Strong Attractors of Hopfield Neural Networks to Model Attachment Types and Behavioural Patterns

Abbas Edalat and Federico Mancinelli

Abstract— **We study the notion of a strong attractor of a Hopfield neural model as a pattern that has been stored multiple times in the network, and examine its properties using basic mathematical techniques as well as a variety of simulations. It is proposed that strong attractors can be used to model attachment types in developmental psychology as well as behavioural patterns in psychology and psychotherapy. We study the stability and basins of attraction of strong attractors in the presence of other simple attractors and show that they are indeed more stable with a larger basin of attraction compared with simple attractors. We also show that the perturbation of a strong attractor by random noise results in a cluster of attractors near the original strong attractor measured by the Hamming distance. We investigate the stability and basins of attraction of such clusters as the noise increases and establish that the unfolding of the strong attractor, leading to its break-up, goes through three different stages. Finally the relation between strong attractors of different multiplicity and their influence on each other are studied and we show how the impact of a strong attractor can be replaced with that of a new strong attractor. This retraining of the network is proposed as a model of how attachment types and behavioural patterns can undergo change.**

I. Introduction

The Hopfield model introduced in [1] was the result of a long term quest to develop an artificial neural network for content addressable memory drawn by the notion of Hebbian rule for learning [2]. This rule which was hypothesised in the middle of the last century found experimental support by the mechanism of Long Term Potentiation in early 1970's. The Hopfield model very quickly attracted interest among researchers in various fields because of the simple form of its unsupervised learning and updating rule, and its applications in pattern recognition and solving optimisation problems. It was also particularly appealing because the stored patterns in the network give rise to attractors of a dynamical system governed by an energy function which always decreases with any random asynchronous rule of updating. The Hopfield model and its stochastic extension to the Boltzmann machine had been inspired by and closely resemble the Ising model of ferromagnetism in statistical physics. This allowed long established and powerful mathematical techniques to be used in the analysis of the model [3]. Since the Hopfield network has a low capacity relative to its size and induces the so-called "spurious patterns" different from the stored patterns, most research for technological applications in this area has been focused on improving this relative capacity for random or correlated patterns [4], [5], [6] and unlearning the spurious patterns [7].

An entirely different type of application of the Hopfield network, one that is related to our work, was sought by biologists, psychologists, psychiatrists and sociologists, who tried to use it to obtain a simple conceptual brain model which is based on the Hebbian rule. In this context the size of the capacity of the network is not a crucial issue.

In [8], Francis Crick, the co-discoverer of the DNA and Graeme Mitchison postulated that the function of dream sleep is to remove some undesirable patterns in the brain network and proposed to use Hopfield like networks to examine this hypothesis. The psychiatrist Ralph Hoffman in [9] proposed that the two basic psychotic disorders of the brain namely mania and schizophrenia can be modelled using the Hopfield network. According to this hypothesis mania results from increased random activity in the brain that corresponds to temperature increase in the model whereas schizophrenia results from an overload in memory, misconception and loose associations that corresponds with the spurious states in an overloaded Hopfield network. Attractor neural networks were proposed by the psychiatrist Avi Peled in [10] as the basis for developing a new diagnostic system for mental illness. More generally, these networks have shown to be a useful conceptual tool in understanding brain functions including in the limbic system [11].

Our focus of application here is attachment types and behavioural patterns. Attachment theory, considered today as a main scientific paradigm in developmental psychology, was introduced by John Bowlby [12]. It classifies the quality and dynamics of the relationship of a child with his/her parent into four kinds: secure attachment and three kinds of insecure attachments, namely, avoidant attachment, anxious attachment and disorganised attachments. The particular type of attachment depends crucially on the kind of response by the parent to the child's needs, which is *repeated* thousands of times during the infant's development. The attachment type will then strongly impact on the emotional, cognitive and social development of the child into adulthood by determining the individual's "working model" of relationships. The theory has been corroborated by the so-called strange situation experiment developed by Mary Ainsworth [13] and has also been supported by the findings in developmental neuroscience in the past twenty years. According to Allan Schore, an academic psychotherapist and a leading researcher in neuropsychology, the orbitofrontal areas of the prefrontal cortex, which are densely connected to the limbic system and develop on the basis of the type of interaction infants

Abbas Edalat and Federico Mancinelli are at the Department of Computing, Imperial College London (email: a.edalat@imperial.ac.uk and federico.mancinelli11@imperial.ac.uk). **In the Proceedings of IJCNN 2013.**

have with their primary care-givers, are critically involved in the attachment processes in the first two years of life [14, page 14]. The same paradigm has been emphasised by Louis Cozolino, a clinical psychologist with research interest in neuroscience: "[A]ttachment schemas are a category of implicit social memory that reflects our early experience with care takers. Our best guess is that these schemas reflect the learning histories that shape experience-dependent networks connecting the orbital frontal cortex, the amygdala, and their many connections that regulate arousal, affect and emotion. It is within these neural networks that interactions with caretakers are paired with feelings of safety and warmth or anxiety and fear." [15, page 139].

Attachment Theory has influenced nearly all forms of psychotherapy including psychoanalysis and psycho-dynamic therapy [16], and its impact on Cognitive Behavioural Therapy, as the most widely used type of therapy today, has led to Schema Therapy [17].

Attachment Theory has been studied in computer science and Artificial Intelligence by Dean Peters and his collaborators; in particular in [18] a reactive agent architecture has been designed to model attachment types. Very recently, in [19], the Hopfield model with a variation of the local and iterative learning rule proposed in [4] has been used to design an attachment model for robots.

In their influential and highly praised interdisciplinary book [20, pages 132-144], the three academic psychiatrists Lewis, Amini and Lannon have used artificial neural networks to point out, in a non-technical language, that our attachment types and key emotional attitudes in relation to others are sculpted by *limbic attractors* as a result of *repeated exposure to similar patterns of interactions in childhood*, which will then profoundly impact our emotional world for the rest of our lives.

A similar argument about repetition of a pattern has been implicitly made by three sociologists, Smith, Stevens and Caldwell in [21, page 222], who proposed the Hopfield network to model behavioural prototypes, including any kind of addictions, and working models in attachment theory. Cognitive and behaviourally patterns (including for example substance abuse, habitual physical exercise, gambling and eating) control the neurophysiology underlying feelings associated with distress and can be considered as: "*prototypes*—deeply learned patterns of thought and social activity. In the sense developed by cognitive psychologists, prototypes are cognitive structures that preserve in memory common or typical features of a person's experience. By matching perceptions and thoughts in prototypes stored in memory, persons categorize and identify objects, form inferences and expectations, and construct predictions about the future. Prototypes thus serve an orienting function, since persons use them to guide their behaviour. In general, a person seeks the closest possible match between ongoing experience and these prototype patterns. When confronted with the unfamiliar, a person will search for the closest match to a learned prototype."[21, page 214].

We now ask the following fundamental question. How can we model *repeated exposure to similar patterns* or *deeply learned patterns* is the Hopfield network as a rudimentary model for learning and retrieval of patterns in the brain?

In this paper, we propose the notion of a *strong pattern* of a Hopfield network, namely one that has been multiply stored, to model attachment types and behavioural patterns. The Hebbian learning paradigm in Hopfield networks provides a biologically somewhat plausible and a mathematically simple rule amongst other choices (see for example [5], [22]). We thus consider a Hopfield network with a training set of random patterns except that some of these patterns are multiply stored while others, called simple patterns, are stored once as usual. We show both mathematically and with our simulations that in this setting strong patterns give rise to strongly stable attractors, which we call *strong attractors*, with a large basin of attraction compared with simple patterns. The strong stability of strong patterns and their large basins of attraction matches the robustness of attachment type in children after the first few years of their development and with its long term persistence and impact throughout adulthood. In the same way that the implicit memory of a child affects and to a large extent determines his/her affective and emotional perception of any interaction with other individuals and defines the prism through which the world is observed, the training of a Hopfield network with a strong pattern builds a strong bias for the retrieval of any random pattern towards the strong pattern.

