



**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** de la comunicación de congreso publicada en:  
This is an **author produced version** of a paper published in:

Advances in Neural Networks – ISNN 2004: International Symposium on Neural Networks, Dalian, China, August 2004, Proceedings, Part I. Lecture Notes in Computer Science, Volumen 3173. Springer 2004. 14-19

**DOI:** [http://dx.doi.org/10.1007/978-3-540-28647-9\\_3](http://dx.doi.org/10.1007/978-3-540-28647-9_3)

**Copyright:** © 2004 Springer-Verlag

El acceso a la versión del editor puede requerir la suscripción del recurso  
Access to the published version may require subscription

# Mutual Information and Topology 1: Asymmetric Neural Network

David Dominguez\*\* \* Kostadin Koroutchev\*\*  
Eduardo Serrano\*\* and Francisco B. Rodríguez \*\*

EPS, Universidad Autonoma de Madrid,  
Cantoblanco, Madrid, 28049, Spain  
david.dominguez@ii.uam.es

**Abstract.** An infinite range neural network works as an associative memory device if both the learning storage and attractor abilities are large enough. This work deals with the search of an optimal topology, varying the (small-world) parameters: the average connectivity  $\gamma$  ranges from the fully linked to a extremely diluted network; the randomness  $\omega$  ranges from purely neighbor links to a completely random network. The network capacity is measured by the mutual information,  $MI$ , between patterns and retrieval states. It is found that  $MI$  is optimized at a certain value  $\gamma_o$  for a given  $0 < \omega < 1$  if the network is asymmetric.

## 1 Introduction

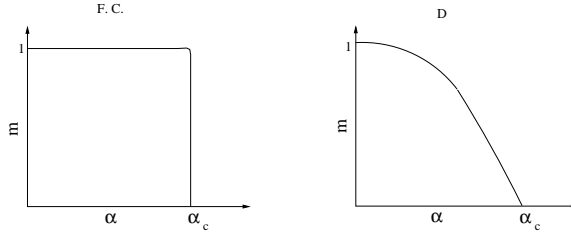
The collective properties of attractor neural networks (*ANN*), such as the ability to perform as an associative memory, has been a subject of intensive research in the last couple of decades[1], dealing mainly with fully-connected topologies. More recently, a renewed interest on ANN has been brought by the study of more realistic architectures, such as small-world or scale-free[2][3] models. The storage capacity  $\alpha$  and the overlap  $m$  with the memorized patterns are the most used measures of the retrieval ability for the Hopfield-like networks[4][5]. Comparatively less attention has been paid to the study of the mutual information  $MI$  between stored patterns and the neural states[6][7].

A reason for this few interest is twofold: first, while  $m$  is a global parameter running only over one site,  $MI$  is a function of the conditional probability, which depends on all states, at input and output; second, the load  $\alpha$  is enough to measure the information if the overlap is close to  $m \sim 1$ , since in this case the information carried by any single (binary, uniform) neuron is almost 1 bit. The first difficulty can be solved for infinite-range connections, because this so called *mean-field networks* satisfies the conditions for the law of large numbers, and  $MI$  is a function only of the macroscopic parameters ( $m$ ) and of the load rate ( $\alpha = P/K$ , where  $P$  is the number of uncorrelated patterns, and  $K$  is the neuron connectivity).

---

\* DD thanks a Ramon y Cajal grant from MCyT.

\*\* Supported by MCyT-Spain BFF-2003-07276 and TIC 2002-572-C02.



**Fig. 1.** The overlap  $m$  vs  $\alpha$  for fully-connected (FC) and extremely diluted (D) networks

The second reason holds for a fully-connected network, for which the critical  $\alpha_c \sim 0.14$ [4], with  $m \sim 0.97$  (it crashes to  $m \rightarrow 0$  for larger  $\alpha$ ): in this case, the information rate is about  $i \sim 0.13$ . Nevertheless, in the case of diluted networks, (the connectivity of each neuron is kept large, but much smaller than  $N$ ), the transition is smooth. In particular, the extremely diluted random network has load capacity  $\alpha_c \sim 0.64$ [8] but the overlap falls continuously to 0, which yields null information at the transition. Such indetermination shows that one must search for the value of  $\alpha$  corresponding to the maximal  $MI$ , instead of  $\alpha_c$ . It is seen in Fig.1. Previous works[3] studied only the overlap  $m(\alpha)$ .

