Biol. Cybernetics 26, 175—185 (1977)

Biological Cybernetics

© by Springer-Verlag 1977

Neural Theory of Association and Concept-Formation

S.-I. Amari*

University of Tokyo, Tokyo, Japan; Center for Systems Neuroscience, University of Massachusetts, Amherst, MA, USA

Abstract. The present paper looks for possible neural mechanism underlying such high-level brain functioning as association and concept-formation. Primitive neural models of association and concept-formation are presented, which will elucidate the distributed and multiply superposed manner of retaining knowledge in the brain. The models are subject to two rules of self-organization of synaptic weights, orthogonal and covariance learning. The convergence of self-organization is proved, and the characteristics of these learning rules are shown. The performances, especially the noise immunity, of the association net and concept-formation net are analyzed.

I. Introduction

A typical "neural theory" builds a model of a specific portion of the brain in the beginning, and predicts the functioning of that portion by computer-simulated experiments. We, however, do not intend to build a model of a specific portion at the present stage. We rather intend to figure out possible neural mechanisms of learning, storing and using knowledge which might be found widely in the brain in various versions. We search for such mechanisms of distributed and multiply-superposed information processing as might underlie neural association and concept-formation. These mechanisms will help us in building more realistic models at the next stage.

The present work is a development of the model, consisting of mutually connected bistable neuron pools, proposed by Amari (1971) and partly analyzed in Amari (1972a). In order to make mathematical analysis tractable, the model is kept as simple as possible as long as the essential features are not missed. This is a common attitude of the author's neural researches (Amari, 1971; 1972a, b; 1974a, b; 1975; 1977a, b).

We consider the following association net. The net, learning from k pairs of stimulus patterns $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_k, \boldsymbol{z}_k)$, self-organizes in such a manner that, when the net receives a key pattern $\boldsymbol{x}_\alpha (\alpha = 1, \ldots, k)$, it correctly outputs the associated partner $\boldsymbol{z}_\alpha$. The self-organization is carried into effect through modification of the synaptic weights of the net. This type of neural association has been investigated by many authors (e.g., Nakano, 1972; Kohonen, 1972; Anderson, 1972; Amari, 1972a; Uesaka and Ozeki, 1972; Wigström, 1973), where the so-called "correlation" of pattern components is memorized in the synaptic weights. The correlational association works well when the patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are mutually orthogonal, but otherwise it does not work well by virtue of the mutual interference of superposed patterns. The present paper considers two methods of neural self-organization, orthogonal and covariance learning, which eliminate the above interference. Orthogonal learning is closely related to Kohonen's generalized inverse approach (Kohonen, 1974; Kohonen and Oja, 1976).

The concept-formation net, on the other hand, is a sequential net having recurrent connections. The net, receiving many stimulus patterns which are distributed in k clusters, self-organizes in such a manner that the net forms an equilibrium state corresponding to each cluster. The equilibrium states correspond to the concept patterns which the net retains. After the learning is completed, the net, receiving a pattern belonging to a cluster, falls into the equilibrium state corresponding to the cluster. The net keeps and reproduces any concept pattern in this manner (short-term memory), while forming many equilibrium states by changing synaptic weights yields the long-term memory. Orthogonal learning and covariance learning again play an important role.

* This research was supported in part by a grant by the Sloan Foundation to the Center for Systems Neuroscience, University of Massachusetts

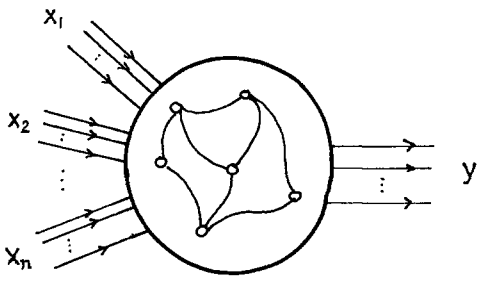


Fig. 1. Neuron pool

The main results of the present paper are: i) a mathematical analysis of the performance, especially noise immunity, of association and concept-formation nets, ii) the formulation of orthogonal and covariance learning, and iii) proof of convergence of neural self-organization including covariance and orthogonal learning.

II. Neural Models of Association and Concept-formation

2.1. Neuron Pool

We first consider the behavior of a neuron pool, which is composed of mutually connected neurons. Neuron pools play the role of building blocks of the models in the present paper. A neuron pool has a bundle of output axons. The output signal y of the neuron pool is represented by the average pulse rate over this bundle of axons. Let $x_1, x_2, \ldots, x_n$ be n signals entering into the neuron pool. A signal x_i is also carried by a bundle of axons (Fig. 1), representing the average pulse rate of the corresponding bundle.

The behavior of a neuron pool in which many neurons are connected in a random manner has been studied in detail (e.g., Amari, 1971, 1972b, 1974; Amari et al., 1977). Let us consider a bistable neuron pool. It has two states, an excited state which emits a high pulse rate output and a resting state which emits a low pulse rate output. It can be in either state. It enters the excited state when a weighted sum

$$u=\sum_{i=1}^{n} w_i x_i$$

of the input signals exceeds a threshold value h_1, and stays in that state. It enters the resting state, when the weighted sum u becomes lower than another threshold value h_2 ($h_2<h_1$).

We represent the high pulse rate output by $y=1$, and the low pulse rate by $y=0$. The behavior of a bistable neuron pool, then, can approximately be described by the following equation

$$y=\varphi[\textstyle\sum w_i x_i+(h_1-h_2)y-h_1], \tag{2.1}$$

where φ is the step-function defined by

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

When $h_1=h_2$ holds, the equation is simplified to yield the input-output relation

$$y=\varphi(\textstyle\sum w_i x_i-h). \tag{2.2}$$

This is the behavior of the so-called McCulloch-Pitts formal neuron. If we attach a feedback connection to a McCulloch-Pitts neuron with weight h_1-h_2, (2.2) coincides with (2.1).

We hereafter use McCulloch-Pitts formal neurons as the constituent elements of our neuron nets, where an element may be a bistable neuron pool. A bundle of input or output axons of a neuron pool will also be treated as a single input or output line. We call w_i the connection weight of input x_i, and the weighted sum u the potential.

2.2. Model Net for Association

Let us consider a net of non-recurrent connections consisting of m elements (m neuron pools). The net receivers n input signals $x_1, \ldots, x_n$ and transforms them into m output signals $y_1, \ldots, y_m$, where y_i is the output from the i-th element (Fig. 2). Every input x_j is connected with all the elements. Let w_{ij} be the connection weight of x_j entering into the i-th element. Let h_i be the threshold of the i-th element.

The output y_i is determined by

$$y_i=\varphi\left(\sum_{j=1}^{n} w_{ij}x_j-h_i\right), \quad i=1,\ldots,m. \tag{2.3}$$

This determines the transformation from $x_1, \ldots, x_n$ to $y_1, \ldots, y_m$. It is convenient to use vector-matrix notation. We represent the input and output signals by column vectors, e.g.,

$$\boldsymbol{x}=(x_1, x_2, \ldots, x_n)^T$$

is the input vector, where the superscript T denotes transposition. The connection weights are also represented by a matrix

$$W=(w_{ij}),$$

which we call the connection matrix. The transformation (2.3) can then be rewritten in the form

$$\boldsymbol{y}=\varphi(W\boldsymbol{x}-\boldsymbol{h}), \tag{2.4}$$

where function φ operates component-wise. We symbolically write

$$\boldsymbol{y}=T\boldsymbol{x} \tag{2.5}$$

to show that $\boldsymbol{x}$ is transformed into $\boldsymbol{y}$.

An input vector $\boldsymbol{x}$ is called an input pattern. It represents a set of n features $x_1, x_2, \ldots, x_n$ of a signal pattern which the net processes. An output vector $\boldsymbol{y}$ is also called an output pattern.

