ON THE PATTERN CLASSIFICATION BY LEARNING

Tanaka, Kensuke
Department of mathematics, Faculty of Science, Niigata University

https://doi.org/10.5109/13050

出版情報：統計数理研究. 14 (3/4), pp.13-25, 1971-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON THE PATTERN CLASSIFICATION BY LEARNING

By

Kensuke TANAKA*

(Received December 25, 1970)

§ 1. Introduction and Summary

This paper is a continuation of our papers ([10], [11], [12]) and is concerned with the pattern classification related to "learning with a teacher" and "learning without a teacher". In the problem related to "learning with a teacher", we tried to find an algorithm by which, for two given categories, we can construct, from a training sequence, a sequence of linear systems of orthonormal functions to get an approximation to a sequence of the Bayes decision functions which are optimum in the sense of minimizing the probability of misclassification at each instant.

Now, we recall the algorithm of learning developed in [13]. In [13], the approximation to the Bayes decision functions for each fixed point of domain is nonparametric and the existence of a training sequence from the independent random variables is assumed. However, it does not seem general enough that the patterns are observed from the independent random variables in the statistical pattern recognition. On this reason, we replace the above assumption by weaker one that the patterns are observed from the dependent random variables with certain conditions. In the above situation, we shall try to extend the algorithm of learning developed in [13].

Next, in the problem related to "learning without a teacher", we pay attention to the algorithm of learning developed in [9]. In [9], the probability density function of the patterns is the following form:

$$p(x)=\sum_{i=1}^{s} q_i f_i(x)\,,$$

where q_i is a priori probability of the category θ_i and $f_i(x)$ is the probability density function of an observed pattern which belongs to the category θ_i.

When $f_i(x)$ is known to the observer for all i, the algorithm is given for estimating a priori probabilities q_i, $i=1, 2, \cdots, s$, on the basis of the unclassified observed patterns. Here, we make the above situation weaker in the following two points: (i) the categories at each instant are from the dependent random variables with certain conditions, (ii) the probability density function is time-variant. In this situation, we shall try to extend the algorithm developed in [9].

Our algorithms in both learning are an application of the method introduced by T. Kitagawa [7] in the successive process of statistical control. This method, which

* Department of Mathematics Faculty of Science, Niigata University, Niigata.

may be called modified stochastic approximation, was investigated by V. Dupač [4] in detail.

This paper consists of six sections. In Section 2, we shall state several lemmas necessary for the proofs of main results of this paper. In Section 3, we shall give the formulation of the problem in this paper. In Section 4, we shall investigate "learning with a teacher" in the case when there are two categories. In Section 5, we shall be concerned with "learning without a teacher" in the case when there are many categories.

§ 2. Preliminaries

In this section, two lemmas are stated without proof in order to prove main results of this paper. Let us consider an m-dimensional 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 defined on R^{nm}. Then accordingly $\{U_n(y^1, \cdots, y^n)\}_{n=1}^{\infty}$, $\{V_n(y^1, \cdots, y^n)\}_{n=1}^{\infty}$ and $\{\zeta_n(y^1, \cdots, y^n)\}_{n=1}^{\infty}$ become again three stochastic processes, respectively. 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 expectations of three stochastic variables U_{n+1}, V_{n+1} and ζ_{n+1} given the random variables $y^1, y^2, \cdots, y^n$ 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) $\sum_{n=1}^{\infty} \gamma_n = \infty \quad (\gamma_n \geqq 0,\ n = 1, 2, \cdots)$,

(A4) $\sum_{n=1}^{\infty} |\mu_n| < \infty$,

(A5) there exists a sequence of positive numbers $\{M_n\}_{n=1}^{\infty}$ such that

$$P[\zeta_n \leqq M_n] = 1 \qquad \text{for all } n,$$

and such that $\sum_{n=1}^{\infty} M_n < \infty$.

The following Lemma 1 and Lemma 2 were essentially proved in [10].

LEMMA 1. *Let the hypotheses for three stochastic processes $\{U_n\}_{n=1}^{\infty}$, $\{V_n\}_{n=1}^{\infty}$ and $\{\zeta_n\}_{n=1}^{\infty}$ be satisfied:* (i) *conditions* (A1)$\sim$(A5) *hold,* (ii) $\lim_{n\to\infty} \gamma_n = 0$, (iii) *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, it holds that*

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

LEMMA 2. *Suppose that a sequence of non-negative real numbers $\{a_n\}_{n=1}^{\infty}$ satisfies the condition: there exist a positive integer n_0, two sequences of non-negative real num-*

bers $\{\gamma_n\}_{n=1}^{\infty}$ *and* $\{A_n\}_{n=1}^{\infty}$ *such that*

$$a_{n+1} \leqq (1-\gamma_{n+1})a_n + A_{n+1} \qquad \textit{for all } n \geqq n_0 , \tag{2.1}$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty , \tag{2.2}$$

$$\lim_{n\to\infty} \gamma_n = 0 , \tag{2.3}$$

$$\sum_{n=1}^{\infty} A_n < \infty . \tag{2.4}$$

Then, it holds that $\lim_{n\to\infty} a_n = 0$.

