Journal of the Physical Society of Japan
Vol. 77, No. 12, December, 2008, 124002
©2008 The Physical Society of Japan

Multiple Stability of a Sparsely Encoded Attractor Neural Network Model for the Inferior Temporal Cortex

Tomoyuki Kimoto*, Tatsuya Uezu[1], and Masato Okada[2,3]

Oita National College of Technology, Oita 870-0152
[1]*Graduate School of Sciences and Humanities, Nara Women's University, Nara 630-8506*
[2]*Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561*
[3]*RIKEN Brain Science Institute, Wako, Saitama 351-0198*

(Received May 10, 2008; accepted September 26, 2008; published November 25, 2008)

We study a neural network model for the inferior temporal cortex, in terms of finite memory loading and sparse coding. We show that an uncorrelated Hopfield-type attractor and some correlated attractors have multiple stability, and examine the retrieval dynamics for these attractors when the initial state is set to a noise-degraded memory pattern. Then, we show that there is a critical initial overlap: that is, the system converges to the correlated attractor when the noise level is large, and otherwise to the Hopfield-type attractor. Furthermore, we study the time course of the correlation between the correlated attractors in the retrieval dynamics. On the basis of these theoretical results, we resolve the controversy regarding previous physiologic experimental findings regarding neuron properties in the inferior temporal cortex and propose a new experimental paradigm.

KEYWORDS: associative memory model, attractor, multiple stability, inferior temporal cortex, stimulus selectivity
DOI: 10.1143/JPSJ.77.124002

1. Introduction

Despite various studies, many questions remain concerning the neurophysiologic properties of the inferior temporal cortex, which is believed to be the final stage of vision. Miyashita and Chang clarified that visual stimuli are memorized within the inferior temporal cortex.[1] In further study, Miyashita had a monkey learn 97 fractal patterns repeated in the same sequence, and then presented the same fractal patterns to the monkey after learning while measuring the firing rate patterns of neurons in the monkey's inferior temporal cortex. He found that when the monkey is presented two fractal patterns that are nearer to each other in the training sequential order, the correlation between corresponding firing rate patterns becomes higher.[2] Thus, they concluded that time correlation regarding the learning sequential orders of visual stimuli is converted to spatial correlation regarding the firing rate patterns of the neuron group.

Griniasty *et al.* and Amit *et al.* proposed an attractor neural network model to explain Miyashita's physiological finding.[3,4] Their models have attractors that are mixtures of consecutively learned memory patterns. The correlation graph between the attractor and consecutively learned memory patterns takes the shape of a Gaussian distribution. In other words, the attractor has a high correlation with one particular memory pattern, and the correlation with other memory patterns gradually becomes smaller. Such an attractor is called a *correlated attractor*. Here, we explain the correlated attractor in detail. When a memory pattern is set in an initial state, the retrieved correlated attractor is highly correlated with the presented memory pattern, and the retrieved correlated attractor changes depending on the presented memory pattern. Furthermore, when examining the correlation between two retrieved correlated attractors, the nearer the learning sequential orders of the presented memory patterns are, the larger the correlation between the retrieved correlated attractors is. They explained Miyashita's physiological finding using these properties. Their models have also been studied by other researchers.[5,6]

These models have a property showing that memory patterns as well as correlated attractors can become the equilibrium state by adjusting the learning parameter. Furthermore, the correlated attractor and memory pattern can become either a monostable state or a bistable state. Clarifying such a complex property could lead to clues that will help us elucidate the physiologic function of the brain.

Previously, the models were studied under the condition of a 50% firing rate. However, it is thought that sparse coding is used in the actual brain based on a physiologic finding and a theoretical viewpoint concerning the brain.[2,7–11] We therefore adopt sparse coding and add a firing rate control operation to the attractor neural network model, and study the stability and retrieval dynamics for memory patterns and correlated attractors. We analyzed these properties using statistical mechanics and computer simulations. Moreover, we interpret the findings of certain physiological experiments concerning neurons in the inferior temporal cortex on the basis of the theoretical results obtained here, and propose a new experimental paradigm.

2. Model

Our model consists of two layers, the associative layer and firing rate control layer, as shown in Fig. 1. The associative layer is a mutual connection network consisting of N neurons, each of which can take either of two states (0 and 1). The firing rate control layer is a feedback loop to control the firing rate of the neurons in the associative layer. For the model of the 50% firing rate, no such firing rate control is necessary to retrieve the memory patterns. In the sparse coding scheme, however, the firing rate control explained in the below is necessary to retrieve memory patterns, as shown by Okada.[11]

*E-mail: kimoto@oita-ct.ac.jp

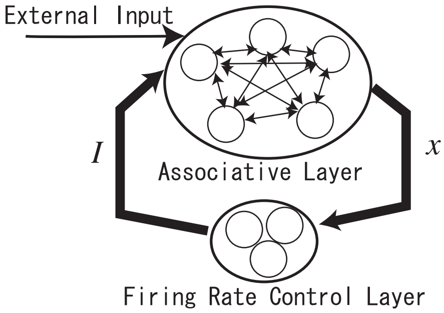


Fig. 1. Model.

