Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory

1994 J. Phys. A: Math. Gen. 27 4783

(http://iopscience.iop.org/0305-4470/27/14/009)

J. Phys. A: Math. Gen. 27 (1994) 4783–4789. Printed in the UK

Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory

Nicolas Brunel†

Laboratoire de Physique Statistique‡, Ecole Normale Supérieure, 24 Rue Lhomond, 75231 Paris Cedex 05, France

Received 7 April 1994

Abstract. The storage capacity in an attractor neural network with excitatory couplings is shown to depend not only on the fraction of active neurons per pattern (or coding rate), but also on the fluctuations around this value, in the thermodynamical limit. The capacity is calculated in the case of exactly the same number of active neurons in every pattern. For every coding level the capacity is increased with respect to the case of random patterns. Results are supported by numerical simulations done with an exhaustive search algorithm, and partly solve in the sparse coding limit the paradox of the discrepancy of the capacity of the Willshaw model with optimal capacity.

The question of the optimal information capacity in attractor neural networks has been widely studied since the original work of Gardner [1]. A particularly interesting case is when the information is coded in a purely excitatory synaptic matrix, and when the allowed synaptic values are discrete (for example, $J = 0, 1$), because in a biological network synaptic plasticity is believed to occur only for purely excitatory synapses, and it seems unlikely that a real synapse is able to maintain a large number of distinct stable states. In this paper I will reconsider the problem of the information capacity in the network with $J = 0, 1$, and show that the capacity depends not only on the coding rate of the memories, but also on the fluctuations around the mean number of active neurons in the memories, for every coding level. This partly solves the apparent paradox of the discrepancy of the capacity of the Willshaw model with the optimal capacity [2–4].

The network I consider is composed of N binary neurons, whose activity is denoted by V_i $(i = 1, \ldots, N)$. If $V_i = 1$, neuron i is active, while if $V_i = 0$, it is inactive. Neurons are interconnected by binary synapses ($J_{ij} = 0, 1$ for very $i \neq j$). The dynamics is discrete and neurons are updated according to the rule

$$V_i(t+1) = \Theta\left(\sum_{j \neq i} J_{ij} V_j(t) - T\right)$$

where Θ is the Heaviside function—$\Theta(x) = 1$ if $x > 0$ and zero otherwise—and T is a fixed threshold. This network will perform as an autoassociative memory if, given a set of p memories $\{\eta_i^\mu = 0, 1\}$ $(i = 1, \ldots, N, \mu = 1, \ldots, p)$ stored in the synaptic matrix, every memory is an attractor of the dynamics of the network, i.e.

$$\eta_i^\mu = \Theta\left(\sum_{j \neq i} J_{ij} \eta_j^\mu - T\right)$$

† Present address: INFN, Dipartimento di Fisica, Università di Roma 'La Sapienza', P. le Aldo Moro, Rome.
‡ Laboratoire associé au CNRS (URA 1306) et aux Universités Paris VI et VII.

0305-4470/94/144783+07$19.50

holds for every μ and i. An example of such a network is provided by the Willshaw model [2] for which the synaptic prescription

$$J_{ij} = \Theta\left(\sum_{\mu} \eta_i^{\mu}\eta_j^{\mu}\right)$$

ensures, together with an appropriate threshold, that the memories are indeed memorized, if their number p does not exceed a critical value.

A measure of the performance of the network is the ratio $\alpha = p_c/N$, where p_c is the largest number of memories that the network is able to store, in the thermodynamical limit. Another measure of performance is the information capacity of the network measured in bits per synapse, which is related to α by the equation

$$I = \alpha[-f \ln_2 f - (1-f)\ln_2(1-f)]$$

where f is the average fraction of unit bits in the memories, or coding rate. This measure is more adequate in the case of sparse coding, i.e. $f \to 0$, because in many cases $\alpha \to \infty$ in this limit, while I remains finite. In the case of binary synapses we have the upper bound $I < 1$.

This quantity has been calculated for the Willshaw matrix in the sparse coding limit $f \sim \log N/N$ and yields $I_W = \ln 2 \sim 0.69$. Another approach has been to study the space of all possible matrices, in order to derive the optimal capacity [1], given by

$$I_{opt} = \max_{\{J_{ij}\}}(I[\{J_{ij}\}]).$$

In the case of $J = 0, 1$ couplings, and in the limit $f \to 0$, one gets $I_{opt} \sim 0.29$ [3]. This is in striking disagreement with the Willshaw calculation since I_W should be smaller than the optimal capacity.

