Neural Networks, Vol. 2, pp. 451–457, 1989
Printed in the USA. All rights reserved.

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

Characteristics of Sparsely Encoded Associative Memory

SHUN-ICHI AMARI

University of Tokyo

(*Received* 18 *November* 1988; *revised and accepted* 10 *February* 1989)

Abstract—*Characteristics of an autocorrelation or crosscorrelation associative memory largely depend on how items are encoded in pattern vectors to be stored. When most of the components of encoded patterns to be stored are 0 and only a small ratio of the components are 1, the encoding scheme is said to be sparse. The memory capacity and information capacity of a sparsely encoded associative memory are analyzed in detail, and are proved to be in proportion of* $\mathrm{n}^2/(\log\mathrm{n}^2$, n *being the number of neurons, which is very large compared to the ordinary non-sparse encoding scheme of about 0.15*n. *Moreover, it is proved that the sparsely encoded associative memory has a large basin of attraction around each memorized pattern, when and only when an activity control mechanism is attached to it.*

Keywords—Associative memory, Sparse encoding, Memory capacity, Information capacity, Basin of attractor.

1. INTRODUCTION

Correlational associative memory models were proposed by Nakano (1972), Anderson (1972) and Kohonen (1972), and its capacity was studied by Uesaka and Ozeki (1972). Its stability as well as the dynamical behavior was studied by Amari (1972). Since then there have been a large number of papers published on this subject, in particular since Hopfield (1982) proposed the spin glass analogy. Hopfield demonstrated by computer simulation that an associative memory model with n neurons can store about $0.15n$ patterns in the form of its equilibria. It is now well-known that the capacity of this model is $n/(2\log n)$ patterns, if exact recalling is required (see, e.g., Weisbuch, 1985; McEliece, Posner, Rodemich, & Venkatesh, 1987), and the capacity is about $0.15n$, if a small noise is permitted (Amit, Gutfreund, & Sompolinsky 1985; Amari & Maginu, 1988). The dynamics of recalling processes was analyzed by Amari and Maginu (1988) and Amari (1988a, 1988b), where interesting dynamical phenomena were found and explained theoretically. See also Meir and Domany (1987).

The binary signal values 1 and -1 are used in most of the above models, instead of 1 and 0. If the binary values 1 and 0 are used, the result is somewhat different, provided the learning rule is Hebbian: The connection weight w_{ij} between the ith and jth neurons increases by

$$\Delta\, w_{ij} = cx_ix_j.$$

In the 1 and 0 case, w_{ij} never increases when $x_i = x_j = 0$, while it increases in the case of 1 and -1 even when $x_i = x_j = -1$, as well as the case with $x_i = x_j = 1$. Amari (1977) showed that the memory capacity decreases drastically in the case of 1 and 0 values. Of course, if the learning rule is correlational, that is,

$$\Delta\, w_{ij} = c(x_i - a)(x_j - a),$$

we have a similar result as the case with 1 and -1, where a is the average firing rate of x_i.

An encoding scheme is said to be sparse, when the number of excited or active components (i.e., those components for which $x_i = 1$) are very small compared with n, the dimension number of vector patterns $\mathbf{x}$ to be memorized. More precisely, the ratio a_n of the number of excited components to n tends to 0 as n tends to infinity. Superiority of the sparse encoding has been remarked by many researchers (Gardner, 1988; Palm, 1980; Willshaw & Longuet-Higgins, 1970). Palm (1980) studied the information capacity of a special model where the connection weight takes on only 0 and 1 values, and showed that the memory capacity increases drastically in the case of sparse encoding. He also extended his idea of sparse encoding to a general associative memory

The present work is partly supported by grant-in-aid #63112005 and #63115008 from the Ministry of Education, Culture and Science of Japan

Requests for reprints should be sent to Shun-ichi Amari, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

model, and showed that the sparse encoding is excellent (Lansner & Ekeberg, 1985; Palm, 1981, 1988). Such results are also remarked recently by spin glass people (Gardner, 1988), where sparse encoding implies a strong applied magnetic field. Rolls (1987) proposed an associative memory model of the hippocampus, and he observed that "sparse encoding" is realized in the hippocampus, asking for its theoretical outcomes. Sparse encoding is supported by the experimental results of memory by Miyashita and Chang (1988).

