Associative Memory: Information and Topology

D. Dominguez, K. Koroutchev, E. Serrano and F.B. Rodriguez

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

(Paper received on August 01, 2006, accepted on September 25, 2006)

Abstract. A neural network works as an associative memory device if it has large storage capacity and the quality of the retrieval is good enough. The learning and attractor abilities of the network both can be measured by the mutual information (MI), between patterns and retrieval states. This paper deals with a search for an optimal topology, of a Hebb network, in the sense of the maximal MI. We use small-world topology. The connectivity γ ranges from an extremely diluted to the fully connected network; the randomness ω ranges from purely local to completely random neighbors. It is found that, while stability implies an optimal $MI(\gamma, \omega)$ at $\gamma_{opt}(\omega) \to 0$, for the dynamics, the optimal topology holds at certain $\gamma_{opt} > 0$ whenever $0 \leq \omega < 0.3$.

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, the interest on ANN has been renewed by the study of more realistic architectures, such as small-world [3] or scale-free [13] models. The storage capacity α_c and the overlap m with the memorized patterns are the most used measures of the retrieval ability for the Hopfield-Hebb networks[4]Comparatively less attention has been paid to the study of the mutual information (MI) between stored patterns and the neural states[5][6], although neural networks are information processing machines.

A reason for this relatively low interest is twofold: on the one hand, it is easier to deal with the global parameter $m[\sigma, \xi]$, than with $MI[p(\sigma|\xi)]$, a function of the conditional probability of neuron states σ given the patterns ξ. This can be solved for the so called *mean-field networks* which satisfy the law of large numbers, hence MI is a function only of the macroscopic parameters m, and the load rate $\alpha = P/K$ (where P is the number of uncorrelated patterns, and K is the neuron connectivity). On the other hand, the load α 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 neuron is almost 1 bit. It is true for a fully-connected (FC) network, for which the critical $\alpha_c^{FC} \sim 0.138$ [4], with $m_c^{FC} \sim 0.97$ (with a sharp transition to $m \to 0$ for larger $\alpha \geq \alpha_c$): in this case, the information rate is about $i_c^{FC} \sim 0.131$, as can be seen in the left panel of Fig.1. There we show the

© H. Sossa and R. Barrón (Eds.)
Special Issue in Neural Networks and Associative Memories
Research in Computing Science 21, 2006, pp. 39-48

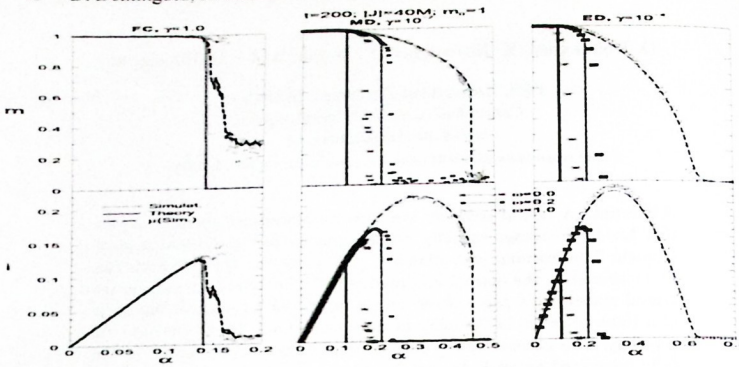

Fig. 1. *The overlap m and the information i vs α for different architectures: fully-connected, $\gamma^{FC} = 1.0$ (left), moderately-diluted, $\gamma^{MD} = 10^{-2}$ (center) and extremely-diluted, $\gamma^{ED} = 10^{-4}$ (right). Symbols represents simulation with initial overlap $m^0 = 1$ and $|J| = 40M$, with local (stars, $\omega = 0.0$), small-world (filled squares, $\omega = 0.2$), and random (circles, $\omega = 1.0$) connections. Lines are for theoretical results: solid, $\omega = 0.0$, dotted, $\omega = 0.2$, and dashed, $\omega = 1.0$. In left, dashed line means averaging the simulation.*

overlap (upper) and information for several architectures. However, in the case of diluted networks the transition is smooth. In particular, the random extremely diluted (RED) network has load capacity $\alpha_c^{RED} \sim 0.64$[8] but the overlap falls continuously to $m_c^{RED} \sim 0$, which yields null information at the transition, $i_c^{RED} \sim 0.0$, as seen in right panel of Fig.1 (dashed line). Such indetermination shows that one must search for the value of α_{max} corresponding to the maximal information $MI_{max} \equiv MI(\alpha_{max})$, instead of α_c.

