Pattern Recognition, Vol. 20, No. 2, pp. 237-244, 1987.
Printed in Great Britain.

0031-3203/87 $3.00 + .00
Pergamon Journals Ltd.

ON THE CONVERGENCE OF "A SELF-SUPERVISED VOWEL RECOGNITION SYSTEM"

AMITA PATHAK and SANKAR K. PAL
Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700035, India

(*Received* 5 *December* 1985; *in revised form* 10 *June* 1986)

Abstract — A self-supervised learning algorithm based on the concept of guard zones was developed by Pal *et al.*[1] for studying the adaptive ability of a recognition system, starting with non-appropriate representative vectors. Guard zones were used to discard unreliable (doubtful) samples from the parameter-updating programme, so that the convergence does not get affected. The algorithm was implemented with success on speech data but no proof of convergence was provided.

The present paper investigates the convergence of this algorithm, using some results on multidimensional stochastic approximation. It is shown that the estimates of the parameters converge strongly to their true values under certain conditions provided the guard zones are effective in discarding mislabelled training samples.

Learning   Guard-zone   Convergence   Stochastic approximation

1. INTRODUCTION

The present work is in connection with the earlier report[1] in which a self-supervised recognition system was developed using the concept of guard zones.

The guard zones are of ellipsoidal shape with dimensions being proportional to the respective standard deviation of features. These were described around the reference vectors of the classes in order to make a restricted updating programme for estimating the class parameters.

For the purpose of supervision, it is assumed that for an input vector falling within the guard zone, the probability of its being misclassified is so low that it would not affect the convergence property of the system in any significant way. The supervisory system, therefore, needs only to check whether the classified input is within the guard zone or not for the purpose of inhibition of the updating programme.

The effectiveness of the adaptive system was demonstrated with success on a set of 871 Vowel Sounds in CNC (Consonant-Vowel Nucleus-Consonant) context with first three vowel formants as features, and non-appropriate initial representative vectors. The representative vector of a vowel class was deliberately chosen just outside the boundary of an ellipsoid having the three axes equal to the respective standard deviations of the features and mean of the classes as the centre. The purpose was to study the adaptive ability of the system in recognizing vowel sounds starting with non-appropriate prototypes. The method used a single pattern training procedure for learning, and maximum value of fuzzy membership function was the basis of recognition. As the system used some inherent properties of the distribution of the same parameters (mean and variance) as used by the classifier itself, it may be called a "self-supervisory" system. The experimental results corroborated the theoretical postulates that such system would basically approach the supervised learning algorithm in so far as the convergence properties are concerned. The system had been found to approach, for certain dimensions of guard zone, the performance of a fully-supervised system which use an extra higher level of knowledge.

In this paper we have investigated theoretically the convergence of this system, and have been able to show that under certain conditions the estimates of the parameters converge strongly to their true values if the guard-zones succeed in weeding out the "wrong" training samples. For this purpose, we have made use of some results on multidimensional stochastic approximation procedures and probability theory. It is to be noted that a training sample is being dubbed "wrong" for updating the parameters of a given class if it is not really a sample from the class but has been assigned to it because of "mislabelling".

2. THE RECOGNITION SYSTEM[1]

Let

$$\mathbf{X} = [x_1, x_2, \ldots, x_N]', \quad \mathbf{X} \in \mathbb{R}^N$$

be an N-dimensional feature vector for a pattern recognition problem of discriminating between m pattern classes $C_1, C_2, \ldots, C_m$. It is assumed that

(A1) the feature vector $\mathbf{X}$ exhibits central tendency in each class C_j, about some point $\bar{\mathbf{X}}_j$ in it, $j = 1(1)m$,

(A2) the feature vector $\mathbf{X}$ admits of second-order moments in each class, with

$$\mathrm{var}(x_n \mid \mathbf{X} \in C_j) = \sigma_{nn}^{(j)}, \; n = 1(1)N, \; j = 1(1)m,$$

(A3) the pattern classes $C_1, C_2, \ldots, C_m$ have ill-defined boundaries, that is, each pattern class is a fuzzy subset of $\mathbb{R}^N$, with corresponding grade of membership $\mu_j(\mathbf{X})$ for any $\mathbf{X} \in \mathbb{R}^N$, where

$$\mu_j(X) \in [0, 1], \; j = 1(1)m.$$

2.1. *The decision rule*

The grade of membership of a pattern with feature vector $\mathbf{X}$, in C_j, $j = 1(1)m$, is defined as

$$\mu_j(\mathbf{X}) = \left(1 + \left[\frac{d(\mathbf{X}, \mathbf{R}_j)}{F_d}\right]^{F_e}\right)^{-1}, \tag{1a}$$

where F_e is the exponential fuzzifier, F_d is the denominational fuzzifier, $\mathbf{R}_j$ is a reference vector for the jth class C_j, and

$$d(\mathbf{X}, \mathbf{R}_j) = \min_l \|\mathbf{X} - \mathbf{R}_j^{(l)}\|, \tag{1b}$$

$\mathbf{R}_j^{(l)}$, $l = 1(1)h_j$ being a set of h_j prototypes from C_j, $j = 1(1)m$, with

$$\|\mathbf{X} - \mathbf{R}_j^{(l)}\| = \left(\sum_{n=1}^{N}\left[\frac{x_n - r_n^{(l)}}{\sigma_{jn}^{(l)}}\right]^2\right)^{0.5} \tag{1c}$$

where $\mathbf{R}_j^{(l)}$ is taken to be equal to $\bar{\mathbf{X}}_j^{(l)} = [\bar{x}_{j1}^{(l)}, \ldots, \bar{x}_{jN}^{(l)}]'$; $\bar{x}_{jn}^{(l)}$ and $\sigma_{jn}^{(l)}$ correspond to the lth prototype and denote respectively the mean and the standard deviation of the nth feature in the jth class.

