Sequential Algorithms for Parameter Estimation Based on the Kullback-Leibler Information Measure

EHUD WEINSTEIN, MEIR FEDER, AND ALAN V. OPPENHEIM

Abstract—**Methods of stochastic approximation are used to convert iterative algorithms for maximizing the Kullback-Leibler information measure into sequential algorithms. Special attention is given to the case of incomplete data, and several algorithms are presented to deal with situations of this kind. The application of these algorithms to the identification of finite impulse response (FIR) systems is considered. Issues such as convergence properties of the proposed algorithms, choice of initial conditions, the limit distribution, and the associated regularity conditions are beyond the scope of this correspondence. However, the existing literature on stochastic approximation, together with the ideas presented in this correspondence should provide the starting point for such analyses.**

I. Introduction

Classical methods of parameter estimation such as maximum likelihood (ML) and maximum *a posteriori* (MAP) generally imply batch algorithms that require processing the received data as a whole. In a variety of applications, it is desirable to process the data sequentially. The advantage of a sequential algorithm over a batch algorithm is not necessarily in the final result, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed. Moreover, if the parameters of interest are subject to changes, e.g., they are time varying, processing all the available data jointly is not desirable, even if we can accommodate the computational and storage load of the batch algorithm, since different data segments correspond to different parameter values. In that case, a sequential algorithm can be designed to be adaptive in nature and track the varying parameters.

In this paper, we use methods of stochastic approximation to convert iterative algorithms for maximizing the Kullback-Leibler information measure into sequential algorithms. Special attention is given in case of incomplete data, and several algorithms are developed to deal with situations of this kind. We then consider the application of these algorithms to the problem of sequentially identifying finite impulse response (FIR) systems.

Important issues such as convergence properties of the proposed algorithms, choice of initial conditions, the limit distribution, and the associated regularity conditions are beyond the scope of this paper. However, the existing literature on stochastic approximation together with the ideas presented in this paper should provide the starting point for such analyses.

Manuscript received February 6, 1989; revised September 21, 1989. This work was supported in part by ONR under Grant N00014-90-J-1109 at Woods Hole Oceanographic Institution, and in part by the Defense Advanced Research Projects Agency monitored by ONR under Contract N00014-89-J-1489 at the Massachusetts Institute of Technology. It is a WHOI contribution number 7334.

E. Weinstein and M. Feder are with the Department of Electrical Engineering-Systems, Faculty of Engineering, Tel-Aviv University, 6978 Tel-Aviv, Israel.

A. V. Oppenheim is with the Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA 02139.

IEEE Log Number 9036891.

II. Sequential Algorithms Based on the Kullback-Leibler Information Measure

Let $y_1, y_2, \cdots, y_n, \cdots$ be an ergodic sequence of observations (vector random variables) whose joint probability density depends on the vector θ of unknown parameters. We want to derive sequential algorithms for estimating θ.

We consider as our objective function the Kullback-Leibler information measure [1]:

$$J(\theta) = E_{\theta_0}\{\log f_Y(y_{n+1}; \theta)\} \tag{1}$$

where $f_Y(y_{n+1}; \theta)$ denotes the marginal probability density of y_{n+1}, and $E_{\theta_0}\{\cdot\}$ denotes the statistical expectation of the bracketed quantity taken with respect to the actual (true) parameter value θ_0. Invoking Jensen's inequality [2]

$$J(\theta) \le J(\theta_0). \tag{2}$$

If $f_Y(y_n; \theta) = f_Y(y_n; \theta_0)$ a.e. y_n implies $\theta = \theta_0$ (identifiability condition), then $J(\theta)$ has a unique maximum at $\theta = \theta_0$. Therefore, by maximizing $J(\theta)$, we get the exact true parameter value. Unfortunately, $J(\theta)$ is not available to us since it involves the unavailable expectation with respect to θ_0. Therefore, given an iterative algorithm for maximizing $J(\theta)$, we shall use the method of stochastic approximation (e.g., [3], [4]) to convert it into a sequential algorithm.

