TR-92-37

Optimal Decay Rate of Connection Weights in Covariance Learning

Shotaro Akaho *
akaho@etl.go.jp

Received 10 November 1992

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba,
Ibaraki, 305, JAPAN

Abstract Associative memory of neural networks can not store items more than its memory capacity. When new items are given one after another, connection weights should be decayed so that the number of stored items does not exceed the memory capacity. This paper presents the optimal decay rate that maximizes the number of stored items, using the method of statistical dynamics.

*Mathematical Informatics Section

1 Introduction

This paper addresses the memory capacity of an associative memory model of neural networks with weight decay.

Neural network is an adaptive system that is trained with sample items that are given from outer environment. It can store items up to some number, which is called *memory capacity*. We consider the on-line type learning scheme (cf. batch type learning) where the learning is proceeded every time a new item is provided. This learning has the advantage that needed memory is a little (memory for all items is not necessary) and it can adapts itself well to the change of the environment. However, when new items are given one after another, it cannot be judged whether the number of stored items is more than the capacity or not, because the stored items are memorized implicitly on connection weights and moreover the network has no memory for the number of items. One method to avoid exceeding the capacity is decaying connection weights to remove older items. If we want to store items as many as possible, connection weights should be decayed as slowly as possible, but if the decay is too slow, old items affects the recall of newer items. Thus there is an optimal decay rate that maximizes the number of stored items.

In this paper, we analyze an associative memory model and present the optimal decay rate. In section 2, we state the learning of associative memory with decay. In section 3 and 4, we define the problem and shows the main theorem. In section 5, we show some simulation result.

2 Associative memory model

The associative memory model is originally proposed in 1972 and analyzed by many researchers[10][4][6][12][3]. Recently, it has attracted more attention because of *sparse coding scheme*, where most of components of pattern vectors are 0 and only a small ratio of those are 1. It was shown that if the ratio of 1s of output pattern is $O(\log n/n)$ where n is the number of neurons, its memory capacity becomes $O(n^2/(\log n)^2)$, while $O(n/\log n)$ in non-sparse coding[13][1][7]. Sparse coding scheme is also supposed to act an important function in the brain memory such as hippocampus[11], and some important experimental results are coming out[9].

The autocorrelation associative memory[1] is trained by a simple Hebbian rule as follows (we consider the case of discrete time).

$$w_{ij}(t+1) = (1-\varepsilon)w_{ij}(t) + cs_i(t)s_j(t), \qquad i \neq j, \tag{1}$$

$$0 < \varepsilon < 1, \quad 0 < c,$$

where w_{ij} is a connection weight from j-th input to i-th neuron, $s_i(t)$ is i-th component of pattern learned at time t, ε and c are time constants (we assume $c = 1$ without loss of

[1]We consider autocorrelation type associative memory, but the result is exactly the same for the crosscorrelation associative memory

generality) and $1-\varepsilon$ denotes the decay rate. By the learning scheme above, connection weights become

$$w_{ij}(t) = \sum_{\mu=0}^{\infty} (1-\varepsilon)^{\mu} s_i(t-\mu) s_j(t-\mu), \qquad i \neq j. \tag{2}$$

Each pattern component $s_i(t)$ takes binary value and there are two possible models. One is 0-1 model and the other is ± 1 model. Amari has shown that 0-1 model is superior in the sparse case and ± 1 model is superior in the non-sparse case[1]. In order to treat both cases, we shall encode patterns as follows.

$$s_i(t) = \begin{cases} 1-a & \text{Prob } a \\ -a & \text{Prob } 1-a \end{cases} \tag{3}$$

where a is called the *activity*. Each $s_i(t)$ takes binary value independently according to the above probability. This coding works as 0-1 model in sparse case and ± 1 model in non-sparse case and it also makes mathematical analysis easier.

Output $\boldsymbol{x}$ for input $\boldsymbol{s}$ is given by

$$x_i = 1_a\left(\frac{1}{n}\sum_{j=0}^{n} w_{ij} s_j - h\right), \tag{4}$$

where h is a threshold value and 1_a is a binary threshold function

$$1_a(u) = \begin{cases} 1-a & u \geq 0 \\ -a & u < 0 \end{cases} \tag{5}$$

3 Memory capacity and optimal decay rate

Let us define the capacity of the model described in the previous section. For some given item $\boldsymbol{s}(t-\mu)$, if it is recalled correctly, namely, if

$$s_i = 1_a\left(\frac{1}{n}\sum_{j=0}^{n} w_{ij} s_j - h\right), \tag{6}$$

holds, pattern $\boldsymbol{s}(t-\mu)$ is said to be stored. A memory capacity M is defined as the maximal number of m, such that most recently learned m items $\boldsymbol{s}(t), \boldsymbol{s}(t-1), \ldots, \boldsymbol{s}(t-m+1)$ can be recalled correctly. Since the patterns are randomly generated, we consider the case that items can be stored with a probability 1 asymptotically for sufficiently large n[5].