The present paper gives a mathematical analysis of associative memory models with sparse encoding, extending the results by Palm (1981). We introduce the sparseness exponent e by

$$a_n = n^{-e}, \qquad (0 < e < 1)$$

where a_n is the probability that a component of x_i of vectors **x** to be stored is equal to 1. Therefore, among n components of **x**, only about $na_n = n^{1-e}$ components are 1 and the other components are 0. We use the convention that $e \rightarrow 1$ implies $a_n = (\log n/n)$, so that log n components are 1 in this case with $e \rightarrow 1$.

The main results of the present paper are summarized as follows: The memory capacity, which is the maximum number $C(e)$ of patterns to be stored in the network in the form of its equilibria is of the order

$$C(e) = \begin{cases} \dfrac{n^{1+e}}{\log n}, & 0 \leqq e < 1 \\ \dfrac{n^2}{(\log n)^2}, & e \longrightarrow 1 \end{cases}$$

for the e-encoding scheme. Therefore, the capacity is maximized as $e \rightarrow 1$ (see Palm, 1981). When $e > 1/3$, the result is the same in order, in the both cases of 1 and -1 signal values and 1 and 0 signal values. However, the 1 and 0 case is much worse when $e < 1/3$.

A sparsely encoded pattern **x** includes a small amount of information compared with a non-sparse encoding case. However, we shall prove that the total amount of information stored in the network is roughly

$$C_I(e) = \begin{cases} \dfrac{n^2}{\log n}, & e = 0 \\ en^2, & \end{cases}$$

which again shows the superiority of sparse encoding.

How large is the basin of attraction? If it is small, sparse encoding is useless even if it has a large memory capacity. Let $\mathbf{x}^*$ be a noisy version of a memorized pattern **x** such that, among n components of **x**, about np components are changed from 0 to 1 or from 1 to 0. We can show that it is difficult to recall the memorized **x** from a noisy $\mathbf{x}^*$, implying that the basin of attraction of **x** is extremely small. However, if we introduce a mechanism which keeps the activity (i.e., the number of excited components) constant, the sparse encoding scheme has a good performance. We consider a p-noisy version $\mathbf{x}^*$ of **x** in the sense that $100p\%$ of the active components ($x_i = 1$) of **x** are changed from 1 to 0, and the same number of inactive components ($x_i = 0$) change their value from 0 to 1, keeping the activity constant. It is then proved that the basins of attractor of one-step recalling are very large for such noisy patterns having a fixed activity. Actually, when the number of stored patterns is $100k\%$ of its capacity, a stored pattern is recalled correctly, via one-step state transition, from its noisy version with $p = 1 - \sqrt{k}$ noise ratio in the above sense.

We treat in the present paper only an autocorrelation associative memory model, which recalls a memorized **x** from its noisy version. However, we can treat a crosscorrelation associative memory model by the same method, which stores pairs of patterns $(\mathbf{s}^\mu, \mathbf{r}^\mu)$, $\mu = 1, 2, \cdots$, such that the model outputs $\mathbf{r}^\mu$ when a noisy version of $\mathbf{s}^\mu$ is given. The memory characteristics of crosscorrelation memory is the same, provided the key patterns $\mathbf{s}^\mu$ are sparsely encoded. They do not depend on the sparsity of the associated $\mathbf{r}^\mu$, so that $\mathbf{r}^\mu$ may be non-sparse. This suggests that, if we encode non-sparse $\mathbf{r}^\mu$ into sparse $\mathbf{s}^\mu$, we can obtain an autoassociative memory of $\mathbf{s}^\mu$ of a large capacity. We then can use a crosscorrelation decoder to obtain the original $\mathbf{r}^\mu$ from $\mathbf{s}^\mu$, without destroying the memory capacity.