The net has another set of inputs $z_1, \ldots, z_m$(Fig. 2). Signal z_i arrives directly at the i-th element, and plays the role of a "teacher" of this element. Since the teacher signal

$$z=(z_1, z_2, \ldots, z_m)^T$$

is strong, the output signal $\boldsymbol{y}$ is always set equal to z, when z arrives. The teacher signal z arrives at the net in the learning phase only. The net usually works without the teacher input.

Let us consider k pattern signals $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. When pattern $\boldsymbol{z}_\alpha$ is associated to pattern $\boldsymbol{x}_\alpha$, we have k pairs of patterns $(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$'s. When the net, receiving input $\boldsymbol{x}_\alpha$, outputs the associated pattern $\boldsymbol{z}_\alpha$, i.e., when $\boldsymbol{z}_\alpha = T\boldsymbol{x}_\alpha$ holds, we say that the net associates $\boldsymbol{z}_\alpha$ with $\boldsymbol{x}_\alpha$. The net self-organizes such that $\boldsymbol{z}_\alpha$ is associated with $\boldsymbol{x}_\alpha$ by receiving k given pattern pairs $(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$'s repeatedly, $\boldsymbol{x}_\alpha$ from the ordinary input and $\boldsymbol{z}_\alpha$ from the second input. This requires that $\boldsymbol{z}_\alpha = T\boldsymbol{x}_\alpha$ eventually holds when the second input is finally allowed to revert to zero.

This is a primitive model of a neural association mechanism. We can construct more complicated and plausible models of association based on the mechanism of this simple model. The mechanism can be incorporated with lateral-inhibition networks such as those of Spinelli (1970) and Malsburg (1973).

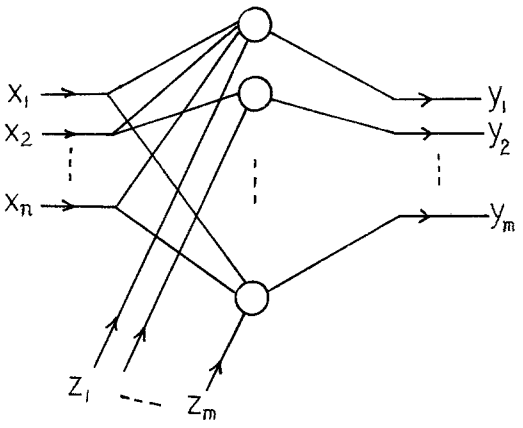


Fig. 2. Net for association

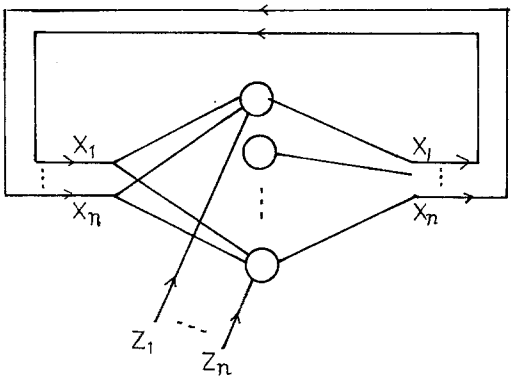


Fig. 3. Net for concept formation

2.3. Model Net for Concept Formation

A recurrent net can be obtained by connecting the outputs of a net to its inputs. We consider the net shown in Figure 3, where the net consists of n elements. Let $x_1(t), \ldots, x_n(t)$ be the outputs of the net at time t. We denote the outputs by

$$\boldsymbol{x}(t)=[x_1(t), \ldots, x_n(t)]^T$$

and call it the state of the net at time t. The output $\boldsymbol{x}(t+1)$ at time $t+1$ is determined by

$$\boldsymbol{x}(t+1)=\varphi[W\boldsymbol{x}(t)-\boldsymbol{h}] \tag{2.6}$$

which we write symbolically as

$$\boldsymbol{x}(t+1)=T\boldsymbol{x}(t), \tag{2.7}$$

and call T the state-transition operator. The second input z plays the role of setting the net in the initial state specified from the outside. When signal z arrives at the second input, the net is set in the state z. The net, then, begins its state transition, if no further signals come from the second input.

A state $\boldsymbol{x}$ which satisfies $T\boldsymbol{x}=\boldsymbol{x}$ is called an equilibrium state of the net. The net can have many equilibrium states. Equilibrium states represent information patterns which the net can retain persistently (the short term memory). We regard an equilibrium state as a concept pattern which the net has formed. The net can thus retain many concept patterns.

Many patterns will arrive at the net in the learning phase from the outside. When the patterns consists of a number of clusters, we consider that there is a concept corresponding to each cluster of patterns. We show in Figure 4 an example of a pattern distribution, where patterns are distributed around $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$, forming three clusters. The abscisa really is an n-dimensional pattern space and the ordinate represents the relative frequency of patterns.

Receiving a distribution of patterns one by one repeatedly, the net must self-organize by modifying the connection weights such that it automatically finds the clusters and forms equilibrium states. Moreover, when a pattern $\boldsymbol{x}'$ belonging to the cluster of $\boldsymbol{x}_\alpha$ is input later, the net, first entering state $\boldsymbol{x}'$, must tend to the equilibrium state $\boldsymbol{x}_\alpha$ thus formed. The net may be regarded as a pattern generator in which patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are coded in the form of equilibria.

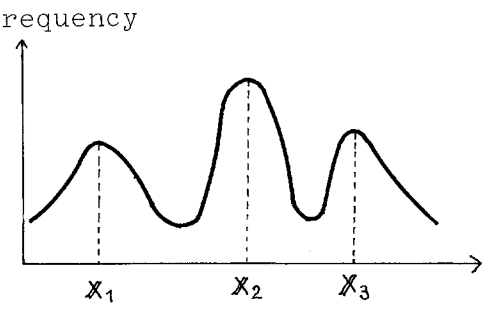


Fig. 4. Distribution of input patterns

III. Self-organization of Nerve Elements

3.1. Convergence of Learning

It is widely believed that a nerve net self-organizes by modifying its synaptic weights. Hebb (1949) proposed the rule that simultaneous firing of presynaptic and postsynaptic neurons brings an increase in that synaptic weight. This hypothesis has widely been accepted in various versions (e.g., Caianiello, 1961; Rosenblatt, 1961; Grossberg, 1969). We also adopt similar but somewhat extended rules of modification of connection weights.

A connection weight w_i of an element increases in proportion to a reinforcement signal r_i in the learning phase. It decays slowly at the same time. The rule of modification of a weight is, hence, formulated as

$$w_i(t+1)=(1-c)w_i(t)+dr_i(t)\,, \tag{3.1}$$

where c and d are constants, and $w_i(t)$ and $r_i(t)$ are respectively the connection weight and reinforcement signal at time t.

The reinforcement signal $r_i(t)$ is determined generally depending on the input signal $\boldsymbol{x}(t)$, the connection weight $\boldsymbol{w}(t)$, the output signal $y(t)$ and the teacher signal $z(t)$ (when it exists). Nerve nets can realize various possibilities of information processing if provided with appropriate reinforcement signals.

When $r_i(t)$ is given by $r_i(t)=x_i(t)y(t)$, it realizes the Hebbian rule. When a teacher signal $z(t)$ arrives at the element, we can consider the rule $r_i(t)=x_i(t)z(t)$. This may be called forced learning, because the teacher signal comes from outside. When the reinforcement signal is given by $r_i(t)=x_i(t)u(t)$, where $u(t)$ is the potential evoked by the input $\boldsymbol{x}(t)$

$$u(t)=\sum w_j(t)x_j(t)\,,$$

we obtain a learning rule based on the potential of the neuron. This plays an important role in orthogonal learning defined later.

