Approximations of unsupervised Bayes learning procedures

U.E. MAKOV
Chelsea College, London

SUMMARY

Computational constrains often limit the practical applicability of coherent Bayes solutions to unsupervised sequential learning problems. These problems arise when attempts are made to learn about parameters on the basic of unclassified observations, each stemming from any one of k classes ($k \geq 2$).

In this paper, the difficulties of the Bayes procedure will be discussed and existing approximate learning procedures will be reviewed for broad types of problems involving mixtures of probability densities. In particular a quasi-Bayes approximate learning procedure will be motivated and defined and its convergence properties will be reported for several special cases.

Keywords: SEQUENTIAL; CLASSIFICATION; ESTIMATION; BAYESIAN; QUASI-BAYES; STOCHASTIC; APPROXIMATION; IDENTIFIABLE MIXTURES.

1. INTRODUCTION

Problems of unsupervised learning arise when attempts are made to learn about parameters on the basis of sequential unclassified observations each stemming from any of k classes ($k \geq 2$). General discussions of such problems in the contexts of Pattern Recognition and Signal Detection are given in Fu (1968), Patrick (1972), Young and Calvert (1974) and references there cited.

In this paper, we shall consider the following special cases. (See a survey in Ho and Agrawala (1968), for a discussion of these and other cases).

Case A. The probabilities, $\pi_1, \ldots, \pi_k$ that an observation belongs to class H_i, i = 1, ..., k, are assumed *unknown*; the conditional probability densities $f_i(x \mid \theta_i) = f(x \mid \theta_i, H_i)$ of an observation x, assuming it to come from class H_i, are assumed completely *known* (i.e. both the functional form and the parameter vectors θ_i are known). These assumptions may be appropriate when large training sets can be made available from each individual class, but there is little initial information regarding the "mix" of observations in the context under study.

Case B. The class probabilities, $\pi_1, \ldots, \pi_k$, are assumed *known*; the conditional $f_i(x|\theta_i) = f(x|\theta_i, H_i)$ are assumed to have known functional forms, depending on parameter vectors θ_i some, or all, of which are *unknown*. For example, in many contexts it may be appropriate to assume that underlying densities are Gaussian with unknown means, while the variances and the class probabilities are known.

Case C. The class probabilities $\pi_1, \ldots, \pi_k$ are assumed *unknown*; the conditional densities $f_i(x|\theta_i)$ are assumed to have a *known* functional form, depending on parameter vectors θ_i, some, or all, of which are *unknown*.

In all the cases, the problem is as follows. A sequence of (possibly vector-valued) observations, $x_1, \ldots, x_n, \ldots$ are received, one at a time, and each has to be classified as coming from one of a known number k of exclusive classes $H_1, \ldots, H_k$ before the next observation is received. Each decision is made on the basis of knowing all the previous observations, but without knowing whether previous classifications were correct or not. We assume that the x's are received at a high rate and that strict computational constraints are imposed. We thus limit ourselves to learning procedures whose demand for computational resources is small.

Defining $\psi = (\pi,\theta)$, where $\pi = (\pi_1, \ldots, \pi_k)$, $\theta = (\theta_1, \ldots, \theta_k)$, we assume that, conditional on ψ, the x_n are independent with probability density

$$f(x_n|\psi) = \sum_{i=1}^{k} \pi_i f_i(x_n|\theta_i), \tag{1}$$

(we shall assume throughout that the f's are such as to make this mixture identifiable (see Yakowitz, 1970)). For a sequence of observations, $x_1, \ldots, x_n$, it follows from (1) that

$$f(x_1, \ldots, x_n|\psi) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_i|\theta_j) \tag{2}$$

This is a sum of k^n products of component densities, each term in the summation having an interpretation as the probability of obtaining a certain partition of the observations among the classes.

