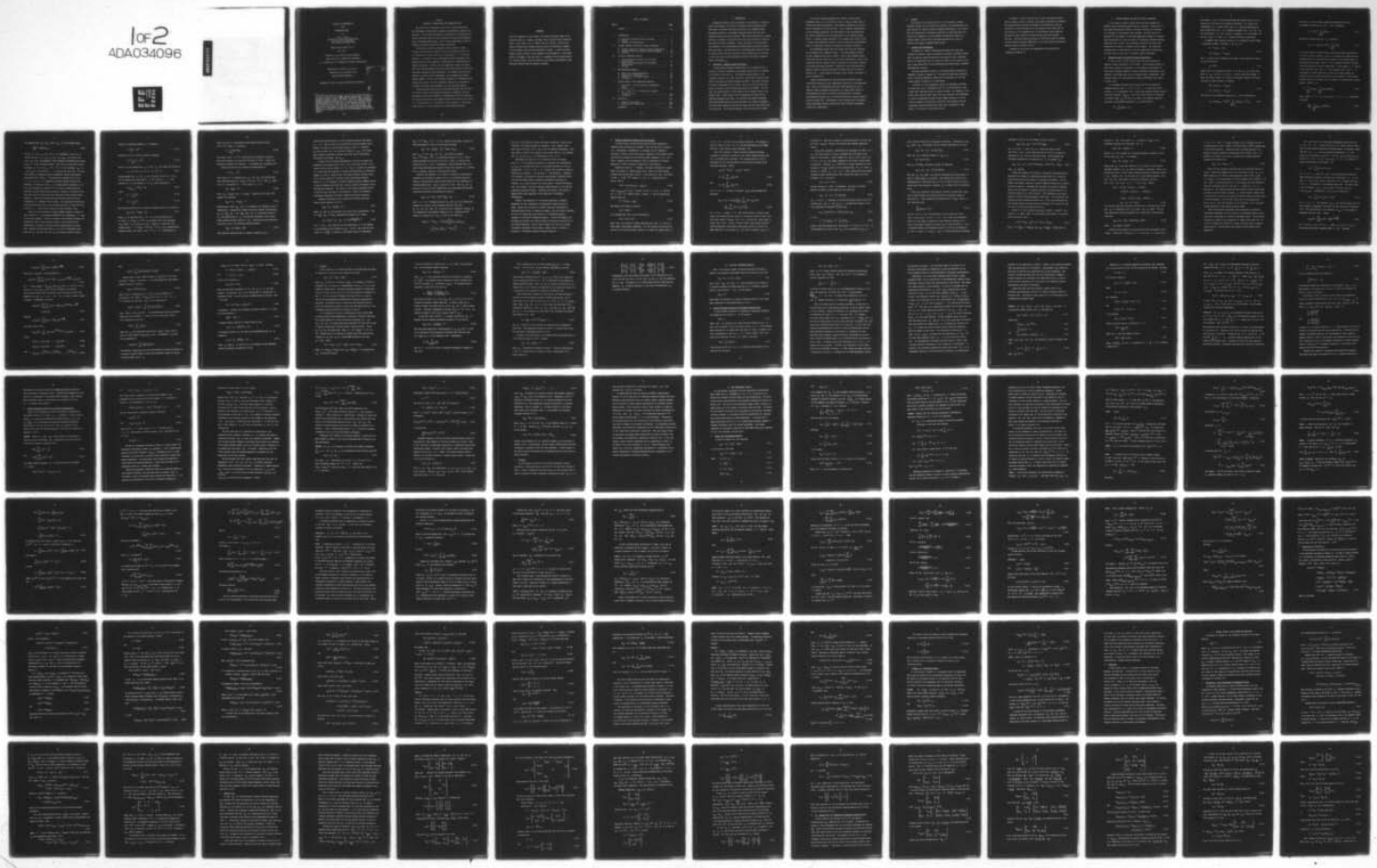
AD-A034 096 COLORADO STATE UNIV FORT COLLINS DEPT OF ELECTRICAL --ETC F/G 12/1
STOCHASTIC APPROXIMATION WITH CORRELATED DATA.(U)
MAY 75 D C FARDEN
UNCLASSIFIED TR-11(ONR) N00014-67-A-0299-0019
NL



STOCHASTIC APPROXIMATION

WITH

CORRELATED DATA[1]

by

David C. Farden
Department of Electrical Engineering
Colorado State University
Fort Collins, Colorado 80523

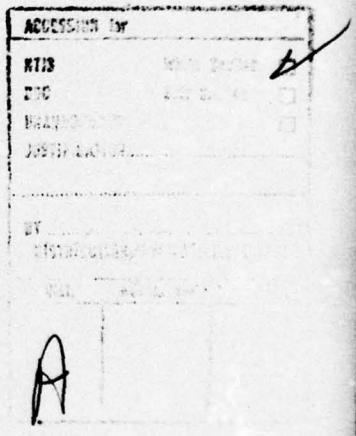
ONR Technical Report No. 11

May, 1975

Prepared for the Office of Naval Research

under Contract No. N00014-67-A-0299-0019;

L. L. Scharf and M. M. Siddiqui, Principal Investigators



Approved for public release; distribution unlimited.

[1]This work was supported in part by the Office of Naval Research, Arlington, Virginia, under Grant N00014-67-A-0299-0019, in part by the Naval Undersea Center, San Diego, California, under contracts N66001-72-C-0479 and N00123-73-C-1375, and by the Advanced Research Projects Agency of the Department of Defense and monitored by the Naval Undersea Center, under Contract No. N66001-74-C-0035. The main body of this report is the author's Ph.D. dissertation, submitted to the Department of Electrical Engineering, Colorado State University, in partial fulfillment of the requirements for the Ph.D. degree.

ABSTRACT

STOCHASTIC APPROXIMATION WITH CORRELATED DATA

New almost sure convergence results for a special form of the multidimensional Robbins-Monro stochastic approximation procedure are developed. The special form treated is motivated by a consideration of several algorithms that have been proposed for discrete time adaptive signal processing applications. Most of these algorithms can also be viewed as stochastic gradient-following algorithms.

Essentially, previous convergence results contain a common "conditional expectation condition" which is extremely difficult (if not impossible) to satisfy when the "training data" is a correlated sequence. In contrast, the new convergence results developed in the present work are easily applied to cases where the "training data" is heavily correlated. In fact, the new convergence results are applicable when certain moments exist and certain "decay rates" on two autocovariance functions can be established. For example, when the data sequence is normal and (i) M-dependent, (ii) autoregressive moving average (ARMA), or (iii) can be viewed as samples of a bandlimited continuous time process, the new convergence results can be applied to establish the almost sure convergence of each algorithm treated.

Several special forms of data correlation matrices that are shown to arise in discrete time signal processing are examined. New computationally efficient procedures are developed for both the inversion of a matrix having one of the treated special forms and for the solution of a corresponding set of simultaneous linear equations. The special forms treated are termed Toeplitz and block Toeplitz matrices.

ADDENDUM

Since the completion of this report, the author has become aware of the excellent paper by H. Akaike, entitled "Block Toeplitz Matrix Inversion" (SIAM J. Appl. Math., Vol. 24, March 1973, pp. 234-241). Most of the results treating block Toeplitz matrices which are developed in Chapter V of the present work have been developed by Akaike. In case the block Toeplitz matrix involved is both symmetric and persymmetric, a case which arises, for example, when each block of a symmetric block Toeplitz matrix is a Toeplitz matrix, then the results of the present work provide a more efficient solution than the results of Akaike.

TABLE OF CONTENTS

I. INTRODUCTION

Although stochastic signal processing can be viewed as a branch of time series analysis, the desire to implement simple sequential real-time signal processing structures motivates one to approach signal processing problems in a decidedly different manner than one would approach related time series problems. This work is devoted to a unified analytical treatment of algorithms that have been proposed for discrete time adaptive signal processing. These algorithms are treated within the framework of the multidimensional Robbins-Monro stochastic approximation procedure. The special form of the Robbins-Monro procedure which is treated herein and the convergence results obtained are of interest in their own right, having applications outside the realm of adaptive signal processing.

A. Motivation: Adaptive Signal Processing

In many signal processing applications, the ultimate goal is to provide an "optimal" estimate of some signal process which is imbedded in an additive noise process. The physical implementation of the "optimal" estimator (or filter structure) requires that certain parameters of the signal and noise processes be known. The filter structure is usually constrained to be a causal, linear structure and the optimality criterion is often minimum mean square error (MMSE). For this case, the optimal filter is well-known to be the Wiener filter or the Kalman filter. These filters can be implemented provided that the required parameters are known. For discrete time signal processing with uncorrelated signal and noise processes, the required parameters are those which completely specify the signal and the noise autocorrelation sequences. The required parameter set may or may not be finite.

In case the required parameters are unknown, identification techniques (see e.g., [1],[2]) can be used, at least in some cases, to estimate the desired parameters. The estimated parameters can then be used to implement the required filter. Due to inherent uncertainties in the estimated parameters, the performance of the resulting filter can differ dramatically from the performance of the desired optimal filter. A closely related approach is to constrain the filter to have a certain fixed suboptimal structure and to estimate the corresponding family of parameters required to implement the simpler structure.

An interesting concept that has evolved from the latter approach is the concept of an "adaptive filter." The term "adaptive filter" is used throughout this work to denote a filter which designs itself, either from the raw input data, or from some training data. Many of the algorithms used for adaptive signal processing are stochastic versions of gradient-following procedures. Significant early contributions to adaptive signal processing were made by Widrow and Hoff [3], and by Sakrison [4]. A more complete treatment of the relevant literature is given in Chapter II.

Primary considerations in the application of adaptive signal processing techniques are the convergence properties of the algorithms used. Most of the algorithms which have been proposed for use in adaptive signal processing applications are slight modifications of multidimensional versions of either the Robbins-Monro stochastic approximation procedure [5] or the Kiefer-Wolfowitz stochastic approximation procedure [6]. Unfortunately, many proposed uses for adaptive signal processing involve processes for which available convergence results are inapplicable.

B. Purpose

The purpose of the present work is to (i) establish a unified framework suitable for the analytical treatment of algorithms which have been proposed for adaptive signal processing applications, (ii) investigate the probabilistic convergence properties of algorithms which fall within this framework, and (iii) examine the detailed structure of several special forms of data correlation matrices that arise in discrete time signal processing applications.

C. Contents and Organization

In Chapter II, several representative systems that have been proposed for adaptive signal processing are reviewed, including systems used for adaptive channel equalization and adaptive array processing. Most of the algorithms that are treated in Chapter II are shown to fall into a specialized form of the multidimensional Robbins-Monro stochastic approximation procedure.

Existing convergence results for the Robbins-Monro procedure are examined in detail in Chapter III. The need for additional analytical work to establish meaningful probabilistic convergence for the algorithms treated in Chapter II is established.

In Chapter IV, new convergence results are developed, providing an almost sure (a.s.) convergence proof for a certain family of algorithms under conditions which are easily verified. For example, in the normal case, when the input signal and noise processes are M-dependent, or stable autoregressive moving average (ARMA) processes, or can be viewed as samples of bandlimited continuous time processes, the new convergence results establish the almost sure convergence of each member of the family of algorithms treated.

In Chapter V, several special forms of data correlation matrices that are shown to arise in discrete time signal processing are examined. New computationally efficient procedures are developed for both the inversion of a matrix having one of the treated special forms and for the solution of a corresponding set of simultaneous linear equations. The special forms treated are termed Toeplitz and block Toeplitz matrices. The new procedures represent an efficient method for designing the desired suboptimal MMSE filter in case the required correlation sequence values are known *a priori*.

A summary of new results and suggestions for future work is presented in Chapter VI.

II. SYSTEMS PROPOSED FOR ADAPTIVE SIGNAL PROCESSING

In this chapter, several systems which have been proposed for adaptive signal processing applications are reviewed. In Section II-A the channel equalization problem is treated; Section II-B is devoted to a treatment of the adaptive array problem. The main point to be developed in this chapter is that many algorithms proposed for adaptive signal processing fall into the realm of "stochastic gradient-following algorithms" and, as such, the convergence properties of these algorithms may be treated in a somewhat unified manner. The literature reviewed here is representative of the most significant contributions in recent years on the topic of "adaptive signal processing."

A. Systems Proposed for Adaptive Channel Equalization

In this section, several systems which have been proposed for adaptive channel equalization are reviewed. The motivating problem, to which these systems are applicable, is the automatic equalization of voice-grade telephone channels to reduce intersymbol interference, thus enabling a much higher data rate for digital signal transmission. Such channels usually are characterized as having a moderately high signal-to-noise ratio.

It is assumed throughout this section that for the equivalent baseband system at time $t = kT$, $k = 0,1,2,\ldots$, a real-valued random variable a_k is transmitted into a linear time-invariant channel having unit pulse response $\{h_k\}_{k=-\infty}^{\infty}$. The output of the channel is corrupted by additive noise, n_k, and fed to the input of an adaptive equalizer. The input to the adaptive equalizer, x_k, is thus given by

$$x_k = \sum_{\ell=-\infty}^{\infty} a_\ell h_{k-\ell} + n_k \ . \tag{2.1}$$

The sequence $\{a_k\}$ is the information-bearing sequence which is to be estimated by the output of the adaptive equalizer. For digital data transmission, a_k is chosen from a set of M discrete amplitudes via some probabilistic rule. It is assumed throughout that $\{a_k\}$ and $\{n_k\}$ are uncorrelated, i.e., that $E\{a_k n_\ell\} = E\{a_k\}E\{n_\ell\}$ for all integer k,ℓ, and that $E\{n_k\} = 0$, where $E\{\cdot\}$ denotes statistical expectation.

A commonly used equalizer structure is a transversal filter having p adjustable weights. Defining W and X_k by

$$W' = (w_1, w_2, \ldots, w_p) \quad ,$$
$$X_k' = (x_k, x_{k-1}, \ldots, x_{k-p+1}) \quad , \tag{2.2}$$

where ' denotes matrix transpose, the output of the transversal equalizer can be written as

$$y_k = W'X_k \quad . \tag{2.3}$$

Suppose that it is desired to choose W so that y_k is a "best estimate" of $a_{k-\alpha}$ for all $k = \alpha, \alpha+1, \ldots$, and for some fixed integer α. There have been a number of "criteria of goodness" proposed for characterizing the "best estimate." Defining

$$H_k' = (h_k, h_{k-1}, \ldots, h_{k-p+1}) \quad ,$$
$$N_k' = (n_k, n_{k-1}, \ldots, n_{k-p+1}) \quad , \tag{2.4}$$

the output of the transversal equalizer, y_k, can be expressed as

$$y_k = W'H_\alpha (a_{k-\alpha} + (W'H_\alpha)^{-1} \sum_{\substack{\ell=-\infty \\ \ell \neq k-\alpha}}^{\infty} a_\ell W'H_{k-\ell}) + W'N_k \quad . \tag{2.5}$$

From (2.5), it can be readily seen that the distortion due to intersymbol interference at time $t = kT$ is given by

$$I_k = (W'H_\alpha)^{-1} \sum_{\substack{\ell=-\infty \\ \ell \neq k-\alpha}}^{\infty} a_\ell W'H_{k-\ell} \quad . \tag{2.6}$$

One easily obtained bound for I_k is given by

$$|I_k| \leq B = \max_\ell |a_\ell| \; |W'H_\alpha|^{-1} \sum_{\substack{m=-\infty \\ m \neq \alpha}}^{\infty} |W'H_m| \quad . \tag{2.7}$$

It is noted that for channels having severe intersymbol interference, B may be infinite; however, in case B is infinite the channel has the interpretation of an unstable linear system in the bounded output for all bounded input sense. Lucky [7] has considered automatic equalization from the point of view of minimizing B with respect to W subject to the constraint that $W'H_\alpha = 1$. The constraint that $W'H_\alpha = 1$ is convenient for digital detection in that the decision regions (or slicing levels) can remain fixed under this constraint. The procedure proposed by Lucky [7] makes use of a sequence of isolated unit training pulses. Define D by

$$D = \sum_{\substack{m=-\infty \\ m \neq \alpha}}^{\infty} |W'H_m| = \sum_{\substack{m=-\infty \\ m \neq \alpha}}^{\infty} W'H_m \; \mathrm{sgn}(W'H_m), \tag{2.8}$$

where $\mathrm{sgn}(\gamma) = 1$ if $\gamma \geq 0$ and $\mathrm{sgn}(\gamma) = -1$ if $\gamma < 0$. Noting that (formally)

$$\frac{\partial D}{\partial w_i} = \sum_{\substack{m=-\infty \\ m \neq \alpha}}^{\infty} h_{m-i+1} \; \mathrm{sgn}(W'H_m) \quad , \tag{2.9}$$

and assuming that $h_k \approx c\delta_{k,o}$, where $\delta_{k,\ell}$ is the Kronecker delta,

$$\frac{\partial D}{\partial w_i} \approx c \ \mathrm{sgn}(W'H_{i-1}) \quad , \tag{2.10}$$

for all $i = 1,2,\ldots,p$ and $i \neq \alpha + 1$. Furthermore, from (2.1) it follows that for $a_\ell = \delta_{\ell,0}$, $x_k = h_k + n_k \approx c\delta_{k,o} + n_k$; hence, from (2.3), $y_k \approx W'H_k$. Consequently, Lucky considers incrementing the weight vector, W, by the following scheme: after each test pulse has arrived increment w_i by $-\mu \ \mathrm{sgn}\ y_{i-1}$ for $i \neq \alpha + 1$, and increment $w_{\alpha+1}$ by $-\mu \ \mathrm{sgn}\ (y_\alpha - 1)$. The constant $\mu > 0$ is termed the step size. For channels capable of supporting binary transmission without equalization, Lucky shows that $|W'H_{i-1}|$ is asymptotically bounded by 2μ for all $i = 1,2,\ldots,p, i \neq \alpha + 1$, assuming an infinite signal-to-noise ratio. Similarly, he shows that $|W'H_\alpha - 1|$ is asymptotically bounded by 2μ. In [8], Lucky extends the results of [7] to obtain a decision-directed adaptive equalizer which does not require a sequence of isolated training pulses and can "track" slow time variations in the channel characteristics. Lucky also investigates what has since been called the "probability of a runaway" for his system. The equalizers introduced by Lucky have also been called "zero forcing equalizers" in that they tend to force $p - 1$ zeros in the overall unit pulse response $W'H_k$.

Gersho [9] has considered a scheme somewhat similar in nature to that of Lucky [7]. He considers minimizing the deterministic ℓ^2 norm of the error sequence. The error sequence is the difference between the deterministic part of the equalizer output and the desired output, assuming that a sequence of known isolated training pulses is being sent. Suppose for the moment that n_k in (2.1) is identically zero. Then with y_k given by (2.3) and d_k the desired equalizer output,

Gersho [9] considers choosing W to minimize

$$\xi = \sum_{\ell}(y_\ell - d_\ell)^2 \quad . \tag{2.11}$$

Motivated by (2.11), Gersho considers minimizing

$$\xi_k = \sum_{\ell \epsilon J_k} (y_\ell - d_\ell)^2 \quad , \tag{2.12}$$

where it is not assumed that $n_k \equiv 0$, and J_k is an index set defined by

$$J_k = \{\ell_o + k\xi,\ \ell_o + k\xi + 1, \ldots, \ell_o + k\xi + \kappa\} \quad . \tag{2.13}$$

Gersho assumes that x_ℓ and d_ℓ are virtually zero for all $\ell \notin J_k$ for an isolated unit pulse sent at $t = k\xi$, and that $\xi > \kappa$. The gradient of ξ_k with respect to $W = W_k$ can be expressed as

$$\nabla_W \xi_k \Big|_{W=W_k} = 2F_k W_k - 2P_k \quad , \tag{2.14}$$

where

$$F_k = \sum_{\ell \epsilon J_k} X_\ell X_\ell' \quad , \tag{2.15}$$

and

$$P_k = \sum_{\ell \epsilon J_k} d_\ell X_\ell \quad . \tag{2.16}$$

The resulting algorithm for "training" the weight vector, W, is given by

$$W_{k+1} = W_k - \mu(F_k W_k - P_k) \quad , \tag{2.17}$$

where W_o is arbitrary and $\mu > 0$. It is worth noting that for $R = E\{F_k\}$, $P = E\{P_k\}$, $w = R^{-1}P$ is the weight vector that minimizes $E\{\xi_k\}$, where ξ_k is given by (2.12). Gersho [9] shows that for a suitably small $\mu > 0$, $E\{(W_k - R^{-1}P)'(W_k - R^{-1}P)\}$ can be asymptotically bounded by some $\epsilon(\mu)$, where $\epsilon(\mu) \to 0$ as $\mu \to 0$. Furthermore, he

points out that for increasingly large signal-to-noise ratios, $R^{-1}P \to A^{-1}P$, where A is given by

$$A = \sum_{\ell \in J_k} E\{X_\ell\}E\{X_\ell'\} \quad . \tag{2.18}$$

The weight vector $W = A^{-1}P$ characterizes the equalizer structure which will minimize the noise-free criterion of (2.11). Gersho also discusses techniques for choosing μ to maximize the convergence rate.

Niessen and Willim [10] consider the minimization of

$$\xi = E\{[y_k - a_k]^2\} \tag{2.19}$$

with respect to W, assuming that $\{a_k\}$ and $\{n_k\}$ are jointly wide-sense stationary and that $E\{a_k n_\ell\} = 0$ for all $k \neq \ell$. With y_k given by (2.3), the gradient of ξ with respect to W is

$$\nabla_W \xi = 2R_{xx}W - 2P \quad , \tag{2.20}$$

where $R_{xx} = E\{X_k X_k'\}$ and $P = E\{a_k X_k\}$. Equations (2.19) and (2.20) suggest the algorithm

$$W_{k+1} = W_k - \mu(R_{xx}W_k - P) \quad . \tag{2.21}$$

Unfortunately, since R_{xx} and P are assumed to be unknown, (2.21) is inapplicable. Consequently, Niessen and Willim consider approximating R_{xx} by $X_k X_k'$ and P by $y_k^* X_k$, where y_k^* is a quantized version of y_k. The quantization of y_k is performed according to the *a priori* known possible discrete levels of $\{a_k\}$. Substituting these approximations into (2.21), the following algorithm results:

$$W_{k+1} = W_k - \mu X_k(y_k - y_k^*) \quad . \tag{2.22}$$

This algorithm represents what is commonly referred to as a

"decision-directed equalizer" in that decisions which are made about $a_{k-\alpha}$ (i.e. y_k^*) are used in the algorithm to train the weight vector. Clearly, in order for the algorithm given by (2.22) to "converge," y_k^* must initially be a very reliable estimate of $a_{k-\alpha}$. The convergence analysis performed by Niessen and Willim [10] is essentially deterministic and assumes $y_k^* = a_{k-\alpha}$.

In order for the strategy represented by (2.22) to be useful for moderately low signal-to-noise ratios (viz., less than 30 dB), a constraint such as used by Lucky [7], i.e., $W'H_\alpha = 1$, seems to be essential. It is noted that the technique of Niessen and Willim does not inherently require an initial "setup period" with known isolated training pulses, and it can be capable of "tracking" slowly time-varying channels. George *et al.* [11] consider a decision feedback strategy somewhat similar to that of Niessen and Willim [10], using an adaptive transversal filter following the quantizer. The output of this second transversal filter is fed back into the input of the quantizer. Monsen [12] presents a performance comparison of decision feedback and linear equalizers.

Schonfeld and Schwartz [13] consider the following algorithm which is quite similar to (2.17):

$$W_{k+1} = W_k - \alpha_k(F_k W_k - P_k) \quad , \tag{2.23}$$

where F_k and P_k are given by (2.15) and (2.16), respectively. With $R = E\{F_k\}$ and $P = E\{P_k\}$, Schonfeld and Schwartz [13] choose

$$\alpha_k = 2[(\lambda_u + \lambda_\ell) - (\lambda_u - \lambda_\ell)\cos(\frac{(2k+1)\pi}{2N})]^{-1} \tag{2.24}$$

for $k = 0,1,\ldots,N-1$, where all of the eigenvalues of R are assumed to be contained in the interval $[\lambda_\ell,\lambda_u]$. In [13], they show that this choice of $\{\alpha_k\}_{k=0}^{N-1}$ is optimal in the minimax sense for minimizing

$E\{W_N - R^{-1}P\}'E\{W_N - R^{-1}P\}$. In [14], Schonfeld and Schwartz extend the above philosophy to obtain a second-order algorithm:

$$W_{k+1} = W_k - \alpha_k(F_kW_k - P_k) + \beta_k(W_k - W_{k-1}) \quad , \tag{2.25}$$

where $\beta_o = 0$ and $\{\alpha_j\}$ and $\{\beta_j\}$ are chosen to minimize $E\{W_k - R^{-1}P\}'E\{W_k - R^{-1}P\}$ in the minimax sense for all k. Both of these algorithms ((2.24) and (2.25)) force $E\{W_k\}$ to converge more rapidly than e.g., (2.17). Consequently, these algorithms seem to be useful when equalizing high signal-to-noise ratio channels in a training mode by sending a sequence of isolated known pulses.

Kosovych and Pickholtz [15] consider a successive overrelaxation technique for training the weight vector of a transversal equalizer during a training period using isolated pulses for the minimization of the mean-squared error $E\{\xi_k\}$, ξ_k given by (2.12). With F_k and P_k given by (2.15) and (2.16), respectively, the overrelaxation algorithm considered by Kosovych and Pickholtz is given by

$$W_{k+1} = W_k - \omega(D_k - \omega E_k)^{-1}(F_kW_k - P_k) \quad , \tag{2.26}$$

where $\omega > 0$ is a "relaxation factor," D_k and E_k are, respectively, diagonal and strictly lower triangular matrices, such that $F_k = D_k - E_k - E_k^+$. Here E_k^+ is strictly upper triangular, leaving D_k to be composed of the diagonal elements of F_k. Denoting the ij^{th} element of a matrix A by $(A)_{i,j}$, (2.26) can be written as

$$(W_{i+1})_{i,1} = (W_k)_{i,1} - \omega(F_k)^{-1}_{i,i}\left\{\sum_{j=1}^{i-1}(F_k)_{i,j}(W_{k+1})_{j,1} + \sum_{j=i}^{p}(F_k)_{i,j}(W_k)_{j,1} - (P_k)_{i,1}\right\} \quad . \tag{2.27}$$

Note that (2.27) does not require any matrix inversions. Kosovych and Pickholtz [15] discuss methods for choosing ω and compare convergence rates of (2.26), (2.17), and (2.25) via computer simulations. They also obtain a bound on the asymptotic mean square error in W_k, assuming that F_k, F_ℓ and P_k, P_ℓ are independent for all $k \neq \ell$.

Recalling that from (2.3) it is desired to train the weight vector, W, of a transversal equalizer to "optimize" the approximation $y_k \approx a_{k-\alpha}$, it is noted that most of the systems discussed so far have assumed that $p = 2N + 1$ and $\alpha = N$. Qureshi [16] presents an adaptive technique for choosing α and training W simultaneously. Kobayashi [17] presents a more general technique using maximum likelihood estimation and the Robbins-Monro stochastic approximation procedure to estimate $\{a_n\}$, sample timing, and carrier phase. Walzman and Schwartz [18], [19] present a discrete frequency domain approach to the adaptive transversal equalizer problem. Benedetto and Biglieri [20] discuss a Kalman filter theory approach to the reduction of intersymbol interference.

Finally, the importance of the Viterbi algorithm to sequence estimation for data transmitted over dispersive channels should be noted. Forney [21] introduces a receiver structure consisting of a whitened matched filter, a symbol-rate sampler, and the Viterbi algorithm. In [21] it is shown that this structure is a maximum-likelihood estimator of the entire transmitted sequence. Qureshi and Newhall [22] and Magee and Proakis [23] discuss adaptive structures which make use of the Viterbi algorithm. Both of these schemes ([22] and [23]) include an adaptive transversal filter having a weight vector, W, which is trained by a stochastic gradient-following algorithm.

B. Systems Proposed for Adaptive Array Processing

In this section, several systems which have been proposed for adaptive array processing are reviewed. Data from an array of sensors (e.g., hydrophones, seismometers, or antennas) can be "optimally" processed to reject certain directional components of the observed field and provide an estimate of some desired signal component (e.g., [24]-[30]). Adaptive array processing is used to compensate for varying degrees of *a priori* statistical ignorance in such problems.