We address the problem of searching for the optimal topology, in the sense of maximizing the mutual information. Using the graph framework [2], one can capture the main properties of a wide range of neural systems, with only 2 parameters: $\gamma \equiv K/N$, which is the average rate of links per neurons, where N is the network size, and ω, which controls the rate of random links (among all neighbors). When γ is large, the clustering coefficient is large ($c \sim 1$) and the mean-length-path between neurons is small ($l \sim \ln N$), whatever ω is. When γ is small, then if ω is too small, $c \sim 1$ and $l \sim N/K$, but if it is about $\omega \sim 0.1$, the network behaves again as if $\gamma \sim 1$, with $c \sim 1$ and $l \sim \ln(N)$. This region, called small-world (SW), is rather usefull when one is interested to built networks where the information transmition is fast and efficient, with high capacity in presence of significant noise, but do not wants to spent too much wiring.Small-world networks may model many biological systems [12]. For

instance, in a brain local connections dominate in intracortex, while there are a few intercortical connections [11].

In Fig.1 we show the overlap (upper) and information for several architectures. In the left panel, it is seen that the maximum information rate, $i \equiv MI/(K.N)$, of FC network is about $i_{max}^{FC} = 0.135$, while in the right panel, we show extremely-diluted networks (ED). The RED network ($\omega = 1.0$) has $i_{max}^{RED} \sim 0.223$. The right panel of Fig.1 plot also the overlap and the information for the local extremely diluted network (LED, $\omega = 0.0$), with $i_{max}^{LEC} = 0.0855$, and a small-world extremely diluted network (SED, $\omega = 0.2$), with $i_{max}^{SED} = 0.165$. We see that the ED transitions are smooth. The central panel of Fig.1 plot moderately diluted (MD) networks, which are commented later. Theoretical results fit well with the simulations, except for small ω, where theory underestimate it. Previous works about small-world attractor neural networks [10] studied only the overlap $m(\alpha)$, so no result about information were known.

Our main goal in this work is to solve the following question: how does the maximal information, $i_{max}(\gamma, \omega) \equiv i(\alpha_{max}; \gamma, \omega)$ behaves with respect to the network topology? To our knowledge, up to now, there were no answer to this question. We will show that, near to the stationary retrieval states, for every value of the randomness $\omega > 0$, the extremely-diluted network, performs the best, $\gamma_{opt} \to 0$. However, regarding the attractor basins, starting far from the patterns, the optimal topology holds for moderate γ_{opt}. For instance, if transients are taken in account, values of $\omega \sim 0.1$ lead to an optimal $i_{opt}(\gamma) \equiv i_{max}(\gamma_{opt}, \omega)$ with $\gamma_{opt} \sim 10^{-2}$.

The structure of the paper is the following: in the next section we review the information measures used in the calculations; in Sec.3, we define the topology and neuro-dynamics model. The results are shown in Sec.4, where we study retrieval by theory and simulation (with random patterns and with images); conclusions are drawn in last section.

2 The Information Measures

2.1 The Neural Channel

The network state at a given time t is defined by a set of binary neurons, $\boldsymbol{\sigma}^t = \{\sigma_i^t \in \{\pm 1\}, i = 1, ..., N\}$. Accordingly, each pattern $\boldsymbol{\xi}^\mu = \{\xi_i^\mu \in \{\pm 1\}, i = 1, ..., N\}$, is a set of site-independent random variables, binary and uniformly distributed: $p(\xi_i^\mu = \pm 1) = 1/2$. The network learns a set of independent patterns $\{\boldsymbol{\xi}^\mu,\ \mu = 1, ..., P\}$.

The task of the neural channel is to retrieve a pattern (say, $\boldsymbol{\xi}$) starting from a neuron state which is inside its attractor basin, $B(\boldsymbol{\xi})$, *i.e.*: $\boldsymbol{\sigma}^0 \in B(\boldsymbol{\xi}) \to \boldsymbol{\sigma}^\infty \approx \boldsymbol{\xi}$. This is achieved through a network dynamics, which couples neighbor neurons $\sigma_i . \sigma_j$ by the *synaptic matrix* $\mathbf{J} \equiv \{J_{i,j}\}$ 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:

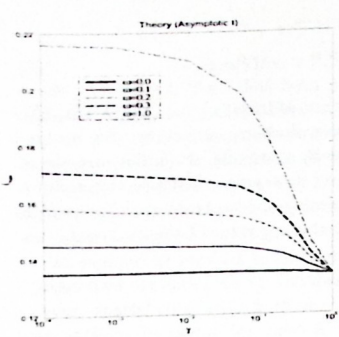

Fig. 2. Maximal information $i_{max} = i(\alpha_{max})$ vs γ. Theoretical results for the stationary states, with several values of randomness ω.

