Neural Processing Letters **16:** 243–257, 2002.

Biologically Plausible Associative Memory: Continuous Unit Response + Stochastic Dynamics

ENRIQUE C. SEGURA MECCIA[1]★ and ROBERTO P. J. PERAZZO[2]

[1]*School of Computing, Information Systems and Mathematics, South Bank University, 103 Borough Road, London SE1 0AA, UK. e-mail: segurae@sbu.ac.uk*

[2]*Departamento de Fisica, Universidad de Buenos Aires, Ciudad Universitaria, (1428) Buenos Aires, Argentina*

Abstract. A neural network model of associative memory is presented which unifies the two historically more relevant enhancements to the basic Little-Hopfield discrete model: the graded response units approach and the stochastic, Glauber-inspired model with a random field representing thermal fluctuations. This is done by casting the retrieval process of the model with graded response neurons, into the framework of a diffusive process governed by the Fokker-Plank equation, which leads to a Langevin system describing the process at a microscopic level, while the time evolution of the probability density function is governed by a multivariate Fokker Planck equation operating over the space of all possible activation patterns. The present unified approach has two notable features: (i) greater biological plausibility and (ii) ability to escape local minima of energy (associated with spurious memories), which makes it a potential tool for those complex optimization problems for which the previous models failed.

Key words. associative memory, Fokker-Planck equation, graded response, Hopfield model, stochastic dynamics

1. Introduction

In seminal papers, Little [9, 10] and Hopfield [5] constructed a content addressable memory as a dense network of artificial 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 configurations to the other as a consequence of the intensity of the local field that acts upon it. The memories, that correspond to fixed points of the dynamics, are stored in the system in a distributed manner through the matrix of synaptic efficacies between the neurons. If this matrix is properly calculated, the above dynamics is enough to ensure a monotonic decrease of an 'energy' function. Thus, starting from an arbitrary configuration the system is led to a local minimum that corresponds to the closest stored memory.

In a later paper, Hopfield [6] aims at a more realistic model by replacing bistable neurons by graded response devices. In fact, the stronger objection to the plausibility of the former model [5, 9, 10] was that a two-state representation of the neural output

★Corresponding author.

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. In either case the retrieval process is again guaranteed by the nature of the matrix of synaptic efficacies.

On the other hand, on a separate line of thought, a number of later investigations [4, 11] have considered more realistic pictures of the neuron response by assuming that the transition between the two stable states of individual neurons is affected by a random field representing thermal fluctuations. However, in this model the random updating of individual neurons prevents them from reaching the exact configuration that corresponds to the fixed point of the dynamics.

In both frameworks the retrieval process has traditionally been analyzed and described making use only of numerical tools or considering configurations that are in thermodynamic equilibrium. A numerical, microscopic description of thermal fluctuations is impractical for models involving graded response neurons, and this is the reason why thermal fluctuations were omitted from these models. On the other hand, a statistical approach emphasizes the role of equilibrium and therefore disregards the transitional pattern that prevails during the retrieval process.

The aim of the present paper is to show that it is possible to merge the advantages of both the graded response neurons approach and the stochastic dynamics approach, getting a model of associative memory that takes account of both the graded response of individual neurons in terms of firing rates and the stochastic behavior of the thermal fluctuations affecting state transitions.

We do this by casting the retrieval process of a Hopfield model with graded response neurons, into the framework of a diffusive process governed by the Fokker-Plank (F-P) equation. We thus provide a description of the transitional regime that prevails during the retrieval process, that is currently disregarded. The possibility of generalizing the non-deterministic, finite temperature Glauber dynamics [2] to the case of graded response neurons has been discussed in [12]. This was attempted for the case in which the network consisted of a non-countable number of neurons organized in a continuous metric space. That approach formally leads to a functional F-P description of the retrieval dynamics but makes very difficult any further analytical treatment.

In the present paper we concentrate on the Hopfield model with a finite number of graded response neurons [6]. Within this framework the individual updating process is formally equivalent to a Langevin process. In addition, a probability density can be defined that represents the average excitation pattern of the ensemble of neurons. This in turn allows a description of its evolution through a multivariate F-P equation whose structure depends upon the temperature of the thermal bath and the pattern of memories stored in the network.

We find that this approach provides a finite temperature description of the retrieval process that agrees with what one would intuitively think. For instance, the average excitation pattern of the system is given by a Gaussian peak in a multidimensional space spanned by the response of all neurons, that is located at the retrieved memory. Its width is given by the temperature of the heat bath and therefore exact retrieval is prevented except for zero temperature.

Besides its biological plausibility, the present approach has the ability to escape spurious states. In fact, if the matrix of synaptic weights and the energy function are defined in the usual way [5, 6], the presence of spurious metastable memories (most of them as combinations of an odd number of stored memories) is unavoidable. However, in the present model the asymptotic probability distribution is proven to always have a peak in a stored memory, which ensures that the system will not get stuck in a spurious memory (remember that these states, being local minima, have higher energy than stored memories). This advantage over the deterministic, graded response Hopfield model [6] is similar to that of the stochastic, Glauber based approaches with bistable units [4, 11] over the discrete Little-Hopfield model (from now on, the 'Hopfield model'). In fact, in those models the equilibrium distribution at finite (non-zero) temperature is the Boltzmann distribution.

