

THEORY OF SELF-ORGANIZING NERVE NETS WITH SPECIAL REFERENCE TO ASSOCIATION AND CONCEPT FORMATION

Shun-ichi Amari

Department of Mathematical Engineering
University of Tokyo
Tokyo, 113 Japan

The present paper proposes a mathematical theory of self-organizing nerve nets, which is applicable to various types of supervised and unsupervised learning, such as learning decision, concept formation, association, etc. Given an environmental information source, a neural system automatically forms a number of separate routines to process the signals in it. This kind of unsupervised self-organization underlies commonly in formation of categories, feature extractors, and content addressable memories. This problem is analyzed mathematically, as well as models of topographic organization of nerve fields and of associative memories, by the proposed method.

1. INTRODUCTION

The brain needs a long period of evolution to get the present information processing manner. Although artificial intelligence is recently investigated rather independently of the behavior of nerve nets, it is plausible that artificial and natural intelligence have some common logic of realizing intelligence at a higher abstract level. They have something to suggest to each other. Neural systems seem, especially, to yield good models of parallel and distributed information processing with self-organizing capabilities of adapting themselves to the environmental information structure.

The present paper proposes a unified mathematical treatment of neural self-organization applicable to wide classes of learning with and without teacher, by summarizing the author's recent researches [1 - 5]. We also present a model which elucidates the mechanism of unsupervised and automatic formation of various routines for processing the signals in the environment [cf. 4, 6, 7], and analyze the model by the proposed method. Recent developments of neurophysiology have revealed the importance of such self-organization. We also touch upon the models of topographic organization of nerve fields [3, 8] as well as of associative memories [9, 10].

2. A GENERAL THEORY OF SYNAPTIC MODIFICATION

We consider a simple mathematical model of a neuron, which receives a vector input signal $\underline{x} = (x_1, \dots, x_n)$ from its environment S , and emits one output signal z . These signals in general

take analog values between 0 and 1, representing the normalized pulse emission rates (they may take on 0 and 1). Let us denote by $\underline{w} = (w_1, \dots, w_n)$ the synaptic efficiencies or weights of the inputs. The input-output relation is given by

$$z = f(\underline{w} \cdot \underline{x} - h) \quad (1)$$

where h is a threshold, f is a monotonically increasing function transforming the average membrane potential $u = \underline{w} \cdot \underline{x}$ into the output pulse emission rate z (Fig.1).

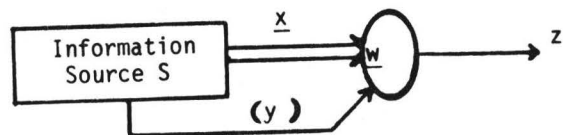


Fig.1 Model of neuron

The neuron self-organizes and adapts to its environment S , by modifying $\underline{w}$ based on an input time sequence $\underline{x}(t)$ from S . In some cases, the so called teacher signal $\underline{y}(t)$ is associated with $\underline{x}(t)$. Hence we regard S as an ergodic information source producing $(\underline{x}(t), \underline{y}(t))$, where teacher signal $\underline{y}(t)$ is absent in the case of unsupervised learning. We consider a simple environment S represented by a probability (density) $p(\underline{x}, \underline{y})$. It chooses a pair $(\underline{x}, \underline{y})$ of signals with the probability $p(\underline{x}, \underline{y})$, each time independently, and outputs it for a fixed time duration, and then choose another, repeating this process.

We propose the following general rule of synapse modification, $\underline{w}$ being changed based on $\underline{x}(t)$ as

$$\dot{\underline{w}}(t) = -\underline{w}(t) + cr(t)\underline{x}(t) \quad (2)$$

where τ is a large time constant, $\dot{\cdot}$ denotes the time derivative d/dt , c is a constant, and r is the reinforcement or learning signal which the neuron produces depending on w , x and y (when it exists) at that time, $r(t) = r[w(\tau), x(t), y(t)]$. This is a generalization of Hebbian type learning, which is obtained by putting $r = z$. Perceptron type learning is obtained by $r = y - f(w \cdot x - h)$. Most of the neural learning rules proposed so far are obtained by choosing appropriate signals r .