Note that $\|\mathbf{X} - \mathbf{R}_j^{(l)}\|$ is the weighted Euclidean distance between $\mathbf{X}$ and $\mathbf{R}_j^{(l)}$ with weights inversely proportional to $\sigma_{jn}^{(l)}$.

A decision rule based on the μ_j-values is as follows: for an unknown pattern with feature vector X, classify it into C_k if $\mu_k(\mathbf{X}) > \mu_j(\mathbf{X})$, $j, k = 1(1)m$, $j \neq k$.

This classified sample is then used as training sample for estimating the parameters of the kth class provided the decision is accepted by the supervisor (described below).

2.2. *Iterative algorithm for parameter estimation*

The components of the reference vector and weight vector for each class, used in the decision rule above, may not be known *a priori* and thus will need to be learned. That is, it may be required to learn $\bar{\mathbf{X}}_j$ and σ_j, $j = 1(1)m$, where

$$\sigma_j = [\sigma_{11}^{(j)}, \sigma_{22}^{(j)}, \ldots, \sigma_{NN}^{(j)}]', \quad j = 1(1)m.$$

Let $\mathbf{X}_{(1)}^{(k)}, \mathbf{X}_{(2)}^{(k)}, \ldots$ be a sequence of learning samples for the class C_k. These are assumed to be independently distributed. Let $\bar{x}_{n(t)}^{(k)}$ and $s_{n(t)}^{(k)}$ be the estimates obtained, of the mean and the variance respectively of the nth feature x_n, by means of the first t training samples. (Subsequently, we shall not be using in many places any suffices to denote classes, wherever there is no scope for confusion.)

The "Decision Parameter of the Supervisor" (DPS) which restricts the updating programme, is defined for the kth class as

$$(\mathrm{DPS})_k = \sum_{n=1}^{N} [(x_n - \bar{x}_n^{(k)})/\hat{\sigma}_{kn}]^2 \tag{2}$$

where $\hat{\sigma}_{kn} = \sqrt{\hat{\sigma}_{nn}^{(k)}}/\lambda$, λ being a positive constant, termed the "zone-controlling parameter", as it controls the dimensions of the hyperellipsoidal regions

$$G_k = \{\mathbf{x} \mid (\mathrm{DPS})_k \leq 1\}, \quad k = 1(1)m,$$

where G_k is the guard zone for the kth class. ($\hat{\sigma}_{kn}$ is some estimate of $\sigma_{nn}^{(k)}$.) Let $c_{n(t)}^{(k)}$ denote the t-th stage estimate of the second-order raw moment for the kth class.

The learning algorithm is as follows – for $k = 1(1)m$ and $n = 1(1)N$,

$$\bar{x}_{n(t)}^{(k)} = x_{n(t)}^{(k)} \tag{3a}$$

$$c_{n(t)}^{(k)} = [x_{n(t)}^{(k)}]^2 \tag{3b}$$

$$s_{n(t)}^{(k)} = 0, \tag{3c}$$

when $t = 1$.

For $t > 1$,

$$\bar{x}_{n(t)}^{(k)} = \frac{t-1}{t}\bar{x}_{n(t-1)}^{(k)} + \frac{1}{t}x_{n(t)}^{(k)} \tag{4a}$$

$$c_{n(t)}^{(k)} = c_{n(t-1)}^{(k)} + [x_{n(t)}^{(k)}]^2 \tag{4b}$$

$$s_{n(t)}^{(k)} = \frac{1}{t}c_{n(t)}^{(k)} - [\bar{x}_{n(t)}^{(k)}]^2 \tag{4c}$$

provided $\mathrm{DPS}_k(t) \leq 1$, i.e. if $\mathbf{X}_{(t)}^{(k)}$ falls within the guard zone for the class C_k.

If not, no updating is done for the parameters of the class, i.e.

$$\bar{x}_{n(t)}^{(k)} = \bar{x}_{n(t-1)}^{(k)} \tag{5a}$$

$$c_{n(t)}^{(k)} = c_{n(t-1)}^{(k)} \tag{5b}$$

$$s_{n(t)}^{(k)} = s_{n(t-1)}^{(k)} \tag{5c}$$

when $\mathrm{DPS}_k(t) > 1$.

Of course, $\mathrm{DPS}_k(t)$ is a suitably modified form of equation (2), i.e.

$$\mathrm{DPS}_k(t) = \sum_{n=1}^{N} [(x_{n(t)}^{(k)} - \bar{x}_{n(t-1)})/d_{n(t-1)}^{(k)}]^2$$

where $d_{n(t-1)}^{(k)} = \sqrt{s_{n(t-1)}^{(k)}}/\lambda$, λ being as before.

2.3. *A model for mislabelled training samples*

The recognition system under consideration is such that the labels of the training samples are determined by the classifier itself, during the training period. Hence it is quite reasonable to expect that a certain proportion of training samples for each class have wrong labels. Moreover, these proportions are modified in some way by the supervisor. Let us therefore assume a simple model, inspired by one assumed by Chittineni,[3] for describing the situation resulting from the joint behaviour of the classifier and the supervisor.