Several explanations have been proposed to account for this discrepancy [4]:

- The calculation in the case of the Willshaw model requires a vanishing signal-to-noise ratio, while the optimal capacity calculation requires that the memories are perfectly recalled, i.e. no errors are made while in the Willshaw calculation a number of errors that vanishes in the limit $N \to \infty$ is allowed.
- In the Willshaw calculation each memory has *exactly* the same number of active neurons, i.e

  $$\sum_i \eta_i^{\mu} = fN$$

  for every μ, while in the optimal capacity calculation memories are drawn according to the distribution

  $$P(\eta_i^{\mu} = \eta) = f\delta(\eta - 1) + (1-f)\delta(\eta) \tag{1}$$

  and thus there are fluctuations of the order of $\sqrt{f(1-f)N}$ around the mean number of active neurons per memory.

Subsequently the information capacity of the Willshaw model has been calculated in two cases [4]:

- The capacity for the criteria of strictly no error, with patterns with exactly the same number of active neurons, is halved with respect to the criteria of vanishing signal-to-noise ratio, and hence in this case $I_W \sim 0.35$.
- When one consider randomly drawn patterns instead of patterns with exactly the same number of active neurons, the capacity becomes $I_W \sim 0.23$. This value is now consistent with the absolute capacity.

More generally it seems interesting to investigate the effects of error tolerance or of the fluctuations in the number of active neurons per pattern on the optimal capacity. In this paper we are concerned with the second issue. The first—more difficult—issue, will be the subject of a future publication [5]. In the following we show that these fluctuations turn out to have quite a drastic effect on the storage capacity, for every coding level f. Furthermore, we show that the optimal capacity of a network with $J = 0, 1$ couplings and patterns with exactly the same number of neurons is equal to the capacity of a network with $J = -1, 1$ and random patterns. For the latter network fluctuations in the number of active neurons per pattern have no effect on the capacity, as is the case in all networks for which the mean synaptic value vanishes in the thermodynamical limit.

In the following we sketch the main steps of the calculation of the optimal capacity for memories with exactly the same number of active neurons. This calculation is performed with standard replica techniques and here we just emphasize the differences from the usual case, performed in [3]. The quantity we consider is the typical 'entropy' per synapse

$$\Omega = \frac{1}{N}\langle \ln T(\{\eta_i^\mu\})\rangle$$

where $T(\{\eta_i^\mu\})$ is the number of couplings such that the p memories $\{\eta_i^\mu\}$ are attractors of the network, and $\langle\cdot\rangle$ is an average over the distribution of patterns. We consider the two following distributions of patterns.

(A) Random patterns with coding rate f: the average is performed with the distribution

$$\langle \ln T\rangle = \mathrm{Tr}_{\eta^\mu}\left[\prod_{\mu,i}\left(f\delta(\eta_i^\mu - 1) + (1-f)\delta(\eta_i^\mu)\right)\ln T(\{\eta_i^\mu\})\right]. \tag{2}$$

The calculation of the optimal capacity in this case [3] gives the typical entropy Ω_A.

(B) Patterns with a fixed number fN of active neurons: the averaging is now done with

$$\langle \ln T\rangle = \frac{\mathrm{Tr}_{\eta^\mu}\prod_\mu \delta\left(\sum_i \eta_i^\mu - fN\right)\ln T(\{\eta_i^\mu\})}{\mathrm{Tr}_{\eta^\mu}\prod_\mu \delta\left(\sum_i \eta_i^\mu - fN\right)} \tag{3}$$

and yields the typical entropy Ω_B.

In both cases the typical entropy depends only in the thermodynamical limit on $\alpha = p/N$ and f. For every f it is a decreasing function of α. One has $\Omega(\alpha = 0) = \ln 2$ and the optimal capacity (and thus the information capacity) is obtained when $\Omega(\alpha) = 0$. At this point replica-symmetry breaking occurs [6, 3].

