The Adaptive Biquadratic Neural Network

David R.C. Dominguez
EPS, Universidad Autonoma de Madrid,
Cantoblanco, 28049 Madrid, Spain,
and
Elka Korutcheva *
Dep. Física Fundamental, Universidad Nacional de Educación a Distancia,
c/Senda del Rey No9, 28080 Madrid, Spain

February 14, 2003

Abstract

We study the macroscopic dynamics of three-state neural network, which are more biologically relevant than binary ones, by using information theory principles and mean-field theory. The results are expressed in terms of the relevant order parameters of the system. By not restricting the learning rule to bilinear weights, an improvement of the storage and information properties of the network is conceived. This is in agreement with the maximization of information by nature evolution. The introduction of a self-adaptive mechanism in the synapses is analyzed.

1 Introduction

Patterns in nature are not binary. Most research on pattern recognition works on black and white statistics, forgetting the richer structure of colors or grey tones. Also, the coding of the information received by neural systems, are not limited to all/nothing, yes/not, but needs fuzzy states (i.e., maybe). This intermediate states comes from, for instance, the variability of the spike train of nervous activity. A first approach to study structured patterns is the use of sparse-coded models[1]. A deeper understanding comes from the use of ternary patterns. In both cases the tools of information theory are useful in the search for the capabilities of the neural network as associative memory device.

In a recent paper we introduced the formalism of the information theory for obtaining the effective Hamiltonian, which maximizes the Mutual Information of the system composed by a three-state neural network [2]. The last is expressed by means of three macroscopic parameters: the overlap between neurons and patterns, the neural activity and the *activity-overlap*, *i.e.*, the overlap restricted to the active neurons. A three-state neural network is defined by a set of $\mu = 1, ..., p$ embedded *ternary* patterns, $\{\xi_i^\mu \in [0, \pm 1]\}$ on sites $i = 1, ..., N$, which are assumed here to be independent random variables that follow the probability distribution

$$p(\xi_i^\mu) = a\delta(|\xi_i^\mu|^2 - 1) + (1-a)\delta(\xi_i^\mu), \qquad (1)$$

where a is the activity of the patterns ($\xi_i^\mu = 0$ are the inactive ones). Accordingly, the neuron states are three-state dynamical variables, defined as $\sigma_i \in \{0, \pm 1\}$, $i = 1, ..., N$ and coupled to other neurons through synaptic connections, for our purpose, of a Hebbian-like form. The active states, $\sigma_i = \pm 1$, become accessible by means of an effective threshold that is built into the model Hamiltonian.

The pattern retrieval task becomes successful if the state of the neuron $\{\sigma_i\}$ matches a given pattern $\{\xi_i^\mu\}$. The measure of the quality of retrieval that we use here is the mutual information, which is a function of the conditional distribution of the neuron states given the patterns, $p(\sigma|\xi)$. The order parameters needed to describe this information are the large-N (thermodynamic) limits of the standard overlap of the μth pattern with the neuron state,

$$m_N^\mu \equiv \frac{1}{aN} \sum_i \xi_i^\mu \sigma_i \to m = \langle \langle \sigma \rangle_{\sigma|\xi} \frac{\xi}{a} \rangle_\xi, \qquad (2)$$

*Permanent Address: G.Nadjakov Inst. Solid State Physics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria

the neural activity,

$$q_{Nt} \equiv \frac{1}{N}\sum_i |\sigma_{it}|^2 \to q = \langle\langle\sigma^2\rangle_{\sigma|\xi}\rangle_\xi, \qquad (3)$$

and the so called *activity-overlap*[1, 3],

$$n^\mu_{Nt} \equiv \frac{1}{aN}\sum_i^N |\sigma_{it}|^2|\xi^\mu_i|^2 \to n = \langle\langle\sigma^2\rangle_{\sigma|\xi}\frac{\xi^2}{a}\rangle_\xi. \qquad (4)$$

In the expressions for the thermodynamic limits, $\lim N \to \infty$, m, q, n, the index μ for the considered pattern were dropped out. The averages are over the conditional distribution, $p(\sigma|\xi)$ and over the pattern distribution, $p(\xi)$, in Eq.(1).

In the next sections we show that, when the patterns are ternary, the network should include biquadratic energy terms in order to perform correctly. Furthermore, the true expansion of the mutual information automatically introduces a mechanism of adapting the synapses according to the neural activity in each time step, which self-controls the evolution of the network and inhibits the appearance of metastable states.

