ON A PATTERN CLASSIFICATION PROBLEM ON THE BASIS OF A TRAINING SEQUENCE ASSOCIATED WITH DEPENDENT RANDOM VARIABLES

Watanabe, Masafumi
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13081

出版情報：統計数理研究. 16 (1/2), pp.35-48, 1974-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON A PATTERN CLASSIFICATION PROBLEM ON THE BASIS OF A TRAINING SEQUENCE ASSOCIATED WITH DEPENDENT RANDOM VARIABLES

By

Masafumi WATANABE*

(Received July 10, 1973)

§ 1. Introduction and Summary.

In this paper we shall be concerned with the pattern classification problem related to "learning with a teacher". Many previous authors have studied this problem under the given situation of a training sequence composed of observed patterns independently sampled from a common population. From the practical point of view, however, the situation that observed patterns are independently sampled is rather restrictive. For this reason, in this paper the author treats the pattern classification problem on the basis of a dependent sequence of observed patterns. In [8], [9] and [10], K. Tanaka treated same problem as the present author, but restricted himself to the parametric case. In this paper, we shall consider the non-parametric case, and appeal to the method which has been developed in [12]. Consequently, our various conditions imposed are different from those in [8], [9] and [10].

This paper consists of five sections. In Section 2, we shall give the formulation of our problem and five assumptions necessary for subsequent arguments. In Section 3, we shall define a recursive algorithm for the pattern classification problem, which is an application of the dynamic stochastic approximation method [3], and investigate the convergence of it. The meaning of the convergence is "in the mean". In Section 4 we shall give two examples.

§ 2. The formulation of the problem and Assumptions.

We consider the two-categories classification problem. Let $\hat{X}^{(n)}$ and $\hat{\Theta}$ denote a pattern space at instant n and the set of categories, respectively. We assume that $\hat{\Theta}$ consists of two categories $\hat{\theta}_1$ and $\hat{\theta}_2$, i.e. $\hat{\Theta}=\{\hat{\theta}_1, \hat{\theta}_2\}$.

An outcome in pattern classification problem is described by a pair (x, θ), where x is an observed pattern in pattern space and θ specifies the category of the observed pattern x. Generally, θ is unknown to the observer. For a sequence of observed patterns $x_1, x_2, \cdots, x_n, \cdots$, we can consider a sequence:

$$(2.1) \qquad (x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n), \cdots$$

* Department of Mathematics, Kyushu University, Fukuoka.

with $x_n \in \hat{X}^{(n)}$ and $\theta_n \in \hat{\Theta}$, where $\theta_n = \hat{\theta}_i$ if x_n is a sample value from a specific category $\hat{\theta}_i$.

Let $X_1, X_2, \cdots, X_n, \cdots$ be a sequence of random variables, where X_n is a $\hat{X}^{(n)}$-valued random variable and can be from either of two classes $\hat{\theta}_1$ and $\hat{\theta}_2$. And let Θ_n be a $\hat{\Theta}$-valued random variable for each n, where $\Theta_n = \hat{\theta}_i$ if X_n is from the class $\hat{\theta}_i$. Thus $(x_1, \theta_1), \cdots, (x_n, \theta_n), \cdots$ are outcomes of $(X_1, \Theta_1), (X_2, \Theta_2) \cdots (X_n, \Theta_n), \cdots$.

In this paper, we shall assume $\hat{X}^{(n)} = R^N$ for all n, where R^N is N-dimensional Euclidean space. And we shall assume for each n, random variable X_n has a probability density function $p_i^{(n)}(\cdot)$ with respect to N-dimensional Lebesgue measure if X_n comes from the class $\hat{\theta}_i$. Furthermore, for each n, suppose that there exists a distribution $(q_1^{(n)}, q_2^{(n)})$ on $\hat{\Theta}$, that is, $q_1^{(n)} = \Pr\{\Theta_n = \hat{\theta}_1\}$ and $q_2^{(n)} = \Pr\{\Theta_n = \hat{\theta}_2\}$. Therefore, for each n, a pair of random variables (X_n, Θ_n) has a probability density function $p^{(n)}(x, \theta)$, $x \in R^N$ and $\theta \in \hat{\Theta}$, where

$$p^{(n)}(x, \hat{\theta}_i) = q_i^{(n)} \cdot p_i^{(n)}(x); \qquad x \in R^N \quad \text{and} \quad i = 1, 2, . \tag{2.2}$$

Let us assume tentatively that $p^{(n)}(\cdot, \hat{\theta}_1)$ and $p^{(n)}(\cdot, \hat{\theta}_2)$ are all known to us at instant n. Let us consider the discriminant function at instant n:

$$D^{(n)}(x) = p^{(n)}(x, \hat{\theta}_1) - p^{(n)}(x, \hat{\theta}_2), \qquad x \in R^N . \tag{2.3}$$

And let us consider the following decision rule based on (2.3); decide X_n to be from the class $\hat{\theta}_1$ if $D^{(n)}(x) \geqq 0$ for the outcome x of X_n, and decide X_n to be from the class $\hat{\theta}_2$ if $D^{(n)}(x) < 0$ for the outcome x of X_n. It is well known that this decision rule minimizes the probability of misclassification at instant n. This decision rule has been called the Bayes decision rule. In this paper, we shall call (2.3) the optimal discriminant function at instant n.

In this paper, we shall treat the case when $p^{(n)}(\cdot, \hat{\theta}_1)$ and $p^{(n)}(\cdot, \hat{\theta}_2)$ are unknown to us for each n, consequently the optimal discriminant function at instant n is unknown to us. In this situation we are supposed to have a training sequence $\{(X_n, \Theta_n)\}_{n=1}^{\infty}$ with the observed R^N-valued random vector X_n and $\hat{\Theta}$-valued random variable Θ_n. It is assumed that the knowldge of the Θ_n is obtained from a "teacher" who classifies (without error) the incoming X_n for each n.