Thus, this unified approach can be tried as a technique for those complex optimization problems in which the application of the previous models of associative memory have not been very successful. One of the classical applications of a neural network with associative memory to a complex optimization problem was due to Hopfield and Tank [7], who applied the continuous activation model [6] to the Travelling Salesman Problem, an archetypical NP-complete combinatorial problem. Its performance was not competitive with respect to other classical stochastic techniques such as the Simulated Annealing nor even in comparison with some deterministic ones, due precisely to its tendency to get stuck in local, sub-optimal solutions associated with spurious states of the energy surface. This suggests the potential value, for the treatment of complex optimization problems, of joining together in the same algorithm the continuity of the function activation and the stochasticity of the overall dynamics.

This paper is organized as follows. In Section 2 we review some basic concepts and definitions (it can be omitted by readers familiar with the Hopfield models). In Section 3 we derive the Langevin approach for the updating dynamics of the neurons and, in 4, the associated multivariate F-P equation. In Section 5 the stationary solutions are analyzed while the dynamics is studied in Section 6, showing the asymptotic stability for an initial Gaussian distribution. Finally, in Section 7 we generalize this result taking advantage of the fact that the initial excitation pattern of the neural network can always be expanded as a superposition of Gaussian distributions.

2. Preliminaries

2.1. ASSOCIATIVE MEMORY. THE ORIGINAL HOPFIELD MODEL

The problem of associative memory is that of storing a set of p patterns $\xi^{\mu}(\mu = 1, 2, \ldots, p)$ in such a way that when presented with a new element ζ as input, the system output is the ξ^{μ} that most resembles ζ. In both the Little [9, 10] and the

Hopfield [6] model, each ξ^{μ} $(\mu = 1, 2, \ldots, p)$ belongs to the set $\{1, -1\}^N$ of all N-tuples whose elements can take on the values 1 (active neuron) or -1 (inactive neuron). The dynamics of the network is

$$S_i(n+1) = \operatorname{sgn} \sum_j \mathbf{T}_{ij} S_j(n) \quad 1 \leqslant i \leqslant N \tag{1}$$

where $S_i(n)$ stands for the state of the i-th unit of the system at time n, $\mathbf{T}_{ij}$ is the synaptic weight between neurons i and j and

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if x} \geqslant 0 \\ -1 & \text{if x} < 0 \end{cases}$$

The stability condition for an element ξ is

$$\xi_i = \operatorname{sgn} \sum_j \mathbf{T}_{ij} \xi_j \quad 1 \leqslant i \leqslant N$$

and the energy function is

$$H[S] = -\frac{1}{2} \sum_{ij} \mathbf{T}_{ij} S_i S_j$$

In [5] it is proven that if the ξ^{μ} are generated pseudo-orthogonally, i.e. from a probability distribution such that $\langle \xi^{\mu} \xi^{v} \rangle = 0$ whenever $\mu \neq v$, the cardinality of p doesn't exceed a critical value p_c and the weight matrix is computed following the Hebb rule, i.e. as $\mathbf{T}_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{\mu} \xi_j^{\mu}$, then the system has the property of associative memory. The energy decreases as the system evolves (provided $\mathbf{T}$ is symmetric), having as minima the stored ξ^{μ} (called *attractors* of the dynamics).

2.2. CONTINUOUS TRANSFER FUNCTIONS: HOPFIELD'84

In [6] Hopfield introduces the following dynamics:

$$\dot{\xi}_i = -\xi_i + g_{\gamma}(h_i^{\xi}) \quad 1 \leqslant i \leqslant N \tag{2}$$

where g_{γ} is a *sigmoid* function, i.e. $g_{\gamma} \in C^1(R)$, non-decreasing and odd and satisfying $\lim_{x \to \pm\infty} g_{\gamma}(x) = \pm 1$, $\lim_{\gamma \to \infty} g_{\gamma}(x) = \operatorname{sgn}(x) \forall x \neq 0$, $|g_{\gamma}(x)| < \min\{1, \gamma x\}$ and $g'_{\gamma}(0) = \gamma$ and $h_i^{\xi} \underline{\underline{\Delta}} \sum_{i=1}^{N} \mathbf{T}_{ij} \xi_j$ is the net input to neuron i when the state of the system is ξ. The stability condition is rewritten as $\xi_i = g_{\gamma}(\sum_j \mathbf{T}_{ij} \xi_j)$, $1 \leqslant i \leqslant N$. Equation 2 is a straightforward differential extension of the difference equation that represents the dynamics of the discrete model (Equation 1). The energy function is now

$$H[\xi] = -\frac{1}{2} \sum_{ij} \mathbf{T}_{ij} \xi_i \xi_j + \sum_i \int_0^{\xi_i} g_{\gamma}^{-1}(\xi) d\xi$$

and is minimized by the dynamics (Equation 2). Moreover, when $\gamma \to \infty$, the attractors tend to be located on the vertices of the hypercube $[-1, 1]^N$, coinciding with those produced by the discrete model [5] for the same $\mathbf{T}$. The existence of spurious attractors is known also in this case.

Hopfield and Tank [7] proposed an application of this model to the Travelling Salesman Problem, the NP-complete problem consisting of finding the shortest tour calling at N cities once and only once each one. However, it is worth noting that the performance was limited precisely by the presence of spurious states coinciding with non-optimal solutions. This strengthes the conjecture about the potential benefit, for the treatment of complex optimization problems, of enhancing its dynamics with some kind of parameter-controllable, theoretically well-founded stochasticity.

