Effect of Wrong Samples on the Convergence of Learning Processes. II. A Remedy

AMITA PAL (PATHAK)

and

SANKAR K. PAL

Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India

ABSTRACT

In our earlier work [1] on the problem of parameter learning in pattern recognition, it was found that estimates converged to nontrue values in the presence of labeling errors. The present work describes a possible remedy to this problem by rejecting those training samples that do not lie within a certain neighborhood of the current estimate of the mean. The convergence of this class of restrictive updating procedure in presence of wrong samples has been studied along with the comparison of its estimates to those in [1]. It is established that, in the presence of labeling errors, the estimates of the proposed restrictive updating procedure are always asymptotically closer to the respective true values than the estimates in [1], provided that certain conditions are satisfied. A set of three-class bivariate data and speech data are also used to demonstrate the above features.

1. INTRODUCTION

This is a continuation of our earlier work on the effect of wrong samples on the convergence of learning processes [1]. We investigated there the convergence of stochastic approximation-based learning algorithms for the problems of parameter learning in pattern recognition when there is a possibility of training samples being mislabeled. It was found that the values the estimates converge to are not the true class parameter values but certain convex linear combinations of true values for all the classes. The general m-class N-feature pattern recognition problem was considered in [1].

The above result is not surprising because one can easily guess that the presence of wrongly labeled samples is bound to affect the behavior of the

655 Avenue of the Americas, New York, NY 10010 0020-0255/92/$03.50

learning system in some way. The work merely confirms this suspicion mathematically by quantifying the effect on the asymptotic behavior of the system.

As this work will seem incomplete without a solution to the problem considered, 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 [2] and Pal et al. [3] in their respective algorithms for parameter learning. Recently, Pathak and Pal [4] generalized these algorithms, called GGA (generalized guard zone algorithm), which consists of modifying the stochastic approximation procedure in such a way 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. The threshold value was again found to lie between certain bounds [5].

The present work concentrates on investigating the convergence properties of the GGA when there is a possibility of certain proportion of learning samples being mislabeled, and the asymptotic improvement of its estimates vis-a-vis the usual recursive unsupervised learning algorithm (i.e., non-GGA) [1]. It is found that in the presence of mislabeled training samples, like non-GGA, GGA also converges strongly to nontrue values which are linear combinations of true parameter values of all the classes. However, the GGA estimates have always an edge over the non-GGA estimates in the sense that they are asymptotically closer to the true parameter values than the non-GGA ones under certain conditions. It is also shown as a special case that if the GGA is effective in weeding out the mislabeled samples, the estimates then become consistent (i.e., the GGA estimates converge strongly to the true parameter values).

Finally, these features are demonstrated on an artificially generated two-dimensional three class data set and on a set of 424 vowel data in CNC (*C*onsonant-vowel *N*ucleus-*C*onsonant) context.

2. THE GENERALIZED GUARD-ZONE ALGORITHM (GGA) [4]

Let us consider a general m-class pattern recognition problem, where C_i, $i = 1, \ldots, n$, denotes the ith class. For this purpose, let the feature vector selected be

$$\mathbf{X}_{N\times 1} = [x_1, x_2, \ldots, x_N]', \qquad \mathbf{X} \in R^N.$$

Let us assume:

(A1) The distribution of $\mathbf{X}$ in each of the subsets of R^N corresponding to the different classes is continuous.

(A2) The probability densities $p_k(\mathbf{X})$ of $\mathbf{X}$ for the classes C_k, $k = 1, \ldots, m$, are of the same family and differ only in respect of the values of parameters.

(A3) The densities $p_k(\cdot)$ involve a q-dimensional parameter vector $\boldsymbol{\theta}_k$, which needs to be learned, either wholly or partly.

(A4) The densities $p_k(\cdot)$ admit of moments of the first two orders, that is,

$$E(\mathbf{X}|C_k) = \boldsymbol{\mu}_k$$

and

$$\mathrm{Var}(\mathbf{X}|C_k) = \boldsymbol{\Sigma}_k$$

exist.

(A5) An unbiased statistic exists for the parameter vector $\boldsymbol{\theta}$.

Let us suppose that for the purpose of learning $\boldsymbol{\theta}^{(k)}$, a set of independent samples

$$\{\mathbf{X}_1^{(k)}, \mathbf{X}_2^{(k)}, \ldots, \mathbf{X}_{nk}^{(k)}\}$$

are provided, $k = 1, \ldots, m$, where the superscripts k denote the labels "given" to the respective samples, as opposed to their true labels.

The generalized Guard-Zone Algorithm (GGA) for estimating $\boldsymbol{\theta}^{(k)}$ recursively is as follows:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_t^{(k)} &= \mathbf{f}\big(\mathbf{X}_t^{(k)}\big) && \text{for } t = 1, \\ &= \hat{\boldsymbol{\theta}}_{t-1}^{(k)} - a_t \mathbf{Y}_t^{(k)} && \text{for } t > 1, \end{aligned} \tag{1a}$$

where

$$\begin{aligned} \mathbf{Y}_t^{(k)} &= \hat{\boldsymbol{\theta}}_{t-1}^{(k)} - \mathbf{f}\big(\mathbf{X}_t^{(k)}\big) && \text{if } \mathbf{X}_t^{(k)} \in G\big(\hat{\boldsymbol{\mu}}_{t-1}^{(k)}, \lambda_t\big) \\ &= 0 && \text{otherwise}, \end{aligned} \tag{1b}$$

where

$\hat{\theta}_t^{(k)}$ = the tth stage estimate of θ_k,
$\{a_t\}$ = a sequence of positive numbers,
f: $R^N \to R^q$ is a continuous map, defining an unbiased statistic for θ,
$\hat{\mu}_{t-1}^{(k)}$ = the $(t-1)$th stage GAA estimate of μ_k,
$G(\hat{\mu}_{t-1}^{(k)}, \lambda_t) = \{\mathbf{X}: \mathbf{X} \in R^N, d(\mathbf{X}, \hat{\mu}_{t-1}^{(k)}) \leqslant \lambda_t\}$,
$d^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'\mathbf{B}_t(\mathbf{x}-\mathbf{y})$,
$\mathbf{B}_t$ = a symmetric positive definite matrix, which may or may not be a function of the training samples $\mathbf{X}_t^{(k)}$ (some examples given in [4]),
λ_t = a positive number, suitably chosen.

Incidentally, $G(\mathbf{a},\mathbf{r})$ is the guard zone and, clearly, is nothing but a closed ball centered at $\mathbf{a}$ and having radius r. In essence, this algorithm allows only those training samples to be used for the updating program which lie within the corresponding guard zone centered at the preceding estimate of the mean. Training samples which lie outside it are ignored and at the corresponding stages, the estimate is kept unchanged.

The choice of the various parameters of the algorithm, namely, $\{a_t\}$, λ_t and $\mathbf{B}_t$, will be governed by a number of factors. For instance, if we insist that the algorithm should converge almost surely, then (as we shall observe in Section 4) a sufficient condition required to hold is

$$\sum_{t=1}^{\infty} a_t^2 < \infty.$$

Clearly then, a set of possible choices of a_t is

$$a_t = t^{-\delta},$$

for any $\delta \in (1/2, \infty)$. In practice, it will be better to choose a small value of δ, or the corrections $\mathbf{Y}_t^{(k)}$ will be too small, otherwise.

Similarly, as also mentioned in [4], $\mathbf{B}_t$ can be chosen from among the following possible values, since $d(\cdot,\cdot)$ is a distance function:

$$\mathbf{B}_t = \mathbf{I} \qquad \text{(the identity matrix of order } N) \tag{2a}$$

or

$$\mathbf{B}_t = \left[\mathrm{Diag}\left(s_{11}^{(t)}, s_{22}^{(t)}, \ldots, s_{NN}^{(t)}\right)\right]^{-1} \tag{2b}$$

or

$$\mathbf{B}_t = \mathbf{S}_t^{-1} = \left(\left(s_{ij}^{(t)}\right)\right)_{N\times N}^{-1}, \tag{2c}$$

where $s_{ij}^{(t)}$ is the "current" estimate of σ_{ij}, the (i,j)th element of the covariance matrix.

The choice is generally governed by the nature of the probability density of the feature vector in any given class, provided, of course, such information is available *a priori*. If the features can be expected to be uncorrelated and to have unit variances, then the first choice (2a) is good enough. If they are uncorrelated but generally do not have unit variances, then the second choice (2b) is suitable. In the most general situation, the third choice (2c) can be used.

Finally, the choice of λ_t can also be made on the basis of some suitable criteria. For instance, we may choose λ_t so that it optimizes some performance index of the algorithm. This is an open problem, and its solution has not yet been attempted by us. In [5], however, we have taken

$$\lambda_t = (1-\alpha)l_t + \alpha L_t \tag{2d}$$

where l_t and L_t are respectively lower and upper bounds for λ_t, whose explicit expressions are given in [5], and α is a number $\in (0,1)$.