Theorem 1 *The optimal decay rate of an associative memory with n neurons is given asymptotically by*

$$1-\varepsilon_{\text{opt}} = 1 - \frac{8e(2+d)a(1-a)\,\log n}{n}, \tag{7}$$

where $d = -\log_n a$ and the memory capacity is given by

$$M_{\text{opt}} = \frac{1}{2\varepsilon_{\text{opt}}} = \frac{n}{16e(2+d)a(1-a)\,\log n}. \tag{8}$$

When $a = 1/2$ (non-sparse case), the capacity is $n/(8e \log n)$, while it is $n/(4 \log n)$ in the case that the network learns only finite number of patterns without decay (e.g. batch type learning). On the proof and more details of this theorem, see section 6. In general, the capacity of this model is $1/2e$ times the capacity in the batch type learning.

4 Some properties of recalling each pattern

Theorem 1 ensures the correctness of recalling patterns up to the capacity $M_{\rm opt}$. However, the correctness does not drop so soon for older patterns than the capacity since the decay is very slow. In this section we investigate the recall of each pattern and estimate the error rate.

Let ε be a form as follows.

$$\varepsilon = \frac{\varepsilon_0 a(1-a)\log n}{n} \tag{9}$$

where ε_0 is a constant order value when the decay rate is optimal.

Let us consider the recall of m-th pattern, where

$$m = \frac{m_0 n}{a(1-a)\log n} = \frac{m_0 \varepsilon_0}{\varepsilon}. \tag{10}$$

When ε_0 is a constant order, m_0 of the capacity is also a constant order.

We can estimate the frequency of incorrectly recalled components of the m-th pattern.

Theorem 2 *The frequency of error components for m-th pattern is asymptotically given by*

$$r(\varepsilon_0, m_0) = O(n^{-r_0(\varepsilon_0, m_0)}), \tag{11}$$

where

$$r_0(\varepsilon_0, m_0) = \frac{\varepsilon_0 \exp(-2\varepsilon_0 m_0)}{8}, \tag{12}$$

and ε_0 and m_0 are defined by (9),(10) respectively.

Figure 1 is a numerical plot of R for some decay rates, where the number of neurons is 1000. It shows that the correctness for recent patterns becomes higher as ε_0 increases, but the correctness for patterns older than memory capacity becomes worse.

Next, we analyze the error correction capability. Consider the noisy version of m-th pattern as follows. Asymptotically, we can say that na components of m-th pattern are $1-a$ and the others are $-a$. To make the noisy version of the pattern with keeping the activity a, we pick up randomly $na\xi$ components of $1-a$ and flip them into $-a$, and similarly flip $na\xi$ components of $-a$ into $1-a$. If $\xi = 0$, the pattern includes no noise.

$$\begin{array}{cccccc} \overbrace{\underbrace{1-a}_{na\xi}\ 1-a\ \cdots\ 1-a}^{na} & & \overbrace{\underbrace{-a}_{na\xi}\ -a\ \cdots\ -a}^{n(1-a)} & & & : \text{original pattern } \boldsymbol{s}^{(m)} \\ \Downarrow & & \Downarrow & & & \\ -a\ \ -a & 1-a & 1-a\ \ 1-a & -a & & : \text{noisy pattern } \hat{\boldsymbol{s}}^{(m)} \end{array} \tag{13}$$

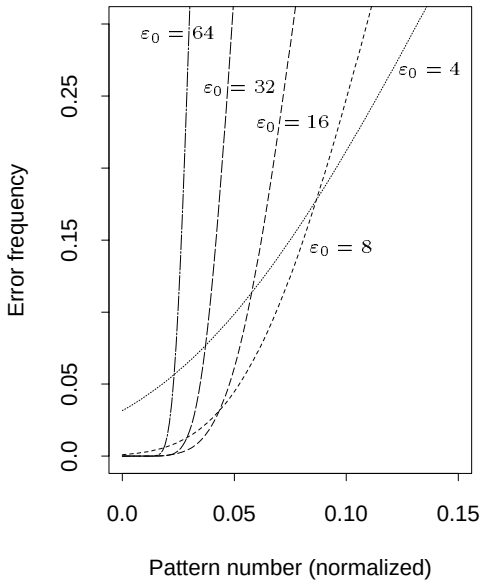

Figure 1: Error frequency versus normalized pattern number(m_0); $\varepsilon_0 = 4, 8, 16, 32, 64$; n = 1000

Theorem 3 *The frequency of error components for noisy version of m-th pattern is asymptotically given by*

$$r(\varepsilon_0, m_0; \xi) = O(n^{-r_\xi(\varepsilon_0, m_0)}), \tag{14}$$

where

$$r_\xi(\varepsilon_0, m_0) = \frac{\varepsilon_0 \exp(-2\varepsilon_0 m_0)}{8} \frac{(1-a-\xi)^2}{(1-a)^2}, \tag{15}$$

ε_0, m_0 are defined by (9),(10), and ξ is a noise parameter defined by (13).

