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
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 (Consonant-vowel Nucleus-Consonant) context.

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

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

$$X_{nx1} = [x_1, x_2, \ldots, x_N], \quad X \in R^N.$$
where 
\[ \hat{\theta}_k^{(t)} \] is the \( t \)th stage estimate of \( \theta_k \), 
\[ \{a_i\} \] is a sequence of positive numbers, 
\[ f: R^N \rightarrow R^p \] is a continuous map, defining an unbiased statistic for \( \theta \), 
\[ \hat{\mu}_k^{(t-1)} \] is the \((t-1)\)th stage GAA estimate of \( \mu_k \), 
\[ G(\hat{\mu}_k^{(t-1)}, \lambda) = \{X: X \in R^N, d(X, \hat{\mu}_k^{(t-1)}) < \lambda \} \], 
\[ d^2(x, y) = (x - y)'B(x - y) \], 
\[ B_i \] is a symmetric positive definite matrix, which may or may not be a function of the training samples \( X_i^{(k)} \) (some examples given in [4]), 
\[ \lambda_i \] is a positive number, suitably chosen.

Incidentally, \( G(a, r) \) is the guard zone and, clearly, is nothing but a closed ball centered at \( a \) 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, \( \{a_i\}, \lambda_i \), and \( B_i \), 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 \( \delta \), or the corrections \( Y_i^{(k)} \) will be too small, otherwise.

Similarly, as also mentioned in [4], \( B_i \) can be chosen from among the following possible values, since \( d^2(\cdot, \cdot) \) is a distance function:
\[ B_i = I \quad \text{(the identity matrix of order } N) \] (2a)

or
\[ B_i = \left[ \text{Diag}(s_{11}^{(1)}, s_{22}^{(1)}, \ldots, s_{NN}^{(1)}) \right]^{-1} \] (2b)

or
\[ B_i = S_i^{-1} = \left( \left( \sigma_i^{(1)} \right) \right)^{-1}_{N \times N} \] (2c)

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,
\[ w, \hat{w} \in \{1, 2, \ldots, m\} = \Omega, \quad \text{say.} \]

Let \( \pi_k = P(w = k) \) denote the a priori probability for the class \( C_k, k = 1, \ldots, m \). Further, let \( p_k(X) = P(X|w = k) \) be the class-conditional probability density of the feature vector \( X \). Also, let \( \alpha_{kj} \) denote the probability that a sample from \( C_j \) is given the label \( k \), i.e.,
\[ \alpha_{kj} = P(\hat{w} = k|w = j), \quad j, k = 1, \ldots, m. \] (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 \( t \)th stage, i.e.
\[
\rho(X^t_k) = \rho(X|\hat{w} = k) = \rho(X|\hat{w} = k)
\]
\[
= \sum_{j=1}^{m} \beta_{kj}(t) \rho(X|w = j) \quad \text{if } X^t_k \in A_k(t), \quad \text{given } \hat{w} = k, \quad (4a)
\]
\[
= \sum_{j=1}^{m} \beta^*_kj(t) \rho(X|w = j), \quad \text{otherwise,} \quad (4b)
\]
where
\[
A_k(t) = \{ x : x \in G(\hat{\pi}_1^{(t)}, \ldots, \hat{\pi}_l^{(t)}) \},
\]
\[
\beta_{kj}(t) = P(A_k(t)|X, \hat{w} = k, w = j) \alpha_{kj}, \quad (4c)
\]
\[
\beta^*_kj(t) = P(A_k(t)|X, \hat{w} = k, w = j) \alpha_{kj}, \quad (4d)
\]
provided that we are prepared to assume:

(A6) \( p(X|\hat{w} = k, w = j) = p(X|w = j) \) for all \( j, k = 1, 2, \ldots, m \).

(A7) \( P(\hat{w} = k, A_k(t)) > 0 \) for all \( k, i \).

(A8) \( P(\hat{w} = k, A_k^*(t)) > 0 \) for all \( k, i \).