3. MODELING MISLABELED TRAINING SAMPLES

A very simple but realistic model, inspired by [6], is adopted for describing the situation in which there may be mislabeled training samples. Let w and $\hat{w}$ denote, respectively, the true and the given labels. Clearly,

$$w, \hat{w} \in \{1,2,\ldots,m\} = \Omega_C, \qquad \text{say.}$$

Let $\pi_k = P(w=k)$ denote the *a priori* probability for the class C_k, $k = 1,\ldots,m$. Further, let $p_k(\mathbf{X}) = p(\mathbf{X}|w=k)$ be the class-conditional probability density of the feature vector $\mathbf{X}$. Also, let α_{kj} denote the probability that a sample from C_j is given the label k, i.e.,

$$\alpha_{kj} = P(\hat{w}=k|w=j), \qquad j,k = 1,\ldots,m. \tag{3}$$

Clearly,

$$\sum_{k=1}^{m} \alpha_{kj} = 1.$$

Under this model, it can be shown that, for any subset $A_k(t)$ of the sample space, the probability density of a sample labeled k at the tth stage, i.e.

$$p\left(\mathbf{X}_t^{(k)}\right) = p\left(\mathbf{X}_t|\hat{w}=k\right) = p(\mathbf{X}|\hat{w}=k)$$

$$= \sum_{j=1}^{m} \beta_{kj}(t)p(\mathbf{X}|w=j) \qquad \text{if } \mathbf{X}_t^{(k)} \in A_k^{(t)}, \qquad \text{given } \hat{w}=k, \quad (4a)$$

$$= \sum_{j=1}^{m} \beta_{kj}^{*}(t)p(\mathbf{X}|w=j), \qquad \text{otherwise,} \qquad (4b)$$

where

$$A_k(t) = \left\{\mathbf{x}: \mathbf{x} \in G\left(\hat{\mu}_{t-1}^{(k)}, \lambda_t\right)\right\},$$

$$\beta_{kj}(t) = P\left(A_k(t)|\mathbf{X}, \hat{w}=k, w=j\right)\alpha_{kj}\pi_j / P(\hat{w}=k) \qquad (4c)$$

$$\beta_{kj}^{*}(t) = P\left(A_k^c(t)|\mathbf{X}, \hat{w}=k, w=j\right)\alpha_{kj}\pi_j / P(\hat{w}=k) \qquad (4d)$$

provided that we are prepared to assume:

(A6) $p(\mathbf{X}|\hat{w}=k, w=j) = p(\mathbf{X}|w=j)$ for all $j,k=1,2,\ldots,m$.
(A7) $P(\hat{w}=k, A_k(t)) > 0$ for all k,t.
(A8) $P(\hat{w}=k, A_k^c(t)) > 0$ for all k,t.

A proof is provided in Appendix A.

It is not difficult to observe that the quantities $\beta_{kj}(t), \beta_{kj}^{*}(t) \in [0,1]$ for all $k,j=1,2,\ldots,m$, as it is known that

$$P(\hat{w}=k) = \sum_{j=1}^{m} P(\hat{w}=k, w=j) = \sum_j \pi_j \alpha_{kj}.$$

We have not studied the problem of estimating the mislabeling probabilities β_{kj} yet. Offhand, it can be said, however, that they can be estimated if some measures of the probability of error of the labeling process involved are available. For instance, if the labeling is done with the help of some statistical classifier, then the error can be measured by its probabilities of misclassification, provided that these can be estimated.

4. CONVERGENCE OF THE LEARNING ALGORITHM

The convergence of a recursive estimate $\hat{\theta}_t$ for estimating a parameter θ, can be defined in various ways. For instance, we say that:

i. The sequence $(\hat{\theta}_t)$ converges to θ *with probability* 1 *or almost surely* (in symbols, $\hat{\theta}_t \overset{a.s.}{\to} \theta$) or *strongly* if

$$P\left[\lim_{t\to\infty} E\|\hat{\theta}_t - \theta\| = 0\right] = 1.$$

ii. $\hat{\theta}$ converges to θ *in the mean square* if

$$\lim_{t\to\infty} E\|\hat{\theta}_t - \theta\|^2 = 0,$$

E being the expectation operator.

For studying the asymptotic behavior of the learning algorithm GGA given in Section 2, use will be made of the following results, due to Schmetterer [7]:

LEMMA 1. *Let $\{a_n\}$ be a sequence of positive real numbers such that*

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

Let $\mathbf{x}_n$ and $\mathbf{y}_n$ be N-dimensional random vectors which satisfy

$$(C2) \qquad \mathbf{x}_{n+1} = \mathbf{x}_n - a_n\mathbf{y}_n, \quad n \geq 1.$$

Let $\mathbf{M}_n$ be a measurable mapping from R^N to R^N such that

$$(C3) \qquad E(\mathbf{y}_n|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathbf{M}_n(\mathbf{x}_n) \quad a.e.$$

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

$$(C4) \qquad E\left(\|\mathbf{y}_n\|^2|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right) \leq a + b\|\mathbf{x}_n\| + c\|\mathbf{x}_n\|^2 \quad a.e.$$

Also, for every $\mathbf{x} \in R^N$ and $n \geq 1$,

$$(C5) \qquad \mathbf{x}'\mathbf{M}_n(\mathbf{x}) \geq 0.$$

If $\mathbf{x}_1^{\circ}$ *is chosen in such a way that*

(C6) $$E\left(\|\mathbf{x}_1\|^2\right) \text{ exists},$$

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

LEMMA 2. *Suppose that the conditions* (C1)–(C6) *hold. If further, there exists for every* $\eta > 0$ *a* $\delta > 0$ *such that for* $n \geq 1$

(C7) $$\inf_{\eta < \|x\| < \eta^{-1}} \left[\mathbf{x}'\mathbf{M}_n(\mathbf{x})\right] \geq \delta,$$

then $\{\mathbf{x}_n\}$ *converges almost surely to the N-dimensional null vector* $\mathbf{0}$.

Let us now state the following theorems.

THEOREM 1. *Consider the setup given in Sections* 2 *and* 3. *If, in addition to assumptions* (A1)–(A8), *we also have*

(A9) $\sum_{n=1}^{\infty} a_n^2 < \infty$, *where* $a_n > 0$ *for all* n

(A10) $\rho_i = E(\|\mathbf{f}(\mathbf{x})\|^2 | w = i)$ *exists, with respect to each class-conditional density* $p_i(\mathbf{X})$,

(A11) $p_t^{(k)} = P[A_k(t) | \hat{w} = k] > \delta_k$ *for some* $\delta_k \in (0,1)$ *for all* t, *then*

$$\phi_t^{(k)} = \hat{\theta}_t^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t)\theta_j \overset{a.s.}{\to} 0, \qquad \textit{the N-dimensional null vector}.$$

Also, $\{E\|\phi_t^{(k)}\|^2\}$ *converges as* $t \to \infty$.

Here $\beta_{kj}(t)$, $k, j = 1(1)m$ is as in Equation (4).

Proof of the theorem follows directly from Lemmas 1 and 2 by verifying that conditions (C1)–(C7) are true for our setup, and is given in Appendix B.

THEOREM 2. *For the setup in Sections* 2 *and* 3, *if assumptions* (A9)–(A11) *also hold, then*

$$\sum_{j=1}^{m} \gamma_{kj}(t)\hat{\theta}_t(j) \overset{a.s.}{\to} \theta^{(k)} \text{ as } t \to \infty \text{ for } k = 1,2,\ldots,m,$$

where $\gamma_{kj}(t)$, $k, j = 1,2,\ldots,m$, *are the elements of the inverse* $\Gamma_{m\times m}(t)$ *of the matrix* $\mathbf{B}_{m\times m}(t) = ((\beta_{ij}(t)))$, *satisfying* $\mathbf{B}(t)\Gamma(t) = \mathbf{I}_m$, *the identity matrix of order* m.

The proof is trivial.

Note. It follows directly from Theorem 1 (and also Theorem 2) that in the case where there is no misclassification, that is, if

$$\beta_{kj}(t) = \delta_{kj}, \qquad \text{the Kronecker delta,} \qquad \text{for all } k, j = 1,2,\ldots,m,$$

then the sequence of estimates $\hat{\theta}_t^{(k)}$ is strongly consistent for $\theta^{(k)}$. This was also seen in [4].