We again use the vector notation, $\boldsymbol{w}=(w_1, \ldots, w_n)$, $\boldsymbol{r}(r_1, \ldots, r_n)$, where $\boldsymbol{w}$ and $\boldsymbol{r}$ are row vectors, so that $\boldsymbol{wx}$ gives the inner product of $\boldsymbol{w}$ and $\boldsymbol{x}$. The reinforcement signal $\boldsymbol{r}$ is a function of input $\boldsymbol{x}$, $\boldsymbol{w}$, and z. We represent it by

$$\boldsymbol{r}=\boldsymbol{r}(\boldsymbol{w}, \boldsymbol{x}, z)\,. \tag{3.2}$$

Let $\Delta\boldsymbol{w}(t)$ be the increment of the connection weight at time t

$$\Delta\boldsymbol{w}(t)=\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,.$$

Then, the rule of weight modification can be written as

$$\Delta\boldsymbol{w}(t)=-c\boldsymbol{w}(t)+d\boldsymbol{r}[\boldsymbol{w}(t), \boldsymbol{x}(t), z(t)] \tag{3.3}$$

for suitable positive constants c and d. This shows the direction of modification of $\boldsymbol{w}$.

We consider the case when there exists a function $R(\boldsymbol{w}, \boldsymbol{x}, z)$ for which

$$\frac{\partial R}{\partial \boldsymbol{w}}=\boldsymbol{w}-\frac{d}{c}\boldsymbol{r}(\boldsymbol{w}, \boldsymbol{x}, z) \tag{3.4}$$

holds[1], where $\partial R/\partial \boldsymbol{w}=(\partial R/\partial w_1, \ldots, \partial R/\partial w_n)$ is the gradient vector. Then

$$\Delta\boldsymbol{w}=-c\frac{\partial R}{\partial \boldsymbol{w}}\,. \tag{3.5}$$

This means that $\boldsymbol{w}$ is modified in the negative direction of the gradient of R with respect to $\boldsymbol{w}$.

The connection weight is modified depending on the $\boldsymbol{x}(t)$ and $z(t)$, $t=1, 2, 3, \ldots$. (We do not have any teacher signals in the case of concept formation.) We consider the situation where $[\boldsymbol{x}(t), z(t)]$'s are independent samples from a fixed distribution. A typical case is that there are k input signals $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, $\boldsymbol{x}_\alpha$ coming with probability or relative frequency p_α. The teacher signal z_α is paired with input signal $\boldsymbol{x}_\alpha$, in the case of association. We show the convergence of the connection weights.

Let $L(\boldsymbol{w})$ be the expected value of R for possible inputs $(\boldsymbol{x}, z)$,

$$L(\boldsymbol{w})=\langle R(\boldsymbol{w}, \boldsymbol{x}, z)\rangle\,, \tag{3.6}$$

where $\langle\ \rangle$ denotes the expectation with respect to $\boldsymbol{x}$ and z. In the above case,

$$L(\boldsymbol{w})=\sum_{\alpha=1}^{k} p_\alpha R(\boldsymbol{w}, \boldsymbol{x}_\alpha, z_\alpha)\,.$$

By taking the expectation of (3.5), we have

$$\langle\Delta\boldsymbol{w}\rangle=-c\partial L(\boldsymbol{w})/\partial\boldsymbol{w}\,.$$

This shows that $\boldsymbol{w}$ is modified on the average in the direction of the steepest descent of $L(\boldsymbol{w})$. Hence, it is expected that $\boldsymbol{w}$ eventually converges to the value which minimizes $L(\boldsymbol{w})$ to within small fluctuations.

[1] The necessary and sufficient condition for the existence of R is given by $\partial r_i/\partial w_j=\partial r_j/\partial w_i$. As will be shown later, the condition is satisfied for most interesting cases

Assume that $L(\boldsymbol{w})$ is a function which has a unique minimum and satisfies some regularity conditions. Then, we can apply the theory of stochastic approximation (e.g., Wasan, 1969). When c is small, we can show that the expected value $\langle \boldsymbol{w}(t) \rangle$ of the connection weights converges to the minimum of $L(\boldsymbol{w})$. Moreover, the deviation of $\boldsymbol{w}(t)$ from the expected value can be made as small as desired, by letting c be sufficiently small.

We assume in the following that the decay constant c is sufficiently small so that the connection weights converge to the minimum of $L(\boldsymbol{w})$. The minimum is attained at $\boldsymbol{w}$ satisfying

$$c\boldsymbol{w} = d\langle \boldsymbol{r}(\boldsymbol{w}, \boldsymbol{x}, z) \rangle . \tag{3.7}$$

3.2. Examples of Simple Learning

A) Hebbian Learning. In the case of Hebbian learning, r_i is given by $r_i = yx_i$, where $y = \varphi(\boldsymbol{w}\boldsymbol{x} - h)$. Let $\Phi(u)$ be the integral of φ

$$\Phi(u) = \int_0^u \varphi(x)dx .$$

The function

$$R(\boldsymbol{w}, \boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{w}\boldsymbol{w}^T - \frac{d}{c}\Phi(\boldsymbol{w}\boldsymbol{x} - h)$$

satisfies (3.4). Therefore, by Hebbian learning, the connection weight converges to $\boldsymbol{w}$ satisfying

$$\boldsymbol{w} = \frac{d}{c}\langle \varphi(\boldsymbol{w}\boldsymbol{x} - h)\boldsymbol{x}^T \rangle . \tag{3.8}$$

B) Forced Learning. The reinforcement signal is given by $r_i = zx_i$ with teacher signal z. It is easy to see that

$$R(\boldsymbol{w}, \boldsymbol{x}, z) = \tfrac{1}{2}\boldsymbol{w}\boldsymbol{w}^T - \frac{d}{c}z\boldsymbol{w}\boldsymbol{x}$$

gives the forced rule (3.4). From

$$2L(\boldsymbol{w}) = \left(\left| \boldsymbol{w} - \frac{d}{c}\langle z\boldsymbol{x}^T \rangle \right| \right)^2 - \left(\frac{d}{c}|\langle z\boldsymbol{x}^T \rangle| \right)^2 ,$$

the minimum is attained at

$$\boldsymbol{w} = \frac{d}{c}\langle z\boldsymbol{x}^T \rangle .$$

This is sometimes called the "correlation" of z and $\boldsymbol{x}$.

C) Learning Based on Potential. The learning rule based on potential is given by

$$R(\boldsymbol{w}, \boldsymbol{x}) = \tfrac{1}{2}\left[\boldsymbol{w}\boldsymbol{w}^T - \frac{d}{c}(\boldsymbol{w}\boldsymbol{x})^2 \right],$$

where the reinforcement signal is $\boldsymbol{r} = (\boldsymbol{w}\boldsymbol{x})\boldsymbol{x}$. We have

$$L(\boldsymbol{w}) = \tfrac{1}{2}\left(\boldsymbol{w}\boldsymbol{w}^T - \frac{d}{c}\boldsymbol{w}\langle \boldsymbol{x}\boldsymbol{x}^T \rangle \boldsymbol{w}^T \right),$$

where $\langle \boldsymbol{x}\boldsymbol{x}^T \rangle$ is the matrix whose (i, j) entry is given by $\langle x_i x_j \rangle$. (It should be remembered that $\boldsymbol{x}$ is a column vector while $\boldsymbol{w}$ is a row vector.)

The function $L(\boldsymbol{w})$, however, has no minimum in this case. Therefore, the connection weight $\boldsymbol{w}(t)$ does not converge under the above rule of learning. If the connection weight is subject to the subsidiary condition $\boldsymbol{w}\boldsymbol{w}^T = \text{const}$, so that $\boldsymbol{w}(t)$ is normalized after each step of learning, we can prove that $\boldsymbol{w}(t)$ converges to the minimum of $L(\boldsymbol{w})$ under the subsidiary condition. It is the direction of the eigenvector of the matrix $\langle \boldsymbol{x}\boldsymbol{x}^T \rangle$ corresponding to the maximum eigenvalue.