Consider an array of L sensors, each sensor followed by a tapped delay line having M equally spaced taps. Denote the delay between adjacent taps on each delay line by D, and denote by $x_\ell(t)$ the output at time t of the ℓ^{th} sensor, $\ell = 1,2,\ldots,L$. Define the $ML \times 1$ matrix $X(t)$ by

$$X'(t) = (x_1(t),x_2(t),\ldots,x_{ML}(t)) \quad , \tag{2.28}$$

where $x_{\ell+(m-1)L}(t) = x_\ell(t - (m-1)D)$ for all $\ell = 1,2,\ldots,L$ and for all $m = 1,2,\ldots M$. Define the $ML \times 1$ matrix W (the so called array weight vector) by

$$W' = (w_1,w_2,\ldots,w_{ML}) \quad . \tag{2.29}$$

The output of the array is given by

$$y(t) = W'X(t) \quad . \tag{2.30}$$

It is assumed that $X(t)$ can be expressed as

$$X(t) = S(t) + N(t) \quad , \tag{2.31}$$

where $S(t)$ is a vector of signal components and $N(t)$ is a vector of noise and/or interference components. For the purposes of this section, the goal of the array processor design is to choose the weight vector, W,

so that $y(t)$ will have certain desired properties. For example, W might be chosen so that $y(t)$ is a minimum mean-square error (MMSE) estimate of some desired signal component, $d(t)$.

Shor [31] considers a simple stochastic gradient-following technique to maximize an estimate of the output signal-to-noise ratio for a narrowband array processor. The technique given in [31] is presented here for the more general array structure defined in the preceding paragraph. Define

$$s_{out}(t) = W'X(t), \quad n_{out}(t) = W'N(t) \quad ,$$

$$s_k = \frac{1}{T} \int_{(k-1)T}^{kT} s_{out}^2(t)dt \quad , \tag{2.32}$$

and

$$n_k = \frac{1}{T} \int_{(k-1)T}^{kT} n_{out}^2(t)dt \quad , \tag{2.33}$$

for $k = 1,2,\ldots$. In order to maximize s_k/n_k, Shor considers the algorithm

$$W_{k+1} = W_k + \lambda(s_k/n_k)\left\{\frac{2}{s_k T} \int_{(k-1)T}^{kT} S(t)\, s_{out}(t)dt \right.$$

$$\left. - \frac{2}{n_k T} \int_{(k-1)T}^{kT} N(t)\, n_{out}(t)dt\right\} \quad , \tag{2.34}$$

for $k = 1,2,\ldots$, where $\lambda > 0$. Shor advises using a "strong" target signal with characteristics similar to the desired signal so that, when the target signal is present, $S(t) \approx X(t)$, and when the target signal is absent, $N(t) \approx X(t)$. Using a "strong" target signal during alternate T-second intervals, and using an approximate version of (2.34), one might hope that on the average, W_k will tend to increase s_k/n_k for

increasing k. Shor also considers an algorithm similar to (2.34) with the factor (s_k/n_k) removed, and presents some computer simulation results.

Lacoss [32] considers a simplified array processor for which $M = 1$, i.e., the array output is simply a weighted sum of the data at the output of the sensors. Lacoss assumes that $x_\ell(t) = s(t) + n_\ell(t)$ for $\ell = 1,2,\ldots,L$; i.e., the signal component at the output of each sensor is identical. Defining $R_{nn} = E\{N(t)N'(t)\}$, Lacoss considers the minimization of $W'R_{nn}W$ subject to the constraint that $W'1_L = 1$, where 1_L is the L x 1 matrix $1_L = (1,1,\ldots,1)'$. This criterion has been termed "minimum variance distortionless look" because the output for such a processor, y(t), is given by

$$y(t) = s(t) + W'N(t) \quad , \tag{2.35}$$

and the variance of $W'N(t)$ is minimized. By using a projected gradient technique, Lacoss shows that the algorithm

$$W_{k+1} = W_k - \mu_k(I - \frac{1}{L} 1_L 1_L')R_{nn}W_k \quad , \tag{2.36}$$

for $k = 0,1,2,\ldots$, converges to the desired optimal weight vector, W*, provided that $W_o'1_L = 1$ and $\sum_k \mu_k = \infty$. An important property of the above algorithm is obtained by noting that

$$R_{xx} = E\{X(t)X'(t)\} = E\{s^2(t)\}1_L 1_L' + R_{nn} \quad , \tag{2.37}$$

so that

$$(I - \frac{1}{L} 1_L 1_L')R_{xx} = (I - \frac{1}{L} 1_L 1_L')R_{nn} \quad , \tag{2.38}$$

where it has been assumed that $E\{s(t)n_\ell(t)\} = 0$ for all $\ell = 1,2,\ldots,L$. The importance of (2.38) is that R_{nn} may be replaced by R_{xx} in

(2.36) without affecting the convergence properties. Consequently, when R_{nn} and/or R_{xx} are unknown, one may consider algorithms of the form

$$W_{k+1} = W_k - \mu_k(I - \frac{1}{L} 1_L 1_L') R_k W_k \quad , \tag{2.39}$$

where R_k is an unbiased estimate of R_{xx}, e.g.,

$$R_k = X(kT)X'(kT) \quad . \tag{2.40}$$

With $y_k = X'(kT)W_k$, one might consider the algorithm

$$W_{k+1} = W_k - \mu_k(I - \frac{1}{L} 1_L 1_L')X(kT)y_k \quad . \tag{2.41}$$

Note that y_k and $X(kT)$ are directly available from the processor, so that no "target signal" is required. One problem which arises in the implementation of algorithms such as (2.41) is that roundoff and quantization errors can accumulate, enabling W_k to wander from the constraint plane.

Frost [33] considers a more general constraint problem than Lacoss, with an added feature that deviations from the constraint plane are corrected for. Frost considers the minimization of $W'R_{xx}W$ subject to the constraints that

$$\sum_{i=1}^{L} w_{i+(m-1)L} = g_m \quad , \tag{2.42}$$

for all $m = 1,2,\ldots,M$. Frost assumes, as does Lacoss [32], that $x_\ell(t) = s(t) + n_\ell(t)$ for all $\ell = 1,2,\ldots,L$, so that the constraints given by (2.42) imply a constraint on the frequency response of the array to any signal component arriving from the same direction as $s(t)$. In obvious notation, the constraints given by (2.42) may be expressed as $C'W = g$, where $g' = (g_1,g_2,\ldots,g_M)$. A projected gradient algorithm,

analogous to (2.36), for the problem at hand is given by

$$W_{k+1} = W_k - \mu_k(I - C(C'C)^{-1}C')R_{xx}W_k \quad , \tag{2.43}$$

for $k = 0,1,2,\ldots$, with $C'W_o = g$. Frost [33] adds the term $C(C'C)^{-1}(g - C'W_k)$ to the right hand side of (2.43) to correct for deviations of W_k from the constraint plane. Frost proposes the following algorithm for the adaptation of W for unknown R_{xx}:

$$W_{k+1} = W_k - \mu(I - C(C'C)^{-1}C')X(kT)y_k + C(C'C)^{-1}(g - C'W_k), \tag{2.44}$$

where $y_k = W_k'X(kT)$.

Winkler and Schwartz [34] propose a stochastic projected gradient algorithm for finding the constrained optimum point for a concave or convex objective function subject to nonlinear constraints. In [35] Winkler and Schwartz consider a similar problem by making use of penalty function techniques. Kobayashi [36] discusses the method of steepest descent and the method of conjugate gradients with projection for the iterative design of an array processor. Such techniques can be quite useful for the off-line processing of array data. It is noted that the adaptive technique proposed by Frost (viz. (2.44)) can be deduced from the steepest descent procedure given by Kobayashi [36] in much the same way that (2.44) can be deduced from (2.43).

Widrow *et al.* [37] consider minimizing $E\{(d(kT) - y(kT))^2\}$ with respect to W, where $d(kT)$ is some desired array output. In terms of obvious notation, define

$$\xi(W) \triangleq E\{(d_k - y_k)^2\} = \sigma^2 - 2P'W + W'R_{xx}W \quad , \tag{2.45}$$

where $\sigma^2 = E\{d_k^2\}$, $P = E\{d_kX_k\}$, and $R_{xx} = E\{X_kX_k'\}$. Noting that the

gradient of $\xi(W)$ with respect to W is given by $2R_{xx}W - 2P$, a reasonable algorithm for minimizing $\xi(W)$ is

$$W_{k+1} = W_k - \mu(R_{xx}W_k - P) \quad . \tag{2.46}$$

Widrow *et al.* [37] consider the following stochastic version of (2.46) for use when R_{xx} and P are unknown:

$$W_{k+1} = W_k - \mu X_k(y_k - d_k) \quad . \tag{2.47}$$

Noting that d_k is the only quantity in (2.47) which is not directly available (indeed, it is d_k which one wishes to estimate), Widrow *et al.* propose the use of a "pilot signal" having statistical properties similar to d_k. Suppose $g(t)$ is the output of a pilot-signal generator, that $g(t)$ and $d(t)$ have similar statistical properties, and that $d(t) = s_1(t - \beta_1) = s_2(t - \beta_2) = \ldots = s_L(t - \beta_L)$. Define

$$\begin{aligned} X_1'(t) = (&g(t+\beta_1),\ g(t+\beta_2),\ldots,\ g(t+\beta_L),\\ &g(t+\beta_1-D),\ g(t+\beta_2-D),\ldots,\ g(t+\beta_L-D),\\ &\ldots\\ &g(t+\beta_1-(M-1)D),\ g(t+\beta_2-(M-1)D),\ldots,\\ &g(t+\beta_L-(M-1)D)). \end{aligned} \tag{2.48}$$

The two-mode adaptation procedure proposed in [37] involves using (2.47) with $d_k \equiv 0$ alternately with $X_k = X_1(kT)$ and $d_k = g(kT)$. The one-mode adaptation procedure proposed in [37] makes use of the following algorithm:

$$W_{k+1} = W_k - \mu(X_k + X_1(kT))(y_k^* - g(kT)) \quad , \tag{2.49}$$

where $y_k^* = W_k'(X_k + X_1(kT))$.

Griffiths [38] proposed an algorithm which does not require a pilot signal. Assume that $E\{d(t)n_\ell(\tau)\} = 0$ for all real t,τ and for all

$\ell = 1,2,\ldots,L$. Then $P = E\{d_k X_k\} = E\{d_k(S_k + N_k)\} = E\{d_k S_k\}$, so that P is appropriately called a signal correlation vector, which is independent of the noise statistics. Considering that if enough statistics are known to be able to generate an appropriate pilot signal $g(t)$, P is probably also known, one is led to consider the following algorithm proposed by Griffiths [38]:

$$W_{k+1} = W_k - \mu(X_k y_k - P). \tag{2.50}$$

Tack [39] has proposed an algorithm that is intimately related to (2.50). Suppose the weight vector is to be trained so that y_k is an MMSE estimate of the additive (nonpropagating) sensor noise $n_1(kT)$. If $\{n_1(kT)\}$ is an uncorrelated or "white" sequence with $E\{n_1^2(kT)\} = \sigma_n^2$ and $n_1(kT)$ is uncorrelated with all other signal and noise components, then the algorithm given by (2.50) with $P' = \sigma_n^2(1,0,0,\ldots,0)$ is appropriate. The resulting array has been termed a spatial innovations processor since the "goal" of making y_k a white sequence implies that a cancellation of all of the spatially correlated signal and noise fields is being attempted. Tack [39] has shown that the resulting weight vector can be a very good indicator of the "bearings" of all the propagating components of the signal and noise fields.

While the previously discussed array processors are inherently time-domain approaches, the next system to be discussed lends an interpretation of processing in "frequency-wavenumber space." The following discussion is based on the presentation of Scharf and Farden [40]. In [40] the treatment was limited to a linear (in-line) array of equally spaced sensors. The discussion here applies to more general array geometries.

Let $\xi_q(t,x,y,z)$ be a real-valued homogeneous random field for $q = 1,2,\ldots,Q$. Let t denote time and (x,y,z) denote spatial coordinates in some suitable cartesian coordinate system. Furthermore, assume that the $\xi_n(.,.,.,.)$ are zero mean and uncorrelated, i.e., that $E\{\xi_n(t_1,x_1,y_1,z_1)\xi_m(t_2,x_2,y_2,z_2)\} = 0$ for all $t_1,t_2,x_1,x_2,y_1,y_2,z_1,z_2$ and for all $n \neq m$. Let $p_\ell = (x_\ell,y_\ell,z_\ell)$, $\ell = 1,2,\ldots,L$, denote the spatial coordinates of the sensors. Let

$$x_\ell(t) = \sum_{q=1}^{Q} \xi_q(t,p_\ell) + n_\ell(t) \quad , \tag{2.51}$$

for $\ell = 1,2,\ldots,L$, where the $n_\ell(t)$ are real-valued zero mean wide sense stationary stochastic processes and $E\{n_\ell(t)n_k(\tau)\} = 0$ for all real t,τ and for all $1\leq k,\ell\leq L$ such that $k \neq \ell$. Suppose that each of the ξ_q corresponds to a propagating plane wave. Then there exists a set of constants $\{\beta_{\ell,q}: 1\leq\ell\leq L, 1\leq q\leq Q\}$ such that $\xi_q(t,p_\ell)$ $= \xi_q(t-\beta_{\ell,q},p_1)$. Consequently, (2.51) can be rewritten as

$$x_\ell(t) = \sum_{q=1}^{Q} \xi_q(t-\beta_{\ell,q},p_1) + n_\ell(t) \tag{2.52}$$

for $\ell = 1,2,\ldots,L$. The constants $\{\beta_{\ell,q}\}$ are clearly functions of the array geometry, propagation velocities, and the "directions of propagation." The relationships of the constants $\{\beta_{\ell,q}\}$ to the concept of wavenumber should be clear. Define

$$z_\ell(f_n,kT) = \sum_{m=0}^{M-1} x_\ell((k-1+\tfrac{m}{M})T)e^{-j2\pi\frac{mn}{M}} \quad , \tag{2.53}$$

for $n = 0,1,2,\ldots,M-1$, $k = 1,2,\ldots,$ where $f_n = n/T$, i.e., $z_\ell(.,.)$ is the discrete Fourier transform (DFT) of $x_\ell(\cdot)$. Defining

$$\eta_\ell(f_n,kT) = \sum_{m=0}^{M-1} n_\ell((k-1+\tfrac{m}{M})T)e^{-j2\pi\frac{mn}{M}} \quad , \tag{2.54}$$

from (2.52) $z_\ell(f_n,kT)$ can be expressed as

$$z_\ell(f_n,kT) = \sum_{m=0}^{M-1}\sum_{q=1}^{Q} \xi_q((k-1+\tfrac{m}{M})T - \beta_{\ell,q},p_1)e^{-j2\pi\frac{mn}{M}} + \eta_\ell(f_n,kT). \tag{2.55}$$

For T "large enough," $E\{z_{\ell_1}(f_{n_1},kT)\bar{z}_{\ell_2}(f_{n_2},kT)\} \approx 0$ for all $n_1 \neq n_2$ and for all $1\leq \ell_1,\ell_2 \leq L$, so that for any criterion of optimality involving only second order statistics, one can process the data independently for each f_n, $n = 0,1,\ldots,M-1$. The $-$ is used to denote complex conjugate. Furthermore, for large T,

$$z_\ell(f_n,kT) \approx \sum_{m=0}^{M-1}\sum_{q=1}^{Q} \xi_q((k-1+\tfrac{m}{M})T,p_1)e^{-j2\pi(f_n\beta_{\ell,q} + \frac{mn}{M})} + \eta_\ell(f_n,kT) \quad . \tag{2.56}$$

Defining

$$y_q(f_n,kT) = \sum_{m=0}^{M-1} \xi_q((k-1+\tfrac{m}{M})T,p_1)e^{-j2\pi\frac{mn}{M}} \quad , \tag{2.57}$$

one easily obtains that

$$z_\ell(f_n,kT) \approx \sum_{q=1}^{Q} y_q(f_n,kT)e^{-j2\pi f_n\beta_{\ell,q}} + \eta_\ell(f_n,kT) \quad . \tag{2.58}$$

Define

$$Z_k'(f_n) = (z_1(f_n,kT), \ \ldots \ , z_L(f_n,kT)) \quad , \tag{2.59}$$

$$N_k'(f_n) = (\eta_1(f_n,kT), \ \ldots \ , \eta_L(f_n,kT)) \quad , \tag{2.60}$$

and

$$\bar{D}_q'(f_n) = (e^{j2\pi f_n\beta_{1,q}}, e^{j2\pi f_n\beta_{2,q}}, \ \ldots \ , e^{j2\pi f_n\beta_{L,q}}) \quad . \tag{2.61}$$

Then

$$Z_k(f_n) \approx \sum_{q=1}^{Q} y_q(f_n,kT)D_q(f_n) + N_k(f_n) \quad . \tag{2.62}$$

Suppose that a linear MMSE estimate of $y_1(f_n,kT)$ of the form $\hat{y}_1(f_n,kT) = W'Z_k(f_n)$ is desired. It is easily shown that the desired weight vector, W^*, is given by

$$W^*(f_n) = R_z^{-1}(f_n)P_1(f_n) \quad , \tag{2.63}$$

where $R_z(f_n) = E\{\overline{Z}_k(f_n)Z_k'(f_n)\}$, $P_1(f_n) = E\{y_1(f_n,kT)\overline{Z}_k(f_n)\} = \sigma_1^2(f_n)\overline{D}_1(f_n)$, and $\sigma_q^2(f_n) = E\{|y_q(f_n,kT)|^2\}$ for $q = 1,2,\ldots,Q$. It is of interest to note that $R_z(f_n)$ can be expressed as

$$R_z(f_n) = \sigma_\eta^2(f_n)I + \sum_{q=1}^{Q} \sigma_q^2(f_n)\overline{D}_q(f_n)D_q'(f_n) \quad , \tag{2.64}$$

where $\sigma_\eta^2(f_n) = E\{|\eta_\ell(f_n,kT)|^2\}$. The Sherman-Morrison matrix inversion lemma [41] can be applied Q times to (2.64) to show that $W^*(f_n)$ can be expressed in the form

$$W^*(f_n) = \sum_{q=1}^{Q} \gamma_q \overline{D}_q(f_n) \quad , \tag{2.65}$$

where the γ_q are complicated functions of $\sigma_\eta^2(f_n)$, $\sigma_q^2(f_n)$, and all pairs of inner products $\overline{D}_q'(f_n)D_p(f_n)$ [40]. Consequently, $\hat{y}_1(f_n,kT)$ can be expressed as

$$\hat{y}_1(f_n,kT) = \sum_{q=1}^{Q} \gamma_q \overline{D}_q'(f_n)Z_k(f_n) \quad . \tag{2.66}$$

The operation $\overline{D}_q'(f_n)Z_k(f_n)$ has the interpretation of being the output of a discrete frequency domain conventional beamformer steered to provide a distortionless look at ξ_q.

Suppose for the moment that the $D_q(f_n)$ are known. Defining

$$D' = (\bar{D}_1(f_n), \bar{D}_2(f_n), \ldots, \bar{D}_Q(f_n)) \quad , \tag{2.67}$$

and

$$\Gamma' = (\gamma_1, \gamma_2, \ldots, \gamma_Q) \quad , \tag{2.68}$$

(2.66) can be rewritten as

$$\hat{y}_1(f_n, kT) = \Gamma' D Z_k \quad , \tag{2.69}$$

where the notational dependence of D, Γ, and Z_k on f_n has been dropped. The operation DZ_k can be interpreted as a spatial DFT, as discussed in [40]. One may now pose the MMSE problem as follows: find Γ such that

$$e(\Gamma) = E\{|\Gamma' D Z_k - y_1(f_n, kT)|^2\} \tag{2.70}$$

is minimized. Invoking the orthogonal projection theorem, Γ^* is seen to be the solution to

$$\bar{D} R_z D' \Gamma - \bar{D} P_1 = 0 \quad . \tag{2.71}$$

A steepest descent solution is readily found as [40]

$$\Gamma_{k+1} = \Gamma_k - \mu_k \bar{D}(R_z D' \Gamma_k - P_1) \quad . \tag{2.72}$$

A stochastic version of (2.72) that can be implemented when R_z is unknown is

$$\Gamma_{k+1} = \Gamma_k - \mu_k \bar{D}(\bar{Z}_k y_k - P_1) \quad , \tag{2.73}$$

where $y_k = Z_k' D' \Gamma_k$. In case the D_q are unknown, one may implement several strategies, as mentioned in [40].

C. Critique

In this section, it is shown that most of the algorithms discussed in Sections II-A and II-B can be written in the form

$$W_{k+1} = W_k + \mu_k(P_k - F_k W_k) \quad , \tag{2.74}$$

where W_k is a real p x 1 matrix, $\{\mu_k\}_{k=1}^{\infty}$ is a sequence of positive constants, P_k is a real p x 1 random matrix, and F_k is a p x p real symmetric random matrix. Detailed convergence results for algorithms that may be cast into the form of (2.74) are presented in Chapters III and IV. It is also shown in Chapter III that (2.74) is a special case of the multidimensional Robbins-Monro stochastic approximation procedure. The purpose here is to show that the algorithm given by (2.74) is sufficiently general to ensure the wide applicability of the convergence results presented in Chapters III and IV.

It is convenient to start by considering a rather general MMSE filtering problem, and establishing a hierarchy of adaptive algorithms for varying degrees of *a priori* statistical ignorance [42]. Let $\{S_k\}$ and $\{N_k\}$ be jointly wide-sense stationary R^p-valued (R^p is used to denote p-dimensional Euclidean space) random processes. Define $X_k = S_k + N_k$, and assume that $E\{N_k\} = 0$ and $E\{S_k N_\ell'\} = 0$ for all k,ℓ. Suppose that it is desired to estimate some real-valued linear function of S_k, say s_k, by a linear MMSE estimate of the form $y_k = W'X_k$. Define

$$\xi(w) = E\{(s_k - y_k)^2\} = E\{s_k^2\} - 2w'P + w'R_{xx}w \quad , \tag{2.75}$$

where $P = E\{s_k X_k\} = E\{s_k S_k\}$, and $R_{xx} = E\{X_k X_k'\}$. It is assumed that R_{xx} is positive definite.

A recursive method for computing the $w = w_o = R_{xx}^{-1} P$ that minimizes $\xi(w)$ is the gradient descent algorithm:

$$w_{k+1} = w_k - \mu_k(R_{xx}w_k - P) \quad , \tag{2.76}$$

where $\mu_k > 0$. This algorithm provides an alternative to computing $w_o = R_{xx}^{-1} P$. The steepest descent algorithm is easily obtained from (2.76) by choosing μ_k to minimize $\xi(w_{k+1})$. The steepest descent algorithm is given by (2.76) with [43]

$$\mu_k = \frac{(R_{xx}w_k - P)'(R_{xx}w_k - P)}{(R_{xx}w_k - P)'R_{xx}(R_{xx}w_k - P)} \quad . \tag{2.77}$$

Note that by letting $P_k = P$, $F_k = R_{xx}$, and μ_k as in (2.77), (2.74) becomes the steepest descent algorithm. In order to make use of gradient descent algorithms such as (2.76), R_{xx} and P must be known *a priori*. Efficient techniques for solving $R_{xx}w = P$ for $w = w_o$ are treated in Chapter V for several special forms of R_{xx}.

In case the "pilot vector," P, is known *a priori* but R_{xx} is unknown, one may consider stochastic versions of (2.76) such as

$$W_{k+1} = W_k - \mu_k(X_kX_k'W_k - P) \quad . \tag{2.78}$$

Note that with appropriate interpretations of μ_k, X_k, and P, (2.78) is the algorithm (2.50) proposed by Griffiths [38], and that with $F_k = X_kX_k'$, $P_k = P$, (2.74) becomes (2.78). Furthermore,

$$F_k = \frac{1}{M} \sum_{\ell=k-M+1}^{k} X_\ell X_\ell' \tag{2.79}$$

and $P_k = P$ in (2.74) is also a reasonable algorithm to consider in this case.

Now, consider the case for which neither R_{xx} nor P is known *a priori*. For this case, one may consider algorithms of the form

$$W_{k+1} = W_k - \mu_k(X_k X_k' W_k - s_k X_k) \quad . \tag{2.80}$$

With suitable interpretations of μ_k, X_k, s_k, (2.80) is the algorithm (2.47) proposed by Widrow *et al.* [37], or algorithm (2.22) proposed by Niessen and Willim [10]. With F_k and P_k given by (2.15) and (2.16), respectively, (2.74) becomes the algorithm (2.17) proposed by Gersho [9], or algorithm (2.23) proposed by Schonfeld and Schwartz [13].

Other algorithms, although not fitting into the MMSE philosophy or directly into the stochastic gradient following philosophy, can, in some cases, be cast into the form of (2.74). With $F_k = (I - \frac{1}{L} 1_L 1_L')R_k$, where $E\{R_k\} = R_{xx}$ and $P_k = 0$, (2.74) becomes the algorithm (2.39) proposed by Lacoss [32]. With $F_k = (D_k - \omega E_k)^{-1} F_k^*$, $P_k = (D_k - \omega E_k)^{-1} P_k^*$, and $\mu_k = \omega$, (2.74) becomes the algorithm (2.26) considered by Kosovych and Pickholtz [15]. With

$$F_k = X_k X_k' - C(C'C)^{-1}C'(X_k X_k' + I) \quad , \tag{2.81}$$

and $P_k = C(C'C)^{-1}g$, (2.74) becomes the algorithm (2.44) proposed by Frost [33]. The algorithms proposed by Lucky [7] and Shor [31] do not fit the class of algorithms given by (2.74).

A simple trick can be used to put complex-valued algorithms such as (2.73) into the form of (2.74). Consider

$$\Gamma_{k+1} = \Gamma_k - \mu_k(R_k \Gamma_k - P) \quad , \tag{2.82}$$

where R_k is Hermitian non-negative definite. Using the superscripts r and i to denote real and imaginary parts, respectively, it is easily shown that

$$\begin{bmatrix} \Gamma_{k+1}^r \\ \Gamma_{k+1}^i \end{bmatrix} = \begin{bmatrix} \Gamma_k^r \\ \Gamma_k^i \end{bmatrix} - \mu_k \left\{ \begin{bmatrix} R_k^r & -R_k^i \\ R_k^i & R_k^r \end{bmatrix} \begin{bmatrix} \Gamma_k^r \\ \Gamma_k^i \end{bmatrix} - \begin{bmatrix} P^r \\ P^i \end{bmatrix} \right\} . \quad (2.83)$$

Consequently, with some obvious definitions, (2.82) (and hence (2.73)) can be put into the form of (2.74), with W_k real, F_k real and symmetric, and P_k real. Furthermore, it is easily shown that the resulting real symmetric F_k is positive definite if and only if the Hermitian R_k is positive definite.

III. EXISTING CONVERGENCE RESULTS

Most of the adaptive signal processing algorithms discussed in Chapter 2 are sequential algorithms which can be written in the form

$$W_{n+1} = W_n + \mu_n (P_n - F_n W_n), \tag{3.1}$$

where $E\{F_n\} = R_{xx}$ and $E\{P_n\} = P$. This algorithm can be viewed as a stochastic gradient-following algorithm or as a stochastic approximation to the solution, $w = w_o$, of the equation

$$R_{xx} w = P. \tag{3.2}$$

This chapter is devoted to a review of existing results on the convergence properties of algorithms similar to (3.1).

A. Strong Convergence Results for Stochastic Approximation

In 1951, Robbins and Monro [5] presented a sequential technique for estimating the solution, θ, of the equation

$$M(x) = \alpha, \tag{3.3}$$

where $M(x)$ is a monotone real valued function defined for all real x and (3.3) is assumed to have the unique solution $x = \theta$. In the Robbins-Monro procedure it is assumed that the nature of $M(x)$ is unknown, and that corresponding to each real x is a random variable $Y(x)$ with distribution function $\Pr[Y(x) \leq y] = H(y|x)$ such that

$$M(x) = \int_{-\infty}^{\infty} y \, dH(y|x) \, . \tag{3.4}$$

The procedure starts with $X_1 = x_1$ an arbitrary real number and proceeds via the recursion

$$X_{n+1} = X_n + a_n(\alpha - Y_n) \ , \tag{3.5}$$

where Y_n is a random variable having the conditional distribution $\Pr[Y_n \leq y | X_n = x_n] = H(y|x_n)$, and $\{a_n\}$ $(n \geq 1)$ is a sequence of positive constants such that

$$\sum_{n=1}^{\infty} a_n = \infty \ , \quad \sum_{n=1}^{\infty} a_n^2 < \infty \ . \tag{3.6}$$

It should be obvious that (3.1) is a multidimensional version of (3.5). Under the additional conditions that $M(x)$ is nondecreasing, $\left.\frac{dM(x)}{dx}\right|_{x=\theta} > 0$, and $Y(x)$ is bounded with probability one for all real x, Robbins and Monro [5] proved that $\lim_{n\to\infty} E\{(X_n-\theta)^2\} = 0$.

