Learning with Mislabeled Training Samples Using Stochastic Approximation

AMITA PATHAK-PAL AND SANKAR K. PAL,
SENIOR MEMBER, IEEE

Abstract —**For the problem of parameter learning in pattern recognition, the convergence of stochastic approximation-based learning algorithms have been investigated for the situation in which mislabeled training samples are present. In the cases considered, it is found that estimates converge to nontrue values in the presence of labeling errors. The general m-class N-feature pattern recognition problem is considered. A possible solution to the problem is also discussed. Some simulation results are provided to support the conclusions drawn.**

I. Introduction

The learning of unknown parameters of classifiers is an indispensible part of pattern recognition problems. If a sufficiently large set of correctly labeled training samples is available, then "reasonably good" estimates of the parameters can generally be obtained. In many real-life situations, however, it is either difficult or expensive to obtain labels, so that mislabeling of training samples can become one of the specters with which a pattern recognition scientist has to contend. It is, therefore, useful to know how this problem can affect the learning procedure. A reasonable amount of work has been done for the two-class classification problem. The effects of random training errors on Fisher's discriminant function have been studied by Lachenbruch [1], [2]. McLachlan [3], Michalek and Tripathi [4], O'Neill [5], Krishnan [6], and Katre and Krishnan [7]. They concluded that the effect is to underestimate distance, overestimate error rate, introduce bias into estimates of the discriminant function, make the maximum likelihood estimates of the discriminant function converge to nontrue values, and change the asymptotic relative efficiency (ARE) relative to a completely correctly classified sample of the same size.

In the context of recursive learning of parameters, the usefulness of stochastic approximation procedures cannot be overemphasized [8].[1] Briefly, a stochastic approximation procedure for recursively estimating a parameter $\hat{\theta}$ by $\hat{\theta}_n$ (at the nth stage) with the help of an unbiased statistic T is

$$\hat{\theta}_{n+1} = \hat{\theta}_n - a_n\left(\hat{\theta}_n - T_{n+1}\right)$$

where $\hat{\theta}_1$ is either a constant or $\hat{\theta}_1 = T_1$, and $\{a_n\}$ is a suitably chosen sequence of positive numbers. For instance, a recursive procedure for estimating the population mean μ of a variable X utilizing the sample mean $\bar{x}$, is

$$\bar{x}_{n+1} = \bar{x}_n - \frac{1}{n}\left(\bar{x}_n - X_{n+1}\right),$$

X_{n+1} being the $(n+1)$th observation on X.

In this correspondence, the particular case in which errors occur in the labeling of training samples is studied for an m-class N-feature pattern recognition problem. The effect of mislabeling is to cause "wrong" samples to be used in the recursive learning of the estimates, for any given class. A simple but realistic model [14] is adopted to describe this sort of situation. Under this model, the authors have investigated the convergence of recursive learning procedures of the type mentioned above. It is found that, under certain conditions, these estimates do converge strongly, that is, with probability one, but to nontrue values, more specifically, to convex linear combinations of true parameters of all m classes. This conclusion is reached using some results on multidimensional stochastic approximation [15].

This result, in itself, is not surprising, because the presence of mislabeled samples in the training set is sure to affect the behavior of the training process in some way. This work merely provides a mathematical description of the effect on its convergence.

As this work will seem incomplete without a solution to the problem considered, we have also discussed in Section V a possible way of countering the effect of the presence of mislabeled samples in the training set. The solution consists of modifying the stochastic approximation procedure in such a fashion that it becomes restrictive, that is, it does not allow all training samples to be used for updating. At any given step in the training process, a sample is used for updating only if it is closer to the preceding estimate of the mean value than some specified threshold. Otherwise, it is excluded from the training set. Some results on the asymptotic behavior of such algorithms are stated. It is found that under certain conditions these algorithms are indeed better than the ones considered earlier. Some simulation results are provided to illustrate the conclusions arrived at in this work.

II. Statement of the Problem

Let us consider a general m-class $(C_i,\ i=1,\cdots,m)$ pattern recognition problem for which an N-dimensional vector

$$X_{N\times 1} = [X_1, X_2, \cdots, X_N]', X \in R^N$$

has been specified. Let us assume that

A1) the distribution of X in each class is continuous;

A2) the probability densities $p(\cdot|C_i)$ of X for the classes C_i, $i=1,\cdots,m$, are of the same family, and they differ only in respect of values parameters;

A3) an unbiased statistic exists for the q-dimensional parameter-vector $\varphi_{q\times 1}$ with respect to the probability density function p.