2.3. ROLE OF NOISE IN ASSOCIATIVE MEMORIES. THE GLAUBER FORMALISM

It is well known that, no matter how many memories are stored in a Hopfield-type associative memory (either discrete or continuous), provided they are more than two, spurious, undesired attractor states appear, introducing the possibility of an error in recall. The probability of retrieving a spurious state instead of a stored memory increases with the number of these memories, until a critical ratio between this number and the size of the network is reached. Then, a ‘phase transition’ takes place: the number of spurious states exponentially increases and the capability of the memory is lost. The system is overload. This phase is called the ‘confusion phase’.

These spurious states have commonly higher energy than stored memories but they are nevertheless local minima. Since the dynamics in the Hopfield model is always ‘dissipative’, in the sense that the energy is monotonically decreasing, they cannot be escaped (they are surrounded by energy barriers). However, this deterministic conception of neural dynamics is not very biologically plausible: most biologists currently consider that noise and randomness are almost universal in living systems. Then, if the neuronal dynamics is stochastic, neurons can make transitions into states which are opposed to the direction of their presynaptic potential (PSP) [1], due to several factors such as the level of the noise and the actual difference between the PSP and the threshold. Some speculations have been made concerning the importance of noise for the sake of making associative memories suitable as models of certain brain disorders [1].

Several authors [4, 11] have introduced noisy dynamics into the discrete Hopfield model via the Glauber formalism. In this approach, the probability distribution of the state of the ith processing unit S_i at an instant $n+1$ is given by

$$\mathbf{P}(S_i(n+1) = \pm 1) = \frac{1}{1 + \exp(\mp 2\beta h_i^{S(n)})}$$

This allows control of the level of noise by means of a unique parameter β, called the ‘inverse temperature’. In fact, for high values of β (low temperature), the noise is not too high, hence the system behaves quasi-deterministically and the spurious states persist. On the other hand, for low β (high temperature), the dynamics is purely ergodic, so there are no attractors at all, either spurious or not. But for some medium range of values of β, it is possible to destabilize the spurious attractors while, at the same time, getting few errors in the retrieval of stored memories.

The activity pattern of the system is defined through a Markov process that can be formulated as a *Master equation* [1]:

$$\mathbf{P}(\xi, n+1) = \mathbf{P}(\xi, n) + \sum_{\zeta \neq \xi} [\mathbf{W}_\beta(\xi \mid \zeta)\mathbf{P}(\zeta, n) - \mathbf{W}_\beta(\zeta \mid \xi)\mathbf{P}(\xi, n)]$$

where

$$W_\beta(\xi \mid \zeta) = \prod_{i=1}^{N} \frac{1}{\exp\{-\beta \xi_i h_i^\zeta\} + 1}$$

and n represents a discrete time. This equation reflects the temporal evolution of the probability distribution associated with the discrete Hopfield model, i.e. $\mathbf{P}(\xi, n)$ is the probability of the system to be in a certain excitation pattern at the nth step of the evolution process.

3. Langevin Processes and the Hopfield Model

As pointed out in the Introduction, in this paper we are interested in a finite temperature dynamics for the associative memory with graded-response units. In other words: would it be possible to find some equivalent equation expressing the evolution of the probability distribution associated to a Hopfield model with graded response neurons and stochastic, noisy dynamics?, that is, a set of processing units labeled as ξ_i, $i = 1, \ldots, N$, each one obeying the dynamics:

$$\dot{\xi}_i = -\xi_i + g_\gamma(h_i^\xi) + L_i(t)$$

where g_γ is a sigmoid function (see Section 2.2), h_i^ξ is the local field on neuron i when the state of the system is ξ and $L_i(t)$ is a stochastic Gaussian process (identical and independent for each i). Consider the equation:

$$\dot{v} = -\rho v + L(t)$$

which describes the motion of a Brownian particle when its mass is taken to be unity [8]. v is the velocity of the particle and the right hand side expresses the force exerted on it, consisting of a damping term linear in v plus the noise $L(t)$. Assume that the following three conditions are satisfied:

1. $L(t)$ is a stochastic process. This means that its properties averaged over a system of many uncorrelated particles (because, for example, their mutual distances are so large that they do not influence each other) are the same as if it were observed acting successively on the same particle (at long enough time intervals). In other words, the system is assumed to be *ergodic*.
2. $L(t)$ is independent of v, so that $L(t)$ acts as an external force. Moreover $\langle L(t) \rangle = 0$.
3. $L(t)$ varies rapidly: $\langle L(t)L(t') \rangle = \Gamma\delta(t - t')$.

Then the above equation is called the Langevin equation and the term $L(t)$, which describes the fluctuations in the system, is a Langevin force. It can be extended to the case of a system with a nonlinear equation of motion $\dot{v} = A(v)$. The corresponding Langevin equation is

$$\dot{v} = A(v) + L(t)$$

also known as the quasilinear Langevin equation. On the other hand, recall from Section 2.2 that in the Hopfield neural network with graded response units, each of them obeys the dynamics given by Equation (2), and assume there is an external force acting on each unit and producing stochastic fluctuations in its state. Then we have a system of stochastic equations

$$\dot{\xi}_i = -\xi_i + g_\gamma(h_i^\xi) + L_i(t) \quad i = 1, \dots, N \tag{3}$$

We assume that the parameter Γ is the same for all units. Thus we are describing the dynamics of the network and representing, at the same time, a multidimensional Langevin process. As we will see, the parameter Γ plays the role of temperature (and that is the reason why it can be considered to be the same for all neurons) in the form of an additive noise.