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

We have not studied the problem of estimating the mislabeling probabilities \( \beta_{kj} \) 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 \( \theta \), can be defined in various ways. For instance, we say that:

i. The sequence \( \{ \hat{\theta}_t \} \) converges to \( \theta \) with probability 1 or almost surely (in symbols, \( \hat{\theta}_t \to \theta \)) or strongly if
\[
P \left( \lim_{t \to \infty} E \| \hat{\theta}_t - \theta \| = 0 \right) = 1.
\]

ii. \( \hat{\theta}_t \) converges to \( \theta \) 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 Schmelterer [7]:

LEMMA 1. Let \( (a_n) \) be a sequence of positive real numbers such that
\[
\sum_{n=1}^{\infty} a_n^2 < \infty.
\]
Let \( x_n \) and \( y_n \) be \( N \)-dimensional random vectors which satisfy
\[
(1) \quad x_{n+1} = x_n - a_n y_n, \quad n \geq 1.
\]
Let \( M_n \) be a measurable mapping from \( R^N \) to \( R^N \) such that
\[
(2) \quad E(y_n|x_1, x_2, \ldots, x_n) = M_n(x_n) \quad \text{a.e.}
\]
Let \( a, b, c \) be nonnegative real numbers and let
\[
(3) \quad E(\|y_n\|^2|x_1, x_2, \ldots, x_n) \leq a + b\|x_n\| + c\|x_n\|^2 \quad \text{a.e.}
\]
Also, for every \( x \in R^N \) and \( n \geq 1 \),
\[
(4) \quad \|xM_n(x)\| \geq 0.
\]
\textbf{EFFECT OF WRONG SAMPLES}

Let us now state the following theorems.

\textbf{Theorem 1.} Consider the setup given in Sections 2 and 3. If, in addition to assumptions (A1)-(A5), we also have

\begin{align}
\text{(A9)} & \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty, \text{ where } \alpha_n > 0 \text{ for all } n \\
\text{(A10)} & \quad \rho_{il} = E(||x||^2|w = i) \text{ exists, with respect to each class-conditional density } \rho_i(X), \\
\text{(A11)} & \quad p_i^{(k)} = P[A_{ik}(t)|\hat{w} = k] > \delta_k \text{ for some } \delta_k \in (0,1) \text{ for all } t, \text{ then}
\end{align}

\begin{align}
\phi_{il}^{(k)} = \hat{\psi}_{il}^{(k)} - \sum_{j=1}^{m} \beta_{ij}^{(k)}(t) \theta_{il}^{(k)} > 0, \quad \text{the } N \text{-dimensional null vector.}
\end{align}

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

Here \(\beta_{jk}(t), k,j = 1,2,\ldots,m\) 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.

\textbf{Theorem 2.} For the setup in Sections 2 and 3, if assumptions (A9)-(A11) also hold, then

\begin{align}
\sum_{j=1}^{m} \gamma_{ik}^{(k)}(t) \hat{\theta}_{ij}^{(k)} \to \theta_{ij}^{(k)} \text{ as } t \to \infty \text{ for } k = 1,2,\ldots,m,
\end{align}

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

The proof is trivial.

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

\begin{align}
\beta_{jk}(t) = \delta_{jk}, \quad \text{the Kronecker delta, for all } k,j = 1,2,\ldots,m,
\end{align}

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

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

\begin{align}
\hat{\theta}_{ij}^{(k)} - \sum_{j=1}^{m} \beta_{jk}(t) \theta_{ij}^{(k)} \overset{a.s.}{\to} 0.
\end{align}

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

\begin{align}
\hat{\theta}_{ij}^{(k)} = f(X_{ik}^{(k)}) \quad \text{for } t = 1 \\
\hat{\theta}_{ij}^{(k)} - \hat{\theta}_{ij}^{(k-1)} - a_{ik} (\hat{\theta}_{ij}^{(k-1)} - f(X_{ik}^{(k)})) \quad \text{for } t > 1,
\end{align}

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

\textbf{RESULT 1.} Under the conditions (A9) and (A10), and the setup considered earlier,

\begin{align}
\hat{\theta}_{ij}^{(k)} - \sum_{j=1}^{m} \epsilon_{kj}^{(k)} \theta_{ij}^{(k)}
\end{align}

where

\begin{align}
\epsilon_{kj}^{(k)} = \pi_k \alpha_k \left( \sum_{i=1}^{m} \pi_i \alpha_{ki} \right).
\end{align}

In both cases, therefore, a.s. convergence takes place but in different forms. In the first situation, the sequence of estimates \(\{\hat{\theta}_{ij}^{(k)}\}\) converges strongly with another sequence \(\{\tilde{\theta}_{ij}^{(k)}\}, \text{ for } t = 1,2,\ldots\), where

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

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

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

where $\beta_{kj} > 0$.

\[(A13) \quad \text{either } \sum_{j=1}^{m} \epsilon_{kj} \theta_{jq} > \sum_{j=1}^{m} \beta_{kj} \theta_{jq} > \theta_{kj}\]

