dependent Statistics of the Ising Model. Journal of Math. Phys. **4** (1963) 294–307

[4] Hebb, D. O.: The Organization of Behavior: A Neuropsychological Theory. New York, Wiley (1949)

[5] Hertz, J., Krogh, A. and Palmer, R. G.: Introduction to the theory of neural computation. Addison-Wesley, Redwood City (1991)

[6] Hopfield, J. J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79** (1982) 2554–2558

[7] Hopfield, J. J.: Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. **81** (1984) 3088–3092

[8] Little, W. A.: The Existence of Persistent States in the Brain. Mathematical Biosciences **19** (1974) 101–120

[9] Little, W. A.: Analytic Study of the Memory Storage Capacity of a Neural Network. Mathematical Biosciences **39** (1978) 281–290

[10] MacLennan, B. J.: Field Computation in Motor Control. In Self-Organization, Computational Maps and Motor Control, ed. by Pietro G. Morasso and Vittorio Sanguineti, Elsevier-North Holland (1997)

[11] MacLennan, B. J.: Field Computation in Natural and Artificial Intelligence. Information Sciences **119** (1999) 73–89

[12] Segura, E. C. and Perazzo, R. P. J.: Associative memories in infinite dimensional spaces. Neural Processing Letters **12** (2) (2000) 129–144

[13] Segura, E. C. and Perazzo, R. P. J.: Biologically Plausible Associative Memory: Continuous Unit Response + Stochastic Dynamics. Neural Processing Letters **16** (3) (2002) 243–257