First, we explain the neurons in the associative layer. Asynchronous updating at a finite temperature is used for the dynamics of associative layer neurons. In asynchronous updating, one neuron is chosen at random from N neurons. For example, let us assume that the ith neuron is chosen. The internal state u_i of the ith neuron is calculated as

$$u_i^{t+1} = \sum_{j\neq i}^{N} J_{ij} x_j^t + h - gI, \tag{1}$$

where J_{ij} is the connection weight from the jth neuron to the ith neuron, and h is the threshold value. I is the input from the firing rate control layer, and g is a positive constant. The probability that the state x_i of the ith neuron becomes 1 is decided by using u_i as

$$\begin{aligned}\mathrm{Prob}[x_i^t = 1] &= 1 - \mathrm{Prob}[x_i^t = 0]\\ &= \frac{1+\tanh(\beta u_i^t)}{2},\end{aligned} \tag{2}$$

where $\beta = 1/T$ depends on the temperature T. Repeating this procedure N times is one Monte Carlo step.

Next, we consider the neurons in the firing rate control layer. The output I of neurons in the firing rate control layer is determined using

$$I = \frac{1}{N}\sum_{j=1}^{N} x_j^t - F. \tag{3}$$

When the firing rate $(1/N)\sum_{j=1}^{N} x_j^t$ in the associative layer becomes greater than F, I becomes positive, and the firing rate in the associative layer becomes lower at a time $t+1$. In contrast, when the firing rate $(1/N)\sum_{j=1}^{N} x_j^t$ in the associative layer becomes smaller than F, it has the opposite effect. Thus, the firing rate control layer serves to make the firing rate of the associative layer close to F.

The memory patterns stored in the model are N-dimensioned vectors consisting of the binary values 0 and 1. The probability that each element η_i^μ of the memory pattern $\boldsymbol{\eta}^\mu$ takes 1 is assumed to be F as the firing rate of the associative layer.

$$\mathrm{Prob}[\eta_i^\mu = 1] = 1 - \mathrm{Prob}[\eta_i^\mu = 0] = F. \tag{4}$$

Therefore, the mean value $\mathrm{E}[\eta_i^\mu]$ of the memory pattern $\boldsymbol{\eta}^\mu$ also becomes F. Reducing this firing rate F results in a sparse coding. The connection weight J_{ij} from the jth neuron to the ith neuron is determined using

$$J_{ij} = \frac{1}{VN}\sum_{\mu,\nu=1}^{s} (\eta_i^\mu - F) A_{\mu\nu} (\eta_j^\nu - F), \tag{5}$$

$$A_{\mu\nu} = \delta_{\mu\nu} + a\delta_{\mu-1,\nu} + a\delta_{\mu+1,\nu}, \tag{6}$$

where s is the number of memory patterns, V is a constant shown in eq. (8), a is a correlation learning coefficient between adjoined memory patterns, and δ is Kronecker's delta. We assume that $J_{ii} = 0$.

The overlap m_μ between the retrieval state $\boldsymbol{x}$ and the memory pattern $\boldsymbol{\eta}^\mu$ is defined as

$$m_\mu = \frac{1}{VN}\sum_{i=1}^{N} (\eta_i^\mu - F)\langle x_i \rangle, \tag{7}$$

$$V = F(1-F), \tag{8}$$

where V is a normalization constant for making $m_\mu = 1$ when the retrieval state $\boldsymbol{x}$ completely corresponds to the memory pattern $\boldsymbol{\eta}^\mu$. $\langle x_i \rangle$ is the thermal mean value in the state x_i. The mean value of state $\boldsymbol{x}$ is defined as

$$M = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{9}$$

3. Theory

To examine the property of the equilibrium state, we discuss the model in terms of statistical mechanics with regard to the following Hamiltonian obtained from eqs. (1), (3), (5), and (6). In this paper, we consider the case where the number N of neurons is infinitely large, and the number s of memory patterns is $O(1)$ irrespective of N.

$$\begin{aligned} H = &-\frac{1}{VN}\sum_{i\neq j}^{N}\sum_{\mu,\nu=1}^{s} (\eta_i^\mu - F) A_{\mu\nu} (\eta_j^\mu - F) x_i x_j \\ &- 2h\sum_{i=1}^{N} x_i + \frac{g}{N}\sum_{i\neq j}^{N} x_i x_j - 2gF\sum_{i=1}^{N} x_i, \end{aligned} \tag{10}$$

$$Z = \mathrm{Tr}_{\boldsymbol{x}} \exp(-\beta H), \tag{11}$$

$$f = -\frac{1}{\beta VN}\log Z, \tag{12}$$

where the Hamiltonian of eq. (10) is twice that of the ±1 spin model. In the 0, 1 spin model that this model uses, since the amount of change in the Hamiltonian caused by a flip of one spin is only half of that with the ±1 spin model, the factor 2 is necessary to be consistent with the probability of eq. (2).

By using a saddle-point method, the free energy f of eq. (12) becomes

$$\begin{aligned} f = &\sum_{\mu,\nu=1}^{s} m_\mu A_{\mu\nu} m_\nu \\ &- \frac{1}{\beta V}\left\langle\!\!\left\langle \log\left(\exp\left\{ 2\beta\left[\sum_{\mu,\nu=1}^{s} m_\mu A_{\mu\nu} (\eta_i^\nu - F) + h - g(M-F) \right]\right\} + 1 \right)\right\rangle\!\!\right\rangle, \end{aligned} \tag{13}$$

and by using $\partial f/\partial m_\mu = 0$ and $\partial f/\partial M = 0$, the order parameter equations that determine m_μ and M that minimize the free energy f become

$$m_\mu = \frac{1}{V}\left\langle\!\!\left\langle(\eta_i^\mu - F)\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle, \tag{14}$$

$$M = \left\langle\!\!\left\langle\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle, \tag{15}$$

$$u_i = \sum_{\nu,\nu'=1}^{s}(\eta_i^\nu - F)A_{\nu\nu'}m_{\nu'} + h - g(M-F), \tag{16}$$

where $\langle\langle\cdots\rangle\rangle$ stands for the average over concerning the memory patterns $\{\boldsymbol{\eta}^\mu\} = \boldsymbol{\eta}^1,\ldots,\boldsymbol{\eta}^s$. The physical meaning of m_μ is the overlap shown in eq. (7). The physical meaning of M is the mean value of state $\boldsymbol{x}$ shown in eq. (9).