4. The Fokker-Planck Equation

The well-known one-dimensional Fokker-Planck equation has the form:

$$\frac{\partial \mathbf{P}(y,t)}{\partial t} = -\frac{\partial}{\partial y}\{a^{(1)}(y)\mathbf{P}\} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\{a^{(2)}(y)\mathbf{P}\}$$

with

$$a^{(v)}(y) = \int_{-\infty}^{\infty} r^v \mathbf{W}(y;r)dr$$

where $\mathbf{W}(y;r)$ is the transition probability matrix of a jump of size r taking place from the state y and hence $a^{(v)}(y)$ is the moment of order v of such a distribution of jumps.

This equation is a type of master equation, often used as a model for Markov processes. The first term on the right hand side is called the ‘transport term’ and the second, the ‘diffusion term’. The following result is crucial in what follows.

THEOREM [8]. *The Langevin equation $\dot{y} = A(y) + L(t)$ with Gaussian noise defined by $\langle L(t)L(t')\rangle = \Gamma\delta(t - t')$ represents the same Markov process as the F-P equation*

$$\frac{\partial \mathbf{P}(y,t)}{\partial t} = -\frac{\partial}{\partial y}\{A(y)\mathbf{P}\} + \frac{\Gamma}{2}\frac{\partial^2}{\partial y^2}\mathbf{P}$$

For a detailed proof, see [8], ch. IX. □

The condition that $L(t)$ be Gaussian is essential, in order to implicitly specify the moments of $L(t)$ of order higher than two, so as to let the Langevin equations fully determine the stochastic process $y(t)$, just as the F-P equation does.

The F-P equation can be generalized for the case of a multidimensional state space:

$$\frac{\partial \mathbf{P}(y,t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial y_i}\{\mathrm{A}_i(y)\mathbf{P}\} + \frac{1}{2}\sum_{i=1,j=1}^{N} \frac{\partial^2}{\partial y_i \partial y_j}\{\mathrm{B}_{ij}(y)\mathbf{P}\}$$

with $y = (y_1, \ldots, y_N)$, A_i, B_{ij} real differentiable functions, B symmetric and positive semi-definite ($x^{\mathrm{T}}\mathrm{B}x \geqslant 0 \ \forall x \in R^N$). These conditions are satisfied if we use

$$A_i(y) = \int_{-\infty}^{\infty} r_i \mathbf{W}(y; r)\mathrm{d}r$$

and

$$\mathbf{B}_{ij}(y) = \int_{-\infty}^{\infty} (r_i - A_i(y))(r_j - A_j(y))\mathbf{W}(y; r)\mathrm{d}r$$

i.e. the jump moments of first and second order respectively (here r stands for $(r_1, r_2, \ldots, r_N)$, $\mathrm{d}r = \mathrm{d}r_1\mathrm{d}r_2 \ldots \mathrm{d}r_N$ and the integration is meant in N real variables).

On the other hand, note that in Equation (3), which governs the dynamics of each processing unit, $\dot{\xi}_i$ represents the jump per unit time. Then in our case

$$A_i(\xi) = \langle r_i \rangle = -\xi_i + g_\gamma(h_i^\xi)$$

as $\langle L(t) \rangle = 0$ by condition 2 (Section 3).

$$B_{ij}(\xi) = \int_{-\infty}^{\infty} (r_i - A_i(\xi))(r_j - A_j(\xi))\mathbf{W}(\xi; r)\mathrm{d}r$$

$$= \langle (r_i - A_i(\xi))(r_j - A_j(\xi)) \rangle_r = \langle L_i(t)L_j(t) \rangle = \Gamma \delta_{ij}$$

Therefore

$$\frac{\partial \mathbf{P}(\xi,t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial \xi_i}\{(-\xi_i + g_\gamma(h_i^\xi))\mathbf{P}\} + \Gamma \sum_{i=1}^{N} \frac{\partial^2}{\partial \xi_i^2}\mathbf{P} \tag{4}$$

is the F-P equation for the Hopfield model with continuous activation functions, macroscopic equivalent of Equation (3).

5. Stationary Solution

5.1. ONE MEMORY

From now on we will assume the synaptic matrix is constructed following the Hebb rule (just as in [5] and [6]) with zero diagonal:

$$\mathbf{T}_{ij} = \frac{1}{N}(1 - \delta_{ij})\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$$

where $\{\xi^\mu\}_{\mu=1}^p$ is the set of p patterns to be stored. Let us consider first the case when only one memory, say ξ^ν, has been stored in the network, i.e. $\mathbf{T}_{ij} = \frac{1}{N}(1-\delta_{ij})\xi_i^\nu\xi_j^\nu$.

We define

$$\mathbf{P}^s(\xi) = \frac{1}{(2\pi\Gamma)^{\frac{N}{2}}}\exp\left(\frac{-\sum_{i=i}^N(\xi_i - \xi_i^\nu)^2}{2\Gamma}\right) \tag{5}$$

that is, a normal (or Gaussian) probability density with parameters (ξ^ν, Γ). This is our candidate stationary solution of Equation (4). Substituting $\mathbf{P}^s$ by $\mathbf{P}$ in the right-hand side of Equation (4):

$$\begin{aligned} &N\mathbf{P}^s + \sum_{i=1}^N(-\xi_i + g_\gamma(h_i^\xi))\frac{(\xi_i - \xi_i^\nu)}{\Gamma}\mathbf{P}^s(\xi) + \Gamma\sum_{i=1}^N\left\{-\frac{1}{\Gamma} + \frac{(\xi_i - \xi_i^\nu)^2}{\Gamma^2}\right\}\mathbf{P}^s(\xi) \\ &\quad = \frac{1}{\Gamma}\sum_{i=1}^N(g_\gamma(h_i^\xi) - \xi_i^\nu)(\xi_i - \xi_i^\nu)\mathbf{P}^s(\xi) \end{aligned} \tag{6}$$

using the hypothesis $\mathbf{T}_{ii} = 0$, which implies that h_i^ξ does not depend on ξ_i.