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

then

$$\|\hat{\theta}(k) - \theta(k)\| - \|\tilde{\theta}(k) - \theta(k)\| \overset{\text{A13}}{\to} C_k \text{ where } C_k > 0.$$

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

$\hat{\theta}(k) \overset{\text{A12}}{\to} \tilde{\theta}(k)$,

where

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

**EFFECT OF WRONG SAMPLES**

This, together with the result stated earlier, implies that

$$\hat{\theta}(k) - \theta(k) \overset{\text{A12}}{\to} \tilde{\theta}(k) - \theta(k)$$

and

$$\hat{\theta}(k) - \theta(k) \overset{\text{A12}}{\to} \tilde{\theta}(k) - \theta(k).$$

Consequently,

$$\|\hat{\theta}(k) - \theta(k)\| - \|\tilde{\theta}(k) - \theta(k)\| \overset{\text{A13}}{\to} \|\tilde{\theta}(k) - \theta(k)\| - \|\tilde{\theta}(k) - \theta(k)\|.$$

However,

$$\|\hat{\theta}(k) - \theta(k)\|^2 - \|\tilde{\theta}(k) - \theta(k)\|^2 = (\tilde{\theta} - \theta)^T(\tilde{\theta} - \theta) - 2(\tilde{\theta} - \theta)^T(\tilde{\theta} - \theta),$$

dropping superscripts, for convenience,

$$= (\tilde{\theta} - \theta)(\tilde{\theta} - \theta) + 2(\tilde{\theta} - \theta)(\tilde{\theta} - \theta) - 2(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)$$

$$= \|\tilde{\theta} - \theta\|^2 + 2(\tilde{\theta} - \theta)(\tilde{\theta} - \theta) > 0$$

because of (A13). Hence the theorem.

**Remarks.**

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

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

$$\tilde{\theta}(k) \overset{\text{A13}}{\to} \tilde{\theta}(k) > \theta(k)$$

or if

$$\theta(k) > \tilde{\theta}(k) > \hat{\theta}(k),$$
where the partial order relation > is defined as follows:

\[ \text{for } a, b \in \mathbb{R}^N, \quad a < b \text{ if } a_i < b_i \text{ for all } i = 1, \ldots, N. \]

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

\[ Q_L(y_0) = \{ y : y_i < y_{0i} \forall i = 1, 2, \ldots, N \} \].

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 \( F_3 \).

5.1. ARTIFICIAL DATA SET

For each of the three classes, the mean vectors and covariance 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 \( 3 \times 3 \) matrix of \( \alpha_{ij} \) 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 \( \overline{\mu}^{(k)} \) and \( \overline{\Sigma}^{(k)} \), (equivalent to the \( \theta^{(k)} \) defined in Section 4) for \( k = 1, 2, \) and \( 3 \).

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

i. \( \alpha = 1/t \) for all \( t > 1 \).

ii. Equal \textit{a priori} probabilities for all the classes.

iii. The Mahalanobis distance for the distance function \( d \).

iv. \( \alpha_{ij} = (1 - \alpha) L_i + \alpha L_j \), where \( L_i \) and \( L_j \) are respectively the lower and upper bounds for \( \alpha_{ij} \), obtained in [5], and \( \alpha \in (0, 1) \).

The optimum value of \( \alpha \) was empirically found to be 0.8. The estimates obtained, with this value of \( \alpha \), 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