2 Mutual Information

The Mutual Information between patterns and neurons, by regarding the patterns as the input and the neuron states as the output of the channel at each time step [4] is :

$$I[\sigma;\xi] = S[\sigma] - \langle S[\sigma|\xi]\rangle_\xi, \qquad (5)$$

where

$$\begin{aligned} S[\sigma] &\equiv -\sum_\sigma p(\sigma)\ln[p(\sigma)], \\ S[\sigma|\xi] &\equiv -\sum_\sigma p(\sigma|\xi)\ln[p(\sigma|\xi)] \end{aligned} \qquad (6)$$

are the entropy and the conditional entropy of the output, respectively. The quantity $\langle S[\sigma|\xi]\rangle_\xi$ is the so-called equivocation term.

The expressions for the entropies follow from the conditional probability [3],[1]

$$p(\sigma|\xi) = (s_\xi + m\xi\sigma)\delta(\sigma^2 - 1) + (1 - s_\xi)\delta(\sigma) \qquad (7)$$

with $s_\xi \equiv s + \frac{n-q}{1-a}\xi^2$, $s \equiv \frac{q-na}{1-a}$.

We search for an energy function which is symmetric in any permutation of the patterns ξ^μ, since they are not known initially to the retrieval process [2]. This requires that the initial retrieval of any pattern ξ^μ is weak, i.e. the overlap $m^\mu \sim 0$. For general a, q, the variable σ^2 is also initially almost independent of $(\xi^\mu)^2$, so that $n^\mu \sim q$. Hence, the parameter

$$l^\mu \equiv \frac{n^\mu - q}{1-a} =< \sigma^2\eta^\mu >, \ \eta^\mu \equiv \frac{(\xi^\mu)^2 - a}{a(1-a)}, \qquad (8)$$

also vanishes initially in this limit. Note that this parameter is a recognition of a fluctuation in $(\xi^\mu)^2$ by the binary state variable σ^2.

An expansion of the mutual information around $m^\mu = 0, l^\mu = 0$ thus gives

$$I^\mu \approx \frac{1}{2}c_1(m^\mu)^2 + \frac{1}{2}c_2(l^\mu)^2, \qquad (9)$$

where $c_1 = a/q$, $c_2 = c_1(1-a)/(1-q)$ and the total information of the network will be given by summing over all the patterns $I_{pN} = N\sum_\mu I^\mu$. The above expression for I^μ will now be used as a measure of the mutual information during the performance of the network through growing values of m^μ and l^μ. The energy function derived from this expansion governs the network learning and retrieval dynamics, and reads $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$, with

$$\begin{aligned} \mathcal{H}_1 &= -\tfrac{1}{2}\textstyle\sum_{ij} J_{ij}\sigma_i\sigma_j, \qquad J_{ij} = \frac{c_1}{a^2N}\sum_{\mu=1}^p \xi^\mu_i\xi^\mu_j, \\ \mathcal{H}_2 &= -\tfrac{1}{2}\textstyle\sum_{i,j} K_{ij}\sigma_i^2\sigma_j^2, \qquad K_{ij} = \frac{c_2}{N}\sum_{\mu=1}^p \eta^\mu_i\eta^\mu_j, \end{aligned} \qquad (10)$$

the bilinear and biquadratic terms, respectively.

3 Macro-dynamics

Because we are mainly interested on the retrieval properties of our network, we take an initial configuration whose retrieval overlaps are only macroscopic of order $O(1)$ for a given pattern, let say the first one.

Supposing a network configuration in a given time step t, $\{\sigma_{i,t}\}$ with parameters m_t, l_t, q_t, the fields $h_{t=0}$ and $\theta_{t=0}$ in this time step are given by:

$$\begin{aligned} h_t &= c_{1t}(\frac{1}{a}\xi m_t + \omega_t); \qquad \omega_t \equiv \sum_{\nu\geq 2}^p \frac{1}{a}\xi^\nu m^\nu_t \qquad (11) \\ \theta_t &= c_{2t}(\eta l_t + \Omega_t); \qquad \Omega_t \equiv \sum_{\nu\geq 2}^p \eta^\nu l^\nu_t, \end{aligned}$$

where the indices $\mu = 1$ where dropped, and the rest of the patterns is regarded as some additive noise. Here we have defined the dynamical variables $c_{1t} = a/q_t$ and $c_2t = c_{1t}(1-a)/(1-q_t)$.