2. ASSOCIATIVE MEMORY

Let us consider a neural network composed of n mutually connected McCulloch-Pitts formal neurons. We assume that all the neurons work synchronously at discrete times $t = 1, 2, \cdots$. A neuron emits output 1 when it is excited and its output is 0 when it is not excited. A neuron is excited when a weighted sum of its inputs exceeds a threshold value h. Let $\mathbf{x} = (x_1, x_2, \ldots x_n)$ be a vector whose component x_i denotes the output of the i-th neuron. This vector is called the state vector of the network. When the present state is **x**, the next state $\mathbf{x}'$ is determined from **x** by

$$x_i' = 1\left(\sum_{j=1}^{n} w_{ij}x_j - h\right), \tag{2.1}$$

where w_{ij} is the weight of connection from the jth neuron to the ith neuron and the function $1(u)$ denotes the unit step function,

$$1(u) = \begin{cases} 1, & u > 0, \\ 0, & u \leqq 0. \end{cases}$$

This equation defines the state transition of the network from the present state $\mathbf{x}$ to the next state $\mathbf{x}'$,

$$\mathbf{x}' = T\mathbf{x} \tag{2.2}$$

where T is the nonlinear operator defined by (2.1) in the component form. Let $\mathbf{x}(t)$ be the state at time t of the network. Its dynamical behavior is then written by the state transition equation

$$\mathbf{x}(t+1) = T\mathbf{x}(t). \tag{2.3}$$

Let us consider m vectors $\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^m$. When a network satisfies

$$\mathbf{s}^\mu = T\mathbf{s}^\mu \tag{2.4}$$

for all $\mu = 1, 2, \cdots, m$, that is all the vectors $\mathbf{s}^\mu$ are equilibrium states of the net, we say that m vectors are memorized or stored in the net in the form of its equilibria. The autocorrelation associative memory proposed by Nakano (1972), Anderson (1972), Kohonen (1972), analyzed mathematically by Amari (1972, 1977) and reformulated by Hopfield (1982), is a network whose connection weight matrix $W = (w_{ij})$ is determined by

$$w_{ij} = \frac{1}{n}\sum_{\mu=1}^{m} s_i^\mu s_j^\mu, \qquad w_{ii} = 0 \tag{2.5}$$

from m patterns $\mathbf{s}^\mu$ to be stored, where s_i^μ is the ith component of $\mathbf{s}^\mu$.

When an initial state $\mathbf{x}^0$ belongs to the basin of attraction of a stored pattern $\mathbf{s}^\mu$, it satisfies

$$\mathbf{s}^\mu = T^k\mathbf{x}^0$$

for some k. Starting with $\mathbf{x}^0$, the state of the net falls in $\mathbf{s}^\mu$ after k steps of state transition. This is interpreted that the network recalls $\mathbf{s}^\mu$ from an initial state $\mathbf{x}^0$. When

$$\mathbf{s}^\mu = T\mathbf{x}^0 \tag{2.6}$$

holds, the network recalls $\mathbf{s}^\mu$ from $\mathbf{x}^0$ by one-step state transition.

Let us define the activity a of a state vector $\mathbf{x}$ by the ratio of excited neurons to all the neurons,

$$a = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.7}$$

When the activity of $\mathbf{x}$ is a, among n components of $\mathbf{x}$, na components are 1 and $n(1-a)$ are 0. The activities of all the patterns $\mathbf{s}^\mu$ to be stored are controlled to be equal to a fixed constant a in the present paper. In other words, we assume that the threshold value h is controlled by some mechanism such that the state $\mathbf{x}$ of the network is always controlled to have a fixed activity a. We do not mention this additional mechanism in detail. It can be easily implemented by a feedback control mechanism with an analog inhibitory neuron (or neuron pool) which is excited by the output activity of the associative memory network and which in turn inhibits its component neurons (see, e.g., Amari & Arbib, 1977). We simply assume here that, among n neurons in the network, those which receive the na largest stimuli (weighted sums of inputs) are excited.

A pattern $\mathbf{s}^\mu$ includes na active components whose values are 1. We treat also a noisy version of $\mathbf{s}^\mu$, such that, among na active components, $100p\%$ (i.e., nap components) are changed to 0, and instead the same number of non-active components whose values are 0 are changed to 1, keeping the total activity constant. Such a vector is said to be a noisy version of $\mathbf{s}^\mu$ with a noise ratio p.

3. SPARSE ENCODING WITH FIXED ACTIVITY

We consider the case where patterns $\mathbf{s}^\mu$ to be stored are generated independently and randomly under the condition that they have a fixed activity. In this case, items to be memorized are encoded in randomly generated $\mathbf{s}^\mu$ and the latter patterns are stored in the network. We treat asymptotic properties of the associative memory, so that we assume that activity a_n of $\mathbf{s}^\mu$ depends on n. When

$$\lim_{n\to\infty} a_n = 0 \tag{3.1}$$

holds, this encoding is said to be sparse, because the number of active components in $\mathbf{s}^\mu$ becomes negligibly small compared to n.