<table>
<thead>
<tr>
<th>Sc No</th>
<th>Sample</th>
<th>True/1</th>
<th>1.00</th>
<th>0.00</th>
<th>Mean Vector</th>
<th>Covariance Matrix (Raw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.20</td>
<td>13.09</td>
<td>0.00</td>
<td>0.00</td>
<td>11.200</td>
<td>13.048</td>
</tr>
<tr>
<td>2</td>
<td>12.38</td>
<td>13.80</td>
<td>0.00</td>
<td>0.00</td>
<td>12.380</td>
<td>13.446</td>
</tr>
<tr>
<td>3</td>
<td>10.82</td>
<td>14.54</td>
<td>1.35</td>
<td>3.65</td>
<td>10.820</td>
<td>14.412</td>
</tr>
<tr>
<td>4</td>
<td>10.56</td>
<td>16.67</td>
<td>4.38</td>
<td>6.30</td>
<td>10.560</td>
<td>16.977</td>
</tr>
<tr>
<td>5</td>
<td>8.94</td>
<td>15.46</td>
<td>4.56</td>
<td>4.90</td>
<td>8.940</td>
<td>15.073</td>
</tr>
<tr>
<td>6</td>
<td>10.36</td>
<td>13.25</td>
<td>2.00</td>
<td>2.24</td>
<td>10.360</td>
<td>13.073</td>
</tr>
<tr>
<td>7</td>
<td>5.41</td>
<td>6.48</td>
<td>14.58</td>
<td>13.25</td>
<td>5.410</td>
<td>6.487</td>
</tr>
<tr>
<td>8</td>
<td>9.42</td>
<td>11.41</td>
<td>3.45</td>
<td>5.81</td>
<td>9.420</td>
<td>11.077</td>
</tr>
<tr>
<td>10</td>
<td>8.07</td>
<td>10.49</td>
<td>3.74</td>
<td>3.71</td>
<td>8.070</td>
<td>10.273</td>
</tr>
<tr>
<td>11</td>
<td>9.66</td>
<td>15.66</td>
<td>0.76</td>
<td>0.76</td>
<td>9.660</td>
<td>15.076</td>
</tr>
<tr>
<td>12</td>
<td>10.33</td>
<td>14.48</td>
<td>0.07</td>
<td>0.09</td>
<td>10.330</td>
<td>14.412</td>
</tr>
<tr>
<td>14</td>
<td>9.41</td>
<td>14.72</td>
<td>0.74</td>
<td>0.66</td>
<td>9.410</td>
<td>14.661</td>
</tr>
<tr>
<td>15</td>
<td>10.57</td>
<td>12.82</td>
<td>1.48</td>
<td>1.29</td>
<td>10.570</td>
<td>12.651</td>
</tr>
<tr>
<td>16</td>
<td>12.20</td>
<td>14.37</td>
<td>1.72</td>
<td>1.43</td>
<td>12.200</td>
<td>14.080</td>
</tr>
<tr>
<td>17</td>
<td>11.99</td>
<td>16.39</td>
<td>0.65</td>
<td>2.36</td>
<td>11.990</td>
<td>16.215</td>
</tr>
<tr>
<td>18</td>
<td>13.60</td>
<td>16.48</td>
<td>1.83</td>
<td>3.34</td>
<td>13.600</td>
<td>16.341</td>
</tr>
<tr>
<td>19</td>
<td>11.03</td>
<td>12.96</td>
<td>1.53</td>
<td>1.20</td>
<td>11.030</td>
<td>12.341</td>
</tr>
<tr>
<td>20</td>
<td>11.51</td>
<td>16.42</td>
<td>0.42</td>
<td>1.83</td>
<td>11.510</td>
<td>16.445</td>
</tr>
</tbody>
</table>

TABLE 2(a) Continued.

Distances from True Parameter Values of the GGA Estimates

<table>
<thead>
<tr>
<th>Means</th>
<th>Distances</th>
<th>Non-GGA Estimation of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiv.</td>
<td>Average</td>
<td>Indiv.</td>
</tr>
<tr>
<td>1.204</td>
<td>1.204</td>
<td>28.656</td>
</tr>
<tr>
<td>1.872</td>
<td>1.576</td>
<td>47.023</td>
</tr>
<tr>
<td>1.581</td>
<td>1.576</td>
<td>40.412</td>
</tr>
<tr>
<td>1.266</td>
<td>1.589</td>
<td>29.608</td>
</tr>
<tr>
<td>0.781</td>
<td>1.387</td>
<td>18.100</td>
</tr>
<tr>
<td>0.781</td>
<td>1.387</td>
<td>18.100</td>
</tr>
</tbody>
</table>

Distances from True Parameter Values of the Non-GGA Estimates of Means

<table>
<thead>
<tr>
<th>Means</th>
<th>Distances</th>
<th>GGA Estimations of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiv.</td>
<td>Average</td>
<td>Indiv.</td>
</tr>
<tr>
<td>1.204</td>
<td>1.204</td>
<td>28.656</td>
</tr>
<tr>
<td>1.872</td>
<td>1.576</td>
<td>47.023</td>
</tr>
<tr>
<td>1.581</td>
<td>1.576</td>
<td>40.412</td>
</tr>
<tr>
<td>1.266</td>
<td>1.589</td>
<td>29.608</td>
</tr>
<tr>
<td>0.781</td>
<td>1.387</td>
<td>18.100</td>
</tr>
<tr>
<td>0.781</td>
<td>1.387</td>
<td>18.100</td>
</tr>
</tbody>
</table>
### Table 2(b)

