Network 1 (1990) 61–74. Printed in the UK

Information storage in sparsely coded memory nets

Jean-Pierre Nadal and Gérard Toulouse

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

Received 7 June 1989

Abstract. We study simple, feedforward, neural networks for pattern storage and retrieval, with information theory criteria. Two Hebbian learning rules are considered, with emphasis on sparsely coded patterns. We address the question: under which conditions is the optimal information storage reached in the error-full regime?

For the model introduced some time ago by Willshaw, Buneman and Longuet-Higgins, the information stored goes through a maximum, which may be found within the error-less or the error-full regimes according to the value of the coding rate. However, it eventually vanishes as learning goes on and more patterns are stored.

For the original Hebb learning rule, where reinforcement occurs whenever both input and output neurons are active, the information stored reaches a stationary value, $1/(\pi \ln 2)$, when the net is overloaded beyond its threshold for errors. If the coding rate f' of the output pattern is small enough, the information storage goes through a maximum, which saturates the Gardner bound, $1/(2 \ln 2)$. An interpolation between dense and sparse coding limits is also discussed.

1. Introduction

At present, in the study of neural networks, either device oriented or biology oriented, several reasons concur to attract attention to the case of sparse coding (i.e. small signal-to-background ratio in the messages to be stored) [1–11]. Here are some of these reasons.

(i) The storage capacity (measured as the number of patterns, or messages, that can be stored without errors) becomes much larger than in the case of dense coding, heretofore also called 'standard coding' (signal-to-background ratio equal to 1, i.e. equal number of active and inactive neurons); in a confrontation with non-neural comparison algorithms, neural nets are at their best with sparsely coded patterns, because of this large capacity.

(ii) Biological recordings, in particular from the mammal cortex, vastly support a picture of 'low activity', namely at any instant of time only a small fraction of the neurons are active.

(iii) In the limit of sparse coding, there exist explicit learning algorithms (one-shot learning) which seem to be excellent, in the sense of coming close to the bounds for optimal storage; this raises one question and one hope: are these learning rules truly optimal? And if so, it may be that sparse coding offers an ideal way to investigate harder problems, such as storage with errors, which could not yet be approached with satisfactory generality in the case of standard coding.

†Laboratoire associé au CNRS (URA 0731) et à l'Université Paris VI.

0954-898X/90/010061 + 14$03.50

As is often the case in neural network science, the history of studies on sparse coding has followed meandering paths that are both puzzling and instructive. In 1969, Willshaw, Buneman and Longuet-Higgins published in *Nature* an elegant note [1], where they defined a learning rule that allowed one to store, with negligible error rate, an information content equal to ln 2 (in bits per adaptive element, i.e. per synapse). Since the synaptic efficacies were restricted to take only two values, 0 and 1, the maximal information content that could ideally be stored was 1 bit per synapse. For about 15 years, the Willshaw model remained somewhat unique, as a solvable model, in the field of memory nets, and the figure of ln 2 surfaced in isolation, the special conditions under which it held true being progressively forgotten.

With the advent of the Hopfield model [12] in 1982, a new group of scientists approached the quantitative study of memory nets from a different side. 'For simplicity', the patterns to be stored were taken as random and uncorrelated (leading to roughly equal numbers of active and inactive neurons in each pattern, i.e. standard coding), and the learning rule was the Hebb rule as modified by Hopfield (leading to synapses taking a wide range of positive and negative values). Since there is no simple bound on the information that can be laid in a continuously variable synapse, as distinct from a discrete synapse, and also because of a special feature of the standard Hopfield model (whereby almost error-free storage can be achieved, until a collapse is reached), the emphasis has been driven towards the objective of maximal storage capacity (optimal number of patterns that can be retrieved without errors), rather than on the goal of maximal information storage.

This orientation was first amplified [13–16], then reversed, by the new strategy introduced by Gardner in 1987. She showed that it was possible to compute absolute figures for the maximal error-free storage capacity, independent of any particular learning prescription. That program was achieved for continuous synapses, first in the case of standard coding, then for arbitrary coding rates. This second step brought attention back to the Willshaw model, and to the notion of information storage: indeed, for sparse coding, the number of stored patterns clearly does not suffice as a measure of storage quality, since the information content of each pattern decreases as the coding rate diminishes.

Let us define, for future purposes, the coding rate f as the fraction of active neurons in a pattern (ratio of signal length to total length). It was found that the maximal information (measured in bits per synapse) in error-free storage decreased from a value of 2, for standard coding ($f = \frac{1}{2}$), to a value of $1/(2 \ln 2)$, for sparse coding (f small). This result came as a blow for those who had come to believe (for no good reason) that the Willshaw value of ln 2 was a general upper bound. But, more positively, it was intriguing to observe how much ln 2 comes close to $1/(2 \ln 2)$. This showed that, for sparse coding, the restriction to discrete synapses has much less drastic effects than for standard coding (where this restriction leads to a drop from a value of 2 to a value estimated around 0.8 [17]). Furthermore, it led to speculation that perhaps the upper bound, for sparse coding *and* discrete synapses, is ln 2, in which case the Willshaw rule would indeed be optimal in its category.