Let us assume that the system takes a state $\boldsymbol{x}$ with a probability $p_t(\boldsymbol{x})$ at a time t, and an element x_i flips from $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots, x_N)$ to $F_i\boldsymbol{x} = (x_1, x_2, \ldots, \tilde{x}_i, \ldots, x_N)$,

$$\tilde{x}_i = \begin{cases} 1 & x_i = 0 \\ 0 & x_i = 1 \end{cases}, \tag{17}$$

with a probability $w_i(\boldsymbol{x})$. Here, F_i is a spin flip operator that changes the ith element x_i to $\tilde{x}_i$. The master equation describing the dynamics of $p_t(\boldsymbol{x})$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{x}) = \sum_{i=1}^{N}[p_t(F_i\boldsymbol{x})w_i(F_i\boldsymbol{x}) - p_t(\boldsymbol{x})w_i(\boldsymbol{x})]. \tag{18}$$

One unit time for t in the above equation corresponds to 1 Monte Carlo step for eq. (2). In the equilibrium state, $p_t(\boldsymbol{x}) = p_{\mathrm{eq}}(\boldsymbol{x})$ and the above equation becomes $(\mathrm{d}/\mathrm{d}t)p_{\mathrm{eq}}(\boldsymbol{x}) = 0$; then the detailed balanced equation is described as

$$\frac{p_{\mathrm{eq}}(F_i\boldsymbol{x})}{p_{\mathrm{eq}}(\boldsymbol{x})} = \frac{w_i(\boldsymbol{x})}{w_i(F_i\boldsymbol{x})}. \tag{19}$$

Equation (19) is satisfied by assuming that the transition probability $w_i(\boldsymbol{x})$ takes the form

$$w_i(\boldsymbol{x}) = \frac{1-(2x_i-1)\tanh(\beta u_i)}{2}, \tag{20}$$

$$u_i = \sum_{j\neq i}^{N}J_{ij}x_j + h - g(M-F). \tag{21}$$

Next, we derive the theoretical equation of retrieval dynamics in the associative layer. Since the equilibrium state is described by the overlap m_μ and the mean value M of the state, the probability that the order parameters take the values $\{m_\mu\} = m_1, \ldots, m_s$ and M at a time t is given as

$$\mathcal{P}_t(\{m_\mu\}, M) = \mathrm{Tr}_{\boldsymbol{x}}\, p_t(\boldsymbol{x})\delta(M - M(\boldsymbol{x})) \times \prod_{\mu=1}^{s}\delta(m_\mu - m_\mu(\boldsymbol{x})), \tag{22}$$

$$m_\mu(\boldsymbol{x}) = \frac{1}{VN}\sum_{i=1}^{N}(\eta_i^\mu - F)x_i, \tag{23}$$

$$M(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}x_i. \tag{24}$$

The amount of change in $\mathcal{P}_t(\{m_\mu\}, M)$ in eq. (22) becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_t(\{m_\mu\}, M) = {}& \sum_{\nu=1}^{s}\frac{\partial}{\partial m_\nu}\mathcal{P}_t(\{m_\mu\}, M) \\ & \times\left[m_\nu - \frac{1}{V}\left\langle\!\!\left\langle(\eta_i^\nu - F)\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle\right] \\ & + \frac{\partial}{\partial M}\mathcal{P}_t(\{m_\mu\}, M) \\ & \times\left[M - \left\langle\!\!\left\langle\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle\right],\end{aligned} \tag{25}$$

$$u_i = \sum_{\nu',\nu''=1}^{s}(\eta_i^{\nu'} - F)A_{\nu'\nu''}m_{\nu''} + h - g(M-F). \tag{26}$$

On the other hand, since the state of the model can be described using only the overlap $\{m_\mu\}$ and M, even in a transitional state, $\mathcal{P}_t(\{m_\mu\}, M)$ can be described as

$$\mathcal{P}_t(\{m_\mu\}, M) = \delta(M - M(t))\prod_{\mu=1}^{s}\delta(m_\mu - m_\mu(t)), \tag{27}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}_t(\{m_\mu\}, M) = {}& \sum_{\nu=1}^{s}\frac{\partial}{\partial m_\nu}\mathcal{P}_t(\{m_\mu\}, M)\left(-\frac{\mathrm{d}m_\nu}{\mathrm{d}t}\right) \\ & + \frac{\partial}{\partial M}\mathcal{P}_t(\{m_\mu\}, M)\left(-\frac{\mathrm{d}M}{\mathrm{d}t}\right).\end{aligned} \tag{28}$$

By comparing eqs. (25) and (28), the evolution equations for the order parameter m_μ and M are obtained as

$$\frac{\mathrm{d}}{\mathrm{d}t}m_\mu = -m_\mu + \frac{1}{V}\left\langle\!\!\left\langle(\eta_i^\mu - F)\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle, \tag{29}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}M = -M + \left\langle\!\!\left\langle\frac{1+\tanh(\beta u_i)}{2}\right\rangle\!\!\right\rangle, \tag{30}$$

$$u_i = \sum_{\nu,\nu'=1}^{s}(\eta_i^\nu - F)A_{\nu\nu'}m_{\nu'} + h - g(M-F). \tag{31}$$