The Bayesian algorithm for learning about ψ (or the components of interest) involves the specification of an a priori density for ψ, and the subsequent recursive computation of the posterior density $p(\psi|x_1, \ldots, x_n)$ using

$$p(\psi|x_1, \ldots, x_n) \propto f(x_n|\psi)p(\psi|x_1, \ldots, x_{n-1}) \tag{3}$$

Classification of x_n is based upon any specified loss structure, and for

$i = 1,\ldots,k$ the values of $p_r(x_n \epsilon H_i | x_1,\ldots, x_n)$ the probability that the n^{th} observation belongs to class H_i, given the observations $x_1,\ldots, x_n$. These probabilities are computed using

$$p_r(x_n \epsilon H_i | x_1, \ldots, x_n) \propto f_i(x_n | x_1,\ldots, x_{n-1}). \\ .p_r(x_n \epsilon H_i | x_1, \ldots, x_{n-1}) \qquad (4)$$

here

$$f_i(x_n | x_1, \ldots, x_{n-1}) = \\ \begin{cases} f_i(x_n | \theta_i), \text{ in Case A} \\ \int f_i(x_n | \theta_i) p(\theta_i | x_1, \ldots, x_{n-1})\, d\theta_i, \text{ in Case B}, \\ \int\int f_i(x_n | \theta_i) p(\theta_i, \pi | x_1, \ldots, x_{n-1})\, d\pi\, d\theta_i, \text{ in Case C}, \end{cases} \qquad (5)$$

and

$$p_r(x_n \epsilon H_i | x_1, \ldots, x_{n-1}) = \\ \begin{cases} \int \pi_i p(\pi_i | x_1, \ldots, x_{n-1}) d\pi_i, \text{ in case A} \\ \pi_i, \text{ in case B} \\ \int\int \pi_i p(\pi_i, \theta | x_1, \ldots, x_{n-1})\, d\theta\, d\pi_i, \text{ in case C} \end{cases} \qquad (6)$$

It is obvious that due to the mixture form inherent in (1) and (2) there exist no reproducting (natural conjugate) densities for unsupervised Bayes learning. This results in an unavoidable increase in computer time and memory requirements, and leads to the solution being impractical in the case of signals arriving at a high rate, where speed of computation and small memory requirements are basic prerequisites for a solution. For this reason, the formal Bayes learning procedure (B) has been regarded as of little practical use. Among the ad hoc solutions proposed in its place, we note the Decision Directed approach, Recursive Moment Estimates and Learning with Probabilistic Teacher, all of which are discussed in the references given above.

As an alternative to these, we propose a Quasi-Bayes procedure which is both highly computationally efficient and retains the *flavour* of the formal

Bayes solution. Our discussion will be in terms of Cases A, B and C as above, but the approach can be extended to more general solutions. The statistical literature abounds with papers on the estimation of parameters of mixture distributions. The proposed methods (maximum likelihood estimators, moment generating function estimator, method of moments) demand considerable computational resources and thus will not be discussed here. (For references, see Quandt and Ramsey, (1978) and the ensuing discussion).

2. APPROXIMATE PROCEDURES FOR CASE A.

For convenience of notation, we shall write $\pi = (\pi_1,\ldots,\pi_k)$ for the unknown class probabilities, and $f_i(x_n)$ for the known densities. Prior knowledge about π is specified in the form of an a priori density $p(\pi)$.

If we denote by $p\,(\pi|X_n)$ the posterior density for π given $X_n = (x_1,\ldots,x_n)$, and by $p_i(\pi|X_n)$ the posterior density for π if it is also known that $x_n \in H_i$, then, by Bayes theorem,

$$p\,(\pi|X_n) = \sum_{i=1}^{k} w_i(X_n)\, p_i(\pi|X_n), \tag{7}$$

where

$$w_i(X_n) = p\,(x_n \in H_i|X_n) = \frac{f_i(x_n)\hat{\pi}_i(X_{n-1})}{\Sigma_{i=1}^{k} f_i(x_n)\hat{\pi}_i(X_{n-1})}, \tag{8}$$

and

$$\hat{\pi}_i(X_{n-1}) = \int \pi_i\, p\,(\pi|X_{n-1})d\pi. \tag{9}$$

We now consider the special case where $p\,(\pi)$ has the form of a Dirichlet density

$$p(\pi) = \frac{\Gamma(\alpha_1^{(0)}+\ldots+\alpha_k^{(0)})}{\Gamma(\alpha_1^{(0)})\ldots\Gamma(\alpha_k^{(0)})} \prod_{i=1}^{k} \pi^{\alpha_i^{(0)}-1} \tag{10}$$

which we denote by $D\,(\pi|\alpha_1^{(0)},\ldots,\alpha_k^{(0)})$, where $\Gamma\,(\cdot)$ is the standard gamma function. Such a form might arise, for example, following a multinomially distributed training sample whose correct classifications were known.