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

<table>
<thead>
<tr>
<th>Ser. No.</th>
<th>Sample</th>
<th>True d</th>
<th>Mean Vector</th>
<th>Covariance Matrix (Raw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.86</td>
<td>4.42</td>
<td>0.42 1.83</td>
<td>4.862 4.416</td>
</tr>
<tr>
<td>2</td>
<td>5.58</td>
<td>4.54</td>
<td>0.42 1.83</td>
<td>5.120 4.476</td>
</tr>
<tr>
<td>3</td>
<td>6.19</td>
<td>4.72</td>
<td>0.42 1.32</td>
<td>5.476 4.560</td>
</tr>
<tr>
<td>4</td>
<td>5.37</td>
<td>4.02</td>
<td>0.16 2.59</td>
<td>5.957 4.360</td>
</tr>
<tr>
<td>5</td>
<td>0.51</td>
<td>0.19</td>
<td>0.28 1.46</td>
<td>4.835 4.487</td>
</tr>
<tr>
<td>6</td>
<td>2.57</td>
<td>4.42</td>
<td>0.93 1.93</td>
<td>4.465 4.454</td>
</tr>
<tr>
<td>7</td>
<td>4.17</td>
<td>2.86</td>
<td>0.92 0.84</td>
<td>4.149 4.454</td>
</tr>
<tr>
<td>8</td>
<td>2.27</td>
<td>6.19</td>
<td>2.20 2.79</td>
<td>1.926 4.671</td>
</tr>
<tr>
<td>9</td>
<td>6.88</td>
<td>3.11</td>
<td>2.12 2.58</td>
<td>4.265 4.475</td>
</tr>
<tr>
<td>10</td>
<td>3.04</td>
<td>0.57</td>
<td>3.00 3.55</td>
<td>4.265 4.497</td>
</tr>
<tr>
<td>11</td>
<td>4.37</td>
<td>4.07</td>
<td>0.13 0.99</td>
<td>4.265 4.497</td>
</tr>
<tr>
<td>12</td>
<td>9.78</td>
<td>5.27</td>
<td>1.67 4.56</td>
<td>4.717 4.812</td>
</tr>
<tr>
<td>13</td>
<td>7.96</td>
<td>5.73</td>
<td>2.73 19.77</td>
<td>4.966 4.832</td>
</tr>
<tr>
<td>14</td>
<td>2.98</td>
<td>5.00</td>
<td>2.36 14.93</td>
<td>4.243 4.951</td>
</tr>
<tr>
<td>15</td>
<td>4.50</td>
<td>8.89</td>
<td>1.40 11.82</td>
<td>4.824 4.851</td>
</tr>
<tr>
<td>16</td>
<td>4.26</td>
<td>5.44</td>
<td>2.10 2.31</td>
<td>4.789 4.925</td>
</tr>
<tr>
<td>17</td>
<td>4.25</td>
<td>9.97</td>
<td>1.55 13.91</td>
<td>4.789 4.925</td>
</tr>
<tr>
<td>18</td>
<td>4.99</td>
<td>3.09</td>
<td>2.61 5.05</td>
<td>4.789 4.925</td>
</tr>
<tr>
<td>19</td>
<td>4.32</td>
<td>1.54</td>
<td>1.86 9.37</td>
<td>4.789 4.925</td>
</tr>
<tr>
<td>20</td>
<td>3.23</td>
<td>1.67</td>
<td>1.26 10.62</td>
<td>4.789 4.925</td>
</tr>
</tbody>
</table>

### Table 2(b) Continued.

Distance from True Parameter Values of the GGA Estimates.

<table>
<thead>
<tr>
<th>Means</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.665</td>
<td>0.665</td>
</tr>
</tbody>
</table>

### Table 2(b) Continued.

Distance from True Parameter Values of the Non-GGA Estimates.

