Capacity and Error Correction Ability of Sparsely Encoded Associative Memory with Forgetting Process

Shotaro Akaho

Electrotechnical Laboratory,
Mathematical Informatics Section,
1–1–4 Umezono, Tsukuba-shi, Ibaraki 305 Japan

Abstract Associative memory model of neural networks can not store items more than its memory capacity. When new items are given one after another, its connection weights should be decayed so that the number of stored items does not exceed the memory capacity (*forgetting process*). This paper analyzes the sparsely encoded associative memory, and presents the optimal decay rate that maximizes the number of stored items. The maximal number of stored items is given by $O(n/a \log n)$ when the decay rate is $1 - O(a \log n/n)$, where the network consists of n neurons with activity a.

1 Introduction

Neural network is an adaptive system that is trained with sample items given from outer environment, and it can store items up to its memory capacity.

In this paper, we consider the on-line type learning scheme where the learning is proceeded every time when a new item is provided (cf. batch type learning). This learning has the advantage that a little memory is needed (not necessary to memorize all patterns) and the network can adapts itself well to the change of the environment. However, when many new items are given one after another, it cannot be judged whether the number of stored items is more than its capacity or not, because the stored items are memorized implicitly on connection weights and the network does not memorize 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. We analyze the associative memory model (Anderson, J.A., 1972) as an example.

Recently the associative memory model 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. If patterns are sparsely encoded, the capacity of network becomes very large (Amari, S., 1989). Sparse coding scheme is also supposed to act an important function in the brain memory such as hippocampus, and some important experimental results are coming out.

2 Associative Memory with Forgetting Process

We consider the autocorrelation associative memory with n input units and n output units. It is trained by a simple Hebbian rule as follows.

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

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

where w_{ij} is a connection weight from j-th input to i-th output units, $s_i(t)$ is i-th component of pattern learned at time t, ε and c are time constants (we assume $c = 1$ without loss of 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 (Amari, S., 1989). 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 a binary value independently according to the probability above. 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

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

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 Optimal Decay Rate and the Capacity

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

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

holds, pattern $\boldsymbol{s}(t-\mu)$ is said to be stored. Memory capacity $M(\varepsilon; n, a)$ for given n, a and ε 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 probability 1 asymptotically for sufficiently large n.

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 the finite number of patterns without decay (batch type learning). In general, the capacity of this model is $1/2e$ times the capacity in the batch type learning.

Next, we investigate the error correction ability of the learning. 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.

$$\begin{array}{cccccc} \overbrace{\underbrace{1-a\ 1-a}_{na\xi}\cdots\ 1-a}^{na} & & & \overbrace{\underbrace{-a\ \ -a}_{na\xi}\cdots\ -a}^{n(1-a)} & & : \text{original pattern} \\ \Downarrow & & \Downarrow & \Downarrow & & \\ -a\ \ -a & & 1-a & 1-a\ 1-a & -a & : \text{noisy pattern} \end{array} \tag{9}$$

Theorem 2 *The maximal number of item patterns whose ξ noisy pattern is corrected is given by*

$$m(\xi) = \frac{(1-a-\xi)^2}{(1-a)^2} M_{\rm opt}, \tag{10}$$

when the decay rate is taken as

$$\varepsilon(\xi) = \frac{(1-a)^2}{(1-a-\xi)^2} \varepsilon_{\rm opt}, \tag{11}$$

where $M_{\rm opt}$ and $\varepsilon_{\rm opt}$ are defined in theorem 1.

Theorem 1 and 2 are proved basically by a similar technique to Amari's (Amari, S., 1989). In the analysis, we only pay attention that the probability that the pattern is recalled correctly is different for each pattern. On the detail of the proof of theorems, see (Akaho, S., 1992).

4 Simulation Results

We show some simulation results.

The following figure shows capacities for some values of ε normalized so that the optimal one becomes 1.0, where we gave the network ten times patterns as many as the theoretical capacity. The capacity is measured as the most recent pattern incorrectly recalled. We can see that the decay rate that maximizes the capacity is close to the theoretical value for all cases.