Vectors $\mathbf{s}^\mu$ are independent random vectors subject to a common probability distribution. More precisely, $\mathbf{s}^\mu$ are generated in such a manner that, given a_n, na_n components are randomly chosen among n components and put equal to 1 and all the other components are put equal to 0. Therefore, the probability distribution of each component s_i^μ of $\mathbf{s}^\mu$ satisfies

$$E[s_i^\mu] = a_n, \quad V[s_i^\mu] = a_n(1-a_n), \tag{3.2}$$

$$Cov[s_i^\mu, s_j^\mu] = -\frac{1}{n}a_n(1-a_n), \qquad i \neq j, \tag{3.3}$$

where E, V, and Cov denote the expectation, variance, and covariance, respectively. Because of the fixed activity constraint, s_i^μ and s_j^μ are not independent. However, their correlation is negligibly small, and careful calculations show that we may disregard this correlation in the first approximation. This implies that we may treat it as if all s_i^μ are independent, s_i^μ taking 1 with probability a_n and 0 with probability $1 - a_n$.

In order to evaluate the sparseness of encoding, we put

$$a_n = cn^{-\epsilon}, \tag{3.4}$$

where c is a constant and e is a power exponent. This is called an e-encoding. When $e = 0$, we have an ordinary non-sparse encoding. If $e = 1$, the number na_n of active components is kept constant even if $n \to \infty$. We exclude this case, because the central limit theorem cannot be applied in this case as will be shown soon. Instead, we use the following convention that

$$\lim_{e \to 1} a_n = c(\log n)n^{-1}, \qquad (3.5)$$

so that the number na_n of active elements grows in a logarithmic order of n with $e \to 1$.

4. MEMORY CAPACITY AND INFORMATION CAPACITY

We now search for capabilities of an associative memory network using sparse encoding with a fixed activity. A memory capacity is defined by the maximum number of patterns which can be stored as equilibria without any confusion, that is, the maximum number m for which

$$T\mathbf{s}^{\mu} = \mathbf{s}^{\mu}, \qquad \mu = 1, 2, \ldots, m \qquad (4.1)$$

hold. Since the patterns are randomly generated, we define the memory capacity $C_n(e)$ of e-encoding by using the probability measure asymptotically in n such that $C_n(e)$ is the maximum m for which (4.1) holds for all $\mu = 1, \ldots, m$ with a probability as close to 1 as desired for sufficiently large n.

Theorem 1. The memory capacity of associative memory with e-encoding is given asymptotically by

$$C_n(e) = \begin{cases} \dfrac{n^{3e}}{8c^3(1 + 3e)\log n}, & e < 0.5 \\ \dfrac{n^{1+e}}{8c(2 + e)\log n}, & e > 0.5 \\ \dfrac{n^2}{24c(\log n)^2}, & e \longrightarrow 1. \end{cases} \qquad (4.2)$$

The proof is outlined in section 6. The theorem shows that the memory capacity increases as e becomes large. Therefore, as encoding becomes sparser, the capacity becomes larger. It should be noted that the memory capacity ratio m/n diverges as n tends to infinity for $e > 1/3$, implying that the possible number of stored patterns grows more rapidly than a linear order of n. This differs from the non-sparse encoding scheme. It is known (McEliece et al., 1987) that the memory capacity is

$$C_n = \frac{n}{4\log n}$$

in non-sparse encoding ($e = 0$, $c = 0.5$), if each component takes on the binary values 1 and -1. In our case of the binary component values 1 and 0, the memory capacity is much worse, if $e < 1/3$. The distinction between these two different binary component values disappears, if we use a little unnatural connection matrix given by

$$w_{ij} = \frac{1}{n}\sum_{\mu=1}^{m} (s_i^{\mu} - a_n)(s_j^{\mu} - a_n) \qquad (4.3)$$

instead of (2.5). In this latter case, we have

$$C_n(e) = \begin{cases} \dfrac{n^{1+e}}{8c(2 + e)\log n}, \\ \dfrac{n^2}{24c(\log n)^2}, & e \longrightarrow 1 \end{cases} \qquad (4.4)$$