Additional questions ensue naturally. Is there a specific learning rule that saturates the bound of $1/(2 \ln 2)$ (thus a kind of 'companion model', for continuous synapses, of the discrete-synapse Willshaw model)? And, insofar as the emphasis has been shifted from capacity to information content, why restrict attention to error-free storage? Perhaps it is not so bad, after all, to overload the network, a little or a lot—the increase in the number of patterns compensating for the decrease in the information per pattern. Furthermore,

as a speculative motivation, perhaps imperfect storage is a mode in which brains, or parts of brains, do manage to function.

The main purpose of this paper is to derive results on information storage for hetero-associative nets, with Hebbian-type learning rules. The discussion below pertains to the simplest feedforward structure that is naturally hetero-associative: in its most general form, the network is an input–output associative net (figure 1), with N input neurons and N' output neurons. An input pattern consists of a string of N bits, with M bits equal to 1, and the others equal to 0, if the coding rate is $f = M/N$. The associated output pattern is a string of N' bits, with coding rate $f' = M'/N'$. During a learning session, the network will try to learn P input–output pairs, i.e. associations of a given input pattern with its desired output pattern. After the learning session the performances are examined; it is checked whether a given input elicits the desired output (error-free storage) or a noisy version of it (error-full storage). Processing of noisy inputs, that is the study of the basins of attraction, is beyond the scope of this study, and by 'association' we mean here simply input–output mapping. In the conclusions, we will focus on the case of networks with similar input and output characteristics: $N' = N$ (square net), and a uniform coding rate $f' = f$. (Note that the Gardner results apply for this case, as well as for self-coupled nets.) However, for the benefit of theoretical transparency, the coding rates f and f' will often be kept as independent variables in the intermediate steps. Moreover, and without loss of generality, we will also consider—in particular for numerical investigations—a simple two-layer single-output perceptron (one output neuron and N adaptive elements). In that geometry, averaged properties over the N' neurons are replaced by probabilities for the single-output neuron.

Let us make a short comment on the choice of net architecture. Another simple structure is the self-coupled net, that feedbacks on itself: starting from a given initial configuration, the activity state of the network evolves freely, until it converges toward an attractor (Hopfield net). In such a case, the storage is auto-associative. A second obvious difference between the two structures is in the effect of retrieval errors; in a self-coupled net, the errors may become amplified during successive iterations of the network state. These points are elaborated later. At this stage, it suffices to say that for a discussion of storage without errors the same considerations apply for both structures.

The paper is organised as follows. In §2, we re-examine the Willshaw model. In §3, we define the information content of a message containing errors, and this allows us to analyse, in §4, the Willshaw model in the error-full regime. Then in §5 we consider the model with the original Hebb learning rule and study it in the sparse coding limit. In §6,

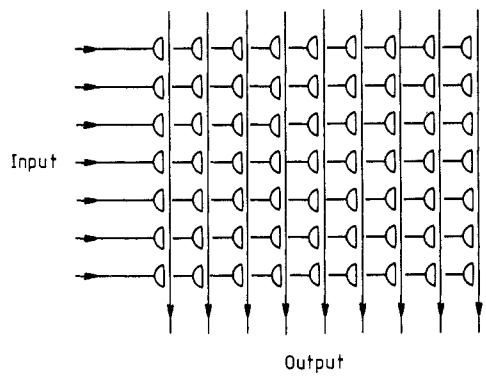

Figure 1. Hetero-associative net with $N = 7$ input neurons and $N' = 9$ output neurons. The synapses at the vertices are the adaptive elements.

we analyse this model in the error-full regime. In §7 we show that there is a natural interpolation from dense to sparse coding. Finally, in the concluding section we draw some perspectives.

2. Reminder on the Willshaw model

As explained above, the network is an input–output associative net (figure 1), with N input neurons and N' output neurons [1]. The coding rate of the input patterns is $f = M/N$, and the coding rate of the associated output patterns is $f' = M'/N'$. In order to learn P input–output pairs, the Willshaw learning rule prescribes that if, in any pair, the input neuron i and the output neuron j are simultaneously active, then the synapse of i on j is made equal to 1. At the end of a learning session, the set of NN' synapses will have acquired a distribution of values 0 and 1. (Note that for a self-coupled net the Willshaw rule would construct a symmetric net, with the synapse of i on j equal to the synapse of j on i.) One can already guess that the optimal information storage will obtain for a balanced distribution, when the proportion of activated synapses is $\frac{1}{2}$.