<table>
<thead>
<tr>
<th>Means</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.665</td>
<td>0.665</td>
</tr>
</tbody>
</table>

### Table 2(b) Continued.

EFFECT OF WRONG SAMPLES


**TABLE 2(e)**

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

<table>
<thead>
<tr>
<th>Ser No.</th>
<th>Sample</th>
<th>True Value of Class -1-1</th>
<th>Lambda</th>
<th>Upon ( \Delta )</th>
<th>Mean Vector</th>
<th>Covariance Matrix (Raw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.70</td>
<td>8.34</td>
<td>3</td>
<td>12.64 10.62</td>
<td>-6.305</td>
<td>38.6581</td>
</tr>
<tr>
<td>2</td>
<td>4.72</td>
<td>8.25</td>
<td>3</td>
<td>12.64 10.62</td>
<td>4.511</td>
<td>38.6581</td>
</tr>
<tr>
<td>3</td>
<td>3.59</td>
<td>11.43</td>
<td>3</td>
<td>9.00   15.63</td>
<td>4.303</td>
<td>38.6581</td>
</tr>
<tr>
<td>4</td>
<td>4.56</td>
<td>11.88</td>
<td>3</td>
<td>8.46   15.74</td>
<td>4.293</td>
<td>42.6926</td>
</tr>
<tr>
<td>5</td>
<td>6.01</td>
<td>9.01</td>
<td>3</td>
<td>6.34   8.88</td>
<td>4.836</td>
<td>43.9790</td>
</tr>
<tr>
<td>6</td>
<td>2.36</td>
<td>13.24</td>
<td>3</td>
<td>2.13   5.27</td>
<td>4.457</td>
<td>21.0654</td>
</tr>
<tr>
<td>7</td>
<td>2.61</td>
<td>12.16</td>
<td>1</td>
<td>3.59   4.79</td>
<td>4.557</td>
<td>21.0654</td>
</tr>
<tr>
<td>8</td>
<td>5.03</td>
<td>7.03</td>
<td>2</td>
<td>1.82   3.38</td>
<td>4.125</td>
<td>99.9284</td>
</tr>
<tr>
<td>9</td>
<td>8.67</td>
<td>10.02</td>
<td>1</td>
<td>41.91  44.10</td>
<td>4.989</td>
<td>28.6738</td>
</tr>
<tr>
<td>10</td>
<td>2.48</td>
<td>10.92</td>
<td>4.04   16.02</td>
<td>4.728</td>
<td>28.6738</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2.34</td>
<td>9.12</td>
<td>12.59  12.30</td>
<td>4.728</td>
<td>28.6738</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5.58</td>
<td>11.53</td>
<td>11.95  10.22</td>
<td>4.958</td>
<td>44.3773</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5.50</td>
<td>10.34</td>
<td>6.92   4.88</td>
<td>4.728</td>
<td>28.6738</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5.53</td>
<td>9.07</td>
<td>18.84  26.01</td>
<td>5.080</td>
<td>30.4727</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>6.71</td>
<td>10.64</td>
<td>8.80   7.49</td>
<td>5.080</td>
<td>30.4727</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>4.27</td>
<td>7.62</td>
<td>10.55  9.40</td>
<td>5.040</td>
<td>30.4727</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4.66</td>
<td>11.12</td>
<td>4.60   5.54</td>
<td>5.088</td>
<td>30.0947</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>6.57</td>
<td>11.69</td>
<td>10.00  8.92</td>
<td>5.006</td>
<td>30.0947</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4.05</td>
<td>6.46</td>
<td>11.47  12.80</td>
<td>5.062</td>
<td>29.8914</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>6.62</td>
<td>10.83</td>
<td>10.78  8.85</td>
<td>5.052</td>
<td>29.8914</td>
<td></td>
</tr>
</tbody>
</table>

**EFFECT OF WRONG SAMPLES**

<table>
<thead>
<tr>
<th>TABLE 2(e). Continued.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance from True Parameter Values of the GGA Estimates</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Non-GGA Estimates</td>
</tr>
<tr>
<td>Index Average</td>
</tr>
<tr>
<td>1.648</td>
</tr>
<tr>
<td>1.648</td>
</tr>
<tr>
<td>1.648</td>
</tr>
<tr>
<td>1.648</td>
</tr>
</tbody>
</table>
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 $\tilde{\mu}_{(k)}$ and $\Sigma_{(k)}$ 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 $\tilde{\mu}_{(k)}$ is greater than $\mu_{(k)}$, $\Sigma_{(k)}$ is less than $\Sigma_{(k)}$. 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 frequencies ($F_1$, $F_2$, and $F_3$) at the 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
EFFECT OF WRONG SAMPLES