Let us suppose that for the purpose of learning we have been given a set of independent samples $X_1^{(k)}, X_2^{(k)}, \cdots, X_{n_k}^{(k)}$, $k = 1,\cdots,m$, where the superscripts k denotes the labels given to the respective samples. For the learning itself, let us utilize a stochastic approximation algorithm as defined below.

Let $\hat{\varphi}_t^{(k)}$ denote the estimate obtained at the tth step for the class C_k. Then

$$\hat{\varphi}_1^{(k)} = f\left(X_1^{(k)}\right) \tag{1a}$$

and for $t>1$,

$$\hat{\varphi}_{t+1}^{(k)} = \hat{\varphi}_t^{(k)} - a_t\left(\hat{\varphi}_t^{(k)} - f\left(X_{t+1}^{(k)}\right)\right), \qquad k=1,\cdots,m \tag{1b}$$

where $\{a_t\}$ is a sequence of positive real numbers such that $a_t \leqslant 1 \forall t$ and f: $R^N \to R^q$ is an unbiased statistic for φ. This algorithm is a generalization of the usual stochastic approximation procedures used for recursive parameter estimation.

III. A Model for Labeling Errors

The model to be used for this purpose was developed by Chittineni [14]. It can be specified as follows. Let w and $\hat{w}$ denote, respectively, the true and the given labels. Clearly,

$$w, \hat{w} \in \{1, 2, \cdots, m\}.$$

Manuscript received July 12, 1986; revised July 15, 1987.

A. Pathak-Pal is with the Electronics and Communications Sciences Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India.

At present, S. K. Pal is with the Centre for Automation Research, University of Maryland, College Park, MD 20742, on leave from the Indian Statistical Institute, Calcutta 700035, India.

IEEE Log Number 8716929.

[1] For instance, there are number of works [9]–[13] by Fu and others in which stochastic approximation techniques, as applied to learning in pattern recognition systems, are discussed. (It may be added, however, that these are not related to the present investigation.)

0018-9472/87/1100-1072$01.00 ©1987 IEEE

Let $\pi_i = P(w = i)$ denote the *a priori* probability for the class C_i, $i = 1, \cdots, m$. Further, let $p_i(X) = p(X|w = i)$ be the class-conditional density of the feature vector X for C_i. Also, let α_{ij} denote the probability that a sample from C_j has been given the label i, i.e.,

$$\alpha_{ij} = P(\hat{w} = i|w = j), \qquad i, j = 1, \cdots, m. \tag{2}$$

Clearly, we must have

$$\sum_{i=1}^{m} \alpha_{ij} = 1, \tag{3a}$$

i.e.,

$$A'_{m \times m} e_{m \times 1} = e_{m \times 1} \tag{3b}$$

where

$$e_{m \times 1} = [1 \quad 1 \quad 1 \cdots 1]'$$

and

$$A = ((\alpha_{ij})).$$

Now,

$$\begin{aligned} p(X|\hat{w} = i) &= \frac{p(X, \hat{w} = i)}{P(\hat{w} = i)} \\ &= \frac{1}{P(\hat{w} = i)} \sum_{j=1}^{m} p(X, \hat{w} = i, w = j) \\ &= \frac{1}{P(\hat{w} = i)} \sum_{j=1}^{m} p(X|\hat{w} = i, w = j) \\ &\quad \cdot P(\hat{w} = i|w = j) P(w = j) \\ &= \frac{1}{P(\hat{w} = i)} \sum_{j=1}^{m} \eta_j \alpha_{ij} p(X|\hat{w} = i, w = j). \end{aligned} \tag{4}$$

However,

$$\begin{aligned} P(\hat{w} = i) &= \sum_{j=1}^{m} P(\hat{w} = i, w = j) \\ &= \sum_{j=1}^{m} P(\hat{w} = i|w = j) P(w = j) \\ &= \sum_{j=1}^{m} \eta_j \alpha_{ij}. \end{aligned} \tag{5}$$

Hence (4) becomes

$$p(X|\hat{w} = i) = \sum_{j=1}^{m} \epsilon_{ij} p(X|\hat{w} = i, w = j) \tag{6}$$

where

$$\epsilon_{ij} = \pi_j \alpha_{ij} \bigg/ \left(\sum_{l=1}^{m} \eta_l \alpha_{il} \right). \tag{7}$$

If we are prepared to assume

A4) $p(X|w = j) = p(X|\hat{w} = i, w = j) \forall i, j$,

then (6) becomes

$$\begin{aligned} p(X|\hat{w} = i) &= \sum_{j=1}^{m} \epsilon_{ij} p(X|w = j) \\ &= \sum_{j=1}^{m} \epsilon_{ij} p_j(X). \end{aligned} \tag{8}$$

It may not be out of place to emphasize here that assumption A4) is perfectly reasonable in the sense that it merely requires that the distribution of X in any class depend not on the given lable $\hat{w}$, but only on the true label w.