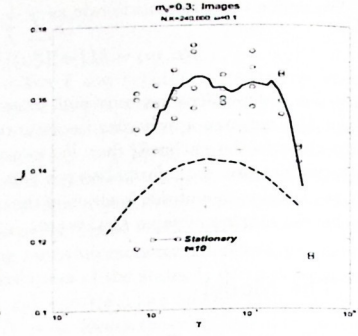

Fig. 3. i_{max} vs γ. Simulation with $\omega = 0.1$ and $m^0 = 0.3$. Dynamics stop at $t = 10$ (Plus dots, Solid line) or at $t = 100$ (Circles, Dashed line).

where x is a normalized random variable and T is the temperature-like environmental noise. In the case of symmetric synaptic couplings, $J_{ij} = J_{ji}$, an energy function $H_s = -\sum_{(i,j)} J_{ij}\sigma_i\sigma_j$ can be defined, whose minima are the stable states of the dynamics Eq.(5).

In the present paper, we work out the asymmetric network by simulation (no constraints $J_{ij} = J_{ji}$). The theory was carried out for symmetric networks. As it is seen in Fig.1, theory and simulation shows similar results, except for local networks (theory underestimate α_{max}, where the symmetry may play some role. We restrict our analysis also for the deterministic dynamics ($T = 0$). The stochastic macro-dynamics comes from the extensive number of learned patterns, $P = \alpha K$.

4 Results

We studied the information for the stationary and dynamical states of the network were studied as a function of the topological parameters, ω and γ. A sample of the results for simulation and theory is shown in Fig.1, where the stationary states of the overlap and information are plotted for the FC, MD and ED architectures. It can be seen that information increases with dilution and with randomness of the network. A reason for this behavior is that dilution decreases the correlation due to the interference between patterns. However, dilution also increases the mean-path-length of the network, thus, if the connections are local, the information flows slowly over the network. Hence, the neuron states can

be eventually trapped in noisy patterns. So, i_{max} is small for $\omega \sim 0$ even if $\gamma = 10^{-4}$.

4.1 Theory: Stationary States

Following to the Gardner calculations[8], at temperature T=0 the MFN approximation gives the fixed point equations:

$$m = \mathrm{erf}(m/\sqrt{r\alpha}), \tag{6}$$

$$\chi = 2\varphi(m/\sqrt{r\alpha})/\sqrt{r\alpha}; \tag{7}$$

$$r = \sum_{k=0}^{\infty} a_k(k+1)\chi^k, \;\; a_k = \gamma Tr[(\mathbf{C}/K)^{k+2}] \tag{8}$$

with $\mathrm{erf}(x) \equiv 2\int_0^x \varphi(z)dz$, $\varphi(z) \equiv e^{-z^2/2}/\sqrt{2\pi}$. The parameter a_k is the probability of existence of cycle of length $k+2$ in the connectivity graph. The a_k can be calculated either by using Monte Carlo [14], or by an analytical approach, which gives $a_k \sim \sum_m \int d\theta[p(\theta)]^k e^{im\theta}$, where $p(\theta)$ is the Fourier transform of the probability of links, $p(C_{ij})$. For an RED and FC networks one recover the known results for $r^{RED} = 1$ and $r^{FC} = 1/(1-\chi)^2$ respectively [1].

The theoretical dependence of the information on the load, for FC, MD and ED networks, with local, small-world and random connections, are plotted in the fat lines in Fig.1. A comparison between theory and simulation is also given in Fig.1. It can be seen that both results agree for most $\omega > 0$, but theory fails for $\omega = 0$. One reason is that theory uses symmetric constraint, while simulation was carried out with asymmetric synapsis. Figure 2 shows their maxima $i(\alpha_{max})$ vs. the parameters (ω, γ). It is seen that the optimal is at $\omega \to 1, \gamma \to 0$. This implies that the best topology for information (stationary states) is the extreme diluted network, with purely random connectivity.

4.2 Simulation: Attractors and Transients

We have studied the behavior of the network varying the range of connectivity γ and randomness ω. We used Eq.(5). Both local and random connections are asymmetric. The simulation was carried out with $N \times K = 36 \cdot 10^6$ synapses, storing an adjacency list as data structure, instead of J_{ij}. For instance, with $\gamma \equiv K/N = 0.01$, we used $K = 600, N = 6 \cdot 10^4$. In [10] the authors use $K = 50, N = 5 \cdot 10^3$, which is far from asymptotic limit.