Our main goal in this work is to solve the following question: how does the maximal  $MI_m(\gamma, \omega) \equiv MI(\alpha_m, m; \gamma, \omega)$  behaves with respect to the network topology? We will show that, for larger values of the randomness  $\omega$ , the extremely-diluted network performs the best. However, with asymmetric connections, smaller values of  $\omega$  lead to an optimal  $MI_o(\gamma) \equiv MI_m(\gamma_o, \omega)$  for intermediate levels of connectivity  $0 < \omega < 1$ .

## 2 The Information Measures

### 2.1 The Neural Channel

The network state in a given time  $t$  is defined by a vector of binary neurons,  $\sigma^t = \{\sigma_i^t \in \{\pm 1\}, i = 1, \dots, N\}$ . Accordingly, each pattern,  $\xi^\mu = \{\xi_i^\mu \in \{\pm 1\}, i = 1, \dots, N\}$ , are site-independent random variables, binary and uniformly distributed:

$$p(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1). \quad (1)$$

The network learns a set of independent patterns  $\{\xi^\mu, \mu = 1, \dots, P\}$ .

The task of the neural channel is to retrieve a pattern (say,  $\xi^\mu$ ) starting from a neuron state which is inside its attractor basin,  $B(\xi^\mu)$ :  $\sigma^0 \in B(\xi^\mu) \rightarrow \sigma^\infty \approx \xi^\mu$ . This is achieved through a network dynamics, which couples neurons  $\sigma_i, \sigma_j$  by the *synaptic matrix*  $\mathbf{J} \equiv \{J_{ij}\}$  with cardinality  $\#\mathbf{J} = N \times K$ .

### 2.2 The Overlap

For the usual binary non-biased neurons model, the relevant order parameter is the *overlap* between the neural states and a given pattern:

$$m_N^{\mu t} \equiv \frac{1}{N} \sum_i \xi_i^\mu \sigma_i^t, \quad (2)$$

in the time step  $t$ . Note that both positive and negative pattern,  $-\xi$ , carries the same information, so the absolute value of the overlap is the measure of the retrieval quality:  $|m| \sim 1$  means a good retrieval.

Alternatively, one can measure the error in retrieving using the (square) Hamming distance:  $E_N^{\mu t} \equiv \frac{1}{N} \sum_i |\xi_i^\mu - \sigma_i^t|^2 = 2(1 - m_N^{\mu t})$ . Together with the overlap, one needs a measure of the load capacity, which is the rate of pattern bits per synapses used to store them. Since the synapses and patterns are independent, the load is given by  $\alpha = \#\{\xi^\mu\}/\#\mathbf{J} = (PN)/(NK) = P/K$ .

We require our network has long-range interactions. That means, it is a mean-field network (MFN), the distribution of the states is site-independent, so every spatial correlation such as  $\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$  can be neglected, which is reasonable in the limit  $N \rightarrow \infty$ . Hence the condition of the law of large numbers, are fulfilled. At a given time step of the dynamical process, the network state can be described by one particular overlap, let say  $m_N^t \equiv m_N^{\mu t}$ . The order parameters can thus be written in the asymptotic limit,  $\lim N \rightarrow \infty$ , as  $m^t = \langle \sigma^t \xi \rangle_{\sigma, \xi}$ . The brackets represent average over the joint distribution of  $\sigma, \xi$  for a single neuron (we can drop the index  $i$ ). These are the macroscopic variables describing the state of the network at a given, unspecified, time step  $t$  of the dynamics.