As observed earlier in the Theorem 1, under certain conditions, viz., (A9)–(A11), we have

$$\hat{\theta}_t^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t)\theta^{(j)} \overset{a.s.}{\to} 0.$$

A similar result may be obtained for the usual recursive (non-GGA) estimate $\hat{\theta}_t^{(k)}$ for $\theta^{(j)}$, obtained without using the GGA. For such an estimate

$$\begin{aligned} \hat{\hat{\theta}}_t^{(k)} &= f\left(\mathbf{X}_t^{(k)}\right) \qquad \text{for } t = 1 \\ &= \hat{\hat{\theta}}_{t-1}^{(k)} - a_{t-1}\left(\hat{\hat{\theta}}_{t-1}^{(k)} - f\left(\mathbf{X}_t^{(k)}\right)\right) \qquad \text{for } t > 1, \end{aligned} \tag{5}$$

the corresponding result may be stated as follows [1].

RESULT 1. Under the conditions (A9) and (A10), and the setup considered earlier,

$$\hat{\hat{\theta}}_t^{(k)} \overset{a.s.}{\to} \sum_{j=1}^{m} \epsilon_{kj}\theta^{(j)}$$

where

$$\epsilon_{kj} = \pi_j \alpha_{kj} \Big/ \left(\sum_{i=1}^{m} \pi_i \alpha_{ki}\right).$$

In both cases, therefore, a.s. convergence takes place but in different forms. In the first situation, the sequence of estimates $\{\theta_t^{(k)}\}$ converges strongly with another sequence $\{\bar{\theta}_t^{(k)}, t = 1,2,\ldots\}$, where

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

In the second situation obviously, $\hat{\theta}_t^{(k)}$ converges strongly to $\bar{\bar{\theta}}^{(k)} = \sum_{j=1}^{m} \epsilon_{kj}\theta^{(j)}$. In order to effect a comparison between the two algorithms with respect of (strong) convergence, it would be logical, therefore, to study how the sequence $\{\hat{\theta}_t^{(k)},\ t = 1, 2, \ldots\}$ behaves with respect to the true value of $\theta^{(k)}$. More specifically, we may wish to know whether $\{\hat{\theta}_t^{(k)}\}$ manages at all to get "closer" eventually to $\theta^{(k)}$ than $\{\hat{\theta}_t^{(k)}\}$ does. The following theorem establishes that, under certain additional conditions, the GGA estimates $\hat{\theta}_t^{(k)}$ do asymptotically approach the true parameter values "closer" than do the usual recursive non-GGA estimates $\hat{\theta}_t^{(k)}$.

THEOREM 3. *If, in addition to the assumptions (A1)–(A11), we also have, for some k,*

$$(A12) \qquad \beta_{kj}(t) \to \beta_{kj} \qquad \text{for all } k, j = 1, \ldots, m \quad \text{as } t \to \infty,$$

where $\beta_{kj} \geq 0$.

$$(A13) \qquad \text{either } \sum_{j=1}^{m} \epsilon_{kj}\theta_{jq} > \sum_{j=1}^{m} \beta_{kj}\theta_{jq} > \theta_{kq}$$

$$\text{or } \theta_{kq} > \sum_{j=1}^{m} \beta_{kj}\theta_{jq} > \sum_{j=1}^{m} \epsilon_{kj}\theta_{jq} \text{ for each } q,$$

then

$$\|\hat{\bar{\theta}}_t^{(k)} - \theta^{(k)}\| - \|\hat{\theta}_t^{(k)} - \theta^{(k)}\| \overset{a.s.}{\to} G_k \text{ where } G_k > 0.$$

Proof. Under the assumption (A12), it follows from Theorem 1 that

$$\hat{\theta}_t^{(k)} \overset{a.s.}{\to} \bar{\theta}^{(k)},$$

where

$$\bar{\theta}^{(k)} = \sum_{j=1}^{m} \beta_{kj}\theta^{(j)}.$$

This, together with the result stated earlier, implies that

$$\hat{\theta}_t^{(k)} - \theta^{(k)} \overset{a.s.}{\to} \bar{\theta}^{(k)} - \theta^{(k)}$$

and

$$\hat{\bar{\theta}}_t^{(k)} - \theta^{(k)} \overset{a.s.}{\to} \bar{\bar{\theta}}^{(k)} - \theta^{(k)}.$$

Consequently,

$$\|\hat{\bar{\theta}}_t^{(k)} - \theta^{(k)}\| - \|\hat{\theta}_t^{(k)} - \theta^{(k)}\| \overset{a.s.}{\to} \|\bar{\bar{\theta}}^{(k)} - \theta^{(k)}\| - \|\bar{\theta}^{(k)} - \theta^{(k)}\|.$$

However,

$$\|\bar{\bar{\theta}}^{(k)} - \theta^{(k)}\|^2 - \|\bar{\theta}^{(k)} - \theta^{(k)}\|^2 = \left(\bar{\bar{\theta}}'\bar{\bar{\theta}} - \bar{\theta}'\bar{\theta}\right) - 2\theta'\left(\bar{\bar{\theta}} - \bar{\theta}\right),$$

dropping superscripts, for convenience,

$$= \left(\bar{\bar{\theta}} - \bar{\theta}\right)'\left(\bar{\bar{\theta}} - \bar{\theta}\right) + 2\bar{\theta}'\left(\bar{\bar{\theta}} - \bar{\theta}\right) - 2\theta'\left(\bar{\bar{\theta}} - \bar{\theta}\right)$$

$$= \|\bar{\bar{\theta}} - \bar{\theta}\|^2 + 2(\bar{\theta} - \theta)'\left(\bar{\bar{\theta}} - \bar{\theta}\right) > 0$$

because of (A13). Hence the theorem. ■

REMARKS.

1. This theorem formalizes some sufficient conditions under which the GGA provides estimates which are asymptotically "closer" to the respective true values than the usual non-GAA estimates.

2. One implication of the condition (A13) is that Theorem 3 will also be true if

$$\bar{\bar{\theta}}^{(k)} > \bar{\theta}^{(k)} > \theta^{(k)}$$

or if

$$\theta^{(k)} > \bar{\theta}^{(k)} > \bar{\bar{\theta}}^{(k)},$$

where the partial order relation $>$ is defined as follows:

$$\text{for } a, b \in R^N, \qquad a \prec b \qquad \text{if } a_i < b_i \qquad \text{for all } i = 1, \ldots, N.$$

Generally speaking, these conditions signify that the theorem will be true only for those learning situations in which the configuration of the m classes is such that, for any given class, either

(a) the true mean $\theta^{(k)}$ is an interior point of the lower quantant of $\theta^{(k)}$ which, in turn, is an interior point of the lower quantant of $\theta^{(k)}$,

or

(b) the inclusion relations are true in the reverse order.

Obviously, then, whether or not GGA estimates are asymptotically "closer" to the true mean than the non-GGA estimates is dependent on the nature of the problem.

(By the lower quantant of any point y_0 in the N-dimensional space $\mathbb{R}^N$, we mean the region

$$Q_L(y_0) = \{y: y_i < y_{i0} \forall i = 1, 2, \ldots, N\}).$$

5. IMPLEMENTATION AND RESULTS

The algorithm described in Section 2 was implemented on the following two different sets of data, for learning the class mean vectors and the covariance matrices:

a. An artificially generated data set for a two-feature three-class PR problem, the feature vector having bivariate normal class-conditional densities.

b. A real data set, consisting of 424 samples of five vowels (a, i, u, e, o). The features considered were the first three formant frequencies F_1, F_2, and F_3.

TABLE I

Some Parameter Values Related to the Artificial Data Set

Class k	Mean Vector $\mu(k)^t$	Covariance Matrix $\Sigma^{(k)}$		α_{ij} value			Modified Mean Vector $\tilde{\bar{\mu}}(k)^t$	Modified Covariance Matrix $\tilde{\bar{\Sigma}}^{(k)}$	
				$j=1$	$j=2$	$j=3$			
1	(10, 15)	103	152	0.85	0.05	0.10	(9.25, 14.00)	92.00	135.35
		152	233					135.35	209.80
2	(5, 5)	29	25	0.05	0.80	0.15	(5.25 6.25)	32.85	34.95
		25	29					34.95	50.30
3	(5, 10)	30	49	0.1	0.15	0.75	(5.50 9.75)	37.15	55.70
		49	103					55.70	104.90

5.1. ARTIFICIAL DATA SET

For each of the three classes, the mean vectors and covarianee matrices were specified first, and, using these, 20 random samples for each class were generated, using standard techniques based on random normal deviates. The samples from the three classes were then mixed in specific proportions to obtain training sets of size 20 for each of the three classes. The true means and covariance matrices for the three classes, as well as the 3×3 matrix of α_{ij} values, where

$$\alpha_{ij} = \text{Prob}[\text{a sample from class } j \text{ is labeled } i]$$

are given in Table 1. This table also gives the values of $\tilde{\bar{\mu}}^{(k)}$ and $\tilde{\bar{\Sigma}}^{(k)}$, (equivalent to the $\bar{\theta}^{(k)}$ defined in Section 4) for $k = 1$, 2, and 3.

