[10] C. R. Johnson, Jr, "Adaptive IIR filtering: Current results and future directions," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 237–250, 1984.
[11] J. M. Mendel, "Discrete techniques of parameter estimation—The equation error formulation," NY: Dekker, 1973.
[12] I. D. Landau, "Near supermartingales for convergence analysis of recursive identification and adaptive control schemes," *Int. J. Contr.*, vol. 35, no. 2, pp. 197–226, 1982.
[13] ——, "Martingales convergence analysis of adaptive schemes—A feedback approach," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 716–720, June 1982.
[14] C. R. Johnson, Jr., "A convergence proof for a hyperstable adaptive recursive filter," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 745–749, Nov. 1979.
[15] J. R. Treichler, M. G. Larimore, and C. R. Johnson, Jr., "Simple adaptive IIR filtering," in *Proc. 1978 ICASSP*.
[16] C. R. Johnson, I. D. Landau, T. Taylor, and L. Dugard, "On adaptive IIR filters and parallel adaptive identifiers with adaptive error filtering," in *Proc. ICASSP '81*, Atlanta, 1981.
[17] C. R. Johnson, M. G. Larimore, J. R. Treichler, and B. D. O. Anderson, "Sharf convergence properties," *IEEE Trans. Circuits Syst.*, vol. CAS-28, June 1981.
[18] I. D. Landau, L. D. Dugard, and S. Cabrerra, "Applications of output error recursive estimation algorithms for adaptive signal processing," in *Proc. ICASSP '82*, Paris, May 1982.
[19] I. D. Landau, "Sur l'utilisation des techniques d'identification recursive en traitement du signal," in *Proc. GRETSI*, Nice, France, May 16–20, 1983.
[20] C. D. Mote and A. Rahimi, "Real time vibration control of rotating circular plates by temperature control and system identification," in *Adaptive Systems in Control and Signal Processing*. New York: Pergamon Press, 1984.
[21] I. D. Landau, *Adaptive Control—The Model Reference Approach*. New York: Dekker, 1979.
[22] B. Friedlander, "System identification techniques for adaptive signal processing," *Circuits, Syst., Signal Process*, vol. 1, no. 1, pp. 1–41, 1982.
[23] L. Ljung, "On positive real transfer functions and the convergence of some recursive schemes," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 539–551, Aug. 1977.
[24] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, "Discrete time multivariable adaptive control," *IEEE Trans. Automat. Control*, vol. AC-25, pp. 449–456, June 1980.
[25] M. De La Sen and I. D. Landau, "An on-line method for improvement of adaptation transients in adaptive control," in *Adaptive Systems in Control and Signal Processing*. New York: Pergamon Press, 1984.
[26] R. Ortega and I. D. Landau, "On the model process mismatch tolerance of various parameter adaptation algorithms in direct control schemes: A sectoricity approach," in *Adaptive Systems in Control and Signal Processing*. New York: Pergamon Press, 1984.
[27] R. Kosut and B. Friedlander, "Performance robustness properties of adaptive control systems," in *Proc. 21st IEEE-CDC*, Orlando, FL, Dec. 8–10, 1982.

Adaptive Linear Procedures Under General Conditions

LÁSZLÓ GYÖRFI

Abstract—**Under mild conditions on the observation processes the almost sure convergence properties of linear stochastic approximation are summarized for least squares and for some of its applications: adaptive filtering, echo cancellation, detection of binary data in Gaussian noise, identification, and linear classification.**

I. Introduction

In MANY adaptive linear communication tasks a least squares problem has to be solved using a realization of a stationary process. Usually, in addition, it is supposed that the corresponding process is ergodic, in which case the problem is equivalent to a mean square minimization and the solution is the root of the Wiener–Hopf equation. For the sake of computational simplicity, recursive procedures are applied, the strong consistency of which were proved under additional conditions on the dependence structure of the process.

Manuscript received January 27, 1983; revised October 7, 1983.
The author is with the Technical University of Budapest, 1111 Budapest, Stoczek u. 2, Hungary.

If the process is nonergodic then we show that the solution is the root of the generalized Wiener–Hopf equation, where in general the root is random and not unique.

The purpose of this paper is to point out that there exists a simple and general stochastic approximation theorem by which the strong consistency of adaptive procedures can be deduced when the observation process is not ergodic. For illustration the consequences of this theorem are shown in the usual examples of communication: adaptive filtering, echo cancellation, detection of binary data in Gaussian noise, identification, and linear classification.