After learning P patterns, according to the Willshaw rule, the fraction q of activated synapses is given by

$$1 - q = (1 - ff')^P = \exp P \ln(1 - ff'). \tag{1}$$

We note that in the following q will appear as a natural parameter (instead of P), and for ff' small one has $Pff' = -\ln(1 - q)$. In the retrieval stage, the threshold of the output neuron is set at $M = Nf$. The learning rule then guarantees that each 'fire' input (input pattern learnt with unit output) is safely retrieved. An error will appear when a 'fail' input (input pattern learnt with null output) elicits a positive response. On average, the error rate is thus q^M, and the noise-to-signal ratio is $((1 - f')/f')q^M$. In the absence of errors, the total information stored (in bits) is

$$\begin{aligned} I_W \ln 2 &= P[-f' \ln f' - (1 - f')\ln(1 - f')] \\ &= (N/M)[-\ln(1 - q)](-\ln f') \qquad \text{if } f' \ll 1. \end{aligned} \tag{2}$$

The errors may be neglected if the noise-over-signal ratio is small, implying:

$$M > (\ln f')/(\ln q). \tag{3}$$

Thus, the maximal information (in bits per synapse) that can be stored in the error-free regime takes the simple expression

$$\hat{i}_W \ln 2 = \ln q \ln(1 - q) \tag{4}$$

which is clearly maximal for $q = \frac{1}{2}$ (equal proportion of activated and unactivated synapses), $\hat{i}_W = \ln 2$.

For a square net, with coding rates $f' = f$, condition (3), taken at the optimal point $q = \frac{1}{2}$, imposes

$$M \simeq a(-\ln f) \qquad a = 1/\ln 2$$

and thus $M \simeq a \ln N$, $f = M/N \simeq a(\ln N)/N$. The coding rate is size dependent and becomes sparser as N grows. The optimal capacity is derived from

$$P_c f^2 = \ln 2$$

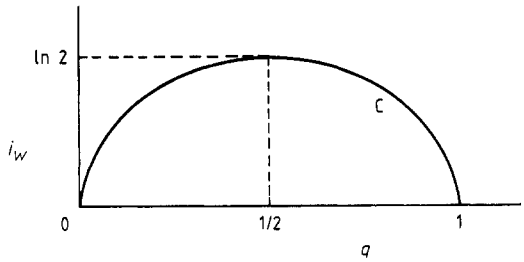

Figure 2. Willshaw model. Shown is the information stored versus the proportion of activated synapses. Below the curve C is the domain accessible in the error-free regime.

and thus is obtained as

$$P_c = (\ln 2)^3 N^2/(\ln N)^2.$$

The number of synapses is N^2, and the total information stored is $N^2 \ln 2$.

Coming back to the general case, we have obtained that, in a plot of i against q, $0 < q < 1$, the curve C (figure 2) defined by (4) delimits the domain accessible in the error-free regime. During a learning session, as P increases, q increases according to (1) and i_W according to (2). Thus the work point, in the (i, q) plane, rises monotonically from the origin, until it reaches curve C. If $M < M^* = -\ln f'/\ln 2$, the contact occurs for $q < \frac{1}{2}$, before the optimum point. We can ask the following question: would not it be advantageous in this case (supposing the value of M is not adjustable) to pursue the learning, despite the emergence of errors, until q reaches $\frac{1}{2}$? In order to be able to address such questions quantitatively, we need a general expression for the information in a net with imperfect storage (i.e. including the information loss due to retrieval errors).

3. Information content of a noisy message

In the case of standard coding, the information content of a noisy message is a simple function of the total number of errors (Hamming distance). No longer is this true for sparse coding, because an error in the signal is not equivalent to an error in the background. This simple fact has non-trivial consequences.

In the absence of errors, the information content of an N-bit message, including M bits of signal, is $I_0 \ln 2 = \ln C_N^M$, where C_N^M is a binomial coefficient.

In the presence of noise, let us define the number of unit bits as M_1, $0 < M_1 < M$, in the signal, and M_2, $0 < M_2 < N - M$, in the background. This implies $M - M_1$ errors on the signal, and M_2 errors on the background (figure 3). The expression for the information content is then

$$I \ln 2 = \ln C_N^{M_1+M_2} - \ln C_M^{M_1} - \ln C_{N-M}^{M_2} \tag{5}$$

$$= \ln C_N^M - \ln C_{M_1+M_2}^{M_1} - \ln C_{N-M_1-M_2}^{M-M_1}. \tag{6}$$

The two negative terms, in the last formula, correspond to the information loss due to errors, namely the information needed to extract the signal from the unit bits, and from the null bits. For N large enough, posing

$$f = M/N \tag{7}$$

one has

$$\ln C_N^M = -N[f \ln f + (1-f)\ln(1-f)]. \tag{8}$$

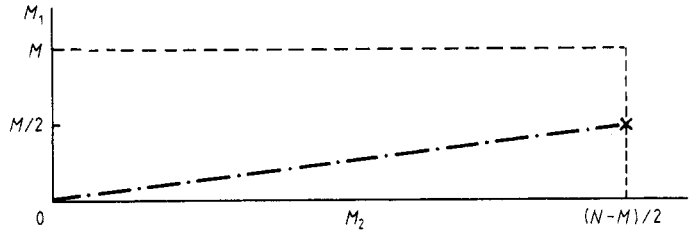

Figure 3. Error plot for a distorted message of length N, with M-bit signal and $(N - M)$-bit background. The number of unit bits, M_1 in the signal and M_2 in the background, are plotted vertically and horizontally respectively. The total number of errors is $M - M_1 + M_2$. An error-less message is represented by the point $(M_1 = M, M_2 = 0)$. The information content vanishes on the chain line.

