ON THE PATTERN CLASSIFICATION PROBLEMS BY LEARNING I

Tanaka, Kensuke
Department of Mathematics, Niigata University | Department of Mathematics, Kyushu

https://doi.org/10.5109/13043

出版情報：統計数理研究. 14 (1/2), pp.31-49, 1970-03. Research Association of Statistical Sciences

ON THE PATTERN CLASSIFICATION PROBLEMS BY LEARNING (I)†

By

Kensuke TANAKA*

Dedicated to Professor Tosio Kitagawa on his 60-th birthday

(Received January 20, 1970)

§ 1. Introduction and Summary.

In recent years considerable attention has been given to the problems of learning in pattern recognition. All of these problems involve to determine an algorithm of classification which patterns become to be collected into a set of groups according to the accumulation of knowledge. In this paper, each group is called a category. Then each pattern is a random sample from the group to which they belong. Therefore, corresponding to each category, there is a probability distribution law by which each observed pattern is drawn. In the case when all of these distribution laws of the categories are known, a classification rule was introduced according to the methods of statistical decision theory by several authors ([12], [19]), and also in the case when all of the distributions have unknown parameters but a priori distributions of the parameters are given to us, a Bayes approach is valid to determine a classification rule. ([20], [28])

However, we are confronted with the case when there is not so much amount of information as we can apply a Bayes approach. When there can be assumed a "training" sequence of observed patterns with their correct classification given by an external indication, it has been tried to find an algorithm by which a function converging to the decision function, which is optimal in the Bayes sense, can be constructed on the basis of a training sequence. ([18], [20], [22]) In general, this approach has been called "learning with a teacher" in the classification problems in pattern recognition. In this approach, many authors used a linear system of linearly independent functions or orthonormal functions and developed learning theories for the pattern classification problems on the basis of a training sequence associated with independent, identically distributed random variables. The algorithm introduced by them is a kind of applications of stochastic approximation theory.

On the other hand, when such a training sequence cannot be assumed, by several authors ([23], [24], [26]) the various decision rules and algorithms were developed

* Department of Mathematics, Kyushu University, Fukuoka and Department of Mathematics, Niigata University, Niigata.

† This paper was prepared while the author was at Kyushu University on leave from Niigata University.

on the basis of a independent sequence of unclassified, observed patterns. This approach has been called "learning without a teacher" or "self-learning" in the problems of pattern recognition.

However, the assumption that a sequence of patterns is observed from the independent distributed random variables does not seem to us to cover the whole real situation in pattern recognition. On this reason, in this paper, we shall treat the pattern classification problems on the basis of a dependent sequence of observed patterns. In the case of learning with a teacher, we give a sequence of linear systems of orthonormal functions as an approximation to the decision function which is optimal in the Bayes sense. In the case of self-learning, we propose a decision function in place of the function optimal in the Bayes sense and try to give an approximation to the decision function proposed. Our algorithm is an application of the method introduced by T. Kitagawa [19] in the successive process of statistical control. This method, which may be called modified stochastic approximation, was investigated by V. Dupač [12] in detail.

This paper consists of six sections. In Section 2, we shall give several lemmas necessary for the proofs of main results in this paper. In Section 3, we shall give preparatory expositions concerning a model of the pattern classification problems and an application of the Bayes decision rule for our model. In Section 4, we shall investigate the pattern classification problems in which there are two categories and is assumed a training sequence. In Section 5, we shall be also concerned with the pattern classification problems when a training sequence cannot be assumed.

§ 2. Preliminaries.

In this section three lemmas are given and proved for the sake of the proofs of main results of this paper. Let us consider a real valued stochastic process $\{y^n\}_{n=1}^{\infty}$ and three sequences of non-negative real valued measurable functions $\{U_n\}_{n=1}^{\infty}$, $\{V_n\}_{n=1}^{\infty}$ and $\{\zeta_n\}_{n=1}^{\infty}$, where each U_n, V_n and ζ_n are measurable functions defined on R^n for every n.

Let us write $U_n = U_n(y^1, \cdots, y^n)$, $V_n = V_n(y^1, \cdots, y^n)$ and $\zeta_n = \zeta_n(y^1, \cdots, y^n)$ for the sake of simplicity. We denote the expected values of three stochastic variables U_n, V_n and ζ_n by $E[U_n]$, $E[V_n]$ and $E[\zeta_n]$. Furthermore, we denote the conditional expectation of three stochastic variables U_{n+1}, V_{n+1} and ζ_{n+1} given the variables $y^1, y^2, y^3, \cdots, y^n$ by $E[U_{n+1}|y^1 \cdots y^n]$, $E[V_{n+1}|y^1 \cdots y^n]$ and $E[\zeta_{n+1}|y^1 \cdots y^n]$.

In what follows, let $\{\gamma_n\}_{n=1}^{\infty}$ and $\{\mu_n\}_{n=1}^{\infty}$ be two sequences of real numbers. Now, we introduce the fundamental conditions for three stochastic processes $\{U_n\}_{n=1}^{\infty}$, $\{V_n\}_{n=1}^{\infty}$ and $\{\zeta_n\}_{n=1}^{\infty}$:

(A1) $E[U_1]$ and $E[V_1]$ exist,

(A2) $E[U_{n+1}|y^1 \cdots y^n] \leqq (1+\mu_n)U_n - \gamma_n V_n + \zeta_n$ hold for all n,

(A3) $\sum_{n=1}^{\infty} \gamma_n = \infty$ $(\gamma_n \geqq 0,\ n=1, 2, \cdots)$ and $\sum_{n=1}^{\infty} |\mu_n| < \infty$,

(A4) There exists a sequence of positive numbers $\{M_n\}_{n=1}^{\infty}$ such that $P[\zeta_n \leqq M_n] = 1$ for $n = 1, 2, \cdots$, (A5) $\sum_{n=1}^{\infty} M_n < \infty$.

LEMMA 1. *Let the following conditions for three stochastic processes* $\{U_n\}_{n=1}^{\infty}$, $\{V_n\}_{n=1}^{\infty}$ *and* $\{\zeta_n\}_{n=1}^{\infty}$ *be satisfied:*

(2.1) *Conditions* (A1)$\sim$(A5) *hold,*

(2.2) $\lim_{n\to\infty} \gamma_n = 0$,

(2.3) *If there exists a subsequence* $\{n_k\}_{k=1}^{\infty}$ *of a sequence* $\{n\}_{n=1}^{\infty}$ *such that* $P[\lim_{k\to\infty} V_{n_k}=0]=1$, *then* $P[\lim_{k\to\infty} U_{n_k}=0]=1$,

Then

$$P[\lim_{n\to\infty} U_n=0]=1 \text{ and for all } 0<\beta<1,\ \lim_{n\to\infty} E[U_n^{\beta}]=0.$$

PROOF. Let us write $\Lambda_n=\prod_{k=n}^{\infty}(1+|\mu_k|)U_n$ for the sake of simplicity. Since by (A2)

$$\begin{aligned}
E[\Lambda_{n+1}] &= E\{E[\Lambda_{n+1}\,|\,y^1,\cdots,y^n]\}\\
&= E\Big\{\prod_{k=n+k}^{\infty}(1+|\mu_k|)E[U_{n+1}\,|\,y^1,\cdots,y^n]\Big\}\\
&\leqq \prod_{k=n+1}^{\infty}(1+|\mu_k|)E[(1+|\mu_n|)U_n-\gamma_n V_n+\zeta_n]\\
&= \prod_{k=n}^{\infty}(1+|\mu_k|)E[U_n]-\gamma_n\prod_{k=n+1}^{\infty}(1+|\mu_k|)E[V_n]\\
&\quad+\prod_{k=n+1}^{\infty}(1+|\mu_k|)E[\zeta_n],
\end{aligned}$$

we have for all n

$$E[\Lambda_{n+1}]\leqq E[\Lambda_n]-\gamma_n E[V_n]+CE[\zeta_n], \tag{2.4}$$

where

$$C=\prod_{k=1}^{\infty}(1+|\mu_k|).$$

Then, the repeated applications of the inequalities (2.4) gives us

$$0\leqq E[\Lambda_{n+1}]\leqq E[\Lambda_1]-\sum_{i=1}^{n}\gamma_i E[V_i]+C\sum_{i=1}^{n}E[\zeta_i]. \tag{2.5}$$

From (2.5), we have for all n

$$0\leqq \sum_{i=1}^{n}\gamma_i E[V_i]\leqq E[\Lambda_1]+C\sum_{i=1}^{n}E[\zeta_i].$$