Since the pioneering work of Robbins and Monro, a great deal of work has been done on establishing conditions for which schemes similar to (3.5) converge. Kiefer and Wolfowitz [6] considered the problem of estimating the value of $x = \theta$ such that $M(x)$ is a maximum. Blum [44] proved almost sure (a.s.) convergence (i.e., $\Pr[\lim_{n\to\infty} x_n = \theta] = 1$) for both the Robbins-Monro and the Kiefer-Wolfowitz procedures under less restrictive conditions than those in [5] and [6]. In 1954, Blum [45] presented multidimensional versions of both the Robbins-Monro and the Kiefer-Wolfowitz procedures, and proved a.s. convergence for each. Dvoretzky [46] presented a general stochastic approximation procedure which contains the Robbins-Monro and the Kiefer-Wolfowitz procedures as special cases. Dvoretzky proved both mean-square (m.s.) and a.s. convergence for his procedure. Wolfowitz [47] presented a vastly simplified proof of Dvoretzky's Theorem. In 1959, Derman and Sacks [48] gave a simple proof for the a.s. convergence of the multidimensional version

of Dvoretzky's procedure. The interested reader is referred to the excellent review papers by Schmetterer [49,50] and Sakrison [51] for a more complete account of the developments in stochastic approximation.

Essentially, all of the above mentioned works contain a common assumption which, for our application to multidimensional adaptive signal processing, severely limits the effectiveness of the results. The assumption under scrutiny is the following: in (3.5) it is assumed that the conditional distribution of Y_n given $X_n=x_n$ coincides with the distribution of $Y(x_n)$ for all real fixed parameter values x_n. In particular, this assumption implies that $E\{Y_n|X_n = x_n\} = E\{Y(x_n)\}$ $(=M(x_n))$. In terms of the algorithm (3.1), this would require that $E\{F_nW_n-P_n|W_n=w\} = E\{F_nw - P_n\}$, for all fixed (parameter) w in p-dimensional Euclidean space, R^p. That this is an unreasonable condition can be seen by noting that W_n is a rather complicated function of $W_1, W_2, \ldots, W_{n-1}$ as well as $P_1, P_2, \ldots, P_{n-1}$, and $F_1, F_2, \ldots, F_{n-1}$; and that, in general, $\{F_\ell\}_{\ell=1}^{\infty}$ is a correlated sequence. Loosely speaking, given the value that the random vector W_n takes on, one is also given some "information" about what values the random matrix F_n is allowed (and possibly also P_n). It is noted that several papers state an alternate assumption which is similar to that above: the conditional distribution of Y_n given $X_1=x_1,\ldots,X_n=x_n$ coincides with the distribution of $Y(x_n)$ for all real fixed parameter values x_n. It is also noted that several stochastic approximation convergence theorems require a weaker condition with the work "distribution" above replaced by "expectation." In practice, such conditions essentially require that either $\{Y_n\}$ is an independent sequence for the distribuiton condition or an uncorrelated

sequence for the expectation condition. Clearly, such conditions severely limit the applicability of the results. Surprisingly, very little has been done to alleviate the restriction due to this assumption. The results of Derman and Sacks [48] will now be discussed to suggest a possible approach for obtaining a more applicable result, as well as to show how the algorithm (3.1) is related to the general stochastic approximation procedure of Dvoretzky.

Derman and Sacks [48] have provided a simple proof to the following multidimensional version of a theorem originally stated by Dvoretzky [46]. The absolute value signs are to be interpreted as the p-dimensional Euclidean length.

THEOREM. *Let* $\{X_n\}$, $\{T_n(X_1,\ldots,X_n)\}$, *and* $\{Y_n(X_1,\ldots,X_n)\}(n\geq 1)$ *be p-dimensional random vectors with* X_1 *arbitrary and*

$$X_{n+1} = T_n(X_1,\ldots,X_n) + Y_n(X_1,\ldots,X_n) \ . \tag{3.7}$$

Assume that

$$E\{Y_n|X_1,\ldots,X_n\} \overset{a.s.}{=} 0 \ , \tag{3.8}$$

$$\sum_{n=1}^{\infty} E\{|Y_n|^2\} < \infty \ , \tag{3.9}$$

and

$$|T_n| \leq max(\alpha_n, \ (1+\beta_n)|X_n| - \gamma_n) \ , \tag{3.10}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, *and* $\{\gamma_n\}$ *are sequences of positive numbers such that*

$$\alpha_n \to 0, \quad \sum_{n=1}^{\infty} \beta_n < \infty \ , \quad \sum_{n=1}^{\infty} \gamma_n = \infty \ . \tag{3.11}$$

Then $|X_n| \overset{a.s.}{\to} 0$.

Making use of a technique suggested by Dvoretzky [46], algorithm (3.1) can be written so that the above theorem can be applied. Defining

$$V_n = W_n - w_o \ , \tag{3.12}$$

(3.1) can be written as

$$V_{n+1} = (I - \mu_n F_n)V_n + \mu_n C_n \ , \tag{3.13}$$

where

$$C_n = P_n - F_n w_o \ . \tag{3.14}$$

Now, defining

$$Y_n(V_n) = \mu_n((R_{xx} - F_n)V_n + C_n) \tag{3.15}$$

and

$$T_n(V_n) = V_n - \mu_n R_{xx} V_n \ , \tag{3.16}$$

(3.13) becomes

$$V_{n+1} = T_n(V_n) + Y_n(V_n) \ . \tag{3.17}$$

Define the matrix norm of a pxp matrix A by

$$||A|| = \sup_{|q|\leq 1} |Aq| \ , \tag{3.18}$$

which for A real and symmetric yields

$$||A|| = \max_i |\lambda_i(A)| \ , \tag{3.19}$$

where $\{\lambda_i(A)\}_{i=1}^{p}$ are the p eigenvalues of A and q is a p-element column vector.

Let $\{\mu_n\}_{n=1}^{\infty}$ and $\{\eta_n\}_{n=1}^{\infty}$ be nonincreasing sequences of positive numbers such that $\mu_n \to 0$, $\eta_n \to 0$, $\sum_{n=1}^{\infty} \mu_n = \infty$, and $\sum_{n=1}^{\infty} \mu_n \eta_n = \infty$.

Since R_{xx} is assumed to be positive definite, there exists an n_o such that for all $n \geq n_o$, $\mu_n < \lambda_{min}^{-1} < \infty$, where $\lambda_{min} = \min_i \lambda_i(R_{xx})$. For all $n \geq n_o$, and for all $u \in R^p$, $|T_n(u)| = |u - \mu_n R_{xx} u| \leq |u|$. $||I - \mu_n R_{xx}|| \leq |u|(1 - \mu_n \lambda_{min})$. For all $|u| \geq \eta_n$, $|u|(1-\mu_n \lambda_{min}) \leq |u| - \mu_n \lambda_{min} \eta_n$; whereas, for $|u| < \eta_n$, $|u|(1-\mu_n \lambda_{min}) \leq \eta_n \cdot (1-\mu_n \lambda_{min})$. It follows that for all $n \geq n_o$,

$$|T_n(u)| \leq \max(\eta_n(1-\mu_n \lambda_{min}), \ |u| - \mu_n \lambda_{min} \eta_n) , \qquad (3.20)$$

so that, with $X_n = V_n$, $\alpha_n = \eta_n(1 - \mu_n \lambda_{min})$, $\beta_n = 0$, and $\gamma_n = \mu_n \lambda_{min} \eta_n$, (3.7), (3.10), and (3.11) are satisfied. The following corollary has thus been established.

COROLLARY. *Let V_n, C_n, Y_n, T_n be p-dimensional random vectors given by (3.12) to (3.17). Let $\{\mu_n\}_{n=1}^{\infty}$ be a nonincreasing sequence of positive numbers with $\mu_n \to 0$ and $\sum_{n=1}^{\infty} \mu_n = \infty$. Assume that (3.8) and (3.9) are satisfied. Then $|V_n| \overset{a.s.}{\to} 0$.*

The difficulty with the above corollary is, of course, the establishment of (3.8) and (3.9). Condition (3.9) can be deleted by requiring that $E\{|\mu_n^{-1} Y_n(u)|^2\}$ be uniformly bounded for all $n \geq 1$ and for all $u \in R^p$, and that $\sum_{n=1}^{\infty} \mu_n^2 < \infty$. This uniformly bounded condition will be discussed in more detail later. Condition (3.8) has the same limitation mentioned previously. Dvoretzky [46] shows that (3.8) may be replaced by

$$\sum_{n=1}^{\infty} \sup_{x_1,\ldots,x_n} |E\{Y_n|x_1,\ldots,x_n\}| < \infty \quad , \tag{3.21}$$

or by the condition that each element of

$$\sum_{n=1}^{\infty} E\{Y_n|x_1,\ldots,x_n\} \tag{3.22}$$

be uniformly bounded and convergent for all sequences $x_1,x_2,\ldots,x_n,\ldots$. Unfortunately, conditions such as (3.21) or (3.22) are extremely difficult to verify in practice.

The method of proof of Derman and Sacks [48] can be modified to obtain yet another alternative to (3.8). Let $P_n = P_{n,x_1,x_2,\ldots,x_n}$ be random orthogonal transformations such that $P_nT_n = (|T_n|,0,\ldots,0)'$ and define $Z_n = P_nT_n$, where $Z_n' = (Z_{n1},Z_{n2},\ldots,Z_{np})$. If

$$\sum_{n=1}^{m} \frac{Z_{n1}(1+\beta_n)^2}{2\alpha_n + \beta_n}$$

and

$$\sum_{n=1}^{m} \frac{Z_{n1}^2(1+\beta_n)^4}{(2\alpha_n+\beta_n)^2}$$

converge a.s. to random variables as $m \to \infty$, then condition (3.8) of the theorem can be deleted and the theorem remains true. Although this may suggest a reasonable approach, the establishment of these conditons appears difficult, even for the special case considered in the corollary. Also, condition (3.9) or its (stronger) alternative of uniform boundedness of $E\{|\mu_n^{-1}Y_n(u)|^2\}$ is somewhat restrictive. In any case, this approach will not be pursued here.

Sakrison [52] presents a continuous Kiefer-Wolfowitz procedure and proves mean-square convergence for an a.s. bounded process and a

requirement on the rate at which the minimum mean-square prediction error approaches its asymptotic value. Sakrison [51] suggests that this condition is applicable to the Robbins-Monro procedure. More recently, some convergence results for algorithms of the form of (3.1) with $\mu_n = \mu$ = constant have appeared.

B. Weaker Convergence Results for Stochastic Approximation

Daniell [53] investigates a kind of mean-square convergence for algorithms similar to (3.1) with $\mu_n = \mu$ = constant. In fact, letting $\{X_n\}_{n=1}^{\infty}$ be a sequence of p-dimensional random vectors, $\mu_n = \mu$, and $F_n = X_n X_n'$, (3.1) is precisely the algorithm considered by Daniell. Rewriting (3.1) in the form of (3.13), with C_n given by (3.14), Daniell [53] proves the following theorem. The trace of a matrix A is denoted by tr{A}.

THEOREM. *Define* $A_i = X_i X_i' - R_{xx}$. *Suppose that (i) there exists a sequence of positive numbers* $\{\eta_k\}$ *converging to zero such that for every pair of positive integers* k *and* l

$$E\{\left|\frac{1}{k}\sum_{j=l+1}^{l+k} C_j\right|^2\} < \eta_k^2 \tag{3.23}$$

and

$$E\{\left|\left|\frac{1}{k}\sum_{j=l+1}^{l+k} A_j\right|\right|^2\} < \eta_k^2 \ ; \tag{3.24}$$

(ii) there exists a constant $\alpha_o > 0$ *such that if for all integer* $i > 1$, *then*

$$E\{|X_i|^4 \ |X_1, C_1, \ldots, X_{i-1}, C_{i-1}\} < \alpha_o \tag{3.25}$$

and
$$E\{|X_i|^2 \; |X_1, C_1, \ldots, X_{i-1}, C_{i-1}\} < \alpha_o \quad ; \tag{3.26}$$

(iii) there exists a sequence of positive real constants $\{\alpha_k\}$ *converging to zero such that if for all integer* $i > 1$ *and for all integer* k, L, M *satisfying* $i < i+k \leq M \leq L$,

$$tr\{E\{A_L' A_M \; | X_1, C_1, \ldots, X_i, C_i\} - E\{A_L' A_M\}\} < \alpha_k \quad ; \tag{3.27}$$

and (iv) there exists a positive constant B *such that*

$$E\{|C_i|^2\} < B^2 \tag{3.28}$$

and
$$E\{|X_i|^4 \; |C_i|^2\} < B^2 \quad . \tag{3.29}$$

Then for all $\delta > 0$ *there exists a* $\mu^* > 0$ *such that for all* $0 < \mu < \mu^*$ *there exists a positive integer* $k_\mu(\delta)$ *such that for all* $k > k_\mu(\delta)$

$$E\{|V_k|^2\} < \delta \; . \tag{3.30}$$

The kind of convergence obtained by Daniell is clearly weaker than mean-square convergence; however, by replacing μ with a nonincreasing sequence of positive constants $\{\mu_\ell\}_{\ell=1}^{\infty}$ converging to zero, it seems reasonable to conclude that the proof could be modified to obtain mean-square convergence. For applications which require the algorithm to track slowly time varying parameters, a fixed step size seems to be a reasonable as well as a widely used technique.

Senne [54] performed a simulation study of an algorithm similar to that treated by Daniell, and noted that when the process $\{X_k\}$ is correlated, a bias is introduced which increases with step size, μ. An analytical justification for this can be obtained by taking the

expectation of both sides of (3.1) to obtain

$$E\{W_{n+1}\} = E\{W_n\} + \mu_n(P-E\{F_nW_n\}) \quad . \tag{3.31}$$

Suppose that $E\{W_n\} = w_o$ and that $\mu_n = \mu$. If $\{F_n\}$ is a correlated sequence, then F_n and W_n are also correlated so that $E\{F_nW_n\} \neq P$, and hence $E\{W_{n+1}\} \neq w_o$. From this simple argument, it should be concluded that in order to have any hope for the algorithm to even be asymptotically unbiased, the condition that $\mu_n \to 0$ is essential. It is interesting to note that if $\{\mu_n\}$ is a sequence of positive constants converging to zero and the variance of each element in the correction vector $\mu_n(P_n-F_nW_n)$ is decreasing with increasing n so that the variance of each element of W_n will also be decreasing, F_n and W_n will "decorrelate."

The main issue here is to determine the limitations of the assumptions made in Daniell's theorem, i.e., to determine the types of correlated processes $\{X_n\}$ for which the theorem is applicable. Daniell [55] provides several examples of processes which satisfy the conditions of the above theorem; however, for the "correlated cases' considered, it is assumed that the process $\{X_k\}$ is bounded. Conditions (3.25) and (3.26) indicate that this bounded assumption is essential for the application of the above theorem.

Kim and Davisson [56] treat another algorithm which fits into the framework of (3.1). Let $\{s_n\}$ and $\{x_n\}$ be jointly stationary M-dependent scalar stochastic processes. A sequence of random variables $\{y_n\}$ is said to be M-dependent if for all index sets I_n, J_m, with $\min_{n\in I_n, m\in J_m} |n-m| > M$, the two sets of random variables $\{y_n : n\in I_n\}$ and $\{y_m : m\in J_m\}$ are statistically independent. Define

$X_n' = (x_n, x_{n-1}, \ldots, x_{n-p-1})$, let $P_n = \frac{1}{K} \sum_{m=nK}^{(n+1)K-1} s_m X_m$, $F_n = \frac{1}{K} \sum_{m=nK}^{(n+1)K-1} X_m X_m'$, and $\mu_n = \mu = \text{constant}$. Substituting into (3.1) yields

$$W_{n+1} = W_n + (\mu/K) \sum_{m=nK}^{(n+1)K-1} X_m (s_m - X_m' W_n) \quad . \tag{3.32}$$

Kim and Davisson [56] show, under the above assumptions, that $E\{|W_n - w_o|^2\}$ can be made arbitrarily small for n large enough by choosing μ small enough and K large enough. Although not explicitly stated by Kim and Davisson [56], their analysis also requires the existence of all fourth-order moments for both $\{s_n\}$ and $\{x_n\}$. The results of Kim and Davisson given above can likely be modified by replacing μ with a nonincreasing sequence $\{\mu_n\}$ of positive constants converging to zero to obtain mean-square convergence.

Schmetterer [50] presents the following theorem, a result which is quite similar in nature to the results discussed above of Daniell, and Kim and Davisson.

THEOREM. *Let a_n be a sequence of positive real numbers, satisfying $\sum_{i=1}^{\infty} a_i = \infty$. Let x_n and y_n be p-dimensional random vectors such that*

$$x_{n+1} = x_n - a_n y_n \tag{3.33}$$

for every $n \geq 1$. Furthermore, for every $n \geq 1$, let $M_n(\cdot)$ be a Borel measurable mapping from R^p to R^p. Assume that $E\{|y_n - M_n(x_n)|^2\}$ exists for every $n \geq 1$, and that there exists a real $C > 0$ such that

$$E\{|y_n - M_n(x_n)|^2\} \leq C \quad , \quad n \geq 1 \quad . \tag{3.34}$$

Furthermore, suppose that there exists a $K > 0$ *which satisfies*

$$a_i < K^{-1} \quad , \tag{3.35}$$

such that for every $n \geq 1$ *and* $x \epsilon R^p$, *the inequality*

$$|x - a_n M_n(x)| \leq (1 - Ka_n)\ |x| \tag{3.36}$$

holds. If $E\{|x_1|^2\}$ *exists, then* $E\{|x_n|^2\}$ *exists for every* $n \geq 2$. *Furthermore,*

$$(E\{|x_n|^2\})^{1/2} \leq C^{1/2}K^{-1} + ((E\{|x_1|^2\})^{1/2} - C^{1/2}K^{-1}) \prod_{i=1}^{n-1} (1-Ka_i) \quad . \tag{3.37}$$

It follows that

$$\overline{\lim_{n\to\infty}}\ (E\{|x_n|^2\})^{1/2} \leq C^{1/2}K^{-1} \quad . \tag{3.38}$$

Although condition (3.34) of the above theorem severely limits its applicability, some comments on the above theorem are in order. First of all, note that no conditional expectation or conditional distribution restriction is made. Secondly, (3.37) gives a bound on the mean norm-squared error for all $n \geq 1$. Hence, if the above theorem could be applied in a practical situation, it would be quite useful. Noting that (3.13) can be written as

$$V_{n+1} = V_n - \mu_n(F_n V_n - C_n) \quad , \tag{3.39}$$

with $C_n = P_n - F_n w_o$, and substituting $x_n = V_n$, $a_n = \mu_n$, $y_n = F_n V_n - C_n$, (3.33) results. Letting $M_n(\nu) = R_{xx}\nu$ for all $\nu \epsilon R^p$, (3.34) requires that there exist a $C > 0$ such that

$$E\{|F_n V_n - C_n - R_{xx} V_n|^2\} \leq C \quad , \quad n \geq 1 \quad . \tag{3.40}$$

Since R_{xx} is assumed to be positive definite with minimum eigenvalue λ_{min}, $K = \lambda_{min}$ and (3.35) establish (3.36). Apparently, (3.40) is difficult to establish unless V_n is uniformly bounded (in n) with probability one, thus suggesting a possible application to truncated algorithms. That is, suppose w_o is known *a priori* to lie within some closed convex parameter space P, and consider the following truncated version of (3.1)

$$W_{n+1} = [W_n + \mu_n(P_n - F_n W_n)]_P \quad , \tag{3.41}$$

where $[x]_P = x$ if $x \in P$, and $[x]_P$ is the boundary point of P closest to x if $x \notin P$. Defining $P_{w_o} = \{x : x + w_o \in P\}$, and with $V_n = W_n - w_o$, (3.41) becomes

$$V_{n+1} = [V_n - \mu_n(F_n V_n + F_n w_o - P_n)]_{P_{w_o}} \quad . \tag{3.42}$$

Clearly, this algorithm is a.s. uniformly bounded, and can be shown to satisfy (3.40). Unfortunately, certain analytic difficulties arise when attempting to establish (3.36) for this algorithm. A result similar to the above theorem of Schmetterer for algorithms such as (3.42) would be highly desirable.

C. Critique

In this chapter, several of the existing convergence results applicable to algorithms having the form of (3.1) have been reviewed in detail. Several suggestions have been made as to how existing results might be modified to obtain reasonable conditions for which $W_n \to w_o$ in

some meaningful probabilistic sense when the sequence $\{F_n\}$ (and possibly also $\{P_n\}$) is correlated.

In summarizing the state of existing stochastic approximation results, it can be said that the conditions imposed by Robbins and Monro, Dvoretzky, and Derman and Sacks, for example, employ ingenious mathematical constructs to permit general applicability of stochastic approximation results. From a practical point of view, however, it cannot be emphasized too strongly that their conditions are easily established for (3.1) only when $\{P_n - F_n w\}_{n=1}^{\infty}$ is an independent sequence of R^p-valued random variables, where w is a fixed parameter. Consequently the existing results are not well-suited to the analysis of structures that must be adapted in correlated environments. As repeatedly mentioned previously, the restrictive assumptions are the "conditional distribution," or the "conditional expectation" assumptions. The only results (known to the author) not making these restrictions are those of Daniell, Kim and Davisson, and Schmetterer, mentioned in Section III-B. In the next chapter, easily verified conditions will be established for which W_n as given by (3.1) will converge a.s. to w_o. These conditions will permit us to relax the "conditional expectation" or "conditional distribution" assumptions of existing theorems and prove convergence in correlated environments of practical interest.

IV. NEW CONVERGENCE RESULTS

In this chapter, new, easily verified conditions are established which ensure the a.s. convergence of W_n to w_o as given by (3.1). Section IV-A contains the main results of this dissertation. The proof of the theorem relies heavily on the techniques presented by Albert and Gardner [57]. The proof of the practically useful result, Corollary 2, makes strong use of the results of Serfling ([58] and [59]). In Section IV-B the results of Section IV-A are applied to the specific algorithms treated in Chapter II, providing analytical justification for existing and proposed applications of these algorithms. In Section IV-C, a highly specialized form of (3.1) is treated which seemingly suggests a "maximum convergence rate" for certain algorithms. Open issues regarding the convergence properties of algorithms fitting the framework of (3.1) are discussed in Section IV-D.

A. Almost Sure Convergence Results

As shown in Chapter III, the algorithm

$$W_{n+1} = W_n + \mu_n(P_n - F_n W_n) \quad , \tag{4.1}$$

can be written in the form

$$V_{n+1} = (I - \mu_n F_n)V_n + \mu_n C_n \quad , \tag{4.2}$$

where $$V_n = W_n - w_o \quad , \tag{4.3}$$

$$w_o = R_{xx}^{-1} P \quad , \tag{4.4}$$

$$C_n = P_n - F_n w_o \quad , \tag{4.5}$$

$$E\{F_n\} = R_{xx} \quad , \tag{4.6}$$

and $$E\{P_n\} = P \quad . \tag{4.7}$$

It is assumed that R_{xx} is a real symmetric positive definite $p \times p$ matrix, W_n and P_n are elements of R^p, $\{\mu_n\}$ is a nonincreasing sequence of positive constants, and that $\{F_k\}_{k=1}^{\infty}$ is a random sequence of real symmetric non-negative definite $p \times p$ matrices. Defining for $\{A_i\}$ a sequence of $p \times p$ matrices

$$\prod_{i=\ell}^{k} A_i = \begin{cases} A_k A_{k-1} \cdots A_{\ell+1} A_\ell \quad , & \text{if } k \geq \ell \ , \\ I, & \text{if } k < \ell; \end{cases} \tag{4.8}$$

and iterating (4.2), one obtains

$$V_{n+1} = \prod_{k=1}^{n} (I - \mu_k F_k) V_1 + \sum_{k=1}^{n} \Big(\prod_{j=k+1}^{n} (I - \mu_j F_j) \Big) \mu_k C_k \quad . \tag{4.9}$$

Defining

$$Q_{\ell m} = \prod_{j=\ell}^{m} (I - \mu_j F_j) \quad , \tag{4.10}$$

$$\Lambda_n = \sum_{k=1}^{n} Q_{k+1,n} \, \mu_k C_k \quad , \tag{4.11}$$

(4.9) becomes

$$V_{n+1} = Q_{1n} V_1 + \Lambda_n \quad . \tag{4.12}$$

Recall that the matrix norm for a $p \times p$ matrix A is defined by

$$\|A\| = \sup_{|x| \leq 1} |Ax| \quad , \ x \in R^p \quad , \tag{4.13}$$

which, for A real and symmetric coincides with

$$\|A\| = \max\{|\lambda_i(A)|\}, \quad i \in \{1,2,\ldots,p\} \tag{4.14}$$

where $\{\lambda_i(A)\}_{i=1}^{p}$ are the p eigenvalues of A. Denote the minimum and maximum eigenvalues of A by $\lambda_{min}(A)$ and $\lambda_{max}(A)$, respectively. With the above notations and definitions ((4.1) - (4.14)) established, which will be assumed throughout the remainder of this section, the main result of this dissertation can now be stated.

THEOREM. *Suppose that the following assumptions (including the structure implied by (4.1) - (4.14)) are satisfied:*

A1) $\{\mu_k\}$ *is a nonincreasing sequence of positive constants converging to zero such that whenever*

$$|k-\ell| < N, \ \mu_k/\mu_\ell < h_N < \infty, \text{ and } \sum_{k=1}^{\infty} \mu_k = \infty,$$

A2) $\mu_k \|F_k\| \xrightarrow{a.s.} 0$ *as* $k \to \infty$,

A3) $n^{-1} \sum_{k=1}^{n} F_k \xrightarrow{a.s.} R_{xx}$ *as* $n \to \infty$,

A4) there exists a random vector $S \in R^p$ *such that*

$$S_n = \sum_{k=1}^{n} \mu_k C_k \xrightarrow{a.s.} S \text{ as } n \to \infty, \text{ and}$$

A5) $|F_n(S - S_{n-1})| \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Then $|V_n| \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Regarding assumptions A1 through A5, assumption A1 is seemingly the only assumption similar in spirit to other stochastic approximation results, and is easily satisfied by $\mu_n = 1/n$, for example.

Assumption A3 is the only other readily recognized assumption, and can be interpreted as a kind of ergodicity assumption. Indeed, assumptions A2 through A5 involve the a.s. convergence of sequences of random variables, and the conclusion of the theorem is the a.s. convergence of still another sequence of random variables. The principal advantage in using such an approach is that assumptions A3 through A5 are in a form suitable for (but not limited to) application of the results of Serfling ([58] and [59]). The end result is sufficient conditions on the "decay rate" of the autocovariance functions of the sequences $\{F_k\}$ and $\{P_k\}$ which imply A3 through A5. Examples in which these results are applied to the algorithms discussed in Chapter II are given in Section IV-B.

As mentioned previously, the proof of the above theorem relies heavily on the techniques of Albert and Gardner [57]. The proof is a direct modification of the proof of Theorem 6.3 of [57]; however, the algorithm treated in Theorem 6.3 of [57] is quite different from (3.1) and the assumptions above are seemingly less restrictive. Before proving the theorem, several useful lemmas will be established. Lemmas 1 and 2, which are similar in nature to Theorem 6.1 of [57], make use of assumptions A1, A2, and A3 to show that $\|Q_{1n}\| \overset{a.s.}{\to} 0$ as $n \to \infty$. The assumption that each F_n is symmetric and non-negative definite can be relaxed by applying Theorem 6.1 of [57]; however, for adaptive signal processing applications, F_n is almost always some form of a sample covariance estimate, hence, the simplification resulting for symmetric F_n seems worthwhile.