In addition

$$h_i^\xi = \sum_{j=1}^N \mathbf{T}_{ij}\xi_j = \sum_{j=1, j\neq i}^N \frac{1}{N}\xi_i^\nu\xi_j^\nu\xi_j = \xi_i^\nu\frac{1}{N}\sum_{j=1, j\neq i}^N \xi_j^\nu\xi_j \overset{N\to\infty}{\longrightarrow} \xi_i^\nu\langle\xi_j^\nu\xi_j\rangle_j$$

Note that ξ^ν must satisfy $\xi_i^\nu = \pm V_*$ for each i, being $V_* \in (0, 1)$ the number such that $g_\gamma(V_*^3) = V_*$ (which exists and depends on γ). The necessity of this condition is derived from the fact that

$$h_i^{\xi^\nu} = \sum_{j=1}^N \mathbf{T}_{ij}\xi_j^\nu = \sum_{j=1, j\neq i}^N \frac{1}{N}\xi_i^\nu\xi_j^\nu\xi_j^\nu = \xi_i^\nu\frac{1}{N}\sum_{j=1, j\neq i}^N (\xi_j^\nu)^2 = \xi_i^\nu\|\xi^\nu\|^2$$

for large N. Hence, for the stability condition $g_\gamma(h_i^{\xi^\nu}) = \xi_i^\nu$ being satisfied, it is required that $\xi_i^\nu = \pm V_*$ with V_* as indicated above. Then $\xi_i^\nu\langle\xi_j^\nu\xi_j\rangle_j = V_*^2\xi_i^\nu$ and therefore $\langle h_i^\xi\rangle = \pm V_*^3$. Finally $g_\gamma(\pm V_*^3) = g_\gamma(\langle h_i^\xi\rangle) = \xi_i^\nu$ and expression (6) vanishes, yielding the stability of $\mathbf{P}^s$. It must be remarked that this is an 'average' stability, in the sense of ensemble averages, i.e. under the assumption of a very large number N of units in the system, and it is in this sense that we will understand stability from now on.

Note that the variance Γ, that determines the width of $\mathbf{P}^s$, also represents the amplitude of the stochastic field in the Langevin equation and, in the neural network model, is related to the temperature of the system. This is the reason why we have assumed the same Γ for all units, and we will use this fact from now on.

5.2. MANY MEMORIES

Now suppose there are p pseudo-orthogonal memories ξ^μ, $\mu = 1, \ldots, p$ to store. Let us call

$$\mathbf{P}_{\xi^\mu,\Gamma}(\xi) = \frac{1}{(2\pi\Gamma)^{\frac{N}{2}}}\exp\left(- \frac{\sum_{i=i}^{N}(\xi_i - \xi_i^\mu)^2}{2\Gamma}\right)$$

that is, a Gaussian function centered at memory ξ^μ with variance Γ.

Now define

$$\mathbf{P}(\xi) = \sum_{\mu=1}^{p} \lambda_\mu \mathbf{P}_{\xi^\mu,\Gamma}(\xi)$$

with the constraint $\sum_{\mu=1}^{p} \lambda_\mu = 1$. $\mathbf{P}(\xi)$ is an extension to many memories of the probability distribution function defined by Equation (5): a convex superposition of Gaussians, each one centered in a different memory. We will show that $\mathbf{P}(\xi)$ cannot, in general, be a stationary solution of Equation (4). Substituting $\mathbf{P}$ in the right-hand side of Equation (4) we obtain now:

$$\frac{1}{\Gamma}\sum_{\mu=1}^{p} \lambda_\mu \sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \xi_i^\mu)\mathbf{P}_{\xi^\mu,\Gamma}(\xi)$$

In order for this expression to vanish, we need[★] $\langle h_i^\xi \rangle = \mathrm{V}_*^2 \xi_i^\mu = \pm \mathrm{V}_*^3$.

On the other hand $\langle \xi_j \rangle = \sum_{\mu=1}^{p} \lambda_\mu \xi_j^\mu$.

But

$$h_i^\xi = \sum_{j=1}^{N} \mathbf{T}_{ij}\xi_j = \sum_{j=1, j\neq i}^{N} \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \xi_j = \sum_{\mu=1}^{p} \xi_i^\mu \frac{1}{N}\sum_{j=1, j\neq i}^{N} \xi_j^\mu \xi_j$$

whose mean value ($N \to \infty$) will never equal $\pm\mathrm{V}_*^3$ (since $\langle \xi_j \rangle$ is a convex combination of the ξ_j^μ's) unless $\lambda_\nu = 1$ for some ν and $\lambda_\mu = 0 \ \ \forall \mu \neq \nu$. If this is the case, we have:

$$h_i^\xi = \xi_i^\nu \left(\frac{1}{N}\sum_{j=1, j\neq i}^{N} \xi_j^\nu \xi_j \right) + \sum_{\mu=1, \mu\neq\nu}^{p} \xi_i^\mu \left(\frac{1}{N}\sum_{j=1, j\neq i}^{N} \xi_j^\mu \xi_j \right)$$

The first term between parentheses tends to V_*^2 when $N \to \infty$ whilst the second vanishes, provided ξ^μ are pseudo-orthogonal. Hence we have $\langle h_i^\xi \rangle = \mathrm{V}_*^2 \xi_i^\mu = \pm\mathrm{V}_*^3$. We conclude that a linear (non trivial) combination of Gaussian functions centered in the memories cannot be a stationary solution of Equation (4).