Next, we mention without proof the lemma given by V. Dupač [4], 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 numbers* A *and* B *such that*

$$a_{n+1} \leqq (1-A/n^s)a_n + B/n^t \qquad \textit{for all } n \geqq n_0 , \tag{2.5}$$

$$t \textit{ real number and } 0 < s < 1 . \tag{2.6}$$

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

§ 3. The formulation of the problem

In the pattern classification problem in this paper, each observed pattern x is a random sample taking value in R^m from a group, to which they belong, and each group is called a category. Therefore, each observed pattern is drawn by a probability distribution law. Now, we consider the case when there exist s categories $\theta_1, \theta_2, \cdots, \theta_s$ and we denote a set of these categories by Θ. Hence, each outcome in pattern classification is described by a pair (x, θ). The element x is an observed pattern in a pattern space R^m and θ specifies the category of an observed pattern. But, generally, θ is unknown to the observer. For a sequence of the observed patterns $x^1, x^2, \cdots, x^n, \cdots$ we can consider a sequence:

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

with $x^n \in R^m$ and $\theta^n \in \Theta$, where $\theta^n = \theta_i$ if x^n is from a specific category θ_i.

For such a sequence, the result of n history is expressed by two sets:

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

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 w.r.t. Lebesgue measure and this is denoted by the following form:

$$p(x^{n+1}, \theta^{n+1} | \alpha^n) = q^{(n+1)}(\theta^{n+1} | \alpha^n) f^{(n+1)}_{\theta^{n+1}}(x^{n+1}) , \tag{3.2}$$

where $q^{(n+1)}(\theta^{n+1} | \alpha^n)$ is the conditional probability of a category θ^{n+1} at instant $n+1$ given a history α^n and $f^{(n+1)}_{\theta^{n+1}}(x^{n+1})$ is a conditional probability density function of the observed pattern x^{n+1} given a category θ^{n+1} at instant $n+1$.

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

$$\Pi_{x^{n+1}}(\alpha^n\theta^{n+1})=\frac{\Pi(\alpha^n)q^{(n+1)}(\theta^{n+1}\mid\alpha^n)f^{(n+1)}_{\theta^{n+1}}(x^{n+1})}{\sum_{\tilde{\alpha}^n\in\Theta^n}\sum_{\tilde{\theta}^{n+1}\in\Theta}\Pi(\tilde{\alpha}^n)q^{(n+1)}(\tilde{\theta}^{n+1}\mid\tilde{\alpha}_n)f^{(n+1)}_{\tilde{\theta}^{n+1}}(x^{n+1})}, \tag{3.3}$$

where $\Theta^n=\Theta\times\Theta\times\cdots\times\Theta$ and $\Pi(\alpha^n)$ is a probability distribution on Θ^n.

Then, if all transition probability density functions at each instant are known to the observer, the classification of an observed pattern at each instant will be determined by the largest of the quantities $\Pi_{x^{n+1}}(\alpha_n\theta_1)$, $\Pi_{x^{n+1}}(\alpha^n\theta_2)$, $\cdots$, $\Pi_{x^{n+1}}(\alpha^n\theta_s)$. From the statistical decision theory, it is well known that this decision rule is optimal, because of the minimum probability of misclassification, and that it is called the Bayes decision rule.

Hence, in the case when there are two categories, we have an optimum decision rule:

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

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

where $D^*(x^{n+1}\mid\alpha^n)=\Pi_{x^{n+1}}(\alpha^n\theta_1)-\Pi_{x^{n+1}}(\alpha^n\theta_2)$.

This decision rule is equivalent to the following decision rule:

x^{n+1} is classified in category θ_1 if $D_0^{(n+1)}(x^{n+1}\mid\alpha^n)\geqq 0$,

is classified in category θ_2 if $D_0^{(n+1)}(x^{n+1}\mid\alpha^n)<0$,

where $D_0^{(n+1)}(x^{n+1}\mid\alpha^n)=q^{(n+1)}(\theta_1\mid\alpha^n)f^{(n+1)}_{\theta_1}(x^{n+1})-q^{(n+1)}(\theta_2\mid\alpha^n)f^{(n+1)}_{\theta_2}(x^{n+1})$.

§ 4. Learning with a teacher

In this section, we treat the case where the amount of a priori information on the transition probability density function at each instant 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 $(x^1, \theta^1), (x^2, \theta^2), \cdots$, , where θ^i is the category indicated by a teacher at instant i.

Now, when $\Theta=\{\theta_1, \theta_2\}$, we consider the problem of finding an algorithm of approximation to the Bayes decision function at each instant, on the basis of a training sequence. This decision which minimizes the probability of misclassification at each instant n is the following form:

$$D_0^{(n)}(x^n\mid\alpha^{n-1})=q^{(n)}(\theta_1\mid\alpha^{n-1})f_{\theta_1}(x^n)-q^{(n)}(\theta_2\mid\alpha^{n-1})f_{\theta_2}(x^n). \tag{4.1}$$

In what follows, each $f_{\theta_i}(x)$ is defined on R^m and all integrals and supremums, unless otherwise indicated, are taken over R^m.

Here, let $K(\cdot)$ be a real-valued function on R^m satisfying the following conditions:

(K1) $\quad K(y)\geqq 0 \quad$ for all $y\in R^m$,

(K2) $\quad \sup K(y)=K<\infty$,

(K3) $$\int K(y)dy=1\,,$$

(K4) $$\int \|y\|K(y)dy=K^*<\infty\,,$$

where the norm $\|y(y_1, y_2, \cdots, y_m)\|^2=\sum_{i=1}^{m}(y_i)^2$.