The pattern classification problem considered here is to find a decision rule to classify the pattern X_n as correctly as possible into either of the two categories for a sufficently large n on the basis of a training sequence $\{(X_i, \Theta_i)\}_{i=1}^{n}$. It is reasonable, therefore, to consider a method of approximation to the limit of $D^{(n)}(\cdot)$, if it exist, by a using a training sequence.

For each n, we assume that (X_n, Θ_n) has a probability density function $p^{(n)}(x, \theta)$, $x \in R^N$ and $\theta \in \hat{\Theta}$, which is defined in (2.2). Furthermore, for each n and m, (X_{n+m}, Θ_{n+m}) has a conditional probability density function given the first m histry $(X_1, \Theta_1) = (x_1, \theta_1)$, $(X_2, \Theta_2) = (x_2, \theta_2), \cdots, (X_m, \Theta_m) = (x_m, \theta_m)$. We denote this transition probability density funcsion by

$$p^{(n+m)}(x, \theta \mid (x_1, \theta_1), \cdots, (x_m, \theta_m)) . \tag{2.4}$$

Next, we shall give assumptions.

ASSUMPTION 1. There exists a positive constant P_0 such that

$$\sup_{(x,\theta),n} p^{(n)}(x,\theta) \leqq P_0. \tag{2.5}$$

ASSUMPTION 2. We put for $n, m=1, 2, \cdots$,

$$H_{n,m} = \sup_{(x,\theta)} |p^{(n)}(x,\theta) - p^{(m)}(x,\theta)|. \tag{2.6}$$

Then

$$\lim_{n,m\to\infty} H_{n,m} = 0. \tag{2.7}$$

ASSUMPTION 3. We put for $n=0, 1, 2, \cdots$ and $m=1, 2, \cdots$,

$$G_{n,m} = \sup_{(x,\theta),(x_1,\theta_1),\cdots,(x_m,\theta_m)} |p^{(n+m)}(x,\theta|(x_1,\theta_1)\cdots(x_m,\theta_m)) - p^{(n+m)}(x,\theta)|. \tag{2.8}$$

Then

$$\lim_{n,m\to\infty} G_{n,m} = 0. \tag{2.9}$$

ASSUMPTION 4. For each n, $p^{(n)}(\cdot, \hat{\theta}_1)$ and $p^{(n)}(\cdot, \hat{\theta}_2)$ are uniformly continuous functions on R^N.

ASSUMPTION 5. For all n, $p^{(n)}(\cdot, \hat{\theta}_1)$ and $p^{(n)}(\cdot, \hat{\theta}_2)$ satisfy an uniform Lipschitz conditions with constants C_1 and C_2, respectively: the following inequality holds for all n,

$$|p^{(n)}(x,\hat{\theta}_i) - p^{(n)}(y,\hat{\theta}_i)| \leqq C_i \cdot \|x-y\| \quad \text{for all } n, \quad x, y \in R^N \text{ and } i=1, 2, \tag{2.10}$$

where $\|y\| = \left(\sum_{i=1}^{N} y_i^2\right)^{\frac{1}{2}}$ for $y=(y_1, \cdots, y_N) \in R^N$.

REMARK. For each n and m, putting

$$H'_{n,m} = \sup_{x} |D^{(n)}(x) - D^{(m)}(x)|, \tag{2.11}$$

$$\begin{aligned} D^{(n+m)}&(x|x_1,\theta_1),\cdots,(x_m,\theta_m)) \\ &= p^{(n+m)}(x,\hat{\theta}_1|(x_1,\theta_1),\cdots,(x_m,\theta_m)) - p^{(n+m)}(x,\hat{\theta}_2|(x_1,\theta_1),\cdots,(x_m,\theta_m)), \end{aligned} \tag{2.12}$$

and

$$G'_{n,m} = \sup_{\substack{x \\ (x_1,\theta_1),\cdots,(x_m,\theta_m)}} |D^{(n+m)}(x|(x_1,\theta_1),\cdots,(x_m,\theta_m)) - D^{(n+m)}(x)|, \tag{2.13}$$

we have

$$H'_{n,m} \leqq 2H_{n,m} \tag{2.14}$$

and

$$G'_{n,m} \leqq 2G_{n,m}. \tag{2.15}$$

Then it is easily seen that

$$\sup_{x,n} |D^{(n)}(x)| \leqq 2P_0, \tag{2.16}$$

$$\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} H_{j,n} = 0, \tag{2.17}$$

and

(2.18) $$\lim_{n\to\infty} n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} G_{i,j}=0$$

under Assumptions 1, 2 and 3, respectively.

And if Assumption 5 holds, then we have

(2.19) $$|D^{(n)}(x)-D^{(n)}(y)|\leqq C_0\cdot\|x-y\|\,, \qquad x, y\in R^N$$

and for all n, where $C_0=\max\{C_1, C_2\}$.

§ 3. Main results.

In this paper, all integrals are interpreted N-dimensional Lebesgue integral on R^N. We denote $\int_{R^N} g(x)d\mu(x)$ simply by $\int g(x)dx$, where μ is N-dimensional Lebesgue measure. And Fubini's theorem is invoked without any comments.

Let $K(\cdot)$ be a real-valued measurable function on R^N satisfying the following conditions:

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

(K2) $\sup_{y\in R^N} K(y)=K_0<\infty$,

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

(K4) $\sup_{y\in R^N} \|y\|^{1+\alpha}\cdot K(y)=K_1<\infty$ for some $\alpha>0$,

(K5) $\int\|y\|\cdot K(y)dy=K_2<\infty$.

Let $\{h_n\}_{n=1}^{\infty}$ be a sequence of positive numbers satisfying the following conditions:

(S1) $1\geqq h_1\geqq h_2\geqq \cdots \geqq h_n\geqq \cdots$,

(S2) $\lim_{n\to\infty} h_n=0$,

(S3) $\lim_{n\to\infty} nh_n^N=\infty$.

Then we can define the sequence $\{K_n(x, y)\}_{n=1}^{\infty}$ by

(3.1) $$K_n(x, y)=h_n^{-N}\cdot K[h_n^{-1}\cdot(x-y)]\,, \qquad x, y\in R^N\,.$$

The following lemma is a modification of Theorem 1A in [5].