It follows from (7) and (10) that after observing x_1 we obtain

$$p\,(\pi\,|\,X_1) = \sum_{i=1}^{k} w_i(X_1)\; D\,(\pi\,|\,\alpha_1^{(0)} + \delta_{i1},\ldots,\alpha_k^{(0)} + \delta_{ik}), \tag{11}$$

where

$$w_i(X_1) = \frac{f_i(x_1)\alpha_i^{(0)}}{\sum_{i=1}^{k} f_i(x_1)\alpha_i^{(0)}} \qquad (i = 1,\ldots,k) \tag{12}$$

and

$$\begin{aligned}\delta_{ij} &= 1 \text{ if } i = j,\\ &= 0 \text{ otherwise.}\end{aligned}$$

Many well-known approximate learning procedures for this problem can be seen as arising from approximations to (11) of the form

$$p\,(\pi\,|\,X_1) \approx D\,(\pi\,|\,\alpha_1^{(0)} + \hat{\delta}_{11},\ldots,\alpha_k^{(0)} + \hat{\delta}_{1k}), \tag{13}$$

where the $\hat{\delta}_{ij}$'s take values according to some specified method. Two approaches are suggested.

I. Averaging. The $\hat{\delta}_{ij}$'s are chosen such that the mean and variance of the approximating density (13) are identical to those of the mixture (11). A similar approach (though in a different context) is taken in Owen (1975); Athans, Whiting and Gruber (1977); Harrison and Stevens (1976).

II. Selection. Here one of the $\hat{\delta}_{ij}$ takes the value one and the others zero according to some decision rules. This approach is akin to the engineering concept of 'learning without a teacher', see Agrawala, (1973); Spragins (1966); Fralick (1967), where the unknown 'teacher', the $\hat{\delta}_{ij}$, is the missing label identifying the observation with its class. Particular examples are the Decision-Directed learning and the Probabilistic Teacher Scheme. A comparative study (in the context of jumps in linear systems) of several averaging and selection methods is given in Smith and Makov (1980).

(i) Decision-Directed Learning (DD)

According to the DD approach one of the $\hat{\delta}_{ij}$ is set equal to one and the others zero in such a way that using (4) and some specified loss function, this results in a minimum expected posterior loss. In other words, by setting $\hat{\delta}_{ij}$ to equal zero or one we regard our own (unconfirmed) classification as if it were true. For example, in Scudder (1965),* the $\hat{\delta}_{ij}$ was set to equal one if $w_i\,(X_1)$ ·

* In context of Case B.

was maximized for $i = j$. The approach in effect assumed that the most likely H_i was, in fact, the true one (and thus zero one loss function assumed).

The DD scheme was further studied in Davisson and Schwartz (1970), where it was shown that the approach did not guarantee asymptotic unbiasedness and could also lead to problems of *runaways*. Runaway occurs when the scheme commits a sequence of errors resulting in a degradation of performance and consequent convergence to biased values. In Davisson and Schwartz (1970), Davisson (1970), the detection of signals in Gaussian noise was considered and bounds on the probability of runaway were provides using random walk theory. It was shown that except for very low signal to noise ratio, the probability of runaway of the class probabilities to the extreme values 0 and 1 was very small.

In Katopis and Schwartz (1972), a modified version of DD (MDD) was proposed in which a bias-removing transformation of the observations was introduced such that the convergence to the true value of the class probability π was ensured. Another modification was given in Schwartz and Katopis (1977). In Kazakos and Davisson (1979), in adittion to a bias-removing transformation, a specific gain function (in the DD recursion) was suggested that guaranteed fastest mean square error convergence of the estimates of the π_i's. All these modifications were shown to avoid the problems associated with the DD scheme, but at the expense of requiring numerical integration after each observation.

(ii) Learning with a Probabilistic Teacher (PT)

According to this scheme, proposed in Agrawala (1970), a randomized choice is made; $\hat{\delta}_{ij}$ being set equal to one with probability $w_j(X_1)$. In Silverman (1979), the theoretical properties of the PT for Case A were discussed; convergence was proved and asymptotic relative efficiency properties were examined.