II. Stochastic Approximation for Dependent Observations

The introduction of Robbins–Monro stochastic approximation [22] resulted in new possibilities for statistical inference from a sequence of random observations. In classical mathematical statistics the only point of view was to get the most efficient inference from the samples with no

0018-9448/84/0300-0262$01.00

regard to their required computational complexity. In real-life problems of communication and control the Robbins–Monro idea opened the research area of defining and analyzing recursive algorithms of much smaller computational complexity, which may therefore work in real time with respect to the generation of the typically high speed observation sequences.

In the classical studies of stochastic approximation ([1], [2], [5], [12], [14], [21], [22]), it was essentially assumed that the observations were independent, which is a very restrictive assumption for almost all practical applications. The main question of interest for proving the strong consistency of stochastic approximation for dependent samples was how to find an efficient technique for evaluating the joint effect of the iteration and the random observation. It turned out that the commonly used devices of conditional expectations, martingales, and other techniques relied upon sufficient conditions, e.g., that the observations are M-dependent or mixing with given rate ([3], [7]). These conditions were thought to be far from necessary in many cases. In that direction Fritz [11], Ljung [18], [19], [20], Kushner and Clark [16], Györfi [13], Eweda and Macchi [6], and Farden [10] made important steps: for linear regression Fritz [11] formulated a completely deterministic problem by which strong consistency can be deduced if for the observation the strong law of large numbers is applied; therefore the conditions of strong consistency of linear stochastic approximation are as general as the conditions of the strong law of large numbers. Thus the effects of the randomness and the iteration were separated. Similar motivations were successful for the celebrated Ljung separation theorem [18] which operates for more general models.

In the sequel applications of linear stochastic approximation are given based on the following theorem.

Theorem: Assume a sequence of $L \times L$ square matrices $A_1, A_2, \cdots$, and a sequence of L-vectors $W_1, W_2, \cdots$, such that

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} A_i = A, \tag{1}$$

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} W_i = W, \tag{2}$$

and

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} \|A_i\|^2 \tag{3}$$

exists and is finite. If A is symmetric and positive definite, and V_n is defined by

$$V_{n+1} = V_n - \frac{1}{n+1}(A_{n+1}V_n - W_{n+1}) \tag{4}$$

(V_0 is arbitrary), then

$$\lim_n V_n = A^{-1}W. \tag{5}$$

This theorem was proved by Györfi [13, th. 1]. Fritz [11] formulated this problem in a Banach space and proved (5) if, for example, in the case of a Hilbert space where (3) is replaced by the restrictive condition that $\|A\| < 1$ and $\|A_i\| < 1$ for $i = 1, 2, \cdots$, Ljung [19, ex. 4] states that as a particular consequence of his separation theorem (5) is valid if (3) is weakened by the condition that $1/n\sum_{i=1}^{n}\|A_i\|$ has finite limit. Although everybody believes that it can be verified, unfortunately there is currently no detailed check of his so called "boundedness condition," therefore we cannot use the more general form of (3). It implies that in the sequel we should assume that the observation process has finite fourth moments instead of second ones.

In some of the applications the recursion (4) has been attempted to accelerate the rate of convergence in the following form:

$$V_{n+1} = V_n - \frac{c_1}{c_2 + n} C_{n+1}(A_{n+1}V_n - W_{n+1}) \tag{6}$$

(V_0 is arbitrary), where c_1 and c_2 are positive constants and C_n is a sequence of matrices. Assume (1), (2), (3), and

$$\lim_n C_n = C. \tag{7}$$

If CA is symmetric and positive definite, then the consistency of (6) is a simple consequence of the theorem (see [13, th. 2]).

In general, (6) is not faster than (4). It may have a better rate of convergence if C has an appropriate relation to A^{-1}, for example, $C = A^{-1}$ ([25], [27]).

The theorem shows the convergence of a deterministic recursive algorithm. For stochastic recursion this theorem guarantees strong consistency on that event on which the conditions of the theorem are met. These conditions are usually verified by a strong law of large numbers.

III. Least Squares

Let $D = (X_1, Y_1), (X_2, Y_2), \cdots$, be a sequence of random pairs such that $\{X_i\}$ are L-dimensional and $\{Y_i\}$ are scalar. An L-vector V is called coefficient vector and the inner product (V, X_i) stands for a linear estimate of Y_i where $i = 1, 2, \cdots$. These estimates are qualified by

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} ((V, X_i) - Y_i)^2 \tag{8}$$

if it exists. The task of least squares problem is to minimize this limit in V.