IV. Convergence of the Learning Algorithm

For studying the asymptotic behavior of the learning algorithm given in Section II, use will be made of the following results, due to Schmetterer [15].

Lemma 1: Let $\{a_n\}$ be a sequence of positive real numbers such that

C1) $\sum_{n=1}^{\infty} a_n^2 < \infty$.

Let x_n and y_n be k-dimensional random vectors that satisfy

C2) $x_{n+1} = x_n - a_n y_n$, $n > 1$.

Let M_n be a measurable mapping from R^k to R^k such that

C3) $E(y_n|x_1, x_2, \cdots, x_n) = M_n(x_n)$ a.e.

Let a, b, c be nonnegative real numbers, and let

C4) $E(\|y_n\|^2|x_1, x_2, \cdots, x_n) \leqslant a + b\|x_n\| + c\|x_n\|^2$ a.e.

Also, for every $x \in R^k$ and $n \geqslant 1$,

C5) $x' M_n(x) \geqslant 0$.

If x_1 is chosen in such a way that

C6) $E(\|x_1\|^2)$ exists,

then the sequence $\{x_n\}$ converges with probability 1, i.e., almost surely and the sequence $E\{\|x_n\|^2\}$ converges also.

Lemma 2: Suppose that conditions C1)–C6) hold. If, further, there exists for every $\eta > 0$ a $\partial > 0$ such that for $n \geqslant 1$

C7) $\inf\limits_{\eta < \|x\| < \eta^{-1}} x' M_n(x) > \partial$,

then x_n converges almost surely to the k-dimensional null vector 0.

Let us now prove the following.

Proposition 1: Consider the setup given in Sections II and III. If, in addition to assumptions A1)–A4), we also have

A5) $\sum_{n=1}^{\infty} a_n^2 < \infty$,
A6) $\rho_i = E(\|f(X)\|^2|w = i)$ exists,

with respect to each class-conditional density $p_i(X)$, then

$$\hat{\varphi}_i^{(k)} \xrightarrow{\text{a.s.}} \sum_{j=1}^{m} \epsilon_{kj} \varphi_j.$$

Also,

$$\left\{ E\| \hat{\varphi}_i^{(k)} - \bar{\varphi}_k \|^2 \right\}$$

converges as $t \to \infty$, where

$$\bar{\varphi}_k = \sum_{j=1}^{m} \epsilon_{kj} \varphi_j.$$

ϵ_{kj}, $k, j = 1, \cdots, m$, is as in (7).

Proof of Proposition: The validity of the proposition can be inferred directly from Lemmas 1 and 2, provided one can show that the conditions C1)–C7) hold for $\psi_i^{(k)}$ where

$$\psi_i^{(k)} = \hat{\varphi}_i^{(k)} - \sum_{j=1}^{m} \epsilon_{kj} \varphi_j.$$

We note that from (1a) and (1b) we have, for $k = 1, \cdots, m$,

$$\psi_{t+1}^{(k)} = \begin{cases} g_k\left(X_1^{(k)}\right), & \text{for } t = 1 \quad (9a) \\ \psi_t^{(k)} - a_t\left(\psi_t^{(k)} - g_k\left(X_{t+1}^{(k)}\right)\right. & \quad (9b) \end{cases}$$

where

$$g_k(X) = f(X) - \sum_{j=1}^{m} \epsilon_{kj}\varphi_j.$$

Condition C1) is seen to be true because of A5). Condition C2) is equivalent to (9b).