By using the above function $K(\cdot)$, we shall construct the following algorithm with a sequence of positive real numbers $\{h_n\}_{n=1}^{\infty}$ satisfying the condition:

(4.2) $$1\geqq h_1\geqq h_2\geqq\cdots \quad \text{and} \quad \lim_{n\to\infty}h_n=0\,.$$

Firstly, 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

(4.3) $$D_1(x\,|\,x^1, \theta^1)=\rho^{(1)}(\theta^1)K_1(x, x^1)-(1-\rho^{(1)}(\theta^1))K_1(x, x^1)\,,$$

where

$$K_1(x, x^1)=h_1^{-m}K[h_1^{-1}(x-x^1)]$$

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

(4.4) $$D_2(x\,|\,\xi^2, \alpha^2)=\frac{1}{2}D_1(x\,|\,\xi^1, \alpha^1)+\frac{1}{2}[\rho^{(2)}(\theta^2)K_2(x, x^2)-(1-\rho^{(2)}(\theta^2))K_2(x, x^2)]\,,$$

where

$$K_2(x, x^2)=h_2^{-m}K[h_2^{-1}(x-x^2)]$$

and

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

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

(4.5) $$D_{n+1}(x\,|\,\xi^{n+1}, \alpha^{n+1})=\frac{n}{n+1}D_n(x\,|\,\xi^n, \alpha^n)+\frac{1}{n+1}[\rho^{(n+1)}(\theta^{n+1})K_{n+1}(x, x^{n+1})$$
$$-(1-\rho^{(n+1)}(\theta^{n+1}))K_{n+1}(x, x^{n+1})]\,,$$

where

$$K_{n+1}(x, x^{n+1})=h_{n+1}^{-m}K[h_{n+1}^{-1}(x-x^{n+1})]$$

and

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

The next lemma is necessary in order to prove the theorem in this section.

LEMMA 4. *Let $f(\cdot)$ be a real-valued function on R^m satisfying a uniform Lipschitz condition:*

$$|f(x)-f(y)| \leqq C\|x-y\| \qquad \textit{for all } x, y \in R^m$$

and $K(\cdot)$ *satisfy* (K1)$\sim$(K4). *Then, it holds that*

$$|f_n(x)-f(x)| \leqq CK^*h_n\,, \tag{4.6}$$

where $f_n(x)=\int K_n(x, y)f(y)dy$.

PROOF.

$$\begin{aligned}|f_n(x)-f(x)| &= \left|\int K_n(x, y)(f(y)-f(x))dy\right| \\ &\leqq \int K(z)|f(x-zh_n)-f(x)|\,dz \\ &\leqq C\int K(z)\|h_nz\|dz \\ &= CK^*h_n\,.\end{aligned}$$

Then, we can prove the following theorem concerning $D_{n+1}(x|\xi^{n+1}, \alpha^{n+1})$ and $D_0^{(n+1)}(x|\alpha^n)$.

THEOREM 4.1. *Let the following hypotheses be satisfied:*

(i) $\int (f_{\theta_i}(x))^2dx < \infty$ *for all* i,

(ii) $f_{\theta_1}(x)$ *and* $f_{\theta_2}(x)$ *satisfy uniform Lipschitz condition,*

(iii) $\sum_{n=1}^{\infty} n^{-1}h_n < \infty$ *and* $\sum_{n=1}^{\infty} n^{-2}h_n^{-m} < \infty$,

(iv) *for each* i, *there exist a non-negative number* q_i $(0 \leqq q_i \leqq 1,\ \sum_{i=1}^{2} q_i = 1)$ *and a sequence of positive numbers* $\{M_n\}_{n=1}^{\infty}$ *such that*

$$n(q^{(n)}(\theta_i|\alpha^{n-1})-q_i)^2 \leqq M_n$$

and such that $\sum_{n=1}^{\infty} M_n < \infty$. *Then, it holds that*

$$P[\lim_{n\to\infty} I_n = 0] = 1 \quad \textit{and} \quad \lim_{n\to\infty} E[I_n^{\beta}] = 0 \qquad \textit{for all } 0 < \beta < 1\,,$$

where

$$I_n = \int [D_n(x|\xi^n, \alpha^n) - D_0^{(n)}(x|\alpha^{n-1})]^2dx\,.$$

PROOF. By the construction of $D_{n+1}(x|\xi^{n+1}, \alpha^{n+1})$, we have

$$\begin{aligned}&D_{n+1}(x|\xi^{n+1}, \alpha^{n+1}) - D_0^{(n+1)}(x|\alpha^n) \\ &= \left(1-\frac{1}{n+1}\right)D_n(x|\xi^n, \alpha^n) + \frac{1}{n+1}[\rho^{(n+1)}(\theta^{n+1})K_{n+1}(x, x^{n+1}) \\ &\quad -(1-\rho^{(n+1)}(\theta^{n+1}))K_{n+1}(x, x^{n+1})] - D_0^{(n+1)}(x|\alpha^n) \\ &= \left(1-\frac{1}{n+1}\right)(D_n(x|\xi^n, \alpha_n) - D_0^{(n)}(x|\alpha^{n-1})) + \frac{1}{n+1}[\rho^{(n+1)}(\theta^{n+1})K_{n+1}(x, x^{n+1}) \\ &\quad -(1-\rho^{(n+1)}(\theta^{n+1}))K_{n+1}(x, x^{n+1}) - D_0^{(n+1)}(x|\alpha^n)] \\ &\quad +\left(1-\frac{1}{n+1}\right)[D_0^{(n)}(x|\alpha^{n-1}) - D_0^{(n+1)}(x|\alpha^n)]\,.\end{aligned} \tag{4.7}$$