Hence, from (A1), (A4) and (A5) we have

$$\sum_{n=1}^{\infty}\gamma_n E[V_n]<\infty. \tag{2.6}$$

Therefore, by (2.6) and (A3), there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ of a sequence $\{n\}_{n=1}^{\infty}$ such that

$$\lim_{k\to\infty} E[V_{n_k}]=0.$$

This implies that, since V_{n_k} is non-negative, there exists a subsequence $\{n_{k_s}\}_{s=1}^{\infty}$ of the sequence $\{n_k\}_{k=1}^{\infty}$ such that

(2.7) $$P[\lim_{s\to\infty} V_{n_{k_s}}=0]=1.$$

From (2.3), this sequence $\{n_{k_s}\}_{s=1}^{\infty}$ satisfies

(2.8) $$P[\lim_{s\to\infty} U_{n_{k_s}}=0]=1.$$

Next, putting

$$W_n=\Lambda_n+C\Big(\zeta_n+\sum_{k=n+1}^{\infty}E[\zeta_k\,|\,y^1,\cdots,y^n]\Big),\qquad C=\prod_{k=1}^{\infty}(1+|\mu_k|),$$

and noting that

$$\begin{aligned}
&E[W_{n+1}\,|\,y^1,\cdots,y^n]\\
&=E\Big[\Lambda_{n+1}+C\Big(\zeta_{n+1}+\sum_{k=n+2}^{\infty}E[\zeta_k|y^1,\cdots,y^{n+1}]\Big)\Big|y^1,y^2,\cdots,y^n\Big]\\
&=E[\Lambda_{n+1}|\,y^1,\cdots,y^n]+C\Big[E[\zeta_{n+1}|y^1,\cdots,y^n]+\sum_{k=n+2}^{\infty}E\{E[\zeta_k|y^1,\cdots,y^{n+1}]\,|\,y^1,\cdots,y^n\}\Big]\\
&=\prod_{k=n+1}^{\infty}(1+|\mu_k|)E[U_{n+1}|y^1,\cdots,y^n]+C\sum_{k=n+1}^{\infty}E[\zeta_k|y^1,\cdots,y^n]\\
&\leqq\prod_{k=n+1}^{\infty}(1+|\mu_k|)\{(1+|\mu_n|)U_n+\zeta_n\}+C\sum_{k=n+1}^{\infty}E[\zeta_k|y^1,\cdots,y^n]\\
&\leqq\Lambda_n+C\Big(\zeta_n+\sum_{k=n+1}^{\infty}E[\zeta_k|y^1,\cdots,y^n]\Big)\\
&=W_n,
\end{aligned}$$

we know that such stochastic process $\{W_n\}_{n=1}^{\infty}$ is a lower semi-martingale. By a property of a lower semi-martingale there exists a random variable Λ^* such that

(2.9) $$P[\lim_{n\to\infty} W_n=\Lambda^*]=1$$

and

(2.10) $$E[\Lambda^*]<\infty.$$

We can easily prove, from (A4) and (A5), that

$$P[\lim_{n\to\infty}\zeta_n=0]=1$$

and

$$P\Big(\lim_{n\to\infty}\sum_{k=n+1}^{\infty}E[\zeta_k|y^1,\cdots,y^n]=0\Big)=1.$$

Therefore, the definition of the stochastic process $\{W_n\}_{n=1}^{\infty}$ gives us

(2.11) $$P[\lim_{n\to\infty}\Lambda_n=\Lambda^*]=1.$$

Using the definition of Λ_n, (A3) and (2.9) we have

(2.12) $$P[\lim_{n\to\infty} U_n=\Lambda^*]=1.$$

Further, from (2.8) and (2.12), we have

(2.13) $$P[\lim_{n\to\infty} U_n = 0] = 1 .$$

Also using (2.13), the uniformely integrable property of $\{U_n\}_{n=1}^{\infty}$, $0<\beta<1$ and the non-negative property of U_n, we have

$$\lim_{n\to\infty} E[U_n^\beta] = E[\lim_{n\to\infty} U_n^\beta] = 0 .$$

Thus, the proof of the lemma is completed.

LEMMA 2. *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of non-negative real numbers. Suppose that there exist a positive integer n_0, two sequences of non-negative real numbers $\{\gamma_n\}_{n=1}^{\infty}$ and $\{A_n\}_{n=1}^{\infty}$ such that*

(2.14) $$a_{n+1} \leqq (1-\gamma_{n+1})a_n + A_{n+1} \quad \textit{for all} \quad n \geqq n_0 ,$$

(2.15) $$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \textit{and} \quad \lim_{n\to\infty} \gamma_n = 0 ,$$

(2.16) $$\sum_{n=1}^{\infty} A_n < \infty .$$

Then, it holds that

$$\lim_{n\to\infty} a_n = 0 .$$

PROOF. The repeated application of the inequalities (2.14) for an integer $m \geqq n_0$ gives us

$$\begin{aligned} a_m &\leqq (1-\gamma_m)a_{m-1} + A_{m+1} \\ &\leqq \prod_{k=n_0+1}^{m} (1-\gamma_k)a_{n_0} + \sum_{l=n_0+1}^{m} \prod_{k=l+1}^{m} (1-\gamma_k)A_l . \end{aligned}$$

Hence, we have

(2.17) $$a_m \leqq F(n_0, m)a_{n_0} + G(n_0, m) ,$$

where

$$F(n_0, m) = \prod_{k=n_0+1}^{m} (1-\gamma_k) \quad \text{and} \quad G(n_0, m) = \sum_{l=n_0+1}^{m} \prod_{k=l+1}^{m} (1-\gamma_k)A_l .$$

Since

$$F(n_0, m) = \prod_{k=n_0+1}^{m} (1-\gamma_k) \leqq \exp\Big(-\sum_{k=n_0+1}^{m} \gamma_k\Big) ,$$

we have from (2.15)

(2.18) $$F(n_0, m) \to 0 \quad \text{as} \quad m \to \infty .$$

Further, noting that for an integer N $(m \geqq N \geqq n_0)$

$$G(n_0, m) = \sum_{l=n_0+1}^{N} \prod_{k=l+1}^{m} (1-\gamma_k)A_l + \sum_{l=N+1}^{m} \prod_{k=l+1}^{m} (1-\gamma_k)A_l ,$$

we have two positive constants C_1 and C_2 for which

(2.19) $$\begin{aligned} G(n_0, m) &\leqq C_1 \prod_{k=N}^{m} (1-\gamma_k) + C_2 \sum_{l=N+1}^{m} A_l \\ &\leqq C_1 \exp\Big(-\sum_{k=N+1}^{m} \gamma_k\Big) + C_2 \sum_{l=N+1}^{m} A_l . \end{aligned}$$

Hence, from (2.19), we have

(2.20) $$G(n_0, m) \to 0 \quad \text{as } m \to \infty \text{ and then } N \to \infty .$$

By (2.18) and (2.20), we have

$$\lim_{n\to\infty} a_n = 0 .$$

Thus, the proof of the lemma is completed.

Next, we mention without proof the lemma given by V. Dupač [9], a modification of the result of K. L. Chung.

LEMMA 3. *Let* $\{a_n\}_{n=1}^{\infty}$ *be a sequence of non-negative real numbers. Suppose that there exist a positive integer* n_0, *two positive constants* A *and* B *such that*

(2.21) $$a_{n+1} \leqq (1 - A/n^s) a_n + B/n^t \qquad \text{for all} \quad n \geqq n_0 ,$$

(2.22) $$t \text{ real number and } 0 < s < 1 .$$

Then, we have

$$\limsup_{n\to\infty} n^{t-s} a_n \leqq B/A .$$

§ 3. The Model and the Bayes decision rule in the Pattern Classification Problems.

In this section, we give a model under which the learning theories will be developed later. Generally, in pattern classification problems, each observed pattern x in pattern space is a random sample from a group to which they belong and each group is called a category. Therefore, corresponding to each category, there is a probability distribution law by which each observed pattern is drawn. We consider the case when there exist s categories $\theta_1, \theta_2, \cdots, \theta_s$ and we denote a set of s categories $\theta_1, \cdots, \theta_s$ by Θ.