Thus formula (5) shows immediately that the information vanishes if the activity level is the same on the signal and on the background (figure 3): signal and background become indistinguishable. That means that M errors in the signal may cost as much as $N - M$ errors in the background. Note the duality:

$$\begin{aligned} M_1 &\to M - M_1 \\ M_2 &\to N - M - M_2 \\ I &\to I \end{aligned} \tag{9}$$

with its fixed point $M_1 = M/2$, $M_2 = (N - M)/2$. Expansions around this point of maximal noise will be useful. Posing $M_1 = (1 + u)M/2$ and $M_2 = (1 + v)(N - M)/2$, the expression for the information density, in bits per message element, becomes

$$\begin{aligned} i \ln 2 = &-\{[1 + (u - v)f + v]\ln[1 + (u - v)f + v] \\ &+ [1 - (u - v)f - v]\ln[1 - (u - v)f - v]\}/2 \\ &+ f\{(1 + u)\ln(1 + u) + (1 - u)\ln(1 - u)\}/2 \\ &+ (1 - f)\{(1 + v)\ln(1 + v) + (1 - v)\ln(1 - v)\}/2. \end{aligned} \tag{10}$$

The total information of a message of length N is an extensive quantity: $I = Ni$, where i is the information density. This is the property that allows us to consider networks with a single output neuron, without loss of generality. In the forthcoming analysis we shall feel free to use extensive or intensive variables, whenever convenient.

4. Error-full regime in the Willshaw model

The Willshaw model guarantees perfect retrieval of the signal, so $M_1 = M'$. Errors occur when an input which should not trigger activity ('fail' input) reaches the threshold. Formula (6) applied to the output pattern can be simplified into

$$I \ln 2 = \ln C_{N'}^{M'} - \ln C_{M' + M_2}^{M'} \tag{11}$$

where M_2 is the number of background errors:

$$i \ln 2 = -(1 - f')\ln(1 - f') - (f' + M_2/N')\ln(f' + M_2/N') + (M_2/N')\ln(M_2/N').$$

In the regime with many errors, $M_2 > M' = N'f'$, and for f' small, the expression simplifies into

$$i \ln 2 \simeq -f' \ln(M_2/N').$$

The total information stored in a network is the product of the number of stored patterns and the information content per output pattern. So we introduce the error rate: $M_2/N' = (1 - f')q^M \simeq q^M$, and we multiply by the number of patterns. The result, in bits per synapse, is

$$i_W \ln 2 \simeq \ln q \ln(1 - q)$$

which is the same as expression (4)! This means that, once the curve C (figure 2) is reached, one enters the error-full regime, and the work point (i, q) *remains* on that curve as q increases. Since, for f small enough, the contact with the curve occurs before the optimal point, this point $(q = \frac{1}{2})$ will be eventually reached if P increases further. However, beyond that value, that is if P increases further, the total information stored decreases monotonously (figure 2), the increase in the number of learnt patterns being unable to compensate for the deterioration in storage quality.

We have thus proved that the answer to the question raised in the end of § 2 is affirmative. If the coding rate, f, is small enough, it is advantageous, as measured by the total information stored, to overload the memory network beyond the error-less limit, and to enter the error-full regime.

5. Model with the original Hebb rule

Let us introduce one modification in the learning rule of the Willshaw model, everything else remaining equal. During the learning session, a synapse gets a unit increment of its efficacy, whenever the corresponding input and output neurons are simultaneously active. (A synapse is no longer a binary element, restricted to the two values 0 or 1.) This learning rule follows directly Hebb's suggestion, as written in his book [18] 40 years ago. It differs from later modifications, such as the one considered by Hopfield, that included inhibitory synapses and mechanisms for efficacy decreases.

With this definition, the Willshaw learning rule can be redefined as a clipped version of the original Hebb rule. The two models are similar, and amenable to simple analytic treatments. They share the virtues of using explicit and local learning rules. In the limit of sparse coding, the model with the original Hebb rule allows for excellent information storage (it saturates the optimal value for unrestricted synaptic efficacies), and its performance is remarkably robust to overloading errors, as shown below.

An essential difference between the two models is in the threshold used in the retrieval process. A constant in the Willshaw model, the threshold is an adjustable parameter in the second model, and it may be modulated to improve discrimination between the two classes ('fire' and 'fail') of inputs.