By replacing $\underline{rx}$ in (2) by its average, we obtain

$$\tau \dot{w} = -w + c \langle r(w, x, y) \rangle_S, \quad (3)$$

where $\langle \rangle_S$ denotes the average with respect to $p(x, y)$ of S , where w is considered as constant. This is not a random equation. Since S is ergodic, we can expect that the real behavior of $w(t)$ of (2) is closely approximated by that of (3). This really is proved by the method of stochastic approximation or by the theorem of Geman [11]. Hence, we consider the average learning equation (3) as the fundamental of neural self-organization.

When r depends on w and x through $w \cdot x$ only, $r = r(w \cdot x, y)$, (3) is rewritten in the form

$$\dot{w} = -\partial L(w) / \partial w \quad (4)$$

where

$$L(w) = w^2 / 2 - c \langle \int_0^{w \cdot x} r(u, y) du \rangle_S. \quad (5)$$

We call $L(w)$ the potential of learning under S . The synaptic weight w converges to one of the minimum of L . In most supervised cases, L has only one minimum, w converging to it. However, in most unsupervised cases, L has a number of minima corresponding to various aspects of S , and w converges to one of them. However, when a pool of neurons receive common inputs from S , every minimum of L is occupied by some neurons in the pool, so that the pool of neurons adapts as a whole to the environment S .

3. FORMATION OF CATEGORY DETECTING CELLS

Information processing routines are formed in the brain to be compatible with the environment by learning. This kind of learning without teacher underlies commonly in the problem of concept formation, formation of feature detectors, etc. Recent developments of neurophysiology have revealed that the feature detecting cells are really formed depending on the visual experiences of an animal. We propose a simple model, and analyze its capabilities.

The model consists of a set of neurons receiving a common input signal x from the environment $S = \{p(x)\}$. They also receive an inhibitory

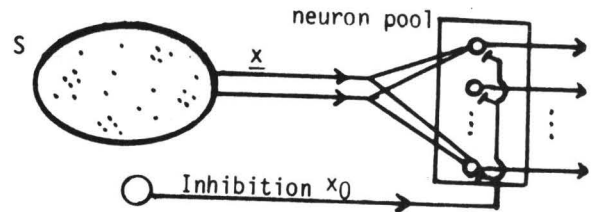


Fig.2 Category formation

input of constant intensity x_0 (Fig.2). Let us take one neuron, and let w be the modifiable synaptic weights of the neuron. Let $-w_0$ be the modifiable synaptic weight for the inhibitory input x_0 .

We assume the following learning rule that the learning signal r is equal to 1 when the neuron is excited ($z > 0$) and 0 otherwise ($z = 0$), i.e.,

$$r = Z(w \cdot x - w_0 x_0), \quad (6)$$

where $Z(u)$ is 1 when $u > 0$ and 0 otherwise, and we put $h = 0$. The average learning equation is

$$\dot{w} = -w + c \langle Z(w \cdot x - w_0 x_0) x \rangle_S, \quad (7)$$

$$\dot{w}_0 = -w_0 + c x_0 \langle Z(w \cdot x - w_0 x_0) \rangle_S. \quad (8)$$

The equations are in general multi-stable, having a number of equilibria.

Let us consider a subset A of the signals in S . A nerve cell is said to be a detector or a representative cell of A , when it is excited by any signal in A but not excited by any others in S . We say that a subset A forms a category under S , when a detector of A is obtained as a stable equilibrium by self-organization. In other words, a subset A is a category, when a set of weights w^A, w_0^A satisfying

$$w^A \cdot x - w_0^A x_0 > 0$$

for all $x \in A$ but not for any $x \in S - A$, is obtained as stable equilibrium of (7) and (8).

Many categories are formed by learning under S . When the neuron pool includes many neurons of different initial weights, they differentiate automatically to be detectors (representatives or processors) of these categories. Mutual interaction among the neurons of lateral inhibition type prevents a large number of neurons becoming the detector of one and the same category. The problem is to know which subsets become categories under S . We give a n.a.s.c. for a subset A to be category under S . Let x^A be the average of signals in A . $x^A = \int p(x) x dx$. Let $g_A(x) = x^A \cdot x$ be the inner product, and $\lambda = c x_0^2 / c$.

Theorem. A necessary and sufficient condition that A is a category under S is

$$\min_{x \in A} g_A(x) > \lambda > \max_{y \in S-A} g_A(y).$$

The theorem shows that a cluster of signals of an adequate diameter forms a category, where the diameter is specified by the parameter λ . In other words, λ specifies the resolution of a category, where λ can be controlled by changing x_0 . This yields a primitive mechanism of formation of information processing routines by learning.