Supposing a given configuration $\{\sigma_{i,t}\}$ as a collection of independently distributed random variable, with zero-mean and variance q_t, the noises above, ω_t and Ω_t in the time step t, according to the central limit theorem are independent Gaussian distributed[1],[5], with zero mean and variance

$$Var[\omega_t] = \frac{1}{a^2}\alpha q_t \equiv \Delta_t^2$$
$$Var[\Omega_t] = \frac{\Delta_t^2}{(1-a)^2}. \tag{12}$$

In the extremely diluted case, where the first time step describes completely the dynamics, and after taking the asymptotic limit $N \to \infty$, the single-step evolution equations for the overlap m_t, the neural activity q_t and the activity overlap n_t, are the following:

$$m_{t+1} = < \frac{\xi}{a}\overline{\sigma_t} >_\xi = \tag{13}$$
$$\int D\Phi(y) \int D\Phi(z) F_\beta(c_{1t}\frac{m_t}{a} + c_{1t}y\Delta_t; c_{2t}\frac{l_t}{a} + c_{2t}\frac{z\Delta_t}{1-a}),$$

$$q_t = < \overline{\sigma_t^2} >_\xi = a n_t + (1-a)s_t,$$
$$s_{t+1} \equiv < \frac{1-\xi^2}{1-a}\overline{\sigma_t^2} >_\xi = \tag{14}$$
$$\int D\Phi(y) \int D\Phi(z) G_\beta(c_{1t}y\Delta_t; -c_{2t}\frac{l_t}{1-a} + c_{2t}\frac{z\Delta_t}{1-a})$$

and

$$n_{t+1} = < \frac{\xi^2}{a}\overline{\sigma_t^2} >_\xi = \tag{15}$$
$$\int D\Phi(y) \int D\Phi(z) G_\beta(c_{1t}\frac{m_t}{a} + c_{1t}y\Delta_t; c_{2t}\frac{l_t}{a} + c_{2t}\frac{z\Delta_t}{1-a}).$$

Here the averages are over the Gaussian distributions ω, Ω, according to Eq(12). The functions F_β, G_β,

$$F_\beta(h,\theta) = \frac{1}{Z}2e^{\beta\theta}\sinh(\beta h), \;\; G_\beta(h,\theta) = \frac{2}{Z}e^{\beta\theta}\cosh(\beta h) \tag{16}$$

with $Z = 1 + 2e^{\beta\theta}\cosh(\beta h)$, are the mean $\overline{\sigma_t}$ and the square mean $\overline{\sigma_t^2}$ of the neurons over the synaptic noise with parameter $\beta = a/T$, respectively.

4 Results and Conclusions

The aspects of interest we present in this work are twofold: first, the dynamics of biquadratic network, and second the enhancement of its capability to retrieve patterns which are far from the initial states.

The complex behavior of the biquadratic neural network, due to the biquadratic term in the energy function can be seen in the Figure 1. Although there is a symmetry for the axis m_t, corresponding to the bilinear term in the energy, there are no symmetry for the axis l_t (except if the pattern are ternary uniform). First we found the usual retrieval phase, the R, with both $m, l > 0$, and the zero phase, Z, with $m = l = 0$, where no information is transmitted. Besides the phase discovered in reference[2], *i.e.*, the quadrupolar, Q, phase, with $m = 0, l > 0$, a new, yet more strange, phase, we named $negative - quadrupolar$ phase, N, with $m = 0, l < 0$, emerges.

In the Q phase no information about the signs of the patterns is achieved, but the total information doesn't vanishes, because some recognition about the localization of the active patterns still remains. In the N phase we can partially retrieve the opposite of the activities, that means, σ is uncorrelated with ξ, but σ^2 is correlated with $1-\xi^2$ (in a complementary way). The resulting information is again finite.

The Figure 2 shows large plateaus in time the system spends around the Q phase, as the network is trying to recognize the signs of ξ, until it succeeds in retrieve the pattern. It has been represented in Fig.1, by the separatrix lines connecting the Q to the R phases. The information starts small, and during $t = 70$ time steps, it remains around $i = 0$. Then, after a while as long as $\Delta_t = 200$ time steps, the informations suddenly increases to a higher value. The comparison is made with the self-control (SC) model[1], where we see that no intermediate plateau appears for models which do not include a biquadratic energy term, like the bilinear Ising models (here we used an extension of the SC model with noise, the SCT).