Let $A_k(t)$ denote the event that the guard zone for C_k, based on $(t-1)$th stage estimates, accepts an observation (given the label k) for updating the estimates of the parameters of C_k, i.e.

$$A_k(t) = \{\mathbf{X} \mid \mathrm{DPS}_k(t, \mathbf{X}) \leq 1\}. \quad (6a)$$

Chittineni's model for labelling errors is specified as follows. Let w and $\hat{w}$ denote respectively the true and the given labels. Clearly,

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

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

$$\alpha_{kj} = P(\hat{w} = k \mid w = j), \quad j, k = 1(1)m. \quad (6b)$$

Clearly,

$$\sum_{k=1}^{m} \alpha_{kj} = 1. \quad (6c)$$

Under this model, it can be shown that

$$p(\mathbf{X}_{(t)}^{(k)}) = p(\mathbf{X}_{(t)} \mid \hat{w} = k)$$

$$= \begin{cases} \sum_{j=1}^{m} \beta_{kj}(t)\, p(\mathbf{X} \mid w = j)/P(A_k(t) \mid \hat{w} = k) & \text{if } \mathbf{X}_{(t)} \in (A_k(t) \mid \hat{w} = k) \quad (6d) \\ \sum_{j=1}^{m} \beta^*_{kj}(t)\, p(\mathbf{X} \mid w = j)/P(A^c_k(t) \mid \hat{w} = k) & \text{if } \mathbf{X}_{(t)} \in (A^c_k(t) \mid \hat{w} = k) \quad (6e) \end{cases}$$

where

$$\beta_{kj}^{(t)} = \frac{P(A_k(t) \mid \mathbf{X}, \hat{w} = k, w = i)}{P(\hat{w} = k, A_k(t))} \alpha_{kj} \pi_j \quad (6f)$$

$$\beta^*_{kj}(t) = \frac{P(A^c_k(t) \mid \mathbf{X}, \hat{w} = k, w = j)}{P(\hat{w} = k, A_k(t))} \alpha_{kj} \pi_j, \quad (6g)$$

provided we are prepared to assume

(A4) $p(\mathbf{X} \mid \hat{w} = k, w = j) = p(\mathbf{X} \mid w = j)\ \forall j, k = 1, 2, \ldots, m$

(A5) $P(\hat{w} = k, A_k(t)) \neq 0 \quad \forall k$ and t

(A6) $P(\hat{w} = k, A_k(t)) \neq 0 \quad \forall k$ and t.

For a proof of equation (6d), see Appendix A.

Also, as noted in Appendix B, finite upper bounds M and M^*, both ≥ 0, exist such that

$$\beta_{kj}(t) \leq M$$

and

$$\beta^*_{kj}(t) \leq M^* \quad \text{for all } k, j = 1(1)m, \text{ and for all } t.$$

3. CONVERGENCE OF THE SYSTEM

The convergence of the recognition system will depend upon the convergence of the learning algorithm. The convergence of learning algorithms can be defined in various ways.[3] For instance, for the problem of estimating θ sequentially by θ_t, we say that

(i) the sequence $\{\hat{\theta}_t\}$ *converges* to θ with *probability* 1 or *almost surely*, if

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

P being the probability measure,

i.e. if $\hat{\theta}_t \xrightarrow{\text{a.s.}} \theta$.

(ii) $\{\hat{\theta}_t\}$ *converges* to θ in the *mean-square* sense if

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

E being the expectation operator.

For the learning algorithm given in Section 2.2, the following theorems can be proved.

Theorem 1. Let $\bar{x}_{n(t)}^{(k)}$ and $c_{n(t)}^{(k)}$ be as in equations (3), (4) and (5), $k = 1(1)m$, $t \geq 1$.

Let

$$\hat{\theta}_t^{(k)} = [\bar{x}_{1(t)}^{(k)}, \bar{x}_{2(t)}^{(k)}, \ldots, \bar{x}_{N(t)}^{(k)}, c_{1(t)}^{(k)}, c_{2(t)}^{(k)}, \ldots, c_{N(t)}^{(k)}]' \quad (7a)$$

and

$$\theta_k = [\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_N^{(k)}, \sigma_{11}^{*(k)}, \sigma_{22}^{*(k)}, \ldots, \sigma_{NN}^{*(k)}]' \quad (7b)$$

where $\bar{x}_n^{(k)}$ = the true mean of the nth feature x_n in C_k,

$$\sigma_{nn}^{*(k)} = E(x_n^2) \quad \text{for the } k\text{th class}$$
$$= \sigma_{nn}^{(k)} - \bar{x}_n^{(k)2}.$$

If

(C1) $p_t^{(k)} = P[\mathrm{DPS}_k(t) \leq 1 \mid \hat{\theta}_{t-1}^{(k)}] > \delta_k\ \forall t$

(C2) $\alpha_n^{(j)} = E(x_n^4)$ exists for each class $C_j\ \forall n = 1(1)N$

then

(a) $\{\hat{\theta}_t^{(k)} - \bar{\theta}_{k(t)}\}$ converges with probability 1 to $\mathbf{0}$, the N-dimensional null-vector, as $t \to \infty$, for each k

(b) $\{E\|\hat{\theta}_t^{(k)} - \bar{\theta}_{k(t)}\|^2\}$ converges, as $t \to \infty$ for each k,

where

$$\bar{\theta}_{k(t)} = \sum_{j=1}^{m} \beta_{kj}(t+1)\,\theta_j.$$

Corollary 1.1. If as $t \to \infty$, $\beta_{kj}(t) \to \beta_{kj}$, $j = 1(1)m$, for some k, then under (C1) and (C2), (where $\beta_{kj} \in [0, \infty)\forall k, j$)

$$\hat{\theta}_t^{(k)} \xrightarrow{\text{a.s.}} \sum_{j=1}^{m} \beta_{kj}\,\theta_j.$$

Corollary 1.2. If, as $t \to \infty$, $\beta_{kj}(t) \to \beta\delta_{kj}$ where $\beta \in [0, \infty)$ and δ_{kj} is the Kronecker delta, $\forall j$ for some k, then under (C1) and (C2),

$$\frac{1}{\beta}\hat{\theta}_t^{(k)} \xrightarrow{\text{a.s.}} \theta_k.$$

Theorem 2. Let $\bar{x}_{n(t)}^{(k)}$ and $s_{n(t)}^{(k)}$ be as in equations (3), (4) and (5). If the conditions (C1) and (C2) hold, then

(i) $\bar{x}^{(k)}_{n(t)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}^{(j)}_{n} \to 0$
almost surely as $t \to \infty$.