Hence, an outcome in pattern classification problem is described by a pair (x, θ). The element x is an observed pattern in pattern space and θ specifies the category of an observed pattern. But generally θ is unknown to the observer. For a sequence of observed patterns $x^1, x^2, \cdots$, we can consider a sequence:

(3.1) $$(x^1, \theta^1), (x^2, \theta^2), \cdots, (x^n, \theta^n), \cdots$$

with $x^n \in X_n$ and $\theta^n \in \Theta$, where X_n is a pattern space at instant n and $\theta^n = \theta_i$ if x^n is a sample value from a specific category θ_i. From such sequence, the result of the first n history is expressed by two sets

$$\xi^n = (x^1, x^2, \cdots, x^n) \quad \text{and} \quad \alpha^n = (\theta^1, \theta^2, \cdots, \theta^n) .$$

Then a history (ξ^n, α^n) at each instant n is an element of the set (X^n, Θ^n), where

$$X^n = X_1 \times X_2 \times \cdots \times X_n \quad \text{and} \quad \Theta^n = \Theta \times \Theta \times \cdots \times \Theta .$$

In what follows, we shall assume that, for each n, the transition probability distribution of an outcome at instant $n+1$ given a history at instant n has the density function. We denote this transition probability density function by

(3.2) $$p(x^{n+1}, \theta^{n+1} \mid \xi^n, \alpha^n) .$$

Next, we can consider "a posteriori" probability density function according to

Bayes formula: after an observed pattern x^{n+1} at instant $n+1$ was known, we have a following "a posteriori" probability density function for $\alpha^n\theta^{n+1}=(\theta^1, \theta^2, \cdots, \theta^{n+1})$,

$$\Pi_{x^{n+1}}(\alpha^n\theta^{n+1})=\frac{\Pi(\alpha^n)p(x^{n+1}, \theta^{n+1}\,|\,\xi^n, \alpha^n)}{\sum_{\tilde{\alpha}^n\in\Theta^n}\sum_{\tilde{\theta}^{n+1}\in\Theta}\Pi(\tilde{\alpha}^n)p(x^{n+1}, \tilde{\theta}^{n+1}\,|\,\xi^n, \tilde{\alpha}^n)}, \tag{3.3}$$

where $\Pi(\alpha^n)$ is a probability distribution of $\alpha^n\in\Theta^n$. Then if all transition probability density functions at each instant are known to the observer, the classification of an observed pattern x^{n+1} at each instant $n+1$ will be determined by the largest of the quantities $\Pi_{x^{n+1}}(\alpha^n\theta_1), \Pi_{x^{n+1}}(\alpha^n\theta_2), \cdots, \Pi_{x^{n+1}}(\alpha^n\theta_s)$ and this decision rule will be called the Bayes decision rule. When $\Theta=\{\theta_1, \theta_2\}$ we can define for an observed pattern x^{n+1} at instant $n+1$ and a given history (ξ^n, α^n) at instant n

$$D^*(x^{n+1}\,|\,\xi^n, \alpha^n)=\Pi_{x^{n+1}}(\alpha^n\theta_1)-\Pi_{x^{n+1}}(\alpha^n\theta_2)\,. \tag{3.4}$$

By the Bayes decision rule we have an optimal decision rule:

x^{n+1} is classified in category θ_1 if $D^*(x^{n+1}\,|\,\xi^n, \alpha^n)\geqq 0$,

classified in category θ_2 if $D^*(x^{n+1}\,|\,\xi^n, \alpha^n)<0$.

This decision rule is equivalent to the following decision rule:

x^{n+1} is classified in category θ_1 if $D(x^{n+1}\,|\,\xi^n, \alpha^n)\geqq 0$,

classified in category θ_2 if $D(x^{n+1}\,|\,\xi^n, \alpha^n)<0$,

where

$$D(x^{n+1}\,|\,\xi^n, \alpha^n)=p(x^{n+1}, \theta_1\,|\,\xi^n, \alpha^n)-p(x^{n+1}, \theta_2\,|\,\xi^n, \alpha^n)\,. \tag{3.5}$$

§ 4. Pattern classification by learning with a teacher.

In this section, we shall investigate the pattern classification problem for the model having two categories. We treat the case where the amount of a priori information on the transition probability density functions in the model is small but an observer is indicated by a teacher the category from which an observed pattern is extracted. By a training sequence we shall imply a sequence $(x^1, \theta^1), \cdots, (x^n, \theta^n), \cdots$, where θ^i is a category indicated by a teacher at instant i. This will correspond to learning with reinforcement indicated by a teacher and may be called "learning with a teacher". The pattern classification problem considered here is to find a decision rule to classify a pattern x^i in a category for $i\geqq n+1$ on the basis of a training sequence up to n.

It is generally known that the Bayes decision rule minimizes the probability of misclassification. It is reasonable, therefore, to consider a method of approximation to the limit of $D(x^{n+1}\,|\,\xi^n, \alpha^n)$, if it exists, by using a training sequence. In what follows, it is assumed that the limit of $D(x^{n+1}\,|\,\xi^n, \alpha^n)$ exists.

Firstly, we give a system of orthonormal functions $\{\varphi_i^{(n+1)}(x^{n+1})\}_{i=1}^N$ defined on each pattern space X_{n+1} at each instant $n+1$ such that

$$\int_{X_{n+1}}\varphi_i^{(n+1)}(x^{n+1})\varphi_j^{(n+1)}(x^{n+1})\,dx^{n+1}=\begin{cases}1 & \text{if } i=j\\ 0 & \text{otherwise}.\end{cases}$$

Secondly, we approximate an unknown decision function $D(x^{n+1}|\xi^n, \alpha^n)$ at this instant by a finite series

(4.1) $$\hat{D}_*(x^{n+1}|\xi^n, \alpha^n)=\sum_{i=1}^{N} c_{i*}^{(n+1)}(\xi^n, \alpha^n)\varphi_i^{(n+1)}(x^{n+1})$$

which minimizes a quantity I_{n+1} defined by

(4.2) $$I_{n+1}=\int_{X_{n+1}}\{D(x^{n+1}|\xi^n, \alpha^n)-\hat{D}(x^{n+1}|\xi^n, \alpha^n)\}^2 dx^{n+1},$$

where

(4.3) $$\hat{D}(x^{n+1}|\xi^n, \alpha^n)=\sum_{i=1}^{N} c_i^{(n+1)}(\xi^n, \alpha^n)\varphi_i^{(n+1)}(x^{n+1})$$

and $\{c_i^{(n+1)}(\xi^n, \alpha^n)\}_{i=1}^N$ are unknown coefficients for a history at instant n.

Hence, we can reduce this problem to the problem of finding an algorithm by which we can construct, on the basis of a training sequence, a random variable $c_j^{(n)}(\xi^n, \alpha^n)$ $(j=1, 2, \cdots, N)$, converging to $c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1})$ as $n\to\infty$ in some sense.

Now by differentiating I_{n+1} with respect to each $c_j^{(n+1)}$, equating the derivatives to zero and using the orthonormal property of $\{\varphi_i^{(n+1)}(x^{n+1})\}_{i=1}^N$, we have for each $j=1, 2, \cdots, N$ at each instant $n+1$

(4.4) $$c_{j*}^{(n+1)}(\xi^n, \alpha^n)=E_{\theta_1}[\varphi_j^{(n+1)}(x^{n+1})|\xi^n, \alpha^n]-E_{\theta_2}[\varphi_j^{(n+1)}(x^{n+1})|\xi^n, \alpha^n],$$

where

$$E_{\theta_i}[\varphi_j^{(n+1)}(x^{n+1})|\xi^n, \alpha^n]=\int_{X_{n+1}}\varphi_j^{(n+1)}(x^{n+1})p(x^{n+1}, \theta_i|\xi^n, \alpha^n)dx^{n+1}.$$

Furthermore, we have for each instant $n+1$ and $j=1, 2, \cdots, N$

(4.5) $$c_{j*}^{(n+1)}(\xi^n, \alpha^n)=E[d^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1})-(1-d^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(x^{n+1})|\xi^n, \alpha^n],$$

where $d^{(n+1)}(\theta^{n+1})$ is a random variable on Θ at instant $n+1$ defined by

$$d^{(n+1)}(\theta^{n+1})=\begin{cases}1 & \text{if } \theta^{n+1}=\theta_1\\ 0 & \text{otherwise}.\end{cases}$$

In view of the above argument, we shall construct the following algorithm for a sequence of non-negative real numbers $\{\gamma_n\}_{n=1}^\infty$ such that

(4.6) $$\sum_{n=1}^{\infty}\gamma_n=\infty \quad\text{and}\quad \sum_{n=1}^{\infty}\gamma_n^2<\infty.$$