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 \( \alpha \) 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 \( \alpha \) 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

\[
p(X|\hat{w} = k) = p(X|\hat{w} = k, A_k(t)) P(A_k(t)|\hat{w} = k)
+ p(X|\hat{w} = k, A'_k(t)) P(A'_k(t)|\hat{w} = k),
\]

where \( A'_k(t) \) stands for the event complementary to \( A_k(t) \). However,

\[
p(X|\hat{w} = k, A_k(t)) = \frac{p(X, \hat{w} = k, A_k(t))}{P(A_k(t), \hat{w} = k)}
= \frac{\sum_{m=1}^{m-1} p(X, \hat{w} = k, A_k(t), w = j)}{P(A_k(t), \hat{w} = k)}
= \frac{\sum_{m=1}^{m-1} P(A_k(t)|X, \hat{w} = k, w = j) p(X|\hat{w} = k, w = j) P(\hat{w} = k, w = j)}{P(A_k(t), \hat{w} = k)}
= \frac{\sum_{m=1}^{m-1} P(A_k(t)|X, \hat{w} = k, w = j) \alpha_{ej} \tau_{ej} p(X|\hat{w} = k, w = j)}{P(A_k(t)|\hat{w} = k) P(\hat{w} = k)}
= \frac{\sum_{m=1}^{m-1} \beta_{ej}(t) \tau_{ej} p(X|w = j)}{P(A_k(t)|\hat{w} = k)}
\]

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.
by assumption (A6). Similarly, we must have

\[ p(X|\phi = k, A_k^l(t)) = \frac{\sum_{j=1}^{n} \beta_\phi^l(t) \rho(X|\omega = j) \rho(A_k^l(t)|\omega = k)}{P(A_k^l(t)|\omega = k)} \]  

(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) \( \psi^{(k)} \to 0 \) with probability 1 as \( t \to \infty \), \( \forall k \)

(ii) \( \{E[||\psi^{(k)}||^2]\} \) converges as \( t \to \infty \), \( \forall k \), where \( \psi^{(k)} = \hat{\theta}^{(k)} - \bar{\theta} \).

These, in turn, follow immediately from Lemmas 1 and 2 if it can be shown that the conditions (C1)-(C7) hold with \( x_n = \psi^{(k)} \).

We first note that

\[ \hat{\theta}^{(k)} = \begin{cases} f(X^{(k)}) & \text{for } t = 1, \\ \hat{\theta}^{(k)} - \frac{1}{t} Y^{(k)} & \text{for } t > 1. \end{cases} \]  

(B.1)

where

\[ Y^{(k)}_{t-1} = \begin{cases} \hat{\theta}^{(k)} - f(X^{(k)}) & \text{if } X^{(k)}(t) \in A_k(t) \\ 0 & \text{otherwise}. \end{cases} \]  

(B.2)

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

Obviously, therefore,

\[ \psi^{(k)} = \begin{cases} g(X^{(k)}) & \text{for } t = 1 \\ \psi^{(k)} - \frac{1}{t} Z^{(k)} & \text{for } t > 1. \end{cases} \]  

(B.4)

where

\[ Z^{(k)} = \begin{cases} \psi^{(k)} - g(X^{(k)}) & \text{if } X^{(k)}(t) \in A_k(t) \\ 0 & \text{otherwise}. \end{cases} \]  

(B.5)

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

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

\[ E[Z^{(k)}|\psi^{(k)}] = E[g(X^{(k)})|\psi^{(k)}, A_k(t+1)], \]

as \( Z^{(k)} = 0 \) in \( A_k(t+1) \).

\[ = \psi^{(k)} - E[g(X^{(k)})|A_k(t+1)], \]

as \( X^{(k)}_{t+1} \) is independent of \( X^{(k)}_1, X^{(k)}_2, \ldots, X^{(k)}_t \) and hence \( \psi^{(k)}, \ldots, \psi^{(k)}, \psi^{(k)} \),

\[ = \psi^{(k)}, \]

since

\[ E[g(X^{(k)}_{t+1})|A_k(t+1)] = E[f(X^{(k)}_1)|A_k(t+1)] - \bar{\theta}^{(k)} \]

and

\[ g(X^{(k)}_i) = f(X^{(k)}_i) - \bar{\theta}^{(k)}, \]

(B.7)

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

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

\[ E[Z^{(k)}|\psi^{(k)}] = E[g(X^{(k)})|\psi^{(k)}, A_k(t+1)], \]

as \( Z^{(k)} = 0 \) in \( A_k(t+1) \).