Hence, from (4.7), we can obtain

$$(4.8)\quad I_{n+1}=\left(1-\frac{1}{n+1}\right)^2 I_n+\left(\frac{n}{n+1}\right)^2\Theta^{(n)}+\left(\frac{1}{n+1}\right)^2\int[Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n)]^2dx$$
$$+2\left(\frac{n}{n+1}\right)\left(\frac{1}{n+1}\right)\int[D_n(x|\xi^n,\alpha^n)-D_0^{(n)}(x|\alpha^{n-1})][Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n)]dx$$
$$+2\left(\frac{n}{n+1}\right)\left(\frac{1}{n+1}\right)\int[D_0^{(n)}(x|\alpha^{n-1})-D_0^{(n+1)}(x|\alpha^n)][Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n)]dx$$
$$+2\left(\frac{n}{n+1}\right)^2\int[D_n(x|\xi^n,\alpha^n)-D_0^{(n)}(x|\alpha^{n-1})][D_0^{(n)}(x|\alpha^{n-1})-D_0^{(n+1)}(x|\alpha^n)]dx,$$

where

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

and

$$\Theta^{(n)}=\int[D_0^{(n)}(x|\alpha^{n-1})-D_0^{(n+1)}(x|\alpha^n)]^2dx.$$

Taking conditional expectation on both sides of (4.8), we have

$$(4.9)\quad E[I_{n+1}|\xi^n,\alpha^n]=\left(1-\frac{1}{n+1}\right)^2 I_n+\left(\frac{n}{n+1}\right)^2\Theta^{(n)}$$
$$+\left(\frac{1}{n+1}\right)^2 E\left[\int(Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n))^2dx|\xi^n,\alpha^n\right]$$
$$+2\left(\frac{n}{n+1}\right)\left(\frac{1}{n+1}\right)E\left[\int(D_n(x|\xi^n,\alpha^n)-D_0^{(n)}(x|\alpha^{n-1}))(Y_{n+1}(x,x^{n+1})\right.$$
$$\left.-D_0^{(n+1)}(x|\alpha^n))dx|\xi^n,\alpha^n\right]$$
$$+2\left(\frac{n}{n+1}\right)\left(\frac{1}{n+1}\right)E\left[\int(D_0^{(n)}(x|\alpha^{n-1})-D_0^{(n+1)}(x|\alpha^n))(Y_{n+1}(x,x^{n+1})\right.$$
$$\left.-D_0^{(n+1)}(x|\alpha^n))dx|\xi^n,\alpha^n\right]$$
$$+2\left(\frac{n}{n+1}\right)^2 E\left[\int(D_n(x|\xi^n,\alpha^n)-D_0^{(n)}(x|\alpha^{n-1}))(D_0^{(n)}(x|\alpha^{n-1})\right.$$
$$\left.-D_0^{(n+1)}(x|\alpha^n))dx|\xi^n,\alpha^n\right].$$

Then, there exist two positive numbers M_1, M_2 such that, from Lemma 4 and (K3),

$$(4.10)\quad |E[(Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n))|\xi^n,\alpha^n]|$$
$$=\left|\int K_{n+1}(x,y)[q^{(n+1)}(\theta_1|\alpha^n)f_{\theta_1}(y)-q^{(n+1)}(\theta_2|\alpha^n)f_{\theta_2}(y)]dy-D_0^{(n+1)}(x|\alpha^n)\right|$$
$$=\left|\int K(z)[q^{(n+1)}(\theta_1|\alpha^n)f_{\theta_1}(x-h_{n+1}z)-q^{(n+1)}(\theta_1|\alpha^n)f_{\theta_2}(x-h_{n+1}z)\right.$$
$$\left.-(q^{(n+1)}(\theta_1|\alpha^n)f_{\theta_1}(x)-q^{(n+1)}(\theta_2|\alpha_n)f_{\theta_2}(x)]dz\right|$$
$$\leqq M_1K^*h_{n+1}$$

and such that, from (K2) and (K3),

$$(4.11)\quad E\left[\int(Y_{n+1}(x,x^{n+1})-D_0^{(n+1)}(x|\alpha^n))^2dx|\xi^n,\alpha^n\right]$$
$$\leqq 2E\left[\int(Y_{n+1}(x,x^{n+1}))^2dx|\xi^n,\alpha^n\right]+2\int(D_0^{(n+1)}(x|\alpha^n))^2dx$$

$$\leqq 2Kh_{n+1}^{-m}E\Big[\int K_{n+1}(x, x^{n+1})dx\,|\,\xi^n, \alpha^n\Big]+4\Big[\sum_{i=1}^{2}\int(f_{\theta_i}(x))^2dx\Big]$$

$$=2Kh_{n+1}^{-m}+4M_2\,.$$

Furthermore, there exist two positive numbers M_3, M_4 such that, from (K3),

$$\text{(4.12)}\qquad \Big|\int[D_n(x\,|\,\xi^n, \alpha^n)-D_0^{(n)}(x\,|\,\alpha^{n-1})]dx\Big|\leqq\int|D_n(x\,|\,\xi^n, \alpha^n)|\,dx+\int|D_0^{(n)}(x\,|\,\alpha^{n-1})|\,dx$$

$$\leqq\frac{1}{n}\sum_{i=1}^{n}\int K_i(x, x^i)dx+\int[q^{(n+1)}(\theta_1\,|\,\alpha^n)f_{\theta_1}(x)+q^{(n+1)}(\theta_2\,|\,\alpha^n)f_{\theta_2}(x)]dx$$

$$\leqq M_3$$

and such that

$$\text{(4.13)}\qquad \Big|\int[D_0^{(n)}(x\,|\,\alpha^{n-1})-D_0^{(n+1)}(x\,|\,\alpha^n)]dx\Big|\leqq M_4\,.$$