The calculation of the typical entropy proceeds along the following lines using the 'replica trick'

$$\frac{\langle \ln T\rangle}{N} = \lim_{n\to 0}\frac{\ln\langle T^n\rangle}{nN}.$$

The calculation of $\langle T^n\rangle$ is performed introducing n replicas $\{J_{ij}\}^\alpha$ $(\alpha = 1, \ldots, n)$

$$\langle T^n\rangle = \left\langle \frac{1}{T_0}\mathrm{Tr}_{J_i^\alpha}\prod_{\mu,\alpha}\Theta(\Delta_\mu^\alpha)\right\rangle$$

where T_0 is the number of possible couplings $T_0 = 2^N$ and

$$\Delta_\mu^\alpha = (2\eta_i^\mu - 1)\left(\frac{1}{\sqrt{N}}\sum_j J_{ij}^\alpha \eta_j^\mu - \theta\right).$$

Then one introduces integral representations for the Heaviside functions

$$\Theta(\Delta_\mu^\alpha) = \int \mathrm{d}x_\mu^\alpha \int_{\lambda_\mu^\alpha > 0} \frac{\mathrm{d}\lambda_\mu^\alpha}{2\pi} \exp\left(\mathrm{i}x_\mu^\alpha[\lambda_\mu^\alpha - \Delta_\mu^\alpha]\right) .$$

This makes possible the averaging over the distribution of the memories. The averaging requires the calculation of

$$\left\langle \prod_{\mu,j} \exp \epsilon_j^\mu \eta_j^\mu \right\rangle$$

where we have introduced

$$\epsilon_j^\mu = -\mathrm{i}\frac{2\eta_i^\mu - 1}{\sqrt{N}} \sum_\alpha J_{ij}^\alpha x_\mu^\alpha .$$

The averaging in case A (equation 2) is immediate and one obtains

$$\left\langle \prod_{\mu,i} \exp \epsilon_i^\mu \eta_i^\mu \right\rangle = \exp\left(f \sum_{\mu,i} \epsilon_i^\mu + \frac{f(1-f)}{2} \sum_{\mu,i} (\epsilon_i^\mu)^2 \right) .$$

In case B (equation 3) one has to introduce an integral representation for the delta function, then a saddle-point method yields

$$\left\langle \prod_{\mu,i} \exp \epsilon_i^\mu \eta_i^\mu \right\rangle = \exp\left(f \sum_{\mu,i} \epsilon_i^\mu + \frac{f(1-f)}{2} \left(\sum_{\mu,i} (\epsilon_i^\mu)^2 - \sum_\mu \left[\sum_i \epsilon_i^\mu \right]^2 \right) \right) .$$

Thus the difference between cases A and B is in the last term of the above equation.

The typical logarithm of the accessible volume is then obtained by a saddle-point method, after order parameters have been introduced and a replica-symmetric ansatz has been done. The result for cases A and B is

$$\Omega_{\mathrm{A,B}} = \mathrm{extr}_{q,p,Q,P,u} G_{\mathrm{A,B}}(q, p, Q, P, u)$$

where G_{A} is given by

$$G_{\mathrm{A}}^{\mathrm{rs}} = \frac{pP + qQ}{2} + \alpha \left(\sum_{\sigma=\pm 1} f_\sigma \int \mathrm{D}t \ln I_0(t) \right) + \int \mathrm{D}t \ln I_1(t)$$

where

$$I_0(t) = H\left(\frac{\sigma u + \sqrt{q}t}{\sqrt{Q-q}} \right)$$

and

$$I_1(t) = 1 + \exp\left(-\frac{p+P}{2} + \sqrt{p}t \right)$$

and G_{B} is simply related to G_{A} by

$$G_{\mathrm{B}}(q, p, Q, P, u) = G_{\mathrm{A}}(q - Q^2, p, Q(1-Q), P, u) .$$

The order parameters, when taken at their saddle-point values, have the following interpretation: q is the typical overlap between the coupling vectors in two different replicas of the system, Q is the typical connectivity, i.e. the fraction of unit couplings, p and P are their respective conjugate parameters, and u is a parameter related to the optimal threshold. The capacity in both cases is obtained when Ω vanishes. The results are as follows.

(A) For the standard coding [3]

$$\alpha_c = 0.59 .$$

In the sparse-coding limit $f \to 0$ the information content per synapse is, at $\alpha = \alpha_c$,

$$i_c \sim 0.29 . \tag{4}$$

However, there is an uncertainty in this value since the limit is very hard to get numerically [3].