At first, using an outcome (x^1, θ^1) of an observed pattern x^1 at instant 1 and a category θ^1, to which x^1 belongs, indicated by a teacher, we make, for $j=1, 2, \cdots, N$, $c_j^{(1)}(\xi^1, \alpha^1)$:

(4.7) $$c_j^{(1)}(\xi^1, \alpha^1)=c_j^{(0)}+\gamma_1\{\rho^{(1)}(\theta^1)\varphi_j^{(1)}(x^1)-(1-\rho^{(1)}(\theta^1))\varphi_j^{(1)}(x^1)-c_j^{(0)}\},$$

where

$$c_j^{(0)}=0 \quad \text{for } j=1, 2, \cdots, N$$

and

$$\rho^{(1)}(\theta^1)=\begin{cases}1 & \text{if } \theta^1=\theta_1\\ 0 & \text{otherwise}.\end{cases}$$

Secondly, using an outcome (x^2, θ^2) of an observed pattern x^2 at instant 2 and a category θ^2, to which x^2 belongs, indicated by a teacher, we make, for $j=1, 2, \cdots, N$, $c_j^{(2)}(\xi^2, \alpha^2)$:

$$(4.8) \qquad c_j^{(2)}(\xi^2, \alpha^2)=c_j^{(1)}(\xi^1, \alpha^1)+\gamma_2\{\rho^{(2)}(\theta^2)\varphi_j^{(2)}(x^2)-(1-\rho^{(2)}(\theta^2))\varphi_j^{(2)}(x^2)-c_j^{(1)}(\xi^1, \alpha^1)\},$$

where

$$\rho^2(\theta^2)=\begin{cases}1 & \text{if } \theta^2=\theta_1 \\ 0 & \text{otherwise}.\end{cases}$$

In general, using an outcome (x^{n+1}, θ^{n+1}) of an observed pattern x^{n+1} at instant $n+1$ and a category θ^{n+1}, to which x^{n+1} belongs, indicated by a teacher, we make, for $j=1, 2, \cdots, N$, $c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1})$:

$$(4.9) \qquad c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1})=c_j^{(n)}(\xi^n, \alpha^n)+\gamma_{n+1}\{\rho^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1}) -(1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(x^{n+1})-c_j^{(n)}(\xi^n, \alpha^n)\},$$

where

$$\rho^{(n+1)}(\theta^{n+1})=\begin{cases}1 & \text{if } \theta^{n+1}=\theta_1 \\ 0 & \text{otherwise}.\end{cases}$$

Concerning these two random variables $c_j^{(n)}(\xi^n, \alpha^n)$ and $c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1})$ we have the following theorems.

THEOREM 4.1. *Let the following condition be satisfied: there are a sequence of non-negative real numbers* $\{M_n\}_{n=1}^{\infty}$ *and three positive constants* K_1, K_2, K_3 *such that*

$$(4.10) \qquad \sum_{n=1}^{\infty} M_n<\infty,$$

$$(4.11) \qquad P[(\theta_j^{(n)})^2 \leqq \gamma_{n+1}M_n]=1 \quad \textit{for each} \quad j=1, 2, \cdots, N \quad \textit{and all } n,$$

where $\theta_j^{(n)}=c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1})-c_{j*}^{(n+1)}(\xi^n, \alpha^n)$ *and a sequence* $\{\gamma_n\}_{n=1}^{\infty}$ *satisfy* (4.6),

$$(4.12) \qquad \mathrm{Var}\,(Y_j^{(n+1)}\,|\,\xi^n, \alpha^n) \leqq K_1(u_j^{(n)})^2+K_2(\theta_j^{(n)})^2+K_3 \quad \textit{for all } j \textit{ and } n,$$

where

$$Y_j^{(n+1)}=\rho^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1})-(1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(\theta^{n+1}),$$

$$u_j^{(n)}=c_j^{(n)}(\xi^n, \alpha^n)-c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}),$$

and $\mathrm{Var}\,[Y_j^{(n+1)}\,|\,\xi^n, \alpha^n]$ *is a conditional variance of* $Y_j^{(n+1)}$ *given a history* (ξ^n, α^n) *at instant* n. *Then, for* $j=1, 2, 3, \cdots, N$, *we have*

$$P[\lim_{n\to\infty} u_j^{(n)}=0]=1$$

$$\lim_{n\to\infty} E[(u_j^{(n)})^{2\beta}]=0 \qquad \textit{for all} \quad 0<\beta<1$$

and

$$\lim_{n\to\infty} E[(u_j^{(n)})^2]=0.$$

PROOF. By the construction of $c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1})$ for $j=1, 2, \cdots, N$, we have

$$
\begin{aligned}
& c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\
& = c_j^{(n)}(\xi^n, \alpha^n) + \gamma_{n+1}[\rho^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1}) - (1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(x^{n+1}) \\
& \quad - c_j^{(n)}(\xi^n, \alpha^n)] - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\
& = c_j^{(n)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) + \gamma_{n+1}[\rho^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1}) \\
& \quad -(1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(x^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\
& \quad + c_{j*}^{(n+1)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) + c_{j*}^{(n)}(\xi^{-1}, \alpha^{n-1}) - c_j^{(n)}(\xi^n, \alpha^n)] \\
& \quad + c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\
& = (1-\gamma_{n+1})[c_j^{(n)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1})] \\
& \quad + (1-\gamma_{n+1})[c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)] \\
& \quad + \gamma_{n+1}[\rho^{(n+1)}(\theta^{n+1})\varphi_j^{(n+1)}(x^{n+1}) - (1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j^{(n+1)}(x^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)] .
\end{aligned} \tag{4.13}
$$

The equality (4.13) can be written in terms of $u_j^{(n+1)}$, $u_j^{(n)}$, $\theta_j^{(n)}$ and $Y_j^{(n+1)}$ as

$$
u_j^{(n+1)} = (1-\gamma_{n+1})u_j^{(n)} + (1-\gamma_{n+1})\theta_j^{(n)} + \gamma_{n+1}[Y_j^{(n+1)} - c_{j*}^{(n+1)}(\xi^n, \alpha^n)] . \tag{4.14}
$$

Squareing both sides of (4.14), we can obtain

$$
\begin{aligned}
(u_j^{(n+1)})^2 \leqq & (1-\gamma_{n+1})^2(u_j^{(n)})^2 + (1-\gamma_{n+1})^2(\theta_j^{(n)})^2 \\
& + \gamma_{n+1}^2[Y_j^{(n+1)} - c_{j*}^{(n+1)}(\xi^n, \alpha^n)]^2 + 2(1-\gamma_{n+1})^2 |u_j^{(n)}| \, |\theta_j^{(n)}| \\
& + 2(1-\gamma_{n+1})\gamma_{n+1}(Y_j^{(n+1)} - c_{j*}^{(n+1)}(\xi^n, \alpha^n))u_j^{(n)} \\
& + 2(1-\gamma_{n+1})\gamma_{n+1}(Y_j^{(n+1)} - c_{j*}^{(n+1)}(\xi^n, \alpha^n))\theta_j^{(n)} .
\end{aligned} \tag{4.15}
$$

Now, taking conditional expectation on both sides of (4.15) and using (4.12), we have for $j=1, 2, \cdots, N$,

$$
\begin{aligned}
E[(u_j^{(n+1)})^2 | \xi^n, \alpha^n] \leqq & (1-\gamma_{n+1})^2(u_j^{(n)})^2 + (1-\gamma_{n+1})^2(\theta_j^{(n)})^2 \\
& + 2(1-\gamma_{n+1})^2 |u_j^{(n)}| \, |\theta_j^{(n)}| + \gamma_{n+1}^2[K_1(u_j^{(n)})^2 + K_2(\theta_j^{(n)})^2 + K_3] \\
\leqq & [1-2\gamma_{n+1} + \gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2 + [1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2 \\
& + 2|u_j^{(n)}| \, |\theta_j^{(n)}| .
\end{aligned} \tag{4.16}
$$

Noting that for $j=1, 2, \cdots, N$ and all n

$$
2|u_j^{(n)}| \, |\theta_j^{(n)}| \leqq \gamma_{n+1}(u_j^{(n)})^2 + (\theta_j^{(n)})^2/\gamma_{n+1} ,
$$

from (4.16) we have for $j=1, 2, \cdots, N$

$$
\begin{aligned}
E[(u_j^{(n+1)})^2 | \xi^n, \alpha^n] \leqq & [1-2\gamma_{n+1} + \gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2 + [1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2 \\
& + \gamma_{n+1}(u_j^{(n)})^2 + (\theta_j^{(n)})^2/\gamma_{n+1} + \gamma_{n+1}^2 K_3 \\
= & [1+\gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2 - \gamma_{n+1}(u_j^{(n)})^2 + [1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2 \\
& + (\theta_j^{(n)})^2/\gamma_{n+1} + \gamma_{n+1}^2 K_3 .
\end{aligned} \tag{4.17}
$$