Noting that, for each instant n,

$$\text{(4.14)}\qquad 2\int|D_0^{(n)}(x\,|\,\alpha^{n-1})-D_0^{(n+1)}(x\,|\,\alpha^n)|\,|D_n(x\,|\,\xi^n, \alpha^n)-D_0^{(n)}(x\,|\,\alpha^{n-1})|\,dx$$

$$\leqq(n+1)^{-1}I_n+(n+1)\Theta^{(n)}\,,$$

from (4.9), (4.10), (4.11), (4.12), (4.13) and (4.14)

$$\text{(4.15)}\qquad E[I_{n+1}\,|\,\xi^n, \alpha^n]\leqq\Big(1-\frac{1}{n+1}\Big)^2I_n+\Theta^{(n)}+\Big(\frac{1}{n+1}\Big)^2(2Kh_{n+1}^{-m}+4M_2)$$

$$+2\Big(\frac{1}{n+1}\Big)M_1M_3K*h_{n+1}+2\Big(\frac{1}{n+1}\Big)M_1M_4K*h_{n+1}+\Big[\Big(\frac{1}{n+1}\Big)I_n+(n+1)\Theta^{(n)}\Big].$$

Since there exists a positive number M_5 such that, from (i) and (ii) in the theorem,

$$(n+1)\Theta^{(n)}\leqq M_5M_{n+1}\,,$$

from (4.15) we can obtain

$$\text{(4.16)}\qquad E[I_{n+1}\,|\,\xi^n, \alpha^n]\leqq\Big[1+\Big(\frac{1}{n+1}\Big)^2\Big]I_n-\Big(\frac{1}{n+1}\Big)I_n+\Big(\frac{1}{n+1}\Big)^2(2Kh_{n+1}^{-m}+4M_2)$$

$$+2\Big(\frac{1}{n+1}\Big)M_1(M_3+M_4)K*h_{n+1}+2M_5M_{n+1}\,.$$

Therefore, letting $U_n=V_n=I_n$, $\mu_n=(1+n)^{-2}$ and

$$\zeta_n=\Big(\frac{1}{n+1}\Big)^2(2Kh_{n+1}^{-m}+4M_2)+2\Big(\frac{1}{n+1}\Big)M_1(M_3+M_4)K*h_{n+1}+2M_5M_{n+1}$$

in (4.16) and using Lemma 1, it follows that

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

THEOREM 4.2. *Let $\{D_n(x\,|\,\xi^n, \alpha^n)\}_{n=1}^{\infty}$ be a sequence of the decision functions obtained by the above argument such that*

$$P\Big\{\lim_{n\to\infty}\int[D_n(x\,|\,\xi^n, \alpha^n)-D_0^{(n)}(x\,|\,\alpha^{n-1})]^2dx=0\Big\}=1\,.$$

Then, it holds that

$$P[\lim_{n\to\infty}(P_{D_n(\cdot|\xi^n,\alpha^n)}(e)-P_{D_0^{(n)}(\cdot|\alpha^{n-1})}(e))=0]=1,$$

where $P_{D(\cdot)}(e)$ *is the probability of misclassification using a decision function* $D(\cdot)$.

PROOF. Choose $\varepsilon>0$ and then a sufficiently large integer N such that

$$P\Big[\int(D_N(x|\xi^N,\alpha^N)-D_0^N(x|\alpha^{N-1}))^2dx<\varepsilon^2/4\int I_{B^N}(x)dx\Big]=1 \tag{4.17}$$

where B^N is a bounded set in R^m such that

$$\int_{B^N}|D_0^N(x|\alpha^{N-1})|dx\geqq 1-\varepsilon/2$$

and $I_A(x)$ is the indicator of A.

Define sets

$$H_0^N=\{x: D_0^N(x|\alpha^{N-1})\geqq 0\} \tag{4.18}$$

and

$$H^N=\{x: D_N(x|\xi^N,\alpha^N)\geqq 0\}. \tag{4.19}$$

$$\begin{aligned}P_{D_0^N(\cdot|\alpha^{N-1})}(e)&=q^N(\theta_1|\alpha^{N-1})\int f_{\theta_1}(x)I_{H_0^N}(x)dx+q^N(\theta_2|\alpha^{N-1})\int f_{\theta_2}(x)I_{H_0^N}(x)dx\\&=q^{(N)}(\theta_1|\alpha^{N-1})+\int[-D_0^{(N)}(x|\alpha^{N-1})]I_{H_0^N}(x)dx\end{aligned} \tag{4.20}$$

and

$$P_{D_N(\cdot|\xi^N,\alpha^N)}(e)=q^{(N)}(\theta_1|\alpha^{N-1})+\int[-D_N(x|\xi^N,\alpha^N)]I_{H^N}(x)dx. \tag{4.21}$$

Then, from (4.20) and (4.21),

$$\begin{aligned}P_{D_N(\cdot|\xi^N,\alpha^N)}(e)-P_{D_0^{(N)}(\cdot|\alpha^{N-1})}(e)&=\int D_0^{(N)}(x|\alpha^{N-1})[I_{H_0^N}(x)-I_{H^N}(x)]I_{B^N}(x)dx\\&\quad+\int D_0^{(N)}(x|\alpha^{N-1})[I_{H_0^N}(x)-I_{H^N}(x)]I_{\overline{B^N}}(x)dx,\end{aligned} \tag{4.22}$$

where $\overline{B^N}$ is the complement of set B^N.