Assume first that D is (strict sense) stationary with $E\|X_i\|^4 < \infty$ and $E|Y_i|^2 < \infty$; then by the strong law of large numbers for stationary processes we get

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} ((V, X_i) - Y_i)^2 = E\{((V, X_1) - Y_1)^2 / \mathcal{F}\} \quad \text{a.s.,} \tag{9}$$

where $\mathcal{F}$ is the σ-algebra of invariant sets of the stationary process D (see Stout [26, th. 3.5.7]).

One has to emphasize that in the nonergodic cases the least squares problem does not result in the usual mean square minimization. On the one hand, the solution of least squares may be random, but the expected asymptotic error

of the least squares is less than the least mean square, since

$$E\Big(\min_V E\big\{((V, X_1) - Y_1)^2/\mathcal{F}\big\}\Big) \le \min_V E\Big(E\big\{((V, X_1) - Y_1)^2/\mathcal{F}\big\}\Big) = \min_V E\big\{((V, X_1) - Y_1)^2\big\}.$$

Denote by Q_{opt} the set of the solutions V of the following Wiener–Hopf equation

$$E(X_1X_1^T/\mathcal{F})V = E(X_1Y_1/\mathcal{F}), \tag{10}$$

then the vectors V from Q_{opt} minimize the error (9). If D is not ergodic then $E(X_1X_1^T/\mathcal{F})$ may be random, therefore there is no use to assume that $E(X_1X_1^T/\mathcal{F})^{-1}$ exists almost surely, or equivalently that the solution of (10) is unique almost surely.

Let (λ_i, φ_i) $i = 1, 2, \cdots, L$ be an eigensystem of the matrix $E(X_1X_1^T/\mathcal{F})$, i.e., $\varphi_1, \cdots, \varphi_L$ are orthogonal solutions of the equation $E(X_1X_1^T/\mathcal{F})\varphi_i = \lambda_i\varphi_i$, and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_L \ge 0$. Let $0 \le L' \le L$ be the integer for which $\lambda_i > 0$ if $i \le L'$ and $\lambda_i = 0$ if $i > L'$; then the conditional distribution of X_1 given $\mathcal{F}$ is concentrated into the subspace $R^{L'}$ of R^L spanned by $\varphi_1, \cdots, \varphi_{L'}$. Put

$$V = \sum_{i=1}^{L} v_i\varphi_i, \qquad E(X_1Y_1/\mathcal{F}) = \sum_{i=1}^{L} u_i\varphi_i;$$

then $u_i = 0$ almost surely for $i > L'$ and V is a solution of (10) if $v_i\lambda_i\varphi_i = u_i\varphi_i$ where $i = 1, \cdots, L'$. Thus

$$Q_{\text{opt}} = \left\{V = \sum_{i=1}^{L'} v_i\varphi_i;\ v_i = \frac{u_i}{\lambda_i},\ i = 1, \cdots, L'\right\} \qquad \text{a.s.}$$

Based on (8), the obvious estimate of a solution is

$$\tilde{V}_n = \left[\frac{1}{n}\sum_{i=1}^{n} X_iX_i^T + B\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_iY_i\right], \tag{11}$$

where B is an arbitrary positive definite matrix ensuring that the inverse in (11) exists. It is surprising that $\tilde{V}_n$ tends to an element of Q_{opt} almost surely in general, even in the case when $E(X_1X_1^T/\mathcal{F})^{-1}$ does not exist. However, in this latter case the inverse factor in (11) is divergent causing many computational difficulties. The proof of the consistency of (11) is similar to that of the proof of the recursive procedure (14) in the sequel. The consistency is guaranteed by the strong law of large numbers:

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} X_iX_i^T = E(X_1X_1^T/\mathcal{F}) \qquad \text{a.s.,} \tag{12}$$

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} X_iY_i = E(X_1Y_1/\mathcal{F}) \qquad \text{a.s.,} \tag{13}$$

and by the fact that the projection of $\tilde{V}_n$ upon the subspace $R^{L'}$ is convergent almost surely.