### 2.3 Mutual Information

For this long-range system, it is enough to observe the distribution of a single neuron in order to know the global distribution. This is given by the conditional probability of having the neuron in a state  $\sigma$ , (we also drop the time index  $t$ ), given that in the same site the pattern being retrieved is  $\xi$ . For the binary network we are considering,  $p(\sigma|\xi) = (1 + m\sigma\xi)\delta(\sigma^2 - 1)/2$  where the overlap is  $m = \langle \langle \sigma \rangle_{\sigma|\xi} \xi \rangle_{\xi}$ , using  $p(\sigma, \xi) = p(\sigma|\xi)p(\xi)$ .

The joint distribution of  $p(\sigma, \xi)$  is interpreted as an ensemble distribution for the neuron states  $\{\sigma_i\}$  and inputs  $\{\xi_i\}$ . In the conditional probability,  $p(\sigma|\xi)$ , all type of noise in the retrieval process of the input pattern through the network (both from environment and over the dynamical process itself) is enclosed.

With the above expressions and  $p(\sigma) \equiv \sum_{\xi} p(\xi)p(\sigma|\xi) = \delta(\sigma^2 - 1)/2$ , we can calculate the *Mutual Information*  $MI$  [7], a quantity used to measure the prediction that the observer of the output ( $\sigma^t$ ) can do about the input ( $\xi^\mu$ ) at each time step  $t$ . It reads  $MI[\sigma; \xi] = S[\sigma] - \langle S[\sigma|\xi] \rangle_{\xi}$ , where  $S[\sigma]$  is the entropy and  $S[\sigma|\xi]$  is the conditional entropy. We use binary logarithms,  $\log \equiv \log_2$ . The entropies are [9]:

$$\langle S[\sigma|\xi] \rangle_{\xi} = -\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2}, \quad S[\sigma] = 1 [bit]. \quad (3)$$

When the network approaches its saturation limit  $\alpha_c$ , the states can not remain close to the patterns, then  $m_c$  is usually small. We avoid this writing the

information rate as

$$i(\alpha, m) = MI[\{\sigma\}|\{\xi\}]/\#\mathbf{J} \equiv \sum_{i\mu} MI[\sigma_i|\xi_i^\mu]/(KN) = \alpha MI[\sigma; \xi]. \quad (4)$$

The information  $i(\alpha, m)$  is a non-monotonic function of the overlap, which reaches its maximum value  $i_m = i(\alpha_m, m)$  at some value of the load  $\alpha$ .

### 3 The Model

#### 3.1 The Network Topology

The synaptic couplings are  $J_{ij} \equiv c_{ij}K_{ij}$ , where the connectivity matrix has a regular and a random parts,  $\{c_{ij} = c_{ij}^n + c_{ij}^r\}$ . The regular part connects the  $K_n$  nearest neighbors,  $c_{ij}^n = \sum_{k \in V} \delta(i - j - k)$ , or  $V = \{1, \dots, K_n\}$  in the asymmetric case, in a closed one-dimensional lattice. The random part consists of independent random variables  $\{c_{ij}^r\}$ , distributed as

$$p(c_{ij}^r) = c_r \delta(c_{ij}^r - 1) + (1 - c_r) \delta(c_{ij}^r), \quad (5)$$

where  $c_r = K_r/N$ , with  $K_r$  = the mean number of random connections of a single neuron. Hence, the neuron connectivity is  $K = K_n + K_r$ . The network topology is then characterized by two parameters: the *average-connectivity*, defined as  $\gamma = K/N$ , and the *randomness* ratio,  $\omega = K_r/K$ , besides its symmetry constraints. The  $\omega$  plays the role of the probability of rewiring in the *small-world* model (SW)[2]. Our analysis of the present topology shows the same qualitative behavior for the clustering and mean-length-path of the original SW. The average-connectivity is normalized with the same scale as the information, since the load rate is  $\alpha = P/K$ .

The learning algorithm,  $K_{ij}$ , is given by the Hebb rule

$$K_{ij}^\mu = K_{ij}^{\mu-1} + \frac{1}{K} \xi_i^\mu \xi_j^\mu. \quad (6)$$

The network start in  $K_{ij} = 0$ , and after  $\mu = P = \alpha K$  learning steps, it reaches a value  $K_{ij} = \frac{1}{K} \sum_{\mu}^P \xi_i^\mu \xi_j^\mu$ . The learning stage is a slow dynamics, being stationary-like in the time scale of the much faster retrieval stage.