The scheme which we propose is as follows:

Quasi-Bayes Learning (QB), see Makov and Smith (1977); Smith and Makov (1978); Makov and Smith (1976), replaces $\hat{\delta}_{ij}$ by $w_j(X_1)$, and so takes

$$p\,(\pi\,|X_1) \approx D\,(\pi\,|\alpha_1^{(1)},\ldots,\alpha_k^{(1)}), \qquad (14)$$

where

$$\alpha_i^{(1)} = \alpha_i^{(0)} + w_i(X_1)\,(i=1,\ldots,k). \qquad (15)$$

Subsequent updating proceeds in the same way, so that with $p\,(\pi\,|X_{n-1})$ having a Dirichlet form with parameters $\alpha_i^{(n-1)}$, it follows that $p\,(\pi\,|X_n)$ will be Dirichlet with parameters

$$\alpha_i^{(n)} = \alpha_i^{(n-1)} + w_i(X_n)\,(i=1,\ldots k), \tag{16}$$

where, corresponding to (12),

$$w_i(X_n) = \frac{f_i(x_n)\,\alpha_i^{(n-1)}}{\Sigma_{i=1}^{k} f_i(x_n)\,\alpha_i^{(n-1)}} \qquad (\mathrm{i}=1,\ldots,\mathrm{k}) \tag{17}$$

In the special case $k = 2$, the Quasi-Bayes procedure leads to recursive estimates of π_1 if the form

$$\hat{\pi}_1^{(n+1)} = \hat{\pi}_1^{(n)} - a_n(\hat{\pi}_1^{(n)} - w_1^{(n+1)}), \tag{18}$$

where

$$a_n^{-1} = (\alpha_1^{(0)} + \alpha_2^{(0)} + n + 1) \tag{19}$$

and

$$w_1^{(n+1)} = \frac{f_1(x_{n+1})\,\hat{\pi}_1^{(n)}}{f_1(x_{n+1})\hat{\pi}_1^{(n)} + f_2(x_{n+1})\hat{\pi}_2^{(n)}} \tag{20}$$

(18) is a typical QB recursion (for this case and others), which corresponds to a Robbins-Monro (Robbins and Monro, 1951) type of Stochastic Approximation. Using existing theorems in this field (e.g. Gladyshev, 1965, and many other) we were able to prove that the QB scheme converges to the true value of π in mean square and with probability one. Convergence properties were established for the case $k = 2$ in Makov and Smith (1977), Makov (1980), and for general k in Smith and Makov (1978). It was also shown in Makov and Smith (1976), that the QB scheme provides a better approximation to the Bayes solution than does the MDD. In Silverman (1979) the QB was proved to be more efficient than the PT.

In Kazakos (1977), a recursive estimation algorithm was provided which was based on the minimization of the Kullback-Leiber information number. The algorithm was shown to be consistent (for any k) and efficient (for $k=2$). In Makov (1980), it was shown that the QB scheme, (18) - (20), is a special case of the one of the discussed in Kazakos (1977).

In Fig. 1, we show the paths of successive estimates of π_1, π_2 for a three-class simulated example ($k=3$), where f_1,f_2,f_3 are circular bivariate Gaussian