Avoiding the computational difficulties of (11), introduce the recursion

$$V_{n+1} = V_n - \frac{1}{n+1}((X_{n+1}, V_n) - Y_{n+1})X_{n+1} \tag{14}$$

(V_0 is arbitrary), the strong consistency of which comes from the theorem for the notations $A_n = X_nX_n^T$, $W_n = X_nY_n$, if $E(X_1X_1^T/\mathcal{F})^{-1}$ exists, since by the strong law of large numbers

$$\lim_n \frac{1}{n}\sum_{i=1}^{n}\|A_i\|^2 = \lim_n \frac{1}{n}\sum_{i=1}^{n}\|X_i\|^4 = E\big(\|X_1\|^4/\mathcal{F}\big) \qquad \text{a.s.} \tag{15}$$

Therefore by (12), (13), and (15) the conditions of the theorem are met, thus

$$\lim_n V_n = E(X_1X_1^T/\mathcal{F})^{-1}E(X_1Y_1/\mathcal{F}) \qquad \text{a.s.}$$

If the inverse of $E(X_1X_1^T/\mathcal{F})$ does not exist then observe that the correction term of (14) is an element of $R^{L'}$ almost surely, since it is a scalar multiplication of X_{n+1} taking values in $R^{L'}$ almost surely given $\mathcal{F}$, therefore by induction

$$V_n = P_{L'}(V_n) + V_0 - P_{L'}(V_0) \qquad \text{a.s.,} \tag{16}$$

where $P_{L'}$ stands for the projection operator upon $R^{L'}$; $P_{L'}(V_n)$ defines an almost sure equivalent of V_n, and the minimization problem has a unique solution in $R^{L'}$. Eq. (16) implies also that the limit V_n has the form

$$\lim_n V_n = \sum_{i=1}^{L'} \frac{u_i}{\lambda_i}\varphi_i + V_0 - P_{L'}(V_0) \qquad \text{a.s.}$$

If D is ergodic and $E(X_1X_1^T)^{-1}$ exists then the rate of convergence of V_n may be much worse than that of $\tilde{V}_n$. In this case the usually proposed accelerated version of (14) is as follows ([25], [27]):

$$V_{n+1}^* = V_n^* - \frac{c_1}{c_2 + n}B_{n+1}^{-1}\big(X_{n+1}X_{n+1}^TV_n^* - X_{n+1}Y_{n+1}\big),$$

where

$$B_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_iX_i^T + B\right)$$

and B is an arbitrary positive definite matrix; B_n^{-1} can be calculated recursively and $\lim_n B_n^{-1} = E(X_1X_1^T)^{-1}$ almost surely. Thus by the consistency of (6) we get $\lim_n V_n^* = E(X_1X_1^T)^{-1}E(X_1Y_1)$ almost surely.

Notice that in the nonergodic case V_n^* tends also to an element of Q_{opt} almost surely, however, B_n^{-1} may be divergent.

The stationary case can be generalized to periodic nonstationary (cyclostationary) observations, for which the multidimensional distributions are invariant for time shifts of multiples of the period T. In this case the sequence D can be decomposed into T stationary processes $D_i = \{(X_{i+jT}, Y_{i+jT})\, j = 1, 2, \cdots\}$ $i = 1, 2, \cdots, T$. Thus, for the error term

$$\lim_n \frac{1}{n}\sum_{i=1}^{n}((X_i, V) - Y_i)^2 = \frac{1}{T}\sum_{i=1}^{T} E\big\{((X_i, V) - Y_i)^2/\mathcal{F}_i\big\} \qquad \text{a.s.,}$$

where $\mathcal{F}_i$ is the σ-algebra of the invariant sets of the stationary process D_i $(i = 1, \cdots, T)$. Here Q_{opt} denotes the

solutions of the equation

$$\left[\frac{1}{T}\sum_{i=1}^{T} E\left(X_i X_i^T/\mathcal{F}_i\right)\right]V = \frac{1}{T}\sum_{i=1}^{T} E\left(X_i Y_i/\mathcal{F}_i\right),$$

and $\tilde{V}_n, V_n, V_n^*$ as defined before tend to an element of Q_{opt} almost surely.