3.3. Orthogonal Learning

An element may have both excitatory and inhibitory channels from the inputs. An input signal, therefore, not only excites it but also inhibits it at the same time through inhibitory interneurons. The connection weight w_i can be decomposed into $w_i = w_i^+ - w_i^-$, where w_i^+ and w_i^- represent, respectively, the excitatory and inhibitory efficiency.

Excitatory and inhibitory neurons may have different self-organization rules. We consider the case where the excitatory w_i^+ is modified subject to forced learning with teacher signal, while the inhibitory weight w_i^- is modified subject to learning based on potential;

$$\Delta \boldsymbol{w}^+ = -c\boldsymbol{w}^+ + dz\boldsymbol{x}^T ,$$

$$\Delta \boldsymbol{w}^- = -c'\boldsymbol{w}^- + d'(\boldsymbol{w}\boldsymbol{x})\boldsymbol{x}^T .$$

The excitatory weight converges to

$$\boldsymbol{w}^+ = \frac{d}{c}\langle z\boldsymbol{x}^T \rangle . \tag{3.9}$$

Define

$$R(\boldsymbol{w}^-, \boldsymbol{w}^+, \boldsymbol{x}) = \tfrac{1}{2}\left[|\boldsymbol{w}^-|^2 + \frac{d'}{c'}|(\boldsymbol{w}^+ - \boldsymbol{w}^-)\boldsymbol{x}|^2 \right]. \tag{3.10}$$

We then have $\Delta \boldsymbol{w}^- = -c'\partial R/\partial \boldsymbol{w}^-$. The function

$$L(\boldsymbol{w}^-, \boldsymbol{w}^+) = \langle R \rangle = \tfrac{1}{2}\left[|\boldsymbol{w}^-|^2 + \frac{d'}{c'}(\boldsymbol{w}^+ - \boldsymbol{w}^-)\langle \boldsymbol{x}\boldsymbol{x}^T \rangle (\boldsymbol{w}^+ - \boldsymbol{w}^-) \right]$$

is quadratic in $\boldsymbol{w}^-$ and its coefficient matrix is positive definite. Hence, it has a unique minimum, and $\boldsymbol{w}^-$ converges to the minimum. Since $\boldsymbol{w}^+$ converges to

(3.9) independently of w^-, the minimum w^- is given by the equation

$$w^- = \frac{d'}{c'}[(w^+ - w^-)\langle xx^T\rangle] .$$

The total synaptic weight $w = w^+ - w^-$, therefore, converges to one satisfying

$$w = \frac{d}{c}\langle zx^T\rangle - \frac{d'}{c'} w\langle xx^T\rangle . \tag{3.11}$$

or

$$w = e\langle zx^T\rangle(\varepsilon E + \langle xx^T\rangle)^{-1} , \tag{3.12}$$

where E is the unit matrix and $e = c'd/cd'$, $\varepsilon = c'/d'$.

Let us consider the case where k signals $x_1, \ldots, x_k$ come to the input ($k < n$) with paired teacher signals z_α. We then have

$$\langle zx\rangle = \sum_\alpha p_\alpha z_\alpha x_\alpha , \qquad \langle xx^T\rangle = \sum_\alpha p_\alpha x_\alpha x_\alpha^T .$$

We assume that k vectors $x_1, \ldots, x_k$ are linearly independent.

Let us introduce the set of dual orthogonal vectors $x_1^*, x_2^*, \ldots, x_k^*$ associated with $x_1, \ldots, x_k$. The x_α^* is a row vector satisfying the two conditions:

1) x_α^* is a linear combination of $x_1^T, \ldots, x_k^T$,

2) $x_\alpha^* x_\beta = \begin{cases} 1, & \alpha = \beta \\ 0, & \alpha \neq \beta . \end{cases}$

In other words, x_α^* belongs to the subspace spanned by $x_1^T, \ldots, x_k^T$ and is orthogonal to all the x_β's except x_α. The dual orthogonal vector is uniquely determined by

$$x_\alpha^* = \frac{1}{n}\sum_{\beta=1}^{k} G_{\alpha\beta}^{-1} x_\beta^T , \tag{3.13}$$

where $(G_{\alpha\beta}^{-1})$ is the inverse matrix of $(G_{\alpha\beta}) = (x_\alpha^T x_\beta)/n$.

Since $\langle xx^T\rangle$ is a singular matrix, it is impossible to expand (3.12) in a power series in ε, even when ε is small. By the method given in Appendix I,

$$w = e\sum z_\alpha x_\alpha^* - \frac{e\varepsilon}{n}\sum_{\alpha,\beta}\frac{1}{p_\beta} z_\alpha G_{\alpha\beta}^{-1} x_\beta^* + O(\varepsilon^2) , \tag{3.14}$$

where $O(\varepsilon^2)$ represents the second-order term in ε.

3.4. Covariance Learning

We propose another possible neural rule of self-organization. The excitatory weight w^+ is modified subject to forced learning as before, while the reinforcement signal r_i for the inhibitory synapse is given, in this case, by

$$r_i = e(bx_i + az - ab) ,$$

where a and b are constants. We then have

$$w = e'\langle (x - a)(z - b)\rangle , \tag{3.15}$$

where $e' = d/c = (d'/c')e$ and $a = (a, a, \ldots, a)^T$.

When a and b are, respectively, the expected values of x_i and z, $\langle (x_i - a)(z - b)\rangle$ is called the covariance of x_i and z. We call this rule of self-organization covariance learning.

When k input signals x_α's come with z_α's we have

$$w = e'\sum_\alpha p_\alpha (z_\alpha - b)(x_\alpha - a)^T . \tag{3.16}$$

When a and b are put equal to zero, this learning reduces to simple forced learning or correlation learning.

IV. Self-organization under Noisy Situations

4.1. Self-organization of Neural Nets

A non-recurrent net model for association receives k pairs of patterns (x_α, z_α) repeatedly, and the connection weight w_i of the i-th element converges to

$$w_i = e\langle z_i x^T\rangle(\varepsilon E + \langle xx^T\rangle)^{-1}$$

under orthogonal learning, where z_i is the teacher signal entering the i-th element. Hence, the connection matrix $W = (w_{ij})$ of the net converges to

$$W = e\langle zx^T\rangle(\varepsilon E + \langle xx^T\rangle)^{-1} . \tag{4.1}$$

This can be evaluated as

$$W = e\sum_\alpha z_\alpha x_\alpha^* - \frac{e\varepsilon}{n}\sum_{\alpha,\beta}\frac{1}{p_\beta} z_\alpha G_{\alpha\beta}^{-1} x_\beta^* + O(\varepsilon^2) . \tag{4.2}$$

When ε is sufficiently small, we have

$$Wx_\beta \doteqdot ez_\beta , \qquad \beta = 1, \ldots, k .$$

Therefore, ε represents the inaccuracy of the orthogonalization effect which suppresses the interference among many patterns.

Under covariance learning, the connection matrix converges to

$$W = e'\sum_\alpha p_\alpha (z_\alpha - \boldsymbol{b})(x_\alpha - \boldsymbol{a})^T , \tag{4.3}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors.

A recurrent net, on the other hand, is forced to be in state x, when it receives an input x from the second or the initial-state setting terminals. The net immediately outputs the x, which is fed back to the ordinary input terminals of the net. Learning takes place under this situation, so that x plays the roles of the input signal and the teacher signal at the same time.