Let us now give the analysis of this model, in the sparse coding limit. For the benefit of simplicity, we shall assume both f and f' (the input and output coding rates) small, so that for instance

$$Nf^2 \ll 1.$$

Then the distribution of values n, for one synapse, after learning P patterns, obeys a Poisson law:

$$Q(n) = e^{-x}(x)^n/n! \tag{12}$$

with

$$x = Pff'. \tag{13}$$

The mean and the variance of this distribution are

$$\langle n \rangle = x$$

$$\langle n^2 \rangle_c = x.$$

An input pattern contains a signal of $M = fN$ unit bits, so the distribution of field values (summed input received by the output neuron) will also be a Poisson law like (12) with

$$Q(n) = e^{-Mx}(Mx)^n/n! \qquad \langle n \rangle = \langle n^2 \rangle_c = Mx \tag{14}$$

for 'fail' input patterns, whereas for 'fire' input patterns, the distribution is shifted towards higher values:

$$\langle n \rangle = M(1 + x) \tag{14'}$$

as a benefit of the learning rule prescription (figure 4). These arguments are valid, provided correlations between synaptic values can be neglected, which is true for $Nf^2 \ll 1$ (see the appendix for details). As x becomes large, these Poisson laws turn into Gauss laws:

$$Q(n) \approx \frac{1}{\sqrt{2\pi\langle n^2 \rangle_c}} \exp - \frac{(n - \langle n \rangle)^2}{2\langle n^2 \rangle_c} . \tag{15}$$

Suppose now the threshold is fixed at $M\lambda$. An estimate of the error probability for 'fail' patterns will be given by (14), evaluated at $n = M\lambda$, provided x is small enough ($x \ll M$).

The discussion of §2 can then be repeated, with expression (2) becoming

$$I_{0h} \ln 2 = (N/M)(x)(-\ln f') \tag{16}$$

This expression, for the information stored, holds provided

$$M[x - \lambda \ln(ex/\lambda)] > (-\ln f'). \tag{17}$$

Therefore, in the error-less regime, the maximal information stored (in bits per synapse) is

$$\hat{\imath}_{0h} \ln 2 = x[x - \lambda \ln(ex/\lambda)]. \tag{18}$$

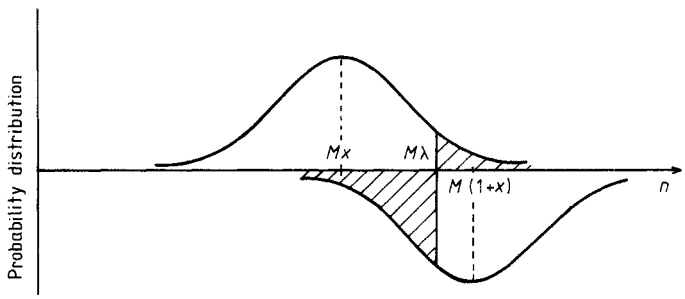

Figure 4. Original Hebb rule. Shown are the shapes of the two field distributions. Upwards: the distribution for 'fail' input patterns, centred at Mx. Downwards: the distribution for 'fire' patterns, centred at $M(1 + x)$. The position of the threshold, at λM, determines the error rates (hatched areas).

This is to be compared with expression (4), for the Willshaw model, that reads, in a similar variable x:

$$\hat{\imath}_W \ln 2 = -x \ln(1 - e^{-x}). \tag{4'}$$

One essential difference between the two expressions comes from the presence of the parameter λ in (18). To choose a value for λ is to fix the threshold. Let us take $\lambda = x + \delta$, with $0 < \delta < 1$, so that the threshold remains between the averages of the two field distributions (14) and (14')—this is a reasonable choice for proper discrimination. Then, condition (17) reads, if x is large,

$$M\delta^2/2x > (-\ln f') \tag{19}$$

and, since $x = Pff'$, the capacity in the error-less regime is bounded by

$$P < P_c = N\delta^2/(-2f' \ln f'). \tag{20}$$

The information per synapse, as given by (18), is a monotonically growing function of x; for x large, one obtains

$$\hat{\imath}_{Oh} \ln 2 \to \delta^2/2. \tag{21}$$

Thus, for $\delta = 1$, the absolute maximal value—as found by Gardner for error-less storage with unrestricted synapses—is actually reached! There is one objection, however, that requires clarification; if δ is strictly equal to 1, the threshold falls in the middle of distribution (14'), and that means that half of the signal inputs will elicit a wrong output, $M_1 = M'/2$, thus yielding an information loss that is not properly accounted for in (19), and which turns out to introduce a drop by a factor $\frac{1}{2}$. The solution out of this difficulty is to make δ a function of x, that tends to 1, as x becomes large, but sufficiently slowly to keep the error rate on signal patterns at a negligible level.