For optimization the knowledge of the period T would be very useful. Applying the decomposition idea let us generate the set of least squares solutions $Q_{i\text{opt}}$ for the observation D_i $(i = 1, \cdots, T)$; then

$$\frac{1}{T}\sum_{i=1}^{T} E\left\{\left(\left(X_i, V_{i\text{opt}}\right) - Y_i\right)^2/\mathcal{F}_i\right\} \le \frac{1}{T}\sum_{i=1}^{T} E\left\{\left(\left(X_i, V_{\text{opt}}\right) - Y_i\right)^2/\mathcal{F}_i\right\} \tag{17}$$

almost surely where $V_{\text{opt}} \in Q_{\text{opt}}$ and $V_{i\text{opt}} \in Q_{i\text{opt}}$, $i = 1, \cdots, T$. For stationary observations the equality in (17) holds almost surely. In contrast to the stationary case for cyclostationary observations with *a priori* known period T it may be worthy to generate T least squares optimization problems instead of one.

The only assumptions on D were (12), (13), and (15) which hold for nonstationary processes, where the nonstationarity is caused by random transients such that (X_n, Y_n) tends in some sense (for example a.s.) to a stationary process.

IV. Examples

Adaptive Filtering

Consider the filtering problem for a communication channel (Fig. 1). Assume that the output process $Z = \{Z_n\}$ of the channel is observed and one has to estimate the input process $Y = \{Y_n\}$ in the form (V, X_n). The ith coordinate of X_n equals to $\phi_i(Z_{n-M+1}, \cdots, Z_{n-M+L^*})$, where the filters ϕ_i are measurable functions of their arguments $(i = 1, \cdots, L)$, $(1 \le M, 1 \le L^*)$. If $M > L^*$, then the filter is called prediction, otherwise it is called interpolation. In typical cases the $\{\phi_i\}$ are linear, for example ([4], [8], [23], [27]), $L = L^*$ and

$$\phi_i(Z_{n-M+1}, \cdots, Z_{n-M+L^*}) = Z_{n-M+i} \qquad i = 1, \cdots, L. \tag{18}$$

If the optimal filtering is formulated as a least square problem, then (14) may have the strong consistency property, provided that (X_n, Y_n) are second order and stationary. However, (14) uses both the input and the output of the channel. An adaptive filter has to operate only with the output of the channel ([15], [23], [27]). For this reason some additional assumptions are made. Suppose that

$$Z_n = Y_n + N_n \tag{19}$$

where $\{Y_n\}$ and $\{N_n\}$ are independent, stationary, and ergodic, and $E(N_n) = 0$. Then

$$E\left(Z_i Y_j\right) = E\left(Y_i Y_j\right), \tag{20}$$

$$E\left(Z_i Z_j\right) = E\left(Z_i Y_j\right) + E\left(N_i N_j\right). \tag{21}$$

For the sake of simplicity assume that ϕ_i $(i = 1, \cdots, L)$ have the form (18). If the statistical properties of the source Y_n are known as far as necessary, namely, $r_{i-j} = E(Y_i Y_j)$ $i, j = 1, 2, \cdots, L$ are known, then using (20) the adaptive filter may be as follows:

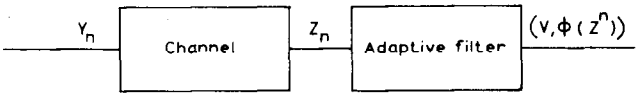

Fig. 1. Adaptive filtering.

$$V_{n+1} = V_n - \frac{1}{n+1}\left(Z^n Z^{n^T} V_n - W^*\right), \tag{22}$$

where $W^* = (E(Y_n Z_{n-M+i}),\ i = 1, 2, \cdots, L)^T = (r_{M-i},\ i = 1, 2, \cdots, L)^T$ and $Z^n = (Z_{n-M+1}, \cdots, Z_{n-M+L})^T$. Eq. (22) depends only on the output of the channel. If the source is unknown and the covariances $\tilde{r}_{i-j} = E(N_i N_j)$ of the channel noise are known, then by (21) V_n has the form (22). Here $W^* = (Z_n Z_{n-M+i} - \tilde{r}_{M-i},\ i = 1, \cdots, L)^T$.

For some communication channels a particular adaptive filtering problem is the adaptive equalization. Here it is usually assumed, that the channel outputs are unknown linear filterings of the inputs with additive noise. If the covariances of the source and noise are known then we have some chance to construct consistent procedures, where the observations are the outputs of the channel. (This is called adaptive equalizer.) However, the problem is still open how to construct an adaptive equalizer of complexity of (22) having strong consistency and not having the run-away property of the usual decision-directed procedures (see Lucky [17]).