Another main objective in this paper is to model change of attachment types or behavioural prototypes, for example in the case of a child as a result of change of environment and a different kind of parenting, or in the case of an adult after a successful course of psychotherapy or self-help therapy, or an addict after undergoing rehabilitation. In terms of neuroscience this change is made possible thanks to the neuroplasticity of the human brain, which can in a process of learning develop new neural circuits connecting the prefrontal cortex to the limbic system that can regulate strong emotions [23], [24]. After such a retraining, an emotionally significant event will more likely be perceived, interpreted and responded to according to the dynamics of the new circuits rather than the old ones.

Here, a new attachment type is modelled by the creation of a new strong pattern which is strengthened with higher and higher multiplicity to progressively challenge and weaken the old strong pattern by competing more and more effectively for a bigger basin of attraction. Consequently, exposure to a random pattern will more likely retrieve the new stronger attractor rather than the old, as in the case of successful psychotherapy.

In this context we study, both mathematically and by computer simulations, the impact of an increasingly stronger pattern on another fixed strong pattern and also the weakening of a strong attractor as induced by a random perturbation, which unfolds the strong pattern into a number of patterns Hamming-close to it.

We show that when a Hopfield network is trained with these perturbed patterns then with high probability *any* pattern close enough to the strong pattern becomes a fixed point of the network. This provides a cluster of new attractors near the strong pattern, which in classical Hopfield network would be regarded as "spurious" since they do not correspond to stored patterns. However, in the context of applications to attachment types and behavioural patterns, these are natural extensions of the perturbed strong pattern. We call them *generalised stored patterns*, similar to the designation used by Hoffman in the above cited work for a fixed point close but not identical to a stored pattern, and we call the union of their basins the generalised basin of the perturbed strong pattern. We show that for small noise the generalised basin is still relatively large, but increasingly smaller than the basin of attraction of the strong attractor as the noise is increased, and eventually the generalised basin breaks up as the level of noise produces random patterns for the perturbed strong pattern.

The basic mathematical properties of the classical Hopfield network is based on the assumption that the stored patterns are random with respect to each other. This allows the use of the partition function in statistical physics and the Central Limit Theorem in probability theory to deduce results about the storage capacity of the network [3, pages 17-20 and chapter 10]. In the presence of strong patterns or perturbed strong patterns, these mathematical tools can no longer be used. Our mathematical results in particular use a theorem of Lyapunov which provides, subject to what is now called the Lyapunov condition, a generalisation of the Central Limit Theorem to a triangular array of random variables that are random with respect to each other but are not identically distributed [25, pages 368-371].

II. Strong Patterns and Strong Attractors

Assume we have a Hopfield network with N neurons $i = 1, \ldots, N$ with values $S_i = \pm 1$ and p stored patterns ξ^μ, with $1 \leq \mu \leq p$, each given by its components ξ_i^μ for $i = 1, \ldots, N$. We use the generalized Hebbian rule for the synaptic couplings:

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \tag{1}$$

for $i \neq j$ with $w_{ii} = 0$ for $1 \leq i, j \leq N$.

In this paper, we assume we have the deterministic updating rule (i.e., with temperature $T = 0$) and zero bias in the local field:

$$\text{If } h_i \geq 0 \text{ then } 1 \leftarrow S_i \text{ otherwise } -1 \leftarrow S_i$$

where $h_i = \sum_{j=1}^{N} w_{ij} S_j$ is the local field at i. The updating is implemented asynchronously in a random way.

We assume that a given pattern ξ^μ can be *multiply* stored or imprinted in the network and we write its multiplicity or more briefly its *degree*, which is a positive integer, as $d_\mu \geq 1$. This means that there are d_μ patterns that are identical with ξ^μ among the p patterns, in other words ξ^μ occurs exactly d_μ times in the set of p patterns. If $d_\mu > 1$, we say ξ^μ is a *strong pattern* and call the corresponding attractor that can be produced in the network as a *strong attractor*. If $d_\mu = 1$, we refer to ξ^μ as a simple pattern and call the corresponding attractor a simple attractor.

III. Stability of Strong Attractors

Assume that we have n patterns $\xi^1, \ldots, \xi^n$ with degrees $d_1, \ldots, d_n \geq 1$ respectively and that the remaining $p - \sum_{k=1}^{n} d_k \geq 0$ patterns are simple, i.e., each has degree one. Let Λ denote the set of all patterns and note that by our assumptions Λ has $p_0 = p + n - \sum_{k=1}^{n} d_k$ elements. We assume that the patterns in Λ are all independent and identically distributed with equal probability ± 1 for each node.

We compute below the local field for ξ^1 at node i by arranging the contributions from the stored patterns $\xi^\mu \in \Lambda$ in three different groups: (i) $\mu = 1$, (ii) $\mu = k$ for $2 \leq k \leq n$ and (iii) $\mu > n$.

$$\begin{aligned} h_i^1 &= \textstyle\sum_{j \neq i} w_{ij} \xi_j^1 \\ &= \tfrac{1}{N} \textstyle\sum_{j \neq i} \sum_{\mu \in \Lambda} d_\mu \xi_i^\mu \xi_j^\mu \xi_j^1 \\ &= \tfrac{N-1}{N} d_1 \xi_i^1 + \tfrac{1}{N} \textstyle\sum_{k=2}^{n} d_k (\sum_{j \neq i} \xi_j^1 \xi_j^k) \xi_i^k \\ &\quad + \tfrac{1}{N} \textstyle\sum_{j \neq i} \sum_{\mu > n} \xi_i^\mu \xi_j^\mu \xi_j^1 \end{aligned}$$

Similar to the standard treatment of the Hopfield model (see, e.g., [3]), we consider the negation $C_i^1(N)$ of the overlap of ξ_i^1 with $h_i^1 - \frac{N-1}{N} d_1 \xi_i^1$ and swap the order of the above two terms to obtain the following sequence of random variables given in terms of the network size N:

$$\left\{ \begin{aligned} C_i^1(N) &= -\xi_i^1 (h_i^1 - \tfrac{N-1}{N} d_1 \xi_i^1) \\ &= -\tfrac{1}{N} \xi^1 (\textstyle\sum_{j \neq i} \sum_{\mu > n} \xi_i^\mu \xi_j^\mu \xi_j^1) \\ &\quad -\tfrac{1}{N} \xi^1 (\textstyle\sum_{k=2}^{n} d_k (\sum_{j \neq i} \xi_j^1 \xi_j^k) \xi_i^k) \end{aligned} \right. \tag{2}$$

In the first term, there are $(p-n)(N-1)$ random variables of the form

$$-\frac{1}{N} \xi_i^1 \xi_i^\mu \xi_j^\mu \xi_j^1 \tag{3}$$

which we denote as Y_{N1t} with $1 \leq t \leq (p-n)(N-1)$. In the second term, for each $k = 2, \ldots, n$, we have $N-1$ random variables of the form

$$-\frac{1}{N} d_k \xi_i^1 \xi_j^1 \xi_j^k \xi_i^k \tag{4}$$

which we denote by Y_{Nkt} with $2 \leq k \leq n$ and $1 \leq t \leq N-1$. Thus,

$$C_i^1(N) = \sum_{t=1}^{(p-n)(N-1)} Y_{N1t} + \sum_{k=2}^{n} \sum_{t=1}^{N-1} Y_{Nkt} \tag{5}$$

The random variables Y_{N1t} for $1 \leq t \leq (p-n)(N-1)$ and Y_{Nkt} for $2 \leq k \leq n$ and $1 \leq t \leq N-1$ are clearly not identically distributed and therefore the Central Limit Theorem cannot be used as in the classical treatment of the

Hopfield model in [1] to deduce that $C_i^1(N)$ has a normal distribution as $N \to \infty$. However all these random variables are independent, and thus $C_i^1(N)$ forms a *triangular array* of random variables and we can check if the Lyapunov condition holds (for $\delta = 1$), which will guarantee that the sequence $C_i^1(N)$ converges to a normal distribution (see [25, pages 368-371]).

