J. Phys. A: Math. Gen. 24 (1991) 1093–1101. Printed in the UK

Associative memory: on the (puzzling) sparse coding limit

Jean-Pierre Nadal

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

Received 13 June 1990, in final form 20 September 1990

Abstract. Recent studies of the information capacity in a sparsely coded memory net has led to some contradictory results. In the Willshaw model, where the couplings are binary (0 or 1), the maximal quantity of information that can be stored is $\ln 2 \approx 0.69$ bits per synapse. On the other hand a calculation *à la* Gardner for (0, 1) couplings gives an upper bound for the maximal capacity of about 0.29 bits per synapse.

In this paper I consider two possible sources for this discrepancy. The first one is that the criterions for defining the maximal capacity are different (with or without a constraint of perfect errorless storage). The second one is a difference in the choice of the probability distribution of the random patterns used to compute this capacity.

This analysis shows in particular that for the Willshaw model the maximal information capacity is much larger when the number of active neurons is exactly the same in every stored pattern, than when it is given only in average. In addition I give an argument showing that this result may be generic, e.g., valid for any activity level and independent of the learning rule.

1. Introduction

In the evaluation of the performance of neural networks in associative tasks, many aspects of the sparse coding limit have been studied, for both attractor neural networks and simple perceptron type networks [1–11]. The regime of interest is when the number M of active neurons, in each pattern to be learned, is negligible with respect to the total number of neurons N (in the large-N limit), the coding rate $f = M/N$ being of order $\ln N/N$. Although the maximal number of such random patterns that can be stored with real valued synapses diverges with N as $(N/\ln N)^2$, the total amount of information stored per synapse tends to a finite limit, $i_m = 1/(2 \ln 2) \approx 0.721$ (bits per synapse) [12]. The general picture is that, when the coding rate goes from $\frac{1}{2}$ to 0, the maximal information capacity, in bits per synapse, decreases from 2 to $1/(2 \ln 2)$. However, in the sparse coding limit simple learning rules, of the Hebb type, give very good performance and in some cases the maximal theoretical capacity i_m can even be reached [11].

One of the most interesting rules was introduced long ago by Willshaw *et al* [1]. In the sparse coding limit it allows the storage of up to $\ln 2 \approx 0.693$ bits per synapse, which is very close to $i_m \approx 0.721$, and moreover this performance is obtained with binary couplings ($J_j = 0, 1$). Recently the maximal capacity for ± 1 couplings has been computed by the replica techniques [13], and the calculation has been extended to other choices of discrete couplings [14, 15]. In the case of (0, 1) couplings in the sparse

† Laboratoire associé au CNRS (URA 1306) et aux Universités Paris VI et Paris VII.

0305-4470/91/051093+09$03.50

coding limit, the maximal capacity is found [14] to be around 0.29, much smaller than ln 2.

Several reasons for this discrepancy are conceivable.

(i) The replica calculation is not reliable, possibly only in the sparse coding limit where the coding rate f is taken of order $\ln N/N$.

(ii) The replica analysis may be reliable, but it gives results which are expected to be true with probability one, and it might be that the Willshaw model is atypical (think of number theory: almost all real numbers are transcendental, but essentially every number one can exhibit is not).

(iii) The critical capacities computed so far with the replica method are associated with the requirement of exactly no error, whereas the maximal capacity of the Willshaw model is reached in a regime where the number of errors is non-zero, but the noise to signal ratio is vanishing in the large-N limit.

(iv) The analysis done for the Willshaw model requires, in fact, that the number of active neurons is *exactly* the same in every pattern, and the fluctuations are neglected.

In the present paper I propose to have a closer look at the Willshaw *et al* model by considering, in sections 2 and 3, the role of the implicit or explicit assumptions—evoked in cases (iii) and (iv) above—that were made in the analysis presented in [11]. Possible extension of the results to non-zero coding rate is discussed in section 4. In section 5, I show that the results obtained in sections 2 and 3 provide a solution for the paradox, and raise new questions.