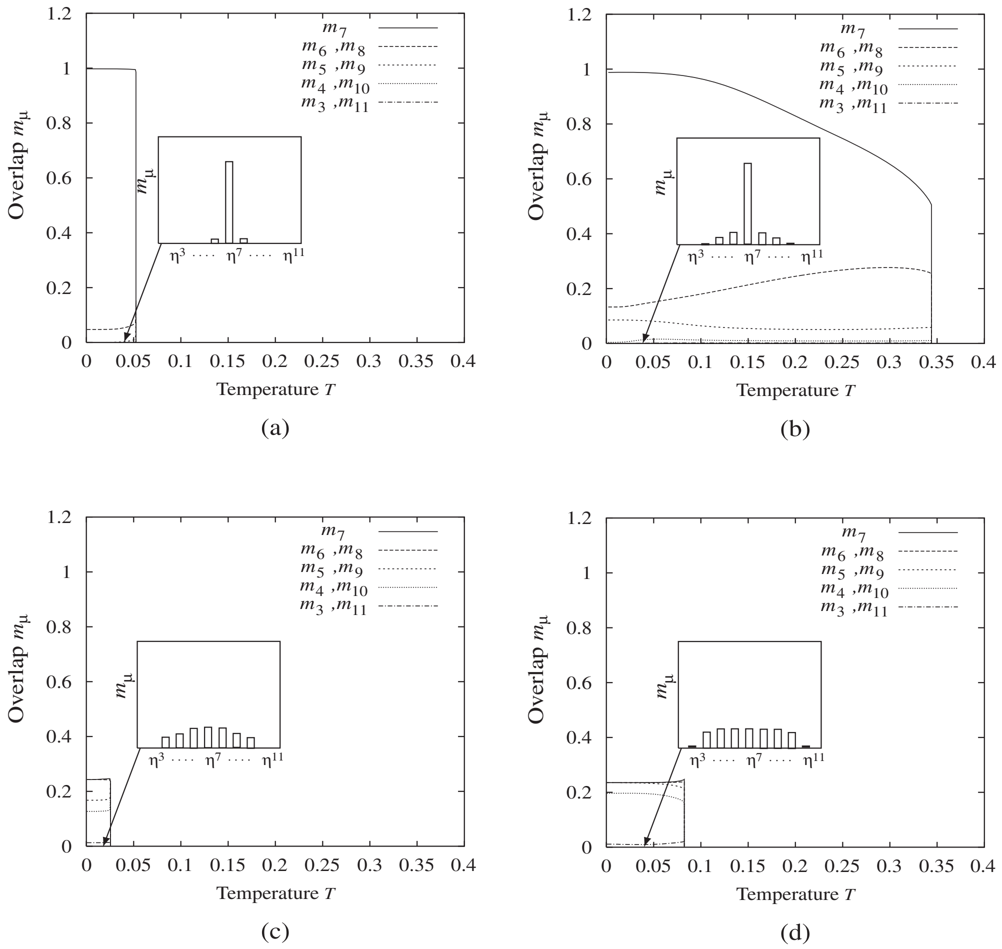
The stationary state of eq. (29) agrees with eq. (14), which describes the equilibrium state, and the stationary state of eq. (30) similarly agrees with eq. (15).

4. Results

4.1 *Equilibrium state of model*

We next examine the equilibrium state of the model, with the parameters set as $F = 0.05$, $s = 13$, $a = 0.7$, $h = -0.7$, and $g = 10$. We chose the feedback coefficient g and the threshold value h so that it would allow the memory pattern to reach the equilibrium state. For these parameters, we examined the overlap m^μ $(1 \le \mu \le s)$ while increasing temperature T. We found that four kinds of attractor were relevant, and these are shown in Fig. 2 where the horizontal axis is the temperature T and the vertical axis is the overlap m^μ $(3 \le \mu \le 11)$. To make the spread of the overlap m^μ $(3 \le \mu \le 11)$ clearer, bar charts are also shown in Fig. 2. The bar charts in Figs. 2(a), 2(b), and 2(d) show the spread of the overlap at $T = 0.04$, and the bar chart in Fig. 2(c) shows the spread of the overlap at $T = 0.02$. The equilibrium state in Fig. 2(a) has a large overlap only with $\boldsymbol{\eta}^7$, while the equilibrium states in the other cases have an overlap spread around $\boldsymbol{\eta}^7$. The number of attractors changes depending on temperature: four kinds of attractor are stable when the temperature is low $(T < 0.02)$, and only one attractor in Fig. 2(b) is stable when the temperature is high $(T > 0.09)$.

We will refer to the attractor that has a large overlap with only one memory pattern [Fig. 2(a)] as a *Hopfield attractor*



(a)

(b)

(c)

(d)

Fig. 2. Four kinds of attractors with multiple stability: (a) Hopfield attractor, (b) correlated attractor 1, (c) correlated attractor 2, and (d) correlated attractor 3.

and to the attractors that have nonzero overlaps with some memory patterns [Figs. 2(b)–2(d)] as *correlated attractors 1–3*. Thus, there are many types of correlated attractor as equilibrium states.