LEMMA 1. *Let Assumption* 1, *Assumption* 2 *and Assumption* 4 *hold. Let* $K(\cdot)$ *be a real-valued measurable function on* R^N *satisfying* (K1), (K2), (K3) *and* (K4), *and let* $\{h_n\}_{n=1}^{\infty}$ *be a sequence of positive numbers satisfying* (S1) *and* (S2). *Define* $g_n(\cdot)$ *by*

(3.2) $$g_n(x)=\int K_n(x, y)D^{(n)}(y)dy \qquad for \quad n=1, 2, \cdots,$$

where $K_n(x, y)$ *is defined by* (3.1). *Then it holds that*

(3.3) $$\sup|g_n(x)-D^{(n)}(x)|\to 0 \qquad as \quad n\to\infty\,.$$

PROOF. Note first that

$$g_n(x)-D^{(n)}(x)=\int\{D^{(n)}(x-y)-D^{(n)}(x)\}\cdot\frac{1}{h_n^N}\cdot K\Big(\frac{y}{h_n}\Big)dy\,. \tag{3.4}$$

Let choose α' satisfying

$$\alpha\cdot(1+\alpha)^{-1}>\alpha'>0\,, \tag{3.5}$$

and split the region of integration into two regions, $\|y\|\leqq\delta_n$ and $\|y\|>\delta_n$, where $\delta_n=h_n^{\alpha'}$. Then

$$\begin{aligned}(3.6)\quad & |g_n(x)-D^{(n)}(x)|\\ &\leqq\sup_{\|y\|\leqq\delta_n}|D^{(n)}(x-y)-D^{(n)}(x)|\cdot\int K(z)dz\\ &\quad+\int_{\|y\|\geqq\delta_n}\frac{|D^{(n)}(x-y)|}{\|y\|}\cdot\frac{\|y\|}{h_n^N}\cdot K\Big(\frac{y}{h_n}\Big)dy+|D^{(n)}(x)|\cdot\int_{\|y\|\leqq\delta_n}\frac{1}{h_n^N}K\Big(\frac{y}{h_n}\Big)dy\\ &\leqq\sup_{\|y\|\leqq\delta_n}|D^{(n)}(x-y)-D^{(n)}(x)|\\ &\quad+\frac{1}{\delta_n}\sup_{\|z\|\geqq\delta_n/h_n}\|z\|\cdot K(z)\cdot\int|D^{(n)}(y)|dy+|D^{(n)}(x)|\cdot\int_{\|y\|\geqq\delta_n/h_n}K(y)dy\,.\end{aligned}$$

From (K4), we have

$$\sup_{\|z\|\geqq\delta_n/h_n}\|z\|\cdot K(z)\leqq K_1\cdot h_n^\alpha/\delta_n^\alpha\,, \tag{3.7}$$

and from the definition of the optimal discriminent function

$$\int|D^{(n)}(y)|dy\leqq q_n^{(1)}+q_2^{(n)}=1\qquad\text{for all}\quad n\,. \tag{3.8}$$

Therefore we have, from (3.6), (3.7), (3.8) and Assumption 1,

$$\begin{aligned}(3.9)\quad \sup_x|g_n(x)-D^{(n)}(x)|&\leqq\sup_{\|x-y\|\leqq\delta_n}|D^{(n)}(x)-D^{(n)}(y)|+K_1\cdot h_n^\alpha\cdot\delta_n^{-(1+\alpha)}\\ &\quad+2P_0\int_{\|y\|\geqq\delta_n\cdot h_n^{-1}}K(y)dy\,.\end{aligned}$$

Noting that

$$\begin{aligned}(3.10)\quad &\sup_{\|x-y\|\leqq\delta_n}|D^{(n)}(x)-D^{(m)}(x)|+\sup_{\|x-y\|\leqq\delta_n}|D^{(m)}(x)-D^{(m)}(y)|\\ &\qquad+\sup_{\|x-y\|\leqq\delta_n}|D^{(m)}(y)-D^{(n)}(y)|\\ &\leqq 2H'_{n,m}+\sup_{\|x-y\|\leqq\delta_n}|D^{(m)}(x)-D^{(m)}(y)|\end{aligned}$$

where m is a arbitrary integer. From Assumption 2, we can choose $\varepsilon>0$ and let N_0 be such that $n, m\geqq N_0$ implies $2H'_{n,m}<\varepsilon$. Let m_0 be some integer satisfying $m_0\geqq N_0$, then we have

$$\begin{aligned}(3.11)\quad &\sup_x|g_n(x)-D^{(n)}(x)|\\ &\leqq\varepsilon+\sup_{\|x-y\|\leqq\delta_n}|D^{(m_0)}(x)-D^{(m_0)}(y)|+K_1\cdot h_n^\alpha\cdot\delta_n^{-(1+\alpha)}\\ &\quad+2P_0\cdot\int_{\|y\|\geqq\delta_n\cdot h_n^{-1}}K(y)dy\qquad\text{for all}\quad n\geqq N_0\,.\end{aligned}$$

By (3.5), it is easily seen that

$$\delta_n = h_n^{\alpha'} \longrightarrow 0 \qquad \text{as} \quad n\to\infty, \tag{3.12}$$

$$h_n^{\alpha}/\delta_n^{1+\alpha} = h_n^{\alpha-\alpha'\cdot(1+\alpha)} \longrightarrow 0 \qquad \text{as} \quad n\to\infty, \tag{3.13}$$

and

$$\delta_n/h_n = h_n^{-(1-\alpha')} \longrightarrow \infty \qquad \text{as} \quad n\to\infty. \tag{3.14}$$

Therefore the right side in (3.11) tends to 0 by letting $n\to\infty$ at first and then $\varepsilon\to 0$. Thus the proof of the lemma is completed.

The following lemma was proved in [12] (Lemma 4).