LEMMA 1. *If A1-A3 are satisfied, then there exists a sequence of integers* $\{\nu_k\}$ *with* $1=\nu_1<\nu_2<\nu_3<\ldots$ *such that, with* $p_k = \nu_{k+1} - \nu_k$,

$J_k = \{\nu_k, \nu_k+1, \ldots, \nu_{k+1}-1\}$, *and* $k = 1,2,\ldots$, *(i)* $p_{min} < p_k < p_{max} < \infty$, *(ii)* $p_k^{-1}\,\lambda_{min}\left(\sum_{j\in J_k} F_j\right) = \alpha_k \overset{a.s.}{>} 0$, *(iii)* $p_k^{-1}\,\lambda_{max}\left(\sum_{j\in J_k} F_j\right) = \gamma_k \overset{a.s.}{<} \gamma < \infty$, *and (iv) there exists a* $\delta > 0$ *such that* $\alpha_k \overset{a.s.}{\geq} \delta$. *The sequences* $\{\nu_k\}$, $\{p_k\}$, $\{\alpha_k\}$, *and* $\{\gamma_k\}$ *may all be random sequences depending on the particular realization of the sequence* $\{F_k\}$.

PROOF. Define

$$R_n^{\ell} = \frac{1}{n-\ell} \sum_{k=\ell+1}^{n} F_k \quad . \tag{4.15}$$

Let $\varepsilon > 0$ be given such that $0<\varepsilon<\lambda_{min}(R_{xx})$. Assumptions A1-A3 imply that for any fixed $\ell\in\{0,1,2,\ldots\}$, $\lim_{n\to\infty} R_n^{\ell} \overset{a.s.}{=} R_{xx}$. It follows that $\lim_{n\to\infty} \lambda_{min}(R_n^{\ell}) \overset{a.s.}{=} \lambda_{min}(R_{xx})$. Hence, it follows that there exists an n_ℓ (possibly random) such that $|\lambda_{min}(R_{xx}) - \lambda_{min}(R_{n_\ell}^{\ell})| \overset{a.s.}{<} \varepsilon$; thus $0 < \lambda_{min}(R_{xx}) - \varepsilon \overset{a.s.}{<} \lambda_{min}(R_{n_\ell}^{\ell})$. Since n_ℓ is finite and ℓ is arbitrary, (i), (ii), and (iv) follow. A similar argument applies to (iii). Q.E.D.

LEMMA 2. *If A1, A2, and A3 are satisfied, then* $\|Q_{1n}\| \overset{a.s.}{\to} 0$ *as* $n \to \infty$.

PROOF. It follows from A2 that there exists a random variable M, $1\leq M<\infty$ such that $\sup_k \|I-\mu_k F_k\| \overset{a.s.}{<} M$. Keeping the same notation as in Lemma 1, for any n, let $K = K(n)$ be the largest integer such that $\nu_K \leq n$ so that $\nu_K \leq n \leq \nu_{K+1}-1$. Then

$$Q_{1n} = \prod_{j=\nu_K}^{n} (I - \mu_j F_j) Q_{1,\nu_K-1} \quad , \tag{4.16}$$

and hence,

$$\|Q_{1n}\| \leq \prod_{j=\nu_K}^{n} \|I - \mu_j F_j\| \cdot \|Q_{1,\nu_K-1}\| \overset{a.s.}{\leq} M^{p_{max}} \|Q_{1,\nu_K-1}\| . \tag{4.17}$$

Consequently, it suffices to show that $\|Q_{1,\nu_K-1}\| \overset{a.s.}{\to} 0$ as $K \to \infty$ with n over some subset of the positive integers. Noting that

$$Q_{1,\nu_K-1} = \prod_{k=1}^{K-1} \prod_{j=\nu_k}^{\nu_{k+1}-1} (I - \mu_j F_j) = \prod_{k=1}^{K-1} Q_{\nu_k,\, \nu_{k+1}-1}, \tag{4.18}$$

and defining $\Gamma_k = Q_{\nu_k,\, \nu_{k+1}-1}$,

$$Q_{1,\nu_K-1} = \prod_{k=1}^{K-1} \Gamma_k . \tag{4.19}$$

Expressing Γ_k as

$$\Gamma_k = \prod_{\ell=\nu_k}^{\nu_{k+1}-1} (I - \mu_\ell F_\ell) = I - \sum_{j \in J_k} \mu_j F_j + \sum_{\substack{\ell_1 > \ell_2 \\ \ell_1, \ell_2 \in J_k}} \mu_{\ell_1}\mu_{\ell_2} F_{\ell_1} F_{\ell_2}$$

$$+ \sum_{q=3}^{p_k} \sum_{\substack{\ell_1 > \ell_2 > \ldots > \ell_q \\ \ell_1, \ell_2, \ldots, \ell_q \in J_k}} (-1)^q \mu_{\ell_1}\mu_{\ell_2}\ldots\mu_{\ell_q} F_{\ell_1} F_{\ell_2} \ldots F_{\ell_q} , \tag{4.20}$$

it follows that (for $\mu_{\nu_k} < 1$)

$$\|\Gamma_k\| \overset{a.s.}{\leq} 1 - \mu_{\nu_{k+1}-1} \lambda_{min}\Big(\sum_{j \in J_k} F_j\Big) + \sum_{q=2}^{p_k} \mu_{\nu_k}^q \lambda_{max}^q \Big(\sum_{j \in J_k} F_j\Big)$$

$$\overset{a.s.}{\leq} 1 - \mu_{\nu_{k+1}-1} p_k \alpha_k + \mu_{\nu_k}^2 \sum_{q=2}^{p_k} (p_{max}\gamma)^q \tag{4.21}$$

from Lemma 1. From A1 and Lemma 1, there exists a positive integer k_o (possible random) such that for all $k \geq k_o$,

$$\|\Gamma_k\| \overset{a.s.}{\leq} 1 - \tfrac{1}{2}\mu_{\nu_{k+1}-1}p_{min}\,\delta \overset{a.s.}{\leq} \exp\{-\tfrac{1}{2}\mu_{\nu_{k+1}-1}p_{min}\,\delta\}, \tag{4.22}$$

since $1 - x \leq e^{-x}$ for all real x. Hence, there exists a random variable M_1 such that for all $K > k_o$,

$$\|Q_{1,\nu_K-1}\| \overset{a.s.}{\leq} M_1 \prod_{k=k_o}^{K-1} \|\Gamma_k\|$$

$$\overset{a.s.}{\leq} M_1 \exp\{-\tfrac{1}{2}p_{min}\,\delta \sum_{k=k_o}^{K-1} \mu_{\nu_{k+1}-1}\} \quad . \tag{4.23}$$

It follows from the above and A1 that $\|Q_{1n}\| \overset{a.s.}{\to} 0$. Q.E.D.

LEMMA 3. (Albert and Gardner)[57]. *Let $\{A_k\}$ be a sequence of square matrices. Then for all $1 \leq k \leq n$ and $n \geq 1$,*

$$\sum_{j=k}^{n} [\prod_{i=j+1}^{n} (I - A_i)]A_j = I - \prod_{i=k}^{n} (I - A_i). \tag{4.24}$$

LEMMA 4. (Toeplitz Lemma)[60]. *If $x_n \to \xi$, and the coefficients $a_{\mu\nu}$ satisfy (i) for fixed $p \geq 1$, $a_{np} \to 0$ as $n \to \infty$, (ii) there exists a K such that for all $n \geq 1$, $\sum_{i=1}^{n} |a_{ni}| < K$, and*

(iii) $\sum_{i=1}^{n} a_{ni} = A_n \to \alpha$ *as* $n \to \infty$, *then* $x_n^* = \sum_{i=1}^{n} a_{ni}x_i \to \alpha\xi$. (4.25)

PROOF of THEOREM. Equation (4.12) expresses V_{n+1} as $V_{n+1} = Q_{1n}V_1 + \Lambda_n$. It has been shown in Lemma 2 that $\|Q_{1n}\| \overset{a.s.}{\to} 0$. It remains to be shown that $|\Lambda_n| \overset{a.s.}{\to} 0$. From (4.11) and A4, with $S_o = 0$ and $Q_{n+1,n} = I$,

$$\Lambda_n = \sum_{k=1}^{n} Q_{k+1,n}\, S_k - \sum_{k=1}^{n} Q_{k+1,n}\, S_{k-1}$$

$$= \sum_{k=1}^{n} (Q_{k,n} - Q_{k+1,n}) S_{k-1} + S_n$$

$$= \sum_{k=1}^{n} (Q_{k+1,n}(I - \mu_k F_k) - Q_{k+1,n}) S_{k-1} + S_n$$

$$= -\sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k S_{k-1} + S_n \quad . \tag{4.26}$$

By assumption A4, there exists a random vector $S \in R^p$ such that $S_n \overset{a.s.}{\to} S$ as $n \to \infty$, hence, (4.26) may be rewritten as

$$\Lambda_n = \sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k (S - S_{k-1}) - \sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k S + S_n. \tag{4.27}$$

From Lemma 3,

$$\sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k S = (I - Q_{1,n}) S \quad , \tag{4.28}$$

so that

$$\Lambda_n = \sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k (S - S_{k-1}) + S_n - S + Q_{1,n}\, S \quad . \tag{4.29}$$

Since $S_n \overset{a.s.}{\to} S$, and $\|Q_{1,n}\| \overset{a.s.}{\to} 0$, it now remains only to show that for

$$b_n \overset{\Delta}{=} \left| \sum_{k=1}^{n} Q_{k+1,n}\, \mu_k F_k (S - S_{k-1}) \right| \quad , \tag{4.30}$$

$b_n \overset{a.s.}{\to} 0$ as $n \to \infty$. Using the same notation as in Lemmas 1 and 2, with $K = K(n)$ the largest integer such that $\nu_K \leq n$ so that $\nu_K \leq n \leq \nu_{K+1}-1$, and $\Gamma_k = Q_{\nu_k, \nu_{k+1}-1}$,

$$b_n = \Big| Q_{\nu_K, n} \sum_{k=1}^{K-1} \sum_{j \in J_k} Q_{j+1, \nu_K - 1} \mu_j F_j(S - S_{j-1}) + \sum_{k=\nu_K}^{n} Q_{k+1,n} \mu_k F_k(S - S_{k-1}) \Big| \quad , \tag{4.31}$$

which can be bounded as

$$b_n \overset{a.s.}{\leq} M^{2p_{max}} p_{max} \sum_{k=1}^{K-1} \prod_{\ell=k+1}^{K-1} \|\Gamma_\ell\| \mu_{\nu_k} d_k + p_{max} M^{p_{max}} \mu_{\nu_K} d_K \quad , \tag{4.32}$$

where d_k is defined by

$$d_k = \max_{j \in J_k} |F_j(S - S_{j-1})| \quad . \tag{4.33}$$

It follows from A5 that $d_k \overset{a.s.}{\to} 0$ as $k \to \infty$, so that it now remains only to show that for

$$e_K \overset{\Delta}{=} \sum_{k=1}^{K-1} \prod_{\ell=k+1}^{K-1} \|\Gamma_\ell\| \mu_{\nu_k} d_k \quad , \tag{4.34}$$

$e_K \overset{a.s.}{\to} 0$ as $K \to \infty$ with n over some subset of the positive integers. Defining $\beta_k = \frac{1}{2} \mu_{\nu_{k+1}-1} p_{min} \delta$, from Lemma 2 there exists a k_o such that for all $k \geq k_o$, $\|\Gamma_k\| \overset{a.s.}{\leq} 1 - \beta_k$. It is assumed that k_o is large enough so that $\beta_k < 1$ for all $k \geq k_o$. Proceeding, for all $K > k_o$,

$$e_K \leq \sum_{k=1}^{k_o-1} \prod_{i=k_o}^{K-1} \|\Gamma_i\| \prod_{\ell=k+1}^{k_o-1} \|\Gamma_\ell\| \mu_{\nu_k} d_k + \sum_{k=k_o}^{K-1} \prod_{\ell=k+1}^{K-1} \|\Gamma_\ell\| \mu_{\nu_k} d_k$$

$$\underset{\sim}{a.s.} M^{k_o} \prod_{i=k_o}^{K-1} (1 - \beta_i) \sum_{k=1}^{k_o-1} \mu_{\nu_k} d_k + \sum_{k=k_o}^{K-1} \prod_{\ell=k+1}^{K-1} (1-\beta_\ell)\beta_k(\mu_{\nu_k} d_k \beta_k^{-1}). \tag{4.35}$$

Define

$$a_{ni} = \prod_{\ell=i+1}^{n} (1 - \beta_\ell)\beta_i \quad . \tag{4.36}$$

Clearly, for all fixed $i \geq k_o$, $a_{ni} \to 0$ as $n \to \infty$.

From Lemma 3,

$$\sum_{i=k_o}^{n} |a_{ni}| = \sum_{i=k_o}^{n} \prod_{\ell=i+1}^{n} (1 - \beta_\ell)\beta_i = 1 - \prod_{i=k_o}^{n} (1 - \beta_i), \tag{4.37}$$

which converges a.s. to 1 as $n \to \infty$, so that by Lemma 4,

$$\lim_{K\to\infty} \sum_{k=k_o}^{K-1} a_{K-1,k}(\mu_{\nu_k} d_k \beta_k^{-1}) \overset{a.s.}{=} \lim_{k\to\infty} (\mu_{\nu_k} d_k \beta_k^{-1}) \quad . \tag{4.38}$$

From A1 and the definition of β_k

$$\mu_{\nu_k} d_k \beta_k^{-1} = \frac{2\mu_{\nu_k} d_k}{\mu_{\nu_{k+1}-1} p_{min} \delta} \leq 2(\delta p_{min})^{-1} h_{p_{max}} d_k \quad , \tag{4.39}$$

and hence, from A5,

$$\lim_{k\to\infty} (\mu_{\nu_k} d_k \beta_k^{-1}) \underset{\sim}{a.s.} 0 \quad . \tag{4.40}$$

Q.E.D.

With the theorem established, considerable attention will now be given to the establishment of corollaries which will guarantee under

extremely realistic conditions, that assumptions A1 through A5 are satisfied. It will be expedient to make use of the order notation, $O(\cdot)$, e.g., $f(n) = O(g(n))$ if $f(n)/g(n)$ is bounded as $n \to \infty$.

A worthwhile simplification of assumptions A1 through A5 results in the case $\|F_n\|$ is a.s. bounded. In this case, the following corollary is easily established.

COROLLARY 1. *If* $\mu_k = O(k^{-1})$, $\lim_{k\to\infty} k\mu_k > 0$, *and* $\|F_k\|$ *is a.s. bounded, then A1, A2, and A5 may be deleted and the theorem remains true.*

PROOF. It suffices to consider $\mu_k = k^{-1}$. Assumption A1 is trivially satisfied. That A4 implies A5 can easily be seen by noting that there exists an $M \overset{a.s.}{<} \infty$ such that $|F_n(S - S_{n-1})| \leq \|F_n\| \cdot |S - S_{n-1}| \overset{a.s.}{<} M \cdot |S - S_{n-1}|$, so that $|S - S_{n-1}| \overset{a.s.}{\to} 0$ implies that $|F_n(S - S_{n-1})| \overset{a.s.}{\to} 0$ as $n \to \infty$. Assumption A2 is easily established by the Borel-Cantelli Lemma and the Chebychev inequality as follows. For all $\varepsilon > 0$, $\Pr\{\mu_k\|F_k\| > \varepsilon\} = \Pr\{\|F_k\| > \varepsilon\mu_k^{-1}\} \leq \mu_k^2 \varepsilon^{-2} E\{\|F_k\|^2\}$, and since k^{-2} is summable, $\mu_k\|F_k\| \overset{a.s.}{\to} 0$ as $k \to \infty$. Q.E.D.

The Borel-Cantelli Lemma, together with probabilistic bounds, such as the Chebychev inequality, the Markov inequality, or the Chernoff bound, provides a frequently used technique for establishing the a.s. convergence of sequences of random variables. Unfortunately, the available probabilistic bounds often approach zero but are not summable (unlike the case presented in Corollary 1). The work of Serfling ([58] and [59]) provides useful techniques by which the above difficulties can be overcome. For a more complete treatment on a.s. convergence, the interested reader is referred to the recent text by Stout [61]. Before

developing the machinery necessary for the proof of Corollary 2, the a.s. convergence of $|S - S_{n-1}|$ is discussed in order to illustrate the concepts involved.

For all $\varepsilon > 0$, the following bound is easily obtained from the Chebychev inequality:

$$\Pr\{|S - S_{n-1}| > \varepsilon\} \leq \varepsilon^{-2} E\{|S - S_{n-1}|^2\} \quad , \tag{4.41}$$

where it has been assumed that $E\{|S - S_{n-1}|^2\} < \infty$. It is noted that $S - S_{n-1}$ is given by (formally)

$$S - S_{n-1} = \sum_{k=n}^{\infty} \mu_k C_k \quad , \tag{4.42}$$

so that

$$E\{|S - S_{n-1}|^2\} = \sum_{k=n}^{\infty} \sum_{\ell=n}^{\infty} \mu_k \mu_\ell E\{C_k^* C_\ell\} \quad . \tag{4.43}$$

Suppose for the moment that $E\{C_k^* C_\ell\} = \delta_{k,\ell}$, and that $\mu_k = O(k^{-1})$, where $\delta_{k,\ell}$ is the Kronecker delta function

$$\delta_{k,\ell} = \begin{cases} 1, & \text{if } k = \ell \\ 0, & \text{if } k \neq \ell \end{cases} \quad . \tag{4.44}$$

Then $E\{|S - S_{n-1}|^2\} = O(n^{-1})$, which is seemingly the fastest rate one can expect, so that it is indeed fruitless to attempt the direct application of the Borel-Cantelli Lemma to (4.41) to obtain the a.s. convergence of $|S - S_{n-1}|$. However, while the summability of $E\{|S - S_{n-1}|^2\}$ seems impossible, it would seem reasonable to require that $E\{|S - S_{n-1}|^2\} \to 0$ as $n \to \infty$. Although mean-square convergence and a.s. convergence are not equivalent, in view of A4 it does not seem unduly restrictive to require that $S_n \xrightarrow{m.s.} S$.

Suppose that $E\{|S - S_{n-1}|^2\} \to 0$ as $n \to \infty$. Then there exists an increasing subsequence $\{n_k\}$ such that $n_k \to \infty$ as $k \to \infty$ and

$$\sum_{k=1}^{\infty} E\{|S - S_{n_k-1}|^2\} < \infty \quad , \tag{4.45}$$

hence $|S - S_{n_k-1}| \overset{a.s.}{\to} 0$ as $k \to \infty$.

This fact can be used by noting that for all $n \in L_k$, with $L_k = \{n_k, n_k+1, \ldots, n_{k+1}-1\}$,

$$|S - S_{n-1}| = \Big| \sum_{i=n}^{n_{k+1}-1} \mu_i C_i + \sum_{i=n_{k+1}}^{\infty} \mu_i C_i \Big|$$

$$\leq \max_{\ell \in L_k} \Big| \sum_{i=\ell}^{n_{k+1}-1} \mu_i C_i \Big| + |S - S_{n_{k+1}-1}|. \tag{4.46}$$

For all sequences $\{n_k\}$ satisfying (4.45) and such that

$$\max_{\ell \in L_k} \Big| \sum_{i=\ell}^{n_{k+1}-1} \mu_i C_i \Big| \overset{a.s.}{\to} 0 \tag{4.47}$$

as $k \to \infty$, $|S - S_{n-1}| \overset{a.s.}{\to} 0$ as $n \to \infty$. The work of Serfling ([58] and [59]) is easily applied to terms like (4.47)

The following lemma, a multidimensional version of Theorem A of [58], will be shown to be invaluable for the establishment of conditions similar to (4.47). The proof of Lemma 5 is a simple modification of that given in [58] and will be omitted.

LEMMA 5. (Serfling) [58]. *Let* $\{x_i\}$ *be a sequence of random vectors,* $x_i \in R^p$ *having finite "variances"* $\sigma_i^2 = E\{(x_i - E\{x_i\})'(x_i - E\{x_i\})\}$. *For each matrix* $X_{a,n} = (x_{a+1}, \ldots, x_{a+n})$ *of* n *consecutive* x_i*'s*

let $F_{a,n}$ *denote the joint distribution function and let*

$$S_{a,n} = \sum_{i=a+1}^{a+n} x_i \quad , \tag{4.48}$$

$M_{a,n} = \max\{|S_{a,1}|, \ldots, |S_{a,n}|\}$, *and let* $g(F_{a,n})$ *be a functional depending on* $F_{a,n}$. *Let* a_o *be an arbitrary but fixed integer and let* $\nu \geq 2$. *Suppose* $g(F_{a,k}) + g(F_{a+k,\ell}) \leq g(F_{a,k+\ell})$ *for all* $a \geq a_o$ *and* $1 \leq k \leq k+\ell$ *such that* $E\{|S_{a,n}|^\nu\} \leq g^{\frac{1}{2}\nu}(F_{a,n})$ *for all* $a \geq a_o$ *and all* $n \geq 1$. *Then* $E\{M_{a,n}^\nu\} \leq (\log_2 2n)^\nu g^{\frac{1}{2}\nu}(F_{a,n})$ *for all* $a \geq a_o$ *and all* $n \geq 1$.

A rather straightforward modification of Lemma 5 will also be needed and is presented below as Lemma 6. The proof of Lemma 6 is virtually identical to that of Lemma 5 and thus will be omitted.

LEMMA 6. *Let* $\{x_i\}$ *be a sequence of random vectors,* $x_i \in R^p$ *having finite "variances"* $\sigma_i^2 = E\{(x_i - E\{x_i\})'(x_i - E\{x_i\})\}$. *For each matrix* $X_{a,n} = (x_{a-n+1}, \ldots, x_a)$ *of* n *consecutive* x_i*'s let* $F_{a,n}$ *denote the joint distribution function and let*

$$S_{a,n} = \sum_{i=a-n+1}^{a} x_i \quad , \tag{4.49}$$

$M_{a,n} = \max\{|S_{a,1}|, \ldots, |S_{a,n}|\}$ *and let* $g(F_{a,n})$ *be a functional depending on* $F_{a,n}$. *Let* a_o *be an arbitrary but fixed integer and let* $\nu \geq 2$. *Suppose* $g(F_{a,k}) + g(F_{a-k,\ell}) \leq g(F_{a,k+\ell})$ *for all* $1 \leq k \leq k+\ell < a-a_o$ *such that* $E\{|S_{a,n}|^\nu\} \leq g^{\frac{1}{2}\nu}(F_{a,n})$ *for all* $1 \leq n < a-a_o$. *Then* $E\{M_{a,n}^\nu\} \leq (\log_2 2n)^\nu g^{\frac{1}{2}\nu}(F_{a,n})$ *for all* $1 \leq n < a-a_o$.

Lemma 7 below makes use of common procedures to obtain bounds on double sums of symmetric functions, such as autocorrelation functions.

The results of Lemma 7 will prove invaluable in establishing "slowest decay rates" of the autocovariance functions of $\{F_k\}$ and $\{P_k\}$ for "gain sequences" $\{\mu_k\}$ of the form $\mu_k = O(k^{-1})$. The technique of proof will allow some flexibility regarding the choice of sequence $\{\mu_k\}$.

LEMMA 7. *Let* $\alpha_{k,\ell} = \alpha_{\ell,k}$ *and* $\rho(k,\ell) = \rho(\ell,k)$ *be real valued functions defined for all non-negative integers* k, ℓ. *Then for* $1 \le n < m$, *define*

$$\gamma_{n,m} = \sum_{k=n}^{m} \sum_{\ell=n}^{m} \alpha_{k,\ell}\, \rho(k,\ell). \tag{4.50}$$

Then

$$(a) \quad \gamma_{n,m} = 2 \sum_{u=1}^{m-n} \sum_{k=n}^{m-u} \alpha_{k,k+u}\, \rho(k,k+u) + \sum_{k=n}^{m} \alpha_{k,k}\, \rho(k,k).$$

Suppose further that there exists a real valued function $f(u)$ *such that for all* $u = 0,1,2,\ldots,$ *and for all* $k = 1,2,\ldots,$ $|\rho(k,k+u)| \le f(u)$, *and* $f(u) = O(u^{-\nu})$. *If* $\alpha_{k,\ell} = 1$, *then, for large* $m - n$ *and* $\nu = 1$,

$$(b) \quad |\gamma_{n,m}| = O((m-n)\ln(m-n)).$$

Finally, if $\alpha_{k,\ell} = \mu_k \mu_\ell$, $\mu_k = O(k^{-1})$, *and* $\nu \ge 1$, *then*

$$(c) \quad |\gamma_{n,\infty}| = O(n^{-\nu/\nu+2}).$$

PROOF. Let $u = k - \ell$ in (4.50). For $u = n-m, n+1-m, \ldots, -1$; $k = n, n+1, \ldots, u+m$. For $u = 0$; $k = n, n+1, \ldots, m$. For $u = 1,2,\ldots,m-n$; $k = u+n, u+n+1, \ldots, m$. Substituting into (4.50),

$$\gamma_{n,m} = \sum_{u=n-m}^{-1} \sum_{k=n}^{m+u} \alpha_{k,k-u}\, \rho(k,k-u) + \sum_{k=n}^{m} \alpha_{k,k}\, \rho(k,k)$$

$$+ \sum_{u=1}^{m-n} \sum_{k=n+u}^{m} \alpha_{k,k-u}\, \rho(k,k-u) \quad . \tag{4.51}$$

Making the transformation $k^* = k - u$ in the last series and making use of the symmetry relations, (a) follows.

Suppose that $\alpha_{k,\ell} = 1$ and $|\rho(k,k+u)| \leq f(u) = O(u^{-1})$ for all $u = 0,1,2,\ldots$, and for all $k = 1,2,\ldots$. Then

$$|\gamma_{n,m}| \leq 2 \sum_{u=1}^{m-n} f(u)(m-u-n+1) + (m-n+1)\, f(o). \tag{4.52}$$

For all $1 \leq \ell < m-n$, for some $C_1 > 0$, and for $C_2 = \max_{1 \leq u \leq m-n} f(u)$,

$$|\gamma_{n,m}| \leq 2C_2(m-n)\ell + 2C_1(m-n+1) \sum_{u=\ell+1}^{m-n} \frac{1}{u}$$

$$- 2C_1(m-n-\ell) + (m-n+1)\, f(o) \quad , \tag{4.53}$$

which, for some $C_3 > 0$, yields

$$|\gamma_{n,m}| \leq 2C_2(m-n)\ell + 2C_1(m-n+1)\ell n\left(\frac{m-n}{\ell}\right) + 2C_1\ell + (m-n)C_3, \tag{4.54}$$

since

$$\sum_{u=\ell+1}^{m-n} \frac{1}{u} \leq \int_{\ell}^{m-n} \frac{dx}{x} = \ell n\left(\frac{m-n}{\ell}\right) \quad . \tag{4.55}$$

It follows that $|\gamma_{n,m}| = O((m-n)\ell n(m-n))$ for large $m-n$ by letting $\ell = \ell n(m-n)$.