The connection matrix of the net converges to

$$W = e\langle xx^T\rangle(\varepsilon E + \langle xx^T\rangle)^{-1} \tag{4.5}$$

by orthogonal learning. It converges to

$$W = e'\langle (x - \boldsymbol{a})(x - \boldsymbol{a})^T\rangle , \tag{4.5}$$

by covariance learning.

4.2. Learning under Noisy Situation

We have so far considered noiseless learning, where patterns $\boldsymbol{x}_\alpha$ and $\boldsymbol{z}_\alpha$ arrive without noise disturbances. In reality, the net receives noisy versions of $\boldsymbol{x}_\alpha$'s and $\boldsymbol{z}_\alpha$'s. Let $\tilde{\boldsymbol{x}}_\alpha$ be a noisy version of $\boldsymbol{x}_\alpha$

$$\tilde{\boldsymbol{x}}_\alpha = \boldsymbol{x}_\alpha + \boldsymbol{n}\,,$$

where $\boldsymbol{n}$ is a noise vector. When the noise disturbs the i-th component of $\boldsymbol{x}_\alpha$, the i-th component n_i of $\boldsymbol{n}$ is equal to 1 or -1, according as $x_i=0$ or 1. The n_i is otherwise equal to 0. We consider a random noise, which independently disturbs every component of a pattern with probability δ. We call δ the noise rate of noisy pattern $\tilde{\boldsymbol{x}}_\alpha$. The noisy $\tilde{\boldsymbol{x}}_\alpha$ differs from $\boldsymbol{x}_\alpha$ in $n\delta$ components on the average.

We consider the case where the net receives noisy pattern pairs $(\tilde{\boldsymbol{x}}_\alpha, \tilde{\boldsymbol{z}}_\alpha)$. The noise independently disturbs the pattern pair $(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$ each time. Let $\bar{\boldsymbol{x}}_\alpha$ be the average of noisy versions of $\boldsymbol{x}_\alpha$, i.e., the expectation of $\tilde{\boldsymbol{x}}_\alpha$ over all the possible noise. We easily have

$$\bar{\boldsymbol{x}}_\alpha = (1-2\delta)\boldsymbol{x}_\alpha + \mathbf{1}, \tag{4.6}$$

where $\mathbf{1}$ is the column vector defined by $\mathbf{1}=(1, 1, \ldots, 1)^T$.

Similarly, by taking the expectation over all possible patterns and noises,

$$\langle \boldsymbol{z}\boldsymbol{x}^T\rangle = \sum_\alpha p_\alpha \bar{\boldsymbol{z}}_\alpha \bar{\boldsymbol{x}}_\alpha^T\,,$$

$$\langle \boldsymbol{x}\boldsymbol{x}^T\rangle = \sum p_\alpha \bar{\boldsymbol{x}}_\alpha \bar{\boldsymbol{x}}_\alpha^T + \sigma E\,,$$

where $\sigma=\delta(1-\delta)$. Therefore, we see from (4.1) that the connection weight converges, under noisy orthogonal learning, to

$$W = e\sum \bar{\boldsymbol{z}}_\alpha \bar{\boldsymbol{x}}_\alpha^* + \frac{e\varepsilon'}{n}\sum \bar{\boldsymbol{z}}_\alpha G_{\alpha\beta}^{-1}\frac{1}{p_\beta}\bar{\boldsymbol{x}}_\beta^* + O(\varepsilon'^2)\,, \tag{4.7}$$

where $\varepsilon'=\varepsilon+\sigma$ and the $\bar{\boldsymbol{x}}_\alpha^*$'s are the dual orthogonal vectors of the $\bar{\boldsymbol{x}}_\alpha$'s.

Noisy learning results in changing $\boldsymbol{x}_\alpha$ and $\boldsymbol{z}_\alpha$ to $\bar{\boldsymbol{x}}_\alpha$ and $\bar{\boldsymbol{z}}_\alpha$, respectively. This is not so important, because they are similar. The important effect of the noise is that it apparently increases ε to $\varepsilon'=\varepsilon+\sigma$, thus increasing the inaccuracy of orthogonalization.

A cluster of patterns can be considered as being composed of noisy versions $\tilde{\boldsymbol{x}}_\alpha$'s of the representative patterns $\boldsymbol{x}_\alpha$ of that cluster. Let δ_α be the noise rate of $\tilde{\boldsymbol{x}}_\alpha$. A typical pattern $\tilde{\boldsymbol{x}}_\alpha$ in the cluster is different from $\boldsymbol{x}_\alpha$ in $n\delta_\alpha$ components on the average, so that δ_α represents the disperion of the cluster. Given a distribution of patterns consisting of k clusters, we have

$$\langle \boldsymbol{x}\rangle = \sum p_\alpha \bar{\boldsymbol{x}}_\alpha\,,$$

$$\langle \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\rangle = \sum p_\alpha \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}_\alpha^T + \sigma E\,,$$

where $\sigma = \sum p_\alpha \delta_\alpha(1-\delta_\alpha)$.

The connection weight matrix of a recurrent net converges, under orthogonal learning of the above distribution, to

$$W = e(\sum p_\alpha \bar{\boldsymbol{x}}_\alpha \bar{\boldsymbol{x}}_\alpha^T + \sigma E)[(\sigma+\varepsilon)E + \sum p_\alpha \bar{\boldsymbol{x}}_\alpha \bar{\boldsymbol{x}}_\alpha^T]^{-1}\,. \tag{4.8}$$

When both ε and σ are small, we can expand it into

$$W = e\left(\frac{\sigma}{\sigma+\varepsilon}E + \frac{\varepsilon}{\sigma+\varepsilon}\sum \bar{\boldsymbol{x}}_\alpha \bar{\boldsymbol{x}}_\alpha^*\right) + O(\varepsilon) \tag{4.9}$$

(see Appendix II). When $\sigma \ll \varepsilon$, i.e., the dispersion of clusters is small compared to ε, we have

$$W \doteqdot e\sum \bar{\boldsymbol{x}}_\alpha \bar{\boldsymbol{x}}_\alpha^*\,.$$

However, when $\sigma \gg \varepsilon$, we have $W \doteqdot eE$ which carries no information about the distribution of patterns. Therefore, orthogonal learning works well only when the dispersion σ of patterns is smaller than ε. Hence, ε represents the degree of noise immunity for orthogonalization.

Covariance learning is not so sensitive to noise disturbances. The connection weight converges to

$$W = e'\sum p_\alpha(\bar{\boldsymbol{z}}_\alpha - \boldsymbol{b})(\bar{\boldsymbol{x}}_\alpha - \boldsymbol{a})^T \tag{4.10}$$

for a non-recurrent net, and to

$$W = e'[\sum p_\alpha(\bar{\boldsymbol{x}}_\alpha - \boldsymbol{a})(\bar{\boldsymbol{x}}_\alpha - \boldsymbol{a})^T + \sigma E] \tag{4.11}$$

for a recurrent net.

V. Behavior of Association and Concept-formation Net

5.1. Simple Illustration of Association Mechanism

The matrix

$$W = e'\sum p_\alpha \bar{\boldsymbol{z}}_\alpha \bar{\boldsymbol{x}}_\alpha^T \tag{5.1}$$

is sometimes called the correlation matrix of k pattern pairs, and many researchers have studied associative memory models by the use of this matrix (e.g., Nakano, 1972; Amari, 1972a; Kohonen, 1972; Anderson, 1972). Consider the very special case where k patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are mutually orthogonal,

$$\boldsymbol{x}_\alpha^T \boldsymbol{x}_\beta = 0 \quad (\alpha \neq \beta)\,. \tag{5.2}$$

In this case, for given input $\boldsymbol{x}_\beta$, the net produces the potential vector

$$\boldsymbol{u} = W\boldsymbol{x}_\beta = (e' p_\beta |\boldsymbol{x}_\beta|^2)\boldsymbol{z}_\beta$$

which is proportional to the associated pattern $\boldsymbol{z}_\beta$. Hence, if the threshold $\boldsymbol{h}$ is adequately chosen, the net outputs the associated pattern $\boldsymbol{z}_\beta$ from input $\boldsymbol{x}_\beta$ $(\beta=1, \ldots, k)$. The orthogonality condition (5.2) makes it possible to extract the necessary pattern from W in which many pattern pairs are superposed. Most studies of the so-called correlation memory have used

mutually orthogonal or nearly orthogonal patterns to take advantage of this fact.