5 Simulation results

In this section, we show some simulation results to ascertain the theorems described in preceding sections.

Figure 2 shows the capacity for some values of ε normalized by the theoretical optimal value, where we stored ten times as much as the theoretical capacity. We take the newest pattern incorrectly recalled as the capacity. We can see that the decay rate that maximizes the capacity is close to the theoretical value for all cases. Figure 3 shows the collection of the maximal number of capacities in the same simulation as above. Theoretical capacities seems rather strict than the simulation, but it fits the simulation well. Figure 4 shows the error frequency plot for each pattern. Pattern number is normalized by theoretical capacity. There seems to be a critical point around 2 where the error suddenly increases.

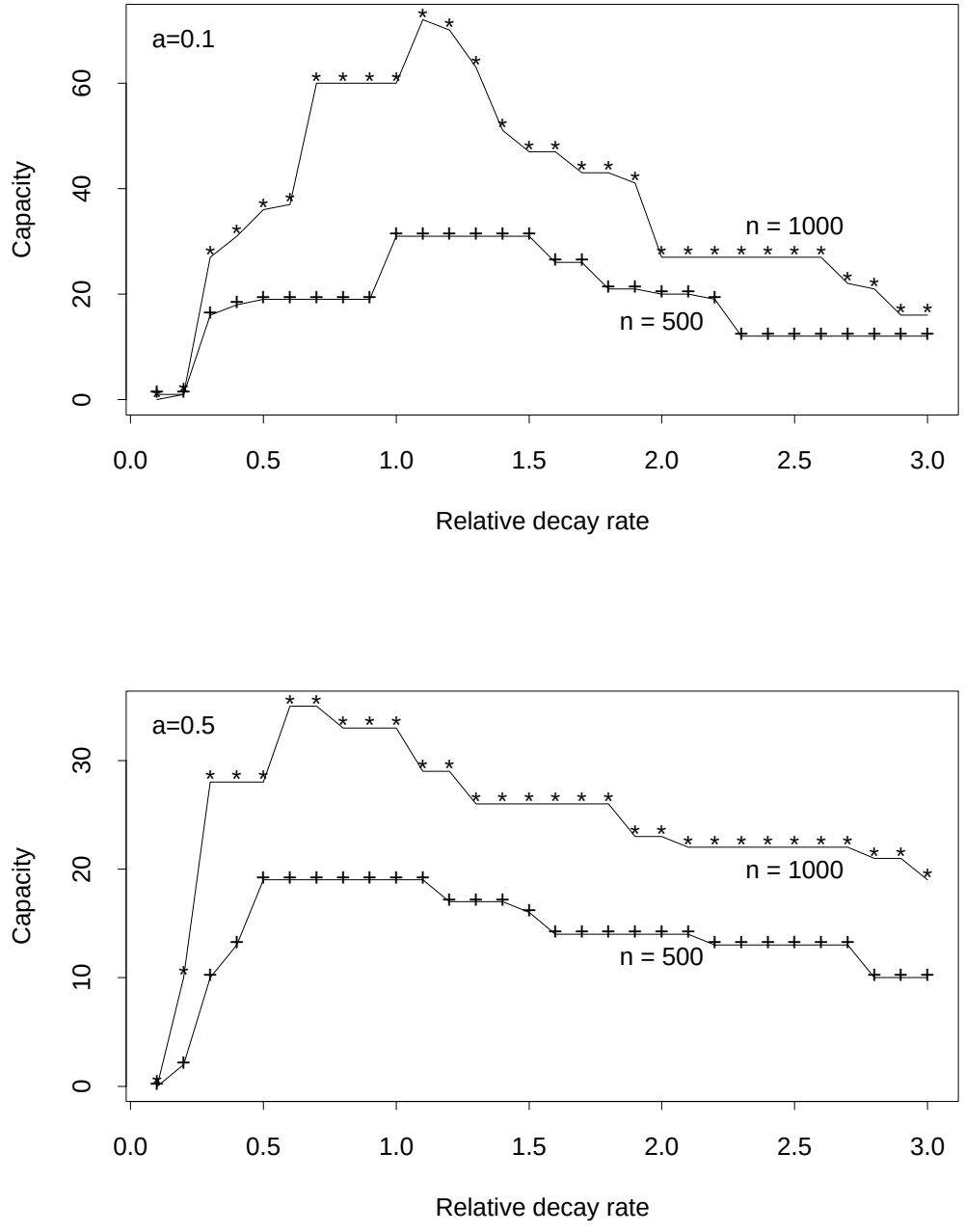

Figure 2: Capacity versus decay rate (ε). Decay rate is normalized by the optimal one. Up: a=0.1 (sparse case); Down: a=0.5 (non-sparse case)