$U_n = (u_j^{(n)})^2$, $V_n = (u_j^{(n)})^2$, $\mu_n = \gamma_{n+1}^2(1+K_1)$ and $\xi_n = [1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2 + (\theta_j^{(n)})^2/\gamma_{n+1} + \gamma_{n+1}^2 K_3$ satisfy (A1)~(A5), (2.1), (2.2) and (2.3) by (4.6), (4.10), (4.11) and (4.17). Therefore, by Lemma 1, it follows that for $j=1, 2, \cdots, N$

$$P[\lim_{n\to\infty} u_j^{(n)}=0]=1$$

and

$$\lim_{n\to\infty} E[(u_j^{(n)})^{2\beta}]=0 \qquad \text{for all} \quad 0<\beta<1.$$

Taking the unconditional expectation on both sides of (4.17) we have for $j=1, 2, \cdots, N$

$$(4.18) \qquad \begin{aligned} E[(u_j^{(n+1)})^2] &\leqq [1-\gamma_{n+1}+\gamma_{n+1}^2(1+K_1)]E[(u_j^{(n)})^2] \\ &\quad +[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2 K_3 \\ &=[1-\gamma_{n+1}\{1-\gamma_{n+1}(1+K_1)\}]E[(u_j^{(n)})^2] \\ &\quad +[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2 K_3 . \end{aligned}$$

$a_n=E[(u_j^{(n)})^2]$ and $A_n=[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2 K_3$ satisfy (2.14), (2.15) and (2.16) by (4.6), (4.10), (4.11) and (4.18). Therefore, by Lemma 2, it follows that for $j=1, 2, \cdots, N$

$$\lim_{n\to\infty} E[(u_j^{(n)})^2]=0.$$

Thus the proof of the theorem is completed.

Next we have the following theorem concerning the order of mean convergence.

THEOREM 4.2. *Suppose that* (4.12) *and the following conditions are satisfied:*

$$(4.19) \qquad \gamma_n=a/n^\alpha, \qquad a>0, \qquad \frac{1}{2}<\alpha<1,$$

$$(4.20) \qquad E[(\theta_j^{(n)})^2]=O(n^{-2\omega}), \quad \omega>\alpha \qquad \text{for} \quad j=1, 2, \cdots, N.$$

Then we have

$$E[(u_j^{(n)})^2]=\begin{cases} O(n^{-2(\omega-\alpha)}) & \text{if} \quad \omega<\frac{3}{2}\alpha \\ O(n^{-\alpha}) & \text{if} \quad \omega\geqq\frac{3}{2}\alpha, \end{cases}$$

where the notation $f(n)=O(g(n))$ *means* $\limsup_{n\to\infty} |f(n)/g(n)|<\infty$.

PROOF. By (4.18), (4.19) and (4.20), there exist a positive integer N and three positive constants C_1, C_2, C_3 such that, for all $n\geqq N$,

$$(4.21) \qquad E[(u_j^{(n+1)})^2]\leqq(1-C_1/n^\alpha)E[(u_j^{(n)})^2]+C_2/n^{2\alpha}+C_3/n^{2\omega-\alpha}.$$

Consequently, we have for $\omega<\frac{3}{2}\alpha$

$$(4.22) \qquad E[(u_j^{(n+1)})^2]\leqq(1-C_1/n^\alpha)E[(u_j^{(n)})^2]+C_4/n^{2\omega-\alpha}$$

and for $\omega\geqq\frac{3}{2}\alpha$

$$(4.23) \qquad E[(u_j^{(n+1)})^2]\leqq(1-C_1/n^\alpha)E[(u_j^{(n)})^2]+C_5/n^{2\alpha},$$

where C_4 and C_5 are some constants. Thus an application of Lemma 3 for $a_n=E[(u_j^{(n)})^2]$ gives us the result of the theorem.

EXAMPLE 4.1. We consider an algorithm for a sequence $\{1/n\}_{n=1}^\infty$ and a model with two categories which satisfies the following condition:

(4.23) each pattern space X_n is equivalent to a space X,

(4.24) $$p(x^{n+1}, \theta_i \mid \xi^n, \alpha^n) = q^{(n+1)}(\theta_i \mid \theta^n) f_{\theta_i}(x^{n+1}) \quad \text{for all } n \text{ and } i,$$

(4.25) $$\sum_{\theta_i \in \Theta} q^{(n+1)}(\theta_i \mid \theta^n) = 1 \quad \text{for all } n$$

that is, $q^{(n+1)}(\theta_i \mid \theta^n)$ is a conditional probability of a category θ_i at instant $n+1$ for a given category θ^n at instant n,

(4.26) for a real number q_1 $(0 \leqq q_1 \leqq 1)$, there are two positive constants C, α such that

$$P[\,|q^{(n+1)}(\theta_1 \mid \theta^n) - q_1| \leqq C/(n+1)^{1+\alpha}] = 1 \quad \text{for all } n,$$

(4.27) $$\int_X f_{\theta_i}(x)dx = 1 \quad \text{for all } i$$

that is, $f_{\theta_i}(x)$ is a conditional probability density function which we can observe a pattern x from a given category θ_i.

Suppose that, at each instant, a system of orthonormal functions $\{\varphi_i(x)\}_{i=1}^N$ defined on the pattern space X is given and, for all i and j, we have

(4.28) $$E_{\theta_i}[\varphi_j^2(x)] = \int_X \varphi_j^2(x) f_{\theta_i}(x)dx < \infty .$$

In this case, we have, from (4.4), for all n

(4.29) $$c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) = c_{j*}^{(n)}(\theta^{n-1})$$
$$= q^{(n)}(\theta_1 \mid \theta^{n-1}) E_{\theta_1}[\varphi_j(x)] - q^{(n)}(\theta_2 \mid \theta^{n-1}) E_{\theta_2}[\varphi_j(x)]$$

and

(4.30) $$\begin{aligned} \theta_j^{(n)} &= c_{j*}^{(n)}(\theta^{n-1}) - c_{j*}^{(n+1)}(\theta^n) \\ &= (q^{(n)}(\theta_1 \mid \theta^{n-1}) - q^{(n+1)}(\theta_1 \mid \theta^n)) E_{\theta_1}[\varphi_j(x)] \\ &\quad -(q^{(n)}(\theta_2 \mid \theta^{n-1}) - q^{(n+1)}(\theta_2 \mid \theta^n)) E_{\theta_2}[\varphi_j(x)] \\ &= (q^{(n)}(\theta_1 \mid \theta^{n-1}) - q^{(n+1)}(\theta_1 \mid \theta^{n-1}))(E_{\theta_1}[\varphi_j(x)] - E_{\theta_2}[\varphi_j(x)]) . \end{aligned}$$

By using (4.26), (4.28) and (4.30), we can easily prove that there exist a sequence of non-negative real numbers $\{M_n\}_{n=1}^{\infty}$, a positive constant K such that

(4.31) $$P[(n+1)(\theta_j^{(n)})^2 \leqq M_n] = 1 \quad \text{for all } n \text{ and } j$$

(4.32) $$\sum_{n=1}^{\infty} M_n < \infty$$

and

(4.33) $$\mathrm{Var}\,[Y_j^{(n+1)} \mid \theta^n] \leqq K \quad \text{for all } n \text{ and } j$$

where

$$Y_j^{(n+1)} = \rho^{(n+1)}(\theta^{n+1})\varphi_j(x^{n+1}) - (1-\rho^{(n+1)}(\theta^{n+1}))\varphi_j(x^{n+1}).$$

Thus, the results of Theorem 4.1 hold in this case

EXAMPLE 4.2 We consider an algorithm for a sequence $\{1/n^{\alpha}\}_{n=1}^{\infty}$, $\left(\frac{1}{2} < \alpha < 1\right)$

and a model with two categories which satisfies (4.23), (4.24), (4.25), (4.27) and the following condition: there exists a real number q_1 ($0 \leqq q_1 \leqq 1$) such that

$$E[(q^{(n+1)}(\theta_1 | \theta^n) - q_1)^2] = O[(n+1)^{-2\omega}], \qquad \omega > \alpha \tag{4.34}$$

is true for all n.