Although the memory capacity increases as encoding becomes sparser, the amount of information of one encoded pattern decreases as e tends to 1. There are ${}_nC_{an}$ vectors whose activity is a, where ${}_nC_k$ is a binominal coefficient. Therefore, such a pattern vector includes $\log {}_nC_{an}$ bits of information. Hence, one e-encoded pattern $\mathbf{s}^{\mu}$ includes asymptotically

$$I(e) = \begin{cases} nH_c, & e = 0 \\ cen^{1-e}\log n, & 0 < e < 1 \\ c(\log n)^2, & e \longrightarrow 1 \end{cases} \qquad (4.5)$$

bits of information (see Appendix A), where

$$H_c = -c\log c - (1 - c)\log(1 - c).$$

It is natural to define the information capacity $C_I(e)$ of a sparsely encoded associative memory by

$$C_I(e) = I(e)C_n(e), \qquad (4.6)$$

which is the total amount of information of stored pattern vectors. This gives the following theorem.

Theorem 2. The information capacity of e-encoded associative memory is

$$C_I(e) = \begin{cases} \dfrac{e}{8c^2(1 + 3e)} n^{1+2e}, & e < 0.5 \\ \dfrac{e}{8(2 + e)} n^2, & e > 0.5 \\ \dfrac{n^2}{24}, & e \longrightarrow 1. \end{cases} \qquad (4.7)$$

This theorem shows that the information capacity again increases as encoding becomes sparser. In the case of (4.3), we have

$$C_I(e) = \begin{cases} \dfrac{H_c n^2}{16c\log n}, & e = 0 \\ \dfrac{e}{8(2 + e)} n^2, & \text{otherwise.} \end{cases}$$

Therefore, the information capacity is in proportion to the number n^2 of synapses, except for the case with non-sparse encoding ($e = 0$).

The fact that $C_I(e)$ does not depend on the coefficient c is interesting. Since $C_I(e)$ becomes larger as $c \to 0$, the memory capacity may become larger if we use a sparser encoding than $e \to 0$, for example,

$a_n = \sqrt{\log n/n}$. However, this does not increase the information capacity, because $C_I(e)$ is fixed even $c \to 0$.

5. ONE-STEP RECALL REGION

It has been shown that the sparse encoding makes it possible for an associative memory model to have a large memory and information capacity. If it has a large basin of attraction, its recalling performance is very good. It is, however, difficult to search for the size and shape of the basin of attraction (see Amari & Maginu, 1988) in the case of non-sparse encoding). We study the region of one-step recalling from a noisy version of a stored pattern.

When

$$T\mathbf{x}_0 = \mathbf{s}^\mu$$

holds, $\mathbf{s}^\mu$ is said to be recalled from $\mathbf{x}_0$ by one-step state transition. When $\mathbf{s}^\mu$ is recalled from any of its p-noisy versions and less noisy versions (q-noisy versions, $q < p$), we say that $\mathbf{s}^\mu$ has a one-step recall region with a radius no less than p. Since all the patterns to be stored are randomly generated, we define the radius p_n of the one-step recalling region of $\mathbf{s}^\mu$ asymptotically by the maximum p_n such that the probability of recalling $\mathbf{s}^\mu$ from a p-noisy version tends to 1 as n goes to infinity. The radius p_n depends on e and m so that we write

$$p_n = p_n(e, m).$$

It is obvious that $p_n \to 0$ when m is larger than the capacity $C_n(e)$. Therefore, we study the radius p_n when m is given by

$$m = kC_n(e).$$

We show that the asymptotic radius p,

$$p = \lim_{n \to \infty} p_n(e, kC_n(e)) \tag{5.1}$$

depends only on the ratio k.

Theorem 3. The radius of the one-step recalling region is given by

$$p = 1 - \sqrt{k}. \tag{5.2}$$

The theorem implies that the radius of the one-step recall region is sufficiently large, if m is kept adequately small compared to its capacity. It should, however, be remarked that this is because of the activity control mechanism. The activity of initial pattern $\mathbf{x}_0$ should be controlled to be equal to a_n. If the activity of an initial pattern is not controlled to be equal to a_n, one-step correct recalling is never guaranteed.