We have:

$$\begin{cases} \mathrm{E}(Y_{N1t}) = 0 & 1 \le t \le (p-n)(N-1) \\ \mathrm{E}(Y_{Nkt}) = 0 & 2 \le k \le n,\ 1 \le t \le N-1 \\ \sigma_{N1t}^2 = \mathrm{E}(Y_{N1t}^2) = 1/N^2 & 1 \le t \le (p-n)(N-1) \\ \sigma_{Nkt}^2 = \mathrm{E}(Y_{Nkt}^2) = d_k^2/N^2 & 2 \le k \le n\, 1 \le t \le N-1. \end{cases} \tag{6}$$

The sum of all the variances is given by:

$$\begin{cases} s_N^2 = \sum_{t=1}^{(p-n)(N-1)} \sigma_{N1t}^2 + \sum_{k=2}^{n} \sum_{t=1}^{N-1} \sigma_{Nkt}^2 \\ = \frac{(p-n)(N-1)}{N^2} + \sum_{k=2}^{n} \frac{(N-1)d_k^2}{N^2} \\ = \frac{N-1}{N^2} \left(p - n + \sum_{k=2}^{n} d_k^2\right). \end{cases} \tag{7}$$

Now put $p_1 = p - n + \sum_{k=2}^{n} d_k^2$ so that $s_N \sim \sqrt{p_1/N}$ as $N \to \infty$. Since $\mathrm{E}(|Y_{N1t}^3|) = 1/N^3$ and $\mathrm{E}(|Y_{Nkt}^3|) = d_k^3/N^3$, we have:

$$\begin{cases} \lim_{N\to\infty} \frac{1}{s_N^3} \sum_{k,t} \mathrm{E}(|Y_{Nkt}|^3) \\ \sim \lim_{N\to\infty} \left(p - n + \sum_{k=2}^{n} d_k^3\right) / (p_1^{3/2}\sqrt{N}) \end{cases} \tag{8}$$

First consider the basic case that p, n and the d_k's (and hence p_1) are all independent of N. Then, the above limit is 0 as $N \to \infty$ and the Lyapunov condition holds. It follows from Lyapunov's theorem that

$$\frac{1}{s_N} C_i^1(N) \sim \mathcal{N}(0,1)$$

as $N \to \infty$, where $\mathcal{N}(0,1)$ is the normal distribution with mean 0 and variance 1; thus for large N:

$$C_i^1 \sim \mathcal{N}(0, p_1/N). \tag{9}$$

Note that when $d_k = 1$ for all $1 \le k \le n$, we obtain $p_1 = p$ and the above distribution is precisely the distribution we obtain for the standard Hopfield network [3, page 18].

Suppose now $d_1 > 1$ and $d_k = 1$ for $2 \le k \le n$ which imply $p_1 = p - 1$. Then the probability Pr_{er} of error, i.e., for ξ_i^1 to change, can be obtained in terms of the normal distribution in Equation 9 or the error function erf as follows.

Theorem 1: (**Stability of strong attractors**) The error probability in the stability of a single strong attractor with degree d_1, as $N \to \infty$, is given by:

$$\begin{aligned} \mathrm{Pr}_{er} &= \sqrt{N/(2\pi p_1)} \int_{d_1}^{\infty} e^{-Nx^2/(2p_1)}\, dx \\ &= \tfrac{1}{2} \left(1 - \mathrm{erf}(d_1 \sqrt{N/2(p-1)}\right) \qquad \square \end{aligned} \tag{10}$$

Note that, for a given p/N, this gives a sharp drop in the probability that ξ_i^1 is changed even for small values of $d_1 > 1$ when compared to $d_1 = 1$. In fact, by Equation 10, even if p is of order of N, so that the classical storage capacity is greatly exceeded, the strong pattern ξ^1 would with high probability be stable if d_1 is large enough. Our first simulation presented in Section VI-B.1 verifies this fact, which shows the strong stability of a single strong attractor is the presence of random attractors. More generally, as long as $\sum_{k=2}^{n} d_k^2$ is not too large compared with p then any strong pattern ξ^1 with $d_1 > 1$ will be strongly stable. For example, if $p_1 = 2p$ then:

$$\mathrm{Pr}_{er} = \frac{1}{2} \left(1 - \mathrm{erf}(d_1 \sqrt{N/4p}\right). \tag{11}$$

On the other hand, we can examine the impact of the presence of stored strong patterns on the stability of a simple stored pattern. For this, we put $d_1 = 1$ with $d_k > 1$ for one or more $k > 1$ so that $p_1 \gg p$. Then, the stability of the simple pattern ξ^1 is compromised as

$$\mathrm{Pr}_{er} = \frac{1}{2} \left(1 - \mathrm{erf}(\sqrt{N/2p_1}\right) \gg \frac{1}{2} \left(1 - \mathrm{erf}(\sqrt{N/2p}\right). \tag{12}$$

Finally, consider the case that p, n and the d_k's do actually depend on N. In this case, the limit in Equation 8 will still be 0 if for example

$$p - n + \sum_{k=2}^{n} d_k^3 = o(\sqrt{N}),$$

as $N \to \infty$ and we can again estimate the error probability in the stability of the patterns.

For the rest of this paper, however, we will always assume for simplicity that p, n and the d_k's are constant, independent of N.

IV. Basin of Attraction

Recall that the Hopfield network has an energy function which in a given state $S \in \{-1, 1\}^N$ has the value

$$H(S) = -\frac{1}{2} \sum_{i \ne j} w_{ij} S_i S_j \tag{13}$$

With the asynchronous updating rule, the energy will always decrease until the network comes to a fixed state at a local minimum of the energy function.