In the model of $F = 0.5$, the firing rate of both the Hopfield attractor and the correlated attractors automatically becomes 0.5 without firing rate control, and the model has a property showing that the Hopfield attractor and correlated attractors can become the multiple stable state.[6] In the sparse coding scheme, it is necessary to control the firing rate of the model by using the firing rate F of the memory pattern as mentioned in §2. Then, it can be easily understood that the Hopfield attractor, which looks like the memory pattern, can become equilibrium state. In this study, it has been clarified that the correlated attractor can also become the equilibrium state, as mentioned above. Therefore, the model holds the property showing that the Hopfield attractor and the correlated attractors can become the multiple stable state. Although a detailed description is omitted, a qualitatively similar multiple stability has also been observed in other values of F.

Let us consider the effect of the coefficient g. If this coefficient g is too small, the effect of bringing the firing rate close to F weakens, then the critical temperatures T_c becomes lower. Oppositely, if g is too large, the amount of the feedback loop of firing rate control becomes excessive and trajectories go out of the basin of the Hopfield attractor or the correlated attractors. Thus each attractor becomes unstable. We set $g = 10$ so that the Hopfield attractor and correlated attractors become stable.

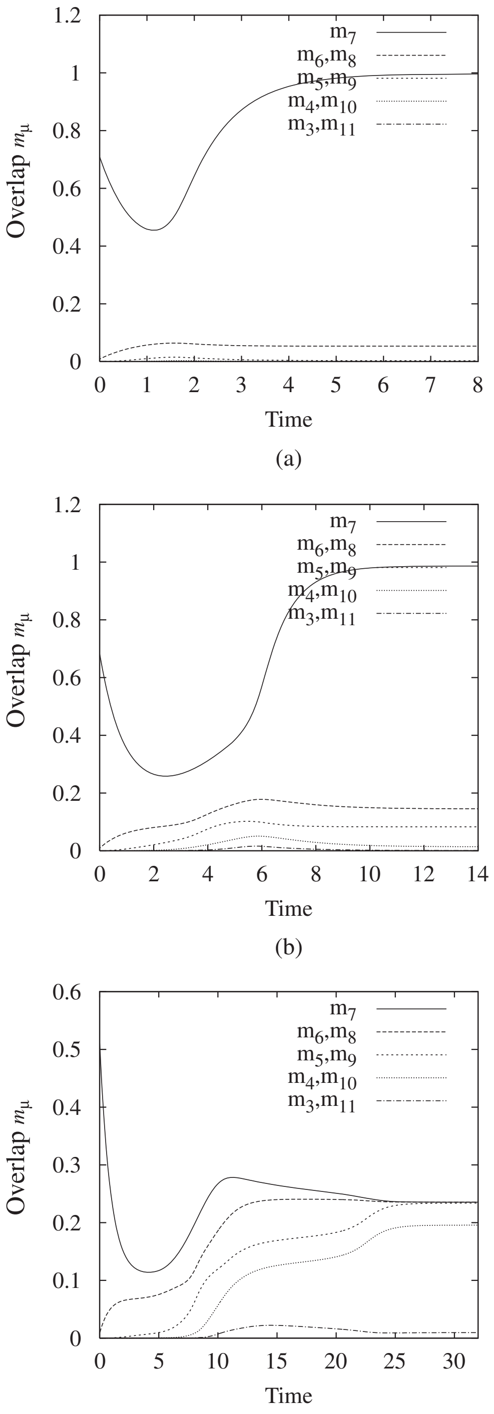
4.2 Model dynamics

We now examine the retrieval dynamics of the attractors that reach an equilibrium state at $T = 0.04$ by numerical calculations. We use a state that has a nonzero overlap only with the memory pattern $\boldsymbol{\eta}^7$ as an initial state $\boldsymbol{x}^0$. The method to generate the initial state is explained in detail in the following. The x_i^0 of the ith neuron where $\eta_i^7 = 1$ is flipped to 0 with a probability f, and that of the ith neuron where $\eta_i^7 = 0$ is flipped to 1 with the same number. Consequently, $\boldsymbol{x}^0$ in this case has a nonzero overlap only with $\boldsymbol{\eta}^7$, and the overlap m_7 becomes

$$m_7 = \frac{1}{VN}\sum_{i=1}^{N}(\eta_i^7 - F)x_i^0 = \frac{1-F-f}{1-F}. \tag{32}$$

Figure 3 shows the results of the retrieval dynamics. Here, the initial overlap was assumed to be $m_7 = 0.76$, 0.74, or 0.56. Figures 3(a)–3(c) show the aspects that the trajectories go toward the Hopfield attractor, the correlated attractor 1 and the correlated attractor 3, respectively. Thus the retrieval state depends on the initial overlap.

We also show the result of retrieval dynamics for various initial overlaps in Fig. 4. Here, the horizontal axis is the overlap m_7 and the vertical axis is the overlap m_6 ($= m_8$).

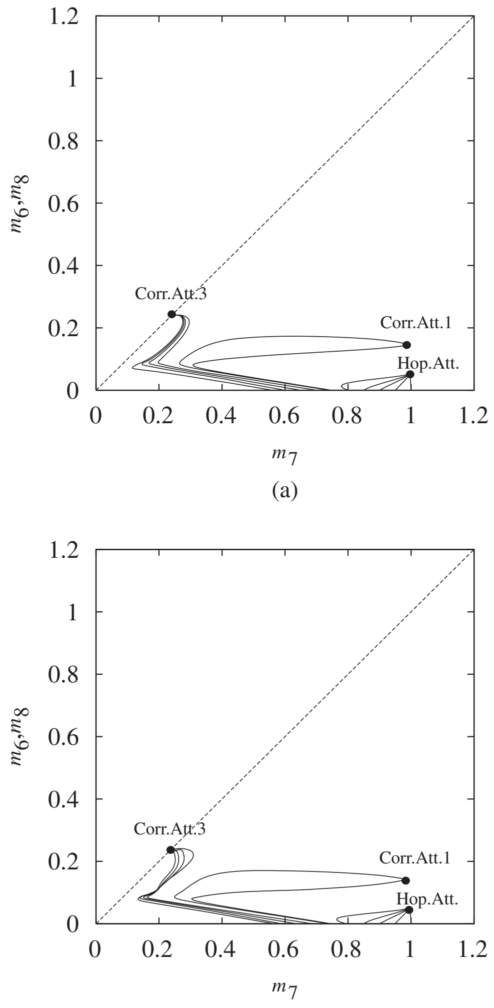


(a)