Condition C3) also holds, with $M_t(x) = x$, as

$$E\left[\psi_t^{(k)} - g_k\left(X_{t+1}^{(k)}\right)|\psi_1^{(k)}, \psi_2^{(k)}, \cdots, \psi_t^{(k)}\right]$$
$$= \psi_t^{(k)} - E\left[g_k\left(X_{t+1}^{(k)}\right)\right],$$

since $X_{t+1}^{(k)}$ is independent of $X_1^{(k)}, \cdots, X_t^{(k)}$ and hence of $\psi_1^{(k)}, \psi_2^{(k)}, \cdots, \psi_t^{(k)}$.

$$\begin{aligned} E&\left[\psi_t^{(k)} - g_k\left(X_{t+1}^{(k)}\right)|\psi_1^{(k)}, \psi_2^{(k)}, \cdots, \psi_t^{(k)}\right] \\ &= \psi_t^{(k)} - E\left[g_k(X)|\hat{w} = k\right] \\ &= \psi_t^{(k)} - E(f(X)|\hat{w} = k) + \sum_{j=1}^{m} \epsilon_{kj}\varphi_j \\ &= \psi_t^{(k)} \end{aligned}$$

as, by (8),

$$E(f(X)|\hat{w} = k) = \sum_{j=1}^{m} \epsilon_{kj}\varphi_j = \bar{\varphi}_k,$$

say. Similarly, we have

$$\begin{aligned} E&\left(\left\|\psi_t^{(k)} - g_k\left(X_{t+1}^{(k)}\right)\right\|^2|\psi_1^{(k)}, \psi_2^{(k)}, \cdots, \psi_t^{(k)}\right) \\ &= E\left(\left\|\psi_t^{(k)}\right\|^2 - 2\psi_t^{(k)\prime} g_k\left(X_{t+1}^{(k)}\right)\right. \\ &\qquad \left. + \left\|g_k\left(X_{t+1}^{(k)}\right)\right\|^2|\psi_1^{(k)}, \psi_1^{(k)}, \cdots, \psi_t^{(k)}\right) \\ &= \left\|\psi^{(k)}\right\|^2 - 2\psi_t^{(k)\prime}E(g_k(X)|\hat{w} = k) + E\left(\left\|g_k(X)\right\|^2|\hat{w} = k\right) \end{aligned}$$

for the same reason as before;

$$= \left\|\psi_t^{(k)}\right\|^2 + E\left(\left\|f(X) - \bar{\varphi}_k\right\|^2|\hat{w} = k\right)$$

since $E(g_k(X)|\hat{w} = k) = 0$;

$$\begin{aligned} &= \left\|\psi_t^{(k)}\right\|^2 + \|\bar{\varphi}_k\|^2 + E\left(\left\|f(X)\right\|^2|\hat{w} = k\right) \\ &\leqslant \left\|\psi_t^{(k)}\right\|^2 + \|\bar{\varphi}_k\|^2 + \sum_{j=1}^{m} \epsilon_{kj}\rho_j \end{aligned}$$

because of our assumption A6);

$$\leqslant \left\|\psi_t^{(k)}\right\|^2 + \sum_{j=1}^{m}\left(\|\varphi_j\|^2 + \rho_j\right) \text{ as } \epsilon_{kj} \leqslant 1 \text{ for all } k, j.$$

Thus C4) is seen to hold with

$$a = \sum_{j=1}^{m}\left(\|\varphi_j\|^2 + \rho_j\right), \qquad b = 0, c = 1.$$

Condition C5) is seen to be true as

$$x'M_t(x) = x'x \geqslant 0.$$

The validity of C6) follows because

$$\begin{aligned} E\left\|\psi_1^{(k)}\right\|^2 &= E\left\|g\left(X_1^{(k)}\right)\right\|^2 \\ &= E\left(\|g(X)\|^2|\hat{w} = k\right) \\ &\leqslant \|\bar{\varphi}_k\|^2 + \sum_{j=1}^{m} \epsilon_{kj}\rho_j < \infty. \end{aligned}$$

Finally, C7) follows because

$$\inf_{\eta < \|x\| < \eta^{-1}}\left[x'M_t(x)\right] = \inf_{\eta < \|x\| < \eta^{-1}} x'x = \eta^2 > 0.$$

Hence the proposition.

Implications of Proposition 1

1) If the matrix A is the identity matrix, i.e., if there is no mislabeling then under our assumptions,

$$\hat{\varphi}_t^{(k)} \xrightarrow{\text{a.s.}} \varphi_k$$

as expected.

2) If $A \neq I_m$, then clearly the estimates $\hat{\varphi}_t^{(k)}$ for the different classes converge to nontrue values

$$\bar{\varphi}_k = \sum_{j=1}^{m} \epsilon_{kj}\varphi_j,$$

i.e., a convex linear combination of the parameter vectors of all the classes, as

$$\sum_{j=1}^{m} \epsilon_{kj} = 1 \forall k = 1, \cdots, m.$$

3) Yet another implication can be stated formally as follows.