Different behaviors are observed for other values of the temperature, as we show in Figure 3. In this figure, we compare the usual biquadratic network (BQN), ($c_1 = c_2 = 1$) with the SCT model, the optimal fixed-threshold model θ_0[5], and the adaptive biquadratic network we studied here (c_1 and c_2 given in the equations, BQS). Although the advantage of the biquadratic model is not seen for the initial conditions used here ($m_0 = l_0 = 1$), we have checked that its basin of attraction of the retrieval region is much larger than this in other models.

Finally, we studied the biquadratic network for the fully-connected architecture. We did it by simulation, using the probabilistic updating according to a Gibbs distribution with energy potential given by the local fields h and θ, in the way: $E = (2-c)h\sigma + c\theta\sigma^2$. The parameter c, plays the role of the $c1, c2$, but now it is kept as a free parameter, since we wish to obtain an optimal c_{opt}, such that the information is maximized for each value of the activity a. The results are presented in Figure 4. The uniform case is optimized just when $c = 1$, as we expected. We found that the values of c_{opt} are small for large a, and large for small a. This means that the closer to the binary patterns, the better the Hopfield (bilinear) model performs. When the patterns have very small activity, a simple sparse-coded network is enough.

The work has shown a series of results comparing different schemes of adaptation of the neural activities to the presence of multi-state patterns. The main result was the appearance of a new phase of information transmission (the N), where the strong states (yes/not) are switched to the fuzzy states. It is plausible in biological learning process, in situations when the environment is very noisy. The origin of our actual model is an indication that natural evolution of neural systems are based on maximization of information. We hope this work can be extended to more states so that realistic images or other patterns can be stored and retrieved without need to compress them in binary patterns.

Acknowledgments: DRCD thanks the Ramon y Cajal fellowship by the MCyT (Spain). EK is financially supported by grant DGI.MCyT BFM2001-291-C02-01 (Spain).

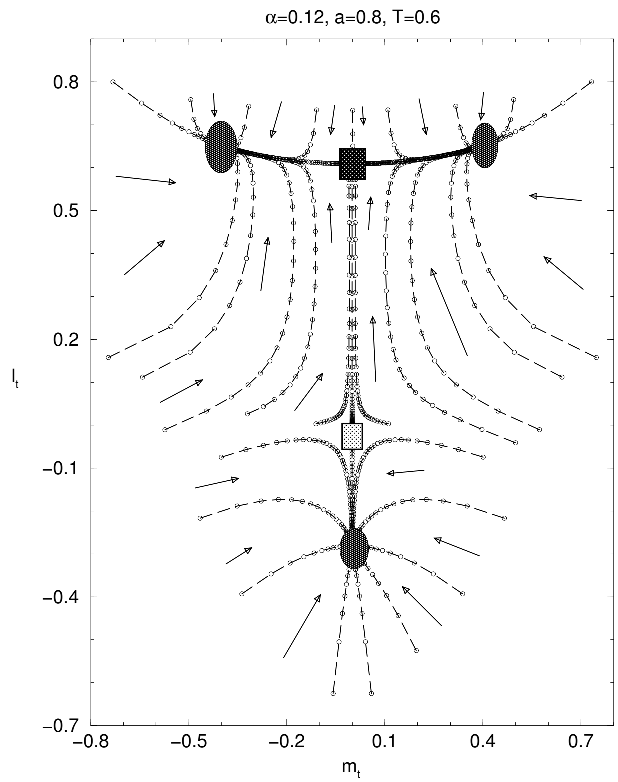

Figure 1: Flow diagram in the order-parameter space (m, l) for a noisy network. The squares are the Q and Z saddle-points; the circles are the R and N attractors.

References

[*] e-mail addresses: david.dominguez@ii.uam.es elka@fisfun.uned.es.

[1] D.Dominguez and D.Bollé, Phys. Rev. Lett. **80**, 2961 (1998).

[2] D.Dominguez and E.Korutcheva, Phys.Rev.E **62(2)**, (2000) 2620-2628

[3] D.Bollé, D.Dominguez and S.Amari, Neural Networks **13**, 455 (2000); D.Bollé and D.Dominguez, Physica A **286**, 401 (2000).

[4] C.E.Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal, v.27 (1948).

[5] D. Bollé, G.M. Shim, B. Vinck, and V. A. Zagrebnov, J. Stat. Phys. **74**, 565 (1994).