(b)

(c)

Fig. 3. Numerical calculation result of overlap dynamics at $T = 0.04$. $m_3, \ldots, m_{11}$ are shown. Retrieval state depends on initial overlap: (a) hopfield attractor [$m_7(t = 0) = 0.76$], (b) correlated attractor 1 [$m_7(t = 0) = 0.74$], and (c) correlated attractor 3 [$m_7(t = 0) = 0.56$].



(a)

(b)

Fig. 4. Result of (a) numerical calculation and (b) computer simulation of overlap dynamics at $T = 0.04$. Only m_7–$m_6(, m_8)$ are shown.

We examine the retrieval dynamics by numerical calculations and confirm these results by computer simulations. In Fig. 4, the initial state $\boldsymbol{x}^0$ has a nonzero overlap with only $\boldsymbol{\eta}^7$, so that the retrieval begins from the horizontal axis. The trajectory goes toward the Hopfield attractor when the initial overlap $m_7 > 0.76$, toward the correlated attractor 1 when the initial overlap is $0.74 < m_7 < 0.76$, and toward the correlated attractor 3 when the initial overlap is $0.56 < m_7 < 0.74$. A critical initial overlap exists and the resultant attractor differs depending on this overlap. These results are consistent with the numerical calculations and computer simulations ($N = 200{,}000$). Incidentally, trajectories do not go toward the correlated attractor 2 because it is already unstable at $T = 0.04$.

At any F, qualitatively similar attractors become multiple stable states as mentioned in chapter 4.1. The system retrieves one of these attractors. If g is too small, the T_c of these attractors decreases and stability becomes weak, then the basins of the attractors become narrow. Oppositely, if g is too large, the amount of the feedback loop of firing rate control becomes excessive and trajectories go out of the basin of the Hopfield attractor or correlated attractors. Thus, each attractor cannot be retrieved.

Next, we examine the correlation between the attractors in this model as well as in the physiological experiment by Miyashita. Before discussing the results, we will explain Miyashita's experiment using the present model. Miyashita had a monkey learn memory patterns $\boldsymbol{\eta}^\mu$ $(1 \leq \mu \leq s)$ repeated in the same sequence. Afterwards, they presented $\boldsymbol{\eta}^\mu$ $(1 \leq \mu \leq s)$ to the monkey, and observed the firing rate pattern $\boldsymbol{\sigma}^\mu$ $(1 \leq \mu \leq s)$ of an inferior temporal cortex neuron. Furthermore, he examined the correlation between the firing rate patterns represented by

$$C(\mu,\nu) = \frac{1}{|C|}\sum_{i=1}^{N}(\sigma_i^\mu - \bar{\sigma}^\mu)(\sigma_i^\nu - \bar{\sigma}^\nu), \qquad (33)$$

where $\boldsymbol{\sigma}^\mu$ is the same as the $\boldsymbol{x}$, $\bar{\boldsymbol{\sigma}}^\mu$ is the average firing rate of the firing rate pattern, and $|C|$ is a normalization constant for the correlation $C(\mu,\nu)$. Through this experiment, he found that the larger the difference $|\mu-\nu|$ of learning sequence numbers is, the lower the correlation between the firing rate patterns $\boldsymbol{\sigma}^\mu$ and $\boldsymbol{\sigma}^\nu$ is.

We also examine the time course of $C(\mu,\nu)$, although Miyashita obtained the above experimental result by analyzing the stationary state of the neuron response. To examine the time course of the correlation $C(\mu,\nu)$, we calculate it using the time course of m_μ $(1 \le \mu \le s)$:

$$C(\mu,\nu) = \frac{1}{|C|}\langle\langle(\sigma_i^\mu - \bar{\sigma}^\mu)(\sigma_i^\nu - \bar{\sigma}^\nu)\rangle\rangle, \qquad (34)$$

$$\sigma_i^\mu = \frac{1+\tanh\left\{\beta\left[\sum_{\nu,\nu'=1}^{s}(\eta_i^\nu - F)A_{\nu\nu'}m_{\nu'} + h - gI\right]\right\}}{2}, \qquad (35)$$

$$\bar{\sigma}^\mu = \left\langle\!\!\left\langle \frac{1+\tanh\left\{\beta\left[\sum_{\nu,\nu'=1}^{s}(\eta_i^\nu - F)A_{\nu\nu'}m_{\nu'} + h - gI\right]\right\}}{2} \right\rangle\!\!\right\rangle. \qquad (36)$$

Figure 5 shows the correlation function with an initial overlap $m_7 = 0.76$, 0.74, or 0.56. The three lines in each graph show the time course of the correlation in the retrieval dynamics. When the state goes toward the Hopfield attractors, the correlation $C(\mu,\nu)$ does not tend to spread, as shown in Fig. 5(a). In particular, the spread of the correlation $C(\mu,\nu)$ is narrow in the equilibrium state $(t=\infty)$. On the other hand, when the state goes toward the correlated attractors, the correlation $C(\mu,\nu)$ spreads so much as shown in Figs. 5(b) and 5(c). Moreover, the correlation $C(\mu,\nu)$ tends to spread depending on elapsed time. This is because the model has an essential property showing that the presented memory pattern excites adjacent consecutive memory patterns one after another owing to the structure of the connection weights of eq. (5) and a correlated attractor is eventually retrieved. Incidentally, we also observed a similar dynamics for a monostable parameter, although Fig. 5 shows only results for a multiple stable parameter.