6. Dynamics

From the previous section we know that each $\mathbf{P}_{\xi^\mu,\Gamma}$, $\mu = 1, \ldots, p$, is a stationary solution of Equation (4). In this section we investigate their stability.

[★] Here again ξ^μ should satisfy $\xi_i^\mu = \pm\mathrm{V}_*$ for each i. But in this case ($p > 1$), the equality $h_i^{\xi^\mu} = \mathrm{V}_*^2 \xi_i^\mu$, which implies $\xi_j^\mu = \pm\mathrm{V}_*$ for each j (without average brackets) only holds if the memories ξ^μ are orthogonal, since only in that case do the crosstalk terms in $h_i^{\xi^\mu}$ vanish. Therefore the condition on the ξ^μ's is now relaxed to $\langle \xi_i^\mu \rangle = \pm\mathrm{V}_*$.

Assume that the probability density function $\mathbf{P}(\xi, t)$ is, at t = 0 (or at any fixed $t > 0$), a normal density function with parameters (ζ, σ^2).

LEMMA. *If* **P** *is Gaussian with parameters* (ζ, σ^2)*, the F-P equation* (*Equation* 4) *has the form:*

$$\frac{\partial \mathbf{P}}{\partial t} = 2(\Gamma - \sigma^2)\frac{\partial \mathbf{P}}{\partial \sigma^2} + \sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \zeta_i)\frac{\partial \mathbf{P}}{\partial \zeta_i}$$

Proof. Differentiating P with respect to its parameters we obtain:

$$\frac{\partial \mathbf{P}}{\partial \sigma^2} = \frac{1}{2}\left\{-\frac{N}{\sigma^2} + \frac{\sum_{i=1}^{N}(\xi_i - \zeta_i)^2}{\sigma^4}\right\}\mathbf{P}$$

and

$$\frac{\partial \mathbf{P}}{\partial \zeta_i} = \frac{(\xi_i - \zeta_i)}{\sigma^2}\mathbf{P}$$

yielding

$$\frac{\partial \mathbf{P}}{\partial t} = \left\{-\frac{\Gamma N}{\sigma^2} + N + \frac{\Gamma}{\sigma^4}\sum_{i=1}^{N}(\xi_i - \zeta_i)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \xi_i)(\xi_i - \zeta_i)\right\}\mathbf{P}$$

On the other hand, substituting **P** in the right-hand side of (2), the same result is obtained. □

The previous lemma states that for a Gaussian **P** the right hand side of the F-P equation can be rewritten as a function of the partial derivatives of **P** with respect to its parameters. This means that we have expressed $\mathbf{P}(\xi, t)$ as $\mathbf{P}(\xi, \zeta(t), \sigma^2(t))$. Then the following expression is valid for the total derivative of **P** with respect to t:

$$\frac{d\mathbf{P}}{dt} = \frac{\partial \mathbf{P}}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial t} + \sum_{i=1}^{N}\frac{\partial \mathbf{P}}{\partial \zeta_i}\frac{\partial \zeta_i}{\partial t}$$

In other words, **P** depends on t only through its parameters ζ and σ. Comparing the last expression with the previous lemma, it follows that $(\partial\sigma^2/\partial t) = 2(\Gamma - \sigma^2)$ and $(\partial\zeta_i/\partial t) = g_\gamma(h_i^\xi) - \zeta_i$.

These identities are true, of course, at the time t when we have assumed **P** is Gaussian. If we could prove that the dynamics determined by the F-P equation evolves Gaussian distributions onto Gaussian distributions, the above identities would be true for any t and we could solve them as ordinary differential equations, yielding $\sigma(t) = \Gamma - \exp(-2t)(\Gamma - \sigma_0)$ and $\zeta_i(t) = \xi_i^\mu - \exp(-t)(\xi_i^\mu - \zeta_i(t_0))$ for some large enough t_0 (keeping in mind that $g_\gamma(h_i^\xi) \xrightarrow{t\to\infty} \xi_i^\mu$ for some stored memory ξ_i^μ). Then, it remains to be proved that, if **P** is a Gaussian distribution at time t, then at time

$t + \Delta t$ it will also be a Gaussian distribution (with slightly different parameters ζ and σ^2). Discretizing t, our lemma can be restated as

$$\begin{aligned}\mathbf{P}(t+\Delta t) &= \mathbf{P}(t) + \Delta t\left\{2(\Gamma - \sigma^2)\frac{\partial \mathbf{P}}{\partial \sigma^2} + \sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \zeta_i)\frac{\partial \mathbf{P}}{\partial \zeta_i}\right\}\\ &= \mathbf{P}(t)\left\{1 + \Delta t\left\{2(\Gamma - \sigma^2)\frac{1}{2}\left\{-\frac{N}{\sigma^2} + \frac{\sum_{i=1}^{N}(\xi_i - \zeta_i)^2}{\sigma^4}\right\}\right.\right.\\ &\quad \left.\left.+\sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \zeta_i)\frac{(\xi_i - \zeta_i)}{\sigma^2}\right\}\right\}\end{aligned}$$

We know that $\mathbf{P}$ is a Gaussian with parameters ζ and σ^2. As for the rest of the above expression, it has the form $1 + \Delta t F(t)$, which can be approximated to first order as $\exp(\Delta t F(t))$.