Suppose that the same system as a given system of orthonormal functions $\{\varphi_i(x)\}_{i=1}^N$ in Example 4.1 is given. By using (4.29) and (4.30), we can easily obtain

$$E[(\theta_j^{(n)})^2] = O(n^{-2\omega}), \tag{4.35}$$

and

$$\mathrm{Var}[Y_j^{(n+1)} | \theta^n] \leqq K \qquad \text{for all } n \text{ and } j, \tag{4.36}$$

where

$$Y_j^{(n+1)} = \rho^{(n+1)}(\theta^{n+1})\varphi_j(x^{n+1}) - (1 - \rho^{(n+1)}(\theta^{n+1}))\varphi_j(x^{n+1})$$

and K is a constant. Thus the result of Theorem 4.2 holds.

§ 5. Pattern classification by self-learning.

In this section, we shall investigate the pattern classification problem for the model having s categories. We treat the case where, in addition to the condition that the amount of a priori information on the transition probability functions and the category set is small, there cannot be assumed a training sequence. In this case, we have to consider the pattern classification problem only on the basis of observed but unclassified patterns. This will correspond to learning without reinforcement indicated by a teacher and may be called "learning without a teacher" or self-learning.

At the first glance, self-learning seems to be impossible, as we are merely observing a sequence of random variables, which may not be even independent. It is reasonable, however, to assign to each mode of the limit of $p(x^{n+1} | \xi^n, \alpha^n)$, if it exists, a category, and to treat the problem in the manner like a problem of pattern recognition, where

$$p(x^{n+1} | \xi^n, \alpha^n) = \sum_{i=1}^{s} p(x^{n+1}, \theta_i | \xi^n, \alpha^n). \tag{5.1}$$

By the above consideration, it is reasonable to treat the method of approximation to $p(x^{n+1} | \xi^n, \alpha^n)$ by using observed and unclassified patterns. In what follows, let us assume that the limit of $p(x^{n+1} | \xi^n, \alpha^n)$ exists.

Firstly, we give a system of orthonormal functions $\{\varphi_i^{(n+1)}(x^{n+1})\}_{i=1}^N$ defined on each pattern space X_{n+1} such that

$$\int_{X_{n+1}} \varphi_i^{(n+1)}(x^{n+1}) \varphi_j^{(n+1)}(x^{n+1}) dx^{n+1} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise}. \end{cases}$$

Secondly, we approximate an unknown $p(x^{n+1} | \xi^n, \alpha^n)$ at this instant by a finite series

$$\hat{p}_*(x^{n+1} | \xi^n, \alpha^n) = \sum_{i=1}^{N} c_{i*}^{(n+1)}(\xi^n, \alpha^n) \varphi_i^{(n+1)}(x^{n+1}) \tag{5.2}$$

which minimizes a quantity J_{n+1} defined by

$$J_{n+1}=\int_{X_{n+1}}[p(x^{n+1}|\xi^n,\alpha^n)-\hat{p}(x^{n+1}|\xi^n,\alpha^n)]^2dx^{n+1}, \tag{5.3}$$

where

$$\hat{p}(x^{n+1}|\xi^n,\alpha^n)=\sum_{i=1}^{N}c_i^{(n+1)}(\xi^n,\alpha^n)\varphi_i^{(n+1)}(x^{n+1}) \tag{5.4}$$

and $\{c_i^{(n+1)}(\xi^n,\alpha^n)\}_{i=1}^N$ are unknown coefficients for a given history (ξ^n,α^n) at instant n. Hence, we can reduce this problem to the problem of finding an algorithm by which we can construct, from an observed sequence, a random variable $c_j^{(n)}(\xi^n,\alpha^n)$ $(j=1,2,\cdots,N)$, converging to $c_{j*}^{(n)}(\xi^{n-1},\alpha^{n-1})$ as $n\to\infty$ in some sense.

Now by differentiating I_{n+1} with respect to each $c_j^{(n+1)}$, equating the derivatives to zero and using the orthonormal property of $\{\varphi_i^{(n+1)}(x^{n+1})\}_{i=1}^N$, we have for each $j=1,2,\cdots,N$ at each instant $n+1$

$$c_{j*}^{(n+1)}(\xi^n,\alpha^n)=E[\varphi_j^{(n+1)}(x^{n+1})|\xi^n,\alpha^n], \tag{5.5}$$

where

$$E[\varphi_j^{(n+1)}(x^{n+1})|\xi^n,\alpha^n]=\sum_{i=1}^{s}\int_{X_{n+1}}\varphi_j^{(n+1)}(x^{n+1})p(x^{n+1},\theta_i|\xi^n,\alpha^n)dx^{n+1}.$$

In view of the above argument, we shall construct a following algorithm with a sequence of non-negative real numbers $\{\gamma_n\}_{n=1}^\infty$ such that

$$\sum_{n=1}^{\infty}\gamma_n=\infty \quad \text{and} \quad \sum_{n=1}^{\infty}\gamma_n^2<\infty. \tag{5.6}$$

At first, using an observed but unclassified pattern x^1 at instant 1, we make, for $j=1,2,\cdots,N$, $c_j^{(1)}(\xi^1,\alpha^1)$:

$$c_j^{(1)}(\xi^1,\alpha^1)=c_j^{(0)}+\gamma_1(\varphi_j^{(1)}(x^1)-c_j^{(0)}), \tag{5.7}$$

where

$$c_j^{(0)}=0 \qquad \text{for all } j.$$

Secondly, using an observed but unclassified pattern x^2 at instant 2, we make, for $j=1,2,\cdots,N$, $c_j^{(2)}(\xi^2,\alpha^2)$:

$$c_j^{(2)}(\xi^2,\alpha^2)=c_j^{(1)}(\xi^1,\alpha^1)+\gamma_2(\varphi_j^{(2)}(x^2)-c_j^{(1)}(\xi^1,\alpha^1)). \tag{5.8}$$

In general, using an observed but unclassified pattern x^n at instant n, we make, for $j=1,2,\cdots,N$, $c_j^{(n)}(\xi^n,\alpha^n)$:

$$c_j^{(n)}(\xi^n,\alpha^n)=c_j^{(n-1)}(\xi^{n-1},\alpha^{n-1})+\gamma_n(\varphi_j^{(n)}(x^n)-c_j^{(n-1)}(\xi^{n-1},\alpha^{n-1})). \tag{5.9}$$

Concerning these two random variables $c_j^{(n)}(\xi^n,\alpha^n)$ and $c_{j*}^{(n)}(\xi^{n-1},\alpha^{n-1})$ we have the following theorems.

THEOREM 5.1. *Let the following condition be satisfied: there are a sequence of non-negative real numbers $\{M_n\}_{n=1}^\infty$ and three positive constants K_1, K_2, K_3 such that*

$$\sum_{n=1}^{\infty}M_n<\infty, \tag{5.10}$$

$$P[(\theta_j^{(n)})^2\leqq\gamma_{n+1}M_n]=1 \qquad \textit{for each } j=1,2,\cdots,N \textit{ and all } n, \tag{5.11}$$

where

$$\theta_j^{(n)} = c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)$$

and a sequence $\{\gamma_n\}_{n=1}^{\infty}$ *is equivalent to* (5.6), *for each* $j=1, 2, \cdots, N$ *and all* n

(5.12) $$\mathrm{Var}[\varphi_j^{(n+1)}(x^{n+1}) \mid \xi^n, \alpha^n] \leqq K_1(u_j^{(n)})^2 + K_2(\theta_j^{(n)})^2 + K_3,$$

where

$$u_j^{(n)} = c_j^{(n)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) \quad \textit{and} \quad \mathrm{Var}[\varphi_j^{(n+1)}(x^{n+1}) \mid \xi^n, \alpha^n]$$

is a conditional variance of $\varphi_j^{(n+1)}(x^{n+1})$ *given a history* (ξ^n, α^n) *at instant* n. *Then, for* $j=1, 2, 3, \cdots, N$, *we have*

$$P[\lim_{n\to\infty} u_j^{(n)} = 0] = 1$$

$$\lim_{n\to\infty} E[(u_j^{(n)})^{2\beta}] = 0 \qquad \textit{for all } 0<\beta<1,$$

and

$$\lim_{n\to\infty} E[(u_j^{(n)})^2] = 0.$$

PROOF. By the construction of $c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1})$, for $j=1, 2, \cdots, N$, we have