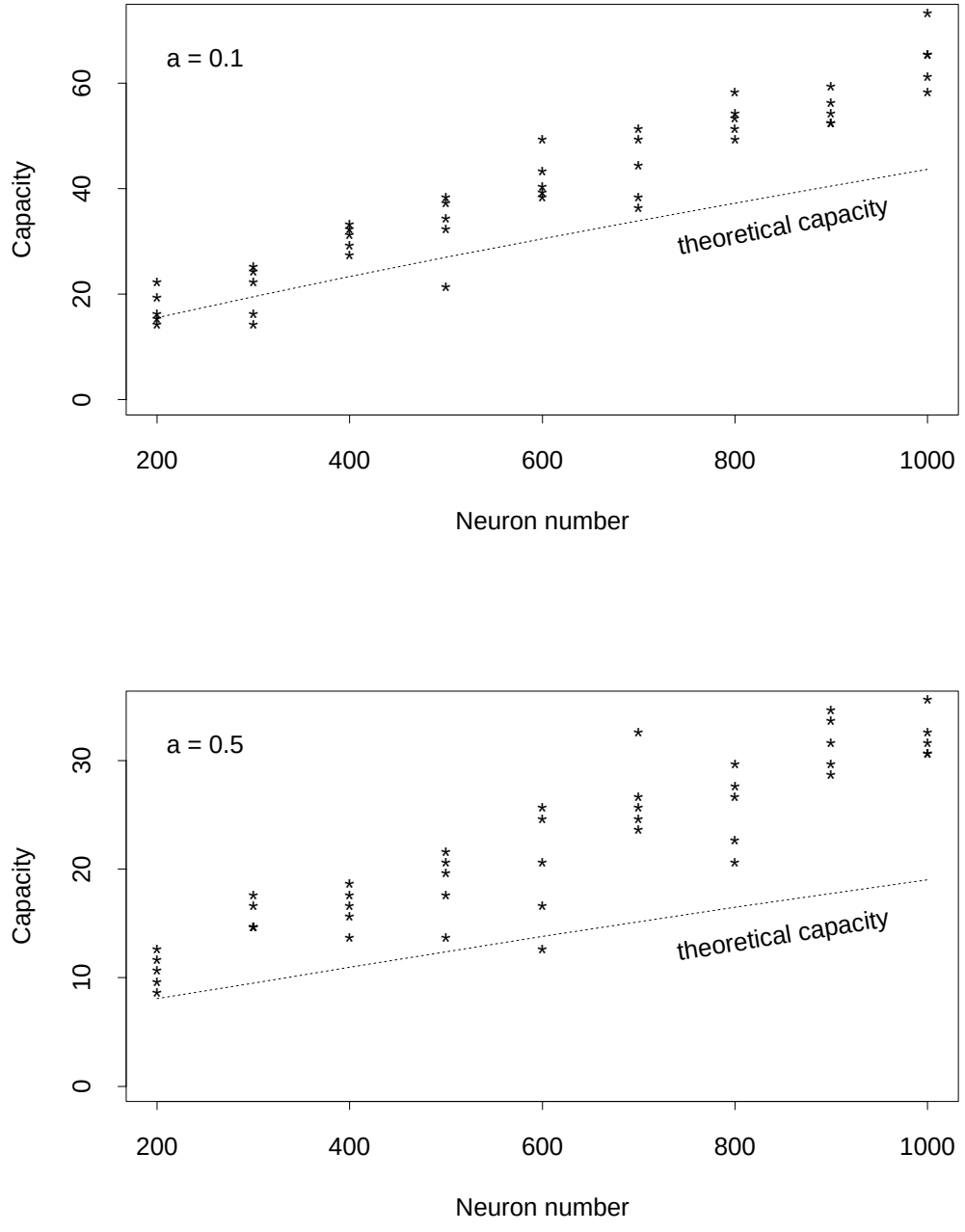

Figure 3: Capacity versus the number of neurons. we tried 5 simulations for each neuron number. Up: $a = 0.1$ (sparse case); Down: $a = 0.5$ (non-sparse case).