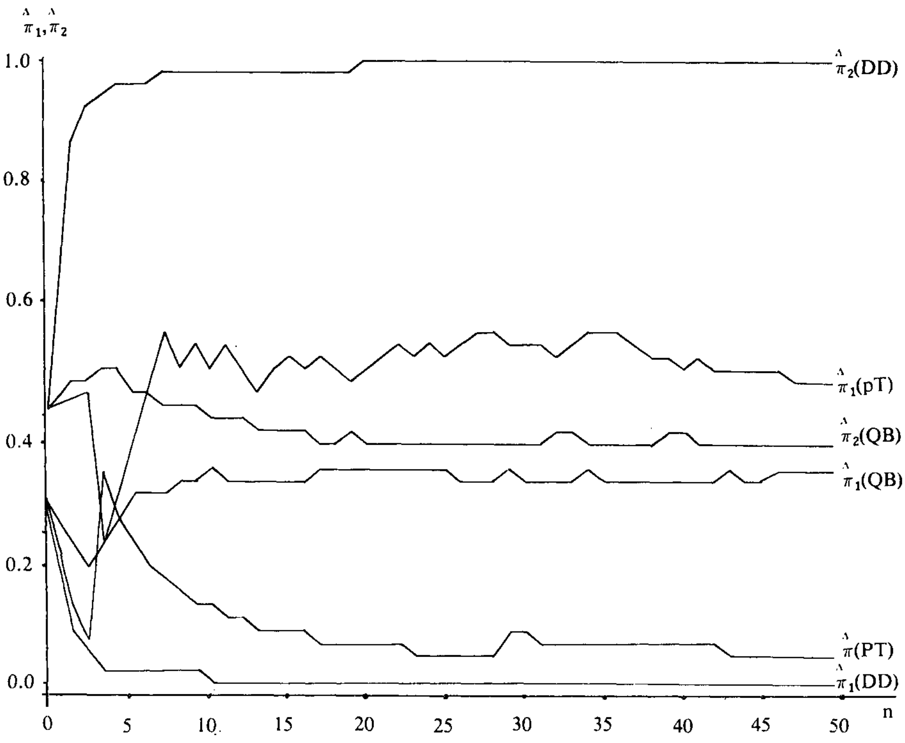

FIG. 1

distributions with all variances equal to one and means given by (-0.5,0), (0,0.5) and (0.5,0) respectively. Comparisons of the QB approach with the B solution have been made in Makov and Smith (1977), Smith and Makov (1978), and we have omitted calculation of B here. Comparison of QB with MDD was made in Makov and Smith (1976), where the latter was shown to be definitely inferior. Since the MDD would, in fact, require successive two-dimensional numerical integration for this example, it is also omitted. The results from the DD, PT and QB schemes are shown for the first 50 simulated observations, where π_1 and π_2 were both equal to 0.33. The estimates for QB were obtained using (9) and (16), which, from the well-known form of the mean of a Dirichlet distribution, implies that

$$\hat{\pi}_i(X_n) = \frac{\alpha_i^{(n)}}{\Sigma_{i=1}^{k}\alpha_i^{(n)}} \qquad (i=1,\ldots,k) \qquad (21)$$

The estimates for PT use successive randomizations as described above, or in Agrawala (1970); the estimates for DD follow the procedure as described above, or in Davisson and Schwartz (1970). The prior parameters used were $\alpha_1^{(0)} = 0.1$, $\alpha_2^{(0)} = 0.15$ and $\alpha_3^{(0)} = 0.1$, representing a very weak form of prior knowledge, implying prior means for π_1, π_2, π_3 of 0.286, 0.428 and 0.286, respectively.

In this and similar examples, where classification is made difficult because of the high overlap of the underlying distributions, the QB method shows marked superiority over the PT method, while the DD method performs very badly indeed. When the underlying distributions have only moderate overlap, there appears little to choose between QB and PT, whereas both are markedly superior to DD.

3. QUASI-BAYES PROCEDURES FOR CASE B

In order to illustrate our approach to problems which fall within the framework of Case B, we shall consider two special cases, both for the case $k=2$, and both involving known π_1, π_2 $(=1-\pi_1)$. The first is that of *Bipolar signal* detection, where $f_1(x|\theta_1)$ is a Gaussian density with unknown mean $\theta > 0$, $f_2(x|\theta_2)$ is a Gaussian density with mean $-\theta$, and the variances are known and equal (to σ^2, say). The second is that of *Signal versus Noise* detection, where $f_1(x|\theta_1)$ is a Gaussian density unknown mean θ, $f_2(x|\theta_2) = f_2(x)$ is a Gaussian density with mean zero, and the variances are known and equal (to σ^2, say).