Suppose now that $\alpha_{k,\ell} = \mu_k \mu_\ell$, $\mu_k = O(k^{-1})$, and that there exists an $f(u) = O(u^{-\nu})$ with the desired properties. Then, since it suffices to consider only $\mu_k = k^{-1}$,

$$|\gamma_{n,m}| \le 2\sum_{u=1}^{m-n} f(u) \sum_{k=n}^{m-u} \frac{1}{k(k+u)} + f(o) \sum_{k=n}^{m} \frac{1}{k^2} \quad . \tag{4.56}$$

For all $1 \le u \le m-n$ $(n \ge 2)$,

$$\sum_{k=n}^{m-u} \frac{1}{k(k+u)} \le \int_{n-1}^{m-u} \frac{dx}{x(x+u)} = \frac{1}{u} \ln\left(\frac{(m-u)(n+u-1)}{m(n-1)}\right) \quad . \tag{4.57}$$

Similarly, for $2 \le n \le m$,

$$\sum_{k=n}^{m} \frac{1}{k^2} \le \int_{n-1}^{m} \frac{dx}{x^2} = \frac{1}{n-1} - \frac{1}{m} \quad . \tag{4.58}$$

For all $1 \le u \le \ell \le n < m$,

$$\ln\left(\frac{(m-u)(n+u-1)}{m(n-1)}\right) \le \ln\left(\frac{n+\ell-1}{n-1}\right) \quad , \tag{4.59}$$

for all $1 \le \ell < u < n < m-n$,

$$\ln\left(\frac{(m-u)(n+u-1)}{m(n-1)}\right) \le \ln 2 \quad , \tag{4.60}$$

and for all $n \le u \le m-n$,

$$\ln\left(\frac{(m-u)(n+u-1)}{m(n-1)}\right) \le \ln u \quad . \tag{4.61}$$

Hence, for all $1 \le \ell \le n-2 < m-n-2$, with $C_1 = \max_u f(u)$,

$$|\gamma_{n,m}| \le 2C_1 \ell \ln\left(\frac{n+\ell-1}{n-1}\right) + 2\ln 2 \sum_{u=\ell+1}^{n-1} \frac{1}{u} f(u)$$

$$+ 2\sum_{u=n}^{m-n} f(u) \frac{\ln(u)}{u} + f(o)\left(\frac{1}{n-1} - \frac{1}{m}\right) \quad . \tag{4.62}$$

Since $f(u) = O(u^{-\nu})$, there exists a $C_2 > 0$ and an ℓ_o such that for all $\ell \ge \ell_o$, $f(u) \le C_2 u^{-\nu}$, so that

$$|\gamma_{n,m}| \leq 2C_1 \ell \ln\left(\frac{n+\ell-1}{n-1}\right) + 2C_2 \ln 2 \sum_{u=\ell+1}^{n-1} \frac{1}{u^{\nu+1}}$$

$$+ 2C_2 \sum_{u=n}^{m-n} \frac{\ln(u)}{u^{\nu+1}} + f(o)\left(\frac{1}{n-1} - \frac{1}{m}\right) . \tag{4.63}$$

Thus, for some fixed C_3, C_4, C_5,

$$|\gamma_{n,\infty}| \leq 2C_1 \ell \ln\left(\frac{n+\ell-1}{n-1}\right) + C_3(\ell)^{-\nu} + C_4(n-1)^{-\nu} + C_5 \frac{\ln(n-1)}{n-1} . \tag{4.64}$$

Substituting $\ell = n^{\beta}$, $\beta > 0$, in (4.64), and using the fact that $\ln(1 + x) \leq x$ for all $x > -1$, one obtains

$$|\gamma_{n,\infty}| = O(2n^{2\beta-1} + n^{-\beta\nu}). \quad \text{Q.E.D.} \tag{4.65}$$

Finally, if $\beta = (\nu + 2)^{-1}$, then $|\gamma_{n,\infty}| = O(n^{-\nu/\nu+2})$.

Enough machinery has now been developed to prove the following useful corollary.

COROLLARY 2. *Define*

$$\rho_C(k,\ell) = E\{C_k' C_\ell\} \tag{4.66}$$

and

$$\rho_F(k,\ell) = \|E\{F_k F_\ell\} - R_{xx}^2\| . \tag{4.67}$$

Suppose that there exists a real-valued function $f(u) = O(u^{-\nu})$ $(\nu \geq 1)$ *such that*

$$\max\{|\rho_C(k,k+u)|, \rho_F(k,k+u)\} \leq f(u) \tag{4.68}$$

for all positive integer k *and for all non-negative integer* u. *Furthermore, suppose that* $\mu_k = O(k^{-1})$, $\lim_{k\to\infty} k\mu_k > 0$, *and* $E\{\|F_k\|^q\}$ $(q > 2\nu^{-1}(\nu + 2))$ *is bounded. Then Assumptions A1 through A5 of the theorem are satisfied and hence,* $|V_n| \overset{a.s.}{\to} 0$ *as* $n \to \infty$.

PROOF. First, consider assumption A3. Define $S_{a,n}$ by

$$S_{a,n} = \sum_{k=a+1}^{a+n} (F_k - R_{xx})w, \tag{4.69}$$

where $w \in R^p$. Clearly, assumption A3 is satisfied if and only if $n^{-1}|S_{a,n}| \overset{a.s.}{\to} 0$ as $n \to \infty$ for all $w \in R^p$ and for all $a = 1,2,\ldots$. Define $M_{a,n} = \max\{|S_{a,1}|,\ldots,|S_{a,n}|\}$. Let $\{n_k\}$ be an increasing sequence of positive integers such that $n_k \to \infty$ as $k \to \infty$. For all $n_k \le n \le n_{k+1}-1$,

$$n^{-1}|S_{a,n}| \le n_k^{-1}|S_{a,n_k-1}| + n_k^{-1} M_{a+n_k-1,n_{k+1}-n_k} . \tag{4.70}$$

Clearly,

$$E\{|S_{a,n}|^2\} = \sum_{k=a+1}^{a+n} \sum_{\ell=a+1}^{a+n} w' E\{F_k F_\ell - R_{xx}^2\} w$$

$$\le |w|^2 \sum_{k=a+1}^{a+n} \sum_{\ell=a+1}^{a+n} \rho_F(k,\ell) = O(n \ell n\ n) , \tag{4.71}$$

from Lemma 7. Letting $n_k = k^2$, $n_k^{-2} E\{|S_{a,n_k}|^2\}$ is summable from (4.71). The Chebychev inequality and the Borel-Cantelli Lemma thus imply that $n_k^{-1}|S_{a,n_k-1}| \overset{a.s.}{\to} 0$ as $k \to \infty$. With $g(F_{a,n}) = E\{|S_{a,n}|^2\}$, Lemma 5 and (4.71) easily yield $E\{n_k^{-2} M_{a+n_k-1,n_{k+1}-n_k}^2\} = O((\ell n\ k/k)^3)$, which is summable. Hence, $n_k^{-1} M_{a+n_k-1,n_{k+1}-n_k} \overset{a.s.}{\to} 0$ as $k \to \infty$ so that, by (4.69), $n^{-1}|S_{a,n}| \overset{a.s.}{\to} 0$ as $n \to \infty$ and A3 is satisfied.

Now consider A5. Let $\{n_k\}$ be an increasing sequence of positive integers such that $n_k \to \infty$ as $k \to \infty$ and let $L_k = \{n_k, n_k+1, \ldots, n_{k+1}-1\}$. For all $n \in L_k$,

$$|F_n(S - S_{n-1})| = |F_n(\sum_{i=n}^{n_{k+1}-1} \mu_i C_i + S - S_{n_{k+1}-1})|$$

$$\leq \|F_n\|(|\sum_{i=n}^{n_{k+1}-1} \mu_i C_i| + |S - S_{n_{k+1}-1}|)$$

$$\leq n_k^{-\beta} \max_{\ell \in L_k} \|F_\ell\| n_k^{\beta} \max_{\ell \in L_k} |\sum_{i=\ell}^{n_{k+1}-1} \mu_i C_i|$$

$$+ n_k^{-\beta} \max_{\ell \in L_k} \|F_\ell\| n_k^{\beta} |S - S_{n_{k+1}-1}|, \qquad (4.72)$$

where as yet, $\beta > 0$ is arbitrary.

Defining

$$S_{a,n} = \sum_{i=a-n+1}^{a} \mu_i C_i \quad , \qquad (4.73)$$

and $M_{a,n} = \max\{|S_{a,1}|,\ldots,|S_{a,n}|\}$, (4.72) becomes

$$|F_n(S - S_{n-1})| \leq n_k^{-\beta} \max_{\ell \in L_k} \|F_\ell\| (n_k^{\beta} M_{n_{k+1}-1, n_{k+1}-n_k} + n_k^{\beta} |S - S_{n_{k+1}-1}|). \qquad (4.74)$$

Since

$$E\{|S_{a,n}|^2\} = \sum_{i=a-n+1}^{a} \sum_{j=a-n+1}^{a} \mu_i \mu_j \rho_C (i,j) \quad , \qquad (4.75)$$

with $g(F_{a,n}) = E\{|S_{a,n}|^2\}$, Lemma 6 applies so that

$$E\{M^2_{n_{k+1}-1, n_{k+1}-n_k}\} \leq (\log_2 2(n_{k+1}-n_k))^2 \, g(F_{n_{k+1}-1, n_{k+1}-n_k}) . \qquad (4.76)$$

From (c) of Lemma 7, $g(F_{n_{k+1}-1,n_{k+1}-n_k}) = O(n_k^{-\nu/\nu+2})$. From (4.43) and Lemma 7, $E\{|S - S_{n_{k+1}-1}|^2\} = O(n_{k+1}^{-\nu/\nu+2})$. If $\{n_k\}$ and $\beta > 0$ can be chosen such that (i) $\sum_{k=1}^{\infty} k^{-q\beta} < \infty$, (ii) $\sum_{k=1}^{\infty} n_k^{2\beta} E\{|S - S_{n_{k+1}-1}|^2\} < \infty$, and (iii) $\sum_{k=1}^{\infty} n_k^{2\beta} E\{M^2_{n_{k+1}-1,n_{k+1}-n_k}\} < \infty$, then the Markov inequality, the Borel-Cantelli Lemma, and (4.74) will show that $|F_n(S - S_{n-1})| \overset{a.s.}{\to} 0$ as $n \to \infty$. It is easily verified that for $n_k = k^{\alpha}$, $q^{-1} < \beta < \nu(2\nu+4)^{-1}$, and $\alpha > (\nu+2)(\nu-2\beta(\nu+2))^{-1}$, (i), (ii), and (iii) are satisfied. Finally, $\mu_k = O(k^{-1})$, $E\{\|F_n\|^q\}$ bounded, and $\lim_{k\to\infty} k\mu_k > 0$ imply A1 and A2; while (ii) and (iii) imply A4. Q.E.D.

B. Application of Corollary 2

In this section, the results of the previous section are applied to the algorithms discussed in Chapter II. In order to apply Corollary 2, it is necessary to establish asymptotic decay rates on $\rho_C(k,\ell)$, and $\rho_F(k,\ell)$, as defined by (4.66) and (4.67). Define $\rho_P(k,\ell) = |E\{P_k'P_\ell\} - P'P|$. From (4.66), (4.4), and (4.5),

$$
\begin{aligned}
|\rho_C(k,\ell)| &= |E\{C_k'C_\ell\}| \\
&= |E\{P_k'P_\ell\} - w_o'(E\{F_\ell P_k\} + E\{F_k P_\ell\}) + w_o'\{F_k F_\ell\}w_o| \\
&\le |E\{P_k'P_\ell\} - P'P| + |P'P - w_o'E\{F_\ell P_k\}| \\
&\quad + |P'P - w_o'E\{F_k P_\ell\}| + |w_o|^2 \|E\{F_k F_\ell\} - R_{xx}^2\| \\
&\le \rho_P(k,\ell) + |w_o|\|R_{xx}P - E\{F_\ell P_k\}| \\
&\quad + |w_o|\cdot|R_{xx}P - E\{F_k P_\ell\}| + |w_o|^2\rho_F(k,\ell) \quad . \qquad (4.77)
\end{aligned}
$$

Hence, by defining

$$\rho_{FP}(k,\ell) = |R_{xx}P - E\{F_k P_\ell\}| \ , \tag{4.78}$$

$\rho_C(k,\ell)$ can be bounded as

$$|\rho_C(k,\ell)| \leq \rho_P(k,\ell) + |w_o|\rho_{FP}(k,\ell) + |w_o|\rho_{FP}(\ell,k) + |w_o|^2\rho_F(k,\ell) \ . \tag{4.79}$$

With (4.79) established, it is easily seen that in order to establish decay rates on $\rho_C(k, \ )$, and $\rho_F(k,\ell)$, it is sufficient to consider $\rho_P(k,\ell)$, $\rho_{FP}(k,\ell)$, and $\rho_F(k,\ell)$. Before treating specific examples, expressions for ρ_P, ρ_{FP} and ρ_F will be developed which are sufficiently general to cover most of the algorithms treated in Chapter II.

Let $\{X_j\}_{j=-\infty}^{\infty}$ and $\{N_j\}_{j=-\infty}^{\infty}$ be sequences of R^p-valued zero-mean random variables, and let $\{s_j\}_{j=-\infty}^{\infty}$ be a sequence of real-valued zero-mean random variables. It is assumed that $E\{X_k N_{k+u}'\} = E\{N_k N_{k+u}'\}$ and $E\{s_k N_{k+u}\} = 0$ for all integers k and u. The ijth element of a matrix A will be denoted by $(A)_{i,j}$. It is assumed that all fourth-order moments correspond to stationarity; e.g., $E\{s_\ell s_{\ell+i}(X_{\ell+j})_{m_1}(X_{\ell+k})_{m_2}\}$ is independent of ℓ. Define

$$R_{xx}(u) = E\{X_k X_{k+u}'\} \ , \tag{4.80}$$

$$P_s(u) = E\{s_k X_{k+u}\} \ , \tag{4.81}$$

and

$$\rho_s(u) = E\{s_k s_{k+u}\} \ . \tag{4.82}$$

Consistent with the notation used previously, define $R_{xx}(0) = R_{xx}$, and $P_s(0) = P$.

The following definitions for P_k and F_k will be sufficient for the purposes of the present analysis. Define

$$P_k = s_k X_k \quad , \tag{4.83}$$

and

$$F_k = X_k X_k' \quad . \tag{4.84}$$

Clearly, $E\{P_k\} = P$ and $E\{F_k\} = R_{xx}$, so that (4.6) and (4.7) are satisfied. Most of the algorithms which have been proposed for use in adaptive signal processing use $F_k = X_k X_k'$ and either $P_k = s_k X_k$ or $P_k = P = E\{s_k X_k\}$. In case $P_k = P$, $\rho_P(k,\ell) \equiv 0$ and $\rho_{FP}(k,\ell) \equiv 0$, so that in this case one need only consider $\rho_F(k,\ell)$.

First consider $\rho_F(k,\ell)$. From (4.84),

$$E\{F_k F_{k+u}\} = E\{X_k X_k' X_{k+u} X_{k+u}'\} \quad . \tag{4.85}$$

In case X_j is a multivariate Gaussian Random Process (GRP), it is easily shown that from (4.80),

$$E\{X_k X_k' X_{k+u} X_{k+u}'\} = R_{xx}^2 + R_{xx}^2(u) + R_{xx}(u)\mathrm{tr}(R_{xx}(u)), \tag{4.86}$$

by recalling that if Y_1, Y_2, Y_3, and Y_4 are jointly normally distributed *zero-mean random* variables, then $E\{Y_1Y_2Y_3Y_4\} = E\{Y_1Y_2\}E\{Y_3Y_4\} + E\{Y_1Y_3\}E\{Y_2Y_4\} + E\{Y_1Y_4\}E\{Y_2Y_3\}$. In general, define $\kappa_1(u)$ such that

$$E\{X_k X_k' X_{k+u} X_{k+u}'\} = R_{xx}^2 + R_{xx}^2(u) + R_{xx}(u)\mathrm{tr}(R_{xx}(u)) + \kappa_1(u), \tag{4.87}$$

so that

$$E\{F_k F_{k+u}\} - R_{xx}^2 = R_{xx}^2(u) + R_{xx}(u)\mathrm{tr}(R_{xx}(u)) + \kappa_1(u). \tag{4.88}$$

Next consider $\rho_p(k,\ell)$. From (4.83),

$$E\{P_k' P_{k+u}\} = E\{s_k X_k' s_{k+u} X_{k+u}\} . \tag{4.89}$$

In case s_k, X_k, s_{k+u}, and X_{k+u} are jointly normal, then

$$E\{s_k X_k' s_{k+u} X_{k+u}\} = P'P + \rho_s(u)\mathrm{tr}(R_{xx}(u)) + P_s'(u)P_s(u). \tag{4.90}$$

In general, define $\kappa_2(u)$ such that

$$E\{s_k X_k' s_{k+u} X_{k+u}\} = P'P + \rho_s(u)\mathrm{tr}(R_{xx}(u)) + P_s'(u)P_s(-u) + \kappa_2(u). \tag{4.91}$$

Then $\rho_p(k,k+u)$ can be determined from

$$E\{P_k' P_{k+u}\} - P'P = \rho_s(u)\mathrm{tr}(R_{xx}(u)) + P_s'(u)P_s(-u) + \kappa_2(u). \tag{4.92}$$

It is important to reiterate that in case $P_k = P$, then $\rho_p(k,k+u) \equiv 0$.

Finally, consider $\rho_{FP}(k,\ell)$. From (4.83) and (4.84),

$$E\{F_k P_{k+u}\} = E\{X_k X_k' s_{k+u} X_{k+u}\} . \tag{4.93}$$

Proceeding as before, (4.93) can be expressed as

$$E\{X_k X_k' s_{k+u} X_{k+u}\} = R_{xx}P + P_s(-u)\mathrm{tr}(R_{xx}(u)) + R_{xx}(u)P_s(-u) + \kappa_3(u), \tag{4.94}$$

where $\kappa_3(u) \equiv 0$ in the normal case. Hence, $\rho_{FP}(k,k+u)$ can be determined from (4.78) and

$$E\{F_k P_{k+u}\} - R_{xx}P = P_s(-u)\mathrm{tr}(R_{xx}(u)) + R_{xx}(u)P_s(-u) + \kappa_3(u) . \tag{4.95}$$

Again, in case $P_k = P = E\{s_k X_k\}$, then $\rho_{FP}(k,\ell) \equiv 0$.

A useful fact for the application of the above results is that for a pxp matrix A,

$$\|A\| \leq \sum_{j=1}^{p} \left(\sum_{i=1}^{p} (A)_{i,j}^2 \right)^{1/2} , \tag{4.96}$$

i.e., the norm of A is bounded by the sum of the Euclidean lengths of its columns (or rows), as shown, e.g., by Rudin [62]. Define

$$g_1(u) = \max_{1 \leq i,j \leq p} |(R_{xx}(u))_{i,j}| , \tag{4.97}$$

and

$$g_2(u) = \max_{1 \leq m \leq p} |(P_s(u))_m| ; \tag{4.98}$$

then, from (4.96), $\|R_{xx}(u)\| \leq p^{3/2} g_1(u)$. From (4.67), (4.88), and (4.97),

$$\rho_F(k,k+u) \leq p^3 \, g_1^2(u) + \|\kappa_1(u)\| . \tag{4.99}$$

From (4.92), (4.97) and (4.98),

$$\rho_P(k,k+u) \leq p \cdot |\rho_s(u)| g_1(u) + p \cdot g_2^2(u) + |\kappa_2(u)| . \tag{4.100}$$

From (4.78), (4.95), (4.97), and (4.98),

$$\rho_{FP}(k,k+u) \leq p^{3/2} g_2(u) g_1(u) + p^2 g_1(u) g_2(u) + |\kappa_3(u)| . \tag{4.101}$$

Now, from (4.79), (4.100), (4.101), and (4.99),

$$\begin{aligned} |\rho_C(k,k+u)| \leq\ & p \cdot |\rho_s(u)| g_1(u) + 2p^2 |w_o| g_1(u) g_2(u) \\ & + p^3 |w_o|^2 g_1^2(u) + p \, g_2^2(u) + |w_o|^2 \, \|\kappa_1(u)\| \\ & + p|\kappa_2(u)| + 2|w_o| \cdot |\kappa_3(u)| , \end{aligned} \tag{4.102}$$

by noting that $g_1(u)$ and $g_2(u)$ are even functions. Finally, by defining

$$g(u) = \max \left(g_1(u), \, g_2(u), \, |\rho_s(u)| \right) , \tag{4.103}$$

there exist positive constants a_1, a_2, a_3, and a_4 such that

$$\max\left(\rho_F(k,k+u),\ |\rho_C(k,k+u)|\right)$$
$$\leq a_1 g^2(u) + a_2\|\kappa_1(u)\| + a_3|\kappa_2(u)| + a_4|\kappa_3(u)| \quad . \tag{4.104}$$

THE NORMAL CASE

In case $\{X_j\}$, $\{N_j\}$, $\{s_j\}$ are GRP's, then $\|\kappa_1(u)\| \equiv |\kappa_2(u)| \equiv |\kappa_3(u)| \equiv 0$, so that

$$\max\left(\rho_F(k,k+u), |\rho_C(k,k+u)|\right) \leq a_1 g^2(u) \quad . \tag{4.105}$$

Also, in the normal case, $E\{\|F_n\|^7\}$ is bounded. Hence, all algorithms of the form of (3.1), with $F_k = X_k X_k'$, $P_k = s_k X_k$ or $P_k = P = E\{s_k X_k\}$, and $\mu_k = O(k^{-1})$, $\lim_{k\to\infty} k\mu_k > 0$, satisfy the hypotheses of Corollary 2 and hence, converge almost surely provided that $g(u)$ in (4.105) is $O(u^{-1/2})$. This result suggests that essentially all one needs to do to establish a.s. convergence for this class of algorithms in the normal case is to ensure that all scalar correlation functions $\gamma(u)$ which can be computed for $\{s_j\}$, $\{X_j\}$, satisfy $\lim_{u\to\infty} u^{1/2}|\gamma(u)| < \infty$.

EXAMPLE 1

Let $\{n(t): -\infty < t < \infty\}$ and $\{s(t): -\infty < t < \infty\}$ be zero mean jointly wide-sense stationary finite variance Gaussian random processes. Define $x(t) = n(t) + s(t)$, and assume that $E\{s(t)n(t+\tau)\} = 0$ for all t,τ. Define the "data vector" $X'(t) = (x(t), x(t-D), \ldots, x(t-(p-1)D))$. Suppose that it is desired to form a linear MMSE estimate of $s(t+\alpha)$ at $t = kT$, $k = 0,1,2, \ldots$, based on the "data vector" $X_k = X(t)|_{t=kT}$, where D is an integer multiple of T. Denoting $s(t+\alpha)|_{t=kT}$ by s_k, it is easily shown that the desired linear MMSE estimate of s_k is given by $\hat{s}_k = w_o' X_k$, where w_o is the (assumed

unique) solution of $R_{xx}w = P$, $R_{xx} = E\{X_k X_k'\}$, and $P = E\{s_k X_k\}$. Defining $\gamma_x(\tau) = E\{x(t)x(t+\tau)\}$, $\gamma_n(\tau) = E\{n(t)n(t+\tau)\}$, $\gamma_s(\tau) = E\{s(t)s(t+\tau)\}$, $R_{xx}(u) = E\{X_k X_{k+u}'\}$ and $P_s(u) = E\{s_k X_{k+u}\}$, it is easily seen that

$$(R_{xx}(u))_{i,j} = \gamma_x(uT + (i-j)D)$$

$$= \gamma_s(uT + (i-j)D) + \gamma_n(uT + (i-j)D) \quad , \tag{4.106}$$

and

$$P_s(u))_m = \gamma_s(uT - \alpha - (m-1)D). \tag{4.107}$$

Define $S_s(f) = F\{\gamma_s(\tau)\}$, $S_n(f) = F\{\gamma_n(\tau)\}$ to be the spectral densities for the processes $s(t)$ and $n(t)$, respectively. Suppose the signal spectral density is the rational density,

$$S_s(f) = \frac{b_2}{f^2 + b_1^2} \quad , \tag{4.108}$$

and the noise spectral density is the ideal lowpass density,

$$S_n(f) = \begin{cases} b_3 \ , & |f| \leq B \\ 0 \ , & |f| > B \end{cases} \quad , \tag{4.109}$$

where b_1, b_2, and b_3 are positive constants. Then

$$\gamma_s(\tau) = \frac{b_2}{b_1} \pi e^{-2\pi b_1 |\tau|} \quad , \tag{4.110}$$

and

$$\gamma_n(\tau) = 2b_3 B \frac{\sin 2\pi B\tau}{2\pi B\tau} \quad . \tag{4.111}$$

It is easily seen that for this example, $g(u)$ defined by (4.103) is $O(u^{-1})$. Suppose that P is known and consider the algorithm

$$W_{k+1} = W_k + \frac{1}{k}(P - X_k X_k' W_k), \tag{4.112}$$

for $k \geq 1$, with W_1 arbitrary. Clearly, all of the assumptions of

Corollary 2 are satisfied and hence, $W_k \overset{a.s.}{\to} w_o$ as $k \to \infty$. Now, suppose that P is unknown but s_k is available. Then the algorithm

$$W_{k+1} = W_k + \frac{1}{k}(s_k X_k - X_k X_k' W_k) \tag{4.113}$$

will converge a.s. to w_o. It is easily shown that algorithms such as

$$W_{k+1} = W_k + \frac{1}{k}(P - \frac{1}{K} \sum_{\ell=k-K+1}^{k} X_\ell X_\ell' W_k) \tag{4.114}$$

and

$$W_{k+1} = W_k + \frac{1}{kK} \sum_{\ell=k-K+1}^{k} (s_\ell X_\ell - X_\ell X_\ell' W_k) \tag{4.115}$$

will also converge a.s. to w_o for any finite positive integer K.

The above example shows the ease with which the assumptions of Corollary 2 can be established for a rather large family of algorithms in the normal case. A straightforward extension of Example 1 to arbitrary rational spectral densities yields identical conclusions; i.e., if $n(t)$ and/or $s(t)$ in Example 1 are finite-order autoregressive moving average processes, the conclusions remain unchanged. Extensions of Example 1 to the adaptive array processing of homogeneous random fields is straightforward, but notationally somewhat cumbersome.

The application of Corollary 2 to the non-normal case is, in general, more difficult than Example 1 suggests for the normal case. Two possible approaches for the non-normal case are as follows: (i) compute bounds on $\rho_C(k,\ell)$ and $\rho_F(k,\ell)$ either directly or via (4.78) and (4.79) and apply Corollary 2 directly, or (ii) compute bounds on the fourth cumulant functions $\kappa_1(u), \kappa_2(u)$, and $\kappa_3(u)$,

apply (4.104) and then apply Corollary 2. Example 2 below considers a rather special case of the former approach. An additional difficulty arises in the non-normal case in establishing that $E\{\|F_n\|^q\}$ is bounded.

EXAMPLE 2

Let $\{n_k\}_{k=-\infty}^{\infty}$, $\{s_k\}_{k=-\infty}^{\infty}$ be independent, zero mean, finite variance, wide-sense stationary stochastic processes. Assume that both $\{n_k\}_{k=-\infty}^{\infty}$ and $\{s_k\}_{k=-\infty}^{\infty}$ are M-dependent. Recall the definition of M-dependence from Chapter III. Define $x_k = s_k + n_k$, and $X_k' = (x_k, x_{k-1}, \ldots, x_{k-p+1})$. Define $F_k = X_k X_k'$ and assume that $E\{\|F_k\|^q\}$, $q>2$, is bounded. Suppose that it is desired to form a linear MMSE estimate of s_k based on the data vector X_k. The desired estimate is easily shown to be $\hat{s}_k = w_o' X_k$, where w_o is the (assumed unique) solution to $R_{xx} w = P$, $R_{xx} = E\{X_k X_k'\}$, and $P = E\{s_k X_k\}$. It is easily seen that $\|\rho_F(k,k+u)\| = \|E\{X_k X_k' X_{k+u} X_{k+u}'\} - R_{xx}^2\| = 0$ for all $u > M_1$ for some $M_1 > M$. Similarly, $\rho_{FP}(k,k+u)$, and $\rho_P(k,k+u)$ are easily shown to be zero for all $|u| > M_2$ (for some $M_2 > M$) for either $P_k = s_k X_k$ or $P_k = P = E\{s_k X_k\}$. Letting $\mu_k = k^{-1}$, all of the assumptions of Corollary 2 have been established. It is not difficult to show that algorithms such as (4.114) and (4.115) will also converge a.s. to w_o.

A slight generalization of the result summarized by (4.104) and (4.105) seems to be useful for algorithms having the form of (3.1) with

$$P_k = \frac{1}{K_k} \sum_{\ell=k-K_k+1}^{k} s_\ell X_\ell \tag{4.116}$$

and

$$F_k = \frac{1}{K_k} \sum_{\ell=k-K_k+1}^{k} X_\ell X_\ell' \quad , \tag{4.117}$$

where K_k is a positive integer-valued function of k. Clearly, $E\{P_k\} = P$ and $E\{F_k\} = R_{xx}$, so that (4.6) and (4.7) are satisfied. In case $K_k = 1$, (4.116) and (4.117) reduce to (4.83) and (4.84), respectively. Denoting the right-hand side of (4.104) by $h(u)$, (4.104) can be restated for the case at hand as

$$\max(\rho_F(k,k+u), |\rho_C(k,k+u)|) \leq \alpha_{k,u} \sum_m \sum_n h(n-m) \quad , \tag{4.118}$$

(4.118)

where $\alpha_{k,u} = (K_k K_{k+u})^{-1}$, and the sums are over the index values $k - K_k + 1 \leq m \leq k$ and $k + u - K_{k+u} + 1 \leq n \leq k + u$. The techniques used in Lemma 7 can be applied to the double sum appearing in (4.118) to obtain

$$\alpha_{k,u} \sum_m \sum_n h(n-m) = \alpha_{k,u} \sum_{\nu=u-K_{k+u}+1}^{u+K_k-1} h(\nu)\, \beta_{k,u,\nu} \tag{4.119}$$

where $\beta_{k,u,\nu} = \min(o,u-\nu) - \max(-K_k, u-\nu-K_{k+u})$. In case $K_k = K$ (a constant), then

$$\alpha_{k,u} \sum_m \sum_n h(n-m) = \frac{1}{K^2} \sum_{\nu=u-K+1}^{u+K-1} h(\nu)(K - |\nu-u|) \quad . \tag{4.120}$$

Another special case of interest is $K_k = k$; then

$$\alpha_{k,u} \sum_m \sum_n h(n-m) = \frac{1}{k(k+u)} \sum_{\nu=1-k}^{\min(u,o)} h(\nu)(k+\nu) + \frac{1}{k+u} \sum_{\nu=1}^{u} h(\nu)$$

$$+ \frac{1}{k} \sum_{\nu=u+1}^{o} h(\nu) + \frac{1}{k(k+u)} \sum_{\max(u+1,1)}^{u+k-1} h(\nu)(u-\nu+k), \tag{4.121}$$

where, by convention, $\sum_a^b = 0$ if $b < a$.