2. *Zero error versus zero noise-to-signal ratio*

For simplicity, I will consider only the case of the simplest architecture that is a perceptron architecture with N inputs and only one output (figure 1). Each neuron (input or output) can be in an active ($V=1$) or quiescent ($V=0$) state. The output V is computed from the input $\{V_j, j=1,\ldots,N\}$ by the standard rule

$$2V-1=\mathrm{sgn}\left(\sum_{j=1,N} J_j V_j-\theta\right) \tag{1}$$

where the J_j are the couplings and θ is the threshold (to avoid ambiguities I chose the convention $\mathrm{sgn}(0)=1$). The task to be performed by this net is to learn a set of p associations

$$\{V_j^\mu, j=1,\ldots,N\}\rightarrow V^\mu \qquad \mu=1,\ldots,p. \tag{2}$$

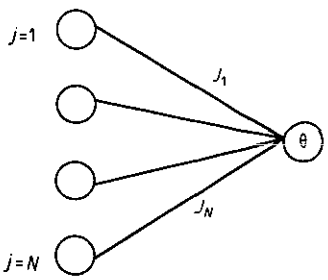

Figure 1. Perceptron type network with N input neurons and one output neuron.

The patterns are chosen at random, with f and f' as coding rates for the inputs and outputs respectively: the mean number of active neurons in each input pattern is $M = Nf$ and the mean number of patterns with an active output is $p_+ = pf'$. As in [11] I will call a 'firing' pattern a pattern with an active desired output ($V^\mu = 1$). The coding rates are low, and in particular if $f = f'$ the regime of interest is $f \approx \ln N/N$ [1]. The performance of the net is evaluated through the quantity of information stored in bits per synapse. If retrieval is perfect,

$$i = (p/N)s(f') \tag{3}$$

where $s(\cdot)$ is the mixing entropy in bits:

$$s(\phi)\ln 2 = -\phi \ln \phi - (1-\phi)\ln(1-\phi) \approx -\phi \ln \phi \tag{4}$$

for ϕ small. In case of errors, with p_1 and p_2 patterns firing among those which should fire and those which should not, respectively,

$$i = (p/N)[s((p_1+p_2)/p) - f's(p_1/p_+) - (1-f')s(p_2/(p-p_+))]. \tag{5}$$

With the Willshaw prescription the couplings are 0 or 1, and the coupling J_j takes the value 1 if for at least one of the patterns which should fire ($V^\mu = 1$) the corresponding input neuron is active ($V_j^\mu = 1$). Let me first recall the analysis done in [11]. As shown in [11] it is convenient to study the properties of the Willshaw model as a function of q, the fraction of activated synapses. After learning p patterns:

$$q = 1 - (1-ff')^p \approx 1 - \exp(pff'). \tag{6}$$

In the retrieval stage the threshold θ is set to the number of active input neurons:

$$\theta = M = Nf. \tag{7}$$

Hence every firing pattern is safely retrieved, but there might be errors on the other patterns: $p_1 = p_+$, $0 \leqslant p_2 \leqslant p - p_+ = (1-f')p$. If p_2/p is negligible compared to f', that is if the noise-to-signal ratio, p_2/pf', is zero, the information stored is equal to its maximal value (3). The critical capacity is reached when this ratio becomes of order unity. Now an input pattern with M active neurons can produce a $+1$ output only if the M couplings linked to these M neurons are non-zero. Hence the probability ζ for one error is

$$\zeta = q^M \tag{8}$$

so that the mean number of errors is

$$p_2 = q^M(1-f')p \approx q^M p. \tag{9}$$

Hence the critical capacity is given by

$$q^M/f' \approx 1 \tag{10}$$

At this threshold, using the expression (6) for q which gives $pff' = -\ln(1-q)$, the critical quantity of information (3) can be written

$$i_c = \ln q \ln(1-q)/\ln 2. \tag{11}$$

Now if instead of the criterion of vanishing noise-to-signal ratio one asks for the limit of no errors, criterion (10) is replaced by

$$q^M p = 1. \tag{12}$$

Then the information stored i_0 can be written

$$i_0 = (-\ln f'/\ln p) i_c(q). \tag{13}$$

The most interesting case is $f = f'$, where the expression (6) for q gives $-\ln f/\ln p = \frac{1}{2} + O(1/\ln p)$. That is, the information stored is simply half the information that can be stored by allowing for errors:

$$i_0 = \tfrac{1}{2} i_c(q). \tag{14}$$

The maximal quantity that can be stored in the regime of exactly no error is thus

$$\ln 2/2 \approx 0.346 \tag{15}$$

(I will comment on this value in the last section). The curves C and C_0 defined by (11) and (14) respectively are plotted in figure 2. When the number of stored patterns increases, the fraction q of activated synapses increases, and i increases according to (3), until reaching the curve C_0. Up to that point no error is made. Then if q increases further i still increases according to (3) (with neglibible corrections, and there is no discontinuity in i, or in its derivatives, when crossing C_0), until it reaches curve C. From C_0 to C the number of errors in the background increases from a finite number up to a number of order $N/\ln N$. In the error-full regime, that is if p and thus q increase further, the point (i, q) remains on this curve $i = i_c(q)$ (see [11]).