(ii) $$s^{(k)}_{n(t)} - \left[\sum_{j=1}^{m} \beta_{kj}(t+1)\, \sigma^{*(k)}_{nn} - \left(\sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}^{(k)}_{n} \right)^2 \right] \to 0$$

almost surely as $t \to \infty$.

The proofs of the Theorems 1 and 2 are given in Appendix B.

Corollary 2.1. If, as $t \to \infty$, $\beta_{kj}(t) \to \beta_{kj} \forall j = 1(1)m$, for some k, $\beta_{kj} \in [0, \infty)$, then under (C1) and (C2),

$$s^{(k)}_{n(t)} \xrightarrow{a.s.} \sum_{j=1}^{m} \beta_{kj}\, \sigma^{*(j)}_{nn} - \left(\sum_{j=1}^{m} \beta_{kj}\, \bar{x}^{(j)}_{n} \right)^2.$$

Corollary 2.2. If, as $t \to \infty$, $\beta^{(t)}_{kj} \to \beta\delta_{kj}$ for all j and some k, where β is some positive quantity $\epsilon[0, \infty)$ and δ_{kj} is the Kronecker delta, then under (C1) and (C2),

$$\frac{1}{\beta} s^{(k)} \xrightarrow{a.s.} \sigma^{(k)}_{nn}.$$

Incidentally, the meaning of the condition (C1) is quite clear. It ensures that the value of λ is such that at no stage the guard-zone is "too small". This serves to emphasize the importance of the choice of λ in ensuring some sort of convergence. The best choice would seem to be that for which, in the long run, the supervisor rejects every wrong decision of the classifier, while the probability of its endorsing a correct decision of the classifier asymptotically becomes as high as possible. Of course, it may be a debatable point whether a fixed choice of λ for all classes and all stages (or even for a given class for all stages) can help ensure this; as such, this point requires further study.

We now state a theorem which establishes that, in general, although individual estimates may not converge strongly to their corresponding true values, certain linear combinations of them converge strongly to the various true parameter values.

Theorem 3. For $k = 1(1)m$, let $\hat{\theta}^{(k)}_t$ and θ_k be as in equations (7a) and (7b) respectively. Then if (C1) and (C2) hold,

$$\sum_{j=1}^{m} \gamma_{kj}(t+1)\, \theta^{(j)}_t \to \theta_k \text{ with probability 1 as } t \to \infty, \quad \text{for } k = 1(1)m,$$

where $\gamma_{kj}(t+1)$, $k, j = 1(1)m$, are the elements of the generalized inverse[5] $\underset{m \times m}{\Gamma}(t+1)$ of the matrix $\underset{m \times m}{\beta}(t+1) = ((\beta_{ij}(t+1)))$, satisfying

$$\beta(t+1)\, \Gamma(t+1) = \mathbf{I}_m \text{ the identity matrix of order } m. \quad (8)$$

Proof of Theorem 3 is given in Appendix B.

Corollary 3.1. If (C1) and (C2) hold and for some $\beta_{kj} \in [0, \infty)$,

$$\beta_{kj}(t) \to \beta_{kj} \text{ as } t \to \infty\ \forall k, j = 1(1)m,$$

then

$$\sum_{j=1}^{m} \gamma_{kj}\, \theta^{(j)}_t \xrightarrow{a.s.} \theta_k,$$

where

$((\gamma_{ij})) = \Gamma$ is the generalized inverse of the matrix $\beta = ((\beta_{ij}))$, satisfying $\beta\Gamma = \mathbf{I}_m$.

Remark. If $\beta(t+1)$ is full-rank then $\Gamma(t+1)$ is just the true inverse $[\beta(t+1)]^{-1}$.

If, however, rank $(\beta(t+1)) = \gamma(\leq m)$, then a g-inverse $\Gamma(t+1)$ satisfying (8) is the Moore–Penrose inverse[5] $\beta^{+}(t+1)$ defined as follows:

$$\underset{m \times m}{\beta^{+}}(t+1) = \sum_{i=1}^{\gamma} \lambda_i^{-1}\, \mathbf{u}_i \mathbf{u}_i'$$

where

λ_i = ith non-zero eigen-value of $\beta(t+1)$, $i = 1(1)\gamma$.

$\mathbf{u}_i$ = the orthonormal eigen-vector of $\beta(t+1)$ corresponding to λ_i.

Another theorem can now be stated.

Theorem 4. Let $s^{(k)}_{n(t)}$ and $\bar{x}^{(k)}_{n(t)}$ be as in equations (3), (4) and (5), $n = 1(1)N$, $k = 1(1)m$, $t \geq 1$. If (C1) and (C2) hold, then

$$\sum_{j=1}^{m} \gamma_{kj}(t+1)\, q^{(j)}_{n(t)} \xrightarrow{a.s.} \sigma^{(j)}_{nn} \text{ as } t \to \infty$$

where

$$q^{(k)}_{n(t)} = c^{(k)}_{n(t)} - \sum_{l=1}^{m} \beta_{kl}(t+1) \left[\sum_{j=1}^{m} \gamma_{lj}(t+1)\, \bar{x}^{(j)}_{n(t)} \right]^2$$

and $\gamma_{lj}(t+1)$, $l, j = 1(1)m$, are as in Theorem 3. Proof of Theorem 4 is given in Appendix B.

Corollary 4.1. If $\beta_{kj}(t) \to \beta_{kj}$ as $t \to \infty\ \forall k, j$ then under (C1) and (C2),

$$\sum_{j=1}^{m} \gamma_{kj}\, q^{*(j)}_{n(t)} \xrightarrow{a.s.} \sigma^{(j)}_{nn} \text{ as } t \to \infty.$$

where

$$q^{*(k)}_{n(t)} = c^{(k)}_{n(t)} - \sum_{l=1}^{m} \beta_{kl} \left[\sum_{j=1}^{m} \gamma_{lj}\, \bar{x}^{(j)}_{n(t)} \right]^2 \text{ and } \beta_{kj} \in [0, 1].$$

Remark. If $\beta_{kj} = \beta\delta_{kj} \forall k, j$, then

$$q^{*(k)}_{n(t)} = s^{(k)}_{n(t)}.$$

4. DISCUSSION

The implications of the different results stated in this paper (and proved in Appendix B) need to be discussed.