The result (4.121) can indeed be used to examine the convergence properties of algorithms having the form of (3.1) with

$$P_k = \frac{1}{k} \sum_{\ell=1}^{k} s_\ell X_\ell \quad , \tag{4.122}$$

and

$$F_k = \frac{1}{k} \sum_{\ell=1}^{k} X_\ell X_\ell' \quad . \tag{4.123}$$

This resulting algorithm seems to be of interest for several reasons and is treated from an alternative viewpoint in detail in the following section.

C. A Simple a.s. Convergence Result

In this section, a simple a.s. convergence result is established which does not require all of the machinery developed in Section IV-A. The result is of interest because of its simplicity and the information provided on the convergence rate for algorithms satisfying the rather restrictive assumptions made in the theorem stated below.

THEOREM. *Let* $\{W_n\}_{n=1}^{\infty}$ *be given by (3.1) with* $W_1 = w_1$ *arbitrary. Suppose that there exists sequences* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$ *of non-negative numbers (possibly random) satisfying*

$$\|F_n - R_{xx}\| \overset{a.s.}{\leq} a_n \quad , \tag{4.124}$$

and

$$\|F_n w_o - P_n\| \overset{a.s.}{\leq} b_n \quad . \tag{4.125}$$

Furthermore, suppose that there exists a positive integer n_o *(possibly random) such that for all* $n \geq n_o$, $0 \overset{a.s.}{<} \mu_n(\lambda_{min} - a_n) \overset{a.s.}{\leq} 1$, *where* $\lambda_{min} = \lambda_{min}(R_{xx})$. *Then for all* $n \geq n_o$,

$$\|V_{n+1}\| \overset{a.s.}{\leq} \|V_{n_o}\| \cdot \prod_{k=n_o}^{n} (1 - \mu_k d_k)$$

$$+ \max_{n_o \leq k \leq n} (b_k/d_k) \cdot (1 - \prod_{j=n_o}^{n} (1 - \mu_j d_j)) \quad , \qquad (4.126)$$

where $d_k = \lambda_{min} - a_n$. *Furthermore, if* $\sum \mu_k d_k \overset{a.s.}{=} \infty$, *and* $b_k d_k^{-1} \overset{a.s.}{\to} 0$ *as* $k \to \infty$, *then* $\|V_n\| \overset{a.s.}{\to} 0$ *as* $n \to \infty$.

PROOF. From (4.2) through (4.7),

$$V_{n+1} = V_n - \mu_n(F_n V_n + F_n w_o - P_n)$$

$$= V_n - \mu_n R_{xx} V_n - \mu_n(F_n V_n + F_n w_o - P_n - R_{xx} V_n), \qquad (4.127)$$

so that for all $n \geq n_o$,

$$\|V_{n+1}\| \leq (1 - \mu_n \lambda_{min})\|V_n\| + \mu_n \|F_n - R_{xx}\| \, \|V_n\|$$

$$+ \mu_n \|F_n w_o - P_n\| \overset{a.s.}{\leq} (1 - \mu_n d_n) \|V_n\| + \mu_n b_n. \qquad (4.128)$$

Iterating (4.128), for all $n \geq n_o$,

$$\|V_{n+1}\| \leq \|V_{n_o}\| \prod_{k=n_o}^{n} (1 - \mu_k d_k) + \sum_{k=n_o}^{n} \prod_{j=k+1}^{n} (1 - \mu_j d_j) \mu_k d_k (b_k d_k^{-1}). \qquad (4.129)$$

Since all terms appearing in the sum in (4.129) are a.s. non-negative, (4.126) follows immediately from (4.129) with the aid of Lemma 3. Furthermore, if $\sum \mu_k d_k \overset{a.s.}{=} \infty$ and $b_k d_k^{-1} \overset{a.s.}{\to} 0$ as $k \to \infty$, (4.129) and Lemma 4 show that $\|V_n\| \overset{a.s.}{\to} 0$ as $n \to \infty$. Q.E.D.

In order for the above theorem to provide useful information regarding convergence rate, the sequences $\{a_n\}$ and $\{d_n\}$ and the integer n_o must be known. As mentioned in the previous section, one application of the above theorem is to algorithms having the form of

(3.1) with P_k and F_k given by (4.122) and (4.123), respectively. For this case, the results of Serfling ([58],[59]) presented in Section IV-A as Lemma 5 can be applied to establish (4.124) and (4.125). It seems that such algorithms should converge with the fastest convergence rate of any stochastic approximation algorithms under consideration, since $F_n \to R_{xx}$ and $P_k \to P$. It seems likely that the above theorem can be used to choose $\{\mu_k\}$ to maximize the convergence rate for such algorithms. An important special case of the above theorem is the (deterministic) steepest descent algorithm.

D. Discussion

In this chapter, new a.s. convergence results are developed, applied, and discussed. In Section IV-A, the main results of this work are developed and the extreme ease with which these results can be applied at least in the normal case is illustrated in Section IV-B. Indeed, it is shown that algorithms (4.112) and (4.113) converge a.s. to w_o in the normal case if X_k, s_k are samples of finite variance finite order autoregressive moving average processes, a case of great practical interest. Although these results seem to be the strongest convergence results yet obtained under the weakest conditions, there are several open issues remaining. In practice, convergence rate is an extremely important issue. This problem is treated in Section IV-C under overly restrictive conditions. The asymptotic distribution of V_n seems to be a topic of considerable theoretical interest. Truncated algorithms such as (3.41) as well as algorithms using a random "gain sequence" also seem to be of interest. Practically, the most important issue is probably an analytical investigation of convergence properties in nonstationary environments.

V. SPECIAL FORMS OF DATA CORRELATION MATRICES

In Chapters II through IV, the stochastic solution of the linear equation

$$R_{xx}w = P \ , \tag{5.1}$$

where R_{xx} is a p x p correlation matrix and P is a p x 1 correlation vector is considered. In case R_{xx} and P are known, the required solution, w_o, of (5.1) can be obtained directly. This chapter is devoted to computationally efficient techniques for solving (5.1) when R_{xx} is either a Toeplitz matrix, i.e., the ij^{th} element of R_{xx} is a function only of i-j, or a "block" Toeplitz matrix, i.e., for $p = ML$, there are M^2 L x L submatrices of R_{xx} arranged in a Toeplitz form. The results of this chapter are computational algorithms which require far less computer storage and computation time than standard numerical techniques for solving (5.1).

A. Motivation: Array Processing of Homogeneous Fields

An important application of linear filtering theory is to the estimation of some component of a scalar-valued homogeneous random field. Let $\xi_n(t,x,y,z,)$ be a scalar homogeneous random field for $n = 1,2,\ldots,$ N. Let t denote time and (x,y,z) denote spatial coordinates in some suitable cartesian coordinate system. Furthermore, assume that the $\xi_n(.,.,.,.)$ are zero mean and uncorrelated, i.e., that $E\{\xi_n(t_1,x_1,y_1,z_1)\bar{\xi}_m(t_2,x_2,y_2.z_2) = 0$ for all $n \neq m$ and for all t_1,t_2, x_1,x_2,y_1,y_2,z_1,z_2, where the $-$ denotes complex conjugate. Then with $(\Delta_t,\Delta_x,\Delta_y,\Delta_z) = (t_2-t_1,x_2-x_1,y_2-y_1,z_2-z_1)$ and

$$\chi(t,x,y,z) = \sum_{n=1}^{N} \xi_n(t,x,y,z) \ , \tag{5.2}$$

the autocorrelation function for χ is given by

$$\rho_\chi(\Delta_t,\Delta_x,\Delta_y,\Delta_z) = \sum_{n=1}^{N} \rho_n(\Delta_t,\Delta_x,\Delta_y,\Delta_z) \quad , \tag{5.3}$$

where $\rho_\chi(\Delta_t,\Delta_x,\Delta_y,\Delta_z) = E\{\chi(t_1,x_1,y_1,z_1)\bar{\chi}(t_2,x_2,y_2,z_2)\}$, and $\rho_n(\Delta_t,\Delta_x,\Delta_y,\Delta_z) = E\{\xi_n(t_1,x_1,y_1,z_1)\bar{\xi}_n(t_2,x_2,y_2,z_2)\}$. From (5.3), $\chi(t,x,y,z)$ is a scalar homogeneous random field.

Suppose that there are L sensors located at coordinates $p_\ell = (x_\ell,y_\ell,z_\ell)(1\leq\ell\leq L)$ and that following the output of each sensor is a tapped delay line having M equally spaced taps. Assume that all L delay lines are identical and have a time delay of D between adjacent taps. Define the p-element "data vector" (p=ML) by

$$\begin{aligned} X'(t) = (&\chi(t,p_1),\ \chi(t,p_2),\ldots,\chi(t,p_L),\\ &\chi(t-D,p_1),\ \chi(t-D,p_2),\ldots,\chi(t-D,p_L),\\ &\vdots\\ &\chi(t-(M-1)D,p_1),\ \chi(t-(M-1)D,p_2),\ldots,\chi(t-(M-1)D,p_L). \end{aligned} \tag{5.4}$$

Note the data is ordered so the first P_L elements correspond to data observed at the input to the array at time t, the second P_L elements correspond to data observed at the input to the array at time $t-D$, and so on.

Suppose that it is desired to form a linear MMSE estimate of

$$s(t) = \xi_1(t-d,\ p_r) \tag{5.5}$$

at time $t = kT$, $k = 0,1,\ldots$, based on the data vector $X_k = X(t)|_{t=kT}$. It is noted that p_r need not correspond to one of the physical sensor locations and that d need not be an integer multiple of T. It is assumed that D is an integer multiple of T. Denoting $s(t)|_{t=kT}$

by s_k, it is easily shown that the desired estimate is given by $\hat{s}_k = w_o' X_k$, where w_o is the (assumed unique) solution of $R_{xx} w = P$, $R_{xx} = E\{\bar{X}_k X'_k\}$, and $P = W\{s_k \bar{X}_k\}$. In order to examine the special forms of R_{xx} that can arise in this application, it is convenient to note that for $m = 1,2,\ldots, ML(=p)$, the m^{th} element of $X(t)$ is given by

$$(X(t))_m = \chi(t - q_m D, p_m - q_m L) \quad , \tag{5.6}$$

where $q_m = [\frac{m-1}{L}]$, and $[\cdot]$ denotes the largest integer part. Then the ij^{th} element of R_{xx} is given by

$$\begin{aligned}(R_{xx})_{i,j} &= E\{\bar{\chi}(kT - q_i D, p_{i-q_i L})\chi(kT - q_i D, p_{j-q_j L})\} \\ &= \bar{\rho}_\chi((q_j - q_i)D, p_{j-q_j L} - p_{i-q_i L}) \quad .\end{aligned} \tag{5.7}$$

Similarly, the m^{th} element of P is given by

$$\begin{aligned}(P)_m &= (E\{s_k \bar{X}_k\})_m = \{\xi_1(kT-d, p_r)\bar{\chi}(kT-q_m D, p_{m-q_m L})\} \\ &= \rho_1(d-q_m D, p_{m-q_m L} - p_r),\end{aligned} \tag{5.8}$$

where the last equality follows from (5.2) and the uncorrelated assumption.

Now, some interpretations of the $\{\xi_n\}_{n=1}^N$ are in order. Suppose that ξ_2, e.g., corresponds to "sensor noise" which is uncorrelated from sensor to sensor. Then

$$\rho_2((q_j - q_i)D, p_{j-q_j L} - p_{i-q_i L}) = \rho_2((q_j - q_i)D, 0)\delta_{j-q_j L, i-q_i L}. \tag{5.9}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker delta. Suppose further that the remaining ξ_n are propagating plane waves. Then

$$\begin{aligned}&\rho_n((q_j - q_i)D, p_{j-q_j L} - p_{i-q_i L}) \\ &\quad = \rho_n((q_j - q_i)D - \tau_n(p_{j-q_j L} - p_{i-q_i L}), 0)\end{aligned} \tag{5.10}$$

for $n\in\{1,3,4,\ldots,N\}$, where $\tau_n(p_{\ell_1} - p_{\ell_2})$ is the propagation time from sensor ℓ_1 to sensor ℓ_2 for ξ_n, which is clearly a function of the propagation velocity, the direction cosines of the propagation direction, and the distance between sensors. From (5.7), (5.3), (5.9), and (5.10),

$$(R_{xx})_{i,j} = \sum_{\substack{n=1 \\ n\neq 2}}^{N} \bar{\rho}_n((q_j - q_i)D - \tau_n(p_{j-q_jL} - p_{i-q_iL}), 0) + \rho_2((q_j - q_i)D, 0)\delta_{j-q_jL,i-q_iL} \quad . \tag{5.11}$$

From (5.11), it is easily seen that the ij^{th} element of R_{xx} is a function of only $q_j - q_i$, $j - q_jL$, and $i - q_iL$. Hence for any $i,j\in\{1,2,\ldots,ML\}$, and any integer u such that $i+uL\in\{1,2,\ldots,ML\}$ and $j+uL\in\{1,2,\ldots,ML\}$, $(R_{xx})_{i+uL,j+uL} = (R_{xx})_{i,j}$. That is, R_{xx} can be expressed as

$$R_{xx} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_{M-1} \\ \alpha_{-1} & \alpha_0 & \alpha_1 & & \\ & \cdots & & & \\ \alpha_{1-M} & & \cdots & \alpha_{-1} & \alpha_0 \end{bmatrix} \quad , \tag{5.12}$$

where each α_ℓ is an $L \times L$ matrix. In other words, R_{xx} is a "block" Toeplitz matrix consisting of M^2 $L \times L$ submatrices arranged in a Toeplitz form. An immediate consequence of (5.12) is that at most $(2M - 1)L^2$ of the $(ML)^2$ elements of R_{xx} are distinct. Furthermore, since R_{xx} is Hermitian, at most $(M - 1)L^2 + (L + 1)L/2$ elements of R_{xx} need be computed.

In obtaining (5.12), no special array geometry is assumed. It is not difficult to see from (5.11) that in case the array geometry is such that $p_{j-q_jL} - p_{i-q_iL}$ is constant for all i,j such that

$(j - q_j L) - (i - q_i L)$ is constant, then each of the α_ℓ in (5.12) is a Toeplitz matrix. In this case, at most $(2M - 1)(2L - 1)$ elements of R_{xx} are distinct. Since R_{xx} is Hermitian, only $(M - 1)(2L - 1) + L$ elements of R_{xx} need be computed.

Finally, in case $L = 1$, (5.12) implies that R_{xx} is a Toeplitz matrix having at most $2M - 1$ distinct elements. Since R_{xx} is Hermitian, only M elements of R_{xx} need be computed. The case $L = 1$ reduces the filtering problem to the filtering of wide-sense stationary processes with an FIR (for Finite Impulse Response; really, finite duration unit pulse response) filter, the Toeplitz nature of which has been exploited in [63].

B. Toeplitz R_{xx}

In case R_{xx} is a Toeplitz matrix, several efficient algorithms are available for either the solution of (5.1) or the computation of R_{xx}^{-1}. Levinson [64] was apparently the first to develop an efficient algorithm for the solution of (5.1) in case R_{xx} is a symmetric Toeplitz matrix. Siddiqui [65] presented a simplified solution for R_{xx}^{-1} for the more specialized case that R_{xx} is a covariance matrix for a stable wide-sense stationary scalar discrete time autoregressive process of order k. Trench [66] obtained an algorithm for finding R_{xx}^{-1} requiring only that R_{xx} be Toeplitz and strongly nonsingular. Zohar [67] presented a much simplified derivation for the result of Trench [66]. Preis [68] explicitly presented the algorithm of Trench for the case that R_{xx} is symmetric and discussed its occurrence in antenna problems. A Fortran routine based on [68] is presented in [63].

Zohar [69] makes use of the algorithm of Trench to solve a set of Toeplitz linear equations. Markel and Gray [70] obtain a similar result

from a different viewpoint. Farden [71] makes use of the techniques used by Zohar [69] to derive a more efficient algorithm in case R_{xx} is Hermitian Toeplitz and P is a "Hermitian vector." A more general version of this latter result is presented below, which provides efficient algorithms for the solution to (5.1) in case R_{xx} is a Toeplitz matrix.

Since the techniques used in this section are inherently related to those used by Zohar [69], an attempt will be made to follow the same notational conventions. Greek letters are used for scalars, capital letters for square matrices, and lower-case letters for column matrices. Subscripts used on matrices will denote the number of elements in one column of the matrix.

With a slight breach of previous notation, define $R_p = R_{xx}$, $w_p = w$, and $d_p = P$ in (5.1). The algorithms developed here make use of Phase 1 of the Trench algorithm [67] which requires that R_p be strongly nonsingular, i.e., that all principal minors of R_p be nonzero. Consequently, it will be assumed that (5.1) has been normalized so that R_p has ones along its main diagonal. It is noted that any nonsingular covariance matrix of interest in the present work is both Hermitian and strongly nonsingular (positive definite implies strongly nonsingular); however, since the results are of more general interest, it will be assumed for now only that R_p is Toeplitz and strongly nonsingular.

Consider the system of equations $R_p w_p = d_p$, where R_p is a $p \times p$ Toeplitz matrix normalized such that $(R_p)_{i,i} = 1$, for $i = 1,2,\ldots,p$. Define the sequences $\{\eta_\ell\}$ and $\{\gamma_\ell\}$ such that $d_p' = (\eta_{\frac{p+1}{2}},\ldots,\eta_2,\gamma_1,\ldots,\gamma_{\frac{p+1}{2}})$ for p odd and $d_p' = (\eta_{\frac{p}{2}},\ldots,\eta_1,\gamma_1,\ldots,\gamma_{\frac{p}{2}})$ for p even. For p even or odd, $d_{i+2}' = (\eta_{[\frac{i+3}{2}}, d_i', \gamma_{[\frac{i+3}{2}]})$ for $i = 0,1,2,\ldots,p-2$,

where $[\cdot]$ denotes the largest integer part, $d_o = \phi$, and $d_1 = \gamma_1$.

The Toeplitz nature of R_p enables one to write $(0 \le i \le p-2)$

$$R_{i+2} = \begin{bmatrix} 1 & a'_{i+1} \\ b_{i+1} & R_{i+1} \end{bmatrix} = \begin{bmatrix} R_{i+1} & \hat{a}_{i+1} \\ \hat{b}'_{i+1} & 1 \end{bmatrix} , \tag{5.13}$$

where the $\hat{}$ denotes the reversed ordering of the elements, e.g., $\hat{b}'_{i+1} = (\beta_{i+1}, \ldots, \beta_1)$. Clearly, (5.13) may be rewritten as

$$R_{i+2} = \begin{bmatrix} 1 & a'_i & \\ & & \hat{a}_{i+1} \\ b_i & R_i & \\ & \hat{b}'_{i+1} & 1 \end{bmatrix} . \tag{5.14}$$

Defining $R_{i+2} w_{i+2} = d_{i+2}$ $(1 \le i \le p-2)$, it follows that

$$R_{i+2} \left\{ w_{i+2} - \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} \right\} = \begin{bmatrix} \theta_i \\ 0_i \\ \phi_i \end{bmatrix} , \tag{5.15}$$

where $\theta_i = \eta_{[\frac{i+3}{2}]} - a'_i w_i$, $\phi_i = \gamma_{[\frac{i+3}{2}]} - \hat{b}'_i w_i$, and 0_i is an $i \times 1$ column matrix of zeros. Defining $B_{i+2} = R_{i+2}^{-1}$, (5.15) yields

$$w_{i+2} = \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} + B_{i+2} \begin{bmatrix} \theta_i \\ 0_i \\ \phi_i \end{bmatrix} . \tag{5.16}$$

From [67], B_{i+2} may be expressed in the form

$$B_{i+2} = \lambda_{i+1}^{-1} \begin{bmatrix} 1 & e'_{i+1} \\ g_{i+1} & M_{i+1} \end{bmatrix} = \lambda_{i+1}^{-1} \begin{bmatrix} N_{i+1} & \hat{e}_{i+1} \\ \hat{g}'_{i+1} & 1 \end{bmatrix} . \tag{5.17}$$

It is not difficult to see from (5.17) that B_{i+2} may be expressed as

$$B_{i+2} = \lambda_{i+1}^{-1} \begin{bmatrix} 1 & (e_{i+1})_1 & \cdots & (e_{i+1})_i & \\ (g_{i+1})_1 & & & & \hat{e}_{i+1} \\ \vdots & & Q_i & & \\ (g_{i+1})_i & & & & \\ & \hat{g}'_{i+1} & & & 1 \end{bmatrix} . \quad (5.18)$$

Substituting (5.18) into (5.16) yields

$$w_{i+2} = \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} + \lambda_{i+1}^{-1} \theta_i \left\{ \begin{bmatrix} 1 \\ g_{i+1} \end{bmatrix} + \phi_i \theta_i^{-1} \begin{bmatrix} \hat{e}_{i+1} \\ 1 \end{bmatrix} \right\} . \quad (5.19)$$

In order to make use of this result, Phase 1 of the Trench algorithm [67] can be applied:

Initial values: $e_1 = -a_1$, $g_1 = -b_1$, $\lambda_1 = 1 - a_1 b_1$

Recursion of λ, g, e $(1 \le i \le p-2)$:

$$\delta_i = -(a_{i+1})_{i+1} - e_i' \hat{a}_i, \quad \omega_i = -(b_{i+1})_{i+1} - b_i' \hat{g}_i ,$$

$$e_{i+1} = \begin{bmatrix} e_i + \delta_i \lambda_i^{-1} \hat{g}_i \\ \delta_i \lambda_i^{-1} \end{bmatrix}, \quad \hat{g}_{i+1} = \begin{bmatrix} \omega_i \lambda_i^{-1} \\ \hat{g}_i + \omega_i \lambda_i^{-1} e_i \end{bmatrix} ,$$

$$\lambda_{i+1} = \lambda_i - \lambda_i^{-1} \delta_i \omega_i .$$

Finally, Phase 1 of the Trench algorithm and (5.19) may be combined by noting that

$$w_1 = \gamma_1 \quad (5.20)$$

and

$$w_2 = (1 - a_1 b_1)^{-1} \begin{bmatrix} \eta_1 - a_1 \gamma_1 \\ \gamma_1 - b_1 \eta_1 \end{bmatrix} . \quad (5.21)$$

Note that efficient use of the above result requires that δ_i, ω_i, e_{i+1}, g_{i+1}, and λ_{i+1} be computed for all $i = 1,2,\ldots,p-2$; whereas, w_{i+2} given by (5.19), $\theta_i = \eta_{[\frac{i+3}{2}]} - a_i'w_i$, and $\phi_i = \gamma_{[\frac{i+3}{2}]} - \hat{b}_i'w_i$ need only be computed for $i = 1,3,5,\ldots,p-2$ if p is odd and for $i = 2,4,\ldots,p-2$ if p is even. Several important specializations of the above result in case R_p is Hermitian.

If R_p is a Hermitian Toeplitz matrix then $b_{i+1} = \bar{a}_{i+1}$, $g_{i+1} = \bar{e}_{i+1}$, and $\omega_i = \bar{\delta}_i$ for all $i = 0,1,\ldots,p-2$. Consequently, the above result simplifies. The simplification is summarized below.

PROBLEM FORMULATION: $R_p w_p = d_p$, $(0 \le i \le p-2)$

$$R_{i+2} = \begin{bmatrix} 1 & a_{i+1}' \\ \bar{a}_{i+1} & R_{i+1} \end{bmatrix},$$

$$d_{i+2}' = (\eta_{[\frac{i+3}{2}]}, d_i', \gamma_{[\frac{i+3}{2}]}), \quad w_p = ?$$

Initial values: $e_1 = -a_1$, $\lambda_1 = 1 - |a_1|^2$,

$$w_1 = \gamma_1, \quad w_2 = \lambda_1^{-1}\begin{bmatrix} \eta_1 - a_1\gamma_1 \\ \gamma_1 - \bar{a}_1\eta_1 \end{bmatrix}.$$

Recursive relations: Compute δ_i, e_{i+1}, and λ_{i+1} for $i = 1,2,\ldots,p-2$. Compute θ_i, ϕ_i, and w_{i+2} for $i = 1,3,5,\ldots,p-2$ if p is odd and for $i = 2,4,6,\ldots,p-2$ if p is even.

$$\delta_i = -(a_{i+1})_{i+1} - e_i'\hat{a}_i,$$

$$e_{i+1} = \begin{bmatrix} e_i + \delta_i\lambda_i^{-1}\overleftarrow{\bar{e}_i} \\ \delta_i\lambda_i^{-1} \end{bmatrix},$$

$$\lambda_{i+1} = \lambda_i - |\delta_i|^2 \lambda_i^{-1} \quad ,$$

$$\theta_i = \eta_{[\frac{i+3}{2}]} - a_i' w_i \quad ,$$

$$\phi_i = \gamma_{[\frac{i+3}{2}]} - \hat{w}_i' \bar{a}_i \quad ,$$

$$w_{i+2} = \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} + \lambda_{i+1}^{-1} \theta_i \left\{ \begin{bmatrix} 1 \\ -\bar{e}_{i+1} \end{bmatrix} + \phi_i \theta_i^{-1} \begin{bmatrix} \hat{e}_{i+1} \\ 1 \end{bmatrix} \right\} \quad .$$

The above results offer no apparent computational advantage (or disadvantage) over the results of Zohar [69]. However, the following results do offer significant computational savings over the results of Zohar [69].

Suppose R_p is a Hermitian Toeplitz matrix and the elements of d_p satisfy a Hermitian symmetry property, i.e., $\hat{d}_p = \bar{d}_p$. Then $d_{i+2}' = (\eta_{[\frac{i+3}{2}]}, d_i', \bar{\eta}_{[\frac{i+3}{2}]})$ $(0 \le i \le p-2)$, i.e., $\gamma_i = \bar{\eta}_i$. Consequently, $\hat{w}_{i+2} = \bar{w}_{i+2}$ and $\phi_i = \bar{\theta}_i$. Hence, only $[\frac{i+3}{2}]$ elements of w_{i+2} need to be computed using the recursive relationship given above, the remaining elements being obtained from the relationship $\hat{w}_{i+2} = \bar{w}_{i+2}$. Making use of these facts, the above algorithm for R_p Hermitian and $\hat{d}_p = \bar{d}_p$ requires approximately $1.5p^2$ additions and $1.5p^2$ multiplications to compute the desired solution, w_p. This compares with $2p^2$ for the case that R_p is Hermitian and d_p arbitrary.

In case R_p, d_p (and hence w_p) are real, R_p is symmetric, and $\hat{d}_p = d_p$, an even further reduction in computational requirements results. For this case the recursion for w_i becomes

$$w_{i+2} = \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} + \lambda_{i+1}^{-1} \theta_i \left\{ \begin{bmatrix} 1 \\ e_{i+1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{i+1} \\ 1 \end{bmatrix} \right\} \quad , \qquad (5.22)$$

and the computation of $a_i' w_i$ in the expression for θ_i may be computed as

$$a_i' w_i = \sum_{\ell=1}^{i/2} (w_i)_\ell ((a_i)_\ell + (a_i)_{i+1-\ell}) \tag{5.23}$$

for i even and

$$a_i' w_i = \sum_{\ell=1}^{\frac{i-1}{2}} (w_i)_\ell ((a_i)_\ell + (a_i)_{i+1-\ell}) + (w_i)_{\frac{i+1}{2}} (a_i)_{\frac{i+1}{2}} \tag{5.24}$$

for i odd. Making use of these expressions, this specialized algorithm requires approximately $1.5p^2$ additions and $1.25p^2$ multiplications. A slightly different form of (5.22) can be easily obtained as

$$w_{i+2} = \begin{bmatrix} 0 \\ w_i \\ 0 \end{bmatrix} + \frac{\theta_i}{\lambda_i - \delta_i} \begin{bmatrix} 1 \\ e_i + \hat{e}_i \\ 1 \end{bmatrix} . \tag{5.25}$$

This final expression (5.25) is slightly more efficient than (5.22). A Fortran routine for this specialized algorithm making use of (5.23) - (5.25) is presented in [72].