Echo Cancellation

Falconer and Mueller [9] investigated the problem of adaptive echo cancellation for the structure illustrated by Fig. 2, which is a model of two-wire full-duplex data transmission. Source_B produces S_n, the output of Channel_B is $\hat{S}_n$, which is corrupted by the echo $\tilde{U}_n$ of the symbol U_n of Source_A. Given $Z_n = \hat{S}_n + \tilde{U}_n$ the task of the Canceller_A is to copy the channel output $\hat{S}_n$ in the form $\tilde{S}_n = Z_n - (V, \phi(U^n))$ and to find the vector V for which

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} (\tilde{S}_i - \hat{S}_i)^2 \tag{23}$$

is minimal. Here the filters ϕ are measurable functions of their arguments $U^n = (U_i,\ i \le n)^T$.

Assume that the sequence $(U_n, \tilde{U}_n, \hat{S}_n)$ is second order and stationary with the invariant σ-algebra $\mathcal{F}$; $(U_n, \tilde{U}_n)$ and $\hat{S}_n$ are independent given $\mathcal{F}$ and $E(\hat{S}_1/\mathcal{F}) = 0$.

$$\tilde{S}_n = \hat{S}_n + \tilde{U}_n - (V, \phi(U^n))$$

and

$$\begin{aligned}
&\lim_n \frac{1}{n}\sum_{i=1}^{n} (\tilde{S}_i - \hat{S}_i)^2 \\
&= E\left((\tilde{S}_1 - \hat{S}_1)^2/\mathcal{F}\right) \\
&= E\left\{\left(\tilde{U}_1 + \hat{S}_1 - (V, \phi(U^1))\right)^2/\mathcal{F}\right\} - E\left(\hat{S}_1^2/\mathcal{F}\right) \\
&= \lim_n \frac{1}{n}\sum_{i=1}^{n} \left(Z_i - (V, \phi(U^i))\right)^2 - E\left(\hat{S}_1^2/\mathcal{F}\right) \qquad \text{a.s.}
\end{aligned} \tag{24}$$

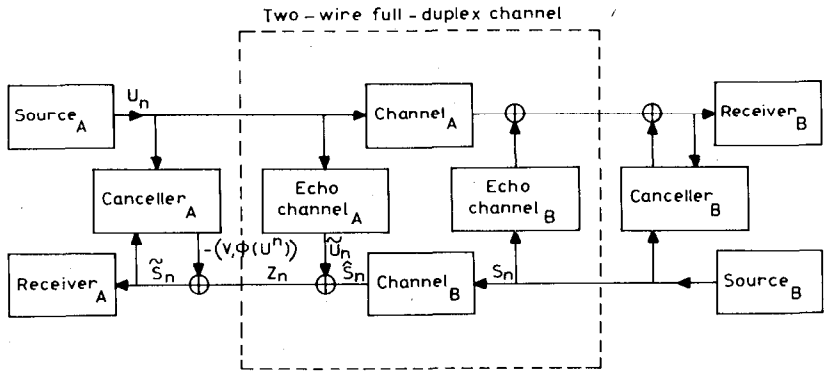

Fig. 2. Adaptive echo canceller.

Thus the minimization of (23) is equivalent to the minimization of the first term of the right-hand side of (24), therefore the procedure (14) has the form

$$V_{n+1} = V_n - \frac{1}{n+1} \cdot \left(\phi(U^{n+1})\phi(U^{n+1})^T V_n - Z_{n+1}\phi(U^{n+1})\right) = V_n + \frac{1}{n+1}\tilde{S}_{n+1}\phi(U^{n+1}),$$

giving a strong consistent estimate for one of the optimal coefficient vectors of Canceller$_A$.

Observe that, as far as the consistency property is concerned, the actual structures of the channels and the echo channels are irrelevant. A good choice of filters ϕ providing an appropriate model for the echo channel is relevant in order to get good cancellation.

Detection of Binary Data in Gaussian Noise

Consider the adaptive detection of binary pulse-amplitude modulation (PAM) signaling in additive Gaussian noise ([15], [23], [28]). The received signal is as follows:

$$r(t) = \sum_n S_n m(t - nT - \tau_n) + \xi(t),$$

where the waveform m is continuous on $(0, T)$ and zero outside of $[0, T]$, $\xi(t)$ is a stationary, ergodic, and zero-mean Gaussian process of almost sure continuous sample functions, $\xi(t)$ is independent of the ± 1 valued, zero-mean source $\{S_n\}$. For the synchronous case take L samples in each bit time T, thus $Z_n = (r(\tau_n + nT + (2i-1)/(2L)T)$ $i = 1, \cdots, L)^T$. If M denotes the vector of the samples of the waveform m