Therefore the last expression becomes, to first order in t:

$$\begin{aligned}&\exp\left\{-\frac{N}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=i}^{N}(\xi_i - \zeta_i)^2}{2\sigma^2} + \Delta t\left\{2(\Gamma - \sigma^2)\frac{1}{2}\left\{-\frac{N}{\sigma^2} + \frac{\sum_{i=1}^{N}(\xi_i - \zeta_i)^2}{\sigma^4}\right\}\right.\right.\\ &\quad \left.\left.+\sum_{i=1}^{N}(g_\gamma(h_i^\xi) - \zeta_i)\frac{(\xi_i - \zeta_i)}{\sigma^2}\right\}\right\} \qquad (7)\end{aligned}$$

Then we should prove that the whole exponent of the last expression has the form

$$-\frac{N}{2}\log(2\pi\sigma'^2) - \frac{\sum_{i=i}^{N}(\xi_i - \zeta'_i)^2}{2\sigma'^2}$$

i.e. the exponent of a Gaussian distribution with parameters $\zeta' = \zeta(t + \Delta t)$ and $\sigma'^2 = \sigma^2(t + \Delta t)$.

But expanding it to first order in t yields precisely the exponent of (7).

It can be concluded that if a solution is Gaussian at a certain time, then it will remain Gaussian. Moreover, for any Gaussian initial condition, the corresponding stationary solution of Equation (4) (in the sense of the 'ensemble average') is $\mathbf{P}_{\xi^\mu,\Gamma}(\xi)$ where ξ^μ is a stored memory. In the next section we analyze how this final asymptotical distribution is reached.

6.1. MULTI GAUSSIAN INITIAL DISTRIBUTION

Now let us consider an initial condition of the form:

$$\mathbf{P}(\xi) = \sum_{m=1}^{M}\alpha_m \mathbf{P}_{\xi^m,\sigma_m^2}(\xi)$$

with the constraint $\sum_{m=1}^{M}\alpha_m = 1$.

This initial condition represents a generalization of the Gaussian initial condition assumed in the previous section, e.g. if the ξ^m's are far away enough from each other, the initial probability distribution has a local maximum at each of those points.

It is easy to see that in this case the right hand side of Equation (4) is the sum of M terms, each of them consisting of a derivative with respect to σ_m^2 and a sum of derivatives with respect to the ξ_i^m's. We can reasonably assume $\sigma_m^2 = \sigma^2 \ \ \forall m$, giving σ^2 the meaning of a temperature or thermal noise (it is plausible to consider it as being the same for all units). Therefore, by virtue of the linearity of Equation (4), the terms involving the derivative with respect to σ_m^2 do not change. As for the derivatives with respect to the ξ_i^m's, note that each term has the form:

$$\alpha_m \sum_{i=1}^{N} (g_\gamma(h_i^\xi) - \xi_i^m) \frac{\partial \mathbf{P}_{\xi^m, \sigma_m^2}}{\partial \xi_i^m} \tag{8}$$

At any given time, it holds that

$$\langle \xi_j \rangle = \sum_{m=1}^{M} \alpha_m \xi_j^m$$

where $\langle \xi_j \rangle$ stands for the jth coordinate of the mean value of $\mathbf{P}(\xi)$. On the other hand:

$$h_i^\xi = \sum_{j=1}^{N} \xi_i^\mu C^\mu$$

with $C^\mu = \langle \xi_j^\mu \xi_j \rangle_j = \frac{1}{N} \sum_{j=1}^{N} \xi_j^\mu \xi_j$. The C^μ's act in the manner of 'Fourier coefficients' of the distribution of ξ with respect to the pseudo-orthogonal system $\{\xi^\mu\}$. We know that $| \ C^\mu \ | \leqslant \mathrm{V}_*^2$ and it holds $C^\mu = \pm \mathrm{V}_*^2$ if and only if ξ is distributed around a particular ξ^μ. Generally speaking, C^μ is a measure of the proximity of $\langle \xi \rangle = \sum_{m=1}^{M} \alpha_m \xi_j^m$ to memory ξ^μ. Thus, h^ξ is the projection of ξ onto the pseudo-orthogonal set $\{\xi^1, \ldots, \xi^p\}$ and, provided g_γ is monotonical, $g_\gamma(h^\xi) - \xi^m$ measures how near to that projection is ξ^m. Note that the quantity expressed by (8) represents the derivative of $\mathbf{P}_{\xi^m, \sigma_m^2}$ (with respect to its parameter ξ^m) in the direction of $g_\gamma(h^\xi) - \xi^m$. Then each ξ^μ attracts ξ^m with a 'strength' proportional to the resemblance between $g_\gamma(h^\xi)$ and ξ^μ. For instance, if at a given time ξ has a mean value very close to a particular ξ^μ, then we have (provided N is large enough) $\langle h^\xi \rangle \approx \xi^\mu \mathrm{V}_*^2$ and hence $g_\gamma(h^\xi) \approx \xi^\mu$; then each ξ^m will tend to ξ^μ. In more precise terms, the only way to cause each of the M terms of the Equation (4) to vanish, so as to obtain an equilibrium solution, is to make $\langle \xi \rangle = \xi^\mu$ so that $g_\gamma(h^\xi) = \xi^\mu$ and $\xi^m = \xi^\mu \ \ \forall m = 1, \ldots, M$ (besides, of course, $\sigma^2 = \Gamma$). In conclusion, all ξ^m are attracted by the same ξ^μ (note that $g_\gamma(h^\xi)$ does not vary from one term to another).