LEMMA 2. *Let Assumption* 5 *holds. Let* $K(\cdot)$ *be a real-valued measurable function on* R^N *satisfying* (K1), (K3) *and* (K5), *and let* $\{h_n\}_{n=1}^{\infty}$ *be a sequence of positive numbers. Then it hold that*

$$\sup_{x\in R^N} |g_n(x)-D^{(n)}(x)| \leqq C_0\cdot K_2\cdot h_n \qquad \textit{for} \quad n=1, 2, \cdots \tag{3.15}$$

where $g_n(\cdot)$ *was defined in Lemma* 1.

Let $\{(X_n, \Theta_n)\}_{n=1}^{\infty}$ be a training sequence which was defined in § 2. Then we define $\{\rho^n\}_{n=1}^{\infty}$ by

$$\begin{aligned} \rho^n = \rho^n(\Theta_n) &= 1 \qquad \text{if} \quad \Theta_n = \hat{\theta}_1, \\ &= -1 \qquad \text{if} \quad \Theta_n = \hat{\theta}_2. \end{aligned} \tag{3.16}$$

for $n=1, 2, \cdots$.

From (3.1) and (3.16), we have

$$\mathrm{E}[\rho^n\cdot K_n(x, X_n)] = \int K_n(x, y) D^{(n)}(y) dy \qquad \text{for} \quad n=1, 2, \cdots. \tag{3.17}$$

In view of the above arguments, we now construct the following algorithm which is an application of the dynamic stochastic approximation method [3].

LEARING ALGORITHM. Let us define the following recursive algorithm:

$$D_0(x) = 0, \qquad x\in R^N, \tag{3.18}$$

$$D_{n+1}(x) = D_n(x) + (n+1)^{-1}\cdot(\rho^{n+1}\cdot K_{n+1}(x, X_{n+1}) - D_n(x)), \tag{3.19}$$

$x\in R^N$ and $n=0, 1, 2, \cdots$.

The above algorithm is equivalent to the following one;

$$D_n(x) = n^{-1}\cdot \sum_{j=1}^{n} \rho^j\cdot K_j(x, X_j), \qquad x\in R^N \quad \text{and} \quad n=1, 2, \cdots. \tag{3.20}$$

From the above algorithm we can obtain the following lemmas.

LEMMA 3. *Let Assumptions* 1, 2 *and* 4 *hold. Let* $K(\cdot)$ *be a real-valued measurable function on* R^N *satisfying* (K1), (K2), (K3) *and* (K4), *and let* $\{h_n\}_{n=1}^{\infty}$ *be a positive sequence satisfying* (S1) *and* (S2). *Then, it holds that*

$$\lim_{n\to\infty} \int (\mathrm{E}[D_n(x)] - D^{(n)}(x))^2 dx = 0. \tag{3.21}$$

PROOF. From (3.20) and (3.17), we have

$$(3.22)\qquad |\mathrm{E}[D_n(x)]-D^{(n)}(x)|\leqq n^{-1}\sum_{j=1}^{n}|g_j(x)-D^{(j)}(x)| + n^{-1}\sum_{j=1}^{n}|D^{(j)}(x)-D^{(n)}(x)|\,,$$

where $g_j(\cdot)$ was defined in (3.2). From Lemma 1, we have

$$(3.23)\qquad \sup_x |g_j(x)-D^{(j)}(x)| \longrightarrow 0 \quad \text{as} \quad j\to\infty\,.$$

And by Assumptions 1 and 2, we have

$$(3.24)\qquad \sup_x n^{-1}\sum_{j=1}^{n}|D^{(j)}(x)-D^{(n)}(x)|\leqq n^{-1}\sum_{j=1}^{n}H'_{j,n}\longrightarrow 0 \quad \text{as} \quad n\to\infty\,.$$

Note that

$$(3.25)\qquad \int(\mathrm{E}[D_n(x)]-D^{(n)}(x))^2dx \leqq \sup_x |\mathrm{E}[D_n(x)]-D^{(n)}(x)|\cdot\int\{|\mathrm{E}[D_n(x)]|+|D^{(n)}(x)|\}\,dx \leqq \sup_x |\mathrm{E}[D_n(x)]-D^{(n)}(x)| \quad \text{for all} \quad n\,.$$

Therefore, by (3.23), (3.24) and (3.25), we have (3.21). Thus the proof of the lemma is completed.

LEMMA 4. *Let Assumption* 3 *hold. Let $K(\cdot)$ be a real-valued measurable function on R^N satisfying* (K1) *and* (K3), *and let $\{h_n\}_{n=1}^{\infty}$ be a positive sequence. Then, it holds that*

$$(3.26)\qquad \int|\mathrm{E}[\rho^m\cdot K_m(x,X_m)\cdot\rho^{n+m}\cdot K_{n+m}(x,X_{n+m})] -\mathrm{E}[\rho^m\cdot K_m(x,X_m)]\cdot\mathrm{E}[\rho^{n+m}\cdot K_{n+m}(x,X_{n+m})]|\,dx \leqq 2G_{n,m}\,,\qquad n=0,1,2,\cdots \ \textit{and}\ \ m=1,2,\cdots\,.$$

PROOF. We put $U_n=\rho^n\cdot K_n(x,X_n)$ for $n=1,2,\cdots$, and $Y_m=((X_1,\Theta_1),\cdots,(X_m,\Theta_m))$ for $m=1,2,\cdots$. Then we have

$$(3.27)\qquad |\mathrm{E}[U_m\cdot U_{n+m}]-\mathrm{E}[U_m]\cdot\mathrm{E}[U_{n+m}]| =|\mathrm{E}[U_m\cdot\{\mathrm{E}[U_{n+m}|Y_m]-\mathrm{E}[U_{n+m}]\}]| \leqq \mathrm{E}[|U_m|\cdot\int K_{n+m}(x,y)\cdot|D^{(n+m)}(y|Y_m)-D^{(n+m)}(y)|\,dy]\,.$$

From Assumption 3 and (3.27)

$$(3.28)\qquad \int|\mathrm{E}[U_m\cdot U_{n+m}]-\mathrm{E}[U_m]\cdot\mathrm{E}[U_{n+m}]|\,dx\leqq 2G_{n,m}\cdot\mathrm{E}\Big[\int K_n(x,X_n)dx\Big]=2G_{n,m}\,.$$

Thus the proof of the lemma is completed.