Remark. The centres of the two field distributions (for 'fail' and 'fire' inputs, figure 4) are at Mx and $M(1 + x)$, respectively, and the width for both distributions is $\sqrt{(Mx)}$. One can therefore distinguish two regimes, according to whether the separation M is larger or smaller than the width $\sqrt{(Mx)}$. The previous analysis pertains to the first case, and the error-less regime. In the second limit, $x \gg M$, the number of errors is large, the distributions take a Gaussian shape, and the calculations are also easy, as shown below. The behaviour in the intermediate domain, $x \simeq M$, is more complex but it turns out that the limits, taken from both sides, join smoothly in the crossover region.

As a conclusion of this section, we have found that the model with the original Hebb rule, in the error-less regime, does saturate the Gardner limit for information storage in memory nets. Thus this model appears as a 'champion' for unrestricted synapses, and a natural 'companion' of the Willshaw model, as far as sparse coding is concerned. This comes as an encouragement to extend the analysis to the error-full regime.

6. Error-full regime for the Hebb rule

Around the limit of maximal noise, it is convenient to use formula (10). With the threshold at λM, and the Gaussian form for the two distributions of input fields, one has

$$\begin{aligned} u &= \mathrm{erf}[M(1 + x - \lambda)/\sqrt{(2Mx)}] \\ v &= -\mathrm{erf}[M(\lambda - x)/\sqrt{(2Mx)}]. \end{aligned} \tag{22}$$

For u and v small, expression (10) simplifies to

$$I \ln 2 = \frac{f(1-f)}{2}(u-v)^2 - \frac{1}{12}[uf + v(1-f)]^4 + \frac{f}{12}u^4 + \frac{1-f}{12}v^4. \quad (23)$$

At lowest order, the information content of an output pattern is independent of the threshold position, and it reaches the asymptotic value

$$i \ln 2 = f'(1-f')M/\pi x. \quad (24)$$

Therefore the information stored, in bits per synapse, does not vanish asymptotically in the error-full regime; rather, it tends to

$$i \ln 2 \simeq 1/\pi. \quad (25)$$

This result is remarkable. Note the difference with the Willshaw model, where the total information stored eventually collapses, as a result of overloading. This second model exhibits a quality of robustness that was not foreseen. (Of course, the existence of an adjustable parameter, the threshold, is an important part of the explanation.)

Computations of formula (23) at fourth order are straightforward though slightly cumbersome, and allow one to address the following question: is the asymptotic limit reached from above or from below, during a learning session? The answer is that the limit can be approached from above, at fixed δ, but only if the coding rate is sparse enough, $f' < f_c$, with

$$f_c = (12\pi - 8 - 3\pi^2)/(12\pi - 8) \simeq 0.003. \quad (26)$$

This finding is in good agreement with the prediction that, during a learning session, the information stored goes through a maximum before reaching the asymptotic value (25), for sparse enough coding. The value obtained for the optimal δ (that fixes the threshold position) is slightly below 1, also in good accord with our qualitative discussion in the previous section.

To illustrate this, we have performed numerical simulations whose results are displayed in figure 5. These simulations have been done for a simple perceptron ($N' = 1$) with patterns having exactly $M = 50$ active neurons among the $N = 1000$ neurons ($f = 0.05$), and for several values of f'. For one single output neuron one can measure the

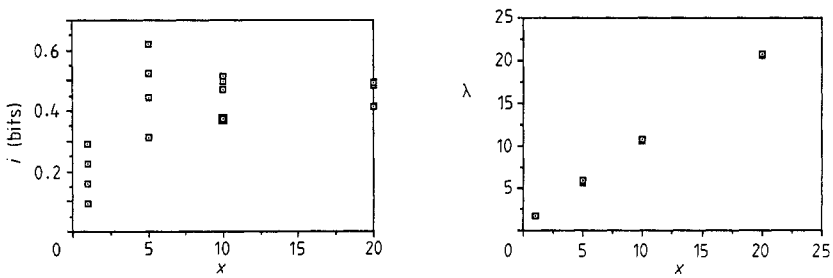

Figure 5. Original Hebb rule: simulations with $N = 1000$, $M = 50$ and for several values of f'. (*a*) Plotted is the quantity of information (in bits) i as a function of x, optimised with respect to threshold position. For each value of x the four points, from bottom to top, correspond to the values of $f' = 10^{-1}, 10^{-2}, 10^{-3}$, and 10^{-4}. (*b*) Threshold value λ as a function of x (all values of f' together).

information content as:

$$I \ln 2 = \ln C_P^{P_1+P_2} - \ln C_{P_+}^{P_1} - \ln C_{P-P_+}^{P_2} \tag{5'}$$

where $P_+ = f'P$ is the number of patterns with an active output, and P_1 and P_2 are respectively the number of patterns which fire among the P_+ (respectively $P - P_+$) patterns. Note that since the values of the couplings depend only on the patterns with an active output, one has to store in the computer only those patterns, and this allows us to deal with low output activity rates. The main qualitative effect is observed (figure 5(*a*)): for large f', i increases continuously up to the asymptotic value, whereas for small enough f', i goes through a maximum before decreasing towards this same asymptotic value. This quantity i was obtained, for each value of P, by looking for the value $\lambda = x + \delta$ of the threshold which maximises the information quantity. In figure 5(*b*) is shown the value of λ as a function of $x = Pff'$: clearly λ does not depend on f' but only on x, and is close to $1 + x$.