$$M = \left(m\left(\frac{2i-1}{2L}T\right) i = 1, \cdots, L\right)^T,$$

then the optimal detector in the sense of error probability is

$$S_n' = \operatorname{sgn}(V^*, Z_n),$$

where V^* is the solution of the equation $KV = M$. Here K stands for the covariance matrix of the sampled noise $(\xi((i/L)T)\ i = 1, \cdots, L)$. Assume that K^{-1} exists. The adaptive detector uses only the output of the channel. It is possible applying the equality $E(Z_n Z_n^T) = MM^T + K$. Thus for the simple recursion

$$V_{n+1} = V_n - \frac{1}{n+1}\left(\left(Z_{n+1}Z_{n+1}^T - MM^T\right)V_n - M\right)$$

we have

$$\lim_n V_n = K^{-1}M \qquad \text{a.s.}$$

For the asynchronous case assume a consistent synchronization, i.e., τ_n' denotes an estimate of τ_n such that

$$\lim_n (\tau_n' - \tau_n) = 0 \qquad \text{a.s.,} \tag{25}$$

then the corresponding samples of the received signals are $Z_n' = (r(\tau_n' + nT + (2i-1)/(2L)T),\ i = 1, \cdots, L)^T$ and the recursion is

$$V_{n+1}' = V_n' - \frac{1}{n+1}\left(\left(Z_{n+1}'Z_{n+1}'^T - MM^T\right)V_n' - M\right).$$

Although Z_n' is nonstationary, V_n and V_n' have the same limit almost surely if

$$\lim_n \frac{1}{n}\sum_{i=1}^n \left(Z_i Z_i^T - Z_i' Z_i'^T\right) = 0 \qquad \text{a.s.}$$

or

$$\lim_n \frac{1}{n}\sum_{i=1}^n \|Z_i - Z_i'\|^2 = 0 \qquad \text{a.s.}$$

This comes from the Toeplitz theorem since by (25) applying the continuity assumptions on m and ξ we have almost surely $\lim_n \|Z_n - Z_n'\|^2 = 0$.

Identification

Denote by $\{Z_n\}$ and $\{Y_n\}$ the inputs and the outputs of a dynamic system, resp. Let the identification task be the best approximation of the output $\{Y_n\}$ in the form $(V, \phi(Z^n, Y^{n-1}))$, where ϕ is an L-vector valued measurable function of its argument $Z^n, Y^{n-1} = \{Z_i, Y_j,\ i \le n,\ j \le n-1\}$. The optimization problem is formulated as the minimization of

$$\lim_n \frac{1}{n}\sum_{i=1}^n \left(Y_i - \left(V, \phi(Z^i, Y^{i-1})\right)\right)^2,$$

which is also a simple least squares problem, and it might be solved by an algorithm like (14) for $X_n = \phi(Z^n, Y^{n-1})$. Here the consistency property does not depend on the actual structure of the system to be identified. However,

for choosing appropriate function ϕ usually some assumptions are made on the structure of the system ([19], [24], [27]). The typical structure is $Y_n = (V^*, \phi(Z^n, Y^{n-1})) + N_n$ where ϕ is either a linear function for a linear model or it is a vector with components being polynomials of Volterra series for nonlinear models.

For the observation it is enough to suppose that $\{\phi(Z^n, Y^{n-1}), Y_n\}$ is second order and stationary, however, for identification it may make sense to deal with the last problem of Section III, for which $\{Z_n, Y_n\}$ is not stationary, but (Z_n, Y_n) tends in some sense to a stationary state.

Linear Classifier

The classification problem is given by the stationary training sample $D = \{(Z_1, Y_1), (Z_2, Y_2), \cdots\}$, where $\{Z_n\}$ are L'-vectors and $\{Y_i\}$ take the values ± 1. For an L-vector V the decision on Y_n is defined $\operatorname{sgn}(V, \phi(Z_n))$. Here ϕ: $R^{L'} \to R^L$ is a measurable function. The task is to minimize the asymptotic error rate

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} I_{\{\operatorname{sgn}(V, \phi(Z_i)) \neq Y_i\}} = P\{\operatorname{sgn}(V, \phi(Z_1)) \neq Y_1 / \mathcal{F}\},$$

where $\mathcal{F}$ is the invariant σ-algebra of D and I stands for the indicator. Unfortunately, this random function of V may have several local minima. The error rate has a simple upper bound

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} I_{\{\operatorname{sgn}(V, \phi(Z_i)) \neq Y_i\}} \leq \lim_n \frac{1}{n} \sum_{i=1}^{n} ((V, \phi(Z_i)) - Y_i)^2,$$

the minimization of which is a least squares problem ([3], [27]).