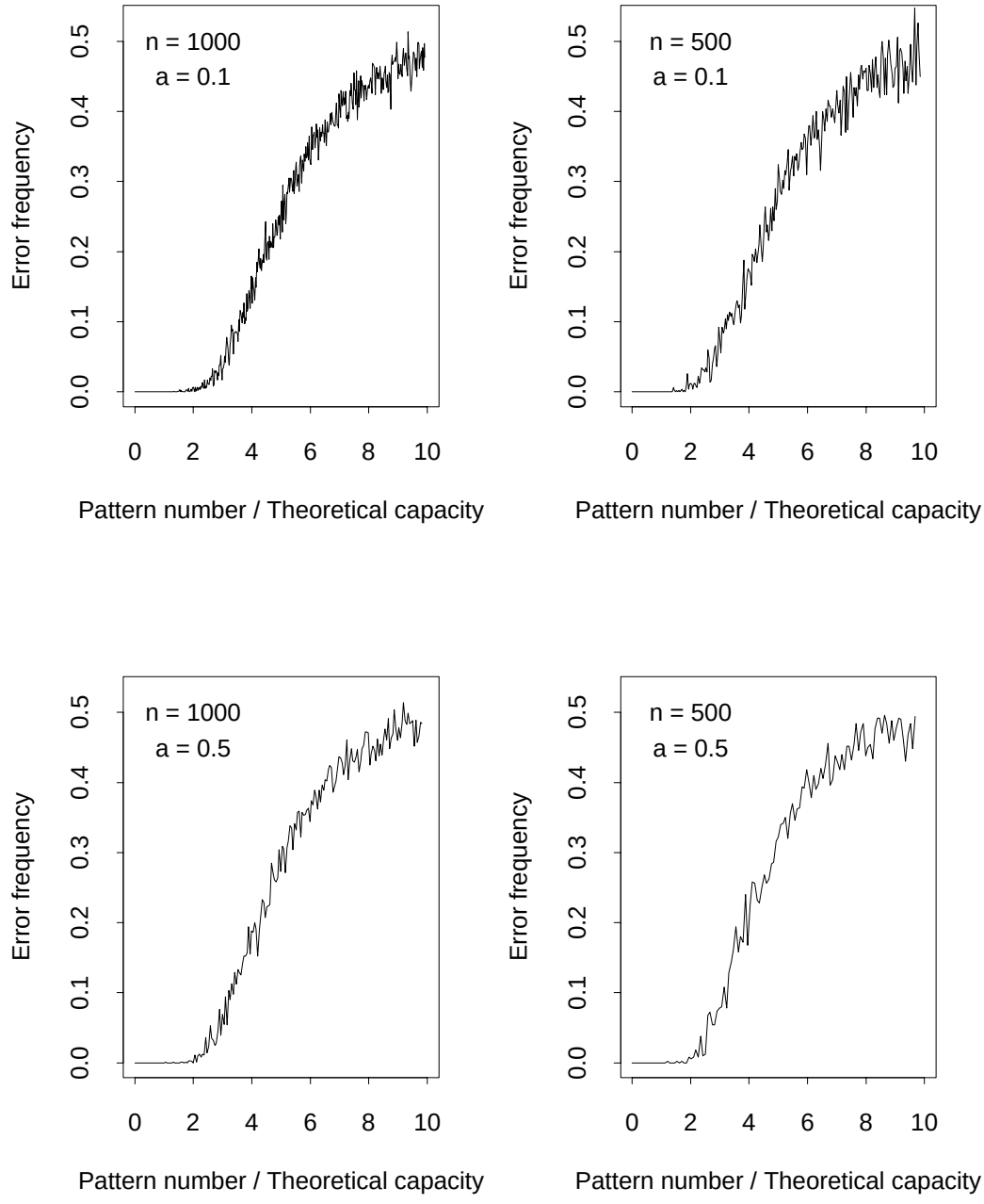

Figure 4: Frequency of error components versus pattern number. Pattern number is normalized by theoretical capacity. Neuron number: 500, 1000

6 Proof of theorems

In this section, we prove the theorems stated in the preceding sections, using the statistical neurodynamics method[3].

To begin with, we shall show that the capacity is $n/\log n$ order.

Lemma 1 *The optimal decay rate of an associative memory with n neurons is bounded asymptotically by*

$$1-\frac{8e(2+d)a(1-a)\,\log n}{n} \leq 1-\varepsilon_{\rm opt} \leq 1-\frac{8ea(1-a)\,\log n}{n} \tag{16}$$

and the memory capacity is bounded by

$$\frac{n}{16e(2+d)a(1-a)\,\log n} \leq M_{\rm opt} \leq \frac{n}{16ea(1-a)\,\log n}, \tag{17}$$

where $d=-\log_n a$.

Proof. Let us consider the recall of item $\boldsymbol{s}(t-m)$. For convenience, let $s_i^{(m)}$ denote $s_i(t-m)$. Since $s_i^{(m)}$ is a random variable with mean 0 and variance $a(1-a)$, a sum of the inputs for i-th neuron is

$$\frac{1}{n}\sum_{j=0}^{n} w_{ij}s_j^{(m)} = (1-\varepsilon)^m s_i^{(m)}(1-a)a + N_i \tag{18}$$

where[2]

$$N_i = \frac{1}{n}\sum_{\mu\neq m}\sum_{j\neq i}(1-\varepsilon)^{\mu} z_{ij}^{(\mu)} \tag{19}$$

$$z_{ij}^{(\mu)} = s_i^{(\mu)}s_j^{(\mu)}s_j^{(m)}. \tag{20}$$

$z_{ij}^{(\mu)}$ is a mutually independent variable with mean 0 and variance $(1-a)^3a^3$. By using the central limit theorem, N_i is asymptotically normally distributed,

$$N_i \sim N[0, \frac{1}{n}\{\frac{1}{\varepsilon}-(1-\varepsilon)^m\}(1-a)^3a^3] = N[0,\sigma^2], \tag{21}$$

where

$$\sigma^2 = \frac{1}{n}\{\frac{1}{\varepsilon}-(1-\varepsilon)^m\}(1-a)^3a^3 \simeq \frac{1}{n\varepsilon}(1-a)^3a^3. \tag{22}$$