The inference from Theorems 1 and 2 is that, in general, if the supervisor fails to weed out (or, at least reduce sufficiently) wrongly labelled training samples, then we can not be assured of the strong convergence of the estimates of means and variances of classes to their corresponding true values. In Theorems 3 and 4, it is inferred that in such cases, if we can impose certain conditions (defined by assumptions A1–A6) then certain linear combinations of these estimates do converge strongly to the true values of the various parameters.

If, however, the supervisor is "asymptotically perfect" or "perfect" in the sense that it can detect *all* wrongly labelled samples eventually or at each stage, then it can be inferred from Theorems 1 and 2 that the estimates converge strongly to the respective true values of parameters. This follows basically from the definition of the $\beta_{kj}(t)$'s (equation 6f) and the fact that $P(A_k(t) \mid \mathbf{x}, \hat{w} = k, w = j)$ can only take either of the two values 0 and 1, with the possibility of its taking the value 1 being considerably higher for correctly labelled samples.

5. SUMMARY

In this paper we have investigated certain aspects of the large-sample behaviour of a self-supervised pattern recognition system which was reported earlier.[1] The problem considered is that of stochastic convergence of the system in the presence of mislabelled training samples. For this purpose, we have adopted a simple model[2] for labelling errors.

The recognition system[1] itself can be characterized as follows. For an m-class pattern recognition problem based on an N-dimensional feature vector $\mathbf{X}$, it is basically a two-stage process for each input sample. In the first stage, the input sample is classified into one of the m classes on the basis of its maximum $\mu_j(\mathbf{X})$-value, defined by equation (1a). In the next stage, the updating of parameters takes place. In this stage a supervisor is appointed by means of the so-called guard zones. These are hyperellipsoidal regions defined with the preceding estimates of the mean as centre and have axes proportional to the preceding values of the standard deviations in the respective directions. The constant of proportionality, called the zone-controlling parameter, controls the dimension of the guard zone. Analytically, the guard zone for a class C_k is defined as the region

$$\{\mathbf{x} \mid \mathrm{DPS}_k(t) \leq 1\}$$

where $\mathrm{DPS}_k(t)$ is defined by equation (2). The current training sample for a given class is used to update the estimates of the parameters only if it falls within the guard zone for the class. Otherwise, the estimates are kept unchanged and the system calls for the next input sample.

This sort of learning algorithm is basically of the stochastic approximation type. Hence we have made use of results on multidimensional stochastic approximation to study its convergence under the model for labelling errors that we have adopted. The inferences made are based on certain assumptions (C1 and C2) and can be summarized thus.

In the presence of labelling errors, the sequence of estimates $\{\hat{\theta}_i^{(k)}\}$ does not converge strongly to the true value of the respective parameters. Rather, it converges strongly with another sequence

$$\left\{\sum_{j=1}^{m} \beta_{kj}(t)\,\theta_j\right\}$$

where $\beta_{kj}(t)$ is as in equation (6f), and $\hat{\theta}_i^{(k)}$ and θ_k are as in equation (7). Also, the sequence

$$\left\{\hat{\theta}_i^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t)\,\theta_j\right\}$$

converges in the mean square as $t \to \infty$ (Theorem 1). Another inference (Theorem 2) which follows from the above, is that certain linear combinations of the estimates, *viz.*

$$\sum_{j=1}^{m} \gamma_{kj}(t+1)\,\hat{\theta}_t(j)$$

converge strongly to $\hat{\theta}_k$, $k = 1(1)m$, as $t \to \infty$, where γ_{kj}, $k, j = 1(1)m$ are as defined in Theorem 3. However, if the parameters to be estimated are the class means and variances, the corresponding results are slightly complicated. While the estimates of the means behave as described above, the behaviour of the estimates $s_{n(i)}^{(k)}$ of the class variances follows a different pattern which is described in Theorems 2 and 4.

Finally, it was seen from Theorems 1 and 2 that if the supervisor can detect all wrongly labelled samples either eventually or at each stage, then the estimates do converge strongly to the respective true values of parameters.

REFERENCES

1. S. K. Pal, A. K. Datta and D. Dutta Majumder, A self-supervised vowel recognition system, *Pattern Recognition* **12**, 27–34 (1980).
2. C. B. Chittineni, Learning with imperfectly labelled samples, *Pattern Recognition* **12**, 281–291 (1980).
3. Ya. Z. Tsypkin, *Foundations of the Theory of Learning Systems*. Academic Press, New York (1973).
4. L. Schmetterer, Multidimensional stochastic approximation, *Multivariate Analysis—II: Proc. 2nd Int. Symp. Multiv. Anal.* Dayton, Ohio (P. R. Krishnaiah, ed.) Academic Press, New York (1968).
5. C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. John Wiley, New York (1971).

About the Author — AMITA PATHAK obtained her B.Sc. degree from Presidency college, Calcutta in 1979, and her M.Sc. degree in Statistics from the University of Calcutta, India in 1981. Presently, she is a senior research fellow in the Electronics and Communication Sciences Unit, Indian Statistical Institute, working towards a Ph.D. degree. Her research interest includes pattern recognition and machine learning.

About the Author — SANKAR K. PAL obtained M. Tech. and Ph.D. degrees in Radiophysics and Electronics from Calcutta University, India in 1974 and 1979 respectively. In 1982 he received another Ph.D. along with DIC in Electrical Engineering from Imperial College, London University, England. He is the recipient of the Commonwealth Scholarship' 1979 and MRC (UK) Post-doctoral Fellowship' 1981 to work at Imperial College, London, and the Fulbright Post-doctoral visiting Fellowship' 1986 to work at the University of California, Berkeley and the University of Maryland, U.S.A.