From the general results given in the introduction, it can be shown that if we take $p^{o}(\theta)$ to be normal with mean μ and variance τ^2, then after observing x_1 we have

$$p^{(1)}(\theta) = \Sigma_{i=1}^{2} w_i^{(1)} N(\theta;\tau^{-2}\mu + \sigma^{-2}\hat{\delta}_{i1}x_1, \tau^{-2} + \sigma^{-2}|\hat{\delta}_{i1}|) \qquad (22)$$

where $w_i^{(1)} = p_r(x_1 \epsilon H_i|x_1)$ is derivable from (4), (5) and (6), $N(\theta;\, c,d)$ denotes that θ has Gaussian distribution with mean c/d, variance d^{-1}, and $\hat{\delta}_{ij} = 1$ or -1 according as $i = 1$ $(x_j \in H_1)$, or not, in the Bipolar signal case, $\hat{\delta}_{ij} = 1$ or 0 according as $i = 1$ $(x_j \in H_1)$, or not, in the Signal versus Noise case.

Our proposal is to replace $\hat{\delta}_{i1}$ by $E(\delta_{i1})$, which is equal to $2w_1^{(1)}$-1 in the Bipolar case, and equal to $w_1^{(1)}$ in the Signal versus Noise case, and to take $p^{(1)}(\theta) = N(\theta;\, \tau^{-2}\mu + \sigma^{-2}E(\delta_{i1})x_1,\ \tau^{-2} + \sigma^{-2}E(|\delta_{i1}|))$. Subsequent updating now takes place entirely within the Gaussian family, and we obtain

$$p^{(n)}(\theta) = N\left(\theta; \tau^{-2}\mu + \sigma^{-2}\Sigma_{j=1}^{n} E(\delta_{ij})x_j, \tau^{-2} + \sigma^{-2}\Sigma_{j=1}^{n} E(|\delta_{ij}|)\right). \qquad (23)$$

The posterior means give a sequence of estimates of θ, and the following recursive relations are obtained:

For the Bipolar case

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} - \frac{\sigma^{-2}}{\tau^{-2} + (n+1)\sigma^{-2}} \{\hat{\theta}^{(n)} - (2w_1^{(n+1)}-1)x_{n+1}\} \qquad (24)$$

For the Signal versus Noise case

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} - \frac{\sigma^{-2}}{\tau^{-2} + \sigma^{-2}\Sigma_{i=1}^{n+1} w_1^{(i)}} \{\hat{\theta}^{(n)} - x_{n+1} w_1^{(n+1)}\} \qquad (25)$$

Various modifications can also be considered for large n, but these are not discussed here. In Smith and Makov (1981), the convergence properties of the Signal versus Noise scheme were discussed for the case where the w's are replaced by $w_j^{*(n)} = p_r(x_n \in H_i | \hat{\theta}^{(n-1)}, \pi)$. The resulting recursion was shown to converge to θ with probability one. In Titterington, 1976, a technique similar to the QB was applied in the context of medical diagnosis where unconfirmed cases (= unsupervised) were incorporated into data banks. The 'fractional updating' was used to estimate the means and covariance matrices of multivariate normal densities.

The performance of the DD scheme and its improved version have been studied in general in Patrick, Costello and Monds (1970); Young and Farjo (1972). At the present time, following the criticism of Agrawala (1970), made in Cooper (1975), there would appear to be no satisfactory account of the theoretical properties of the PT scheme for this case.

In Fig. 2, we show the paths of successive estimates for a simulated example of the Signal versus Noise problem. A comparison is given, for the first 50 observations, of the Decision Directed, Improved Decision Directed, Probabilistic Teacher and Quasi-Bayes methods. The underlying parameters were as follows: $\theta = 2.0$, $\sigma^2 = 4.0$, $\pi_1 = 0.5$, $\mu = 5.0$, $\tau^2 = 25.0$; the latter represent very vague prior knowledge about θ. Again the pattern shown by this example is typical. Both the Probabilistic Teacher and the Quasi-Bayes procedure perform better than the Decision Directed schemes.