Assume the same network as in Section III. Then, expanding the value of the matrix w_{ij} and organising the terms into three groups ($\mu = 1$, $\mu = k$ for $2 \le k \le n$, and $\mu > n$) as in Equation 1, the energy level for pattern ξ^1 is given by:

$$\begin{cases} H(\xi^1) &= -\frac{d_1(N-1)}{2} - \frac{1}{2N} \sum_{i \ne j} \sum_{\mu > n} \xi_i^\mu \xi_j^\mu \xi_i^1 \xi_j^1 \\ & \quad -\frac{1}{2N} \sum_{k=2}^{n} d_k \left(\sum_{i \ne j} \xi_i^k \xi_j^k \xi_i^1 \xi_j^1\right) \end{cases} \tag{14}$$

The two sums have terms similar to Equation 3, denoted by Y_{N1t}, and Equation 4, denoted by Y_{Nkt} with $2 \le k \le n$, respectively; however the number of terms in each sum is now multiplied by N. Thus, we have:

$$H(\xi^1) + d_1(N-1)/2 = \sum_{t=1}^{(p-n)N(N-1)} Y_{N1t} + \sum_{k=2}^{n} \sum_{t=1}^{N(N-1)} Y_{Nkt} \tag{15}$$

We again have a triangular array of random variables each with mean zero. This time the sum of variances yields:

$$\begin{cases} s_N^2 &= \sum_{t=1}^{(p-n)N(N-1)} \sigma_{N1t}^2 + \sum_{k=2}^{n} \sum_{t=1}^{N(N-1)} \sigma_{Nkt}^2 \\ &= \frac{N(N-1)}{4N^2} p_1 \sim p_1/4 \end{cases} \tag{16}$$

as $N \to \infty$, where as before $p_1 = p - n + \sum_{k=1}^{n} d_k^2$ and we have put $\sigma_{N1t}^2 = \mathrm{E}(Y_{N1t}^2)$ and $\sigma_{Nkt}^2 = \mathrm{E}(Y_{Nkt}^2)$ for $2 \leq k \leq n$. The Lyapunov condition for $\delta = 1$ can be easily checked to hold when p, n and d_k's are independent of N. We thus obtain:

Theorem 2: (**Energy distribution of attractors**) The probability distribution of the energy $H(\xi^1)$ of a strong attractor with degree d_1, as $N \to \infty$, is given by

$$H(\xi^1) + \frac{d_1 N}{2} \sim \mathcal{N}(0, p_1/4) \qquad \square \tag{17}$$

Therefore, for large N, the energy level of a strong pattern is on average lower than that of a simple pattern by a multiplicative factor d, the degree of the strong pattern. We can now show that the strongest strong pattern (i.e., one with the highest degree) gives rise to a large basin of attraction for the induced strong attractor.

Assume d_k's are in descending order of magnitude with $d_1 > d_2 \geq \ldots$. Given a pattern ξ^* and a positive probability $q < 1$, we say ξ^0 is a *random perturbation* of ξ^* by q if $\Pr(\xi_i^0 \neq \xi_i^*) = q$ (and thus $\Pr(\xi_i^0 = \xi_i^*) = 1 - q$) for each $i = 1, \ldots, N$. Note that on average the Hamming distance between ξ^* and ξ^0 (i.e., the number of nodes with different values in the two) is Nq. We call ξ^* the root pattern for ξ^0; we will extensively use these notions in the next section as well.

Assume ξ^0 is a random perturbation of the strong pattern ξ^1 by $q \ll 1$. We will show that ξ^0 is, with high probability, in the basin of attraction of ξ^1. For this, we compute the energy function at the configuration ξ^0 and, as before, split the terms into three groups:

$$\begin{cases} H(\xi^0) &= -\frac{d_1}{2N} \sum_{i \neq j} \xi_i^1 \xi_j^1 \xi_i^0 \xi_j^0 \\ & \quad -\frac{1}{2N} \sum_{i \neq j} \sum_{\mu > n} \xi_i^\mu \xi_j^\mu \xi_i^0 \xi_j^0 \\ & \quad -\frac{(N-1)}{2} \sum_{k=2}^{n} \sum_{i \neq j} d_k \xi_i^k \xi_j^k \xi_i^0 \xi_j^0 \end{cases} \tag{18}$$

Consider the first sum. For $i \neq j$, the two random variables $\xi_i^1 \xi_i^0$ and $\xi_j^1 \xi_j^0$ are independent with equal distribution given by

$$\begin{cases} \Pr(\xi_i^1 \xi_i^0 = 1) = 1 - q \\ \Pr(\xi_i^1 \xi_i^0 = -1) = q \end{cases} \tag{19}$$

We thus have:

$$\begin{cases} \mathrm{E}(\xi_i^1 \xi_i^0) = (1-q) - q = 1 - 2q \\ \mathrm{E}(\xi_i^1 \xi_i^0 \xi_j^1 \xi_j^0) = (1-2q)^2 \text{ for } i \neq j. \end{cases} \tag{20}$$

We subtract the sum of the means of the terms in the first sum from both sides of Equation 18 and, as in Equation 5, rewrite the terms of the three sums as:

$$\begin{cases} H(\xi^0) + \frac{(N-1)d_1(1-2q)^2}{2} = \sum_{t=1}^{N(N-1)} Y_{N0t} \\ + \sum_{t=1}^{(p-n)N(N-1)} Y_{N1t} + \sum_{k=2}^{n} \sum_{t=1}^{N(N-1)} Y_{Nkt} \end{cases} \tag{21}$$

where Y_{N0t} for each t is of the form $-\frac{d_1}{2N} \xi_i^1 \xi_j^1 \xi_i^0 \xi_j^0 + \frac{d_1(1-2q)^2}{2N}$. The required expected values relating to the three sums in Equation 21 for checking the Lyapunov condition (for $\delta = 1$) are given by:

$$\mathrm{E}(Y_{N0t}) = 0,\ \mathrm{E}(Y_{N1t}) = 0,\ \mathrm{E}(Y_{Nkt}) = 0$$

$$\mathrm{E}(Y_{N0t}^2) = \tfrac{d_1^2 q(1-q)}{N^2},\ \mathrm{E}(Y_{N1t}^2) = \tfrac{1}{4N^2},\ \mathrm{E}(Y_{Nkt}^2) = \tfrac{d_k^2}{4N^2}$$

$$\mathrm{E}(|Y_{N0t}|^3) = O(\tfrac{1}{N^3}),\ \mathrm{E}(|Y_{N1t}|^3) = \tfrac{1}{8N^3},\ \mathrm{E}(|Y_{Nkt}|^3) = \tfrac{d_k^3}{8N^3}$$

This gives as $N \to \infty$

$$s_N^2 \sim q(1-q)d_1^2 + \frac{1}{4}\left((p-n) + \sum_{k=2}^{n} d_k^2\right) \tag{22}$$

It is easily checked that

$$\frac{1}{s_N^3} \sum_{k,t} \mathrm{E}(|Y_{Nkt}|^3) = O(1/N).$$

Thus the Lyapunov condition holds and for large N we have:

Theorem 3: (**Energy of perturbations of attractors**) The probability distribution of the energy $H(\xi^0)$ of a random perturbation ξ^0 by q of a strong attractor with degree d_1, as $N \to \infty$, is given by:

$$H(\xi^0) + \frac{d_1(1-2q)^2 N}{2} \sim \mathcal{N}(0, p_2/4), \tag{23}$$

where

$$p_2 = 4q(1-q)d_1^2 + ((p-n) + \sum_{k=2}^{n} d_k^2) \qquad \square$$

Now if q is small enough so that $((1-2q)^2 d_1 > d_2$, then

$$m = \lfloor (d_1(1-2q)^2 - d_2)N/(\sqrt{p_1} + \sqrt{p_2}) \rfloor$$

is positive and is large for large N. It also follows that

$$\frac{d_1(1-2q)^2 N}{2} + m\frac{\sqrt{p_2}}{2} \leq -m\frac{\sqrt{p_1}}{2} - \frac{d_2 N}{2}. \tag{24}$$

Note that $\mathrm{E}(H(\xi^2)) = -d_2 N/2$ by Equation 17 with d_1 replaced with d_2, whereas $\mathrm{E}(H(\xi^0)) = -d_1(1-2q)^2 N/2$ by Equation 23. Therefore, by the above two Equations, we deduce that the energy of a random perturbation ξ^0 of ξ^1 by q is less than the energy of ξ^2 and thus that of any other stored patterns with probability at least

$$\begin{aligned} &\Pr\left(H(\xi^0) \leq \mathrm{E}(H(\xi^0)) + m(\sqrt{p_2}/2)\right) \times \\ &\Pr\left(H(\xi^2) \geq \mathrm{E}(H(\xi^2)) - m(\sqrt{p_1}/2)\right) \\ &= \left(\tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{m} e^{-x^2/2}\, dx\right)^2 = \tfrac{1}{4}(1 + \mathrm{erf}(m/\sqrt{2}))^2 \end{aligned} \tag{25}$$

Since the energy of any state decreases at each step of updating, we conclude that with at least the above probability ξ^0 is in the basin of attraction of ξ^1. Putting $\ell = \lfloor Nq \rfloor$, we deduce:

Corollary 4: (**Size of basin of attraction**) The size of the basin of the strong pattern ξ^1, as $N \to \infty$, is at least

$$\sum_{r=1}^{\ell} \binom{N}{r}, \tag{26}$$

with probability at least $\frac{1}{4}(1 + \mathrm{erf}(m/\sqrt{2}))^2$. $\square$

V. Cluster of Attractors

We now consider the random perturbation of a strong pattern and show how this gives rise to a cluster of attractors close to the strong pattern with respect to the Hamming distance.