6. OUTLINE OF PROOFS

Without loss of generality, we study the one-step recalling process of the first pattern $\mathbf{s}^1$ from a p-noisy version $\mathbf{x}$ of $\mathbf{s}^1$. To this end, we evaluate the probability that

$$T\mathbf{x} = \mathbf{s}^1$$

holds. We put

$$P(e, n, m, p) = \text{Prob}\,\{T\mathbf{x} = \mathbf{s}^1 | \mathbf{x} \text{ is a } p\text{-noisy version of } \mathbf{s}^1\}.$$

When $p = 0$, this gives the probability that $\mathbf{s}^1$ is an equilibrium. Let us define u_i by

$$\begin{aligned} u_i &= \sum w_{ij} x_j - h \\ &= \frac{1}{n} s_i^1 \mathbf{s}^1 \cdot \mathbf{x} + N_i - h, \end{aligned}$$

where

$$N_i = \frac{1}{n} \sum_{\mu=2}^{m} \sum_{j \neq i}^{n} s_i^\mu s_j^\mu x_j \tag{6.1}$$

is the crosstalk term due to other patterns disturbing correct recalling of $\mathbf{s}^1$. Since $\mathbf{x}$ is a p-noisy version of $\mathbf{s}^1$, we have

$$\mathbf{s}^1 \cdot \mathbf{x} = na_n(1 - p).$$

Therefore, the ith component of $T\mathbf{x}$ is given by

$$(T\mathbf{x})_i = 1[u_i] = 1[a_n(1 - p) + N_i - h]$$

when $s_i^1 = 1$, and when $s_i^1 = 0$,

$$(T\mathbf{x})_i = 1[N_i - h].$$

The threshold value h is determined such that na_n components of $T\mathbf{x}$ become 1. Recalling is correct when

$$h - a_n(1 - p) < N_i \text{ and } N_i < h$$

for $s_i^1 = 1$ and $s_i^1 = 0$, respectively. In order to obtain the probability of the above inequalities, we evaluate the probability distribution of N_i, where $\mathbf{s}^\mu$ ($\mu = 2, \ldots, m$) are assumed to be random variables (Appendix B).

Lemma 1. The probability distribution of N_i is asymptotically normal with mean ma_n^3 and variance

$$\sigma^2 = \begin{cases} ma_n^5, & e < 0.5 \\ mn^{-1}a_n^3, & e > 0.5 \end{cases} \tag{6.2}$$

The threshold value h is then asymptotically given by

$$h = ma_n^3 + 0.5\, a_n(1 - p). \tag{6.3}$$

We now evaluate the probability q of an error occuring in the ith component. When $s_i^1 = 1$, an error occurs if $N_i < h - a_n(1 - p)$, and when $s_i^1 = 0$, an error occurs if $N_i > h$. The error probability is given in both cases by

$$\begin{aligned} q &= \text{Prob}\,\{N_i < m\, a_n^3 - 0.5\, a_n(1 - p)\} \\ &= F\left(\frac{1 - p}{2\sigma} a_n\right), \end{aligned} \tag{6.4}$$

where $F(u)$ is the error integral given by

$$F(u) = (2\pi)^{-1/2} \int_{-\infty}^{-u} \exp\{-x^2/2\} dx.$$

Therefore, the probability that no error occurs in any components, that is, $T\mathbf{x} = \mathbf{s}^1$, is given by

$$Q = (1 - q)^n.$$

More precisely, N_i are not independent so that we need to take their correlations into account. However, if we do so, the probability can be evaluated asymptotically by the above Q.

The probability that

$$T\mathbf{x}^\mu = \mathbf{s}^\mu$$

holds for fall μ, where $\mathbf{x}^\mu$ is a p-noisy version of $\mathbf{s}^\mu$, is given by

$$P = Q^m = (1 - q)^{nm}.$$

We have

$$\log P = nm \log(1 - q) = -\exp\left\{ -\frac{(1-p)^2}{8\sigma^2} a_n^2 + \log nm - \log(a_n/\sigma) \right\}, \quad (6.5)$$

where we use the following asymptotic evaluation of $F(u)$ when u is large,

$$F(u) = (\surd 2\, \pi u)^{-1} \exp\{-u^2/2\}.$$

In order that P converges to 1 as n tends to infinity, the number m should not be larger than the following limit (see Appendix C),

$$(1 - p)^2 a_n^2 \approx 8\sigma^2 \log nm. \quad (6.6)$$

In order to determine the capacity, we put $p = 0$ and solve the above equation to evaluate m. By using (3.4), (3.5), and (6.2), we hae the capacity (4.2). By using (4.5), we have the information capacity.