4. TOPOGRAPHIC ORGANIZATION

Topographic organization is found in many parts of nerve fields in the brain. For example, the continuous projection from the retina to the striate cortex is known as the retinotopy. It is known that, when part of the retina (or striate cortex, or both) is removed, the regeneration of connections takes place such that the topological map is completed between the remaining parts. This suggests that topographic organization is formed based, at least partly, on the self-organizing ability of nerve cells.

Let us consider a simple model consisting of a presynaptic nerve field X and postsynaptic nerve field Y connected by modifiable synapses. Stimuli are supplied from the environmental information source S to the presynaptic field X , exciting neurons of X , and then those of Y . The self-organization takes place based on these excitations. In this case, S brings the topological information of X , in such a way that two nearby neurons of X are frequently stimulated at the same time. When the neurons in Y have fixed mutual connections of lateral-inhibition type and when they have additional modifiable inhibitory synapses, we can prove that continuous connections from X to Y (topographic organization) is formed by Hebbian type unsupervised learning. Moreover, we can prove that some microstructures are also formed in some cases, as is shown in the columnar structures in the cerebrum. The fundamental equations of the model are much complicated, because we should solve the field equations of neural excitation together with the average learning equations. We can solve them.

5. PRIMITIVE MODEL OF ASSOCIATION

Let us consider a pool of neurons, which receive a common input signal x and output a vector signal z (the i -th neuron emitting the i -th component z_i). We consider such an environment S that consists of a finite number of signal pairs (x_j, y_j) , where the i -th component y_{ji} of y_j acts as the teacher signal to the i -th neuron. The problem is to modify the synaptic

weights w_{ij} of the i -th neuron (or synaptic weight matrix $W = (w_{ij})$) such that, receiving an input x_j , the neuron pool outputs the associated signal y_j . In this case, one may say that the net recalls y_j from the associated x_j , and that the associated pairs (x_j, y_j) 's are memorized in the synaptic weights W in a distributed and superimposed manner [10].

The correlation learning [10] is obtained, if each neuron self-organizes with $r = y$. In this case, W converges to $W = \sum y_j x_j^T$, T denoting the transposition. This learning works well only when x_j 's are mutually orthogonal. The synaptic weights converges to $W = \sum y_j x_j^*$ by orthogonal learning [2] with $r = y - W \cdot x$, where $\{x_j^*\}$ is the dual orthogonal system of $\{x_j\}$ (i.e., $x_j^* \cdot x_i = \delta_{ji}$), and associative recall can be done even when x_j 's are not orthogonal. We can also treat the dynamic recall of signal sequences $x_1, x_2, \dots, x_t$ [9].

REFERENCES

- [1] S. Amari "A Method of Statistical Neurodynamics". *Kybernetik*, 14 (1974) 201-215
- [2] S. Amari "Neural Theory of Association and Concept-formation". *Biol. Cybernetics*, 26 (1977) 175-185
- [3] S. Amari "Topographic Organization of Nerve Fields". to appear
- [4] S. Amari and A. Takeuchi "Mathematical Theory of Formation of Category Detecting Nerve Cells". *Biol. Cybernetics*, 29 (1978) 127-136.
- [5] A. Amari and M.A. Arbib "Competition and Cooperation in Neural Nets". In *Systems Neuroscience* (J. Netzler ed.), Academic Press, (1977) 119-165
- [6] C. von der Malsburg "Self-organization of Orientation Sensitive Cells in the Striate Cortex". *Kybernetik*, 14, (1973) 85-100
- [7] K. Fukushima "Cognitron: A Self-organizing Multilayered Neural Network". *Biol. Cybernetics*, 20 (1975) 121-136
- [8] A. Takeuchi and S. Amari; to appear
- [9] S. Amari "Learning Patterns and Pattern Sequences by Self-organizing Nets of Threshold Elements". *IEEE Trans.*, C-21 (1972) 1197-1206
- [10] T. Kohonen *Associative Memory*. Springer-Verlag, 1977
- [11] S. Geman "Some Averaging and Stability Results for Random Differential Equations". *SIAM J. Math.* 36 (1979) 86-105