We assume initially there are $p+1-d$ random patterns such that one is a single strong pattern ξ^* with degree $d>1$ and the other $p-d$ patterns are all simple, i.e., $n=1$ in the setting of Section III. Consider d independent random perturbations of ξ^* by $q<1$ which, by relabelling we denote by ξ^μ with $1\leq\mu\leq d$. These d simple patterns form a cluster of patterns near the root of the cluster which is the strong pattern ξ^*. The Hopfield network is trained with the simple patterns in this cluster as well as the other original $p-d$ simple patterns which are labelled as ξ^μ for $d+1\leq\mu\leq p$. Our objective is to study in this trained network the stability of any random perturbation ξ^0 of ξ^* by $q_0<1$.

We evaluate the overlap $h_i^0\xi_i^0$ where h_i^0 is the local field at node i for ξ^0 and and separate the contributions of the perturbed patterns ξ^μ with $1\leq\mu\leq d$ in the synaptic couplings w_{ij} from the rest:

$$\left\{\begin{array}{l} h_i^0\xi_i^0=\sum_{j\neq i}\frac{1}{N}(\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu)\xi_j^0\xi_i^0\\ =\frac{1}{N}\sum_{\mu=1}^{d}\sum_{j\neq i}(\xi_j^\mu\xi_j^0)(\xi_i^\mu\xi_i^0)\\ +\frac{1}{N}\sum_{j\neq i}\sum_{\mu>d}(\xi_j^\mu\xi_j^0)(\xi_i^\mu\xi_i^0)\\ =\sum_{t=1}^{d(N-1)}Y_{N1t}+\sum_{t=1}^{(p-d)(N-1)}Y_{N2t}\end{array}\right. \tag{27}$$

For each j and $1\leq\mu\leq d$, using the identity

$$\left\{\begin{array}{lll}\Pr(\xi_j^\mu\xi_j^0=1) & = & \Pr(\xi_j^\mu\xi_j^0=1|\xi_j^*=1)\Pr(\xi_j^*=1)+\\ & & \Pr(\xi_j^\mu\xi_j^0=1|\xi_j^*=-1)\Pr(\xi_j^*=-1),\end{array}\right. \tag{28}$$

we can easily deduce that

$$\left\{\begin{array}{l}\Pr(\xi_j^\mu\xi_j^0=1)=qq_0+(1-q)(1-q_0)\\ \Pr(\xi_j^\mu\xi_j^0=-1)=q(1-q_0)+(1-q)q_0\end{array}\right. \tag{29}$$

Thus, we have

$$\left\{\begin{array}{lll}\mathrm{E}(\xi_j^\mu\xi_j^0) & = & qq_0+(1-q)(1-q_0)\\ & & -(q(1-q_0)+(1-q)q_0)\\ & = & (1-2q)(1-2q_o),\end{array}\right. \tag{30}$$

which by the independence of the random variables $\xi_j^\mu\xi_j^0$ for different j's, implies

$$\mathrm{E}(\xi_j^\mu\xi_j^0\xi_i^\mu\xi_i^0)=(1-2q)^2(1-2q_0)^2. \tag{31}$$

Thus, the variance is given by:

$$\mathrm{E}(\xi_j^\mu\xi_j^0\xi_i^\mu\xi_i^0)^2-(\mathrm{E}(\xi_j^\mu\xi_j^0\xi_i^\mu\xi_i^0))^2=1-(1-2q)^4(1-2q_0)^4. \tag{32}$$

We can now compute the sum of variances in Equation 27:

$$s_N^2\sim p_3/N, \tag{33}$$

where

$$p_3=d(1-(1-2q)^4(1-2q_0)^4)+p-d.$$

It follows that the Lyapunov condition (with $\delta=1$) holds again as:

$$\frac{1}{s_N^3}(\sum_{t=1}^{d(N-1)}E(|Y_{N1t}|^3+\sum_{t=1}^{(p-d)(N-1)}E(|Y_{N2t}|^3))=O(N^{-1/2})$$

We conclude that

Theorem 5: (**Fixed points near perturbed attractors**) The error probability of stability of a random perturbation ξ^0 by q_0 of a strong pattern with degree d when the pattern is replaced with d independent random perturbations of it by q is, as $N\to\infty$, given by:

$$h_i^0\xi_i^0-(1-2q)^2(1-2q_0)^2d\sim\mathcal{N}(0,p_3/N) \qquad \square \tag{34}$$

Assume now that $q,q_0\ll 1$ so that we have:

$$h_i^0\xi_i^0-(1-4(q+q_0))d\sim\mathcal{N}(0,((p-8d(q+q_0)/N). \tag{35}$$

The random perturbation ξ^0 will be a fixed point if $h_i^0\xi_i^0>0$ which holds with a high probability even for small values of $d>1$. For example if $q+q_0=1/8$ then even with $d=2$ we obtain

$$h_i^0\xi_i^0-1\sim\mathcal{N}(0,(p-2)/N) \tag{36}$$

and from the classical Hopfield model we know that ξ^0 will be a fixed point with high probability if $p/N\leq 0.13$. This probability increases sharply for higher values of d.

How many fixed points ξ^0 do we obtain with high probability as random perturbations by q_0 of ξ^* for a given value of $q\ll 1$? For any $q_0\leq q$, on average ξ^0 will differ from ξ^* by Nq_0 nodes. Thus if $\ell=\lfloor Nq\rfloor$ then the number of random perturbations is again given by Equation 26, showing that we indeed have a cluster of attractors near the perturbed strong attractor. We call these attractors *generalised stored patterns*, following a similar designation in [9], and we call the union of the basins of these attractors the *generalised basin* of the perturbed strong pattern ξ^*. As pointed out in the Introduction, for applications in attachment theory and behavioural patterns these generalised memory states are quite natural.

Our result in this section can be extended to a general Hopfield network that has both a number of strong attractors and a number of clusters of attractors each being a random perturbation of a different strong attractor. Due to space limitation here, this is done in the full version of the paper.

VI. Simulations

In this section we give a description of the simulations carried out during our study. Our experiments involve Hopfield networks with asynchronous updating rule that we train with both simple and strong attractors, perturbed as well as unperturbed.

A. Methodology

All our simulations were performed using custom software written in MatLab[1].

[1]MatLab R2012b (The MathWorks)

In all simulations, we introduce a variable number of simple attractors, which we specify individually. This is to underline that all results on strong attractors are resilient to the introduction of other random patterns in the network. We can see how the presence of random simple attractors affects strong attractors in the simulation presented in Fig. 1.