Output value is given by

$$r_i = 1_a[(1-\varepsilon)^m s_i^{(m)}(1-a)a + N_i - h]. \tag{23}$$

This pattern is recalled correctly if

$$(1-\varepsilon)^m(1-a)^2a + N_i - h > 0 \tag{24}$$

[2]first term of right hand side of this equation is not exact and it must be multiplied by $1-u_j/n$ where u_j is an asymptotically normal variable with mean 0 and variance 1. However it is negligibly small, when n is large.

when $s_i^{(m)} = 1 - a$, and if

$$-(1-\varepsilon)^m(1-a)a^2 + N_i - h < 0 \tag{25}$$

when $s_i^{(m)} = -a$. We shall take the threshold value h at the midpoint between $(1-\varepsilon)^m(1-a)^2 a$ and $-(1-\varepsilon)^m(1-a)a^2$. Namely, h satisfies

$$h = (1-\varepsilon)^m(1-a)^2 a - A_m = -(1-\varepsilon)^m(1-a)a^2 + A_m, \tag{26}$$

hence, the value A_m becomes

$$A_m = \frac{(1-\varepsilon)^m(1-a)a}{2}. \tag{27}$$

Remark that the exact value of h cannot be given since it depends on unknown m. However, if it is determined such that na components of output vector become $1-a$, it satisfies (26) asymptotically [3].

Then the probability that $\boldsymbol{s}^{(m)}$ is recalled correctly is given by

$$p_m \equiv \{1 - \Phi(\frac{A_m}{\sigma})\}^n, \tag{28}$$

where Φ is the error integral function,

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-u^2/2}\, du. \tag{29}$$

Thus the probability P_M that $\boldsymbol{s}^{(0)}, \ldots, \boldsymbol{s}^{(M-1)}$ can be recalled correctly is

$$P_M = p_0 p_1 \cdots p_{M-1}. \tag{30}$$

Since $1 > p_0 > \cdots > p_{M-1}$, P_M is bounded by

$$(p_{M-1})^M \leq P_M \leq p_{M-1} \tag{31}$$

Let us evaluate $p_M (\simeq p_{M-1})$ at first. Using the fact that

$$\Phi(u) \simeq \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2} - \log u), \tag{32}$$

for large u, we can get

$$\begin{aligned} \log p_M &= n \log\{1 - \Phi(\frac{A_M}{\sigma})\} && (33)\\ &\simeq -n\Phi(\frac{A_M}{\sigma}) && (34)\\ &\simeq -\frac{1}{\sqrt{2\pi}} \exp\{\log n - \frac{1}{2}(\frac{A_M}{\sigma})^2 - \log\frac{A_M}{\sigma}\}. && (35) \end{aligned}$$

[3] This activity control can be done by a simple competitive network[2].

If we consider only the higher order of the capacity, the last term inside the exponential function of equation (35) is negligible. Substituting A_M and σ by their values, we get

$$\log p_M \simeq \exp\{\log n - \frac{1}{8(1-a)a} n\varepsilon(1-\varepsilon)^{2M}\}. \tag{36}$$

In order to p_M converges to 1 as n tends to infinity,

$$\log n \leq \frac{1}{8(1-a)a} n\varepsilon(1-\varepsilon)^{2M} \tag{37}$$

should be satisfied. Thus,

$$\begin{aligned} 2M &\leq \frac{1}{\log(1-\varepsilon)} \log\{8(1-a)a\frac{\log n}{n\varepsilon}\} &(38)\\ &\simeq \frac{1}{\varepsilon}\log\{\frac{n\varepsilon}{8(1-a)a\log n}\}. &(39)\end{aligned}$$

Using the fact that the function $\log(bx)/x$ takes maximal value b/e at $x = e/b$, we get the optimal value of ε is given by

$$\varepsilon = \frac{8e(1-a)a\log n}{n}, \tag{40}$$

and the capacity is given by

$$M = \frac{\log n}{16e(1-a)an}. \tag{41}$$

This provides the upper bound of the capacity.

On the other hand, we evaluate $(p_M)^M (\simeq (p_{M-1})^M)$. Analogously, we get

$$2M \leq \frac{1}{\varepsilon}\log\{\frac{n\varepsilon}{8(1-a)a\log(nM)}\}. \tag{42}$$

This equation can be solved asymptotically and if $a = n^{-d}$,

$$2M \leq \frac{1}{\varepsilon}\log\{\frac{n\varepsilon}{8(1-a)a(2+d)\log n}\}. \tag{43}$$

The optimal value of ε is given by

$$\varepsilon = \frac{8e(1-a)a(2+d)\log n}{n} \tag{44}$$

and the capacity is given by

$$M = \frac{\log n}{16e(1-a)a(2+d)n} \tag{45}$$

This provides the lower bound of the capacity. Thus the lemma has proved. □

Next, we shall prove the main theorem including higher order terms. Theorem 1 is presented as a corollary of this theorem.