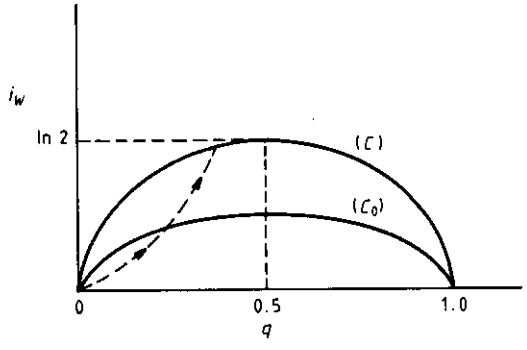

Figure 2. Willshaw prescription rule: information stored as a function of the proportion q of non-zero synaptic efficacies. Below the curve (C) is the domain accessible in the regime of vanishing noise-to-signal ratio [11]. Below the curve (C_0) is the domain accessible in the regime of strictly zero error. The dotted line give a typical trajectory that would be followed in a learning experiment where the number of stored patterns is progressively increased.

3. The role of fluctuations

The analysis of section 2 neglects the fluctuations in the number of active input neurons: they are valid for patterns chosen at random with a number of active neurons M exactly equal to Nf. However, one would expect to have, in the large N limit, the same results for a number of active neurons fixed on average only. This is not the case.

In fact, the results of section 2 have been obtained with the tacit hypothesis that the correlations between synapses can be neglected. We will see that it is these correlations which cannot be neglected if the number of active inputs fluctuates. In the following I reconsider the calculation of the probability ζ for one error, paying attention to the precise choice of the distribution for the input patterns.

There are $p_+ = pf'$ patterns with a +1 output, and only these matter for the learning stage. After learning, the number of activated synapses is $L = q_1 N$. In the large-N limit the distribution for q_1 is sharply peaked around its average value $q = 1-(1-f)^p + \simeq 1-\exp(-p_+ f)$, so we can consider that $L = qN$. For the retrieval stage, we set the threshold θ at $M = Nf$ as before, so that every firing pattern will be correctly retrieved. Now we need to evaluate the typical number p_2 of errors made on patterns which should not activate the output neuron. The probability ζ for an error on one of these patterns is the probability that at least $\theta = M$ input neurons, among the L neurons connected to the output with an activated synapse, are active in that pattern.

First I consider the case where the number of active neurons is exactly M. Then

$$\zeta = \frac{C_L^M}{C_N^M} \tag{16}$$

where

$$C_N^M = \frac{N!}{M!(N-M)!}.$$

For large L and N this gives

$$\zeta \alpha \exp fN\{\ln q - [(1-x)/x]\ln(1-x) + [(1-f)/f]\ln(1-f)\} \tag{17}$$

where

$$x = f/q. \tag{18}$$

For f and x small (that is $M \ll L$, which means that the number of stored patterns p_+ is large),

$$\zeta \alpha \exp fN[\ln q + O(f)]. \tag{19}$$

Thus the correlations are negligible if Nf^2 is small, but Nf large: this was also the condition found in [11] for the case of a Hebb model.

Now if for every pattern the activity of each input neuron is taken 1 or 0 with probability f, $(1-f)$, the probability for one error is:

$$\zeta = \sum_{n>\theta} C_L^n f^n (1-f)^{L-n}. \tag{20}$$

For large L ($L = qN$), this gives, using the Stirling formula, and omitting non-exponential prefactors,

$$\zeta = \int_x^1 dt \exp L[-t\ln(t/f) - (1-t)\ln((1-t)/(1-f))] \tag{21}$$

where $x = f/q$ as above. This integral is dominated by the smallest value of t, $t = x$, that is:

$$\zeta \alpha \exp fN\{\ln q - [(1-x)/x]\ln(1-x)/(1-f)\}. \tag{22}$$

For f small,

$$\zeta \alpha \exp fN[\ln q + 1 - q + O(f)] \tag{23}$$

instead of $\zeta\alpha \exp fN \ln q = q^M$. To try to understand better this result one may find it useful to reformulate the preceding calculation in the following way. The probability (20) is a sum of strictly decreasing terms, and is dominated by its first term:

$$\zeta = C_L^M f^M (1-f)^{L-M}.$$

One can rewrite the RHS as a product of three terms:

$$\zeta = C_N^M f^M (1-f)^{N-M} \cdot \{C_L^M / C_N^M\} \cdot (1-f)^{-(N-L)}.$$

The first term in parentheses is of order 1 in the large-N limit: it expresses the fact that in the large N limit the number of active neurons is almost surely $M = Nf$. The second term in parentheses is equal to the probability (16) for a fixed number of active inputs. The last term gives the correction, that is $1-q$ in (23). The origin of the difference between (16) and (23) is that in the present case, even though the number of active neurons is almost surely Nf, one does not care for the activities of the $N-L$ neurons for which the coupling is 0 (if ζ had been the probability for having at least M among L active neurons and *all* the other neurons inactive, the two choices of patterns distributions would have given the same result).

We can now express the condition (10) for vanishing noise-to-signal ratio, which gives here the maximal capacity:

$$i_c \ln 2 = \ln(1-q)[\ln(q) + 1 - q]. \qquad (24)$$

erratum:
q*=0.244
i1=0.264

Its maximum i_1 is reached at $q^* = 0.389\ldots$ and its value is:

$$i_1 = 0.236\ldots. \qquad (25)$$

Thus, in the case of a fluctuating number of active inputs, the general picture, as described at the end of the preceding section, is the same, with the expression for i_c being given by (24) instead of (11). The maximal capacity is strongly diminished, to about one third of the maximal capacity in the non-fluctuating case.

I thank N. Brunel for pointing out this surprising numerical error (with no consequence on the qualitative results) - JPN, mai 2012

4. Genericity of the Willshaw model

From the results of section 2 one may ask whether the difference in the capacities is a specific feature of the Willshaw model and/or of the sparse coding limit. In this section I will leave aside the Willshaw model and come back to the question of the theoretical upper limit for the storage capacity. I will show that the results obtained by the replica techniques implies that one should expect different optimal capacities for the two distributions of patterns whatever the coding rate f.

If the patterns were chosen at random but with exactly the same number $M = Nf$ of active neurons for all the patterns, then one would get the same maximal capacity for 0, 1 couplings and for ± 1 couplings. Indeed, consider the p inequalities which have to be solved in the former case:

$$(2V^\mu - 1)\left(\sum_{j=1,N} J_j V_j^\mu - \theta\right) > 0 \qquad \mu = 1, \ldots, p. \qquad (26)$$

If we make the change of variable

$$J_j \to W_j = 2J_j - 1 = \pm 1 \qquad (27)$$

the inequalities (26) can be rewritten

$$(2V^{\mu}-1)\left(\sum_{j=1,N} W_j V_j^{\mu}-\theta'\right)>0 \qquad \mu=1,\ldots,p \tag{28}$$

where the new threshold θ' does not depend on the pattern μ:

$$\theta'=2\theta-\sum_{j=1,N} V_j^{\mu}=2\theta-M. \tag{29}$$

Note that this is true for any value of the coding rate f.

In the replica calculations [12-14], the learned patterns are randomly chosen, with the probability distribution

$$P(\{V_j^{\mu}, j=1,\ldots,N\})=\prod_{j=1}^{N}[f\delta_{V_j^{\mu},1}+(1-f)\delta_{V_j^{\mu},0}] \tag{30}$$

and thus the number of active neurons is fixed only on average. The computed optimal capacities for ± 1 couplings and 0, 1 couplings are indeed different: for $f=f'=\frac{1}{2}$, $\alpha_c=0.83$ in the first case [13] and 0.59 in the second case [14].

Could one compute directly the optimal capacity for patterns with exactly the same number of active neurons? This is not clear since, *a priori*, the replica method cannot distinguish between the two types of distributions† a constraint such as

$$\delta\left(\sum_{j=1}^{N} V_j^{\mu}-M\right) \tag{31}$$

would be taken into account by introducing a field h:

$$\exp h\left(\sum_{j=1}^{N} V_j^{\mu}-M\right)=\mathrm{e}^{-hM}\prod_{j=1}^{N}[\mathrm{e}^{h}\delta_{V_j^{\mu},1}+\delta_{V_j^{\mu},0}] \tag{32}$$

and one would be back to the same formulation as for the usual distribution (30). This equivalence of the two distributions in the large-N limit is similar to the equivalence in statistical mechanics between the micro-canonical and the canonical free energies in the thermodynamic limit.