By substituting $m = kC_n(e)$ in (6.6), we easily have the radius of one-step recalling (5.2).

REFERENCES

Amari, S. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, **C-21**, 1197–1206.

Amari, S. (1977). Neural theory of association and concept-formation. *Biological Cybernetics*, **26**, 175–185.

Amari, S. (1988a). Associative memory and its neurodynamical analysis. In H. Haken (Ed.), *Neural and synergetic computers (pp. 85–99)*. Springer Series in Synergetics. New York: Springer Verlag.

Amari, S. (1988b). Various versions of associative memory. *Proceedings of the 2nd IEEE ICNN Conference*, **I**, 633–640.

Amari, S., & Arbib, M. A. (1977). Competition and cooperation in neural nets. In Systems neuroscience J. Metzler (Ed.), (pp. 119–165). New York: Academic Press.

Amari, S., & Maginu, K. (1988). Statistical neurodynamics of associative memory. *Neural Networks*, **1**, 63–73.

Amit, D. J., Gutfreud, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review*, **A2**, 1007–1018.

Anderson, J. A. (1972). A simple neural network generating interactive memory. *Mathematical Biosciences*, **14**, 197–220.

Gardner, E. (1988). The space of interactions in neural network models, *Journal of Physics*, **A21**, 257–270.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, USA, **79**, 2445–2558.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, **C-21**, 353–359.

Lansner, A., & Ekeberg, Ö. (1985). Reliability and speed of recall in an associative network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI** 7, 490–498.

McEliece, R. J., Posner, E. C., Rodemich, E. R., & Venkatesh, S. S. (1987). The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory*, **IT-33**, 461–482.

Meir, R., & Domany, E. (1987). Exact solution of a layered neural network memory. *Physical Review Letters*, **59**, 359–362.

Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, **331**, 68–70.

Nakano, K. (1972). Associatron—a model of associative memory. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-2**, 381–38.

Palm, G. (1980). On associative memory. *Biological Cybernetics*, **36**, 646–658.

Palm, G. (1981). On the storage capacity of an associative memory with randomly distributed storage elements. *Biological Cybernetics*, **39**, 125–127.

Palm, G. (1988). Local synaptic rules with maximal information storage capacity. In H. Haken (Ed.), *Neural and synergetic computers*, (pp. 100–110). Springer Series in Synergetics. New York: Springer Verlag.

Rolls, E. T. (1987). Information representation, processing, and storage in the brain: Analysis at the single neuron level. In J.-P. Changeux & M. Konishi (Eds.), *The neural and molecular bases of learning* (pp. 503–539). New York: Wiley.

Uesaka, Y. & Ozeki K. (1972). Some properties of associative type memories. *Journal of Institute of Electrical and Communication Engineers of Japan*, **55-D**, 323–330.

Weisbuch, G. (1985). Scaling laws for the attractors of Hopfield networks. *Journal de Physique Letters*, **46**, 623–630.

Willshaw, D. J., & Longuet-Higgins, H. C. (1970). Associative memory models. In B. Meltzer & O. Michie (Eds.), *Machine intelligence* (Vol. 5). Edinburgh: Edinburgh University Press.

APPENDIX A. INFORMATION AMOUNT IN A SPARSELY ENCODED PATTERN

The number of vectors in which k components are 1 and $(n - k)$ components are 0 is given by the binomial coefficient ${}_nC_k$. Hence, the amount of information included in such a pattern is $I = \log {}_nC_k$. By using Sterling's formula, we have

$$\begin{aligned} I = \log {}_nC_k &= n\log n - k\log k \\ &\quad - (n - k)\log(n - k) + O(\log n) \\ &= k\log(n/k) + O(\log n) + O(k), \end{aligned}$$