We studied the network by searching for the stability properties and transients of the neuron dynamics. To look for stability, we started the network at some pattern (with initial overlap $m^0 = 1.0$), and wait until it stays or leave it after a flag time step $t = t_f$ (unless it converges to a fixed point m^* before $t = t_f$). When we check transients, we start with $m^0 = 0.1$, and stop the dynamics at the time t_f. Usually, $t_f = 20$ parallel (all neurons) updates is a large enough delay for retrieval. Indeed in most case far before the saturation, after

$t_f = 4$ the network end up in a pattern, however, near α_{max}, even after $t_f = 100$ the network has not yet relaxed.

In first place, we checked for the stability properties of the network: the neuron states start precisely at a given pattern $\boldsymbol{\xi}^\mu$ (which changes at each learned step μ). The initial overlap is $m_0^\mu = 1.0$, so, after $t_m \leq 20$ time steps in retrieving, the information $i(\alpha, m; \gamma, \omega)$ for final overlap is calculated. We plot it as a function of α, and its maximum $i_{max} \equiv i(\alpha_{max}; \gamma, \omega)$ is evaluated. Second, we checked for the retrieval properties: the neuron states start far from a learned pattern, but inside its basin of attraction, $\boldsymbol{\sigma}^0 \in B(\boldsymbol{\xi}^\mu)$. The initial configuration is chosen with distribution: $p(\sigma_i^0 = \pm \xi_i^\mu | \xi_i^\mu) = (1 \pm m^0)/2$, for all neurons (so we avoid a bias between local/random neighbors). The initial overlap is now $m^0 = 0.1$, and after $t_f \leq 20$ steps, the information $i(\alpha, m; \gamma, \omega)$ is calculated.

Each maximal $i_{max}(\gamma; \omega)$ is plotted in Fig.4. We see that, for intermediate values of the randomness parameter $0 \leq \omega < 0.3$ there is an optimal information respect to the dilution γ, if dynamics is truncated. We observe that the optimal $i_{opt} \equiv i_{max}(\gamma_{opt}; \omega)$ is shifted to the left (stronger dilution) when the randomness ω of the network increases. For instance, with $\omega = 0.1$, the optimal is at $\gamma \sim 0.020$ while with $\omega = 0.2$, it is $\gamma \sim 0.005$. This result does not change qualitatively with the flag time, but if the dynamics is truncated early, the optimal γ_{opt}, for a fixed ω, is shifted to more connected networks. However, the behavior depends strongly on the initial condition: respect to $m_0 = 0.1$, where the maximal are pronounced, with $m_0 = 1.0$, the dependence on the topology becomes almost flat. We see also that for $\omega \geq 0.3$ there is no intermediate optimal topology. It is worth to note that the simulation converges to the theoretical results if $m_0 = 1.0$ when $t \to \infty$.

One can understand this non-monotonic behavior of the information in terms of the basins of attraction. Random topologies have very deep attractors, specially if the network is diluted enough, while regular topologies almost lose their retrieval abilities with dilution. However, since the basins becomes rougher with dilution, then network takes longer to reach the attractor. Hence, the competition between depth-roughness is won by the more robust MD networks.

4.3 Simulation with Images

The simulations presented so far use artificial patterns randomly generated. In order to check if our results are robust against possibly correlations existent in realistic patterns, we test the algorithm with images. We see that the same non-monotonic behavior for $i_{max}(\gamma)$ is observed here.

We have checked the results by using data derived from the Waterloo image database. We are working with square shaped patches. In order to use Hebb-like non-sparse code binary network and still preserve the structure of the image we process the images preserving the edges, by applying edge filter. Each pixel of the patch represents a different neuron. The number of connections is up to $N \times K = 3 \cdot 10^5$ and the feasible connectivities (more than 3 patterns) are $\gamma > 0.002$.

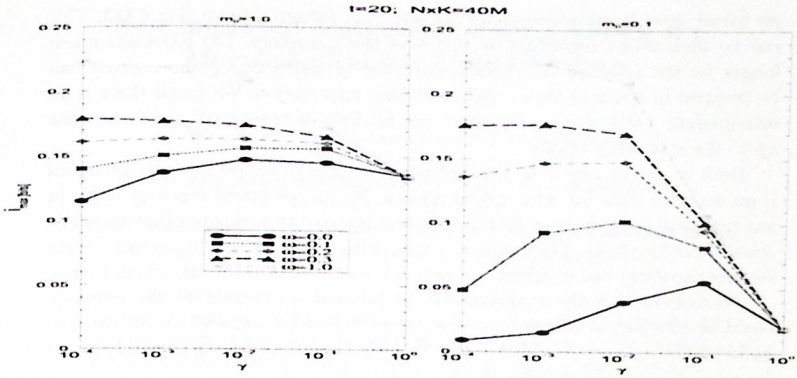

Fig. 4. Maximal information $i_{max} = i(\alpha_{max})$ vs γ, for simulations with $N.K = 4.10^7$, and several ω. Initial overlap $m^0 = 1.0$ (left) and $m^0 = 0.1$ (right); the retrieval stops after $t_f = 20$ steps.