#### 3.2 The Neural Dynamics

The neurons states,  $\sigma_i^t \in \{\pm 1\}$ , are updated according to the stochastic dynamics:

$$\sigma_i^{t+1} = \text{sign}(h_i^t + Tx), \quad i = 1 \dots N, \quad h_i^t \equiv \sum_j J_{ij} \sigma_j^t, \quad (7)$$

where  $x$  is a random variable and  $T$  is the temperature-like environmental noise. In the case of symmetric synaptic couplings,  $J_{ij} = J_{ji}$  it can be defined an

energy function  $H_s = -\sum_{(i,j)} J_{ij} \sigma_i \sigma_j$  whose minima are the stable states of the dynamics Eq.(7).

In the present paper, we work out the asymmetric network (no constraints  $J_{ij} = J_{ji}$ ), for which no energy function can be written. We restrict our analysis also for the deterministic dynamics ( $T = 0$ ). The only stochastic effects comes from the large number of learned patterns ( $P = \alpha K$ , cross-talk noise).

## 4 Results

We have studied the behavior of the network varying the range of connectivity  $\gamma$  and randomness  $\omega$ . In all cases we used the parallel dynamics, Eq.(7). The simulation was carried out with  $N \times K = 25 \cdot 10^6$  synapses, storing  $K_{ij}$  and  $c_{ij}$ , instead of  $J_{ij}$ , i.e., we use adjacency list as data structure. For instance, with  $\gamma \equiv K/N = 0.01$ , we used  $K = 500, N = 5 \cdot 10^4$ . In [3] the authors use  $K = 50, N = 5 \cdot 10^3$ , which is far from asymptotic limit.

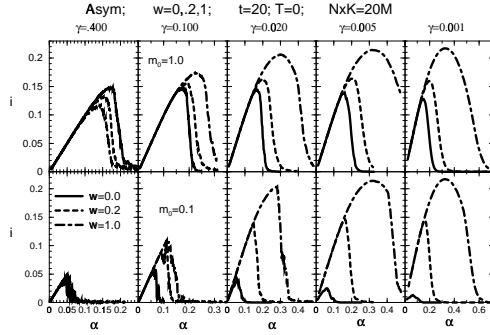
We studied the asymmetric network by searching for the stationary states of the network dynamics; we suppose  $t = 20$  parallel (all neurons) updates are enough for convergence, which is true in mostly cases. We look at the behavior of fixed-point of the overlap  $m^t$ , with the memory loading  $P = \alpha K$ . We averaged over a window in the axis of  $P$ , usually  $\delta P = 25$ .

In first place, we checked for the stability properties of the network: the neuron states start precisely at a given pattern  $\xi^\mu$  (which changes at each learned step  $\mu$ ). The initial overlap is  $m_0^\mu = 1$ , so, after  $t \leq 20$  time steps in retrieving, the information  $i(\alpha, m; \gamma, \omega)$  for final overlap is calculated. We plot it as a function of  $\alpha$ , and its maximum  $i_m \equiv i(\alpha_m, m; \gamma, \omega)$  is evaluated. This is repeated for various values of the average-connectivity  $\gamma$  and randomness  $\omega$  parameters. The results are in the upper part of Fig.2.