The GGA and the non-GGA were implemented on the data, using

i. $a_t = 1/t$ for all $t > 1$.
ii. Equal *a priori* probabilities for all the classes.
iii. The Mahalanobis distance for the distance function d.
iv. $\lambda_t = (1 - \alpha) l_t + \alpha L_t$, where l_t and L_t are respectively the lower and upper bounds for λ_t, obtained in [5], and $\alpha \in (0, 1)$.

The optimum value of α was empirically found to be 0.8. The estimates obtained, with this value of α, are given in Tables 2(a), 2(b) and 2(c), respectively, for the three different classes. At each iteration, the distances of the estimates, both individual and (cumulative) average, from the true value, have been calculated separately for the mean vector and the vector consisting of the distinct elements of the covariance matrix.

TABLE 2(a)

Learning of Means and Covariances of Class 1 Using GGA and Non-GGA

Ser No.	Sample		True d Class	(,)	Lamda	Upd?	GGA Estimates of Mean Vector		Covariance Matrix (Raw)	
1	11.20	15.09	1	0.00	0.00	—	11.200	15 088	125.4469	168.9856
									168.9856	227.6351
2	12.38	13.80	1	0.00	0 00	—	11.788	14 446	139 3107	169 9111
									169 9111	209 0852
3	10.82	14.34	1	3.35	3 65	Y	11 467	14 412	131.9301	165 0289
									165.0289	207.9721
4	10 58	16 67	1	4.74	6.39	Y	11.245	14 977	126.9330	167.8718
									167 8718	225.4735
5	8.91	15.46	1	4.56	4 99	Y	10 777	15.073	117.4074	161.8245
									161 8245	228.1524
6	10.38	13.25	1	2.50	2.44	N	10 777	15.073	117.4074	161.8245
									161.8245	228.1524
7	5.41	6 48	2	14.58	13 25	N	10 777	15 073	117.4074	161.8245
									161 8245	228 1524
8	9.42	11 41	1	5.45	5.11	N	10 777	15.073	117.4074	161.8245
									161 8245	228 1524
9	8 48	14.39	1	3 02	3 13	Y	10.522	14 997	112 3541	157 4067
									157.4067	225.8188
10	8.07	10.19	3	0.74	3 71	Y	10 277	14 516	107 6277	149 8856
									149.8856	213.6163
11	9.66	13.66	1	0.16	0 76	Y	10 221	14 438	106.3306	148 2601
									148.2601	211.1649
12	10 33	14.48	1	0.07	0 09	Y	10.230	14 442	106.3605	148.3653
									148.3653	211.0301
13	6.06	9.49	3	1.33	5.12	Y	9.909	14 061	101.0079	141.3792
									141.3792	201.7237
14	9.41	14.72	1	0 74	0.66	N	9.909	14.061	101 0079	141.3792
									141.3792	201 7237
15	10.91	12.82	1	1.44	1 26	N	9.909	14.061	101 0079	141 3792
									141.3792	201.7237
16	12.20	14.37	1	1.72	1 83	Y	10 052	14.080	103 9917	143.4929
									143.4929	202.0130
17	11.99	16.39	1	0.65	2 38	Y	10.166	14.215	106.3263	146.6062
									146.6062	205.9249
18	13.60	16.48	1	1 83	3.34	Y	10.357	14.341	110.6910	150.9139
									150.9139	209.5808
19	11.03	12.96	1	1.23	1.20	N	10 357	14.341	110.6910	150.9139
									150.9139	209.5808
20	11.41	16.42	1	0.42	1 83	Y	10.409	14.445	111.6657	152.7373
									152.7373	212.5871

TABLE 2(a) Continued.

Distance from True Parameter Values of the GGA Estimates				Non-GGA Estimates of				Distances from True Parameter Values of the Non-GGA Estimates of			
Means		Dispersions		Mean Vector		Covariance Matrix (Raw)		Means		Dispersions	
Indiv.	Average	Indiv.	Average					Indiv.	Average	Indiv.	Average
1.204	1.204	28.656	28 656	11.200	15.088	125.4469	168.9856	1.204	1 204	28.656	28.656
						168.9856	227.6351				
1.872	1 574	47.023	38.938	11.788	14.446	139.3107	169.9111	1.872	1.574	47.023	38.938
						169.9111	209.0852				
1 581	1.576	40.412	39.435	11.467	14.412	131.9301	165 0289	1.581	1.576	40.412	39.435
						165.0289	207 9721				
1 246	1 500	29.688	37.238	11 245	14.977	126.9330	167.8718	1.246	1.500	29.688	37.238
						167.8718	225.4735				
0 781	1 387	18 100	34.277	10.777	15.073	117.4074	161.8245	0.781	1.387	18.100	34.277
						161.8245	228.1524				
0 781	1 305	18.100	32 151	10.712	14.768	115.8125	157 7820	0.749	1.302	19.575	32 294
						157.7820	219.3767				
0 781	1 244	18.100	30.542	9 954	13.584	103.4459	140.2449	1.417	1.319	40.708	33.626
						140.2449	194 0284				
0.781	1.196	18 100	29.277	9.887	13.312	101 6067	136.1449	1.692	1.371	49.586	36.010
						136 1449	186.0379				
0 522	1 141	12 973	27 939	9.731	13 432	98.3090	134 5804	1.591	1.397	48.125	37.549
						134.5804	188.3837				
0.557	1 097	20 040	27.253	9 565	13 107	94.9871	129 3419	1 942	1.461	58 263	40.105
						129 3419	179 9247				
0 603	1 061	22 402	26 848	9.574	13.158	94.8391	129 5841	1.891	1.505	57.633	42.002
						129.5841	180 5361				
634	1 031	22.521	26.514	9 637	13 268	95.8267	131.2455	1.770	1 529	54.652	43.198
						131.2455	182.9538				
0.944	1.024	33.090	27.077	9.362	12.977	91.2844	125 5764	2.121	1.583	64.082	45.148
						125.5764	175.8071				
0 944	1.019	33.090	27.550	9.365	13.101	91.0835	126.4963	2.002	1.616	61.139	46.474
						126.4963	178.7267				
0.944	1.014	33.090	27.954	9.468	13.083	92 9530	127.3942	1.989	1.644	61.288	47.605
						127.3942	177.7748				
0 922	1.009	32 149	28.234	9.639	13.163	96.4402	130.3820	1.872	1.659	58.018	48.321
						130.3820	179.5610				
0.802	0 998	27 807	28.209	9.777	13.353	99.2190	134.2665	1.662	1.659	51.504	48.514
						134.2665	184.7935				
0.749	0.985	24.674	28.024	9.989	13.527	103.9786	139.2598	1.473	1.649	45.219	48.337
						139.2598	189.6234				
0.749	0.974	24.674	27 858	10.044	13.497	104.9119	139.4564	1.504	1.642	46.288	48.232
						139.4564	188.4855				
0.689	0.962	22.188	27 602	10.112	13.643	106.1756	141.8527	1.361	1.629	41.827	47.932
						141.8527	192.5466				

TABLE 2(b)

Learning of Means and Covarianees of Class 2 Using GGA and Non-GGA

Ser No.	Sample		True d Class	$(\cdot,)$	Lamda	Updt?	GGA Estimates of Mean Vector		Covariance Matrix (Raw)	
1	4.68	4 42	2	0.42	1.83	—	4 682	4 416	21 9211	20.6757
									20.6757	19 5011
2	5.56	4 54	2	0.42	1 83	—	5 120	4 478	26 4062	22.9545
									22 9545	20.0563
3	6.19	4.72	2	2 46	3.32	Y	5.476	4 560	30.3679	25.0470
									25 0470	20.8096
4	5.37	4.02	2	1.64	1.59	N	5 476	4.560	30 3679	25.0470
									25.0470	20.8096
5	0.51	4 19	2	8.28	14 46	Y	4 483	4.487	24 3464	20 4654
									20.4654	20 1656
6	2.57	4.29	2	0.93	1.93	Y	4 165	4 454	21 3929	18.8949
									18.8949	19 8720
7	4.11	3.68	2	0 92	0.84	N	4 165	4 454	21 3929	18.8949
									18 8949	19.8720
8	2.27	6.19	2	2.70	2 79	Y	3 928	4.671	19.3618	18 2874
									18 2874	22 1744
9	6.88	3.11	2	2.12	2.38	Y	4 256	4 497	22.4759	18 6327
									18.6327	20 7839
10	3.04	9.57	3	3.80	3.55	N	4 256	4 497	22 4759	18 6327
									18 6327	20 7839
11	4.37	4.07	2	0.32	0.30	N	4 256	4 497	22 4759	18 6327
									18 6327	20.7839
12	9.78	8.27	1	3.67	4.56	Y	4 717	4.812	28.5765	23 8237
									23.8237	24.7555
13	7.96	5.73	2	5.71	19.77	Y	4.966	4.882	31.2474	25.4967
									25 4967	25.3751
14	2.98	5.00	2	2.36	14 93	Y	4.824	4.891	29.6489	24 7399
									24.7399	25.3511
15	4.50	8.89	3	14.40	11.82	N	4 824	4 891	29 6489	24.7399
									24.7399	25.3511
16	4.26	5.44	2	2 10	2.31	Y	4.789	4.925	28 9322	24.6434
									24.6434	25.6163
17	4.25	9.97	3	16.58	13 91	N	4.789	4.925	28.9322	24.6434
									24.6434	25.6163
18	4.99	3.09	2	6.01	5.05	N	4.789	4.925	28.9322	24.6434
									24.6434	25.6163
19	4.32	1.54	2	11.38	9.37	N	4.789	4 925	28.9322	24.6434
									24.6434	25.6163
20	5 23	1.07	2	12.64	10.62	N	4.789	4.925	28.9322	24 6434
									24.6434	25.6163

TABLE 2(b) Continued.