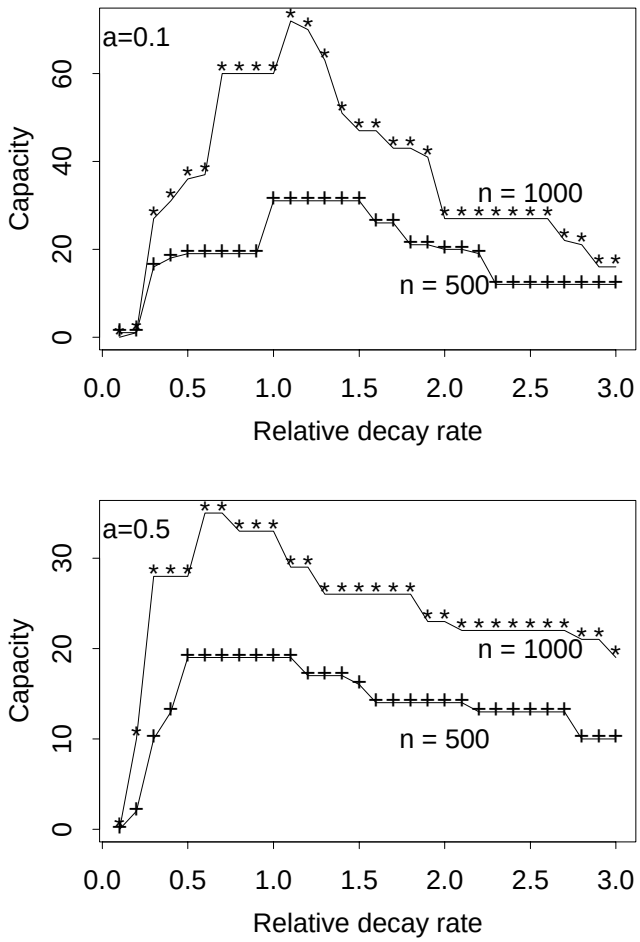

Figure: Capacity versus decay rate (ε). Decay rate is normalized by the optimal one. Up: a=0.1 (sparse case); Down: a=0.5 (non-sparse case)

5 Concluding Remarks

First of all, we give some remarks for the preceded sections.

We analyzed the case of autocorrelation associative memory for the simplicity, but the result is the same for the crosscorrelation case. In the later case, the activity can be different between input patterns and output patterns. We can show that the capacity and the decay rate does not depend on the activity of input. Thus only if output is sparsely encoded, the network has a large capacity, whether the input is sparsely encoded or not (but it is necessary that the activity of input is fixed). Moreover, the input does not need to be a binary pattern if the activity is fixed.

Some paper says that the capacity of associative memory is $O(n/\log n)$ and some paper says that it is $O(n)$ (the present paper is the former one; cf. (Mézard, M. and Nadal, J.P. and Toulouse, G., 1986)). This difference is caused by the difference of the definition of the capacity. While the former permits just $O(1)$ error components among n units, the later permits $O(n)$ error components. They are almost the same for a practical size of n.

Next, let us refer to the future problems. We have given attention only to the capacity as the measure of the network ability. However, there are two other aspects of the network as a learning machine. One is the adaptation ability for the change of environments (e.g. $f(\boldsymbol{x}) = y_1$ in the past but $f(\boldsymbol{x}) = y_2$ $(\neq y_1)$ at present), which the on-line type learning is expected to have. The other is the generalization ability from the finite number of samples, i.e. the capability of describing unknown data only from the given sample data. (e.g. a sample indicates $f(\boldsymbol{x}_1) = y$, hence $f(\boldsymbol{x}_2)$ may be also y, where $\boldsymbol{x}_2 \simeq \boldsymbol{x}_1$).

Those three aspects are related to each other. Generalization theory of learning (such as VC dimension (Vapnik, V.A., 1984)) teaches us that the generalization ability of the network becomes higher, as the number of samples increases and the capacity of the network decreases. However it is the result induced under the situation that the environment does not change. If the environment changes gradually with time, old samples are not so reliable as newer items. Thus we should assign weights according to the "age" of each sample to ignore old samples, which corresponds to the forgetting process in this paper.

Thus new problem can be stated as follows:

Future Problem How should we assign the optimal weights for samples so that the network can achieve the desired generalization ability?

In order to solve this problem, we will have to synthesize and develop the generalization theory of learning and the adaptive control theory.

References

Akaho, S.(1992). Optimal decay rate of connection weights in covariance learning. Technical Report 92–37, Electrotechnical Laboratory.

Amari, S.(1989). Characteristics of sparsely encoded associative memory. *Neural Networks*, **2**, 451–457.

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

Mézard, M., Nadal, J., and Toulouse, G.(1986). Solvable models of working memories. *J. Physique*, **47**, 1457–1462.

Vapnik, V.(1984). *Estimation of Dependences Based on Empirical Data.* Springer-Verlag.