Consider first the gradient-search method for maximizing $J(\theta)$:

$$\begin{aligned}\theta^{(l+1)} &= \theta^{(l)} + \beta_l DJ(\theta)\big|_{\theta=\theta^{(l)}} \\ &= \theta^{(l)} + \beta_l E_{\theta_0}\{D \log f_Y(y_{n+1}; \theta)\}\big|_{\theta=\theta^{(l)}}\end{aligned} \tag{3}$$

where $DJ(\theta)$ denotes the gradient (vector partial derivatives) of $J(\theta)$, and l denotes the index of iteration. In the transition from the first version of (3) to its second version, we have assumed that the regularity conditions for interchanging the expectation with differentiation operations are satisfied ([2, pp. 136-137]). Since the expectation in (3) is not available to us, it is approximated by its current realization. Setting $l = n$, we obtain the following sequential algorithm:

$$\theta^{(n+1)} = \theta^{(n)} + \beta_n D \log f_Y(y_{n+1}; \theta). \tag{4}$$

Invoking the ergodic nature of the $\{y_n\}$ sequence, the next iteration is performed using the next realization and thus achieves a time averaging that approximates the unavailable ensemble average. If $\{\beta_n\}$ is chosen to be a sequence of positive numbers such that

$$\lim_{n\to\infty} \beta_n = 0, \quad \sum_{n=1}^{\infty} \beta_n = \infty, \quad \sum_{n=1}^{\infty} \beta_n^2 < M < \infty \tag{5}$$

(e.g., $\beta_n = \beta/n$), then, under the stated regularity conditions ([5], [6]), the algorithm in (4) converges almost surely (a.s.) and in the mean square (m.s.) to the maximum of $J(\theta)$, that is the true parameter value. Using well-known results from the theory of stochastic approximation (e.g., [7]-[9]), the limit distribution of the parameter estimate at the point of convergence can also be derived.

If the observed sequence is not stationary, e.g., when the vector parameters exhibit changes in time, and we want an adaptive algorithm, choosing a constant gain $\beta_n = \beta$ is recommended. This corresponds to exponential weighting that reduces the effect of past observations relative to the new input data in order to track the varying parameters.

The same method can be applied to convert iterative Newton-Raphson methods into sequential algorithms.

III. Sequential Algorithms Using Incomplete Data

In many situations of interest, $\log f_Y(y_n; \theta)$ and its derivatives are complicated to express analytically. Motivated by the considerations leading to the expectation-maximization (EM) algorithm

[10], suppose we can find a vector x_n (the so-called complete data) that is related to the observed y_n (the so-called incomplete data) by

$$H_n(x_n) = y_n \tag{6}$$

where $H_n(\cdot)$ is a noninvertible (many-to-one) transformation. Express densities

$$f_X(x_n; \theta) = f_{X/Y=y_n}(x_n; \theta) \cdot f_Y(y_n; \theta), \qquad \forall H_n(x_n) = y_n \tag{7}$$

where $f_X(x_n; \theta)$ is the probability density associated with x_n, and $f_{X/Y=y_n}(x_n; \theta)$ is the conditional probability density of x_n given $Y = y_n$. Taking the logarithm on both sides of (7)

$$\log f_Y(y_n; \theta) = \log f_X(x_n; \theta) - \log f_{X/Y=y_n}(x_n; \theta). \tag{8}$$

Taking the conditional expectation given $Y = y_n$ at a parameter value θ', the left side of (8) remains unchanged, and we obtain

$$\log f_Y(y_n; \theta) = E_{\theta'}\{\log f_X(x_n; \theta) \mid Y = y_n\} - E_{\theta'}\{\log f_{X/Y=y_n}(x_n; \theta) \mid Y = y_n\}. \tag{9}$$

Define for convenience

$$Q_n(\theta, \theta') \equiv E_{\theta'}\{\log f_X(x_n; \theta) \mid Y = y_n\} \tag{10}$$

$$P_n(\theta, \theta') \equiv E_{\theta'}\{\log f_{X/Y=y_n}(x_n; \theta) \mid Y = y_n\}. \tag{11}$$

Then (9) reads

$$\log f_Y(y_n; \theta) = Q_n(\theta, \theta') - P_n(\theta, \theta'). \tag{12}$$

Taking the expectation on both sides of (12) with respect to the true parameters value

$$J(\theta) = E_{\theta_0}\{Q_n(\theta; \theta')\} - E_{\theta_0}\{P_n(\theta, \theta')\} \equiv \bar{Q}(\theta, \theta') - \bar{P}(\theta, \theta') \tag{13}$$

where $J(\theta)$ is the objective function defined in (1). Invoking Jensen's inequality

$$P_n(\theta, \theta') \le P_n(\theta', \theta'). \tag{14}$$

Therefore

$$\bar{P}(\theta, \theta') \le \bar{P}(\theta', \theta'). \tag{15}$$

Hence, recall (13)

$$\bar{Q}(\theta, \theta') > \bar{Q}(\theta', \theta') \quad \text{implies } J(\theta) > J(\theta'). \tag{16}$$