From Lemma 4 we have the following lemma.

LEMMA 5. *Let the conditions in Lemma* 4 *be satisfied. Moreover, suppose that $K(\cdot)$ satisfies* (K2), *and $\{h_n\}_{n=1}^{\infty}$ satisfies* (S1) *and* (S3). *Then it holds that*

$$(3.29)\qquad \lim_{n\to\infty}\int\mathrm{E}(D_n(x)-\mathrm{E}[D_n(x)])^2dx=0\,.$$

PROOF. At first note that

$$\text{(3.30)} \qquad \begin{aligned} &\mathrm{E}[D_n(x)-\mathrm{E}[D_n(x)]]^2 \\ &=\mathrm{E}\Big[n^{-1}\cdot\sum_{j=1}^{n}(U_j-\mathrm{E}[U_j])\Big]^2 \\ &=n^{-2}\cdot\sum_{j=1}^{n}\mathrm{E}[U_j-\mathrm{E}[U_j]]^2+n^{-2}\sum_{j\neq i}^{n}\{\mathrm{E}[U_i\cdot U_j]-\mathrm{E}[U_i]\cdot\mathrm{E}[U_j]\}\,. \end{aligned}$$

From Lemma 4 we have

$$\text{(3.31)} \qquad \begin{aligned} &n^{-2}\int\sum_{i\neq j}^{n}\{\mathrm{E}[U_i,\,U_j]-\mathrm{E}[U_i]\cdot\mathrm{E}[U_j]\}\,dx \\ &\leqq 2\cdot n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int|\mathrm{E}[U_{i+j}\cdot U_j]-\mathrm{E}[U_{i+j}]\cdot\mathrm{E}[U_j]|\,dx \\ &\leqq 4n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}G_{i,j} \qquad \text{for all} \quad n\,. \end{aligned}$$

Next we have

$$\text{(3.32)} \qquad \begin{aligned} &n^{-2}\int\sum_{j=1}^{n}\mathrm{E}[U_j-\mathrm{E}[U_j]]^2dx \\ &\leqq n^{-2}\cdot\int\sum_{j=1}^{n}\mathrm{E}[K_j(x,\,X_j)]^2dx \\ &=n^{-2}\sum_{j=1}^{n}h_j^{-N}\cdot\int\Big\{\int h_j^{-N}\cdot K^2[h_j^{-1}(x-y)]\cdot p^{(j)}(y)dy\Big\}dx \qquad \text{for all} \quad n\,, \end{aligned}$$

where $p^{(j)}(y)=p^{(j)}(y,\,\hat{\theta}_1)+p^{(j)}(y,\,\hat{\theta}_2)$, $y\in R^N$ and $j=1, 2, \cdots$. By (K2) and (S1), we have

$$\text{(3.33)} \qquad n^{-2}\sum_{j=1}^{n}\int\mathrm{E}[U_j-\mathrm{E}[U_j]]^2dx\leqq K_1\cdot(n\cdot h_n^N)^{-1} \qquad \text{for} \quad n=1, 2, \cdots.$$

Therefore by (3.31), (3.33), Assumption 3 and (S3), we have

$$\int\mathrm{E}[D_n(x)-\mathrm{E}D_n(x)]^2dx\longrightarrow 0 \qquad \text{as} \quad n\to\infty\,.$$

Thus the proof of the lemma is completed.

From Lemma 3 and Lemma 5, we have the following theorem.

THEOREM 1. *Let Assumption* 1, 2, 3 *and* 4 *hold. Let* $K(\cdot)$ *be a real-valued measurable function on* R^N *satisfying* (K1), (K2), (K3) *and* (K4), *and let* $\{h_n\}_{n=1}^{\infty}$ *be a sequence of positive numbers satisfying* (S1), (S2) *and* (S3). *Then it holds that*

$$\text{(3.34)} \qquad \lim_{n\to\infty}\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big]=0\,.$$

PROOF. Noting that

$$\text{(3.35)} \qquad \begin{aligned} &\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big] \\ &=\int\mathrm{E}[D_n(x)-\mathrm{E}(D_n(x))^2]dx+\int(\mathrm{E}[D_n(x)]-D^{(n)}(x))^2dx\,, \end{aligned}$$

it is easily seen that (3.34) is direct from Lemmas 3 and 5.

Next, we shall give a theorem concerning the order of the convergence. The following lemma was given by the present author in [11].

LEMMA 6. *Let* $\{A_n\}_{n=1}^{\infty}$ *be a sequence of non-negative numbers. Suppose that there exist a position integer* n_0*, two sequences of positive numbers* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$*, and two positive constants* C *and* α_0 *such that*

$$A_{n+1} \leqq (1-a_{n+1})A_n + C\cdot a_{n+1}\cdot b_{n+1} \qquad \textit{for all} \quad n \geqq n_0\,, \tag{3.36}$$

$$\sum_{n=1}^{\infty} a_n = \infty \qquad \textit{and} \qquad \lim_{n\to\infty} a_n = 0\,, \tag{3.37}$$

$$0 < \alpha_0 < 1\,, \tag{3.38}$$

and

$$(1-a_{n+1})\cdot b_n/b_{n+1} \leqq (1-\alpha_0 a_{n+1}) \qquad \textit{for all} \quad n \geqq n_0\,. \tag{3.39}$$