It is obvious that

$$\int[-D_N(x|\xi^N,\alpha^N)][I_{H_0^N}(x)-I_{H^N}(x)]I_{B^N}(x)dx\geqq 0. \tag{4.23}$$

Adding (4.23) to (4.22) and recalling that $D_0^{(N)}(x|\alpha^{N-1})$ is the Bayes decision function, we can obtain that

$$\begin{aligned}0&\leqq P_{D_N(\cdot|\xi^N,\alpha^N)}(e)-P_{D_0^{(N)}(\cdot|\alpha^{N-1})}(e)\\&\leqq\int[D_0^{(N)}(x|\alpha^{N-1})-D_N(x|\xi^N,\alpha^N)][I_{H_0^N}(x)-I_{H^N}(x)]I_{B^N}(x)dx\\&\quad+\int|D_0^{(N)}(x|\alpha^{N-1})|I_{\overline{B^N}}(x)dx\\&\leqq\Big\{\int[D_0^{(N)}(x|\alpha^{N-1})-D_N(x|\xi^N,\alpha^N)]^2dx\cdot\int I_{B^N}(x)dx\Big\}^{1/2}+\varepsilon/2.\end{aligned} \tag{4.25}$$

Since ε was arbitrary, the theorem is proved.

§ 5. Learning without a teacher

This section is concerned with the problem of "learning without a teacher" in statistical pattern recognition. We treat the case when there exist s categories $\theta_1, \theta_2, \cdots, \theta_s$, and the case when there can not be assumed a training sequence.

In what follows, it is assumed that, for each instant n, the transition probability density function has the following form:

$$p^{(n)}(x^n|\alpha^{n-1}) = \sum_{i=1}^{s} q^{(n)}(\theta_i|\alpha^{n-1}) f_{\theta_i}^{(n)}(x^n), \tag{5.1}$$

where $\alpha^{n-1} = (\theta^1, \theta^2, \cdots, \theta^{n-1})$ and each $f_{\theta_i}^{(n)}(x^n)$ is defined on R^m.

Now, we consider the problem of finding an algorithm of the estimation of the probabilities $q^{(n)}(\theta_i|\alpha^{n-1})$, $i=1, 2, \cdots, s$, in the mixture density function $p^{(n)}(x^n|\alpha^{n-1})$, when $f_{\theta_i}^{(n)}(x)$ is known to the observer for every i and n, on the basis of the observed but unclassified patterns. Here, we reduce this problem to the problem of finding an algorithm by which, at each instant n, we can construct, from the observed patterns, $q_{i*}^{(n)}$, $i=1, 2, \cdots, s$, which minimizes a quantity defined by

$$I_n = \int [\sum_{i=1}^{s} (\hat{q}_i - q^{(n)}(\theta_i|\alpha^{n-1})) f_{\theta_i}^{(n)}(x^n)]^2 dx^n + 2\lambda[\sum_{i=1}^{s} \hat{q}_i - 1], \tag{5.2}$$

where λ is a Lagrange multiplier.

Differentiating I_n with respect to $\hat{q}_i$, $i=1, 2, \cdots, s$, and equating the derivatives to zero, we have

$$W^{(n)} Q_*^{(n)}(\alpha^{n-1}) = E[\boldsymbol{f}^{(n)}(x^n)|\alpha^{n-1}] - \lambda U, \tag{5.3}$$

where $W^{(n)}$ is the matrix with elements $w_{ij}^{(n)} = \int f_{\theta_i}^{(n)}(x) f_{\theta_j}^{(n)}(x) dx$, $i, j = 1, 2, \cdots, s$; $E[\boldsymbol{f}^{(n)}(x^n)|\alpha^{n-1}]$ is a column vector of with the i-th component equal to $E[f_{\theta_i}^{(n)}(x^n)|\alpha^{n-1}] = \int f_{\theta_i}^{(n)}(x^n) p^{(n)}(x^n|\alpha^{n-1}) dx^n$, $i=1, 2, \cdots, s$; U is a column vector of s components all equal to one and $Q_*^{(n)}(\alpha^{n-1})$ is a column vector with the i-th component $q_{i*}^{(n)}$, $i=1, 2, \cdots, s$. When $\det W^{(n)}$, the determinant of $W^{(n)}$, is not equal to zero at each instant n, the i-th component of $Q_*^{(n)}(\alpha^{n-1})$ which satisfies (5.3) is

$$q_{i*}^{(n)} = \sum_{k=1}^{s} \left(E[f_{\theta_k}^{(n)}(x^n)|\alpha^{n-1}] - \frac{\sum_{l=1}^{s}\sum_{j=1}^{s} E[f_{\theta_j}^{(n)}(x^n)|\alpha^{n-1}] W_{jl}^{(n)} - \det W^{(n)}}{\sum_{l=1}^{s}\sum_{j=1}^{s} W_{jl}^{(n)}} \right) \frac{W_{ki}^{(n)}}{\det W^{(n)}} \tag{5.4}$$

where $W_{jl}^{(n)}$ is the adjunct of $w_{jl}^{(n)}$ in the matrix $W^{(n)}$. Then, we can write (5.4) to the following form:

$$q_{i*}^{(n)} = E[F_i^{(n)}(x^n)|\alpha^{n-1}], \tag{5.5}$$

where

$$F_i^{(n)}(x^n)=\sum_{k=1}^{s}\left(f_{\theta k}^{(n)}(x^n)-\frac{\sum_{l=1}^{s}\sum_{j=1}^{s}f_{\theta j}^{(n)}(x^n)W_{jl}^{(n)}-\det W^{(n)}}{\sum_{l=1}^{s}\sum_{j=1}^{s}W_{jl}^{(n)}}\right)\frac{W_{ki}^{(n)}}{\det W^{(n)}}\,.$$