Currently, he is working as an Associate Professor in Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta. He is also a guest teacher in Computer Science, Calcutta University. His research interest includes pattern recognition, image processing, artificial intelligence and fuzzy sets and systems. He is the author of the book *Fuzzy Mathematical Approach to Pattern Recognition* (John Wiley, 1986) and has more than seventy research papers, including seven in edited books, to his credit. Dr. Pal is one of the reviewers of the *Mathematical Reviews* (American Mathematical Society), a Senior Member of the IEEE, a Fellow of the IETE and Treasurer of the Indian Society for Fuzzy Mathematics and Information Processing (ISFUMIP).

APPENDIX A

Proof of equation (6d)

We know that for any event A in the sample space of the random variable $\mathbf{x}$

$$p(\mathbf{x}\,|\,\hat{w}=k)=\begin{cases} p(\mathbf{x}\,|\,\hat{w}=k,A) & \text{if } x\in A \\ p(\mathbf{x}\,|\,\hat{w}=k,A^c) & \text{otherwise} \end{cases} \tag{A.1}$$

Now, for any event A for which $P(\hat{w}=k, A)\neq 0$ and $P(\hat{w}=k, A^c)\neq 0$, we have

$$\begin{aligned} p(\mathbf{x}\,|\,\hat{w}=k,A) &= \frac{p(\mathbf{x},\hat{w}=k,A)}{P(\hat{w}=k,A)} \\ &= \frac{\left(\sum_{j=1}^{m} p(\mathbf{x},\hat{w}=k,w=j,A)\right)}{P(\hat{w}=k,A)} \\ &= \sum_{j=1}^{m} a_{kj}\, p(\mathbf{x}\,|\,\hat{w}=k,w=j)/P(A_k(t)\,|\,\hat{w}=k), \text{ say,} \end{aligned} \tag{A.2}$$

where

$$a_{kj}=\frac{P(A\,|\,\mathbf{x},\hat{w}=k,w=j)}{P(\hat{w}=k)}\alpha_{jk}\pi_j \tag{A.3}$$

using various well-known results from the theory of conditional probability.

Similarly we have

$$\begin{aligned} p(\mathbf{x}\,|\,\hat{w}=k,A^c) &= \frac{\sum_{j=1}^{m} P(A\,|\,\mathbf{x},\hat{w}=k,w=j)\,p(\mathbf{x}\,|\,w=k,w=j)\alpha_{jk}\pi_j}{P(\hat{w}=k,A^c)} \\ &= \sum_{j=1}^{m} a^*_{kj}\, p(\mathbf{x}\,|\,\hat{w}=k,w=j)/P(A^c_k(t)\,|\,\hat{w}=k), \text{ say,} \end{aligned} \tag{A.4}$$

where

$$a^*_{kj}=\frac{P(A\,|\,\mathbf{x},\hat{w}=k,w=j)}{P(\hat{w}=k)}\alpha_{jk}\pi_j \tag{A.5}$$

$$p(\mathbf{x}\,|\,\hat{w}=k)=\begin{cases} \sum_{j=1}^{m} a_{kj}\,p(\mathbf{x}\,|\,w=j) & \text{if } x\in A \\ \sum_{j=1}^{m} a^*_{kj}\,p(\mathbf{x}\,|\,w=j) & \text{otherwise} \end{cases} \tag{A.6}$$

Note. Under assumptions (A5) and (A6), the a_{kj}'s and the a^*_{kj}'s can easily be seen to be lying in the interval [0, 1] as shown below.

As

$$\begin{aligned} P(\hat{w}=k) &= \sum_{j=1}^{m} P(\hat{w}=k, w=j) \\ &= \sum_{j=1}^{m} P(\hat{w}=k\,|\,w=j)\,P(w=j)=\sum_{j=1}^{m}\alpha_{jk}\pi_j \end{aligned}$$

we must have

$$0\leqslant\frac{\alpha_{jk}\pi_j}{P(\hat{w}=k)}=\frac{\alpha_{jk}\pi_j}{\sum_{j=1}^{m}\alpha_{jk}\pi_j}\leqslant 1.$$

However, as $P(A\,|\,\mathbf{x},\hat{w}=k,w=j)\in[0,1]$ it follows that

$$a_{kj}\in[0,1],\ \forall\ k,j.$$

Similarly, it can be seen that

$$a^*_{kj}\in[0,1]\ \forall\ k,j.$$

APPENDIX B

For proving Theorems 1, 2, we shall require the following lemmas.

Lemma 1.[3] Let $\{a_n\}$ be a sequence of positive real numbers such that

$$\sum_{n=1}^{\infty} a_n^2<\infty. \tag{B1}$$

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

$$\mathbf{x}_{n+1}=\mathbf{x}_n-a_n\mathbf{y}_n,\quad n\geqslant 1. \tag{B2}$$

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

$$E(\mathbf{y}_n\,|\,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=\mathbf{M}_n(\mathbf{x}_n)\text{ a.e.} \tag{B3}$$

Let a, b, c be non-negative real numbers and let

$$E(\|\mathbf{y}_n\|^2\,|\,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)\leqslant a+b\|\mathbf{x}_n\|+c\|\mathbf{x}_n\|^2\text{ a.e.} \tag{B4}$$

Also, for every $\mathbf{x}\in\mathbb{R}^k$ and $n\geqslant 1$,

$$\mathbf{x}'\,\mathbf{M}_n(\mathbf{x})\geqslant 0. \tag{B5}$$

If $\mathbf{x}_1$ is so chosen that

$$E(\|\mathbf{x}_1\|^2)\text{ exists} \tag{B6}$$

then the sequence $\{\mathbf{x}_n\}$ converges with probability 1 and the sequence $\{E\|x_n\|^2\}$ converges also.