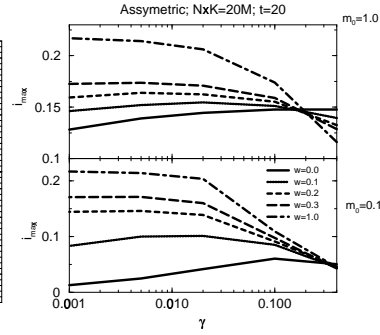
Second, we checked for the retrieval properties: the neuron states start far from a learned pattern, but inside its basin of attraction,  $\sigma^0 \in B(\xi^\mu)$ . The initial configuration is chosen with distribution:  $p(\sigma^0|\xi) = (1 + m^0)/2\delta(\sigma^0 - \xi) + (1 - m^0)/2\delta(\sigma^0 + \xi)$ , for all neurons (so we avoid a bias between regular/random neighbors). The initial overlap is now  $m^0 = 0.1$ , and after  $t \leq 20$  steps, the final information  $i(\alpha, m; \gamma, \omega)$  is calculated. The results are in the lower part of Fig.2. The first observation in both parts is that the maximal information  $i_{max}(\gamma; \omega)$  increases with dilution (smaller  $\gamma$ ) if the network is more regular,  $\omega \simeq 1$ , while it decreases with dilution if the network is more random,  $\omega \simeq 0$ .

The comparison between stability and retrieval properties, looking at upper and lower parts of Fig.2, shows clearly that the relation between dilution and randomness is powered respect to the basins of attraction. Random topologies have very robust attractors only if the network is diluted enough; otherwise, they are strongly damaged by bad initial conditions (while regular topologies almost lost their retrieval abilities with extreme dilution).

Each maxima of  $i_{max}(\gamma; \omega)$  in Fig.2 is plotted in Fig.3. We see that, for intermediate values of the randomness parameter  $0 < \omega < 1$  there is an optimal information respect to the dilution  $\gamma$ . We observe that the optimal  $i_o \equiv i_m(\gamma_o; \omega)$



**Fig. 2.** The information  $i \times \alpha$  with  $\omega = 0, 0.2, 1$  and several  $\gamma$ ; simulation with  $N \times K = 20M$ ,  $t \leq 20$ . Upper panel: initial overlap  $m^0 = 1.0$ . Lower panel:  $m^0 = 0.1$ .



**Fig. 3.** The maximal information  $i_{max} \times \gamma$  with  $\omega = 0, .1, .2, .3, 1$ ; simulation. The initial overlap  $m_0 = 1.0$  (upper) and  $m_0 = 0.1$  (lower).

is shifted to the left (stronger dilution) when the randomness  $\omega$  of the network increases. For instance, with  $\omega = 0.1$ , the optimal is at  $\gamma \sim 0.02$  while with  $\omega = 0.3$ , it is  $\gamma \sim 0.005$ . This result does not change qualitatively with the initial condition, but respect to the attractors ( $m_0 = 0.1$ ), even the regular topology presents an optimum at  $\gamma \sim 0.1$ .

## 5 Conclusions

In this paper we have studied the dependence of the information capacity with the topology for an attractor neural network. We calculated the mutual information for a Hebbian model, for storing binary patterns, varying the connectivity and randomness parameters, and obtained the maxima respect to  $\alpha$ ,  $i_m(\gamma, \omega)$ . Finally we also presented results for the basins of attraction.

We found there is an optimal  $\gamma_o$  for which an optimized topology, in the sense of the information,  $i_m(\gamma_o)$ , holds. We believe that the maximization of information respect to the topology could be a biological criterion to build real neural networks. We expect that the same dependence should happens for more structured networks and learning rules.

## References

1. Hertz, J., Krogh, J., Palmer, R.: Introduction to the Theory of Neural Computation. Addison-Wesley, Boston (1991)
2. Strogatz, D., Watts, S.: Nature **393** (1998) 440
3. McGraw, P., Menzinger, M.: Phys. Rev. E **68** (2004) 047102
4. Amit, D., Gutfreund, H., Sompolinsky, H.: Phys. Rev. A **35** (1987) 2293
5. Okada, M.: Neural Network **9/8** (1996) 1429
6. Perez-Vicente, C., Amit, D.: J. Phys. A, **22** (1989) 559
7. Dominguez, D., Bolle, D.: Phys. Rev. Lett **80** (1998) 2961
8. Derrida, B., Gardner, E., Zippelius, A.: Europhys. Lett. **4** (1987) 167
9. Bolle, D., Dominguez, D., Amari, S.: Neural Networks **13** (2000) 455