Proposition 2: Consider the setup specified in Sections II and III. If assumptions A1)–A6) hold, then

$$\sum_{j=1}^{m} \gamma_{kj}\hat{\varphi}_t^{(j)} \xrightarrow{\text{a.s.}} \varphi_k, \qquad k = 1, \cdots, m$$

where

$$\Gamma_{m \times m} = ((\gamma_{ij}))$$

is a generalized inverse [16] of the matrix

$$E_{m \times m} = ((\epsilon_{ij}))_{\substack{i = 1, \cdots, m \\ j = 1, \cdots, m}}$$

satisfying

$$E\Gamma = I_m \tag{10}$$

Proof: Firstly, we note that the matrix E is not full-rank as shown by (3b). Consequently,

$$\operatorname{rank}(E) = r \leqslant m - 1.$$

From proposition 1, it is known that if E' denotes the transpose of E, then

$$\left(\hat{\varphi}_t^{(1)} \,|\, \hat{\varphi}_t^{(2)} \,|\, \cdots \,|\, \varphi_t^{(m)}\right) \xrightarrow{\text{a.s.}} E'\left(\varphi_1 \,|\, \varphi_2 \,|\, \cdots \,|\, \varphi_m\right) \text{element-wise}$$

(i.e., every element of the matrix on the left-hand side converges a.s. to the corresponding element on the right-hand side).

By well-known results on almost sure convergence it follows that

$$\left(\hat{\varphi}_t^{(1)} \,|\, \hat{\varphi}_t^{(2)} \,|\, \cdots \,|\, \hat{\varphi}_t^{(m)}\right) \xrightarrow{\text{a.s.}} E'\left(\varphi_1 \,|\, \varphi_2 \,|\, \cdots \,|\, \varphi_m\right) \text{column-wise}$$

TABLE I
PARAMETER VALUES FOR THE THREE CLASSES

Class k	π_k	μ_k'	Σ_k
1	$\frac{1}{3}$	(10, 20)	$\begin{bmatrix} 5 & 2 \\ 2 & 8 \end{bmatrix}$
2	$\frac{1}{3}$	(0, 5)	$\begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}$
3	$\frac{1}{3}$	(10, 10)	$\begin{bmatrix} 10 & 5 \\ 5 & 8 \end{bmatrix}$

(i.e., every column of the matrix on the left-hand side converges a.s. to the corresponding column on the right-hand side) by (10)).

It may be mentioned in passing that one such Γ satisfying (10) is the Moore–Penrose inverse [16] of E, viz. E^+ defined as

$$E^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} u_i u_i'$$

where λ_i is the ith nonzero eigenvalue of E, and u_i the corresponding eigenvector, $i = 1, \cdots, r$.

Some Simulation Results

To make an empirical study of the problem considered in this correspondence, we simulated the problem of learning of the mean vector in a three-class two-feature pattern recognition problem. The mean vector μ_j and disperson matrices Σ_j were prespecified and samples were generated with the help of the routines G05EAF and G05EZF in the Numerical Algorithms Group (NAG) package, so that actually pseudorandom samples were obtained. Tables I and II give the details of the parameters and the training sets for each class. n_{kj} denotes the numbers of samples in the training set which have true labels j but are (mis)labeled k.

We considered three different cases, using a different value of the matrix $A = ((a_{ij}))$ defined in Section III. In each case, we obtained a training set of size 20 for each class by combining samples from all three classes in suitable proportions determined by the elements of the respective A-matrix. The results obtained with these training sets are given in Table III. The distances of the estimates from the respective true values are also given to facilitate comparison of the effects of different sets of a_{ij} values. The inferences are obvious. An increase in the proportion of mislabeled samples in the training set causes the estimates to recede even further from the true values.

V. DISCUSSION

For the general m-class N-feature pattern recognition problem, it is found that in the presence of labeling errors for training samples, the recursive estimates for class parameters φ_k, defined by means of (1), do converge strongly under certain conditions. However, the values they converge to are not the true class-parameter values but certain convex linear combinations of true values for all the m classes.

This result is not surprising because one can easily guess that the presence of wrongly labeled training samples is bound to affect the behavior of the learning system in some way. This work merely confirms this suspicion mathematically by quantifying the effect on the asymptotic behavior of the system.