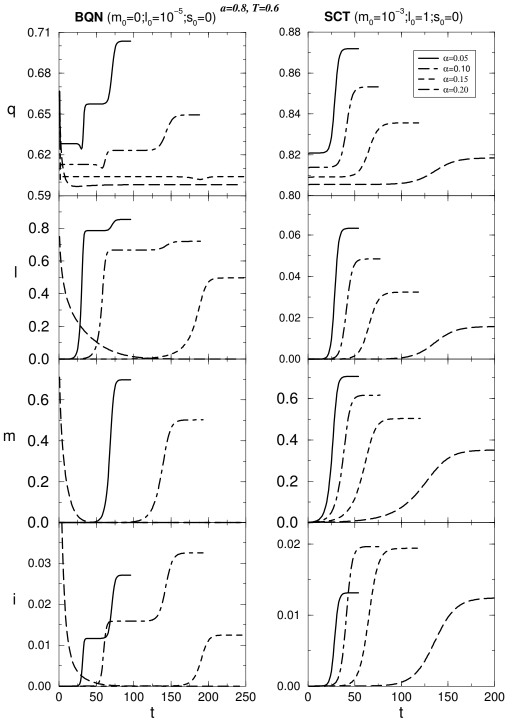

Figure 2: Evolution behavior of the information i and the parameters m, l, q, for the biquadratic network compared with the self-control network, for several values of the load α.

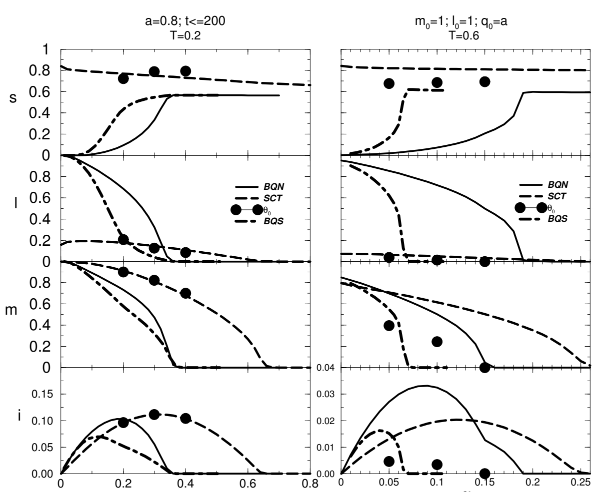

Figure 3: The information i and the parameters m, l, s as a function of the load α, for the BQN, the SCT, θ_0, and BQS, and two values of the noise.

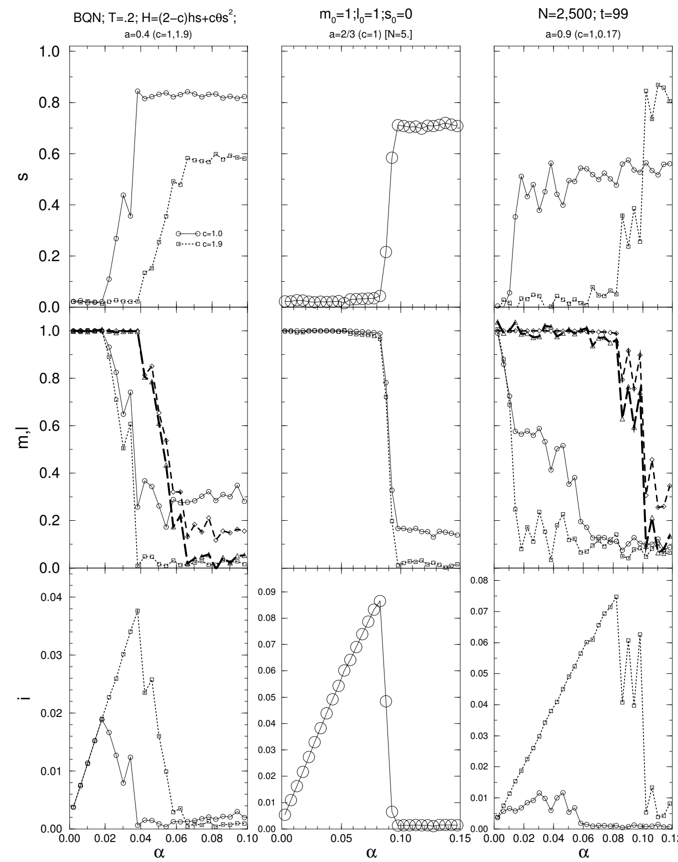

Figure 4: Simulation for the BQS network with free parameters $c = 1$ (solid line) and the optimal c (dotted line), for different values of the activity.