The networks we use for our simulations consist of 500 units. There is no specific pattern we choose as the root for a strong attractor: roots are always random binary strings. When we have two strong attractors stored in our network, we choose one to be random and the other one to have a certain hamming distance from the first (in all our simulations, the roots differ in 150 random locations).

In the section devoted to competing strong attractors, we look at the recalled pattern (the pattern the network converges to) and take its Hamming distance from the root of two strong attractors; in doing this we start the network from a random initial configuration and then take the average distance recorded over 10 trials.

The details for each simulation are given in each individual description.

B. One strong attractor

1) Basin size for strong attractors: The basin size for a single strong attractor rises very steeply even in the presence of other random patterns in the network. In this first simulation we trace the dependency between the degree of a strong attractor and its basin size, that is, the size of the basin of attraction of the root pattern r. We use an increasing number of simple, random patterns (namely, 200, 400, 800 and 1600) showing that even in the presence of a very large number of other random attractors, strong attractors with a large enough degree can still be learnt (Fig. 1). This is in accordance with our mathematical results in section III, where we prove the stability of strong attractors in the presence of simple attractors. The details of the simulation in Fig. 1 are as follows:

(i) For all d values in $\{1, 2, ..., 40\}$
(ii) For $r = \{200, 400, 800, 1600\}$
(iii) Train the network with a strong attractor with degree d, and a fixed number r of random simple attractors
(iv) Measure the size of the basin of attraction for the strong attractor

2) Basin size for perturbed strong attractors: Recall that a strong attractor of degree d is perturbed when each node in each of its d identical copies is flipped with some probability q (noise). We consider values of probabilities ranging from $q = 0$, equivalent to an unperturbed strong attractor, to $q = 0.5$, where patterns become random.

The simulation in Fig. 2 was carried out as follows:

(i) For all d values in $\{1, 2, ..., 15\}$
(ii) For all q values from 0 to 0.5, with a step size of 0.025
(iii) Train the network with a strong attractor with degree d, perturbed with noise q, and a fixed number of random simple attractors (100 for this simulation)

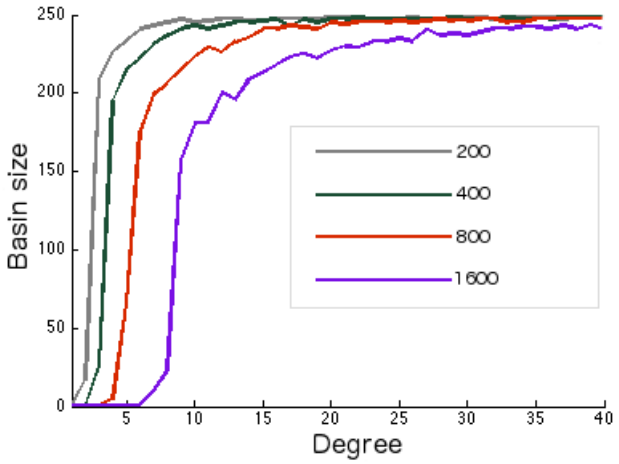


Fig. 1. *Basin size for a strong attractor in presence of other random patterns* We show the basin size for a strong attractor in presence of a variable number of random simple attractors drawn in four different colours. Strong attractors can always be learnt, provided they have a large enough degree, even in the presence of very high number of random patterns.

(iv) Take the average size of the basin of attraction for all patterns belonging to the strong attractor.

The simulation shows a sharp decrease in the basin size when $q > 0$, due to the strong disruptive interference arising among very similar patterns. However, this does not tell us anything about the number of generalised stored patterns that are Hamming-close to the root and are stored in our network. These patterns may have a very low basin size, but they are quite a few. In fact, when we start the network from a random initial configuration, the recalled pattern will be, with high probability, very close to the root (figure 5). This is in accordance with the result shown in Section V.

After shrinking to small values, the basins start broadening again around $q \sim 0.3$. This is because the patterns belonging to the strong attractor interfere with each other less and less as the value of q builds up, and thus the network offers more space for their basins. However, as q increases, the strong attractor starts losing its "identity" , that is the Hamming distance of the patterns from the root pattern starts being considerably high (figure 2).

C. Two strong attractors

In this section we explore the dynamics of a network storing two competing strong attractors. We will see that both the degrees and the values of noise affect the dominance of a strong attractor over another.

1) Competing for basin size - the role of degrees: When two strong attractors are stored, their degrees play a major role in the size of their basins of attraction. As intuition suggests, in the absence of noise, the greater the degree the larger the basin size. The simulation carried out confirms this statement (Fig. 3). The details are the same as for simulation 1, except for the presence of two strong attractors with degrees in $\{1, 2, ...20\}$ and a fixed number of random patterns (100).

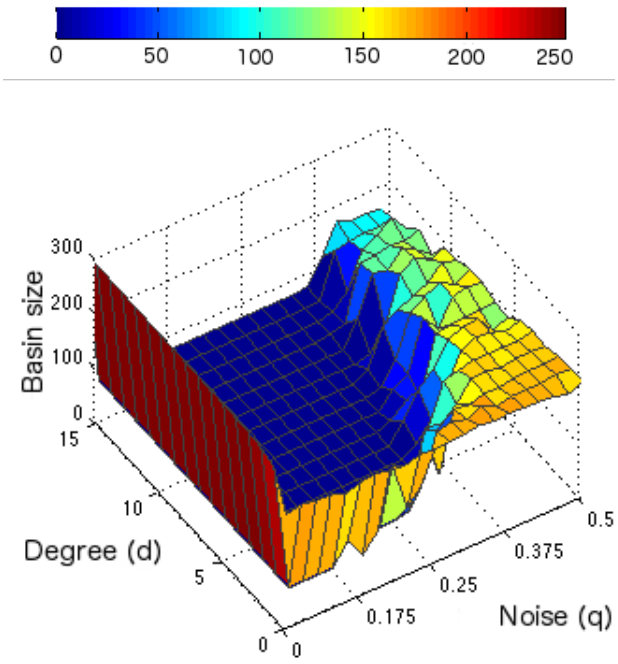


Fig. 2. *Basin size for a perturbed strong attractor in the presence of other random patterns* We show the basin size for a perturbed strong attractor in the presence of a variable number of random simple attractors with varying degrees. A perturbed strong attractor has a very low basin size, but as we will see, this does not compromise its retrieval. Notice how the average basin size is lower for an unfolded strong attractor with higher degree, and how when the degree equals 1 the noise does not affect basin size. The reader is invited to consult the colour scale placed above when inspecting this and the remaining plots.

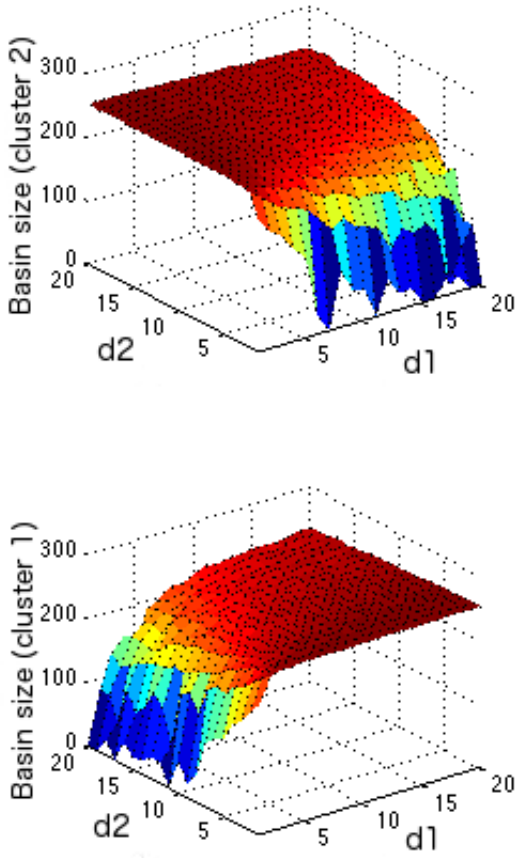


Fig. 3. *The role of degrees in competing strong attractors*. The higher degree provides a larger basin size. On both figures, *d1* and *d2* are the respective degrees of the two strong attractors. The z axis indicates, for the top figure, the basin size of cluster 2, while for the bottom Fig. the basin size of cluster 1.