Distance from True Parameter Values of the GGA Estimates				Non-GGA Estimates of				Distances from True Parameter Values of the Non-GGA Estimates of			
Means		Dispersions		Mean Vector		Covariance Matrix (Raw)		Means		Dispersions	
Indiv.	Average	Indiv.	Average					Indiv	Average	Indiv.	Average
0 665	0.665	12.611	12.611	4.682	4 416	21 9211	20.6757	0.665	0.665	12.611	12.611
						20.6757	19.5011				
0.536	0.604	9.534	11.179	5.120	4.478	26.4062	22.9545	0.536	0.604	9 534	11.179
						22 9545	20 0563				
0 648	0 619	8 304	10 310	5.476	4 560	30.3679	25 0470	0.648	0.619	8.304	10.310
						25.0470	20.8096				
0 648	0 626	8.304	9 847	5.450	4 426	29.9852	24.1848	0.729	0.648	9.436	10.099
						24 1848	19.6513				
0.729	0.648	10.967	10.081	4.462	4.379	24.0402	19.7756	0.822	0.687	12.131	10.537
						19.7756	19.2390				
0 998	0.718	13.359	10 697	4 147	4.364	21 1377	18.3201	1.064	0.762	14.299	11.251
						18 3201	19.0998				
0.998	0.765	13.359	11 117	4.141	4 267	20 5265	17.8627	1.129	0.825	15.397	11.932
						17 8627	18.3080				
1.122	0.818	13 585	11 454	3.907	4.507	18 6037	17.3842	1.199	0 880	15.272	12.399
						17.3842	20.8059				
0.898	0.827	12 272	11 548	4.238	4.352	21.8021	17.8299	1.001	0.895	13.863	12.570
						17.8299	19.5675				
0.898	0.834	12 272	11 622	4 118	4.873	20.5454	18.9538	0.891	0.894	10.633	12.390
						18.9538	26.7608				
0.898	0.840	12 272	11.683	4 141	4.800	20 4121	18.8453	0.882	0.893	11.031	12.272
						18 8453	25.8309				
0.340	0.811	4 425	11 258	4.611	5.089	26 6847	24.0186	0.399	0.863	2.544	11.773
						24.0186	29.3819				
0.123	0.780	4.294	10 882	4.868	5 138	29.5011	25.6766	0.191	0 831	1.061	11.315
						25.6766	29.6456				
0.207	0.753	3.715	10.533	4.733	5.129	28.0273	24.9069	0.296	0.804	1.027	10.907
						24.9069	29.3166				
0 207	0.730	3.715	10 221	4.717	5 379	27.5083	25.9116	0.473	0.787	4.025	10.588
						25.9116	32.6255				
0.224	0.709	3.403	9 933	4.689	5.383	26.9254	25.7419	0.493	0.772	4.082	10.303
						25.7419	32.4360				
0.224	0 690	3.403	9 672	4.663	5.653	26.4045	26.7217	0.735	0.769	8.010	10.182
						26 7217	36.3799				
0.224	0 672	3.403	9.433	4.681	5.511	26.3187	26.0942	0.602	0.761	6.564	10.015
						26.0942	34.8906				
0.224	0.656	3.403	9.215	4.662	5.302	25.9167	25.0707	0.453	0.748	5.194	9.821
						25.0707	33.1788				
0.224	0.642	3.403	9.014	4.691	5.090	25.9906	24.0982	0.322	0.733	4.064	9.615
						24.0982	31.5775				

TABLE 2(c)

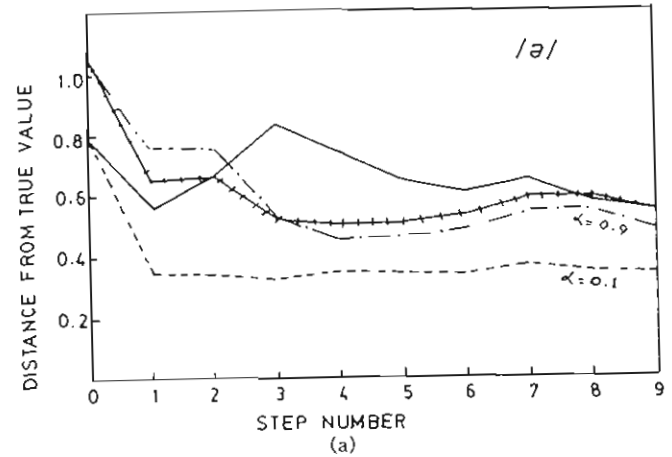
Learning of Means and Covariances of Class 3 Using GGA and Non-GGA

Ser No.	Sample		True d Class	$(\cdot,\cdot)$	Lamda	Updt?	GGA Estimates of Mean Vector		GGA Estimates of Covariance Matrix (Raw)	
1	4.30	8.51	3	12.64	10.62	—	4.305	8.511	18.5294	36.6361
									36.6361	72.4361
2	4.72	8.25	3	12.64	10.62	—	4.511	8.380	20.3957	37.7789
									37.7789	70.2424
3	3.59	11.43	3	9.00	15.65	Y	4.203	9.398	17.8855	38.8563
									38.8563	90.4062
4	4.56	11.88	3	8.46	15.74	Y	4.293	10.017	18.6216	42.6926
									42.6926	103.0644
5	7.01	7.01	3	6.34	8.88	Y	4.836	9.416	24.7196	43.9790
									43.9790	92.2789
6	2.56	13.24	3	2.15	5.27	Y	4.457	10.053	21.6905	42.2933
									42.2933	106.1049
7	7.61	12.16	1	5.59	4.79	N	4.457	10.053	21.6905	42.2933
									42.2933	106.1049
8	5.03	7.03	2	1.82	3.88	Y	4.528	9.676	22.1367	41.4257
									41.4257	99.9284
9	8.67	10.02	1	41.91	44.10	Y	4.989	9.713	28.0388	46.4776
									46.4776	99.1712
10	2.48	10.92	3	6.64	16.02	Y	4.738	9.834	25.8499	44.5373
									44.5373	101.1728
11	2.54	9.12	3	13.59	12.36	N	4.738	9.834	25.8499	44.5373
									44.5373	101.1728
12	5.58	11.55	3	11.98	10.22	N	4.738	9.834	25.8499	44.5373
									44.5373	101.1728
13	5.50	10.34	3	5.92	4.88	N	4.738	9.834	25.8499	44.5373
									44.5373	101.1728
14	9.53	9.07	3	18.84	26.01	Y	5.080	9.779	30.4972	47.5341
									47.5341	99.8240
15	6.71	10.64	3	8.80	7.49	N	5.080	9.779	30.4972	47.5341
									47.5341	99.8240
16	4.27	7.62	2	10.55	9.40	N	5.080	9.779	30.4972	47.5341
									47.5341	99.8240
17	4.86	11.12	3	4.03	5.54	Y	5.068	9.858	30.0947	47.9198
									47.9198	101.2279
18	6.57	11.69	3	10.90	8.92	N	5.068	9.858	30.0947	47.9198
									47.9198	101.2279
19	4.95	6.46	2	11.47	12.80	Y	5.062	9.679	29.8014	47.0809
									47.0809	98.0951
20	6.62	10.81	3	10.78	8.85	N	5.062	9.679	29.8014	47.0809
									47.0809	98.0951

TABLE 2(c) Continued.