Theorem 4 *When n is sufficiently large, the optimal decay rate of an associative memory with n neurons is given by*

$$1-\varepsilon_{\rm opt} = 1 - 8e(2+d-\frac{5\log\log n + C}{2\log n})\frac{a(1-a)\,\log n}{n}, \tag{46}$$

and the memory capacity is given by

$$M_{\rm opt} = \frac{1}{2\varepsilon_{\rm opt}} = \frac{1}{16e(2+d-\dfrac{5\log\log n + C}{2\log n})}\frac{n}{a(1-a)\,\log n}, \tag{47}$$

where $d = -\log_n a$, C is asymptotically constant.

Proof. We evaluate R, the expected number of incorrectly recalled components among M patterns. Since the probability that each component s_i^m is recalled incorrectly is given by $\Phi(A_m/\sigma)$ where Φ, A_m, σ is defined in the proof of lemma 1. Thus

$$R = n\sum_{m=0}^{M}\Phi(A_m/\sigma) \tag{48}$$

Approximating summation by integration,

$$R \simeq n\int_0^M \Phi(A_m/\sigma)\,dm, \tag{49}$$

and using the approximation (32), we get

$$R \simeq n\int_0^M \frac{1}{\sqrt{4\pi D}(1-\varepsilon)^m}\exp\{-D(1-\varepsilon)^{2m}\}\,dm. \tag{50}$$

where

$$D = \frac{n\varepsilon}{8(1-a)a}. \tag{51}$$

According to the result of lemma 1, we can let

$$\varepsilon = \frac{\varepsilon_0 a(1-a)\log n}{n}, \qquad M = \frac{M_0 n}{(1-a)a\log n} = \frac{\varepsilon_0 M_0}{\varepsilon}, \tag{52}$$

where ε_0 and M_0 are constant order values (see also (9),(10)). From (51) and (52), we get

$$D = \frac{\varepsilon_0\log n}{8}. \tag{53}$$

Since

$$(1-\varepsilon)^M \le (1-\varepsilon)^m \le 1, \qquad 0 \le m \le M \tag{54}$$

and $(1-\varepsilon)^M \simeq e^{-\varepsilon_0 M_0}$ is constant order, we can write

$$R \simeq \frac{n}{\sqrt{4\pi D}f}\int_0^M \exp\{-D(1-\varepsilon)^{2m}\}\,dm, \tag{55}$$

where f is a constant order variable bounded by

$$(1-\varepsilon)^M \leq f \leq 1 \tag{56}$$

Then

$$R = \frac{n}{\sqrt{4\pi D} f} \frac{\mathrm{Ei}(-D(1-\varepsilon)^{2M}) - \mathrm{Ei}(-D)}{2\log(1-\varepsilon)} \tag{57}$$

where Ei is the exponential integral function

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t}\, dt, \tag{58}$$

and when x is sufficiently large,

$$\mathrm{Ei}(-x) \simeq -\frac{e^{-x}}{x}. \tag{59}$$

Using this fact we get

$$R \simeq \frac{n}{4\sqrt{\pi} f \varepsilon D^{3/2}} \{ \frac{e^{-D(1-\varepsilon)^{2M}}}{(1-\varepsilon)^{2M}} - e^{-D} \}. \tag{60}$$

Because from (53) D is $\log n$ order that goes to infinity as n increases and $(1-\varepsilon)^{2M}$ is constant order,

$$\frac{e^{-D(1-\varepsilon)^{2M}}}{(1-\varepsilon)^{2M}} \gg e^{-D}. \tag{61}$$

We shall find the condition that

$$R \leq \delta, \tag{62}$$

for some δ, which denotes the expected number of error components asymptotically. From the equation

$$\frac{n}{4\sqrt{\pi} f \varepsilon D^{3/2} (1-\varepsilon)^{2M}} e^{-D(1-\varepsilon)^{2M}} = \delta, \tag{63}$$

we get

$$2M = \frac{1}{\log(1-\varepsilon)} \log \frac{\dfrac{n}{4\sqrt{\pi} f \varepsilon D^{3/2} (1-\varepsilon)^{2M} \delta}}{D} \tag{64}$$

$$\simeq -\frac{1}{\varepsilon} \log \frac{\dfrac{n}{4\sqrt{\pi} f \varepsilon D^{3/2} (1-\varepsilon)^{2M} \delta}}{D}. \tag{65}$$