C. R_{xx} Having M^2 L x L Submatrices Arranged in Toeplitz Form

In this section, the solution of (5.1) for the general situation with R_{xx} expressed by (5.12) is given. An important by-product of this development will be an efficient algorithm for computing R_{xx}^{-1}. Again, all covariance matrices of interest are Hermitian; however, since the results seem to be of more widespread interest, the Hermitian restriction will not be made. As assumed in (5.12), let $p = ML$. Throughout this section, capital letters will be used to denote square matrices and lower case letters will be used to denote "vectors" with LxL matrix "elements." Subscripts on these quantities will be used to

denote the number of elements in each column of the matrix. Greek letters will be used to denote L x L matrices. These conventions will be violated with $R_p = R_{xx}$, $w_p = w$, and $d_p = P$ as in the previous section. The ℓ x ℓ identity matrix will be denoted by I_ℓ. An a x b matrix with all zero elements will be denoted by $O_{a,b}$.

From (5.12), $R_{xx} = R_p = R_{ML}$ can be expressed as

$$R_{ML} = \begin{bmatrix} \alpha_o & a''_{(M-1)L} \\ b_{(M-1)L} & R_{(M-1)L} \end{bmatrix} , \tag{5.26}$$

where $a''_{(M-1)L} = (\alpha_1, \alpha_2, \ldots, \alpha_{M-1})$, $b''_{(M-1)L} = (\alpha_{-1}, \alpha_{-2}, \ldots, \alpha_{1-M})$, and the " is used to denote an obvious matrix operation similar to matrix transpose. Define $B_{mL} = R_{mL}^{-1}$ for m = 1,2,...,M. Express $B_{(m+1)L}$ as

$$B_{(m+1)L} = \begin{bmatrix} \beta_m & \beta_m e'_{mL} \\ f_{mL}\beta_m & A_{mL} \end{bmatrix} . \tag{5.27}$$

Then $I_{(m+1)L} = B_{(m+1)L} R_{(m+1)L}$ yields

$$\begin{bmatrix} I_L & O_{L,mL} \\ O_{mL,L} & I_{mL} \end{bmatrix} = \begin{bmatrix} \beta_m \alpha_o + \beta_m e'_{mL} b_{mL} & \beta_m a''_{mL} + \beta_m e'_{mL} R_{mL} \\ f_{mL}\beta_m \alpha_o + A_{mL} b_{mL} & f_{mL}\beta_m a''_{mL} + A_{mL} R_{mL} \end{bmatrix} . \tag{5.28}$$

Solving (5.28) to obtain $A_{mL} = B_{mL} + f_{mL}\beta_m e'_{mL}$, and substituting into (5.27) yields

$$B_{(m+1)L} = \begin{bmatrix} \beta_m & \beta_m e'_{mL} \\ f_{mL}\beta_m & B_{mL} + f_{mL}\beta_m e'_{mL} \end{bmatrix} . \tag{5.29}$$

Define the "block exchange matrix" E^*_{mL} by

$$E^*_{mL} = \begin{bmatrix} & & & I_L \\ & 0 & \iddots & I_L \\ & I_L & & 0 \\ I_L & & & \end{bmatrix} . \tag{5.30}$$

Note that $E^*_{mL} E^*_{mL} = I_{mL}$ and that the block Toeplitz nature of R_{ML} given by (5.26) enables one to easily verify that $E^*_{mL} R''_{mL} E^*_{mL} = R_{mL}$. Then, by defining $B^*_{mL} = (R''_{mL})^{-1}$, it follows that $I_{mL} = R''_{mL} B^*_{mL}$ $= R_{mL}(E^*_{mL} B^*_{mL} E^*_{mL})$. Hence, $B_{mL} = E^*_{mL} B^*_{mL} E^*_{mL}$ and $B^*_{mL} = E^*_{mL} B_{mL} E^*_{mL}$. It will be convenient to use the symbol $\hat{}$ to denote a reversal in the ordering of the L x L "elements," e.g., $\hat{a}''_{mL} = (\alpha_m, \alpha_{m-1}, \ldots, \alpha_1) = (E^*_{mL} a_{mL})''$ $= a''_{mL} E^*_{mL}$. Expressing $B^*_{(m+1)L}$ as

$$B^*_{(m+1)L} = \begin{bmatrix} \gamma_m & \gamma_m \hat{g}''_{mL} \\ \hat{h}_{mL} \gamma_m & D_{mL} \end{bmatrix} , \tag{5.31}$$

and noting that $I_{mL} = B^*_{mL} R''_{mL}$ yields

$$\begin{bmatrix} I_L & 0_{L,mL} \\ 0_{mL,L} & I_{mL} \end{bmatrix} = \begin{bmatrix} \gamma_m \alpha_o + \gamma_m \hat{g}''_{mL} a_{mL} & \gamma_m b''_{mL} + \gamma_m \hat{g}''_{mL} R''_{mL} \\ \hat{\hat{h}}_{mL} \gamma_m \alpha_o + D_{mL} a_{mL} & \hat{\hat{h}}_{mL} \gamma_m b''_{mL} + D_{mL} R''_{mL} \end{bmatrix} . \tag{5.32}$$

Solving (5.32) for $D_{mL} = B^*_{mL} + \hat{\hat{h}}_{mL} \gamma_m \hat{g}''_{mL}$ and substituting into (5.31) yields

$$B^*_{(m+1)L} = \begin{bmatrix} \gamma_m & \gamma_m \hat{g}''_{mL} \\ \hat{h}_{mL} \gamma_m & B^*_{mL} + \hat{\hat{h}}_{mL} \gamma_m \hat{g}''_{mL} \end{bmatrix} . \tag{5.33}$$

Pre- and post-multiplying (5.33) by $E^*_{(m+1)L}$ and combining the result with (5.29), one obtains (since $E^*_{mL} B^*_{mL} E^*_{mL} = B_{mL}$)

$$B_{(m+1)L} = \begin{bmatrix} \beta_m & \beta_m e'_{mL} \\ f_{mL}\beta_m & B_{mL} + f_{mL}\beta_m e'_{mL} \end{bmatrix}$$

$$= \begin{bmatrix} B_{mL} + h_{mL}\gamma_m g''_{mL} & h_{mL}\gamma_m \\ \gamma_m g''_{mL} & \gamma_m \end{bmatrix} . \qquad (5.34)$$

Using techniques analogous to those used by Zohar [67], it can be shown that all of the elements of $B_{(m+1)L}$ can be generated from β_m, e_{mL}, f_{mL}, γ_m, g_{mL}, and h_{mL}. Denote the ij^{th} L x L "block" of a matrix, say A_{mL}, by $(A_{mL})_{i,j}$. From the first equality in (5.34), it follows directly that

$$(B_{(m+1)L})_{1,1} = \beta_m, \qquad (5.35)$$

$$(B_{(m+1)L})_{i+1,1} = (f_{mL})_{i,1}\beta_m, \quad 1 \le i \le m, \qquad (5.36)$$

$$(B_{(m+1)L})_{1,i+1} = \beta_m (e'_{mL})_{1,i}, \quad 1 \le i \le m, \qquad (5.37)$$

and

$$(B_{(m+1)L})_{i+1,j+1} = (B_{mL})_{i,j} + (f_{mL}\beta_m e'_{mL})_{i,j}, \quad 1 \le i,j \le m. \qquad (5.38)$$

From the second equality in (5.34), it follows that

$$(B_{(m+1)L})_{i,j} = (B_{mL})_{i,j} + (h_{mL}\gamma_m g''_{mL})_{i,j}, \quad 1 \le i,j \le m . \qquad (5.39)$$

Combining (5.38) and (5.39) to eliminate $(B_{mL})_{i,j}$,

$$(B_{(m+1)L})_{i+1,j+1} = (B_{(m+1)L})_{i,j} + (f_{mL}\beta_m e'_{mL})_{i,j} - (h_{mL}\gamma_m g''_{mL})_{i,j}, \quad 1 \le i,j \le m. \qquad (5.40)$$

Equation (5.40) is a recursive relationship for generating the elements of $B_{(m+1)L}$ starting with the initial conditions given by (5.35) - (5.37). It is important to note that it is the property that $E^*_{mL} B^*_{mL} E^*_{mL} = B_{mL}$ that enabled the derivation of (5.40).

In order for the above result to be of practical use, recursive relationships for β_m, e_{mL}, f_{mL}, γ_m, g_{mL}, and h_{mL} must be developed. Solving the upper right equation of (5.32) yields $\hat{g}''_{mL} = -b''_{mL} B^*_{mL}$ or

$$g''_{mL} = -\hat{b}''_{mL} B_{mL} \quad . \tag{5.41}$$

Solving the lower left equation of (5.32), and substituting $D_{mL} = B^*_{mL} + \hat{h}_{mL}\gamma_m\hat{g}''_{mL}$ yields $\hat{h}_{mL}\gamma_m\alpha_o = -B^*_{mL} a_{mL} - \hat{h}_{mL}\gamma_m\hat{g}''_{mL}a_{mL}$. Solving the upper left equation of (5.32) for $\gamma_m\hat{g}''_{mL}a_{mL}$ and substituting yields $\hat{h}_{mL} = -B^*_{mL}a_{mL}$ or

$$h_{mL} = -B_{mL}\hat{a}_{mL} \quad . \tag{5.42}$$

The upper right equation of (5.28) is easily solved for

$$e'_{mL} = -a''_{mL} B_{mL} \quad . \tag{5.43}$$

Solving the lower left of (5.28) for $f_{mL}\beta_m\alpha_o$ and substituting $A_{mL} = B_{mL} + f_{mL}\beta_m e'_{mL}$ and $\beta_m e'_{mL} b_{mL} = I_L - \beta_m\alpha_o$ yields

$$f_{mL} = -B_{mL} b_{mL} \quad . \tag{5.44}$$

Equations (5.41) - (5.44) can now be used with (5.34) to derive recursive relationships for g''_{mL}, h_{mL}, e'_{mL}, and f_{mL}. From (5.41) and the first equality in (5.34),

$$g''_{(m+1)L} = -(\alpha_{-(m+1)}, \hat{b}''_{mL}) \begin{bmatrix} \beta_m & \beta_m e'_{mL} \\ f_{mL}\beta_m & B_{mL} + f_{mL}\beta_m e'_{mL} \end{bmatrix} , \tag{5.45}$$

or $g''_{(m+1)L} = (0_{L,L}, g''_{mL}) - \varepsilon_m\beta_m(I_L, e'_{mL})$, where

$$\varepsilon_m = \alpha_{-(m+1)} + \hat{b}''_{mL} f_{mL} \quad . \tag{5.46}$$

From (5.42) and the first equality in (5.34),

$$h_{(m+1)L} = \begin{bmatrix} 0_{L,L} \\ \\ h_{mL} \end{bmatrix} - \begin{bmatrix} I_L \\ \\ f_{mL} \end{bmatrix} \beta_m \delta_m \quad , \tag{5.47}$$

where

$$\delta_m = \alpha_{m+1} + e'_{mL}\, \hat{\hat{a}}_{mL} \quad . \tag{5.48}$$

From (5.43) and the second equality in (5.34),

$$e'_{(m+1)L} = (e'_{mL},\ 0_{L,L}) - \eta_m \gamma_m (g''_{mL},\ I_L) \quad , \tag{5.49}$$

where

$$\eta_m = \alpha_{m+1} + a''_{mL}\, h_{mL} \quad . \tag{5.50}$$

Also, from (5.44) and the second equality in (5.34),

$$f_{(m+1)L} = \begin{bmatrix} f_{mL} \\ \\ 0_{L,L} \end{bmatrix} - \begin{bmatrix} h_{mL} \\ \\ I_L \end{bmatrix} \gamma_m \omega_m \quad , \tag{5.51}$$

where

$$\omega_m = \alpha_{-(m+1)} + g''_{mL}\, b_{mL} \quad . \tag{5.52}$$

Finally, equating the four L x L "corner blocks" of (5.34), and using (5.45) - (5.52), β_m can be expressed as

$$\beta_m = \beta_{m-1} + \beta_{m-1} \delta_{m-1} \gamma_m \varepsilon_{m-1} \beta_{m-1} \quad , \tag{5.53}$$

or

$$\beta_m \eta_{m-1} \gamma_{m-1} = \beta_{m-1} \delta_{m-1} \gamma_m \quad . \tag{5.54}$$

Substituting (5.54) into (5.53) and solving for β_m yields

$$\beta_m = \beta_{m-1} (I_L - \eta_{m-1} \gamma_{m-1} \varepsilon_{m-1} \beta_{m-1})^{-1} \quad . \tag{5.55}$$

Furthermore, γ_m can be expressed as

$$\gamma_m = \gamma_{m-1} (I_L + \omega_{m-1} \beta_m \eta_{m-1} \gamma_{m-1}) \quad . \tag{5.56}$$

Now, consider the equation $R_{mL} w_{mL} = d_{mL}$, for $m = 1,2,\ldots,M$. Recall that w_{mL} and d_{mL} are $mL \times 1$ matrices. Define the $L \times 1$

AD-A034 096 COLORADO STATE UNIV FORT COLLINS DEPT OF ELECTRICAL --ETC F/G 12/1
STOCHASTIC APPROXIMATION WITH CORRELATED DATA.(U)
MAY 75 D C FARDEN

UNCLASSIFIED TR-11(ONR) N00014-67-A-0299-0019

NL

2 OF 2
ADA034096

END

DATE FILMED
2 - 77

matrix Δ_m by $(\Delta_m)_i = (d_{mL})_{(m-1)L+i}$ for $i = 1,2,\ldots,L$. Recall that the block Toeplitz nature of R_{mL} enables one to write

$$R_{(m+1)L} = \begin{bmatrix} \alpha_o & a''_{mL} \\ b_{mL} & R_{mL} \end{bmatrix} = \begin{bmatrix} R_{mL} & \hat{a}_{mL} \\ \hat{b}''_{mL} & \alpha_o \end{bmatrix} . \tag{5.57}$$

Using the last expression for $R_{(m+1)L}$ in (5.57) it is easily shown that

$$R_{(m+1)L} \left\{ w_{(m+1)L} - \begin{bmatrix} w_{mL} \\ 0_{L,1} \end{bmatrix} \right\} = \begin{bmatrix} 0_{mL,1} \\ \Delta_{m+1} - \hat{b}''_{mL} w_{mL} \end{bmatrix} . \tag{5.58}$$

Defining $\Gamma_m = \Delta_{m+1} - \hat{b}''_{mL} w_{mL}$ and premultiplying both sides of (5.58) by $B_{(m+1)L}$,

$$w_{(m+1)L} = \begin{bmatrix} w_{mL} \\ 0_{L,1} \end{bmatrix} + B_{(m+1)L} \begin{bmatrix} 0_{mL,1} \\ \Gamma_m \end{bmatrix} . \tag{5.59}$$

Making use of the second equality in (5.34),

$$w_{(m+1)L} = \begin{bmatrix} w_{mL} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} h_{mL} \\ I_L \end{bmatrix} \gamma_m \Gamma_m . \tag{5.60}$$

Once initial conditions are found, the development will be complete. From (5.28) with $m = 1$, it follows that $e'_{1L} = -\alpha_1 \alpha_o^{-1}$, $f_{1L} = -\alpha_o^{-1} \alpha_{-1}$, and $\beta_1 = (\alpha_o + e'_{1L} \alpha_{-1})^{-1}$. From (5.32), with $m = 1$, $g_{1L} = -\alpha_{-1} \alpha_o^{-1}$ and $h_{1L} = -\alpha_o^{-1} \alpha_1$. With $m = 1$, (5.34) yields $\gamma_1 = \alpha_o^{-1} + f_{1L} \beta_1 e'_{1L}$, so that the set of initial conditions is complete. The complete algorithm is summarized below.

PROBLEM FORMULATION: $R_{ML} w_{ML} = d_{ML}$,

$$R_{ML} = \begin{bmatrix} \alpha_o & a''_{(M-1)L} \\ b_{(M-1)L} & R_{(M-1)L} \end{bmatrix},$$

$$a''_{ML} = (\alpha_1, \alpha_2, \ldots, \alpha_m), \quad (1 \leq m \leq M-1),$$

$$b''_{mL} = (\alpha_{-1}, \alpha_{-2}, \ldots, \alpha_{-m}), \quad (1 \leq m \leq M-1),$$

$$w_{ML} = ?$$

Initial values: $e'_{1L} = -\alpha_1 \alpha_o^{-1}$, $f_{1L} = -\alpha_o^{-1} \alpha_{-1}$,

$$g_{1L} = -\alpha_{-1} \alpha_o^{-1}, \quad h_{1L} = -\alpha_o^{-1} \alpha_1, \quad \beta_1 = (\alpha_o + e'_{1L} \alpha_{-1})^{-1},$$

$$\gamma_1 = \alpha_o^{-1} + f_{1L} \beta_1 e'_{1L}, \quad w_{1L} = \alpha_o^{-1} d_{1L},$$

$$w_{2L} = \begin{bmatrix} w_{1L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} h_{1L} \\ I_{1L} \end{bmatrix} \gamma_1 \Gamma_1.$$

Recursive relations: $(1 \leq m \leq M-2)$,

$$\varepsilon_m = \alpha_{-(m+1)} + \hat{b}''_{mL} f_{mL},$$

$$g''_{(m+1)L} = (0_{L,L}, g''_{mL}) - \varepsilon_m \beta_m (I_L, e'_{mL}),$$

$$\delta_m = \alpha_{m+1} + e'_{mL} \hat{a}_{mL},$$

$$h_{(m+1)L} = \begin{bmatrix} 0_{L,L} \\ h_{mL} \end{bmatrix} - \begin{bmatrix} I_L \\ f_{mL} \end{bmatrix} \beta_m \delta_m,$$

$$\eta_m = \alpha_{m+1} + a''_{mL} h_{mL},$$

$$e'_{(m+1)L} = (e'_{mL}, 0_{L,L}) - \eta_m \gamma_m (g''_{mL}, I_L),$$

$$\omega_m = \alpha_{-(m+1)} + g''_{mL}\, b_{mL},$$

$$f_{(m+1)L} = \begin{bmatrix} f_{mL} \\ 0_{L,L} \end{bmatrix} - \begin{bmatrix} h_{mL} \\ I_L \end{bmatrix} \gamma_m \omega_m,$$

$$\beta_{m+1} = \beta_m (I_L - \eta_m \gamma_m \varepsilon_m \beta_m)^{-1},$$

$$\gamma_{m+1} = \gamma_m (I_L + \omega_m \beta_{m+1} \eta_m \gamma_m),$$

$$\Gamma_{m+1} = \Delta_{m+2} - \hat{b}''_{(m+1)L}\, w_{(m+1)L},$$

$$w_{(m+2)L} = \begin{bmatrix} w_{(m+1)L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} h_{(m+1)L} \\ I_L \end{bmatrix} \gamma_{m+1} \Gamma_{m+1}.$$

Note that if it is desired to compute the inverse of R_{ML}, the above algorithm can be used with the expressions for Γ_{m+1} and $w_{(m+2)L}$ deleted. After β_{M-1}, $e_{(M-1)L}$, $f_{(M-1)L}$, $g_{(M-1)L}$, $h_{(M-1)L}$, and γ_{M-1} have been computed, equations (5.35) - (5.37) and (5.40) with $m = M - 1$ can be used to generate B_{ML}.

The above algorithm requires approximately $6M \cdot L^2 + 2ML$ storage locations, which can represent a considerable savings for large M. The algorithm requires approximately $4M^2$ matrix (L x L) multiplications, $4M^2$ matrix (L x L) additions, and M matrix (L x L) inversions. Considering that an L x L matrix multiplication requires approximately L^3 scalar additions and L^3 scalar multiplications, and that standard routines for an L x L matrix inversion require approximately L^3 multiplications and additions, the above algorithm requires approximately $4M^2L^3$ operations compared with approximately $(ML)^3$ for standard algorithms. These remarks, of course, do not include the operations

necessary to actually compute R_{ML}^{-1}, which requires an additional $4M^2L^3$ operations (approximately). It can certainly be concluded that the above algorithm can offer a substantial computational advantage over standard algorithms for large M. Recall that M is the number of taps on a delay line realization of a FIR filter.

As noted previously, all covariance matrices of interest in this work are Hermitian (in fact, most are real and symmetric). Consequently, the simplification of the above algorithm for Hermitian block Toeplitz R_{xx} will now be undertaken.

For the case that R_{ML} is Hermitian, it follows easily that $\bar{\gamma}_m' = \gamma_m$, $\bar{\beta}_m' = \beta_m$, $e'_{mL} = \bar{f}'_{mL}$, and $g''_{mL} = \bar{h}'_{mL}$. Substituting these identities into (5.34) easily yields

$$B_{(m+1)L} = \begin{bmatrix} \beta_m & \beta_m \bar{f}'_{mL} \\ f_{mL}\beta_m & B_{mL} + f_{mL}\beta_m \bar{f}'_{mL} \end{bmatrix}$$

$$= \begin{bmatrix} B_{mL} + h_{mL}\gamma_m \bar{h}'_{mL} & h_{mL}\gamma_m \\ \gamma_m \bar{h}'_{mL} & \gamma_m \end{bmatrix} . \qquad (5.61)$$

Making the required substitutions into the general algorithm, the simplified algorithm for the case at hand is easily obtained. A summary of the algorithm for Hermitian R_{xx} is presented below.

PROBLEM FORMULATION: $R_{ML} w_{ML} = d_{ML}$,

$$R_{ML} = \begin{bmatrix} \alpha_o & \bar{b}'_{(M-1)L} \\ b_{(M-1)L} & R_{(M-1)L} \end{bmatrix},$$

$$b'_{mL} = (\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_m), \quad (1 \leq m \leq M-1),$$

$$w_{ML} = ?$$

Initial values: $f_{1L} = -\alpha_o^{-1}\bar{\alpha}'_1$, $h_{1L} = -\alpha_o^{-1}\alpha_1$,

$$\beta_1 = (\alpha_o + \bar{f}'_{1L}\bar{\alpha}'_1)^{-1},\ \gamma_1 = \alpha_o^{-1} + f_{1L}\beta_1\bar{f}'_{1L},\ w_{1L} = \alpha_o^{-1}d_{1L} ,$$

$$w_{2L} = \begin{bmatrix} w_{1L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} h_{1L} \\ I_1 \end{bmatrix} \gamma_1\Gamma_1 .$$

Recursive relations: $(1 \le m \le M-2)$,

$$\delta_m = \alpha_{m+1} + \bar{f}'_{mL}(\hat{\bar{b}}'_{mL})'' ,$$

$$h_{(m+1)L} = \begin{bmatrix} 0_{L,L} \\ h_{mL} \end{bmatrix} - \begin{bmatrix} I_L \\ f_{mL} \end{bmatrix} \beta_m\delta_m ,$$

$$\omega_m = \bar{\alpha}'_{m+1} + \bar{h}'_{mL}\, b_{mL} ,$$

$$f_{(m+1)L} = \begin{bmatrix} f_{mL} \\ 0_{L,L} \end{bmatrix} - \begin{bmatrix} h_{mL} \\ I_L \end{bmatrix} \gamma_m\omega_m ,$$

$$\beta_{m+1} = (I_L - \beta_m\delta_m\gamma_m\omega_m)^{-1}\beta_m ,$$

$$\gamma_{m+1} = \gamma_m(I_L + \omega_m\beta_{m+1}\bar{\omega}'_m\gamma_m) ,$$

$$\Gamma_{m+1} = \Delta_{m+2} - \hat{b}''_{(m+1)L}w_{(m+1)L} ,$$

$$w_{(m+2)L} = \begin{bmatrix} w_{(m+1)L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} h_{(m+1)L} \\ I_L \end{bmatrix} \gamma_{m+1}\Gamma_{m+1} .$$

In case the inverse of R_{ML} is desired, the above algorithm can be used with the expressions for Γ_{m+1} and $w_{(m+2)L}$ deleted. After β_{M-1}, $f_{(M-1)L}$, $h_{(M-1)L}$, and γ_{M-1} have been computed, equations (5.34)-(5.36) and (5.39) with $m = M-1$, $e_{mL} = \bar{f}_{mL}$, and $g''_{mL} = \bar{h}'_{mL}$ can be

used to generate B_{ML}. Efficient use of this procedure of course demands that the Hermitian property of B_{ML} be used. The above algorithm requires approximately half the computations required by the previous algorithm.

Another interesting case for block Toeplitz matrices arises when R_{ML} is persymmetric, i.e., symmetric about the main cross diagonal. Define the exchange matrix E_m by

$$E_m = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} . \tag{5.62}$$

A persymmetric matrix A satisfies $E_m A' E_m = A_m$. It is easily shown that a block Toeplitz matrix as given by (5.12) is persymmetric if and only if α_ℓ is persymmetric for all $\ell = 0,\pm1,\ldots,\pm(M-1)$. Hence, in case α_ℓ is Toeplitz for all $\ell = 0,\pm1,\ldots,\pm(M-1)$, R_{ML} is persymmetric. Also, it is easily shown that the inverse of a persymmetric matrix is persymmetric. With R_{ML} persymmetric and given by (5.26), $B_{(m+1)L}$ can still be expressed as in (5.29). Computing $E_{(m+1)L} B'_{(m+1)L} E_{(m+1)L}$ from (5.29), $B_{(m+1)L}$ may be expressed as

$$\begin{aligned} B_{(m+1)L} &= \begin{bmatrix} \beta_m & \beta_m e'_{mL} \\ f_{mL}\beta_m & B_{mL} + f_{mL}\beta_m e'_{mL} \end{bmatrix} \\ &= \begin{bmatrix} B_{mL} + E_{mL} e_{mL} \beta'_m f'_{mL} E_{mL} & E_{mL} e_{mL} \beta'_m E_L \\ E_L \beta'_m f'_{mL} E_{mL} & E_L \beta'_m E_L \end{bmatrix} . \end{aligned} \tag{5.63}$$

In full analogy to the development of (5.35) - (5.40),

$$(B_{(m+1)L})_{1,1} = \beta_m , \tag{5.64}$$

$$(B_{(m+1)L})_{i+1,1} = (f_{mL})_{i,1}\beta_m, \quad 1 \le i \le m , \tag{5.65}$$

$$(B_{(m+1)L})_{1,i+1} = \beta_m (e'_{mL})_{1,i} \ , \ 1 \le i \le m \ , \tag{5.66}$$

$$(B_{(m+1)L})_{i+1,j+1} = (B_{(m+1)L})_{i,j} + (f_{mL}\beta_m e'_{mL})_{i,j} - (E_{mL} e_{mL} \beta'_{mL} f'_{mL} E_{mL})_{i,j} \ , \ 1 \le i,j \le m \ . \tag{5.67}$$

Thus, in this case, all of the elements of B_{ML} can be generated from β_{M-1}, $e_{(M-1)L}$, and $f_{(M-1)L}$. Equations (5.43) and (5.44), which are still valid for this case, can be used with (5.63) to obtain recursive relationships for β_m, e_{mL}, and f_{mL}. As in the previous section, it will be convenient to use the symbol $\hat{}$ to denote a reversal in the vertical ordering of the rows of a matrix, e.g., $\hat{a}_{mL} = E_{mL} a_{mL}$, and $a'_{mL} E_{mL} = (E_{mL} a_{mL})' = \hat{a}'_{mL}$. From (5.43) and the second equality in (5.63),

$$e'_{(m+1)L} = (e'_{mL}, \ 0_{L,L}) - \zeta_m \beta'_m (\hat{f}'_{mL}, \ E_L), \tag{5.68}$$

where

$$\zeta_m = \alpha_{m+1} E_L + a''_{mL} \hat{e}_{mL} \ . \tag{5.69}$$

From (5.44) and the second equality in (5.63),

$$f_{(m+1)L} = \begin{bmatrix} f_{mL} \\ 0_{L,L} \end{bmatrix} - \begin{bmatrix} \hat{e}_{mL} \\ E_L \end{bmatrix} \beta'_m \phi_m \ , \tag{5.70}$$

where

$$\phi_m = E_L \alpha_{-(m+1)} + f'_{mL} \hat{b}_{mL} \ . \tag{5.71}$$

Solving the upper left L x L block of (5.63), and substituting (5.68) - (5.71), one obtains

$$\beta_m = \beta_{m-1} + \beta_{m-1} \zeta_{m-1} \beta'_m \phi'_{m-1} \beta_{m-1} \ . \tag{5.72}$$

Similarly, from the lower left L x L block of (5.63), one obtains

$$\beta'_{m-1} \phi_{m-1} \beta_m = \beta'_m \phi'_{m-1} \beta_{m-1} \ . \tag{5.73}$$

Substituting (5.72) into (5.71),

$$\beta_m = (I_L - \beta_{m-1}\zeta_{m-1}\beta'_{m-1}\phi_{m-1})^{-1}\beta_{m-1} \ . \tag{5.74}$$

Finally, in order to solve the equation $R_{ML}w_{ML} = d_{ML}$, note that (5.57) through (5.59) are still valid. Recall that $(\Delta_m)_i = (d_{mL})_{(m-1)L+i}$ for $i = 1,2,\ldots,L$, and $\Gamma_m = \Delta_{m+1} - \hat{b}''_{mL}w_{mL}$. Making use of the second equality of (5.63) and (5.59), one easily obtains

$$w_{(m+1)L} = \begin{bmatrix} w_{mL} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{mL} \\ E_L \end{bmatrix} \beta'_m E_L \Gamma_m \ . \tag{5.75}$$

The following is a summary of the algorithm for the solution of $R_{ML}w_{ML} = d_{ML}$, for the special case that R_{ML} is a persymmetric block Toeplitz matrix.