Distance from True Parameter Values of the GGA Estimates				Non-GGA Estimates of				Distances from True Parameter Values of the Non-GGA Estimates of			
Means		Dispersions		Mean Vector		Covariance Matrix (Raw)		Means		Dispersions	
Indiv	Average	Indiv	Average					Indiv.	Average	Indiv.	Average
1.643	1.643	34.908	34.908	4.305	8.511	18.5294	36.6361	1.643	1.643	34.908	34.908
						36.6361	72.4361				
1.692	1.668	35.933	35.425	4.511	8.380	20.3957	37.7789	1.692	1.668	35.933	35.425
						37.7789	70.2424				
0.999	1.479	20.205	31.188	4.203	9.398	17.8855	38.8563	0.999	1.479	20.205	31.188
						38.8563	90.4062				
0.707	1.329	13.010	27.782	4.293	10.017	18.6216	42.6926	0.707	1.329	13.010	27.782
						42.6926	103.0644				
0.607	1.219	12.963	25.516	4.836	9.416	24.7196	43.9790	0.607	1.219	12.963	25.516
						43.9790	92.2789				
0.546	1.135	11.121	23.731	4.457	10.053	21.6905	42.2933	0.546	1.135	11.121	23.731
						42.2933	106.1049				
0.546	1.071	11.121	22.369	4.907	10.353	26.8682	49.4686	0.365	1.060	9.592	22.268
						49.4686	112.0547				
0.573	1.022	11.619	21.324	4.922	9.938	26.6673	47.7042	0.099	0.992	3.782	20.873
						47.7042	104.2325				
0.287	0.968	4.987	20.173	5.339	9.947	32.0660	52.0584	0.343	0.942	3.776	19.719
						52.0584	103.7988				
0.310	0.924	6.362	19.243	5.053	10.044	29.4744	49.5600	0.069	0.894	2.461	18.723
						49.5600	105.3377				
0.310	0.886	6.362	18.448	4.825	9.960	27.3826	47.1632	0.179	0.854	3.214	17.878
						47.1632	103.3274				
0.310	0.853	6.362	17.758	4.888	10.093	27.6946	48.6017	0.145	0.819	3.671	17.150
						48.6017	105.8287				
0.310	0.824	6.362	17.152	4.935	10.112	27.8881	49.2345	0.129	0.788	3.604	16.507
						49.2345	105.9110				
0.235	0.796	3.533	16.555	5.263	10.037	32.3898	51.8958	0.266	0.762	3.949	15.942
						51.8958	104.2236				
0.235	0.772	3.533	16.020	5.359	10.077	33.2286	53.1926	0.368	0.742	5.597	15.469
						53.1926	104.8218				
0.235	0.749	3.533	15.536	5.292	9.924	32.2935	51.9025	0.301	0.723	3.861	15.009
						51.9025	101.8956				
0.157	0.728	2.078	15.081	5.266	9.994	31.7853	52.0313	0.266	0.704	3.522	14.586
						52.0313	103.1777				
0.157	0.709	2.078	14.664	5.339	10.088	32.4173	53.4088	0.350	0.689	5.427	14.232
						53.4088	105.0428				
0.327	0.694	5.271	14.324	5.318	9.897	32.0018	52.2809	0.335	0.675	4.054	13.884
						52.2809	101.7093				
0.327	0.680	5.271	14.011	5.383	9.943	32.5907	53.2436	0.388	0.664	5.000	13.579
						53.2436	102.4679				

A careful inspection of Table 2 reveals that the performance of the GGA is uniformly better (with respect to the "closeness-to-the-true-value" criterion) for classes 1 and 2, but *not* for class 3. This is to be expected, in view of Theorem 3, for it can be readily seen from Table 1 that although the means and covariances of classes 1 and 2 satisfy the conditions of the theorem, those of class 3 do not. That is, for $k = 1, 2$, the modified parameters $\bar{\bar{\mu}}_{(k)}$ and $\bar{\bar{\Sigma}}_{(k)}$



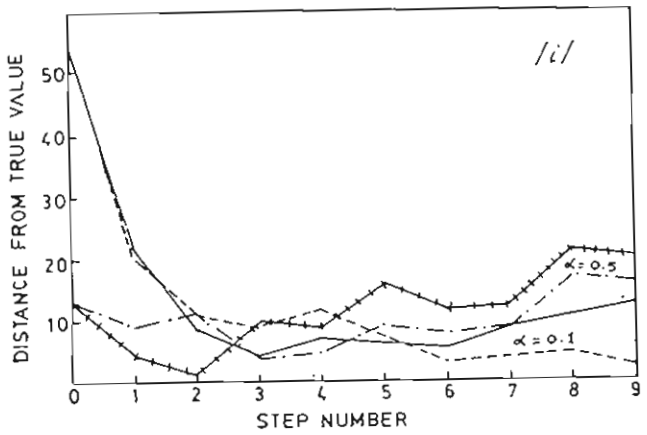
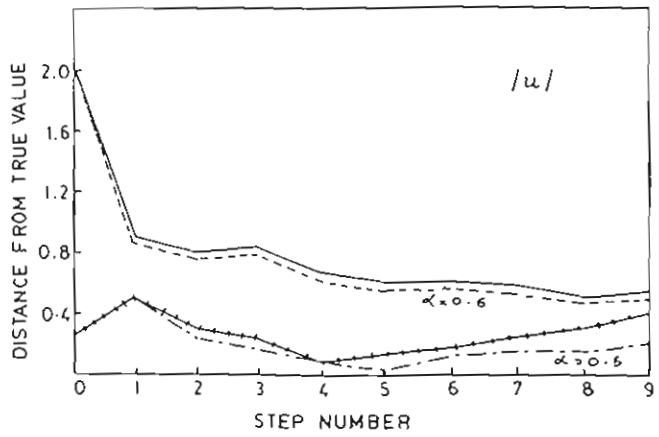
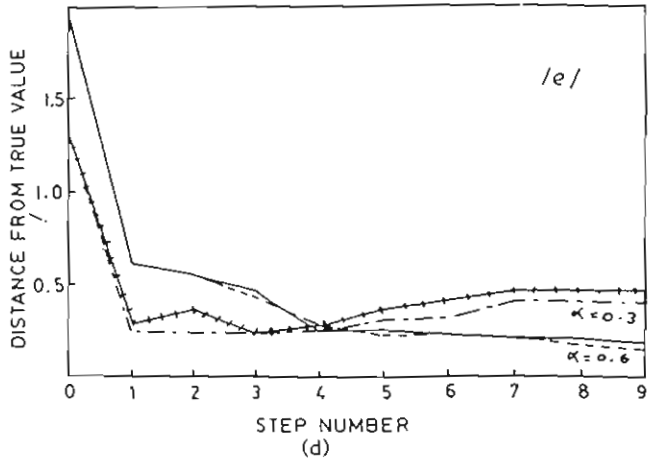


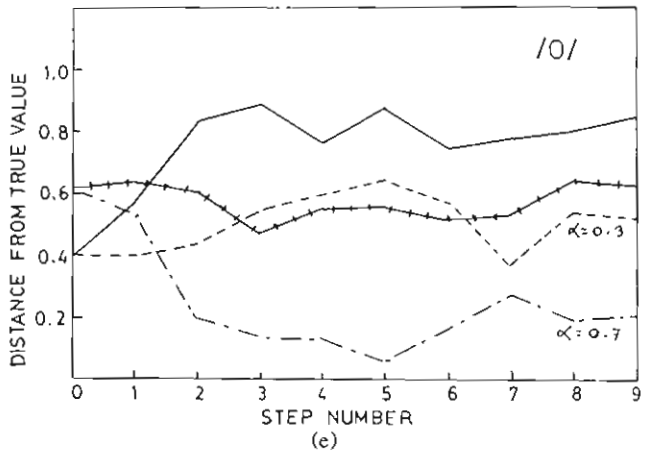
Fig. 1. Distance of estimated mean vectors from their true values. (a) /a/; (b) /i/; (c) /u/; (d) /e/; (e) /o/; (——) non-GGA with sequence 1; (---) GGA with sequence 1; (——) non-GGA with sequence 2; (-.-) GGA with sequence 2.

are either strictly less (for $k = 1$) or strictly greater (for $k = 2$) elementwise than the corresponding elements of the true parameters $\mu^{(k)}$ and $\Sigma^{(k)}$. However, for class 3, this is not true; although $\bar{\bar{\mu}}_{(k)_1}$ is greater than $\mu_1^{(k)}$, $\bar{\bar{\mu}}_{(k)_2}$ is less than $\mu_2^{(k)}$. By virtue of the theorem, therefore, the GGA estimates for classes 1 and 2 are expected to be "closer" to their true values in the long run, but not those for class 3.

These features of the GGA have been demonstrated further on some real data below.