In the case of strong attractors, a larger size of the basin of attraction ensures that the network will generally converge to a pattern close to the root (Fig. 4, the details of the simulation follow).

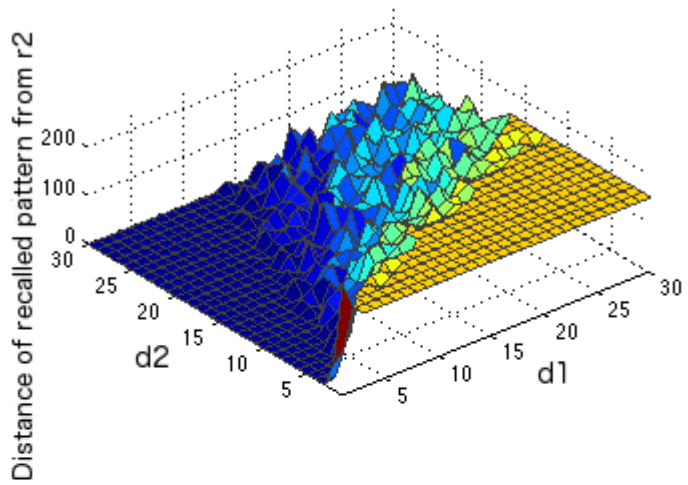
(i) For all d_1 values in $\{1, 2, ..., 20\}$ and d_2 values in $\{1, 2, ..., 20\}$
(ii) Train the network with two strong attractors with degrees $d1$ and $d2$ plus a fixed number of random simple attractors (100 for this simulation)
(iii) Measure the Hamming distance between the recalled pattern and the root of both strong attractors, when the network is started with a random configuration. Repeat for 10 runs, and take the average.

2) Perturbed strong attractors - how noise affects basin size and recalled pattern: When a Hopfield network learns two strong attractors, it will favour the one with higher degree, or when the degree is the same, the one endowed with a lower value of noise q. This means that starting with a random initial configuration and over a sufficiently large amount of trials, the network will converge to a pattern that is closer to the root of the less noisy perturbed strong attractor. This can be seen through our simulations in Fig. 4 and Fig. 5. It does not matter, as expected, that the basins of the strong attractors are very small. The simulation shows that when the network is bombarded with random patterns, it will still converge to a pattern that is close to the root of the less noisy perturbed strong attractor. The details of the simulation in Fig. 5 are as follows:

(i) For all q_1 values in $\{0, 0.01, ..., 0.5\}$ and q_2 values in $\{0, 0.01, ..., 0.5\}$ with $d_1 = 30$, and $d_2 = 30$
(ii) Train the network with two strong attractors with degree $d1$ and $d2$, perturbed with values q_1 and q_2, plus a fixed number of random simple attractors (100 for this simulation)
(iii) Measure the Hamming distance between the recalled pattern and the root of both strong attractors, when the network is started from a random configuration. Repeat for 10 runs and take the mean

D. Unfolding phases of a strong attractor

From Fig. 2 we observe that, when the network learns a strong pattern, the basin size of the strong attractor is very high. On the other hand, when it learns a cluster of patterns



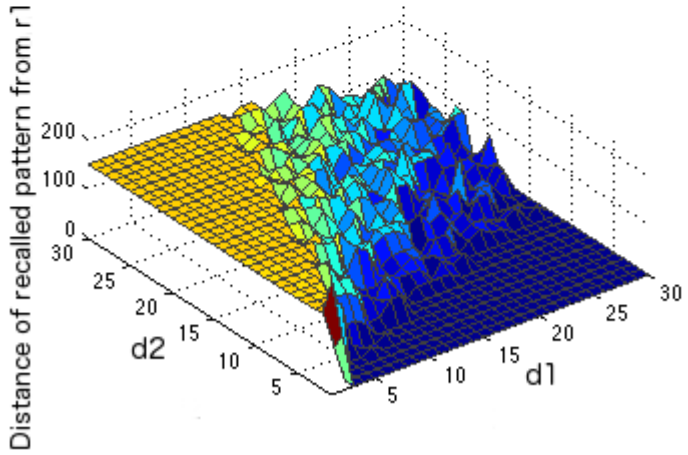
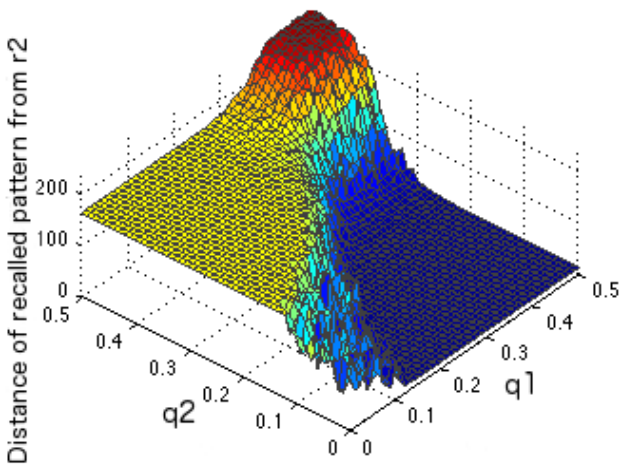


Fig. 4. *The role of degrees in competing strong attractors*. The multiply stored pattern with higher degree will win the competition. On both figures, *d1* and *d2* are the degrees of the two strong attractors. The z axis indicates, for the top figure, the distance from $r2$, the root of cluster 2, while for the bottom figure it gives the distance from $r1$, the root of cluster 1.



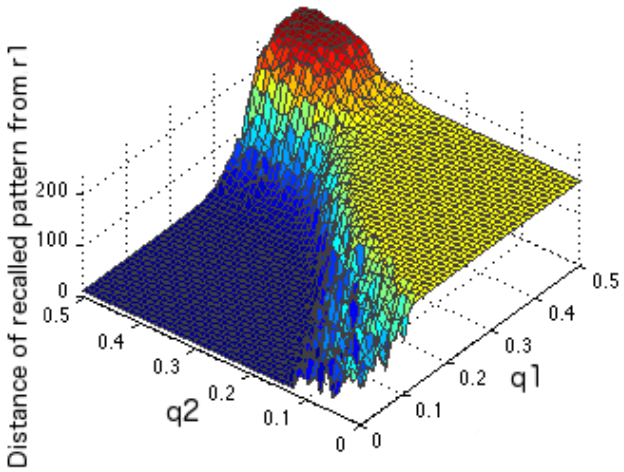


Fig. 5. *Competition between perturbed strong attractors*. The multiply stored pattern with lower value of q will win the competition. On both figures, *q1* and *q2* are the noise values for the two strong attractors. The z axis indicates, for the top figure, the distance from $r2$, the root of cluster 2, while for the bottom figure it gives the distance from $r1$, the root of cluster 1. Note the steeper transition from one strong attractor to the other (yellow to blue), with respect to the transition for varying degrees (Fig. 4).

that are random perturbations of a strong pattern with a value of q that is comprised between 0 and a value γ (between 0.25 and ≈ 0.375 depending on the degree of the strong attractor, see Fig. 2) it will in fact learn a significant number of generalised stored patterns that are very close to the strong attractor's root. So we have three phases for the unfolding of a strong attractor in the presence of other random patterns:

(i) When $q = 0$. The basin size of the strong attractor is very large (depending on how many other random patterns or strong patterns are stored).