Lemma 2.[3] Suppose that assumptions (B1)–(B6) hold. If there exists, for every $\eta > 0$ a $\delta > 0$ such that for $n \geq 1$

$$\inf_{\eta \leq \|x\| \leq \eta^{-1}} x' M_n(x) \geq \delta, \quad \text{(B7)}$$

then $\{x_n\}$ converges to the k-dimensional null vector 0 almost surely, that is, with probability one.

Lemma 3. Let $g: \mathbb{R}^p \to \mathbb{R}^q$ be a continuous map, $p, q \geq 1$. If

$$P\left[\lim_{n\to\infty} \|x_n - a\| = 0\right] = 1, \quad x_n, a \in \mathbb{R}^p$$

then

$$P\left[\lim_{n\to\infty} \|g(x_n) - g(a)\| = 0\right] = 1.$$

i.e.

$$x_n \xrightarrow{a.s.} a \Rightarrow g(x_n) \xrightarrow{a.s.} g(a).$$

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 $t \to \infty$, $\forall k$

and

(ii) $\{E[\|\varphi_t^{(k)}\|^2]\}$ converges as $t \to \infty$, $\forall k$,

where

$$\varphi_t^{(k)} = \theta_t^{(k)} - \bar{\theta}_{k(t)}$$

These, in turn, follow immediately from Lemmas 1 and 2 if it can be shown that the conditions (B1)–(B7) hold with $x_n = \theta_n^{(k)}$.

We first note that

$$\theta_t^{(k)} = \begin{cases} f(X_{(t)}^{(k)}) & \text{for } t = 1 \quad \text{(B.1)} \\ \theta_{t-1}^{(k)} - \frac{1}{t} Y_{t-1}^{(k)}, & t > 1 \quad \text{(B.2)} \end{cases}$$

where

$$Y_{t-1}^{(k)} = \begin{cases} \theta_{t-1}^{(k)} - f(X_{(t)}^{(k)}) & \text{if } DPS_k(t) \leq 1 \\ 0 & \text{otherwise.} \end{cases} \quad \text{(B.3)}$$

And $f: \mathbb{R}^N \to \mathbb{R}^{2N}$ is a continuous map defined as

$$f(x) = [x_1, x_2, \ldots, x_N, x_1^2, x_2^2, \ldots, x_N^2]',$$

where

$$x = [x_1, x_2, \ldots, x_N]' \in \mathbb{R}^N.$$

Obviously, therefore,

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

where

$$Z_{t-1}^{i} = \begin{cases} \varphi_t^{(k)} - g(X_{(t)}^{(k)}) & \text{if } DPS_k(t) \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{(B.6)}$$

and

$$g(X_{(t)}^{(k)}) = f(X_{(t)}^{(k)}) - \bar{\theta}_{k(t)},$$

$$\bar{\theta}_{k(t)} = \sum_{j=1}^{m} \beta_{kj}(t+1)\,\theta_j \quad \text{(B.7)}$$

We now proceed to verify the conditions (B1)–(B7) for $\varphi_t^{(k)}$. As $a_n = \frac{1}{n} \forall n$ here, and $\sum_n \frac{1}{n^2} < \infty$, (B1) is satisfied. (B2) holds, because of equation (B.5).

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

$$E[Z_t^{(k)} \mid \varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}]$$
$$= E[\varphi_t^{(k)} - g(X_{(t+1)}^{(k)}) \mid \varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}, A_k(t+1)],$$

as $Z_t^{(k)} = 0$ in $A_k^c(t+1)$.

$$= \varphi_t^{(k)} - E[g(X_{(t+1)}^{(k)}) \mid A_k(t+1)]$$

as $X_{(t+1)}^{(j)}$ is independent of $X_{(1)}^{(k)}, X_{(2)}^{(k)}, \ldots, X_{(t)}^{(k)}$ and hence $\varphi_1^{(k)}, \ldots, \varphi_t^{(k)}$.

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

since

$$E[g(X_{(t+1)}^{(k)}) \mid A_k(t = 1)]$$
$$= E[f(X_{(t+1)}^{(k)}) \mid A_k(t+1)] - \theta_k(t)$$
$$= \sum_{j=1}^{m} \beta_{kj}(t+1)\,E(X \mid w = j) - \theta_k(t)$$

on account of Equation (6d);

$$= \sum_{j=1}^{m} \beta_{kj}(t+1)\theta_j - \bar{\theta}_k(t)$$
$$= 0.$$

This verifies (B3) with $M_t^{(k)}(x) = x$, $\forall x \in \mathbb{R}^N$.

Also,

$$E[\|Z_t^{(k)}\|^2 \mid \varphi_1^{(k)}, \varphi_2^{(k)}, \ldots, \varphi_t^{(k)}]$$
$$= E[\|\varphi_t^{(k)} - g(X_{(t+1)}^{(k)})\|^2 \mid A_k(t+1)]$$

for the same reason as before.

$$= [\|\varphi_{(t)}^{(k)}\|^2 - 2\varphi_t^{(k)'}\{E\,g(X_{(t+1)}^{(k)})\} + E\|g(X_{(t+1)}^{(k)})\|^2]$$
$$\leq \|\varphi_{(t)}^{(k)}\|^2 + R,$$

R being a finite positive constant independent of $\varphi_1^{(k)}, \ldots, \varphi_t^{(k)}$, since $E\,g(X_{(t+1)}^{(k)}) = 0$ (as seen above) in the sub-space

$$A_k(t+1) = \{x \mid DPS_k(t+1) \leq 1\}.$$

and

$$E\|g(X_{(t+1)}^{(k)})\|^2$$
$$= E\|f(X_{(t+1)}^{(k)}) - \bar{\theta}_{k(t)}\|^2$$
$$\leq E\|f(X_{(t+1)}^{(k)})\|^2 - \|\bar{\theta}_{k(t)}\|^2 \text{ as } E\,f(X_{(t+1)}^{(k)}) = \bar{\theta}_{k(t)}$$
$$\leq E\|f(X_{(t+1)}^{(k)})\|^2$$
$$= \sum_{j=1}^{m} \beta_{kj}(t+1)\,(\sigma_{nn}^{(j)} + \alpha_n^{(j)}), \text{ by (C2)}$$
$$\leq \sum_{j=1}^{m} (\sigma_{nn}^{(j)} + \alpha_n^{(j)}) = R, \text{ say.}$$

Thus (B4) holds with $a = R$, $b = 0$, $c = 1$.