The next step, therefore, is to see how the learning procedure may be modified so that such deviant behavior is taken care of. One obvious method is to screen the training samples and weed out "doubtful" or "spurious" samples from among them. This approach was adopted by Chien [17] and Pal *et al.* [18] in their respective algorithms for parameter learning. Essentially, both algorithms reject those samples that do not lie within a certain neighborhood of the current estimate of the mean. We investigated the large-sample behavior of this class of restrictive-updating algorithms in [19] and arrived at the following conclusions.

Let $\hat{\varphi}_t^{(k)}$ denote the estimate for φ_k obtained at the tth step for the kth class, using this family of algorithms. Clearly, then

$$\hat{\varphi}_t^{(k)} = \begin{cases} f(X_t^{(k)}), & \text{for } t = 1 \\ \hat{\varphi}_{t-1}^{(k)} - a_t Y_{t-1}^{(k)}, & \text{for } t > 1 \end{cases} \tag{11}$$

where

$$\begin{aligned} Y_{t-1}^{(k)} &= \left(\hat{\varphi}_{t-1}^{(k)} - f(X_t^{(k)}\right), && \text{if } X_t^{(k)} \in G\left(\hat{\mu}_{t-1}^{(k)}, \lambda_t\right) \\ &= 0, && \text{otherwise} \end{aligned} \tag{12}$$

$\{a_t\}$ sequence of positive numbers.
f $R^N \to R^q$ is a continuous map defining an unbiased statistic for φ.
$\hat{\mu}_{t-1}^{(k)}$ $(t-1)$th step estimate of μ_k;
$G(\mu_{t-1}^{(k)}, \lambda_t) = \{X: X \in R^N, d(X_t^{(k)}, \mu_{t-1}^{(k)}) < \lambda_t\}$
$d^2(x, y)$ $= (x-y)'B_t(x-y)$.
B_t symmetric positive definite matrix, which may or may not be a function of the training samples $X_t^{(k)}$.
λ_t positive number suitably chosen.

Also, let $A_k(t)$ denote the event $\{w: X(w) \in G(\mu_{t-1}^{(k)}, \lambda_t)\}$.

Result 1: Under the setup considered in Section III and defined by assumptions A1)–A6), if we also have

A7) $p_t^{(k)} = P[A_k(t)|\hat{w} = k] > \partial_k$
for some $\partial_k \in (0,1)$ for all t.

then

$$\hat{\varphi}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\varphi_j \xrightarrow{\text{a.s.}} 0.$$

the N-dimensional null vector. Also, $E\{\|\hat{\varphi}_{t-1}^{(k)} - \Sigma_{j=1}^{m}\beta_{kj}(t+1)\varphi_j\|^2\}$ converges as $t \to \infty$. Here,

$$\begin{aligned} \beta_{kj}(t) &= \frac{P(A_k(t)|X, \hat{w} = k, w = j)\alpha_{kj}\pi_j}{P(\hat{w} = k, A_k(t))}, \qquad k, j = 1, \cdots, m \\ &= P(A_k(t)|X, \quad \hat{w} = k, w = j)\epsilon_{kj}P(A_k(t)/\hat{w} = k). \end{aligned} \tag{13}$$

Result 2: If, in addition to assumptions A1)–A7), we also have for some k,

A8) $\beta_{kj}(t) \to \beta_{kj}$ for all $k, j = 1(1)m$ as $t \to \infty$,

where $\beta_{kj} \in (0,1]$ and

A9) either $\Sigma_{j=1}^{m}\epsilon_{kj}\varphi_{jq} > \Sigma_{j=1}^{m}\beta_{kj}\varphi_{jq} > \varphi_{kq}$
or $\varphi_{kq} > \Sigma_{j=1}^{m}\beta_{kj}\varphi_{jq} > \Sigma_{j=1}^{m}\epsilon_{kj}\varphi_{jq}$, for each q.

then $\{\|\bar{\bar{\varphi}}_t^{(k)} - \varphi_k\| - \|\hat{\varphi}^{(k)} - \varphi_k\|\}$ converges almost surely to some strictly positive quantity l_k which is dependent on the parameters of the class C_k.