5.2. *SPEECH DATA SET*

The data were prepared from a set of nearly 600 discrete phonetically balanced speech units in consonant-vowel-consonant Telugu (a major Indian language) vocabulary, uttered by three male speakers in the age group of 30–35 years. The first three vowel formant frequences (F_1, F_2, and F_3) at the

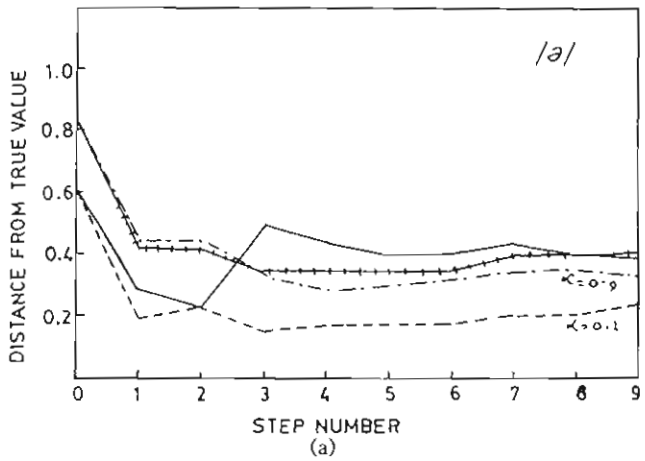
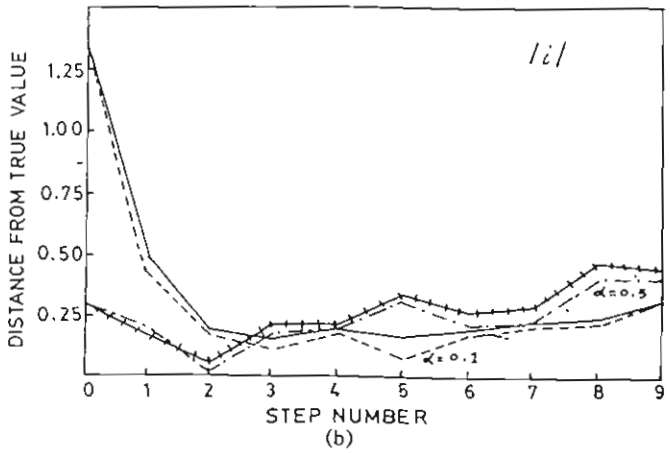
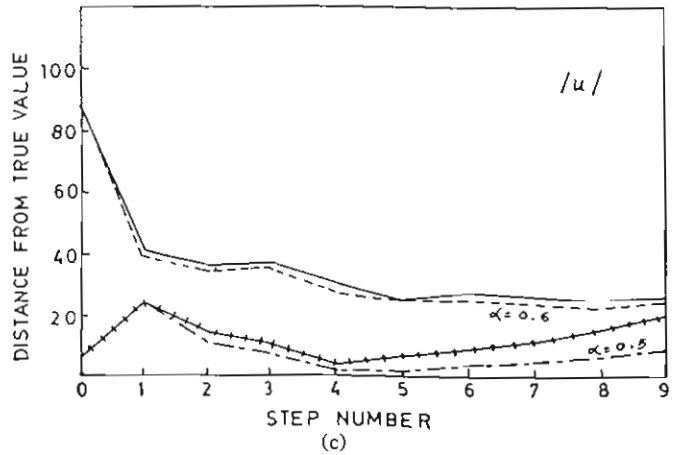


Fig. 2. Distance of estimated variance vectors from their true values: (a) /a/; (b) /i/; (c) /u/; (d) /e/; (e) /o/; (——) non-GGA with sequence 1; (---) GGA with sequence 1; (——) non-GGA with sequence 2; (-.-) GGA with sequence 2.

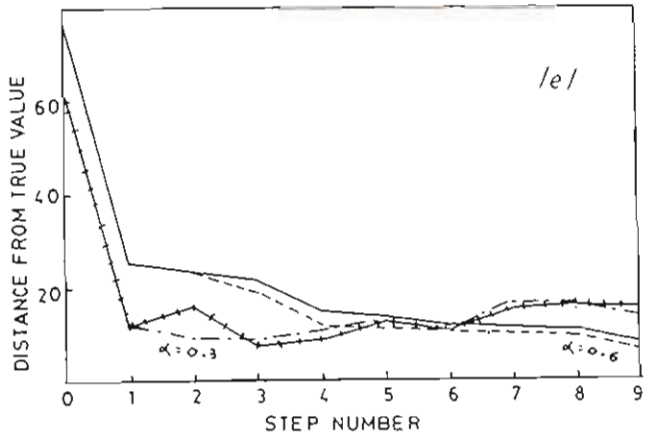




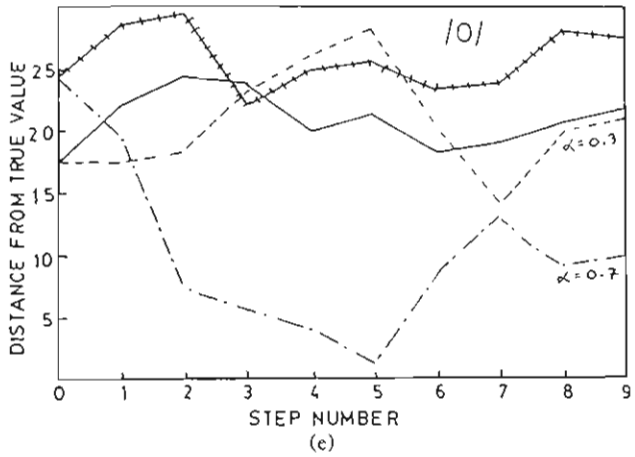
steady state were obtained through spectrum analysis. The details of processing and formant extraction are available in [8, 9].

For the problem of learning, the procedure used was to take a 10% random sample to obtain the initial estimates for each class. An adaptive Bayes classifier was then used on the entire data set to provide labels to the samples. The GGA (or the non-GGA) were then applied, with the same (algorithm) parameters as for the artificial data set.

Since an adaptive procedure is dependent on the sequence of incoming samples and the initial estimates, the algorithms were applied to various



(d)



(e)

permutations of the vowel data set. For illustration, we have given here the result corresponding to two sequences, with different initial estimates. The results are presented for different classes in the form of graphs as shown in Figures 1 and 2. Here, the distance of the estimates from their true values are plotted at intervals of 50 iterations. Figure 1 corresponds to mean vector whereas, Figure 2 corresponds to variance vector for the same sequences of input.

From the curves, the estimates obtained by the GGA are always seen, in the long run, to be closer to their true values than those of the non-GGA. Furthermore, as the initial estimates tend to be poorer (i.e., distance of the estimates from their true values increases), the value of α for obtaining better GGA estimates is seen to be higher [except for Figure 1(b)]. This conforms to our earlier investigation [5] where the similar effect of α on recognition score was observed. This means that the guard zone needs to be flexed more as the estimates tend to be weaker, in order to strengthen the estimates by allowing a higher proportion of correct to incorrect samples to be available for learning.

APPENDIX A: PROOF OF EQUATION (4)

From well-known results in probability theory, we have

$$p(\mathbf{X}|\hat{w}=k) = p(\mathbf{X}|\hat{w}=k,\, A_k(t))\, P(A_k(t)|\hat{w}=k) + p(\mathbf{X}|\hat{w}=k,\, A_k^c(t))\, P(A_k^c(t)|\hat{w}=k), \qquad (A.1)$$

where $A_k^c(t)$ stands for the event complementary to $A_k(t)$. However,

$$p(\mathbf{X}|\hat{w}=k, A_k(t)) = \frac{p(\mathbf{X}, \hat{w}=k, A_k(t))}{P(A_k(t), \hat{w}=k)} = \frac{\sum_{j=1}^{m} p(\mathbf{X}, \hat{w}=k, A_k(t), w=j)}{P(A_k(t), \hat{w}=k)} = \frac{\sum_{j=1}^{m} P(A_k(t))|\mathbf{X}, \hat{w}=k, w=j)\, p(\mathbf{X}|\hat{w}=k, w=j)\, P(\hat{w}=k, w=j)\, P(w=j)}{P(A_k(t), \hat{w}=k)}$$

$$= \frac{\sum_{j=1}^{m} P(A_k(t))|\mathbf{X}, \hat{w}=k, w=j)\, \alpha_{kj}\pi_j\, p(\mathbf{X}|\hat{w}=k, w=j)}{P(A_k(t)|\hat{w}=k)\, P(\hat{w}=k)} = \frac{\sum_{j=1}^{m} \beta_{kj}(t)\, p(\mathbf{X}|w=j)}{P(A_k(t)|\hat{w}=k)} \qquad (A.2)$$

by assumption (A6). Similarly, we must have

$$p(\mathbf{X}|\hat{w}=k, A_k^c(t)) = \frac{\sum_{j=1}^{m}\beta_{kj}^*(t)p(\mathbf{X}|w=j)}{P(A_k^c(t)|\hat{w}=k)} \tag{A.3}$$

Hence the equation. ■

APPENDIX B: PROOF OF THEOREM 1

The theorem can be shown to be true if it can be established that under the conditions (C1) and (C2),

(i) $\varphi_t^{(k)} \to 0$ with probability 1 as $t \to \infty$, $\forall k$
(ii) $\{E[\|\varphi_t^{(k)}\|^2]\}$ converges as $t \to \infty$, $\forall k$, where $\varphi_t^{(k)} = \hat{\theta}_t^{(k)} - \bar{\theta}$.