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

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

Firstly, using an observed but unclassified pattern x^1 at instant 1, we make for $i=1, 2, \cdots, s$

$$g_i^{(1)}(\xi^1, \alpha^1)=g_i^{(0)}+\gamma_1[F_i^{(1)}(x^1)-g_i^{(0)}]\,, \tag{5.7}$$

where $g_i^{(0)}=0$ for all i.
Secondly, using an observed but unclassified pattern x^2 at instant 2, we make for $i=1, 2, \cdots, s$

$$g_i^{(2)}(\xi^2, \alpha^2)=g_i^{(1)}(\xi^1, \alpha^1)+\gamma_2[F_i^{(2)}(x^2)-g_i^{(1)}(\xi^1, \alpha^1)]\,. \tag{5.8}$$

In general, using an observed but unclassified pattern x^{n+1} at instant $n+1$, we make for $i=1, 2, \cdots, s$

$$g_i^{(n+1)}(\xi^{n+1}, \alpha^{n+1})=g_i^{(n)}(\xi^n, \alpha^n)+\gamma_{n+1}[F_i^{(n+1)}(x^{n+1})-g_i^{(n)}(\xi^n, \alpha^n)]\,. \tag{5.9}$$

Then, we can prove the following theorem concerning $g_i^{(n+1)}(\xi^{n+1}, \alpha^{n+1})$ and $q_{i*}^{(n+1)}$.

THEOREM 5.1. *Let the following hypotheses be satisfied:*

(i) *for each instant n, $\det W^{(n)}$ is not equal to zero,*

(ii) *there exist a set of positive numbers $\{q_i\}_{i=1}^{s}$, $(0\leqq q_i\leqq 1, \sum_{i=1}^{s}q_i=1)$ and a sequence of positive numbers $\{M_n\}_{n=1}^{\infty}$ such that, for all i,*

$$\gamma_n^{-1}[q^{(n)}(\theta_i\,|\,\alpha^{n-1})-q_i]^2\leqq M_n$$

and such that $\sum_{n=1}^{\infty}M_n<\infty$.

(iii) *$\{f_{\theta_i}(x)\}_{i=1}^{s}$ is a set of the conditional probability density functions defined on R^m and satisfying $\int f_{\theta_i}(x)f_{\theta_j}(x)dx<\infty$, for $i, j=1, 2, \cdots, s$.*

(iv) *there exists a sequence of positive numbers $\{N_n\}_{n=1}^{\infty}$ such that, for each n,*

$$\gamma_n^{-1}\Big(\int f_{\theta i}^{(n)}(x)f_{\theta j}^{(n)}(x)dx-\int f_{\theta_i}(x)f_{\theta_j}(x)\Big)^2\leqq N_n$$

and such that $\sum_{n=1}^{\infty}N_n<\infty$.

Then, it holds that, for all i,

$$P[\lim_{n\to\infty}u_i^{(n)}=0]=1 \quad \textit{and} \quad \lim_{n\to\infty}E[u_n^{2\beta}]=0 \qquad \textit{for all } 0<\beta\leqq 1\,,$$

where $u_i^{(n)}=g_i^{(n)}(\xi^n, \alpha^n)-q_{i*}^{(n)}$.

PROOF. By the construction $g_i^{(n+1)}(\xi^{n+1}, \alpha^{n+1})$, $i=1, 2, \cdots, s$, we have

$$(5.10)\quad g_i^{(n+1)}(\xi^{n+1}, \alpha^{n+1})-q_{i*}^{(n+1)} = g_i^{(n)}(\xi^n, \alpha^n)+\gamma_{n+1}[F_i^{(n+1)}(x^{n+1})-g_i^{(n)}(\xi^n, \alpha^n)]-q_{i*}^{(n+1)}$$
$$=(1-\gamma_{n+1})(g_i^{(n)}(\xi^n, \alpha^n)-q_{i*}^{(n)})+(1-\gamma_{n+1})(q_{i*}^{(n)}-q_{i*}^{(n+1)})$$
$$+\gamma_{n+1}[F_i^{(n+1)}(x^{n+1})-q_{i*}^{(n+1)}]\,.$$

The equality (5.10) can be written in terms of $u_i^{(n+1)}$, $u_i^{(n)}$ and $\theta_i^{(n)}$ as

$$(5.11)\quad u_i^{(n+1)}=(1-\gamma_{n+1})u_i^{(n)}+(1-\gamma_{n+1})\theta_i^{(n)}+\gamma_{n+1}[F_i^{(n+1)}(x^{n+1})-q_{i*}^{(n+1)}]\,,$$

where $\theta_i^{(n)}=q_{i*}^{(n)}-q_{i*}^{(n+1)}$.