The implication of Result 1 is that the estimates $\hat{\varphi}^{(k)}$ converge strongly with the sequence, say,

$$\sum_{j=1}^{m} \beta_{kj}(t+1)\varphi_j = \bar{\bar{\varphi}}_t^{(k)}.$$

In particular, if A8) is also true, then this implies that these

TABLE II
α_{kj} AND $\bar{\varphi}_k$ VALUES

	α_{kj} for First Set				α_{kj} for Second Set				α_{kj} for Third Set			
k	$j=1$	$j=2$	$j=3$	$\bar{\varphi}_k'$	$j=1$	$j=2$	$j=3$	$\bar{\varphi}_k'$	$j=1$	$j=2$	$j=3$	$\bar{\varphi}_k'$
1	0.9	0.0	0.1	(10,19)	0.80	0.05	0.15	(9.5,17.75)	0.70	0.10	0.20	(9,16.5)
2	0.0	0.9	0.1	(1,5.5)	0.05	0.80	0.15	(2,6.5)	0.05	0.70	0.25	(3,7)
3	0.1	0.1	0.8	(9,10)	0.10	0.20	0.70	(8,10)	0.25	0.20	0.55	(8,11.5)

$n_{kj} = 20\alpha_{kj}$

TABLE III
LEARNING OF THE MEAN VECTOR USING (1)

Class	First Set		Second Set		Third Set	
k	$\hat{\varphi}_{20}^{(k)'}$	$\|\hat{\varphi}_{20}^{(k)} - \varphi_k\|$	$\hat{\varphi}_{20}^{(k)'}$	$\|\hat{\varphi}_{20}^{(k)} - \varphi_k\|$	$\hat{\varphi}_{20}^{(k)'}$	$\|\hat{\varphi}_{20}^{(k)} - \varphi_k\|$
1	(9.98, 18.97)	1.03	(9.76, 18.46)	1.56	(9.18, 17.62)	2.52
2	(0.16, 6.23)	1.24	(0.56, 6.41)	1.52	(1.09, 6.48)	1.84
3	(9.39, 11.02)	1.19	(8.86, 10.46)	1.23	(8.76, 11.42)	1.89

TABLE IV
LEARNING OF CLASS-1 MEAN VECTOR

Iteration	$x_i^{(1)}$	W	$\hat{\varphi}_i^{(1)}$	$\|\hat{\varphi}_i^{(1)} - \varphi_1\|$	$d(\cdot)$	λ_i	Update	$\hat{\xi}_i^{(1)}$	$\|\hat{\xi}_i^{(1)} - \varphi_1\|$
1	9.08, 24.61	1	9.08, 24.61	4.70	—	—	y	9.08, 24.61	4.70
2	10.50, 18.20	1	9.79, 21.41	1.43	—	—	y	9.79, 21.41	1.42
3	11.67, 10.91	3	10.42, 17.91	2.13	4.21	9.18	y	10.42, 17.91	2.13
4	11.21, 20.58	1	10.62, 18.58	1.55	0.88	1.56	y	10.62, 18.58	1.55
5	9.16, 19.11	1	10.32, 18.68	1.36	1.49	0.95	n	10.62, 18.58	1.55
6	8.87, 16.61	1	10.08, 18.34	1.66	1.46	1.54	y	10.33, 18.25	1.78
7	5.15, 12.51	2	9.38, 17.51	2.57	5.22	4.68	n	10.33, 18.25	1.78
8	9.87, 14.57	3	9.43, 17.14	2.92	0.99	2.25	y	10.26, 17.79	2.22
9	8.57, 15.92	1	9.34, 17.00	3.07	1.92	1.72	n	10.26, 17.79	2.22
10	10.76, 24.65	1	9.48, 17.77	2.29	1.90	4.69	y	10.32, 18.48	1.55
11	10.39, 20.62	1	9.57, 18.03	2.02	0.55	1.58	y	10.32, 18.67	1.37
12	8.65, 17.62	1	9.49, 17.99	2.07	2.16	1.52	n	10.32, 18.67	1.37
13	4.53, 10.68	3	9.11, 17.43	2.72	7.71	7.62	n	10.32, 18.67	1.37
14	11.34, 12.63	3	9.27, 17.09	3.00	2.05	4.73	y	10.40, 18.24	1.80
15	10.68, 21.65	1	9.36, 17.39	2.69	0.99	2.77	y	10.41, 18.47	1.58
16	7.97, 22.56	1	9.27, 17.72	2.39	3.56	3.95	y	10.26, 18.72	1.31
17	2.88, 8.97	2	8.90, 17.20	3.01	8.44	8.29	n	10.26, 18.72	1.31
18	13.51, 16.22	1	9.15, 17.15	2.97	3.60	2.78	n	10.26, 18.72	1.31
19	8.95, 18.66	1	9.14, 17.23	2.90	1.42	0.89	n	10.26, 18.72	1.31
20	9.90, 25.05	1	9.18, 17.62	2.52	1.75	4.30	y	10.25, 19.04	0.99

$d(\cdot)$: $d(x_i^{(1)}, \hat{\theta}_{i-1}^{(1)})$. y: updating done; n: no updating.

estimates too, converge strongly to nontrue values, viz,

$$\bar{\bar{\varphi}}_k = \sum_{j=1}^{m} \beta_{kj}\varphi_j$$

which are linear combinations of the true parameter values for all the classes. Result 2, however, establishes that the estimates $\hat{\xi}^{(k)}$ are asymptotically closer (in the sense of Euclidean distance) to the true value φ_k than the sequence of estimates $\hat{\varphi}_i^{(k)}$.