(B) In this case the capacity turns out to be equal to the capacity of a network with $J = \pm 1$ at the same coding level. The capacity in this case is well known [6, 3]. For standard coding [6]

$$\alpha_c = 0.83$$

while in the sparse-coding limit [3]

$$i_c \sim 0.45 . \tag{5}$$

The two main conclusions are the following.

(i) The fluctuations of the number of active neurons per pattern thus have an important effect on the storage capacity even in the limit of a large network, for every coding level.

(ii) The estimates for the bounds on the storage capacity are now consistent with the estimates obtained for the Willshaw synaptic matrix in the sparse-coding limit [4]. These estimates are $i_c \sim 0.23$ and $i_c \sim 0.35$ for the fluctuating and non-fluctuating cases, to be compared with (4) and (5).

The calculation can be repeated for any particular constraint on the synaptic matrix. It is found that taking distribution A and B makes a difference in all cases in which the average synaptic value does not vanish. For example, when all or a finite fraction of elements of the synaptic matrix are constrained to be positive, the capacity will be improved when one uses patterns with exactly the same number of active neurons.

The dependence of the capacity in the fluctuations in the number of active neurons has been checked by numerical simulations in the case of the standard coding ($f = 0.5$). Since for discrete couplings there exists no polynomial algorithm guaranteed to converge to the optimal solution, I resort, as in [7], to enumeration of all possible couplings for small systems (up to $N = 20$), speeded up by the use of the Gray code. As already emphasized in [7], the use of binary patterns is problematic and a Gaussian distribution was used instead of the binary one. For each pattern drawn, the deviation of its total activity to the mean activity $N/2$

$$D = \left| \sum_i \eta_i^\mu - \frac{N}{2} \right|$$

was calculated. The pattern was rejected if D was greater than some fixed value $\gamma\sqrt{N}/2$, where $\sqrt{N}/2$ is the variance of the distribution of the total activities. Rejecting patterns with a large deviation D makes the variance of the resulting distribution of total activity decrease. The standard deviation σ of the distribution becomes

$$\sigma = \frac{\sqrt{N}}{2} \sqrt{1 - \frac{\gamma G(\gamma)}{\int_0^\gamma G(t)\mathrm{d}t}}$$

where $G(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Decreasing the parameter γ one gets a distribution of total activities more peaked around the mean, and in the limit $\gamma \to 0$ one gets a set of

patterns with exactly the same total activity. Thus for $\gamma = \infty$ the distribution is the usual one, while for $\gamma = 0$ the variance vanishes, $\sigma = 0$.

The dependence of the capacity on the parameter γ was investigated. Since the storage capacity of an attractor neural network is equal to the capacity of a perceptron which has to classify the same set of input patterns [1], I simulated a perceptron whose coupling vector was denoted by J. Each set of patterns Ξ was separated into two classes, Ξ^+ and Ξ^- with the same number of patterns. For every pattern η^μ the network produces the output h^μ given by

$$h^\mu = \frac{J \cdot \eta^\mu}{\sqrt{J \cdot J}}.$$

Learning the set of patterns means that the network has to separate the two classes, i.e.

$$\min_{\mu \in \Xi^-} h^\mu > \max_{\mu \in \Xi^+} h^\mu .$$

The quality of learning is measured by the parameter

$$\Delta = \min_{\mu \in \Xi^-} h^\mu - \max_{\mu \in \Xi^+} h^\mu .$$

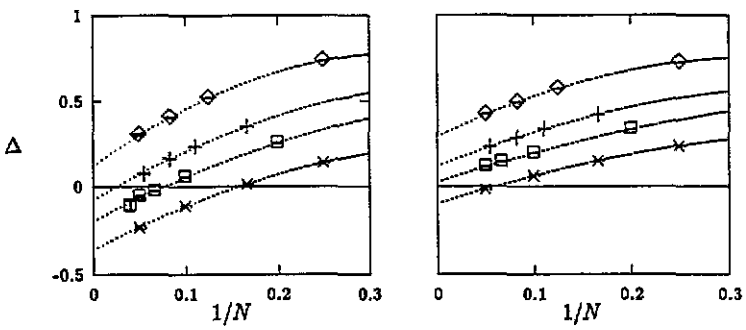

Figure 1. Parameter Δ versus $1/N$, for different values of γ and α. Left: $\gamma = \infty$. Right: $\gamma = 0.2$. For both curves, $\alpha = 0.5$ ($\diamond$), $\alpha = 0.666$. (+), $\alpha = 0.8$ ($\square$), $\alpha = 1$ ($\times$).