When the orthogonality condition does not exactly hold, and when a noisy version $\boldsymbol{x}_\beta+\boldsymbol{n}$ is input as a key pattern, the induced potential vector is written as

$$W(\boldsymbol{x}_\beta+\boldsymbol{n})=(e' p_\beta|\boldsymbol{x}_\beta|^2)\boldsymbol{z}_\beta+\boldsymbol{N}\,,$$

where

$$\boldsymbol{N}=e' \sum_{\alpha\neq\beta} p_\alpha(\boldsymbol{x}_\alpha^T\boldsymbol{x}_\alpha)\boldsymbol{z}_\alpha+W\boldsymbol{n}$$

represents the term due to the interference from the other superposed patterns and the noise. When $\boldsymbol{N}$ is small enough, the interference is completely eliminated by operating the non-linear threshold function φ. The neural association is realized by the two mechanisms: 1) elimination of interference of patterns by orthogonalization; and 2) noise elimination by operating the non-linear function φ. Amari (1972a) has studied the condition of noise elemination by the use of the stability concept of threshold-element nets. Uesaka and Ozeki (1972) have shown the probability of noise elimination by using randomly generated patterns $\boldsymbol{x}_\alpha$'s, where each component takes on 1 and -1 with equal probability. In this case, two vectors are orthogonal on the average. Kohonen (1972) has studied the linear aspects of the correlation matrix memory. He also proposed the generalized-inverse approach, which uses the matrix W minimizing

$$L(W)=\sum_{\alpha=1}^{k}|W\boldsymbol{x}_\alpha-\boldsymbol{z}_\alpha|^2\,.$$

An iterative method of obtaining W was also shown (Kohonen, 1974). Orthogonal learning of the present paper has a close relation with the generalized inverse approach.

It should be noted that holography-like memory such as the holophone (Willshaw and Longuet-Higgins, 1969; Pfaffelhuber, 1975) also uses correlation memory. However, the dimensionality of the memory is as small as that of the patterns, so that it is difficult to superpose many patterns in the holography-like case.

5.2. *Noise Immunity of Association Net*

Let $\boldsymbol{n}$ be a noise disturbing a pattern $\boldsymbol{x}_\alpha$ into $\boldsymbol{x}'_\alpha=\boldsymbol{x}_\alpha+\boldsymbol{n}$. We call

$$\delta=\frac{1}{n}\sum_{i=1}^{n}|x'_{\alpha,i}-x_{\alpha,i}|=\frac{1}{n}\sum|n_i| \tag{5.3}$$

the noise rate of $\boldsymbol{x}'_\alpha$. The noise rate of an output pattern $\boldsymbol{z}'_\alpha$ is defined similarly. An association net is said to have a high noise immunity, if the net, receiving a noisy input $\boldsymbol{x}'_\alpha$, outputs the correct or less noisy pattern $\boldsymbol{z}'_\alpha$. The behavior of an association net is evaluated by the noise immunity shown by the relation between the noise rates of the input and output.

This relation, of course, depends on the whole set of pattern pairs $(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$ in a very complicated manner, so that it is almost impossible to obtain the relation. We, therefore, use the stochastic technique to evaluate the noise immunity. To this end, we assume that k pattern pairs $(\boldsymbol{x}_\alpha, \boldsymbol{z}_\alpha)$ are randomly generated. Let a be the probability that every component $x_{\alpha,i}$ of every $\boldsymbol{x}_\alpha$ is set equal to 1. Let b be the probability of $z_{\beta,j}$ being set equal to 1. All these components are independently determined.

We treat the case where n, m, and k are sufficiently large. Then, the law of large numbers holds and we can get an asymptotic evaluation of the noise immunity. The net must memorize k patterns $\boldsymbol{z}_\alpha$ which altogether include mk components. The net has nm modifiable connection weights w_{ij}. Therefore,

$$s=mk/nm=k/n \tag{5.4}$$

represents the ratio of the number of memorized components to that of the modifiable synapses.

We first consider a net subject to covariance learning of k pattern pairs, with $p_\alpha=1/k$. By evaluating the probability distribution of the interference term $\boldsymbol{N}$ (where we use the central limit theorem, the law of large numbers and assume that h_i's are adequately chosen), we have the following relation between the input noise ratio δ and the output noise ratio δ':

$$\delta'=F\left[\frac{\sqrt{a(1-a)(1-2\delta)}}{2\sqrt{sb(1-b)(a+\delta-2a\delta)}}\right] \tag{5.5}$$

where

$$F(s)=\frac{1}{\sqrt{2\pi}}\int_s^\infty \exp(-u^2/2)du\,.$$

The relation is shown in Figure 5, where $a=b=0.2$. When $s=0.1$, i.e., when the number of superposed pattern pairs is 10% of the dimensionality of input patterns, the net has an ability of eliminating an input noise of $\delta=0.1$ almost completely, outputting the precise associated pattern. When $s=0.3$, noise reduction also takes place, but the output pattern includes a noise of rate 2%, even when the input pattern is noiseless. When $s=0.4$, there is no noise reduction. The output noise rate is similar to that of the input. Therefore, the association net with covariance learning works well for $s<0.3$ or $k<0.3\,n$. It should be noted that the net works better for smaller a and b. Mere correlation learning (5.1) does not work well, even if the thresholds are adequately tuned.

Association by covariance learning, however, has the following shortcomings: 1) Even if $\delta=0$, i.e., the correct cue pattern is input, the net outputs a noisy

pattern, $\delta' \neq 0$. 2) When the set of pattern pairs cannot be regarded as a typical sample from randomly generated patterns, it does not work well.

Association by orthogonal learning does not share these shortcomings. It works well for an arbitrary (non-random) set of pattern pairs, as long as the x_α's are linearly independent. When $\delta=0$, the output patterns are noiseless $\delta'=0$. However, the above performances are guaranteed only when ε is neglisibly small. When the netself-organizes under the influence of noise, ε increases to $\varepsilon'=\varepsilon+\sigma$. Therefore, it is not realistic to ignore the term of order ε.

Since x_α^*'s depend on all the x_α's in a very complicated manner, it is difficult to evaluate the noise immunity in the general situation. We again consider the case where the (x_α, z_α)'s are randomly generated, though the net works well for non-randomly generated pattern sets. We treat the case where ε and δ are so small that their higher order terms can be neglected. We then have the following noise relation

$$\delta' \leqq F\left[\frac{a(1-a)-\varepsilon' s+\delta}{2\sqrt{abs\delta(1-a+s+sa)}}\right], \qquad (5.6)^2$$

which is shown in Figure 6, where $a=b=0.2$, $\varepsilon'=0.1$. For small δ, the performance is very good, even when $s=0.5$. In the case of $s=0.1$, even the fifteen percent noise is completely reduced. However, this evaluation holds only when $\varepsilon'=\varepsilon+\sigma$ is not so large.

5.3. Dynamics of a Concept Formation Net

When each cluster of patterns is so sharply distributed that $\sigma \fallingdotseq 0$, $\bar{x}_\alpha \fallingdotseq x_\alpha$ hold, the connection matrix (4.9) reduces to

$$W = e\sum x_\alpha x_\alpha^* + O(\varepsilon).$$

Therefore, for small ε, $Tx_\alpha = x_\alpha$ $(\alpha=1, \ldots, k)$ holds. This shows that each pattern x_α becomes an equilibrium state of the net. By receiving an input pattern x, the net will tend to the state x_α which is closest to it. The net can store k patterns x_α in this manner. It can reproduce any of them correctly from a noisy input. The net may be regarded as a pattern generator, in which k patterns are coded.