In fact, in the simulation studies made in [19] using the same data sets as in Tables I and II, it was found, even without verifying conditions A8) and A9), that a large majority of the estimates $\hat{\xi}^{(k)}$ were closer to the true values φ_k than the respective $\hat{\varphi}_i^{(k)}$'s were. The results for Class 1, using the *third* set of α_{kj}-values (from Tables I and II) are given in Table IV. The distance function d used was a weighted distance (from the preceding estimate of the mean), the weights being the preceding estimates of the standard deviations for the respective features. Also, we had taken

$$\lambda_i = 0.5(\lambda_{\min} + \lambda_{\max})$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the lower and upper bounds to λ_i, derived in [19].

REFERENCES

[1] P. A. Lachenbruch, "Discriminant functions when the initial samples are misclassified," *Technometrics*, vol. 8, pp. 657-662, 1966.

[2] ——, "Discriminant functions when the initial samples are misclassified II: Nonrandom misclassification models," *Technometrics*, vol. 16, pp. 419-424, 1974.

[3] G. J. McLachlan, "Estimating the linear discriminant function from initial samples containing a small number of unclassified observations," *J. Amer. Statist. Assoc.*, vol. 72, pp. 403-406, 1977.

[4] J. E. Michalek and R. C. Tripathi, "The effect of errors in diagnosis and measurement on the estimation of probability of an event," *J. Amer. Statist. Assoc.*, vol. 75, pp. 713-721, 1980.

[5] T. J. O'Neill, "Normal discrimination with unclassified observations," *J. Amer. Statist. Assoc.*, vol. 73, pp. 821-826, 1978.

[6] T. Krishnan, "Efficiency of normal discrimination with misclassified initial samples," Indian Statist. Inst., Calcutta, Tech. Rep. ASC 85 1, 1985.

[7] U. A. Katre and T. Krishnan, "Pattern recognition with an imperfect teacher," in *Pattern Recognition in Practice*, vol. II, E. S. Gelsema and L. N. Kanal, Eds. Amsterdam: North Holland, 1985.

[8] M. B. Nevelson and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*. Providence, RI: Amer. Math. Soc., 1973.

[9] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*. New York: Academic, 1968.

[10] K. S. Fu, "Relationships among various learning techniques in pattern recognition systems," in *Pattern Recognition*, L. Kanal, Ed. Washington, DC: Thomson Books, 1968, pp. 399–408.

[11] G. N. Saridis, Z. J. Nikolic, and K. S. Fu, "Stochastic approximation algorithms for system identification, estimation and decomposition of mixtures," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, pp. 8–15, 1969.

[12] K. S. Fu, "Learning control systems—review and outlook," *IEEE Trans. Automat. Contr.*, vol. AC-15, pp. 210–221, 1970.

[13] D. C. Farden, "Stochastic approximation with correlated data," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 105–113, 1981.

[14] C. B. Chittineni, "Learning with imperfectly labeled samples," *Pattern Recognition*, vol. 12, pp. 281–291, 1980.

[15] L. Schmetterer, "Multidimensional stochastic approximation," in *Multivariate Analysis — II: Proc. 2nd Int. Symp. Multivariate Analysis*, (P. R. Krishnaiah, Ed.) New York: Academic, 1968.

[16] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*. New York: Wiley, 1971.

[17] Y. T. Chien, "The threshold effect of a non-linear learning algorithms for pattern recognition," *Inform. Sci.*, vol. 2, pp. 351–358, 1970.

[18] S. K. Pal, A. K. Dutta, and D. Dutta Majumder, "A self-supervised vowel recognition system," *Pattern Recognition*, vol. 12, pp. 27–34, 1980.

[19] A. Pathak-Pal and S. K. Pal, "Generalized guard zone algorithm (GGA) for learning: Asymptotic and dynamic behavior," *Patt. Recog. Lett.*, submitted.