The relation in (16) forms the basis to the following iterative algorithm:

$$\max_{\theta} \bar{Q}(\theta, \theta^{(l)}) \Rightarrow \theta^{(l+1)}. \tag{17}$$

This algorithm for maximizing $J(\theta)$ is completely analogous to the iterative EM algorithm for maximizing log-likelihood functions [10]. Following essentially the same considerations as in [11], if $\bar{Q}(\theta, \theta')$ is continuous in both θ and θ', the algorithm in (17) converges to a stationary point of $J(\theta)$, where the maximization operation ensures that each iteration increases the value of $J(\theta)$. Now, observing that $\bar{Q}(\theta, \theta') = E_{\theta_0}\{Q_n(\theta, \theta')\}$, the iterative algorithm in (17) gives rise to the following sequential algorithm:

$$\max_{\theta} Q_{n+1}(\theta, \theta^{(n)}) \Rightarrow \theta^{(n+1)} \tag{18}$$

where from (10)

$$Q_{n+1}(\theta, \theta^{(n)}) = E_{\theta^{(n)}}\{\log f_X(x_{n+1}; \theta) \mid Y = y_{n+1}\}. \tag{19}$$

We may consider the following modification of the algorithm in (18):

$$\max_{\theta} \Psi_{n+1}(\theta) \Rightarrow \theta^{(n+1)} \tag{20}$$

where

$$\Psi_{n+1}(\theta) = \gamma_n \Psi_n(\theta) + Q_{n+1}(\theta, \theta^{(n)}). \tag{21}$$

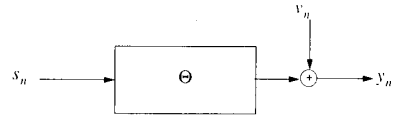

Fig. 1. FIR system identification.

The cumulative average indicated in (21) may improve the statistical stability of the resulting parameter estimates. For $\gamma_n = 1$ the algorithm in (20) coincides with the sequential algorithm proposed by Titterington [12]. Given the appropriate regularity, the resulting sequence estimates converges a.s. and in m.s. to the true parameter value, and the associated limit distribution is readily available (see [12], [13]). If we choose $\gamma_n < 1$, it corresponds to exponential weighting that reduces the effect of past observations relative to the new input data. However, it may affect the statistical stability and rate of convergence of the algorithm. These issues must be explored in depth.

The notion of complete data can also be incorporated into the gradient-based algorithms by invoking the following identity, first presented by Fisher [14], and more recently in [15]–[17]:

$$D \log f_Y(y_{n+1}; \theta) = E_{\theta}\{D \log f_X(x_{n+1}; \theta) \mid Y = y_{n+1}\} = DQ(\theta, \theta')\big|_{\theta'=\theta}. \tag{22}$$

Using (22), the log-likelihood gradient (score) of the observed data can be computed by taking the conditional expectation of the complete data score. We may find this identity very helpful in situations where the direct computation of the observed (incomplete) data score is complicated.

IV. FIR System Identification

Let y_n be the noise contaminated output of an unknown causal discrete pth-order FIR filter Θ driven by the input signal s_n (see Fig. 1), that is

$$y_n = \sum_{i=0}^{p-1} \theta_i s_{n-i} + v_n = \theta^T s_n + v_n \tag{23}$$

where $s_n = [s_n\ s_{n-1} \cdots s_{n-p+1}]^T$ and $\theta = (\theta_0\ \theta_2 \cdots \theta_{R-1})^T$ are the unit sample response coefficients of the filter. Assuming that s_n is a known signal and that v_n is a realization from a zero-mean Gaussian random variable with a known variance σ^2

$$\log f_Y(y_n; \theta) = C - \frac{1}{2\sigma^2}(y_n - \theta^T s_n)^2 \tag{24}$$

where C is a constant independent of θ. Substituting (24) into (4) and carrying out the indicated differentiation operation, we obtain the following sequential algorithm:

$$\theta^{(n+1)} = \theta^{(n)} + \frac{\beta_n}{\sigma^2} e_{n+1} s_{n+1} \tag{25}$$

where

$$e_{n+1} = y_{n+1} - \theta^{(n)T} s_{n+1}. \tag{26}$$

We recognize the algorithm in (25) as the LMS algorithm. This should not be surprising; the LMS method is, in fact, a stochastic gradient algorithm applied to the mean-square error (MSE) criterion, that is

$$E_{\theta_0}\{e^2\} = E_{\theta_0}\{(y_n - \theta^T s_n)^2\}. \tag{27}$$

Hence, minimizing the MSE in this case is equivalent to maximizing the Kullback–Leibler information measure.