Then, there exists a positive constant C_0 *such that*

$$A_n \leqq C_0\cdot b_n \qquad \textit{for all} \quad n\,. \tag{3.40}$$

REMARK. In Lemma 6, when $a_n = n^{-1}$ and $b_n = n^{-\beta}$ $(0<\beta<1)$ the conditions (3.37) and (3.39) are satisfied by α_0 such that $0<\alpha_0+\beta<1$. And when $A_n = n^{-1}\sum_{j=1}^{n} b_j$ the condition (3.36) is satisfied by $C=1$. Therefore, there exists a position constant C_0 such that

$$n^{-1}\sum_{j=1}^{n} b_j \leqq C_0\cdot b_n \qquad \text{for all} \quad n\,.$$

THEOREM 2. *Let Assumption* 1, 2, 3 *and* 5 *hold. Let* $K(\cdot)$*, be a real-valued measurable function satisfying* (K1), (K2), (K3) *and* (K5)*, and let* $\{h_n\}_{n=1}^{\infty}$ *be a sequence of positive numbers satisfying* (S1), (S2) *and* (S3)*. Moreover, let the following conditions be satisfied: there exist positive constant* H, G *and* α_0*, and positive integer* n_0 *such that*

$$n\cdot H_{n-1,n} \leqq H\cdot b_n \qquad \textit{for} \quad n=1, 2, \cdots, \tag{3.41}$$

$$\sup_m G_{n,m} \leqq G\cdot b_n \qquad \textit{for} \quad n=1, 2, \cdots, \tag{3.42}$$

$$0<\alpha_0<1\,, \tag{3.43}$$

and

$$b_n/b_{n+1} \leqq 1+\alpha_0 n^{-1} \qquad \textit{for all} \quad n \geqq n_0 \tag{3.44}$$

where

$$b_n = \max\{h_n, (n\cdot h_n^N)^{-1}\} \qquad \textit{for} \quad n=1, 2, \cdots. \tag{3.45}$$

Then, there exists a positive constant C^* *such that*

$$\mathrm{E}\Big[\int (D_n(x)-D^{(n)}(x))^2 dx\Big] \leqq C^*\cdot b_n \qquad \textit{for all} \quad n\,. \tag{3.46}$$

PROOF. From (3.22) and (3.25) in the proof of Lemma 3, and from Lemma 2 we have

(3.47) $$\int(\mathrm{E}[D_n(x)]-D^{(n)}(x))^2dx \leqq C_0\cdot K_2\cdot\sum_{j=1}^{n}h_j+n^{-1}\sum_{j=1}^{n}H'_{j,n} \quad \text{for} \quad n=1,2,\cdots.$$

From (3.31) and (3.33) in the proof of Lemma 5 we have

(3.48) $$\int \mathrm{E}(D_n(x)-\mathrm{E}[D_n(x)])^2dx \leqq 4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}G_{i,j}+K_1\cdot(nh_n^N)^{-1} \quad \text{for} \quad n=1,2,\cdots.$$

Therefore,

(3.49) $$\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big] \leqq C_0\cdot K_2\cdot\sum_{j=1}^{n}h_j+2n^{-1}\cdot\sum_{j=1}^{n}H'_{j,n}+4n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}G_{i,j}+K_1\cdot(nh_n^N)^{-1} \quad \text{for} \quad n=1,2,\cdots.$$

Noting that

(3.50) $$H'_{j,n}\leqq 2(H_{j,j+1}+H_{j+1,j+2}+\cdots+H_{n-1,n}),$$

by (3.41) we have

(3.51) $$n^{-1}\cdot\sum_{j=1}^{n}H'_{j,n}\leqq 2H\cdot n^{-1}\sum_{j=1}^{n-1}\Big(\frac{b_{j+1}}{j+1}+\cdots+\frac{b_n}{n}\Big) \leqq 2H\cdot n^{-1}\sum_{j=1}^{n}b_j \quad \text{for} \quad n=1,2,\cdots.$$

By (3.42)

(3.52) $$n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}G_{i,j}\leqq G\cdot n^{-1}\sum_{j=1}^{n}b_j \quad \text{for} \quad n=1,2,\cdots.$$

There by (3.45), (3.49), (3.51) and (3.52), we have

(3.53) $$\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big]\leqq\tilde{C}\cdot\Big(n^{-1}\sum_{j=1}^{n}b_j+b_n\Big) \quad \text{for} \quad n=1,2,\cdots,$$

where $\tilde{C}=\max\{C_0\cdot K_2, 4H, 4G, K_1\}$. And by (3.44) we have

$$(1-(n+1)^{-1})b_n/b_{n+1}\leqq(1-\alpha'_0\cdot(n+1)^{-1}) \quad \text{for all} \quad n\geqq n_0,$$

where $\alpha'_0=1-\alpha_0$. Therefore from Lemma 6 and its remark, there exists a positive constant C' such that

(3.54) $$n^{-1}\sum_{j=1}^{n}b_j\leqq C'\cdot b_n \quad \text{for all} \quad n.$$

By (3.53) and (3.54), we have

(3.55) $$\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big]\leqq C^*\cdot b_n \quad \text{for all} \quad n,$$

where $C^*=\max\{C_0\cdot K_2\cdot C', 4H\cdot C', K_1\}$. Thus the proof of the theorem is completed.

REMARK. When $h_n=n^{-\beta}$ $(0<\beta<N^{-1})$, we can remove the conditions (3.43) and (3.44). In this case, we have

$$\mathrm{E}\Big[\int(D_n(x)-D^{(n)}(x))^2dx\Big]\leqq C^*\cdot n^{-\beta_0} \quad \text{where} \quad \beta_0=\min\{\beta, 1-N\beta\}.$$

Let $P_d^{(n)}(e)$ be the probability of misclassification by using a discriminant function $d(\cdot)$ at instant n, and let $P^{(n)}(e)$ be the probability of misclassification by using the optimal discriminant function $D^{(n)}(\cdot)$ at instant n.