Note that the procedure, strictly speaking, does not guarantee the conditions for the distribution of ξ, because neither $p(\xi = \pm 1)$ is uniform (due to the threshold in large blocks), nor ξ_i are uncorrelated (due to image edges).

We are choosing at random the origin of the patch and the image to be used from the available 12 images. The topology of the network is a ring with small world topology. The results of the simulation, using Chen filter, are shown in Fig.3. The optimal connectivity with $\omega = 0.1$ and $t_f = 10$ is found to be $\gamma_{opt} \sim 0.03$. The fluctuation now are much larger than with random patterns, due to correlation and small network size. In the stationary states, $t_f \to \infty$, the optimal connectivity remains at $\gamma_{opt} \sim 0.03$, with $i_{opt} \sim 0.165$. The results agree qualitatively with simulation for random patterns, Fig.4, where the initial overlaps are $m^0 = 0.1$ and $m^0 = 1.0$ (in Fig.3 it is always $m^0 = 0.3$).

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 Hebb model, for storing binary patterns, varying the connectivity (γ) and randomness (ω) parameters, and obtained the maximal respect to α, $i_{max}(\gamma, \omega) \equiv i(\alpha_{max}; \gamma, \omega)$. Then we look at the optimal topology, γ_{opt} in the sense of the information, $i_{opt} \equiv i_{max}(\gamma_{opt}, \omega)$. We presented stationary and transient states. The main result is that larger ω always leads to higher information i_{max}.

From the stability calculations, the stationary optimal topology, is the extremely diluted (RED) network. Dynamics shows, however, that this is not true:

we found there is an intermediate optimal γ_{opt}, for any fixed $0 \leq \omega < 0.3$. This can be understood regarding the shape of the attractors. The ED waits much longer for the retrieval than more connected networks do, so the neurons can be trapped in spurious states with vanishing information. We found there is an intermediate optimal γ_{opt}, whenever the retrieval is truncated, and it remains up to the stationary states.

Both in nature and in technological approaches to neural devices, dynamics is an essential issue for information process. So, an optimized topology holds in any practical purpose, even if no attemption is payed to wiring or other energetic costs of random links. The reason is a competition between the broadness (larger storage capacity) and roughness (slower retrieval speed) of the attraction basins.

We believe that the maximization of information respect to the topology could be a biological criteria (where non-equilibrium phenomena are relevant) to build real neural networks. We expect that the same dependence should happens for more structured networks and learning rules.

Acknowledgments Work supported by grants TIC01-572, TIN2004-07676-C01-01, BFI2003-07276, TIN2004-04363-C03-03 from MCyT, Spain.

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. Masuda, N. and Aihara, K. Biol. Cybernetics, **90**: 302 (2004)
4. Amit, D., Gutfreund, H., Sompolinsky, H.: Phys. Rev. A **35** (1987) 2293
5. Perez-Vicente, C., Amit, D.: J. Phys. A, **22** (1989) 559
6. Dominguez, D., Bolle, D.: Phys. Rev. Lett **80** (1998) 2961
7. Bolle, D., Dominguez, D., Amari, S.: Neural Networks **13** (2000) 455
8. Canning, A. and Gardner, E. *Partially Connected Models of Neural Networks*, J. Phys. A, **21**, 3275-3284, 1988
9. Kupermann, M. and Abramson, G. Phys. Rev. Lett. **86**: 2909, 2001
10. McGraw, P.N. and Menzinger, M. Phys. Rev. E **68**: 047102-1, 2003
11. Rolls, E., Treves, A., Neural Network and Brain Function. Oxford U. Press, 2004
12. Sporns, O. et al., Cognitive Sciences, **8(9)**: 418-425, 2004
13. Torres, J. et al., Neurocomputing, **58-60**: 229-234, 2004
14. Dominguez, D. et al., LNCS **3173**: 14-29, 2004
15. Li, C., Chen, G.: Phys. Rev. E **68** (2003) 52901