Now suppose that s_n is unknown, we only know that it is a realization from a wide sense stationary (WSS) zero-mean Gaussian random process with a prespecified correlation/spectrum function.

We assume that s_n and v_n are statistically independent. The LMS algorithm cannot be applied to this case since the input (reference) signal is not available to us. But, we can still use the algorithm in (4). To derive the score, we shall use (22), where the complete data $\boldsymbol{x}_{n+1}$ is specified by

$$\boldsymbol{x}_{n+1} = \begin{pmatrix} y_{n+1} \\ \boldsymbol{s}_{n+1} \end{pmatrix}. \tag{28}$$

Now

$$\begin{aligned} \log f_X(\boldsymbol{x}_{n+1}; \boldsymbol{\theta}) &= \log f_S(\boldsymbol{s}_{n+1}) + \log f_{Y/S}(y_{n+1}/\boldsymbol{s}_{n+1}; \boldsymbol{\theta}) \\ &= C - \frac{1}{2\sigma^2}(y_{n+1} - \boldsymbol{\theta}^T \boldsymbol{s}_{n+1})^2 \\ &= C - \frac{1}{2\sigma^2}(y_{n+1}^2 - 2\boldsymbol{\theta}^T \boldsymbol{s}_{n+1} y_{n+1} \\ &\quad + \boldsymbol{\theta}^T \boldsymbol{s}_{n+1} \boldsymbol{s}_{n+1}^T \boldsymbol{\theta}) \end{aligned} \tag{29}$$

where C is a constant independent of $\boldsymbol{\theta}$. Substituting (29) into (22) and performing the indicated differentiation and expectation operations, we obtain

$$D \log f_Y(y_{n+1}; \boldsymbol{\theta}^{(n)}) = \frac{1}{\sigma^2}[\hat{\boldsymbol{s}}_{n+1} y_{n+1} - \widehat{\boldsymbol{s}_{n+1}\boldsymbol{s}_{n+1}^T}\boldsymbol{\theta}^{(n)}] \tag{30}$$

where $\hat{\boldsymbol{s}}_{n+1} = E_{\boldsymbol{\theta}^{(n)}}\{\boldsymbol{s}_{n+1}/y_{n+1}\}$, and $\widehat{\boldsymbol{s}_{n+1}\boldsymbol{s}_{n+1}^T} = E_{\boldsymbol{\theta}^{(n)}}\{\boldsymbol{s}_{n+1}\boldsymbol{s}_{n+1}^T/y_{n+1}\}$.

Since $\boldsymbol{s}_{n+1}$ and y_{n+1} are jointly Gaussian, these conditional expectations are readily available:

$$\hat{\boldsymbol{s}}_{n+1} = \frac{y_{n+1}}{\boldsymbol{\theta}^{(n)} P \boldsymbol{\theta}^{(n)} + \sigma^2} P\boldsymbol{\theta}^{(n)} \tag{31}$$

$$\widehat{\boldsymbol{s}_{n+1}\boldsymbol{s}_{n+1}^T} = P + \frac{y_{n+1}^2 - \boldsymbol{\theta}^{(n)T} P \boldsymbol{\theta}^{(n)} - \sigma^2}{(\boldsymbol{\theta}^{(n)T} P \boldsymbol{\theta}^{(n)} + \sigma^2)^2} P\boldsymbol{\theta}^{(n)}\boldsymbol{\theta}^{(n)T} P \tag{32}$$

where $P \triangleq E\{\boldsymbol{s}_n \boldsymbol{s}_n^T\}$. Substituting (31) and (32) into (30)

$$D \log f_Y(y_{n+1}; \boldsymbol{\theta}^{(n)}) = \frac{y_{n+1}^2 - \boldsymbol{\theta}^{(n)T} P \boldsymbol{\theta}^{(n)} - \sigma^2}{(\boldsymbol{\theta}^{(n)T} P \boldsymbol{\theta}^{(n)} + \sigma^2)^2} \cdot P\boldsymbol{\theta}^{(n)}. \tag{33}$$