References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, *Potential Function Method in the Theory of Machine Learning*, (in Russian). Moscow: Nauka, 1972.

[2] J. R. Blum, "Multidimensional stochastic approximation methods," *Ann. Math. Stat.*, vol. 25, pp. 737–744, 1954.

[3] S. Csibi, *Stochastic Processes with Learning Properties*. Wien–New York: Springer-Verlag, 1975.

[4] L. D. Davisson, "A theory of adaptive filtering," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 97–102, Mar. 1966.

[5] A. Dvoretzky, "On stochastic approximation," in *Proc. Third Berkeley Symp. Math. Stat. and Prob.*, Univ. of Calif., vol. 1. pp. 39–55, 1956.

[6] E. Eweda and O. Macchi, "Convergence of an adaptive linear estimation algorithm," *IEEE Trans. Automat. Contr.*, Feb. 1984.

[7] ——, "Quadratic mean and almost-sure convergence of unbounded stochastic approximation algorithms with correlated observations," *Annales Institut Henri Poincaré*, vol. 19. no. 3, pp. 235–255, Sept. 1983.

[8] D. D. Falconer, "Adaptive filter theory and applications," in *Analysis and Optimization of Systems*, A. Bensoussan and J. L. Lions, Eds. Berlin: Springer-Verlag, 1980, pp. 163–188.

[9] D. D. Falconer and K. H. Mueller, "Adaptive echo cancellation/AGC structures for two-wire, full-duplex data transmission," *Bell. Syst. Tech. J.*, vol. 58. pp. 1593–1616, Sept. 1979.

[10] D. C. Farden, "Stochastic approximation with correlated data," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 105–113, Jan. 1981.

[11] J. Fritz, "Learning from an ergodic training sequence," in *Limit Theorems of Probability Theory*, P. Révész, Ed. Amsterdam: North-Holland, 1974, pp. 79–91.

[12] E. G. Gladysev, "On the stochastic approximation," *Theory Prob. Appl.*, vol. 10, pp. 297–300, 1965.

[13] L. Györfi, "Stochastic approximation from ergodic samples for linear regression," *Z. für Wahrscheinlichkeitstheorie*, vol. 54, pp. 47–55, 1980.

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

[15] H. J. Kushner, "Adaptive techniques for the optimization of binary detections signals," *IEEE Intern. Conv. Rec.*, vol. 11, pt. 4, 1963.

[16] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer-Verlag, 1978.

[17] R. W. Lucky, "A survey of the communication theory literature: 1968–1973," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 725–739, Nov. 1973.

[18] L. Ljung, "Convergence of recursive stochastic algorithms," Rep. 7403, Division of Automatic Control, Lund Institute of Technology, Sweden, Feb. 1974.

[19] ——, "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 551–575, Aug. 1977.

[20] ——, "Strong convergence of a stochastic approximation algorithm," *Ann. Statist.*, vol. 6, pp. 680–696, 1978.

[21] M. B. Nevel'son and R. Z. Khasminskii, *Stochastic Approximation and Recursive Estimation*, (in Russian). Moscow: Nauka, 1972.

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

[23] D. J. Sakrison, "Stochastic approximation: a recursive method for solving regression problems," in *Adv. in Comm. Systems*, A. V. Balakrishnan, Ed. New York: Academic, 1966, pp. 51–106.

[24] G. N. Saridis, "Stochastic approximation methods for identification and control—A survey," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 798–809, Dec. 1974.

[25] G. N. Saridis, Z. J. Nikolic, and K. S. Fu, "Stochastic approximation algorithms for system identification, estimation and decomposition of mixtures," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, pp. 8–15, 1969.

[26] W. F. Stout, *Almost Sure Convergence*. New York: Academic Press, 1974.

[27] Y. Z. Tsypkin, *Adaptation and Learning in Automatic Systems*. New York: Academic Press, 1971.

[28] I. Vajda, "Adaptive Gaussian detection without *a priori* symbol synchronization," *Prob. Control and Inform. Theory*, vol. 9, pp. 133–140, 1980.