7. Gaussian Functions as Universal Approximators

We have seen that, given an initial condition in the form of a linear combination of Gaussians, the system evolves according to Equation (4) approaching an equilibrium Gaussian distribution whose mean coincides with one of the p stored memories and whose variance is Γ, the parameter of the stochastic process affecting the dynamics of each individual unit. In this section we will prove that the same result holds for any arbitrary initial distribution, provided it is continuous.

The capacity of Gaussian functions as universal approximators for continuous functions (and, in particular, of neural networks with Gaussian activation functions) has been widely studied. We will refer here to the approach of [3]. Let K be a compact convex subset of R^N and define a two-parameter family of restricted Gaussians as follows:

$$F = \left\{ f_{\zeta,\sigma^2}(\xi) : K \to R;\ f_{\zeta,\sigma^2}(\xi) = \exp\left\{ \frac{-\| \xi - \zeta \|^2}{\sigma^2} \right\}; \sigma^2 > 0, \zeta, \xi \in K \right\}$$

In [3] it is noted that L, the set of all finite linear combinations of elements in F and real coefficients, is an algebra of Gaussians on K (it is closed with respect to multiplication), that L separates points of K and that L does not vanish at any point of K. In those conditions, it is concluded by applying Stone's theorem, that the uniform closure of L contains all real valued continuous functions on K. Moreover, it also follows that L is dense in $C(K)$ and, as a corollary, that any element of $C(K)$ (the set of all continuous functions in K) can be uniformly approximated with an arbitrary precision by elements of span*(F)*.

It is easy to see that the uniform approximability extends to every function in $C(R^N)$, provided it is integrable. Take, for example, $K = S_D$ the N-dimensional sphere of radius D and note that

$$\int_{R^N \setminus S_D} f(\xi) d\xi \overset{D\to\infty}{\longrightarrow} 0$$

if f is continuous and integrable over R^N. Then, in particular, any continuous initial condition for Equation (4) is arbitrarily close, in the uniform topology, to a linear, finite combination of Gaussians. Finally, by virtue of the continuity of equation 4, the system will reach its equilibrium at a Gaussian distribution whose mean coincides with one of the p stored memories, no matter the initial distribution.

8. Conclusions

We have presented a neural network model of associative memory that unifies the two historically more relevant enhancements to the basic Hopfield discrete model [5]: the graded response units approach [6] and the stochastic, Glauber inspired model with a random field representing thermal fluctuations.

A finite temperature dynamics was introduced for the Hopfield model of associative memory with graded-response units, in the same way as the Hopfield discrete model is extended as to include finite temperature effects through the Glauber dynamics [4, 11]. This leads, in the present case, to a Langevin system describing the process at a microscopic level, while the time evolution of the probability density distribution is governed by a multivariate Fokker-Planck equation operating over the space of all possible activation patterns.

In practical terms, the present unified approach has two notable features: (i) greater biological plausibility, resulting from the conjunction of the graded response transfer function for individual units and the stochastic, noisy dynamics of the whole system; (ii) usefulness as a global optimization technique, since it can escape spurious memories, due to the form of equilibrium probability distributions, that assign maximum probability to stored memories.

Although the results presented here from a theoretical point of view are sufficiently rigorous by themselves, it will be interesting to carry out a computational simulation of the whole recall process, simultaneously at both the microscopical level (retrieval of a stored memory from an arbitrary initial state) and the macroscopical one (convergence to equilibrium from an initial distribution). This work is currently in progress. It will also be worthwhile to try this approach on concrete, complex optimization problems.

References

1. Amit, D. J.: *Modeling Brain Function*, Cambridge University Press, Cambridge, 1989.
2. Glauber, R. J.: Time-dependent Statistics of the Ising Model, *Journal of Mathematical Physics* **4** (1963), 294–307.
3. Hartman, E. J., Keeler, J. D. and Kowalsky, J. M.: Layered Neural Networks with Gaussian Hidden Units as Universal Approximations, *Neural Computation* **2**, 210–215.
4. Hinton, G. E., Sejnowsky, T. J.: Optimal Perceptual Inference. In: *Proc. IEEE Conf. Comp. Vision and Patt. Recognition* (Washington, 1983) New York, IEEE. 448–453, 1983.
5. Hopfield, J. J.: Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci.* **79** (1982), 2554–2558.
6. 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.
7. Hopfield, J. J. and Tank, D. W.: 'Neural' Computation of Decisions in Optimization Problems, *Biological Cybernetics* **52** (1985), 141–152.
8. Kampen, N. G. van: *Stochastic Processes in Physics and Chemistry*, Elsevier, Amsterdam, 1997.
9. Little, W. A.: The Existence of Persistent States in the Brain, *Mathematical Biosciences* **19** (1974), 101–120.
10. Little, W. A.: Analytic Study of the Memory Storage Capacity of a Neural Network, *Mathematical Biosciences* **39** (1978), 281–290.
11. Peretto, P.: Collective Properties of Neural Networks: A Statistical Physics Approach, *Biological Cybernetics* **50** (1984), 51–62.
12. Segura, E. C. and Perazzo, R. P. J.: Associative memories in infinite dimensional spaces, *Neural Processing Letters* **12** (2000), 129–144.