For each set of patterns the optimal parameter Δ_{max} was obtained by enumeration of all the couplings, i.e.

$$\Delta_{max} = \max_J \Delta .$$

For a given set of parameters N, p and γ the results were averaged over many samples of patterns (from 1000 for $N = 20$ to 50000 for $N = 5$). Then, for a given $\alpha = p/N$ and γ, we perform an $1/N$ extrapolation by best quadratic fit of the experimental values. This extrapolation of Δ_{max} at $N = \infty$ is shown in figure 1, for $\gamma = \infty$ (i.e. the usual distribution), 1 and 0.2, and $\alpha = 1, \frac{4}{5}, \frac{2}{3}, \frac{1}{2}$. Then for each value of γ we plot the obtained value of Δ_{max} versus α. A new extrapolation with a quadratic fit gives the value of α_c at the point $\Delta_{max}(\alpha) = 0$. It is shown in figure 2, for $\gamma = \infty$ and $\gamma = 0.2$. We obtain $\alpha_c \sim 0.60$ for $\gamma = \infty$ (to be compared with the analytical estimate $\alpha_c = 0.59$), $\alpha_c \sim 0.75$ for $\gamma = 1$, and $\alpha_c = 0.82$ for $\gamma = 0.2$. This shows than there is already a substantial increase in the critical capacity when the fluctuations around the mean number of active neurons are halved, and if one takes a sharp distribution for the number of active neurons ($\gamma = 0.2$) one obtains a result very close to the analytic estimate for patterns with exactly the same number of active neurons (i.e. $\gamma = 0$).

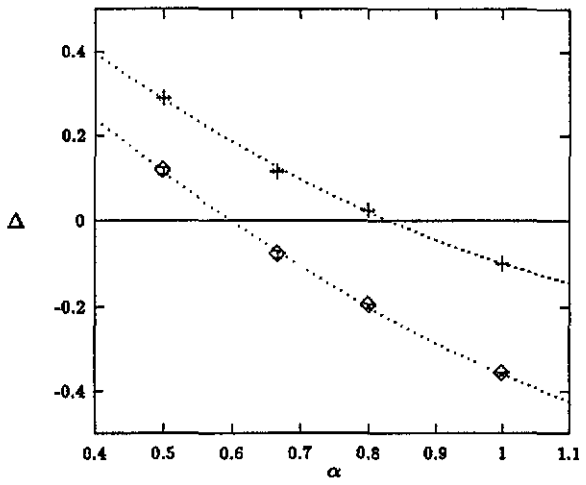

Figure 2. Δ versus α, for $\gamma = \infty$ (+), and $\gamma = 0.2$ ($\diamond$). The intersection with $\Delta = 0$ yields $\alpha_c \sim 0.6$ and 0.82, respectively.

Interestingly, a very simple scenario of learning dynamics [8] in an attractor neural network similar to the one discussed here (where the asymptotic values of the synaptic couplings are restricted to 0 and 1) leads to attractors where the number of active neurons fluctuate much less than the number of active neurons in the patterns presented to the network [9].

Acknowledgment

I am indebted to Jean-Pierre Nadal for many helpful discussions and a careful reading of the manuscript.

References

[1] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257
[2] Willshaw D, Buneman O P and Longuet-Higgins H 1969 Non-holographic associative memory *Nature* **222** 960
[3] Gutfreund H and Stein Y 1990 Capacity of neural networks with discrete synaptic couplings *J. Phys. A: Math. Gen.* **23** 2613
[4] Nadal J P 1991 Associative memory: on the (puzzling) sparse coding limit *J. Phys. A: Math. Gen.* **24** 1093
[5] Brunel N 1994 in preparation
[6] Krauth W and Mézard M 1989 Storage capacity of memory networks with binary couplings *J. Physique* **50** 3057
[7] Krauth W and Opper M 1989 Critical storage capacity of the $J = \pm 1$ neural network *J. Phys. A: Math. Gen.* **22** 1983
[8] Amit D J and Fusi S 1994 Learning in neural networks with material synapses *INFN preprint*
[9] Amit D J and Brunel N 1994 A learning attractor neural networks (in preparation)