(5.13)
$$\begin{aligned} & c_j^{(n+1)}(\xi^{n+1}, \alpha^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\ &= c_j^{(n)}(\xi^n, \alpha^n) + \gamma_{n+1}(\varphi_j^{(n+1)}(x^{n+1}) - c_j^{(n)}(\xi^n, \alpha^n)) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\ &= c_j^{(n)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) + \gamma_{n+1}(\varphi_j^{(n+1)}(x^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\ &\quad + c_{j*}^{(n+1)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) + c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) \\ &\quad - c_j^{(n)}(\xi^n, \alpha^n)) + c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n) \\ &= (1-\gamma_{n+1})(c_j^{(n)}(\xi^n, \alpha^n) - c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1})) \\ &\quad + (1-\gamma_{n+1})(c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)) \\ &\quad + \gamma_{n+1}(\varphi_j^{(n+1)}(x^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)). \end{aligned}$$

The equality (5.13) can be written in terms of $u_j^{(n+1)}$, $u_j^{(n)}$ and $\theta_j^{(n)}$ as

(5.14) $$u_j^{(n+1)} = (1-\gamma_{n+1})u_j^{(n)} + (1-\gamma_{n+1})\theta_j^{(n)} + (\varphi_j^{(n+1)}(x^{n+1}) - c_{j*}^{(n+1)}(\xi^n, \alpha^n)).$$

Squareing both sides of (5.14) and then taking conditional expectation with respect to a history (ξ^n, α^n) at instant n, we can obtain from (5.12), for $j=1, 2, \cdots, N$,

(5.15)
$$\begin{aligned} E[(u_j^{(n+1)})^2 \mid \xi^n, \alpha^n] &\leqq (1-\gamma_{n+1})^2(u_j^{(n)})^2 + (1-\gamma_{n+1})^2(\theta_j^{(n)})^2 \\ &\quad + 2(1-\gamma_{n+1})^2 |u_j^{(n)}|\,|\theta_j^{(n)}| + \gamma_{n+1}^2[K_1(u_j^{(n)})^2 + K_2(\theta_j^{(n)})^2 + K_3] \\ &\leqq [1-2\gamma_{n+1} + \gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2 + [1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2 \\ &\quad + 2|u_j^{(n)}|\,|\theta_j^{(n)}| + \gamma_{n+1}^2 K_3. \end{aligned}$$

Noting that, for $j=1, 2, \cdots, N$ and all n,

$$2|u_j^{(n)}|\,|\theta_j^{(n)}| \leqq \gamma_{n+1}(u_j^{(n)})^2 + (\theta_j^{(n)})^2/\gamma_{n+1},$$

from (5.15) we have for $j=1, 2, \cdots, N$

$$(5.16)\qquad E[(u_j^{(n+1)})^2 \mid \xi^n, \alpha^n] \leqq [1-2\gamma_{n+1}+\gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2 \\ +[1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2+\gamma_{n+1}(u_j^{(n)})^2 \\ +(\theta_j^{(n)})^2/\gamma_{n+1}+\gamma_{n+1}^2K_3 \\ =[1+\gamma_{n+1}^2(1+K_1)](u_j^{(n)})^2-\gamma_{n+1}(u_j^{(n)})^2 \\ +[1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2+(\theta_j^{(n)})^2/\gamma_{n+1}+\gamma_{n+1}^2K_3 .$$

$U_n=(u_j^{(n)})^2$, $V_n=(u_j^{(n)})^2$, $\mu_n=\gamma_{n+1}^2(1+K_1)$ and $\zeta_n=[1+\gamma_{n+1}^2(1+K_2)](\theta_j^{(n)})^2+(\theta_j^{(n)})^2/\gamma_{n+1}+\gamma_{n+1}^2K_3$ satisfy (A1)$\sim$(A5), (2.1), (2.2) and (2.3) by (5.6), (5,10), (5.11) and (5.16), Therefore, by Lemma 1, it follows that for $j=1, 2, \cdots, N$

$$P[\lim_{n\to\infty} u_j^{(n)}=0]=1$$

and

$$\lim_{n\to\infty} E[(u_j^{(n)})^{2\beta}]=0 \qquad \text{for all } 0<\beta<1 .$$

Taking the unconditional expectation on both side of (5.16)

$$(5.17)\qquad E[(u_j^{(n+1)})^2] \leqq [1-\gamma_{n+1}+\gamma_{n+1}^2(1+K_1)]E[(u_j^{(n)})^2] \\ +[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2K_3 \\ =[1-\gamma_{n+1}\{1-\gamma_{n+1}(1+K_1)\}]E[(u_j^{(n)})^2] \\ +[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2K_3 .$$

$a_n=E[(u_j^{(n)})^2]$ and $A_n=[1+\gamma_{n+1}^2(1+K_2)]E[(\theta_j^{(n)})^2]+E[(\theta_j^{(n)})^2]/\gamma_{n+1}+\gamma_{n+1}^2K_3$ satisfy (2.14), (2.15) and (2.16) by (5.6), (5.10), (5.11) and (5.17). Therefore, by Lemma 2, it follows that for $j=1, 2, \cdots, N$

$$\lim_{n\to\infty} E[(u_j^{(n)})^2]=0 .$$

Thus the proof of the theorem is completed.

Next we have the following theorem concerning the order of mean convergence.

THEOREM 5.2. *Suppose that* (5.12) *and the following conditions are satisfied:*

$$(5.18)\qquad \gamma_n=a/n^\alpha , \qquad a>0 , \qquad \frac{1}{2}<\alpha<1$$

$$(5.19)\qquad E[(\theta_j^{(n)})^2]=O(n^{-2\omega}) , \qquad \omega>\alpha , \qquad \textit{for } j=1, 2, \cdots, N .$$

Then, we have

$$E[(u_j^{(n)})^2]=\begin{cases} O(n^{-2(\omega-\alpha)}) & \textit{if } \omega<\frac{3}{2}\alpha \\ O(n^{-\alpha}) & \textit{if } \omega\geqq\frac{3}{2}\alpha . \end{cases}$$

PROOF. By (5.17), (5.18) and (5.19) there exist a positive integer N and three positive constant C_1, C_2, C_3 such that for all $n\geqq N$

$$(5.20)\qquad E[(u_j^{(n+1)})^2] \leqq (1-C_1/n^\alpha)E[(u_j^{(n)})^2]+C_2/n^{2\alpha}+C_3/n^{2\omega-\alpha} .$$

Consequently we have for $\omega<\frac{3}{2}\alpha$

(5.21) $$E[(u_j^{(n+1)})^2] \leqq (1-C_1/n^{\alpha})E[(u_j^{(n)})^2]+C_4/n^{2\omega-\alpha}$$

and for $\omega \geqq \frac{3}{2}\alpha$

(5.22) $$E[(u_j^{(n+1)})^2] \leqq (1-C_1/n^{\alpha})E[(u_j^{(n)})^2]+C_5/n^{2\omega-\alpha},$$

where C_4 and C_5 are some constants. Thus an application of Lemma 3 for $a_n=E[(u_j^{(n)})^2]$ gives us the results of the theorem.

Now we shall show two examples extremely similar to those in the previous section.

EXAMPLE 5.1. We consider an algorithm for a sequence $\{1/n\}_{n=1}^{\infty}$ and a model with s categories which satisfies the following conditions:

(5.23) each pattern space X_n is equivalent to a space X,

(5.24) $$p(x^{n+1}, \theta_i \mid \xi^n, \alpha^n)=q^{(n+1)}(\theta_i \mid \theta^n)f_{\theta_i}(x^{n+1}) \qquad \text{for all } n \text{ and } i,$$

(5.25) $$\sum_{\theta_i \in \Theta} q^{(n+1)}(\theta_i \mid \theta^n)=1 \qquad \text{for all } n,$$

(5.26) for a set of real numbers $\{q_i\}_{i=1}^{s}$, there are two positive real numbers C, α such that

$$p[\,|q^{(n+1)}(\theta_i \mid \theta^n)-q_i| \leqq C/(n+1)^{1+\alpha}]=1 \qquad \text{for } i=1, 2, \cdots, s-1 \text{ and all } n,$$

where

$$0 \leqq q_i \leqq 1 \qquad \text{for each } i=1, 2, \cdots, s$$

and

$$\sum_{i=1} q_i=1,$$

(5.27) $$\int_X f_{\theta_i}(x)dx=1 \qquad \text{for all } i.$$

Suppose that, at each instant, a system of orthonormal functions $\{\varphi_i(x)\}_{i=1}^{N}$ defined on the pattern space X is given and, for all i and j, we have