These, in turn, follow immediately from Lemmas 1 and 2 if it can be shown that the conditions (C1)–(C7) hold with $\mathrm{x}_n = \varphi_n^{(k)}$.

We first note that

$$\hat{\theta}_t^{(k)} = \begin{cases} \mathbf{f}(\mathbf{X}_t^{(k)}) & \text{for } t=1, \quad (B.1) \\ \hat{\theta}_{t-1}^{(k)} - \frac{1}{t}\mathbf{Y}_t^{(k)}, & t>1, \quad (B.2) \end{cases}$$

where

$$\mathbf{Y}_{t-1}^{(k)} = \begin{cases} \hat{\theta}_{t-1}^{(k)} - \mathbf{f}(\mathbf{X}_t^{(k)}) & \text{if } \mathbf{X}_k^{(t)} \in A_k(t) \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

And f: $\mathbb{R}^N \to \mathbb{R}^q$ is a continuous map defining an unbiased statistic for $\theta^{(k)}$. Obviously, therefore,

$$\varphi_t^{(k)} = \begin{cases} \mathbf{g}(\mathbf{X}_t^{(k)}) & \text{for } t=1 \quad (B.4) \\ \varphi_{t-1}^{(k)} - \frac{1}{t}\mathbf{Z}_t^{(k)} & \text{for } t>1 \quad (B.5) \end{cases}$$

where

$$\mathbf{Z}_{t-1} = \begin{cases} \varphi_{t-1}^{(k)} - \mathbf{g}(\mathbf{X}_t^{(k)}) & \text{if } \mathbf{X}_t^{(k)} \in A_k(t), \\ 0 & \text{otherwise} \end{cases} \tag{B.6}$$

and

$$\begin{aligned} \mathbf{g}(\mathbf{X}_t^{(k)}) &= \mathbf{f}(\mathbf{X}_t^{(k)}) - \bar{\theta}_t^{(k)}, \\ \bar{\theta}_t^{(k)} &= \sum_{j=1}^{m} \beta_{kj}(t+1)\theta^{(k)}. \end{aligned} \tag{B.7}$$

We now proceed to verify the conditions (C1)–(C7) for $\varphi_t^{(k)}$. (C1) is satisfied, on account of (A9). (C2) holds, because of Equation (B.5).

By equations (B.6) and (B.7), we have

$$\begin{aligned} &E\left[\mathbf{Z}_t^{(k)}|\varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}\right] \\ &\quad = E\left[\varphi_t^{(k)} - \mathbf{g}(\mathbf{X}_{t+1}^{(k)})\middle|\varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}, A_k(t+1)\right], \end{aligned}$$

as $\mathbf{Z}_t^{(k)} = 0$ in $A_k^c(t+1)$,

$$= \varphi_t^{(k)} - E\left[\mathbf{g}(\mathbf{X}_{t+1}^{(k)})\middle|A_k(t+1)\right],$$

as $\mathbf{X}_{t+1}^{(k)}$ is independent of $\mathbf{X}_{(1)}^{(k)}, \mathbf{X}_{(2)}^{(k)}, \ldots, \mathbf{X}_{(t)}^{(k)}$ and hence $\varphi_1^{(k)}, \ldots, \varphi_t^{(k)}$,

$$= \varphi_t^{(k)},$$

since

$$\begin{aligned} &E\left[\mathbf{g}(\mathbf{X}_{t+1}^{(k)})\middle|A_k(t+1)\right] \\ &\quad = E\left[\mathbf{f}(\mathbf{X}_{t+1}^{(k)})\middle|A_k(t+1)\right] - \bar{\theta}_t^{(k)} \\ &\quad = \sum_{j=1}^{m} \beta_{kj}(t+1)E(\mathbf{X}|w=j) - \bar{\theta}_t^{(k)} \qquad \text{on account of Equation (4)} \\ &\quad = \sum_{j=1}^{mn} \beta_{kj}(t+1)\theta^{(k)} - \bar{\theta}_t^{(k)} \\ &\quad = 0. \end{aligned}$$

This verifies (C3) with $\mathbf{M}_t^{(k)}(\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in \mathbb{R}^N$. Also,

$$E\left[\|\mathbf{Z}_t^{(k)}\|^2 \mid \varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}\right]$$

$$= E\left[\left\|\varphi_t^{(k)} - \mathbf{g}\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2 \mid A_k(t+1)\right] \qquad \text{(for the same reason as before)}$$

$$= \left[\|\varphi_{(t)}^{(k)}\|^2 - 2\varphi_t^{(k)\prime}\left\{E\mathbf{g}\left(\mathbf{X}_{t+1}^{(k)}\right)\right\} + E\left\|\mathbf{g}\left(\mathbf{X}_{t-1}^{(k)}\right)\right\|^2\right]$$

$$\leqslant \|\varphi_{(t)}^{(k)}\|^2 + R,$$

R being a finite positive constant independent of $\varphi_1^{(k)}, \ldots, \varphi_t^{(k)}$,

since $E\mathbf{g}(\mathbf{X}_{t+1}^{(k)}) = \mathbf{0}$ (as seen above) in the subspace $A_k^{(t+1)}$, and

$$E \quad \left\|\mathbf{g}\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2$$

$$= E\left\|\mathbf{f}\left(\mathbf{X}_{t+1}^{(k)}\right) - \bar{\theta}_t^{(k)}\right\|^2$$

$$\leqslant E\left\|\mathbf{f}\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2 - \|\bar{\theta}_t^{(k)}\|^2 \qquad \text{as } E\mathbf{f}\left(\mathbf{X}_{t+1}^{(k)}\right) = \bar{\theta}_t^{(k)}$$

$$\leqslant E\left\|\mathbf{f}\left(\mathbf{X}_{t+1}^{(k)}\right)\right\|^2$$

$$= \sum_{j=1}^{m} \beta_{kj}(t+1)\rho_j, \qquad \text{by (A10)},$$

$$\leqslant \sum_{j=1}^{m} \rho_d = R, \qquad \text{say}.$$

Thus (C.4) holds with $a = R$, $b = 0$, $c = 1$. Finally, as

(i) $\mathbf{x}'\mathbf{M}_t^{(k)}(\mathbf{x}) = \mathbf{x}'\mathbf{x} \geqslant 0$,
(ii) $E[\|\varphi_1^{(k)}\|^2] < R < \infty$, as seen before,
(iii) $\inf_{\eta \leqslant \|\mathbf{x}\| \leqslant \eta^{-1}} \mathbf{x}'\mathbf{M}_t^{(k)}(\mathbf{x}) > \delta_k \eta^2 > 0$ because of (A11), the conditions (C5), (C6), and (C7) are respectively seen to be true. Hence the theorem. ■

REFERENCES

1. A. Pal (Pathak) and S. K. Pal, Effect of wrong samples on the convergence of learning process, *Inform. Sci.* 53:191–201 (1991).
2. Y. T. Chien, The threshold effect of a nonlinear learning algorithm for pattern recognition, *Inform. Sci.* 2:351–358 (1970).

3. S. K. Pal, A. K. Datta, and D. Dutta Majumder, A self-supervised vowel recognition system, *Patt. Recogn.* 12:27–34 (1980).
4. A. Pathak and S. K. Pal, A generalized learning algorithm based on guard zones, *Patt. Recogn. Lett.* 4:63–69 (1986).
5. S K. Pal, A. Pathak, and C. Basu, Dynamic guard zone for self-supervised learning, *Patt. Recog. Lett.* 7:135–144 (1988).
6. C. B. Chittineni, Learning with imperfectly labelled samples, *Patt. Recogn.* 12:281–291 (1980).
7. L. Schmetterer, Multidimensional stochastic approximation, in *Multivariate Analysis — II: Proc. 2nd Int. Symp. Multiv. Anal.*, Dayton, Ohio (P. R. Krishnaiah, Ed.), Academic, New York, 1968.
8. S. K. Pal and D. Dutta Majumder, Fuzzy Sets and decision making approaches in vowel and speaker recognition, *IEEE Trans. Syst. Man Cybern.*, SMC-7:625–629 (1977).
9. A. K. Datta, N. R. Ganguli, and S. Ray, Recognition of unaspirated plosives—a statistical approach, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-28:85–91 (1980).

Received 12 September 1988; revised 30 June 1989