Thus

$$2M_0 = \frac{1}{\varepsilon_0} \log \frac{\varepsilon_0}{8(2 + d - \dfrac{5\log\log n + C}{2\log n})}, \tag{66}$$

where C is asymptotically constant. ε_0 that maximizes M_0 is given by

$$\varepsilon_0 = 8e(2 + d - \frac{5\log\log n + C}{2\log n}), \tag{67}$$

and M_0 is given by

$$M_0 = \frac{1}{2\varepsilon_0} = \frac{1}{16e(2+d-\dfrac{5\log\log n + C}{2\log n})}. \tag{68}$$

These values ε_0, M_0 are the values of theorem 4 and they converges to the values of theorem 1 very slowly. □

Asymptotically constant value C in the proof above converges to

$$C = 2\log\{32(2+d)^{5/2}\sqrt{\pi}e^{3/2}(1-a)f\delta\}. \tag{69}$$

as n tends to infinity, where f is bounded by

$$-\frac{1}{2} \leq \log f \leq 0. \tag{70}$$

from (56). However, the convergence rate of C is so slow as M_0 and ε_0 tends to the values of theorem 1.

Next, we shall prove theorem 3 that includes theorem 2 (the case of noise zero).

Proof of theorem 3. For the noisy pattern $\hat{s}_j^{(m)}$ defined in (13), a sum of the inputs for i-th neuron is

$$\frac{1}{n}\sum_{j=0}^{n} w_{ij}\hat{s}_j^{(m)} = (1-\varepsilon)^m s_i^{(m)}(1-a-\xi)a + N_i, \tag{71}$$

where N_i is a random variable that behaves just the same as (19). Thus the error frequency for $\hat{\boldsymbol{s}}$ becomes

$$r(\varepsilon_0, m_0) = \Phi(A'_m/\sigma), \tag{72}$$

where σ is defined by (22) and

$$A'_m = \frac{(1-\varepsilon)^m a(1-a-\xi)}{2}, \tag{73}$$

(cf. (27)). Substituting variables by their values ((9),(10)) and using the approximation $(1+1/x)^x \simeq e$ when x is large, we get

$$r(\varepsilon_0, m_0) \simeq \Phi\left(\sqrt{\frac{\varepsilon_0 e^{-\varepsilon_0 m_0}(1-a-\xi)^2}{2(1-a)^2}\log n}\right). \tag{74}$$

Then from (32), we get the formula of the theorem. □

7 Conclusion

We presented the optimal decay rate of associative memory model that maximizes the memory capacity, and also we showed some simulation in order to ascertain the theoretical results.

Mézard et al. got a similar result using the replica method under the assumption of replica-symmetry[8], whose reliability is not assured enough by the mathematical analysis.

In learning from examples such as neural network learning, the most necessary property is the generalization rather than the capacity. As a future subject, we will analyze not only associative memory model but also more general type neural networks, and estimate the optimal weight decay in terms of the generalization property.

Acknowledgement

The author would like to thank K.Tamura, Director of Information Science Division of ETL, for affording an opportunity of this study. He is also deeply indebted to K.Kurata in Osaka University for the fruitful discussions with him. He also expresses his thanks to all members of Mathematical Informatics Section of ETL for their helpful discussions.

References

[1] S. Amari: Characteristics of sparsely encoded associative memory. *Neural Networks*, Vol. 2, No. 6, pp. 451–457, 1989.

[2] S. Amari and M.A. Arbib: Competition and cooperation in neural nets. In J. In Metzler (ed.), *Systems neuroscience*. Academic Press, New York, 1977.

[3] S. Amari and K. Maginu: Statistical neurodynamics of associative memory. *Neural Networks*, Vol. 1, No. 1, pp. 63–73, 1988.

[4] J.A. Anderson: A simple neural network generating interactive memory. *Mathematical Biosciences*, Vol. 14, pp. 197–220, 1972.

[5] T.M. Cover: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.*, Vol. 14, pp. 326–334, 1965.

[6] T. Kohonen: Correlation matrix memories. *IEEE Trans. Comput.*, Vol. 21, pp. 343–359, 1972.

[7] C. Meunier, H. Yanai, and S. Amari: Sparsely coded associative memories: capacity and cynamical properties. *Network*, Vol. 2, pp. 469–487, 1991.

[8] M. Mézard, J.P. Nadal, and G. Toulouse: Solvable models of working memories. *J. Physique*, Vol. 47, pp. 1457–1462, 1986.

[9] Y. Miyashita and H.s. Chang: Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, Vol. 331, pp. 68–70, 1988.

[10] K. Nakano: Associatron — a model of associative memory. *IEEE Trans. Sys. Man Cybern.*, Vol. 2, pp. 381–388, 1972.

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

[12] Y. Uesaka and K. Ozeki: Some properties of associative type memories. *J. IEICE Japan*, Vol. 55-D, pp. 323–330, 1972.

[13] D.J. Willshaw and Longuest-Higgins H.C.: Associative memory models. In B. Meltzer and O. Michie (eds.), *Machine Intelligence,* Vol. 5. Edinburgh University Press, 1970.