7. A Hebbian interpolation from dense to sparse coding

For dense coding ($f = \frac{1}{2}$, unbiased patterns), many results have been derived [10], with the generalised Hebb learning rule as in the Hopfield scheme, using $+1$ or -1 neural variables (rather than 1 or 0 variables):

$$\xi_j^\mu = \pm 1 \qquad \sigma_i^\mu = \pm 1$$

$$J_{ij} = \Sigma_\mu \xi_j^\mu \sigma_i^\mu$$

$$h_i = \Sigma_j J_{ij} S_j.$$

In the error-less regime, the maximal capacity is

$$P_c \sim N/\ln N \tag{27}$$

and the information storage is poor

$$i < (\ln N)^{-1}. \tag{28}$$

For a self-coupled net, it is well known [10, 19] that the information density i increases appreciably, in the regime with errors: it rises to a value close to 0.15, then a collapse occurs. For a feedforward net, however, we find that the information density increases smoothly in the error-full regime, toward the asymptotic value

$$i = 1/(\pi \ln 2) \tag{29}$$

i.e. precisely the same value as found above for sparse coding, in our second model. Moreover there exists a simple learning rule that interpolates nicely for arbitrary coding rates, between the dense and sparse limits. This rule, which we shall call for short the Hebbian rule, has been properly defined recently [11] and its properties for self-coupled nets have been studied. It turns out that the analysis of §§6 and 7 can be easily generalised for this rule, because the relative distance between the two field distributions (M) and their common width (Mx) are simply renormalised according to the recipe:

$$f \to f(1-f) \qquad f' \to f'(1-f'). \tag{30}$$

The information content of an output pattern becomes independent of the precise threshold position in the asymptotic error-full regime (P large). Formula (24) is then

generalised into

$$i \ln 2 = f'(1-f')M(1-f)/\pi x(1-f)(1-f') = f'M/\pi x \tag{31}$$

and (25) into

$$i \ln 2 = 1/\pi \qquad \text{for all } f, f'. \tag{32}$$

This asymptotic invariance of the Hebbian rule allows us to draw a simple picture of the evolution of its storage performances, as one interpolates between dense and sparse coding. Taking $f = f'$ for simplicity, we consider the variation of the information storage as a function of the number of patterns presented (learning profile). For $f = \frac{1}{2}$, the initial linear rise, corresponding to the error-less regime, has a short extension, with a maximal capacity given by (28), and the information storage is far below the Gardner limit $i = 2$. However, in the error-full regime, the information stored rises up to $i = 1/(\pi \ln 2) \simeq 0.46$ (figure 6(*a*)). As f decreases, that asymptotic limit remains unchanged, while the error-less regime acquires a larger domain of existence. Eventually, for f small enough, the learning profile exhibits a maximum, reached at, and around, the error threshold. We see now more clearly the particular way whereby the (improving) Hebbian storage manages to touch the (decreasing) Gardner limit, for f small.

8. Conclusion and perspectives

Several interesting results—and good surprises—arise from this analysis of the information storage abilities of Hebbian models. They bring some clarifications on the comparison between sparse and dense coding, and on the possible advantages of going into the error-full regime.

Perhaps the most striking result is the finite asymptotic value, $1/(\pi \ln 2)$, that is reached in the error-full regime, with Hebbian learning. This result has been found analytically and confirmed by numerical simulations. In many cases comparison between this asymptotic value and the information stored at the error threshold is enough to judge whether it is advantageous to enter the error-full regime. Note, in this context, that it would be interesting to have a calculation similar to that of Gardner, asking for maximal information content (without the restriction to error-less retrieval;

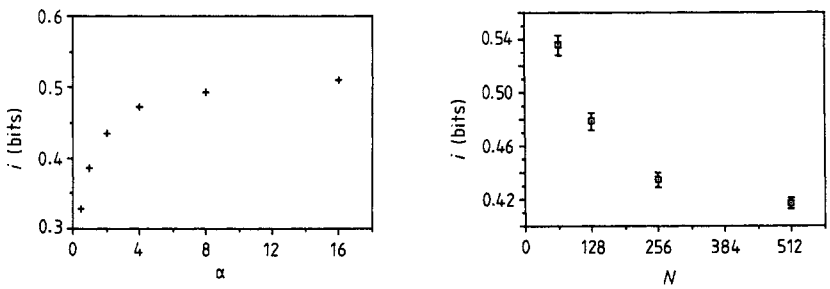

Figure 6. Hebbian rule, for $N = 256$ and $f = f' = 0.5$. The optimal quantity of information per synapse is shown as a function of $\alpha = P/N$. The asymptotic value reached for large α is slightly above the theoretical limit, $1/(\pi \ln 2)$. This is due to finite-size effects, as can be seen in (*b*), where i is plotted as a function of N, for same value of α, $\alpha = 2$.

and in contrast with the case of error-less retrieval, one would expect different results for self-coupled networks and feedforward networks).