Finally, as

(i) $x' M_t^{(k)}(x) = x'x \geq 0$

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

and

(iii) $\inf_{\eta \leq \|x\| \leq \eta^{-1}} x' M_t^{(k)}(x) > \delta_k \eta^2 > 0$ because of (C1),

the conditions (B5), (B6) and (B7) are respectively seen to be true. Hence the theorem.

Proof of Theorem 2. This theorem follows directly from Lemma 3 and Theorem 1.

As (C1) and (C2) hold, we must have, by Theorem 1,

$$\bar{x}_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\,\bar{x}_n^{(j)} \xrightarrow{a.s.} 0,$$

and

$$c_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\,\sigma_{nn}^{*(j)} \xrightarrow{a.s.} 0 \quad \text{as } t \to \infty.$$

Also,

$$s_{n(t)}^{(k)} - \left[\sum_{j=1}^{m} \beta_{kj}(t+1)\, \sigma_{nn}^{*(j)} - \left(\sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}_n^{(j)} \right)^2 \right]$$

$$= \left[c_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \sigma_{nn}^{*(j)} \right]$$

$$- \left[(\bar{x}_{n(t)}^{(k)})^2 - \left(\sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}_n^{(j)} \right)^2 \right]. \qquad \text{(B.8)}$$

By virtue of Lemma 4 given below, it follows that

$$\bar{x}_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}_n^{(j)} \overset{a.s.}{\to} 0$$

implies

$$(\bar{x}_{n(t)}^{(k)})^2 - \left(\sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}_n^{(j)} \right)^2 \overset{a.s.}{\to} 0$$

since $\sum_{j=1}^{m} \beta_{kj}(t+1)\, \bar{x}_n^{(j)}$ is bounded above by $\sum_{j=1}^{m} \bar{x}_n^{(j)}$, a finite quantity.

The right hand side of equation (B.8) converges surely to zero. Hence the theorem.

Lemma 4. Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on the probability space (Ω, F, P). If

$$X_n - Y_n \overset{a.s.}{\to} 0,$$

and $|Y_n| < K$, a finite quantity, then

$$X_n^2 - Y_n^2 \overset{a.s.}{\to} 0.$$

Proof of Theorem 3. This follows directly from Theorem 1 and Lemmas 3 and 5.

Lemma 5. Let $\mathbf{X}_n$, $n = 1, 2 \ldots$ be a sequence of matrices of order $p \times q$ whose elements $x_{ij}^{(n)}$ are random variables over a probability space (Ω, F, P). Let $\mathbf{A}$ be another matrix of order $p \times q$ such that every element $x_{ij}^{(n)}$ of $\mathbf{X}_n$ converges with probability 1 to the corresponding element a_{ij} of $\mathbf{A}$, i.e.

$$x_{ij}^{(n)} \overset{a.s.}{\to} a_{ij},\ i = 1(1)p,\ j = 1(1)q, \text{ as } n \to \infty.$$

Let $\mathbf{P}$ and $\mathbf{Q}$ be matrices of order $m \times p$ and $q \times l$ respectively, and define

$$\mathbf{Z}_{n_{m \times l}} = \mathbf{P}\mathbf{X}_n\mathbf{Q}, \quad \mathbf{B}_{m \times l} = \mathbf{P}\mathbf{A}\mathbf{Q}.$$

Then

$$z_{ij}^{(n)} \overset{a.s.}{\to} b_{ij}, \quad i = 1(1)m,\ j = 1(1)l, \text{ as } n \to \infty.$$

Proof of Theorem 4. As seen in Theorem 1,

$$c_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \sigma_{nn}^{*(j)} \overset{a.s.}{\to} 0 \quad \text{as } t \to \infty.$$

Also, from Theorem 3, we have

$$\sum_{j=1}^{m} \gamma_{lj}(t+1)\, \bar{x}_{n(t)}^{(j)} \overset{a.s.}{\to} \bar{x}_n^{(l)}, \quad \forall l,$$

from which it follows by Lemma 3, that

$$\left[\sum_{j=1}^{m} \gamma_{lj}(t+1)\, \bar{x}_{n(t)}^{(j)} \right]^2 \overset{a.s.}{\to} [\bar{x}_n^{(l)}]^2$$

and hence

$$\sum_{l=1}^{m} \beta_{kl}(t+1) \left[\sum_{j=1}^{m} \gamma_{lj}(t+1)\, \bar{x}_{n(t)}^{(j)} \right]^2 - \sum_{l=1}^{m} \beta_{kl}(t+1)\, [\bar{x}_n^{(l)}]^2 \overset{a.s.}{\to} 0.$$

This, coupled with the first statement, implies that

$$q_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \sigma_{nn}^{*(j)} + \sum_{l=1}^{m} \beta_{kl}(t+1)\, [\bar{x}_n^{(l)}]^2 \overset{a.s.}{\to} 0.$$

i.e.

$$q_{n(t)}^{(k)} - \sum_{j=1}^{m} \beta_{kj}(t+1)\, \{\sigma_{nn}^{*(j)} - (\bar{x}_n^{(j)})^2\} \overset{a.s.}{\to} 0.$$

As $\sigma_{nn}^{(j)} = \sigma_{nn}^{*(j)} - (\bar{x}_n^{(j)})^2$, the theorem follows.