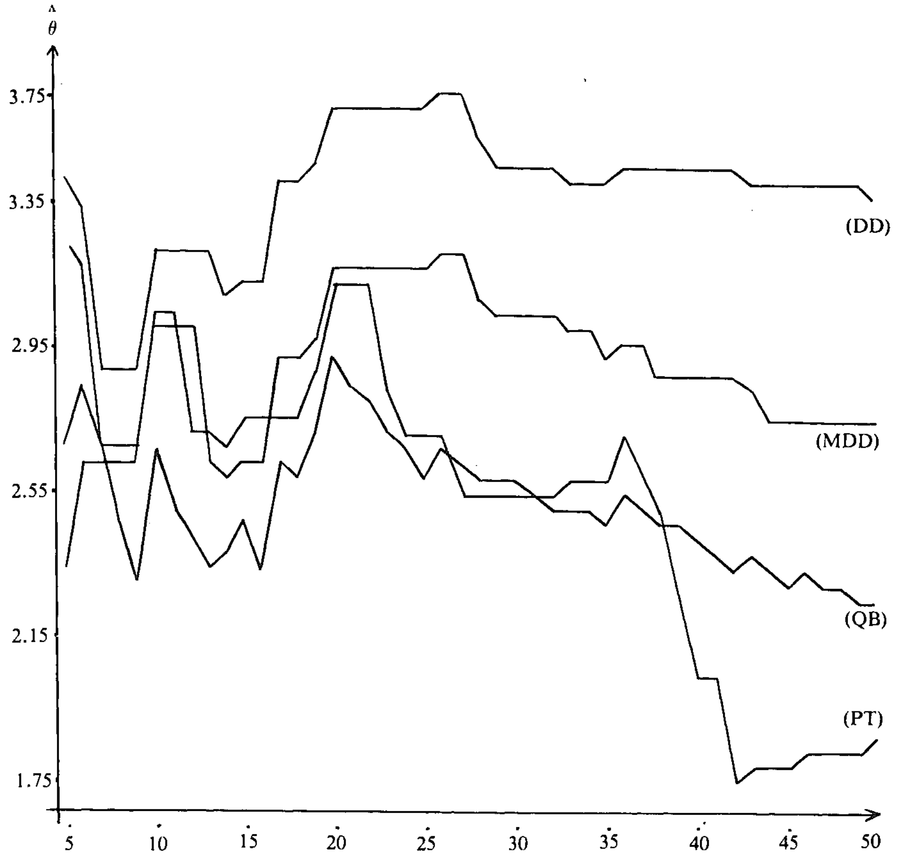

FIG. 2

4. QUASI-BAYES PROCEDURE FOR CASE C

Few results are available in this rather difficult case. In Young and Coraluppi (1970), stochastic estimation of a mixture of normal densities using an information criterion is discussed. In Katopis and Schwartz (1972); Schwartz and Katopis (1977), modified DD schemes proved to be consistent in a two-class decision problem where the mixture consisted of two normal densities, the mean of one of which was unknown (as well as the mixing parameter). In Makov (1980a), the QB scheme was attempted in a Kalman filter context in which an attempt was made to track a process when there was a non-zero probability that the observation contained nothing but pure noise. Simulations results showed that the QB scheme is by far more reliable than the PT or DD so long as the process is going through the contaminated environment. Work is in progress on the mathematical properties of the QB in

Case C. Preliminary results indicate that convergence may be guaranteed if certain restrictions are imposed on the parameter space. This will not be discussed here.

ACKNOWLEDGEMENT

I am grateful to Professor A.F.M. Smith for his useful comments on the manuscript.

REFERENCES

AGRAWALA, A.K. (1970). Learning with a probabilistic teacher. *IEEE Trans. Inform. Theory* **IT-16**, 373-379.

— (1973). Learning with various types of teachers. *Proc. 1st. Int. Joint Conf. Pattern Recognition,* 453-461.

ATHANS, M., WHITING, R.H. & GRUBER, M. (1977). A suboptimal estimation algorithm with probabilistic editing for false measurements with application to target tracking with wake phenomena. *IEEE Trans. on Automat. Contr.* **AC-22**, 372-384.

COOPER, D.B. (1975). On some convergence properties of 'learning with a probabilistic teacher' algorithms. *IEEE Trans. Inform. Theory* **IT-21**, 699-702.

DAVISSON, L.D. (1970). Convergence probability bounds for stochastic approximation. *IEEE Trans. Inform. Theory* **IT-16**, 680-685.

DAVISSON, L.D. and SCHWARTZ, S.C. (1970). Analysis of a decision directed receiver with unknown priors. *IEEE Trans. Inform. Theory* **IT-16**, 270-276.

FRALICK, S.C. (1967). Learning to recognize patterns without a teacher. *IEEE Trans. on Inform. Theory* **IT-13**, 57-64.

FU, K.S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning.* New York: Academic Press.