\[ = \psi^{(k)} - E[g(X^{(k)})|A_k(t+1)], \]

as \( X^{(k)}_{t+1} \) is independent of \( X^{(k)}_1, X^{(k)}_2, \ldots, X^{(k)}_t \) and hence \( \psi^{(k)}, \ldots, \psi^{(k)}, \psi^{(k)} \),

\[ = \psi^{(k)}, \]

since

\[ E[g(X^{(k)}_{t+1})|A_k(t+1)] = E[f(X^{(k)}_1)|A_k(t+1)] - \bar{\theta}^{(k)} \]

and

\[ g(X^{(k)}_i) = f(X^{(k)}_i) - \bar{\theta}^{(k)}, \]

(B.7)

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

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

\[ E[Z^{(k)}|\psi^{(k)}] = E[g(X^{(k)})|\psi^{(k)}, A_k(t+1)], \]

as \( Z^{(k)} = 0 \) in \( A_k(t+1) \).

\[ = \psi^{(k)} - E[g(X^{(k)})|A_k(t+1)], \]

as \( X^{(k)}_{t+1} \) is independent of \( X^{(k)}_1, X^{(k)}_2, \ldots, X^{(k)}_t \) and hence \( \psi^{(k)}, \ldots, \psi^{(k)}, \psi^{(k)} \),

\[ = \psi^{(k)}, \]

since

\[ E[g(X^{(k)}_{t+1})|A_k(t+1)] = E[f(X^{(k)}_1)|A_k(t+1)] - \bar{\theta}^{(k)} \]

and

\[ g(X^{(k)}_i) = f(X^{(k)}_i) - \bar{\theta}^{(k)}, \]

(B.7)

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

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

\[ E[Z^{(k)}|\psi^{(k)}] = E[g(X^{(k)})|\psi^{(k)}, A_k(t+1)], \]

as \( Z^{(k)} = 0 \) in \( A_k(t+1) \).

\[ = \psi^{(k)} - E[g(X^{(k)})|A_k(t+1)], \]

as \( X^{(k)}_{t+1} \) is independent of \( X^{(k)}_1, X^{(k)}_2, \ldots, X^{(k)}_t \) and hence \( \psi^{(k)}, \ldots, \psi^{(k)}, \psi^{(k)} \),

\[ = \psi^{(k)}, \]

since

\[ E[g(X^{(k)}_{t+1})|A_k(t+1)] = E[f(X^{(k)}_1)|A_k(t+1)] - \bar{\theta}^{(k)} \]

and

\[ g(X^{(k)}_i) = f(X^{(k)}_i) - \bar{\theta}^{(k)}, \]

(B.7)
This verifies (C3) with $M^{(k)}(x) = x$, $\forall x \in \mathbb{R}^N$. Also,

$$E \left[ \|Z^{(k)}_i\|^2 \right] = E \left[ \|\psi^{(k)}_i - \mu(X^{(k)}_{i+1})\|^2 + A_k(t+1) \right]$$

(for the same reason as before)

$$= \|\psi^{(k)}_i\|^2 + 2\psi^{(k)}_i \{ E\mu(X^{(k)}_{i+1}) \} + E\|\mu(X^{(k)}_{i+1})\|^2$$

$$\leq \|\psi^{(k)}_i\|^2 + R,$$

$R$ being a finite positive constant independent of $\psi^{(k)}_1, \ldots, \psi^{(k)}_i$, since $E\mu(X^{(k)}_{i+1}) = 0$ (as seen above) in the subspace $A^{(i+1)}$, and

$$E \left[ \|f(X^{(k)}_{i+1})\|^2 \right]$$

$$= E \left[ \|f(X^{(k)}_{i+1}) - \bar{\theta}^{(k)}_i\|^2 \right]$$

$$\leq E \left[ \|f(X^{(k)}_{i+1})\|^2 - \|\bar{\theta}^{(k)}_i\|^2 \right] \quad \text{as } Ef(X^{(k)}_{i+1}) = \bar{\theta}^{(k)}$$

$$\leq E \left[ \|f(X^{(k)}_{i+1})\|^2 \right]$$

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

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

Thus (C4) holds with $a = R$, $b = 0$, $c = 1$. Finally, as

(i) $\langle x' M^{(k)}_X x \rangle = x' x > 0$,

(ii) $E[\|\psi^{(k)}_i\|^2] < R < \infty$, as seen before,

(iii) $\inf_{\|x\| = \eta} \langle x' M^{(k)}_X x \rangle > \delta \eta^2 > 0$ because of (A11), the conditions (C5), (C6), and (C7) are respectively seen to be true. Hence the theorem. ■

REFERENCES