Squareing both sides of (5.11) and then taking conditional expectation, we can obtain, for all i and a sufficiently large n,

$$(5.12)\quad E[(u_i^{(n+1)})^2|\xi^n, \alpha^n]\leqq(1-\gamma_{n+1})^2(u_i^{(n)})^2+(1-\gamma_{n+1})^2(\theta_i^{(n)})^2$$
$$+2(1-\gamma_{n+1})^2|u_i^{(n)}||\theta_i^{(n)}|+\gamma_{n+1}^2\sigma^2\,,$$

where the positive number σ^2 satisfies a condition $\mathrm{Var}\,[F_i^{(n+1)}(x^{n+1})|\alpha^n]\leqq\sigma^2$. Noting that, for $i=1, 2, 3, \cdots, s$ and all n,

$$2|u_i^{(n)}||\theta_i^{(n)}|\leqq\gamma_{n+1}(u_i^{(n)})^2+\gamma_{n+1}^{-1}(\theta_i^{(n)})^2\,,$$

from (5.12) we have for all i

$$(5.13)\quad E[(u_i^{(n+1)})^2|\xi^n, \alpha^n]\leqq(1-\gamma_{n+1})^2(u_i^{(n)})^2-\gamma_{n+1}(u_i^{(n)})^2+\gamma_{n+1}^{-1}(\theta_i^{(n)})^2+(\theta_i^{(n)})^2+\gamma_{n+1}^2\sigma^2\,.$$

Then, from (ii), (iii) and (iv), there exists a positive number M such that

$$(5.14)\quad \gamma_{n+1}^{-1}(\theta_i^{(n)})^2+(\theta_i^{(n)})^2\leqq MM_n\,.$$

From (5.14), we can write (5.13) to the following form:

$$(5.15)\quad E[(u_i^{(n+1)})^2|\xi^n, \alpha^n]\leqq(1+\gamma_{n+1}^{2})(u_i^{(n)})^2-\gamma_{n+1}(u_i^{(n)})^2+MM_n+\gamma_{n+1}^2\sigma^2\,.$$

Therefore, by Lemma 1, it follows that, for all i,

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

Also, taking the unconditional expectation on both sides of (5.15) and using Lemma 2, it follows that, for all i,

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

Thus, the proof of the theorem is completed.

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

THEOREM 5.2. *Let the following hypotheses be satisfied:*

(i) $\gamma_n=a/n^\alpha$, $a>0$, $(1/2)<\alpha<1$,

(ii) $\mathrm{Var}\,[F_i^{(n)}(x^n)|\alpha^{n-1}]\leqq\sigma^2$ *for all* i *and* n,

(iii) $E[(\theta_i^{(n)})^2]=O(n^{-2\omega})$, $\omega>\alpha$, *for all* i,

Then, it holds that

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

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

PROOF. By (i), (ii), (iii) and (5.13), there exist a positive integer N and three positive numbers C_1, C_2, C_3 such that, for all $n \geqq N$,

$$E[(u_i^{(n+1)})^2] \leqq (1-C_1/n^\alpha)E[(u_i^{(n)})^2]+C_2/n^{2\alpha}+C_3/n^{2\omega-\alpha}. \tag{5.16}$$

Consequently, we can obtain for $\omega < (3/2)\alpha$

$$E[(u_i^{(n+1)})^2] \leqq (1-C_1/n^\alpha)E[(u_i^{(n)})^2]+C_4/n^{2\omega-\alpha} \tag{5.17}$$

and for $\omega \geqq (3/2)\alpha$

$$E[(u_i^{(n+1)})^2] \leqq (1-C_1/n^\alpha)E[(u_i^{(n)})^2]+C_5/n^{2\alpha}, \tag{5.18}$$

where C_4 and C_5 are some positive numbers.

Thus, an application of Lemma 3 for $a_n = E[(u_i^{(n)})^2]$ gives us the result of the theorem.

§ 6. Acknowledgement

The author is deeply indebted to Professor T. Kitagawa for his valuable advices. The author is also grateful to Professors S. Kanō and N. Furukawa for their advices and encouragements.

References

[1] Браверман, Э. М. и Розоноэр, Л. И.: Сходнмость случайных процессов в теории обучения машин. 1, Автоматика и телемеханика, Ио. 1, (1969), 57–77.

[2] Браверман, Э. М. и Розоноэр, Л. И.: Сходимость случайных процессов в теории обучения машин. 2, Автоматика и телемеханика, Ио. 3, (1969), 87–103.

[3] Chung, K. L.: *On a stochastic approximation method*, Ann. Math. Stat., vol. 25, (1954), 463–483.

[4] Dupač, V.: *A dynamic stochastic approximation method*, Ann. Math. Stat., vol. 36, (1965), 1695–1702.

[5] Fu, K. S.: "Sequential Method in Pattern Recognition and Machin Learning", Academic Press, New York, 1968.

[6] Kitagawa, T.: *Successive process of statistical controls.* 1, Mem. Fac. Sci. Kyushu Univ., Ser. A., vol. 7, (1952), 13–28.

[7] Kitagawa, T.: *Successive process of statistical controls.* 2, Mem. Fac. Sci. Kyushu Univ., Ser. A., vol. 13, (1959), 1–16.

[8] Parzen, E.: *On estimation of a probability density and mode*, Ann. Math. Stat., vol. 33, (1962), 1065–1076.

[9] Saridis, S. N., Nikolic, Z. J. and Fu, K. S.: *Stochastic approximation algorithms for system identification, estimation and decomposition of mixture*, IEEE Trans. Systems Science and Cybernetics, vol. 5, No. 1, January (1969), 8–15.

[10] Tanaka, K.: *On the pattern classification problems by learning.* 1, Bull. Math. Stat., vol. 10, (1970), 31–49.

[11] Tanaka, K.: *On the pattern classification problems by learning.* 2, Bull. Math. Stat., vol. 10, (1970), 61–73.

[12] Tanaka, K.: *On the pattern classification problems by learning.* 3, Mem. Fac. Sci. Kyushu Univ., Ser. A., vol. 24, (1970), 249–273.

[13] Wolverton, C. T. and Wagner, T. J.: *Asymptotically optimal discriminant functions for pattern classification*, IEEE Trans. Information Theory, vol. IT-15, (1969), 258–265.