GLADYSHEV, E.G. (1965). On stochastic approximation. *Theory of Prob. and its Appl.* **10**, 275-278.

HARRISON, P.J. and STEVENS, C.P. (1976). Bayesian forecasting. *J.R. Statist. Soc. B.* **38**, 205-247.

HO, Y.C. and AGRAWALA, A.K. (1968). On pattern classification algorithms; introduction and survey. *Proc. IEEE* **56**, 2102-2114.

KATOPIS, A. and SCHWARTZ, S.C. (1972). Decision directed learning using stochastic approximation. *Proc. Modelling and Simulation Conf.,*- 473-481.

KAZAKOS, D. (1977). Recursive estimation of prior probabilities using a mixture. *IEEE Trans. on Inform Theory* **IT-23**, 203-211.

KAZAKOS, D. and DAVISSON, L.D. (1979). An improved decision-directed detector. *IEEE Trans. on Inform. Theory.* Submitted to publication.

MAKOV, U.E. (1980). On the choice of gain functions in recursive estimation of prior probabilities. *IEEE Trans. on Inform. Theory.* **IT-26**, 497-498.

— (1980a). A quasi-Bayes approximation for unsupervised filters. *IEEE Trans. on Autom. Contr.* **AC-25**, 842-847.

MAKOV, U.E. and SMITH, A.F.M. (1976). Quasi Bayes procedures for unsupervised learning. *Proc. IEEE Conf. on Decision and Control.* 408-412. New York: IEEE Inc.

— (1977). A quasi-Bayes unsupervised learning procedure for priors. *IEEE Trans. Inform. Theory* **IT-23**, 761-764.

OWEN, J.R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *J. Amer. Statist. Assoc.* **70**, 351-356.

PATRICK, E.A. (1972). *Fundamentals of Pattern Recognition.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

PATRICK, E.A., COSTELLO, J.P. and MONDS, F.C. (1970). Decision directed estimation for a two class decision boundary. *IEEE Trans. on Computers* **C-19**, 197-205.

QUANDT, R.E. and RAMSEY, J.B. (1978). Estimating mixture of normal distributions and swithching regressions. *J. Amer. Statist. Assoc.* **73**, 730-738.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **25**, 400-407.

SCHWARTZ, S.C. and KATOPIS, A. (1977). Modified stochastic approximation to enhance unsupervised learning. *Proc. of the IEEE Conf. on Decision and Control.* 1067-1069.

SCUDDER, H.J. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inform. Theory* **IT-11**, 363-371.

SILVERMAN, B.W. (1979). Some asymptotic properties of the probabilistic teacher. *IEEE Trans. on Inform. Theory.* **IT-26**, 296-249.

SMITH, A.F.M. and MAKOV, U.E. (1978). A quasi-Bayes sequential procedure for mixtures. *J. Roy. Statist. Soc. B* **40**, 106-112.

— (1980). Bayesian detection and estimation of jumps in linear systems. *Proceedings of the IMA Conference on the Analysis and Optimization of Stochastic Systems,* (O.L.R. Jacobs *et.al.,* eds), 333-346. New York: Academic Press.

— (1981). Approximation to Bayes learning procedures. *IEEE Trans. on Inform. Theory.* To appear.

SPRAGINS, J. (1966). Learning without a teacher. *IEEE Trans. on Inform. Theory* **IT-12**, 223-230.

TITTERINGTON, D.M. (1976). Updating a diagnostic system using unconfirmed cases. *Appl. Statist.* **25**, 238-347.

YAKOWITZ, S.J. (1970). Unsupervised learning and the identification of finite mixtures. *IEEE Trans. on Inform. Theory* **IT-16**, 330-338.

YOUNG, T.Y. and CORALUPPI, G. (1970). Stochastic estimation of a mixture of normal density functions using an information criterion. *IEEE Trans. on Inform. Theory* **IT-16**, 258-263.

YOUNG, T.Y. and FARJO, A.A. (1972). On decision directed estimation and stochastic approximation. *IEEE Trans. on Inform. Theory* **IT-18**, 671-673.

YOUNG, T.Y. and CALVERT, T.W. (1974). *Classification, Estimation and Pattern Recognition.* New York: American Elsevier.