(ii) When $0 < q < \gamma$. The basin size of each perturbed pattern stored collapses to a small value, but we have a multitude of basins of generalised stored states which guarantees that a pattern Hamming-close to the strong pattern will be retrieved. The number γ depends on the degree of the strong attractor, and on how many simple random patterns are stored in the network.

(iii) When $\gamma < q < 0.5$. The basin size of each perturbed pattern stored starts getting larger on average but the pattern recalled when the network is exposed to random patterns will have further and further Hamming distance from the root be the strong pattern is weaker and the generalised basin eventually breaks up completely (at $q = 0.5$).

E. Discussion

Past studies on Hopfield models using Hebbian learning have shown how networks with correlated patterns provide less capacity (see for example, [26]). In this study we have pointed that when strong patterns are perturbed with low values of q, the network forms a multitude of generalised stored patterns that are very Hamming-close to the patterns presented in the learning phase. These are recalled with high probability when the network is initialised in a random configuration, confirming the results of Section V.

This can be inferred by combining the results shown in Fig. 2 and in Fig. 5. The first figure gives us the basin size for a strong attractor with low q values and the second shows that, when exposed to random patterns, the network recalls, with a very high probability, a pattern very near the root of the strong attractor with the lower value of q.

VII. Conclusion

We have shown, by deriving some simple mathematical properties and running various computer simulations, that the notion of strong attractors and their random perturbations in Hopfield neural networks can be employed as a useful conceptual tool to model attachment types and behavioural patterns as well as to model changes that they can experience.

For future work, we will examine the variance of the size of basins of attraction of the generalised stored patterns when a strong pattern is perturbed by small noise. This information will provide us more understanding of how strong attractors unfold as a result of random perturbation. Another essential task is to consider the properties of strong patterns and their random perturbations in sparse networks with low level of neural activity as in [27], which is more biologically realistic, and where the basins of attractors have very different shapes [28]. We will also extend our results to stochastic Hopfield networks and Boltzmann machines. One area for further work is to examine the impact of strong attractors in the Boltzmann machine developed in [29] to model neuroses. A more challenging task is to integrate these conceptual tools with the current work on modelling cognitive-emotional decision making using attractor neural networks as in [30].

Acknowledgement

We thank Wael Al Jishi, Niklas Hambuechen, Razvan Marinescu, Mihaela Rosca and Lukasz Severyn who reconfirmed some of the earlier results of simulations of this paper in their undergraduate group project in Autumn 2012.

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Science, USA*, vol. 79, pp. 2554–2558, 1982.

[2] D. O. Hebb, *The organization of behavior.* Wiley, 1949.

[3] J. A. Hertz, A. S. Krogh, and R. G. Palmer, *Introduction To The Theory Of Neural Computation*. Westview Press, 1991.

[4] S. Diederich and M. Opper, "Learning of correlated patterns in spin-glass networks by local learning rules," *Phys. Rev. Lett*, vol. 58, p. 949952, 1987.

[5] A. J. Storkey and R. Valabregue, "The basins of attraction of a new hopfield learning rule," *Neural Networks*, vol. 12, no. 6, pp. 869–876, 1999.

[6] N. Davey, S. P. Hunt, and R. Adams, "High capacity recurrent associative memories," *Neurocomputing*, pp. 459–491, 2004.

[7] D. Kleinfeld and D. B. Pendergraft, ""unlearning" increases the storage capacity of content addressable memories," *Biophys J*, vol. 51, no. 1, pp. 47–53, 1987.

[8] F. Crick and G. Mitchison, "The function of dream sleep," *Nature*, vol. 304, pp. 111–114, 1983.

[9] R. E. Hoffman, "Computer simulations of neural information processing and the schizophrenia-mania dichotomy," *Arch Gen Psychiatry.*, vol. 44, no. 2, pp. 178–88, 1987.

[10] A. .Peled, "A new diagnostic system for psychiatry," *Med Hypotheses*, vol. 54, no. 3, 2000.

[11] J. J. Knierim and K. Zhang, "Attractor dynamics of spatially correlated neural activity in the limbic system," *Annual Review of Neuroscience*, vol. 35, no. 267-85, 2012.

[12] J. Bowlby, *Attachment: Volume One of the Attachment and Loss Trilogy*. Pimlico, second revised edition ed., 1997.

[13] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum, Hillsdale NJ, 1978.

[14] A. N. Schore, *Affect Dysregulation and Disorders of the Self*. W. W. Norton, 2003.

[15] L. Cozolino, *The Neuroscience of Human Relationships*. W. W. Norton, 2006.

[16] P. Fonagy, *Attachment Theory and Psychianalysis*. Other Press, 2001.

[17] J. Young, J. Klosko, and M. Weishaar, *Schema Therapy: A Practitioner's Guide*. Guildford Press, 2006.

[18] D. Petters, "Building agents to understand infant attachment behaviour," *International Joint Conference on Artificial Intelligence*, vol. , pp. 158–165, 2005. .

[19] A. Hiolle, L. Caamero, M. Davila-Ross, and K. A. Bard, "Eliciting caregiving behavior in dyadic human-robot attachment-like interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, p. Article number 3, 2012. Special Issue on Affective Interaction in Natural Environments.

[20] T. Lewis, F. Amini, and R. Richard, *A General Theory of Love*. Vintage, 2000.

[21] T. S. Smith, G. T. Stevens, and S. Caldwell, "The familiar and the strange: Hopfield network models for prototype-entrained," in *Mind, brain, and society: Toward a neurosociology of emotion* (T. S. E. Franks, David D. (Ed); Smith, ed.), vol. 5 of *Social perspectives on emotion*, Elsevier Science/JAI Press, 1999.

[22] L. Personnaz, I. Guyon, and G. Dreyfus, "Information storage and retrieval in spin-glass like neural networks," *Physical Review A*, vol. 46, pp. 359–365, 1985.

[23] L. Cozolino, *The Neuroscience of Psychotherapy: Healing the Social Brain*. W. W. Norton, second edition ed., 2010.

[24] A. N. Schore, *The Science of the Art of Psychotherapy*. Norton, 2012.

[25] P. Billingsley, *Probability and Measure*. John Wiley & Sons, second edition ed., 1986.

[26] M. Lowe, "On The Storage Capacity of Hopfield Models with Correlated Patterns," *Annals of Appplied Probability*, vol. 8, no. 4, pp. 1216–1250, 1998.

[27] M. Tsodyks and M. Feigelman, "Enhanced storage capacity in neural networks with low level of activity," *Europhysics Letters,*, vol. 6, pp. 101–105, 1988.

[28] A. Knoblauch, "Neural associative memory with optimal bayesian learning," *Neural Computation*, vol. 23, no. 6, pp. 1393–1451, 2011.

[29] R. S. Wedemann, R. Donangelo, and L. a. V. de Carvalho, "Generalized memory associativity in a network model for the neuroses.," *Chaos (Woodbury, N.Y.)*, vol. 19, no. 1, p. 015116, 2009.

[30] D. S. Levine, "Brain pathways for cognitive-emotional decision making in the human animal," *Neural Networks*, vol. 22, pp. 286–293, 2009.