As an alternative, we may use the algorithm specified by (20) (the indicated cumulative averaging is necessary here). Following straightforward algebraic manipulations, the resulting algorithm is:

$$\boldsymbol{\theta}^{(n+1)} = G_{n+1}^{-1} \boldsymbol{g}_{n+1} \tag{34}$$

where $\boldsymbol{g}_n$ and G_n are computed using the terms calculated in (31) and (32) via the following recursions:

$$\boldsymbol{g}_{n+1} = \gamma_n \boldsymbol{g}_n + \hat{\boldsymbol{s}}_{n+1} y_{n+1} \tag{35}$$

$$G_{n+1} = \gamma_n G_n + \widehat{\boldsymbol{s}_{n+1}\boldsymbol{s}_{n+1}^T}. \tag{36}$$

We may simplify the form of the algorithm by successive substitutions of (31) and (32) into (35) and (36) and then (35) and (36) into (34).

Given the appropriate regularity, these algorithms converge to the true parameter values, and the limit distribution can also be derived by using the results developed in [12] and [13]. As a by-product of these algorithms, we also obtain an on-line estimate of the input signal using (31). The algorithm can easily be extended to include on-line estimation of unknown signal and noise spectral parameters.

References

[1] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1959.

[2] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, 1973.

[3] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," in *Ann. Math. Stat.*, vol. 23, pp. 462-466, 1952.

[4] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400-407, 1951.

[5] J. R. Blum, "Approximation methods which converge with probability one," *Ann. Math. Stat.*, vol. 25, pp. 382-386, 1954.

[6] A. Dvoretzky, "On stochastic approximation, in *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, 1956, pp. 35-56.

[7] V. Dupac, "On the Kiefer Wolfowitz approximation method," *Casopis Pest. Mat.*, vol. 82, pp. 47-75, 1957.

[8] H. J. Kushner, *Stochastic Approximation Methods of Constrained and Unconstrained Systems*. New York: Springer, 1978.

[9] V. Fabian, "On asymptotically efficient recursive estimation," *Ann. Stat.*, vol. 6, pp. 854-866, 1978.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, ser. 39, pp. 1-38, 1977.

[11] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, pp. 95-103, 1983.

[12] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Stat. Soc.*, vol. 46(B), pp. 257-267, 1984.

[13] M. Feder, "Iterative algorithms for parameter estimation with applications to signal processing," Sc.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[14] R. A. Fisher, "Theory of statistical estimation," *Proc. Cambridge Phil. Soc.*, vol. 22, pp. 700-725, 1925.

[15] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Roy. Stat. Soc.*, vol. 44(B), pp. 226-233, 1982.

[16] I. Meilijson, "A fast improvement to the EM algorithm on its own terms," *J. Amer. Stat. Assoc.*, to be published.

[17] M. Segal and E. Weinstein, "A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix of linear dynamic systems," *IEEE Trans. Inform. Theory*, vol. 35, no. 3, pp. 682-687, May 1989.

Weight Adjustment Rule of Neural Networks for Computing Discrete 2-D Gabor Transforms

HONG YAN AND JOHN C. GORE

***Abstract*—Daugman has recently proposed a neural network model for computing the discrete 2-D Gabor transform. We prove here that the weight adjustment rule used in the neural network is equivalent to the use of Jacobi iteration for solving simultaneous linear equations, and we propose more efficient algorithms for solving the problem.**

Introduction

The Gabor transform has proven to be very useful for image compression and analysis [1]-[3]. The computation of the Gabor transform is, however, very complicated since the Gabor elementary functions are not orthogonal to each other. Daugman has recently developed a neural network method for computing the Gabor transform [1]. The network consists of two fixed layers and one adjustable layer. The weights of the fixed layers are related to the Gabor elementary functions only, but the weights of the adjustable layer need to be determined iteratively in order to find an optimal representation of the image. Daugman used a least squares error criterion and a gradient based weight adjustment rule, which may be implemented by using an adaptive control signal that is the difference between a feedforward signal and a feedback signal. We show here that the neural network actually solves a set of simul-

Manuscript received May 28, 1989; revised September 25, 1989.

H. Yan is with the School of Electrical Engineering, University of Sydney, NSW 2006, Australia.

J. C. Gore is with the Department of Diagnostic Radiology, Yale University, New Haven, CT 06510.

IEEE Log Number 9036897.