We put $p^{(n)}(x)=p^{(n)}(x, \hat{\theta}_1)+p^{(n)}(x, \hat{\theta}_2)$, $x\in R^N$. Choose $\varepsilon>0$. Let $\{S_\varepsilon^{(n)}\}_{n=1}^\infty$ be a sequence of bounded sets in R^N such that for each n, the following inequality is satisfied

$$\int_{S_\varepsilon^{(n)}} p^{(n)}(x)dx \geqq 1-\varepsilon . \tag{3.56}$$

THEOREM 3. *Let the hypotheses in Theorem* 1 *hold. In addition, let $\varepsilon>0$ be given then there exists a positive constant M_ε such that*

$$\mu(S_\varepsilon^{(n)}) \leqq M_\varepsilon \qquad \textit{for all} \quad n , \tag{3.57}$$

where μ is N-dimensional Lebesgue measure on R^N and $\{S_\varepsilon^{(n)}\}_{n=1}^\infty$ is sequence of bounded sets in R^N such that for each n, (3.56) *is satisfied. Then it holds that*

$$\lim_{n\to\infty} \mathrm{E}\,|P_{D_n}^{(n)}(e)-P^{(n)}(e)|=0 .$$

PROOF. For each n, define sets

$$H^{(n)}=\{x\in R^N ;\ D^{(n)}(x)\geqq 0\} \tag{3.58}$$

and

$$H_n=\{x\in R^N ;\ D_n(x)\geqq 0\} . \tag{3.59}$$

And for a set A in R^N denote the complement of A by $\tilde{A}$, and by I_A the indicator of A.

Now, we have

$$\begin{aligned} p^{(n)}(e) &= \int_{\tilde{H}^{(n)}} P^{(n)}(x, \hat{\theta}_1)dx+\int_{H^{(n)}} p^{(n)}(x, \hat{\theta}_2)dx \\ &= q_1^{(n)}+\int[-D^{(n)}(x)]\mathrm{I}_{H^{(n)}}(x)dx . \end{aligned} \tag{3.60}$$

and

$$P_{D_n}^{(n)}(e)=q_1^{(n)}+\int[-D^{(n)}(x)]\mathrm{I}_{H_n}(x)dx . \tag{3.61}$$

Therefore

$$\begin{aligned} P_{D_n}^{(n)}(e)-P^{(n)}(e) &= \int D^{(n)}(x)[\mathrm{I}_{H^{(n)}}(x)-\mathrm{I}_{H_n}(x)]\cdot \mathrm{I}_{S_\varepsilon^{(n)}}(x)dx \\ &\quad +\int D^{(n)}(x)[\mathrm{I}_{H^{(n)}}(x)-\mathrm{I}_{H_n}(x)]\cdot \mathrm{I}_{\tilde{S}_\varepsilon^{(n)}}(x)dx . \end{aligned} \tag{3.62}$$

Noting that

$$\int[-D_n(x)]\cdot[\mathrm{I}_{H^{(n)}}(x)-\mathrm{I}_{H_n}(x)]\cdot \mathrm{I}_{S_\varepsilon^{(n)}}(x)dx \geqq 0 \tag{3.63}$$

for each n, and $0\leqq P_{D_n}^{(n)}(e)-P^{(n)}(e)$ for each n, we have

$$\begin{aligned} 0 &\leqq P_{D_n}^{(n)}(e)-P^{(n)}(e) \\ &\leqq \int[D^{(n)}(x)-D_n(x)]\cdot[\mathrm{I}_{H^{(n)}}(x)-\mathrm{I}_{H_n}(x)]\,\mathrm{I}_{S_\varepsilon^{(n)}}(x)dx \end{aligned} \tag{3.64}$$

$$+\int D^{(n)}(x)\cdot[\mathrm{I}_{H^{(n)}}(x)-\mathrm{I}_{H_n}(x)]\,\mathrm{I}_{\tilde{S}_\varepsilon^{(n)}}(x)dx$$

$$+\int|D^{(n)}(x)-D_n(x)|\cdot\mathrm{I}_{S_\varepsilon^{(n)}}(x)dx+\int_{\tilde{S}_\varepsilon^{(n)}}p^{(n)}(x)dx\,.$$

By (3.64)

$$\text{(3.65)}\qquad \mathrm{E}|P_{D_n}^{(n)}(e)-P^{(n)}(e)|\leqq\mathrm{E}\Big[\int(D^{(n)}(x)-D_n(x))^2dx\cdot\mu(S_\varepsilon^{(n)})\Big]^{\frac{1}{2}}+\frac{\varepsilon}{2}$$

$$\leqq\mathrm{E}\Big[\int(D^{(n)}(x)-D_n(x))^2dx\Big]^{\frac{1}{2}}\cdot M_\varepsilon^{\frac{1}{2}}+\frac{\varepsilon}{2}\,.$$

Thus, the theorem is proved.

The following theorem can be obtained by modified arguments of the proof in Theorem 3.

THEOREM 4. *Let the hypotheses of Theorem 2 hold. In addition, let, for each* n, $p^{(n)}(\cdot,\hat{\theta}_1)$ *and* $p^{(n)}(\cdot,\hat{\theta}_2)$ *have bounded supports. We put*

$$\text{(3.67)}\qquad S_i^{(n)}=\{x\in R^N\,;\,p^{(n)}(x,\hat{\theta}_i)\neq 0\}\qquad i=1,2,\ \text{and for}\ \ n=1,2,\cdots.$$

If there exist a positive constant M *such that*

$$\text{(3.68)}\qquad \mu(S_i^{(n)})\leqq M\qquad \text{for all}\ \ i\ \text{and}\ n\,,$$

then it holds that

$$\text{(3.69)}\qquad \mathrm{E}|P_{D_n}^{(n)}(e)-P^{(n)}(e)|\leqq(C^*\cdot b_n)^{\frac{1}{2}}\cdot M^{\frac{1}{2}}\qquad \text{for all}\ \ n\,,$$

where b_n *was defined in Theorem 2 and* C^* *is a positive constant which satisfies* (3.46).

§ 4. Examples.

EXAMPLE 1. Let $\{(X_n,\Theta_n)\}_{n=1}^{\infty}$ be a strictly stationary Markov process satisfying the condition D_0 (See [2], p. 221), and let $p(x,\theta)$ and $p^{(n)}(x,\theta\,|\,(x',\theta'))$ be a initial probability density function and a n-step transition probability density function, respectively. From [2].

$$\text{(4.1)}\qquad \sup_{\substack{(x,\theta)\\(x',\theta')}}|p^{(n)}(x,\theta\,|\,(x',\theta'))-p(x,\theta)|\leqq 2\nu\cdot\lambda^n\qquad \text{for all}\ \ n=1,2,\cdots,$$

where $0<\nu$ and $0<\lambda<1$.