5. Discussion

I have considered two possible sources of discrepancy between the maximal capacity estimated from the computation *à la* Gardner [16] and the maximal capacity of the Willshaw model. The maximal capacity for 0, 1 couplings (and an adjustable threshold) is, in the sparse coding limit, less or equal to 0.29 [16]. This is still below the maximal capacity of the Willshaw model in the first case I have considered: the capacity is $i_0=0.346$ if the criterion is no error at all. It suggests that the maximal capacity might be reached in the error regime. For unbiased patterns ($f=f'=\frac{1}{2}$), this can be shown to be true [17]. For biased patterns, a replica calculation of the minimal number of errors has been done [16]. This can be generalized to the computation of the information capacity for arbitrary coding rates, and it is presently under study [18]; preliminary results indicate that, although for finite coding rates the maximum is indeed reached

† I thank M Mézard for this remark.

in the error regime, in the sparse coding limit the maximum is the one obtained in the error-free regime.

The value 0.29 is, on the other hand, above the value $i_1 = 0.236$ obtained in the Willshaw model when one allows for fluctuations in the number of active neurons. Note, however, that from the replica calculation at several values of f the numerical extrapolation at $f = 0$ is very hard to get [14], (the value 0.29 is only an upper bound) and thus we do not know exactly how 0.236 compares with the theoretical limit.

It was remarkable that the ln 2 capacity of the Willshaw model was so close to the optimal capacity $1/(2 \ln 2)$ as computed by E Gardner for continuous couplings. I have shown that this comparison was in fact inadequate: ln 2 should be compared with the (unknown) capacity for patterns with exactly the same number of active neurons. The capacity of the Willshaw model for a fluctuating number of active neurons remains however remarkably good, when compared to the relevant upper limit, the one for 0, 1 couplings.

As shown in section 4, the effects of the choice of the patterns distribution on the capacity is likely to be generic. Clearly it would be interesting to find a general method to compute directly the maximal capacity for patterns with exactly the same number of active neurons.

The present analysis has been made for a perceptron architecture. It would be interesting to study the case of an attractor neural networks: when errors are allowed, the capacity might be different due to the dynamics which can emplify the errors.

Acknowledgments

I thank Hanoch Gutfreund, Marc Mézard and Gérard Toulouse for many fruitful discussions. I thank Daniel Amit and Gérard Toulouse for a critical reading of the manuscript. This work has been supported by the European initiative BRAIN, contract number ST2J-0422-C(EDB).

References

[1] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Non-holographic associative memory* *Nature* **222** 960

[2] Palm G 1980 On associative memory *Biol. Cybern.* **36** 19

[3] Palm G 1987 Technical comment on computing with neural networks *Science* **235** 1227

[4] Palm G 1988 On the asymptotic information storage capacity of neural networks *Neural Computers* ed R Eckmiller and C von der Malsburg (Berlin: Springer) p 271

[5] Horner H 1989 Neural networks with low levels of activity: Ising versus McCulloch-Pitts neurons *Z. Phys.* B **75** 133

[6] Buhmann J, Divko R and Schulten K 1989 On sparsely coded associative memories *Neural Networks from Models to Applications* ed L Personnaz and G Dreyfus (Paris: IDSET) p 360; *Phys. Rev.* A **39** 2689-92

[7] Tsodyks M V and Feigel'man M V 1988 The enhanced storage capacity in neural networks with low activity level *Europhys. Lett.* **6** 101

[8] Tsodyks M V 1988 *Associative memory in asymmetric diluted network with low level of activity* *Europhys. Lett.* **7** 203

[9] Perez Vicente C J and Amit D J 1989 Optimised network for sparsely coded patterns *J. Phys. A: Math. Gen.* **22** 559

[10] Amari S 1989 Characteristics of sparsely encoded associative memory *Neural Networks* **2** 451

[11] Nadal J-P and Toulouse G T 1990 Information storage in sparsely coded memory nets *Network* 1 61

[12] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257
[13] Krauth W and Mézard M 1989 Storage capacity of memory networks with binary couplings *J. Physique* **50** 3057
[14] Gutfreund H and Stein Y 1990 Capacity of neural networks with discrete synaptic couplings *J. Phys. A: Math. Gen.* **23** 2613
[15] Bouten M, Komoda A and Serneels R 1990 Storage capacity of a diluted neural network with Ising couplings *Preprint*
[16] Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271
[17] Toulouse G 1990 unpublished
[18] Brunel N and Nadal J-P 1990, in preparation