Clearly it is of practical importance, for brain studies, if optimal storage properties can be obtained with a simple Hebbian (local, one-shot) learning rule. Further questions, however, request investigation. These performances are obtained with a fine tuning of the threshold. How could such tuning be monitored biologically? An even more pressing task is to imagine how the decoding could be efficiently performed. One can envisage redundant channels followed by an error-correcting stage, but clearly this deserves a separate study.

We have considered a Hebbian rule that interpolates smoothly from dense to sparse coding. It is tempting to construct a similar interpolation scheme for the case of binary synapses, that would extend the Willshaw model for arbitrary coding rates. A proper way of clipping the Hebbian synaptic efficacies seems the natural strategy. But not much has yet been done in this direction.

Acknowledgments

We should like to thank Bernard Derrida, Werner Krauth and Marc Mézard for useful discussions. This work has been supported by the European initiative BRAIN, contract number ST2J-0422-C(EDB).

Appendix. Corrections on the field distributions due to synaptic correlations

These corrections to formulae (12) and (13) originate from the synaptic correlations introduced by the learning mode. Negligible in the limit of sparse coding, the correction terms increase as a function of the coding rate(s). Therefore the study of these corrections sheds light on learning performance limitations, and helps to define properly the conditions inside which a coding rate is safely sparse.

Consider a set of L synapses randomly selected on the output neuron. We wish to compute the probability distribution for the sum of the L synaptic efficacies, after a learning session (P random patterns with input and output coding rates f and f', respectively). For $L = 1$, we shall recover the one-synapse distribution. For the study of retrieval, we shall set $L = M = Nf$.

The patterns are presented one after the other. Only a pattern with a positive output (probability f') has a chance to leave a trace on the set S, and its contribution h will be equal to the number of its input signal neurons that feed into the set S (according to the definition of the original Hebb learning rule). Therefore the probability distribution for the integrated synaptic value of set S, after one learning event, is the sum of a δ-function and a binomial distribution:

$$Q(h) = (1 - f')\delta(h) + f' C_L^h f^h (1 - f)^{L-h} \tag{A1}$$

whose mean and variance are

$$\begin{aligned} \langle h \rangle &= Lff' \\ \langle h^2 \rangle_c &= Lff'[1 - f + Lf(1 - f')]. \end{aligned} \tag{A2}$$

Different learning events are independent events; and for P large the cumulative distribution, according to the central-limit theorem, will be a Gauss law with its two first cumulants obtained from (A1) and (A2) and multiplied by P.

For $L = 1$, one gets the expected result for one synapse. For $L = M$, the bracket in (A2) is a correction factor K, that leads to a widening of the two field distributions important in the retrieval process. It can be checked that this corrective term leads to a reduction of the information storage by a factor $1/K$, in the error-full regime, both near the threshold for errors and in the asymptotic regime (P large).

On this account, and for this particular learning rule, the condition for 'safely sparse' coding is found to rest mainly on the input coding rate:

$$Nf^2 \ll 1 \qquad \text{or} \qquad M^2 \ll N. \tag{A3}$$

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] Shaw G L and Palm G (eds) 1988 *Brain Theory—Reprint Volume* (Singapore: World Scientific)

[6] Buhmann J, Divko R and Schulten K 1987 Associative memory with high information content *Preprint* Munich Technical University; 1989 On sparsely coded associative memories *Neural Networks from Models to Applications* ed L Personnaz and G Dreyfus (Paris: IDSET) p 360

[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] Sompolinsky H 1988 Statistical mechanics of neural networks *Phys. Today* **42** (12) 70

[10] Amit D J 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)

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

[12] Hopfield J J 1982 Neural networks and physical systems with emergent collective abilities *Proc. Natl Acad. Sci. USA* **79** 2554

[13] Gardner E 1987 Maximum storage capacity in neural networks *Europhys. Lett.* **4** 481

[14] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257

[15] Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271

[16] Gardner E and Derrida B 1989 Three unfinished works on the optimal storage capacity of networks *J. Phys. A: Math. Gen.* **22** 1983

[17] Krauth W and Opper M 1989 Critical storage capacity of the $J = \pm 1$ neural network *J. Phys. A: Math. Gen.* **22** L519

[18] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley)

[19] Amit D J, Gutfreund H and Sompolinsky H 1987 Neural networks with correlated patterns: towards pattern recognition *Phys. Rev.* A **35** 2293