If $p(x,\hat{\theta}_1)$ and $p(x,\hat{\theta}_2)$ are uniformly continuous functions on R^N, then Assumptions 1, 2, 3 and 4 hold. And it is easily seen that the hypotheses of Theorem 3 are satisfied. Thus, the results of Theorem 1 and Theorem 3 hold.

If $p(x,\hat{\theta}_1)$ and $p(x,\hat{\theta}_2)$ satisfy the uniform Lipschitz conditions:

$$\text{(4.2)}\qquad |p(x,\hat{\theta}_1)-p(y,\hat{\theta}_i)|\leqq C_i\cdot\|x-y\|\,,\qquad i=1,2,\cdots$$

and let $h_n=n^{-1/N+1}$, then Assumptions 1, 2, 3 and 5 hold, and there exists positive constant G such that

$$\text{(4.3)}\qquad \lambda^n\leqq G\cdot n^{-1/N+1}\qquad \text{for all}\ \ n=1,2,\cdots,$$

and $b_n=\max\{h_n,\,n^{-1}h_n^{-N}\}=h_n=n^{-1/N+1}$. Therefore, Theorem 2 hold. In addition, if $p(\cdot,\hat{\theta}_1)$ and $p(\cdot,\hat{\theta}_2)$ have bounded supports in R^N, then Theorem 4 also hold.

EXAMPLE 2. Let the following conditions be satisfied:

$$p^{(n+m)}(x, \hat{\theta}_i \mid (x_1, \theta_1), \cdots, (x_m, \theta_m)) = q^{(n+m)}(\hat{\theta}_i \mid \theta_1, \cdots, \theta_m)\cdot p_i(x) \qquad i=1,2 \text{ and } n, m=1, 2, \cdots, \tag{4.4}$$

$$\sum_{i=1}^{2} q^{(n+m)}(\hat{\theta}_i \mid \theta_1, \cdots, \theta_m)=1 \qquad \text{for } n, m=1, 2, \cdots, \tag{4.5}$$

that is, $q^{(n+m)}(\hat{\theta}_i \mid \theta_1, \cdots, \theta_m)$ is a conditional probability of a category $\hat{\theta}_i$ at instant $n+m$ for given categories $\theta_1, \theta_2, \cdots, \theta_m$,

$$\sup_{\substack{i\\ \theta_1,\cdots,\theta_m}} |q^{(n+m)}(\hat{\theta}_i \mid \theta_1, \theta_2, \cdots, \theta_m)-q_i^{(n+m)}| \longrightarrow 0 \qquad \text{as } n, m\to\infty. \tag{4.6}$$

And

$$|q_i^{(n)}-q_i^{(m)}| \longrightarrow 0 \qquad \text{as } n, m\to\infty \tag{4.7}$$

for $i=1, 2$.

If $p_1(\cdot)$ and $p_2(\cdot)$ are uniformly continuous functions on R^N, then the hypothese of Theorem 1 and Theorem 3 hold. Therefore, the results of Theorem 1 and Theorem 3 hold.

If $p_1(\cdot)$ and $p_2(\cdot)$ satisfy uniform Lipschitz conditions,

$$\sup_{\substack{i\\ \theta_1,\cdots,\theta_m}} |q^{(n+m)}(\hat{\theta}_i \mid \theta_1, \cdots, \theta_m)-q_i^{(n+m)}| \leqq G'\cdot h_m \qquad \text{for all } m, \tag{4.8}$$

where G' is some positive constant and

$$|q_i^{(n)}-q_i^{(n+1)}| \leqq n^{-1}\cdot h_n \qquad \text{for } n=1, 2, \cdots, \tag{4.9}$$

then Theorem 2 hold. In addition, if $p_1(\cdot)$ and $p_2(\cdot)$ have bounded supports in R^N, then Theorem 4 also hold.

§ 5. Acknowledgement.

The author is deeply indebted to Professor T. Kitagawa of Kyushu University, Professor N. Furukawa of Kyushu University and Professor K. Tanaka of Niigata University for their helpful advices and encouragements.

References

[1] E. M. BRAVERMAN and E. S. PYATNITSKI, *Estimation of the rate of convergeace of algorithm based on the potential function method*, Auto. and Remote Control, **27**, 1, (1966), 80-100.

[2] J. DOOB, *Stochastic processes*, Wiley (1953), New York.

[3] V. DUPAC, *A dynamic stochastic approximation method*, Ann. Math. Stat., **36** (1965), 1695-1702.

[4] N. V. LOGINOV, *The method of stochastic approximation*, Auto. and Remote Control, **27**, 4 (1966), 706-728.

[5] E. PARZEN, *On the estimation of a probability density function and mode*, Ann. Math. Stat., **33** (1962), 1065-1076.

[6] G. G. ROUSSAS, *Nonparametric estimation of the transition distribution function of a Markov process*, Ann. Math. Stat., **40**, 4 (1969), 1386-1400.

[7] J. V. RYZIN, *A stochastic a posteriori updating algorithm for pattern recognition*, J. Math. Anal. Appl., **20**, 2 (1967), 359-379.

[8] K. TANAKA, *On the pattern classification problem by learning (I)*, Bull. Math. Stat., **14**, 1-2, (1970), 31-49.

[9] K. TANAKA, *On the pattern classification problem by learning (III)*, Mem. of Faculty of Science, Kyushu Univ., series A, Math., **XXIV**, 2, (1972), 249-273.

[10] K. TANAKA, *On the pattern classification by learning*, Bull. Math. Stat., **14**, 3-4 (1971), 13-25.

[11] M. WATANABE, *On asymptotically optimal algorithms for pattern classification problems*, Bull. Math. Stat., **15**, 3-4 (1973), 31-48.

[12] C. T. WOLVERTON and T. J. WAGNER, *Asymptotically optimal discriminant function for pattern classification*, IEEE Trans. on Informatioh Theory, **IT-15**, 2, (1969), 258-265.