5. Discussion

Here, we discuss the current understanding of the findings of certain physiological experiments concerning neurons of the inferior temporal cortex on the basis of the obtained theoretical results. First, we consider the findings regarding the contradiction of the results of two physiologic experiments using the model with multiple stability. Tanaka *et al.* and Fujita *et al.* used single-cell recording, and showed that the neurons of the inferior temporal cortex respond selectively to a partial feature of visual stimuli.[12,13] On the other hand, Gochin *et al.*[14] examined neuron response using a stimulus set that is different from the visual stimuli used by Tanaka and Fujita, and showed that the selectivity of the response of the neuron is weak, that many neurons respond to many visual stimuli, and that only firing strength changes depending on the type of visual stimulus. If the visual system of the inferior temporal cortex is composed of a dynamical system and it executes visual processing using the attractors, there is no question even if a few differences in the visual stimulus cause quite different responses. To explain this contradiction, we propose three hypotheses: the inferior temporal cortex has the same structure as that in the Amit model,[3,4] which has some stable attractors; the response of Tanaka–Fujita-type neurons corresponds to the Hopfield attractor; and the response of Gochin-type neurons corresponds to a correlated attractor.

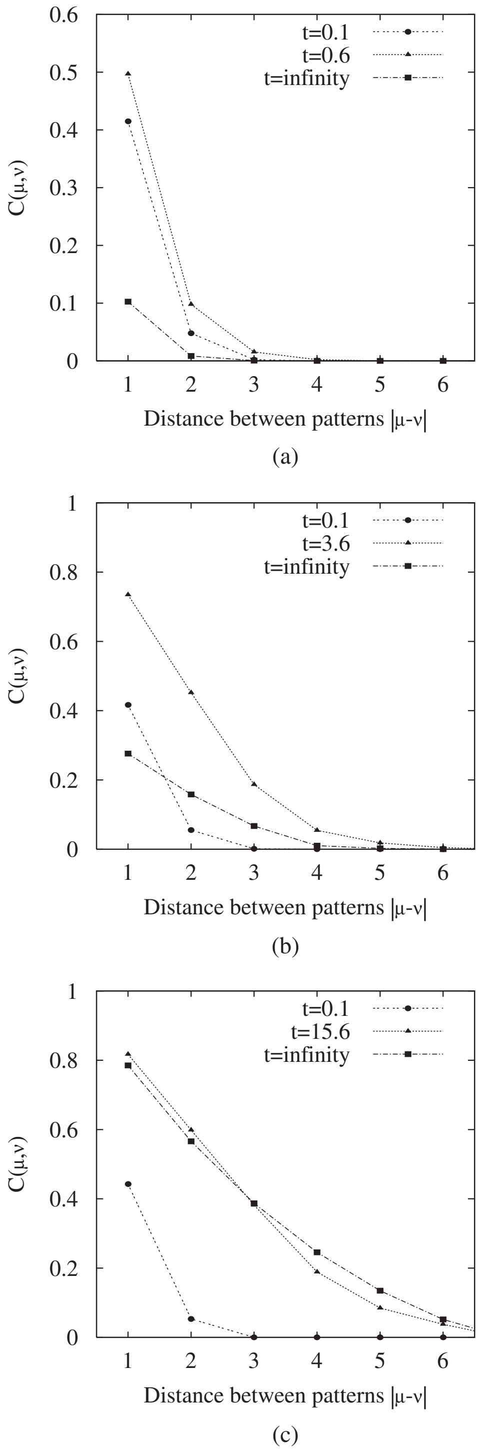


Fig. 5. Time course of correlation between attractors at $T = 0.04$: (a) $m_7 = 0.76$, (b) $m_7 = 0.74$, and (c) $m_7 = 0.56$.

We then explain these hypotheses in detail. Some physiological findings and theoretical viewpoints suggest that the sparse coding scheme is used in the brain.[2,7–11] If a neuron group is in a state corresponding to the Hopfield attractor under sparse coding, the neurons in the group seem to selectively respond to a specific feature. That is, each neuron shows the feature selectivity of the Tanaka–Fujita type. However, if a neuron group is in a state corresponding to a correlated attractor, the feature selectivity of the neurons seems weak because neurons respond to many features. That is, each neuron shows the feature selectivity of the Gochin-type. Furthermore, the similarity between the characteristics of these attractors and the response of the physiologic neurons is consistent with the corresponding experimental paradigms. Tanaka *et al.* and Fujita *et al.* determined the feature selectivity of neurons by gradually reducing the visual stimulus from complex figures to simple figures. Apparently, the visual stimuli they use, that is, visual stimuli with a large overlap, are stimuli to which the inferior temporal cortex neurons respond significantly. On the other hand, the method by Gochin *et al.* differs from that by Tanaka and Fujita in the sense that they examined neuron response using visual stimuli prepared beforehand. In other words, Gochin *et al.* chose visual stimuli without considering what images neurons have learned, while Tanaka *et al.* and Fujita *et al.* chose visual stimuli through the reduction method. Therefore, Gochin *et al.*'s visual stimuli were not the most appropriate for these neurons.