In the case where the dispersion σ of the clusters is not neglisibly small, the net works only when $\sigma<\varepsilon$. Otherwise, the term $[\sigma/(\sigma+\varepsilon)]E$ dominates in W, which then gives no information about the necessary patterns. We may say in this sense that ε gives the degree of resolution of patterns: Only a pattern cluster whose dispersion σ is smaller than ε is treated as comprising versions of only one pattern. It is difficult for the net to recognize a cluster of patterns whose dispersion is larger than ε. The parameter ε, on the other hand, must be small, because otherwise we cannot neglect the harmful interference term $O(\varepsilon)$. Therefore, orthogonal learning works well for a trainable pattern generator in which the patterns to be generated are precisely taught. The cerebellum might have this mechanism in which the patterns are taught by the cerebrum.

When the net treats clusters of largely dispersed patterns, covariance learning works better. We hence analyze the behavior of the net with covariance learning. We consider again the case where k patterns x_α are randomly generated. We moreover assume for the sake of simplicity in calculation that the connection weight w_{ii} has a fixed self-inhibition term of amount $-e'/4$.

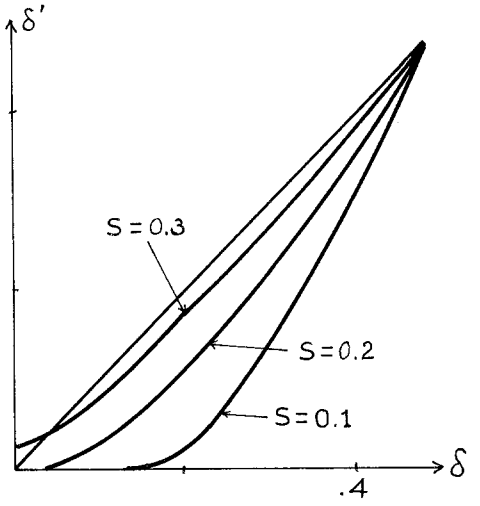


Fig. 5. Noise reduction rate of an association net with covariance learning

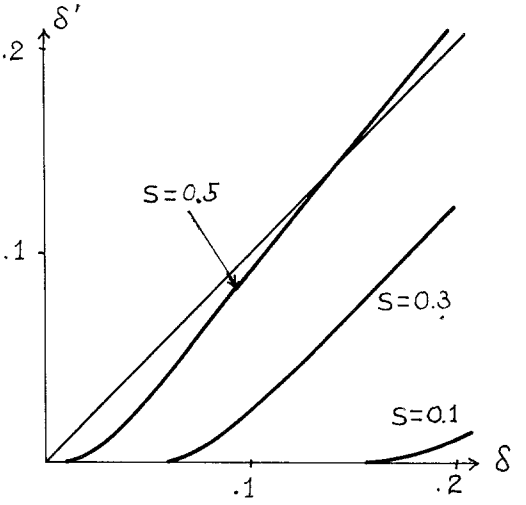


Fig. 6. Noise reduction rate of an association net with orthogonal learning

[2] The derivation is complicated and involves some stochastic assumptions

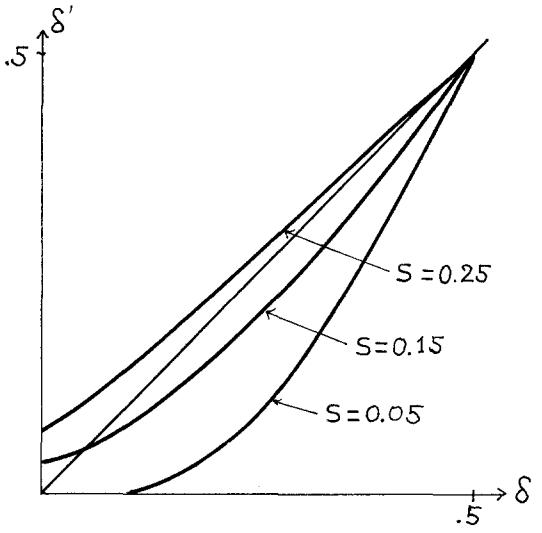


Fig. 7. Noise reduction rate of a concept-formation net with covariance learning

Assume that, in the neighborhood of a pattern x_α, the relation

$$\delta' = g(\delta), \tag{5.7}$$

holds, where δ and δ' are, respectively, the noise rates of states before and after state transition. Then, the noise rate δ_t at time t obeys the dynamical equation

$$\delta_{t+1} = g(\delta_t), \tag{5.8}$$

because the state $x(t+1)$ of the net is determined from $x(t)$. We can analyze the dynamical behavior of the net through the change in the noise rate. When δ_t of (5.8) converges to an equilibrium δ_0, the net tends to and remains in states close to x_α whose noise rate is δ_0. We can say in this case that the net can retain pattern x_α persistently within an error of noise rate δ_0. Unfortunately, there does not exist an exact relation like (5.7). We obtain instead the relation

$$\delta' \leqq g(\delta),$$

where

$$g(\delta) = F\left[\frac{k p_\alpha (0.5-\delta)(0.5-\delta_\alpha)^2}{\sqrt{s\gamma(0.5+\delta)}}\right] \tag{5.9}$$

for patterns around x_α, where we put $a=0.5$,

$$\gamma = k \sum_\beta p_\beta^2 (0.5-\delta_\beta)^4$$

and δ_α represents the dispersion of patterns around x_α.

We show the function $g(\delta)$ in Figure 7, where $p_\alpha = 1/k$, $\delta_\alpha = 0.2$ for all α. Since we have put $a=b=0.5$, the result is not so good compared to covariance association learning. For $s=0.1$, dynamical Equation (5.8) has a stable equilibrium at $\delta=0.01$. Therefore, the net forms the corresponding concept within a 1% noise rate. For $s=0.15$, the equation has a stable equilibrium at $\delta=0.06$, so that the concept pattern is formed in the net within a 6% noise rate. However, for $s=0.2$, there are no equilibria except for $\delta=0.5$. This means that the net fails to form concept patterns in the case when $s>0.2$, or $k>0.2n$. The situation seems better for for small a and b.

Appendix I. Derivation of (3.14)

Let X be a $k \times n$ matrix consisting of k column vectors x

$$X = [x_1, x_2, \ldots, x_k].$$

Let P be a $k \times k$ diagonal matrix, whose diagonal entries are p_α. We denote by $\sqrt{P}$ the diagonal matrix whose diagonal entries are $\sqrt{p_\alpha}$. Let z be the row vector

$$z = (z_1, z_2, \ldots, z_k).$$

We easily have

$$\langle x x^T \rangle = XPX^T = (X\sqrt{P})(X\sqrt{P})^T,$$
$$\langle z x^T \rangle = zPX^T = z\sqrt{P}(X\sqrt{P})^T,$$
$$G = X^T X/n.$$

Equation (3.12) is rewritten as

$$w = ezPX^T(\varepsilon E + XPX^T)^{-1}.$$

We can use the identity

$$(X\sqrt{P})^T(\varepsilon E + XPX^T)^{-1} = (\varepsilon E_k + \sqrt{P}X^T X\sqrt{P})^{-1}\sqrt{P}X^T,$$

where E_k is the $k \times k$ identity matrix, and the expansion formula

$$(\varepsilon E_k + A)^{-1} = A^{-1} - \varepsilon A^{-2} + O(\varepsilon^2),$$

where A is a non-singular $k \times k$ matrix. (Since $k<n$, the $n \times n$ matrix XPX^T is singular, while the $k \times k$ matrix $\sqrt{P}X^T X\sqrt{P}$ is non-singular.) We obtain

$$w = ez(X^T X)^{-1}X^T - e\varepsilon z(X^T X)^{-1}P^{-1}(X^T X)^{-1} + O(\varepsilon^2).$$

Taking account of $(X^T X)^{-1}X^T = G^{-1}X^T$, and rewriting the above w in the vector notation, we have (3.14). It should be noted that the $k \times n$ matrix $X^+ = G^{-1}X^T$, which is composed of k row vectors x_α^*'s, is the generalized inverse of X [Albert (1972), see also Kohonen (1972)].