when $k \ll n$. By substituting $k = na_n = cn^{1-e}$, we have

$$I(e) = cen^{2-e}\log n, \qquad 0 < e < 1.$$

In the limit $e \rightarrow 1$, $\quad k = c\log n$ so that

$$I(e) = c(\log n)^2.$$

When $e = 0$, we have the well known formula

$$I(e) = nH.$$

APPENDIX B. PROBABILITY DISTRIBUTION OF N_i

Since the noise or crosstalk term N_i is a sum of $n(m - 1)$ variables $s_i^\mu s_j^\mu x_j$ divided by n, it is not difficult to show that N_i is asymptotically normally distributed if an adequate scaling factor is chosen. It should be noted that these $n(m - 1)$ variables are not independent. Let us put

$$r^\mu = \mathbf{s}^\mu \cdot \mathbf{x} = \sum s_j^\mu x_j.$$

Given a p-noisy version $\mathbf{x}$ of $\mathbf{s}^1$ of which na_n components are 1, r^μ is written as a sum of s_i^μ,

$$r^\mu = \sum{}' s_j^\mu,$$

where Σ' implies summation over such j for which $x_j = 1$. Hence, r^μ is a sum of na_n random variables s_i^μ. Therefore

$$E[r^\mu] = na_n^2.$$

The variance of r^μ is a little complicated, because s_j are not independent, of (3.3). Neglecting small order terms, we have an asymptotic evaluation

$$\begin{aligned} V[r^\mu] &= V\left[\sum{}' s_j^\mu\right] = na_n V[s_j^\mu] + na_n(na_n - 1)Cov[s_j^\mu, s_k^\mu] \\ &= na_n^2(1 - a_n) - na_n^3(1 - a_n) \\ &= na_n^2(1 - a_n)^2 \doteq na_n^2. \end{aligned}$$

From

$$N_i = \frac{1}{n}\sum_{\mu=2}^{m} s_i^\mu r^\mu,$$

we have asymtotically

$$E[N_i] = ma_n^3.$$

Similarly, since $s_i^\mu r^\mu$ and $s_i^\lambda r^\lambda$ $(\mu \neq \lambda)$ are almost independent,

$$\begin{aligned} n^2V[N_i] &= V\left[\sum s_i^\mu r^\mu\right] = mV[s_i^\mu r^\mu] \\ &= m\{E[s_i^\mu(r^\mu)^2] - (E[s_i^\mu r^\mu])^2\} \\ &= m\{a_n[na_n^2(1 - a_n)^2 + n^2a_n^4] - n^2a_n^6\} \\ &= mna_n^3\{(1 - a_n)^2 + na_n^2(1 - a_n^2)\} \\ &\doteq mna_n^3(1 + na_n^2). \end{aligned}$$

When $e < 0.5$, the term na_n^2 dominates 1, and when $e > 0.5$, $na_n^2 \to 0$. Therefore, we have

$$\sigma^2 = \begin{cases} ma_n^5, & e < 0.5 \\ mn^{-1}a_n^3 & e > 0.5 \end{cases}$$

This proves (6.2).

The covariance between N_i and N_j is given similarly by

$$Cov[N_i, N_j] = mn^{-1}a_n^4.$$

In order to evaluate Prob $\{T\mathbf{x} = \mathbf{s}^1\}$, we need a little complicated procedure, because of this covariance.

When the connection matrix (w_{ij}) is given by the covariance (4.3) of excitations of the neurons, the variance is

$$\sigma^2 = mn^{-1}a_n^3$$

irrespective of the value of e. This is the same as the case where the output of each neuron takes on 1 and -1.

APPENDIX C: EVALUATION OF PROBABILITY

When and only when

$$-\frac{(1 - p)^2}{8\sigma^2}a_n^2 + \log nm - \log\frac{a_n}{\sigma} \longrightarrow -\infty$$

the probability P converges to 1. Since σ^2 is proportional to m, this shows that m cannot be greater than some order of n. When m satisfies

$$8\sigma^2\log nm = (1 - p)^2a_n^2,$$

$a_n/\sigma \to \infty$, so that $P \to 1$. However, when m is larger than this limit, P cannot tends to 1. By substituting (6.2) and (3.4), the above equation is easily solved in the asymptotic sense, giving the result (4.2). See Amari and Maginu (1988) or McEliece et al. (1987) for more detailed discussions.