Here, we propose a physiological experiment that may verify this hypothesis. First, the optimum visual stimulus for inferior temporal cortex neurons is determined through the reduction method used by Tanaka *et al.* and Fujita *et al.* Next, the noise is superimposed on the visual stimulus, as carried out by Shidara *et al.*[15] and Amit *et al.*[16] Then, the noise level is changed and the feature selectivity of the neuron is measured. Thus, it can be determined whether the firing strength of an initially excited neuron decreases and instead other neurons begin to respond suddenly when the noise level exceeds a certain boundary. This will enable us to estimate the possibility that the state will change into a different attractor when the noise level exceeds a certain boundary. To determine whether the state is drawn toward another attractor, without quantitatively measuring the change in the firing strength, the autocorrelation function of the response fluctuations of a neuron can be examined. Tanimoto *et al.* reported that the autocorrelation function of neuron response fluctuations differs between the Hopfield attractor and the correlated attractor in the Amit model.[17]

Next, we propose the analysis of the transition property of the neurons in the Miyashita's experiment, on basis of our finding regarding the retrieval dynamics of the correlated attractor. The correlation function diverges as time, as shown in Fig. 5. This is because the correlated attractor is progressively generated by iteration dynamics due to the connection matrix, as shown in eqs. (5) and (6). It is possible that the correlated attractor is created in one shot throught a feed-forward connection from a visual stimulus without iteration dynamics. However, the change in the correlation function as time, as shown in Fig. 5, would not occur without iteration dynamics. Therefore, measurement of the time course of the correlation is one way to test whether the Amit model is appropriate. The time course of the correlation has not been measured in Miyashita's experiment, so it will be useful to examine the time course of the neuron response in detail. If the structure of the inferior temporal cortex can be described by the Amit model, the transition from the response of the Hopfield attractor type to the response of the correlated attractor type would be observed by adding external noise to stimuli. Moreover, the time course of the neuron group response can be analyzed through principal component analysis, the clustering technique or mutual information.[18,19] Examining the time course of the neuron group response enables us to search for evidence supporting the Amit model and indicating that iteration dynamics is used in the brain.

6. Conclusions

We examined the sparsely encoded attractor neural network model with a firing rate control operation, and analyzed the stability and retrieval dynamics for the Hopfield attractor and correlated attractors. We found the parameter region where the Hopfield attractor and correlated attractors are stable. In addition, small differences in the initial state change the retrieval attractor. Furthermore, we examined not only the static correlations between the correlated attractors, as in Miyashita's experiment, but also the time course of those correlations in the retrieval dynamics. Then we found that the correlation function spreads gradually as time. Our theoretical results suggest a consistent explanation for physiologic findings regarding neurons of the inferior temporal cortex, which seem contradictory, and we have proposed new paradigms for physiological experiments.

Acknowledgements

This work was partially supported by a Grant-in-Aid for the Encouragement of Young Scientists (B) No. 18700302, a Grant-in-Aid for Scientific Research on Priority Areas No. 18079003, and a Grant-in-Aid for Scientific Research (A) No. 20240020 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

1) Y. Miyashita and H. S. Chang: Nature **331** (1988) 68.
2) Y. Miyashita: Nature **335** (1988) 817.
3) M. Griniasty, M. V. Tsodyks, and D. J. Amit: Neural Comput. **5** (1993) 1.
4) D. J. Amit, N. Brunel, and M. V. Tsodyks: J. Neurosci. **14** (1994) 6435.
5) T. Fukai, T. Kimoto, M. Doi, and M. Okada: J. Phys. A **32** (1999) 5551.
6) T. Uezu, A. Hirano, and M. Okada: J. Phys. Soc. Jpn. **73** (2004) 867.
7) M. V. Tsodyks and M. V. Feigel'man: Europhys. Lett. **6** (1988) 101.
8) J. Buhmann, R. Divko, and K. Schulten: Phys. Rev. A **39** (1989) 2689.
9) S. Amari: Neural Networks **2** (1989) 451.
10) C. J. Perez-Vicente and D. J. Amit: J. Phys. A **22** (1989) 559.
11) M. Okada: Neural Networks **9** (1996) 1429.
12) K. Tanaka, H. Saito, Y. Fukada, and M. Moriya: J. Neurophysiol. **66** (1991) 170.
13) I. Fujita, K. Tanaka, M. Ito, and K. Cheng: Nature **360** (1992) 343.
14) P. M. Gochin, M. Colombo, G. A. Dorfman, G. L. Gerstein, and C. G. Gross: J. Neurophysiol. **71** (1994) 2325.
15) M. Shidara, S. Liu, and B. J. Richmond: Soc. Neurosci. Abstr. **22** (1996) 1614.
16) D. J. Amit, S. Fusi, and V. Yakovlev: Neural Comput. **9** (1997) 1071.
17) S. Tanimoto, M. Okada, T. Kimoto, and T. Uezu: J. Phys. Soc. Jpn. **75** (2006) 104004.
18) Y. Sugase, S. Yamane, S. Ueno, and K. Kawano: Nature **400** (1999) 869.
19) N. Matsumoto, M. Okada, Y. Sugase-Miyamoto, S. Yamane, and K. Kawano: Cereb. Cortex **15** (2005) 1103.