Appendix II. Derivation of (4.9)

Let $\bar{X}$ be the matrix defined by $\bar{X} = [\bar{x}_1, \ldots, \bar{x}_k]$. We have

$$W = e(\sigma E + \bar{X}P\bar{X}^T)[(\sigma+\varepsilon)E + \bar{X}P\bar{X}^T]^{-1}.$$

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the non-zero eigenvalues of the matrix XPX^T, and let $e_1, \ldots, e_k$ be the corresponding eigenvectors. We add $n-k$ vectors $e_{k+1}, \ldots, e_n$ such that $\{e_i\}$ $i=1, \ldots, n$ forms a set of orthonormal basis vectors. We then have

$$\bar{X}P\bar{X}^T = \sum_{i=1}^{k} \lambda_i e_i e_i^T,$$

$$E = \sum_{i=1}^{n} e_i e_i^T.$$

By the use of the relations

$$\left(\sum_{i=1}^{n} \mu_i e_i e_i^T\right)^{-1} = \sum_{i=1}^{n} \frac{1}{\mu_i} e_i e_i^T ,$$

$$(\sum \mu_i e_i e_i^T)(\sum \nu_i e_i e_i^T) = \sum \mu_i \nu_i e_i e_i^T ,$$

we obtain

$$W = e\left(\sum_{i=1}^{k} \frac{\sigma+\lambda_i}{\sigma+\varepsilon+\lambda_i} e_i e_i^T + \sum_{i=k+1}^{n} \frac{\sigma}{\sigma+\varepsilon} e_i e_i^T\right).$$

Expanding the above, where $\sigma+\varepsilon$ is assumed to be small, we obtain

$$W = e\sum_{i=1}^{n} \frac{\sigma}{\sigma+\varepsilon} e_i e_i^T + e\sum_{i=1}^{k} \frac{\varepsilon}{\sigma+\varepsilon} e_i e_i^T - \varepsilon e \sum_{i=1}^{k} \frac{1}{\lambda_i} e_i e_i^T + O[\varepsilon(\varepsilon+\sigma)] .$$

Here, we put $\sigma = 0$ and compare the result carefully with (4.2) or (4.7). We then have (4.9).

Conclusion

We have used two types of nets with recurrent connections and non-recurrent connections as models of association and concept-formation, respectively. We proposed two types of self-organization, one called orthogonal learning and the other covariance learning. We have analyzed the behavior of these nets and elucidated possible neural mechanism of information processing including association and concept-formation.

The present models are still too simple to be compared with the actual brain. However, the actual brain might have the mechanisms analyzed here commonly in its various portions. The present analysis will help in building more realistic models of the brain because it provides a possible neural logic.

Acknowledgement. This research was initiated at the University of Tokyo and completed during the author's visit to the Center for Systems Neuroscience of the University of Massachusetts on sabbatical leave from the University of Tokyo. The author would like to thank Professor M. A. Arbib for helpful comments on the manuscript.

References

Albert, A.: Regression and the Moore-Penrose pseudoinverse. New York: Academic Press 1972

Amari, S.: Characteristics of randomly connected threshold element networks and network systems. Proc. IEEE **59**, 35—47 (1971)

Amari, S.: Learning patterns and pattern sequences by self-organizing nets of threshold elements. IEEE Trans. Comp. C-**21**, 1197—1206 (1972a)

Amari, S.: Characteristics of random nets of analog neuron-like elements. IEEE Trans. on Systems, Man, and Cybernetics SMC-**2**, 643—657 (1972b)

Amari, S.: A method of statistical neurodynamics. Kybernetik **14**, 201—215 (1974a)

Amari, S.: A mathematical theory of nerve nets. In: Kotani, M. (Ed.): Advances in biophysics, Vol. 6, pp. 75—120. Tokyo: University Press 1974b

Amari, S.: Homogeneous nets of neuron-like elements. Biol. Cybernetics **17**, 211—220 (1975)

Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. To be published (1977a)

Amari, S.: A mathematical approach to neural systems. In: Arbib, M.A., Metzler, J. (Eds.): Systems neuroscience. I. New York: Academic Press 1977b

Amari, S., Yoshida, K., Kanatani, K.: Mathematical foundation of statistical neurodynamics. SIAM J. Appl. Math. **33** (1977)

Anderson, J.A.: A simple neural networks generating interactive memory. Math. Biosc. **14**, 197—220 (1972)

Caianiello, E.R.: Outline of a theory of thought processes and thinking machines. J. theor. Biol. **1**, 204—235 (1961)

Grossberg, S.: Onlearning, information lateral inhibition, and transmitters. Math. Biosc. **4**, 255—310 (1969)

Hebb, D.O.: The organization of behavior. New York: Wiley 1949

Kohonen, T.: Correlation matrix memories. IEEE Trans. Comp. C-**21**, 353—359 (1972)

Kohonen, T.: An adaptive associative memory principle. IEEE Trans. Comp. C-**23**, 444—445 (1974)

Kohonen, T., Oja, E.: Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. Biol. Cybernetics **21**, 85—95 (1976)

Malsburg von der, C.: Self-organization of orientation sensitive cells in the striate cortex. Kybernetik **14**, 85—100 (1973)

Nakano, K.: Associatron—a model of associative memory. IEEE Trans. on Systems, Man, and Cybernetics SMC-**2**, 380—388 (1972)

Pfaffelhuber, E.: Correlation memory models—a first approximation in a general learning scheme. Biol. Cybernetics **18**, 217—223 (1975)

Rosenblatt, F.: Principles of Neurodynamics. Washington: Spartan 1961

Spinelli, N.: OCCAM: a computer model for a content-addressable memory in the central nervous systems. Biology of memory 293—306. New York: Academic Press 1970

Uesaka, Y., Ozeki, K.: Some properties of association-type memories. JIEECE of Japan (in Japanese), **55**-D, 323—330 (1972)

Wasan, M.T.: Stochastic approximation. Cambridge: University Press 1969

Wigström, H.: A neuron model with learning capability and its relation to mechanism of association. Kybernetik **12**, 204—215 (1973)

Willshaw, D.J., Longuet-Higgins, H.C.: The holophone—recent developments. In: Metzler, B., Michie, D. (Eds.): Machine intelligence, Vol. 4, pp. 349—357. Edinburgh: University Press 1969

Received: December 12, 1976

Dr. Shun-ichi Amari
Dept. of Mathematical Engineering
and Instrumentation Physics
University of Tokyo
Tokyo 113 Japan

Verantwortlich für den Textteil: Prof. Dr. W. Reichardt, Max-Planck-Institut für biologische Kybernetik, Spemannstr. 38, D-7400 Tübingen. Verantwortlich für den Anzeigenteil: L. Siegel, G. Sternberg, Springer-Verlag, Kurfürstendamm 237, D-1000 Berlin 15, Fernsprecher: (030) 882 1031, Telex: 01-85411. Springer-Verlag, Berlin · Heidelberg · New York.
Druck der Brühlschen Universitätsdruckerei, Lahn-Gießen. Printed in Germany. —© by Springer-Verlag Berlin Heidelberg 1977