(5.28) $$E_{\theta_i}[\varphi_j^2(x)]=\int_X \varphi_j^2(x)f_{\theta_i}(x)dx<\infty.$$

In this case, we have, from (5.5), for all n and j,

(5.29)
$$\begin{aligned} c_{j*}^{(n)}(\xi^{n-1}, \alpha^{n-1}) &= c_{j*}^{(n)}(\theta^{n-1}) \\ &= \sum_{i=1}^{s} E_{\theta_i}[\varphi_j(x^n) \mid \theta^{n-1}] \\ &= \sum_{i=1}^{s} q^{(n)}(\theta_i \mid \theta^{n-1})E_{\theta_i}[\varphi_j(x^n)] \end{aligned}$$

and

(5.30)
$$\begin{aligned} \theta_j^{(n)} = c_{j*}^{(n)}(\theta^{n-1})-c_{j*}^{(n+1)}(\theta^n) &= \sum_{i=1}^{s} (q^{(n)}(\theta_i \mid \theta^{n-1})-q^{(n+1)}(\theta_i \mid \theta^n))E_{\theta_i}(\varphi_j(x)) \\ &= \sum_{i=1}^{s-1} (q^{(n)}(\theta_i \mid \theta^{n-1})-q^{(n+1)}(\theta_i \mid \theta^n))[E_{\theta_i}(\varphi_j(x)) \\ &\qquad -E_{\theta_s}(\varphi_j(x))]. \end{aligned}$$

By using (5.26), (5.28) and (5.30), we can easily show that there exist a sequence of non-negative real numbers $\{M_n\}_{n=1}^{\infty}$, a positive constant K such that

(5.31) $$P[(n+1)(\theta_j^{(n)})^2 \leqq M_n] = 1 \quad \text{for all } n \text{ and } j$$

(5.32) $$\sum_{n=1}^{\infty} M_n < \infty$$

and

(5.33) $$\text{Var}\,[\varphi_j(x^{n+1}) \mid \theta^n] \leqq K \quad \text{for all } n \text{ and } j.$$

Thus Theorem 5.1 holds in this case.

EXAMPLE 5.2. We consider an algoritum for a sequence $\{1/n^{\alpha}\}_{n=1}^{\infty}$, $\left(\frac{1}{2} < \alpha < 1\right)$ and a model with s categories which satisfies (5.23), (5.24), (5.25), (5.27) and a following condition: there exists a set of real numbers $\{q_i\}_{i=1}^{s}$ such that

(5.34) $$E[(q^{(n+1)}(\theta_1 \mid \theta^n) - q_i)^2] = O((n+1)^{-2\omega}), \qquad \omega > \alpha$$

is true for all n and i, where $0 \leqq q_i \leqq 1$ for $i = 1, 2, \cdots, s$ and $\sum_{i=1}^{\infty} q_i = 1$. Further, it is assumed that the same system as a given system of orthonormal functions $\{\varphi_i(x)\}_{i=1}^{N}$ in Example 5.1 is given. By using (5.30) and (5.34), we can easily obtain

(5.35) $$E[(\theta_j^{(n)})^2] = O(n^{-2\omega})$$

and

(5.36) $$\text{Var}\,[\varphi_j(x^{n+1}) \mid \theta^n] \leqq K \quad \text{for all } n \text{ and all } j,$$

where K is a certain constant. Thus the result of Theorem 5.2 holds in this case.

§ 6. Acknowledgement

The author is deeply indebted to Professor T. Kitagawa for his helpful advices and critical readings of the original manuscript. The author is also grateful to Professors S. Kanō and N. Furukawa for their advices and encouragements.

References

[1] BLUM, J. R.: *Approximation methods which converge with probability one,* Ann. Math. Stat., Vol. 25, (1954), 382-386.

[2] BLUM, J. R.: *Multidimensonal stochastic approximation methods,* Ann. Math. Stat., Vol. 25, (1954), 737-744.

[3] BRAVERMAN, E. M.: *On the method of potential functions,* Autom. and Remote Control, Vol. 26, No. 12, (1965), 2130-2138.

[4] BRAVERMAN, E. M. and PYATNITSKII, E. S.: *Estimation of the rate of convergence of algorithms based on the potential function method,* Autom. and Remote Control, Vol. 27, No. 1, (1966), 80-100.

[5] Еравермаи, Э. М. и Розонозр, П. И.: Сходимость случайных процессов в теории обучения машин. I. Автоматика и телемеханика, No. 1, (1969), 57-77.

[6] Еравермаи, Э. М. и Розонозр, П. И.: Сходимость случайных процессов в теории обучения машин. II, Автоматика и телемеханика, No. 3, (1969), 87-103.

[7] CHIEN, Y. T. and FU, K. S.: *On bayesian learning and stochastic approximation,* Proceed-

ings of the fourth Annual Allerton Conference on circuit and system, (1966), 107–115.

[8] CHUNG, K. L.: *On a stochastic approximation method*, Ann. Math. Stat., Vol. 25, (1954), 463–483.

[9] DUPAČ, V.: *A dynamic stochasic approximation method*, Ann. Math. Stat., Vol. 36, (1965), 1695–1702.

[10] DVORETZKY, A.: *On stochastic approximation*, Proceedings of the third Berkeley symposium on Mathematical Statistics and Probability, Vol. 1, (1956), 39–55.

[11] DYNKIN, E. B.: *Controlled random sequence*, Theory of probability and its application, Vol. 10, No. 1, (1965), 1–14.

[12] FELDBAUM, A. A.: *"Optimal Control System"*, Academic Press, N. Y., 1965.

[13] GLADYSHEV, E. G.: *On stochastic approximation*, Theory of probability and its application, Vol. 10, No. 2, (1965), 275–278.

[14] KITAGAWA, T.: *Successive process of statistical controls I*, Mem. Fac. Sci. Kyushu Univ., Ser. A., Vol. 7, (1952), 13–28.

[15] KITAGAWA, T.: *Successive process of statistical controls II*, Mem. Fac. Sci. Kyushu Univ., Ser. A., Vol. 13, (1959), 1–16.

[16] LITVAKOV, B. M.: *Convergence of recurrent algorithms for pattern recognition learning*, Autom. and Remote Control, Vol. 29, No. 1, (1968), 121–128.

[17] LOGINOV, N. V.: *The methods of stochastic approximation*, Autom. and Remote Control, Vol. 27, No. 4, (1966), 706–728.

[18] MEISEL, W. S.: *Least-square methods in abstract pattern recognition*, Information Science, Vol. 1, No. 1, (1968), 43–54.

[19] MIDDLETON, D.: *"An introduction to statistical comunication theory"*, McGraw-Hill, N.Y., 1960.

[20] PATTERSON, J. D., WAGNER, T. J. and WOMACK, B. F.: *A mean-square performance criterion for adaptive pattern classification systems*, IEEE Trans. Automatic Control, Vol. AC–12, No. 2, (1967), 195–197.

[21] ROBBINS, H. and MONRO, S.: *A stochastic approximation method*, Ann. Math. Stat., Vol. 22, (1951), 400–407.

[22] SCHUMPERT, J. M. and YAU, S. S.: *Design of pattern classifiers with the updating property using stochastic approximation techniques*, IEEE Trans. Computers, Vol. C–17, No. 9, (1968), 861–872.

[23] SPRAGINS, J.: *Learning without a teacher*, IEEE Trans. Information Theory, Vol. IT–12, (1966), 223–230.

[24] TSYPKIN, Ya. Z.: *Use of the stochastic approximation method in estimating unknown distribution densities from observations*, Autom. and Remote Control, Vol. 27, No. 3, (1966), 432–434.

[25] TSYPKIN, Ya. Z.: *Algorithms for the estimation of distribution densities and mements from observations*, Vol. 28, No. 7, (1967), 1084–1087.

[26] TSYPKIN, Ya. Z.: *All the same, does a theory of synthesis of optimal adaptive system exist?*, Autom. and Remote Control, Vol. 29, No. 1, (1968), 93–98.

[27] TSYPKIN, Ya. Z.: *Self-learning what is it*, IEEE Trans. Automatic Control, Vol. AC–13, No. 6, (1968), 608–612.

[28] WAGNER, T. J. and WOLVERTON, C. T.: *Asymptotically optimal discriminant functions for pattern classification*, IEEE Trans. Information Theory, Vol. IT–15, No. 2, (1969), 258–265.

[29] WASAN, M.T.: *"Stochastic Approximation"*, Cambridge Univ. Press, 1969.