PROBLEM FORMULATION: $R_{ML}w_{ML} = d_{ML}$,

$$R_{ML} = \begin{bmatrix} \alpha_o & a''_{(M-1)L} \\ b_{(M-1)L} & R_{(M-1)L} \end{bmatrix} \ , \ E_{ML}R'_{ML}E_{ML} = R_{ML},$$

$$a''_{mL} = (\alpha_1,\alpha_2,\ldots,\alpha_m), \ 1 \le m \le M-1 \ ,$$

$$b''_{mL} = (\alpha_{-1},\alpha_{-2},\ldots,\alpha_{-m}) \ , \ 1 \le m \le M-1 \ ,$$

$$w_{ML} = ?$$

Initial values: $e'_{1L} = -\alpha_1\alpha_o^{-1}$, $\beta_1 = (\alpha_o + e'_{1L}\alpha_{-1})^{-1}$,

$$f_{1L} = -\alpha_o^{-1}\alpha_{-1}, \ w_{1L} = \alpha_o^{-1}d_{1L}, \ \Gamma_1 = \Delta_2 - \alpha_{-1}w_{1L} \ ,$$

$$w_{2L} = \begin{bmatrix} w_{1L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{1L} \\ E_L \end{bmatrix} \beta'_1 E_L \Gamma_1 \ .$$

Recursive relations: $(1 \le m \le M-2)$

$$\zeta_m = \alpha_{m+1} E_L + a''_{mL} \hat{e}_{mL} ,$$

$$e'_{(m+1)L} = (e'_{mL}, 0_{L,L}) - \zeta_m \beta'_m (\hat{f}'_{mL}, E_L),$$

$$\phi_m = E_L \alpha_{-(m+1)} + f'_{mL} \hat{b}_{mL} ,$$

$$f_{(m+1)L} = \begin{bmatrix} f_{mL} \\ 0_{L,1} \end{bmatrix} - \begin{bmatrix} \hat{e}_{mL} \\ E_L \end{bmatrix} \beta'_m \phi_m ,$$

$$\beta_{m+1} = (I_L - \beta_m \zeta_m \beta'_m \phi_m)^{-1} \beta_m ,$$

$$\Gamma_{m+1} = \Delta_{m+2} - \hat{\hat{b}}''_{(m+1)L} w_{(m+1)L} ,$$

$$w_{(m+2)L} = \begin{bmatrix} w_{(m+1)L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{(m+1)L} \\ E_L \end{bmatrix} \beta'_{m+1} E_L \Gamma_{m+1} .$$

If the inverse of R_{ML} is desired, the above algorithm can be used with the expressions for w_{mL} and Γ_m deleted. After β_{M-1}, $e_{(M-1)L}$, and $f_{(M-1)L}$ have been computed, equation (5.67) can be used with $m = M-1$ to generate $R_{ML}^{-1} = B_{ML}$, using (5.64) - (5.66) as initial conditions. The computational requirements of the above algorithm are virtually identical with those of the previous algorithm for Hermitian block Toeplitz R_{ML}. The reason for this similarity is clear: a Hermitian matrix has conjugate symmetry about the main diagonal, and a persymmetric matrix has symmetry about the main cross diagonal.

Finally, in case R_{ML} is a Hermitian, persymmetric block Toeplitz matrix, the computational requirements of the above algorithm can be approximately halved. The simplification is easily obtained by

substituting $a''_{mL} = \bar{b}'_{mL}$, $f_{mL} = \bar{e}_{mL}$, and $\phi_m = \bar{\zeta}'_m$ into the above algorithm. The resulting algorithm is summarized below.

PROBLEM FORMULATION: $R_{ML} w_{ML} = d_{ML}$,

$$R_{ML} = \begin{bmatrix} \alpha_o & \bar{b}'_{(M-1)L} \\ b_{(M-1)L} & R_{(M-1)L} \end{bmatrix} , \quad E_{ML}, R'_{ML} E_{ML} = R_{ML} ,$$

$$b'_{mL} = (\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_m) , \quad (1 \leq m \leq M-1) ,$$

$$w_{ML} = ?$$

Initial values: $e'_{1L} = -\alpha_1 \alpha_o^{-1}$, $\beta_1 = (\alpha_o + e'_{1L} \bar{\alpha}'_1)^{-1}$,

$$w_{1L} = \alpha_o^{-1} d_{1L} , \quad \Gamma_1 = \Delta_2 - \bar{\alpha}'_1 w_{1L} ,$$

$$w_{2L} = \begin{bmatrix} w_{1L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{1L} \\ E_L \end{bmatrix} \beta'_1 E_L \Gamma_1 .$$

Recursive relations: $(1 \leq m \leq M-2)$

$$\zeta_m = \alpha_{m+1} E_L + \bar{b}'_{mL} \hat{e}_{mL} ,$$

$$e'_{(m+1)L} = (e'_{mL}, 0_{L,L}) - \zeta_m \beta'_m (\hat{\bar{e}}'_{mL}, E_L) ,$$

$$\beta_{m+1} = (I_L - \beta_m \zeta_m \beta'_m \bar{\zeta}'_m)^{-1} \beta_m ,$$

$$\Gamma_{m+1} = \Delta_{m+2} - \hat{\bar{b}}''_{(m+1)L} w_{(m+1)L} ,$$

$$w_{(m+2)L} = \begin{bmatrix} w_{(m+1)L} \\ 0_{L,1} \end{bmatrix} + \begin{bmatrix} \hat{e}_{(m+1)L} \\ E_L \end{bmatrix} \beta'_{m+1} E_L \Gamma_{m+1} .$$

D. Discussion

In this chapter, special forms of data correlation matrices which can arise in discrete-time stochastic signal processing applications have been considered. Computationally efficient algorithms have been presented for the solution of $R_{xx}w = P$ as well as for obtaining R_{xx}^{-1} for these special forms. These results, the development of which relies heavily on generalizations of the results of Zohar [67], [69], are of interest in their own right. The application of these results to filter design problems for which R_{xx} and P are known is straightforward.

In case R_{xx} and/or P are unknown and a fixed (nonadaptive) filter is desired, the results of this chapter are still applicable. The obvious approach is to use estimates of R_{xx} and/or P. An alternative approach to the design of FIR filters, with the signal and noise structure of Example 1 in Section IV-B, involves the utilization of "approximate" spectral density functions, and has been treated by Farden and Scharf [63]. Extensions of the concepts treated in [63] to the design of multidimensional FIR filters can be accomplished with the aid of Section V-C.

The results of this chapter are also useful for performing simulations of adaptive structures. In performing such simulations, it is advantageous to generate data having known covariance functions in order to evaluate the performance of the adaptive processor by computing, e.g., $\|W_k - w_o\|$ with W_k obtained from some form of (3.1) and w_o the solution to $R_{xx}w = P$. The results of this chapter are ideally suited for the computation of w_o for several cases of practical interest, as discussed in Section V-A.

Finally, the results of Section V-A suggest modifications of the algorithms discussed in Chapter II which should result in an increased convergence rate without a severe increase in storage requirements. Consider the algorithm

$$W_{k+1} = W_k + \frac{1}{k}(P - X_k Y_k) \quad , \tag{5.76}$$

where $Y_k = W_k' X_k$, and X_k is a p-element data vector as in Example 1 of Section IV-B. Under conditions established in Chapter IV, $W_k \overset{a.s.}{\to} w_o$ as $k \to \infty$, and hence, $Y_k \overset{a.s.}{\to} \hat{s}_k$ as in Example 1. In order to implement algorithm (5.76), the only storage needed is for W_k, P, X_k, Y_k, and $1/k$, or $3p + 2$ words. This small storage requirement (as well as the minimal computational requirement) is indeed a practical advantage. In many applications, convergence rate is an extremely important issue. One would certainly expect an algorithm of the form

$$W_{k+1} = W_k + \frac{1}{k}\left(P - K^{-1} \sum_{\ell=k-K+1}^{k} X_\ell X_\ell' W_k\right) \tag{5.77}$$

for any integer $K > 1$ to converge faster than (5.76). Note that in order to implement (5.77), the p x p matrix

$$F_k = K^{-1} \sum_{\ell=k-K+1}^{k} X_\ell X_\ell' \tag{5.78}$$

must be computed and stored. For large P, the storage requirement alone can preclude the use of algorithms such as (5.77). For the case being considered, R_{xx} is a Toeplitz matrix having only p distinct elements. Consequently, one is led to consider algorithms of the form

$$W_{k+1} = W_k + \frac{1}{k}(P - F_k^* W_k) \quad , \tag{5.79}$$

where F_k^* is an unbiased estimate of R_{xx} and constrained to be Toeplitz, e.g., consider F_k^* with the ij^{th} element given by

$$(F_k^*)_{i,j} = (p - |j-i|)^{-1} \sum_{\ell=1}^{p-|j-i|} (X_k)_\ell (X_k)_{\ell+|j-i|} \quad . \tag{5.80}$$

The idea here is that the ij^{th} element of F_k^* is the average of all terms on the $|i-j|^{th}$ diagonal of $X_k X_k^T$. Clearly, $E\{F_k^*\} = R_{xx}$ and F_k^* is Toeplitz. An obvious alternative to (5.79) with F_k^* given by (5.80) is

$$W_{k+1} = W_k + \frac{1}{k}(P - K^{-1} \sum_{\ell=k-K+1}^{k} F_\ell^* W_k) \quad , \tag{5.81}$$

with F_ℓ^* given by (5.80). Algorithms (5.80) and (5.81) can be implemented with virtually no increase in storage requirements over (5.76). It seems reasonable to conclude that algorithms such as (5.80) and (5.81) are viable alternatives to (5.77) in cases where storage requirements are an important issue and algorithm (5.76) converges too slowly to be of interest. Extensions of algorithms such as (5.80) and (5.81) for block Toeplitz R_{xx} are immediate. It is obvious that additional analytical work on the issue of convergence rate is necessary to evaluate the above remarks.

VI. CONCLUSION

In this work, new almost sure convergence results for a special form of the multidimensional Robbins-Monro stochastic approximation procedure are given. The form treated has been motivated by adaptive signal processing applications. Several types of data correlation matrices (e.g., Toeplitz and "block" Toeplitz) have been examined and new computationally efficient procedures have been given for both the inversion of a matrix having this special form and for solving a corresponding set of simultaneous linear equations. In this chapter, these new results are summarized, and suggestions for future work are presented.

A. Summary of New Results

The new convergence results of this work, presented in Chapter IV, are applicable to any algorithm that may be cast into the form of equation (3.1). It is shown in Chapter II that this particular form is applicable to many of the algorithms that have been proposed for adaptive signal processing application. Although many proposed algorithms make use of a constant gain sequence, i.e., $\mu_k = \mu$, it is pointed out in Section III-B that in order for these algorithms to be asymptotically unbiased when used with correlated data, the condition that $\mu_k \to 0$ is essential. The theorem which is stated and proved in Section IV-A transforms the convergence problem from consideration of the a.s. convergence of a stochastic difference equation to the a.s. convergence of several stochastic sequences. Corollary 2 of Section IV-A provides sufficient conditions on the decay rates of the autocovariance functions of the sequences $\{F_k\}$ and $\{C_k\}$ to

establish the conditions of the theorem. In Section IV-B the results of Corollary 2 are applied to several specific algorithms that have been proposed for adaptive signal processing. In particular, in the normal case with $F_k = X_k X_k'$, $P_k = s_k X_k$ or $P_k = P = E\{s_k X_k\}$, and $\mu_k = O(k^{-1})$, $\lim_{k\to\infty} k\mu_k > 0$, and W_k given by (3.1), if $\{s_k\}$, $\{X_k\}$ are jointly wide-sense stationary and all scalar correlation functions $\gamma(u)$ which can be computed for $\{s_j\}$, $\{X_j\}$ satisfy $\lim_{u\to\infty} u^{1/2}|\gamma(u)| < \infty$, then $|V_k| \overset{a.s.}{\to} 0$ as $k \to \infty$. For example, if $\{s_j\}$ and $\{X_j\}$ are finite-order autoregressive moving average processes, or can be viewed as samples of strictly bandlimited continuous time processes, then $|V_k| \overset{a.s.}{\to} 0$ as $k \to \infty$. Furthermore, even in the non-normal case, if $\{F_k\}$ and $\{P_k\}$ are M-dependent and $E\{\|F_k\|^q\}$, $q>2$, is bounded, then $|V_k| \overset{a.s.}{\to} 0$ as $k \to \infty$ for suitable $\{\mu_k\}$.

In Chapter V, special forms of data correlation matrices, R_{xx}, that are shown to arise in discrete time signal processing applications are examined. New computationally efficient procedures are developed for both the computation of R_{xx}^{-1} and the solution of $R_{xx}w = P$ in case R_{xx} is Toeplitz or block Toeplitz. The new procedures can result in a significant savings in storage requirements and computation time over standard solution techniques. For example, when R_{xx} is an $ML \times ML$ symmetric matrix having M^2 $L \times L$ submatrices arranged in Toeplitz form, the appropriate new procedure for the solution of $R_{xx}w = P$ requires approximately $2M^2L^3$ operations compared with approximately $(ML)^3/3$ for standard algorithms.

B. Suggestions for Future Work

Although the new convergence results presented in Chapter IV for algorithms of the form (3.1) seem to be the strongest convergence

results yet obtained under the weakest conditions, there are several important issues remaining. The results of Chapter IV apply to algorithms of the form of (3.1) with $E\{F_k\}$ symmetric and positive definite. Extensions to $E\{F_k\}$ nonsymmetric and positive definite are straightforward in view of Theorem 6.1 of [57]. Algorithms for which a.s. convergence was explicitly developed in Chapter IV can be interpreted as stochastic gradient-following algorithms, such as proposed by Widrow *et al.* [37], Griffiths [38], and Gersho [9]. Although stochastic projected gradient algorithms such as proposed by Lacoss [32] and Frost [33] can be cast into the form of (3.1), $E\{F_k\}$ is only positive semidefinite. It is the author's opinion that the results of Chapter IV can be easily extended to the analysis of these stochastic projected gradient algorithms.

An extremely important issue that warrants serious analytical treatment is the issue of convergence rate and the tradeoffs involved between convergence rate and computational requirements. In this regard, a treatment of truncated algorithms such as (3.41), algorithms which use a data-dependent gain sequence $\{\mu_k\}$, and decision-directed and decision feedback strategies would certainly seem to be of great interest. Other areas that merit additional work include (i) the effects of quantization errors on the convergence properties of these algorithms, (ii) strategies for use in nonstationary environs, and (iii) the asymptotic distribution of the "weight vector" for algorithms used with correlated training data.

Finally, the new results obtained in this work have applications to areas outside the realm of the adaptive signal processing schemes discussed in Chapter II. For example, the algorithms proposed by

Saridis and Stein [73], and Graupe and Perl [74] for the identification of systems fall directly into the framework of the new convergence results treated in Chapter IV. In fact, the results of Chapter IV provide analytical justification for an even broader family of algorithms than proposed in [73] and [74]. The new results of Chapter V are also applicable to the system identification problem.

REFERENCES

1. Graupe, D., *Identification of Systems*, Van Nostrand Reinhold Co., New York, 1972.

2. Sage, A.P., and J.L. Melsa, *System Identification*, Academic Press, New York and London, 1971.

3. Widrow, B., and M.E. Hoff, Jr., "Adaptive Switching Circuits," *1960 IRE WESCON Conv. Record*, pt. 4, pp. 96-104.

4. Sakrison, D.J., "Application of Stochastic Approximation Methods to System Optimization," Technical Rept. No. 391, Research Laboratory of Electronics, Massachusetts Institute of Technology, July 10, 1962.

5. Robbins, H., and S. Monro, "A Stochastic Approximation Method," *Ann. Math. Statist.*, *22*, 1951, pp. 400-407.

6. Kiefer, J., and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," *Ann. Math. Statist.*, *23*, 1952, pp. 462-466.

7. Lucky, R.W., "Automatic Equalization for Digital Communication," *Bell Syst. Tech. J.*, *44*, April 1965, pp. 547-588.

8. Lucky, R.W., "Techniques for Adaptive Equalization of Digital Communication Systems," *Bell Syst. Tech. J.*, *45*, Feb. 1966, pp. 255-286.

9. Gersho, A., "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," *Bell Syst. Tech. J.*, *48*, Jan. 1969, pp. 55-70.

10. Niessen, C.W., and D.K. Willim, "Adaptive Equalizer for Pulse Transmission," *IEEE Trans. Commun. Technol.*, *COM-18*, Aug. 1970, pp. 377-395.

11. George, D.A., R.R. Bowen, and J.R. Storey, "An Adaptive Decision Feedback Equalizer," *IEEE Trans. Commun. Technol.*, *COM-19*, June 1971, pp. 281-293.

12. Monsen, P., "Adaptive Equalization of the Slow Fading Channel," *IEEE Trans. Commun.*, *COM-22*, Aug. 1974, pp. 1064-1075.

13. Schonfeld, T.J., and M. Schwartz, "A Rapidly Converging First-Order Training Algorithm for an Adaptive Equalizer," *IEEE Trans. Inform. Theory*, *IT-17*, July 1971, pp. 431-439.

14. Schonfeld, T.J., and M. Schwartz, "Rapidly Converging Second-Order Tracking Algorithms for Adaptive Equalization," *IEEE Trans. Inform. Theory*, *IT-17*, Sept. 1971, pp. 572-579.

15. Kosovych, O.S., and R.L. Pickholtz, "Automatic Equalization Using a Successive Overrelaxation Iterative Technique," *IEEE Trans. Inform. Theory, IT-21,* Jan. 1975, pp. 51-58.

16. Qureshi, S.U.H., "Adjustment of the Position of the Reference Tap of an Adaptive Equalizer," *IEEE Trans. Commun., COM-21,* Sept. 1973, pp. 1046-1052.

17. Kobayashi, H., "Simultaneous Adaptive Estimation and Decision Algorithm for Carrier Modulated Data Transmission Systems," *IEEE Trans. Commun. Technol., COM-19,* June 1971, pp. 268-280.

18. Walzman, T., and M. Schwartz, "Automatic Equalization Using the Discrete Frequency Domain," *IEEE Trans. Inform. Theory, IT-19,* Jan. 1973, pp. 59-68.

19. Walzman, T. and M. Schwartz, "A Projected Gradient Method for Automatic Equalization in the Discrete Frequency Domain," *IEEE Trans. Commun., COM-21,* Dec. 1973, pp. 1442-1446.

20. Benedetto, S. and E. Biglieri, "On Linear Receivers for Digital Transmission Systems," *IEEE Trans. Commun., COM-22,* Sept. 1974, pp. 1205-1215.

21. Forney, G.D., Jr., "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference," *IEEE Trans. Inform. Theory, IT-18,* May 1972, pp. 363-378.

22. Qureshi, S.U.H., and E.E. Newhall, "An Adaptive Receiver for Data Transmission over Time-Dispersive Channels," *IEEE Trans. Inform. Theory, IT-19,* July 1973, pp. 448-457.

23. Magee, F.R., Jr., and J.G. Proakis, "Adaptive Maximum-Likelihood Sequence Estimation for Digital Signaling in the Presence of Intersymbol Interference," *IEEE Trans. Inform. Theory, IT-19,* Jan. 1973, pp. 120-124.

24. Bryn, F., "Optimum Signal Processing of Three-Dimensional Arrays Operating on Gaussian Signals and Noise," *J. Acoust. Soc. Am., 52,* 1962, pp. 39-51.

25. Burg, J.P., "Three-Dimensional Filtering with an Array of Seismometers," *Geophysics, XXIX,* Oct. 1964, pp. 693-713.

26. Middleton, D., and Groginsky, H.L., "Detection of Random Acoustic Signals by Receivers with Distributed Elements: Optimum Receiver Structures for Normal Signal and Noise Fileds," *J. Acoust. Soc. Am., 38,* 1965, pp. 727-737.

27. Capon, J., R.J. Greenfield, and R.J. Kolker, "Multidimensional Maximum-Likelihood Processing of a Large Aperture Seismic Array," *Proc. IEEE, 55,* Feb. 1967, pp. 192-211.

28. Schweppe, F.C., "Sensor-Array Data Processing for Multiple-Signal Sources," *IEEE Trans. Inform. Theory, IT-14,* Mar. 1968, pp. 294-305.

29. Capon, J., "Applications of Detection and Estimation Theory to Large Array Seismology," *Proc. IEEE, 58,* May 1970, pp. 760-770.

30. Young, G.O., and J.E. Howard, "Applications of Space-Time Decision and Estimation Theory to Antenna Processing System Design," *Proc. IEEE, 58,* May 1970, pp. 771-778.

31. Shor, S.W.W., "Adaptive Technique to Discriminate Against Coherent Noise in a Narrow-Band System," *J. Acoust. Soc. Am., 39,* 1966, pp. 74-78.

32. Lacoss, R.T., "Adaptive Combining of Wideband Array Data for Optimal Reception," *IEEE Trans. Geosci. Elect., GE-6,* May 1968, pp. 78-86.

33. Frost, O.L., "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proc. IEEE, 60,* Aug. 1972, pp. 922-935.

34. Winkler, L.P., and M. Schwartz, "Adaptive Nonlinear Optimization of the Signal-to-Noise Ratio of an Array Subject to a Constraint," *J. Acoust. Soc. Am., 52,* 1972, pp. 39-51.

35. Winkler, L.P., and M. Schwartz, "Constrained Array Optimization by Penalty Function Techniques," *J. Acoust. Soc. Am., 55,* May 1974, pp. 1042-1048.

36. Kobayashi, H., "Iterative Synthesis Methods for a Seismic Array Processor," *IEEE Trans. Geosci. Elect., GE-8,* July 1970, pp. 169-178.

37. Widrow, B., P.E. Mantey, L.J. Griffiths, and B.B. Goode, "Adaptive Antenna Systems," *Proc. IEEE, 55,* Dec. 1967, pp. 2143-2159.

38. Griffiths, L.J., "A Simple Algorithm for Real-Time Processing in Antenna Arrays," *Proc. IEEE, 57,* Oct. 1969, pp. 1696-1704.

39. Tack, D.H., private communication with L.L. Scharf, Dec. 1972.

40. Scharf, L.L., and D.C. Farden, "Optimum and Adaptive Array Processing in Frequency-Wavenumber Space," *Proc. 1974 IEEE Conf. on Decision and Control,* Nov. 1974, pp. 604-609.

41. Sage, A.P., and J.L. Melsa, *Estimation Theory with Applications to Communications and Control,* McGraw-Hill, New York, 1971.

42. Scharf, L.L., "On Stochastic Approximation and the Hierarchy of Adaptive Array Algorithms," *Proc. 1972 IEEE Conf. on Decision and Control,* New Orleans, Dec. 13-15, 1972, pp. 258-261.

43. Luenberger, D.G., *Optimization by Vector Space Methods,* John Wiley and Sons, Inc., New York, 1969.

44. Blum, J.R., "Approximation Methods which Converge with Probability One," *Ann. Math. Statist., 25,* 1954, pp. 382-386.

45. Blum, J.R., "Multidimensional Stochastic Approximation Methods," *Ann. Math. Statist., 25,* 1954, pp. 737-744.

46. Dvoretzky, A., "On Stochastic Approximation," *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, 1,* University of California Press, Berkeley and Los Angeles, 1956, pp. 39-55.

47. Wolfowitz, J., "On Stochastic Approximation Methods," *Ann. Math. Statist., 27,* 1956, pp. 1151-1156.

48. Derman, C., and J. Sacks, "On Dvoretzky's Stochastic Approximation Theorem," *Ann. Math. Statist., 30,* 1959, pp. 601-606.

49. Schmetterer, L., "Stochastic Approximation," *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability,* vol. 1, University of California Press, Berkeley and Los Angeles, 1961, pp. 587-609.

50. Schmetterer, L., "Multidimensional Stochastic Approximation," *Multivariate Analysis II, Proc. 2nd Int. Symp.,* P.R. Krishnaiah, editor, Academic Press, New York and London, 1969, pp. 443-460.

51. Sakrison, D.J., "Stochastic Approximation: A Recursive Method for Solving Regression Problems," *Advances in Comm. Sys. 2,* A.V. Balakrishnan, editor, Academic Press, New York and London, 1966, pp. 51-106.

52. Sakrison, D.J., "A Continuous Kiefer-Wolfowitz Procedure for Random Processes," *Ann. Math. Statist., 35,* 1964, pp. 590-599.

53. Daniell, T.P., "Adaptive Estimation with Mutually Correlated Training Sequences," *IEEE Trans. Syst. Sci. Cybernet., SSC-6,* Jan. 1970, pp. 12-19.

54. Senne, K.D., "Adaptive Linear Discrete-Time Estimation," Rept. SU-SEL-68-090 (Tech. Rept. 6778-5),Stanford Electronics Lab, Stanford, California, June 1968.

55. Daniell, T.P., "Adaptive Estimation with Mutually Correlated Training Samples," Rept. SU-SEL-68-083 (Tech. Rept. 6778-4), Stanford Electronics Lab., Stanford, California, Aug. 1968.

56. Kim, J.K., and L.D. Davisson, "Adaptive Linear Estimation for Stationary M-dependent Processes," *IEEE Trans. Inform. Theory, IT-21,* Jan. 1975, pp. 23-31.

57. Albert, A.E., and L.A. Gardner, Jr., *Stochastic Approximation and Nonlinear Regression,* Research Monograph No. 42, M.I.T. Press, Cambridge, Massachusetts, 1967.

58. Serfling, R.J., "Moment Inequalities for the Maximum Cumulative Sum," *Ann. Math. Statist., 41,* 1970, pp. 1227-1234.

59. Serfling, R.J., "Convergence Properties of S_n Under Moment Restrictions," *Ann. Math. Statist., 41,* 1970, pp. 1235-1248.

60. Knopp, K., *Theory and Applications of Infinite Series,* Hafner, New York, 1947.

61. Stout, W.F., *Almost Sure Convergence,* Academic Press, New York and London, 1974.

62. Rudin, W., *Principles of Mathematical Analysis,* McGraw-Hill, New York, 1964.

63. Farden, D.C., and L.L. Scharf, "Statistical Design of Nonrecursive Digital Filters," *IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-22,* June 1974, pp. 188-196.

64. Levinson, N., "The Wiener RMS Criterion in Filter Design and Prediction," *J. Math. and Phys., 25,* Jan. 1947, pp. 261-278.

65. Siddiqui, M.M., "On the Inversion of the Sample Covariance Matrix in a Stationary Autoregressive Process," *Ann. Math. Statist., 29,* 1958, pp. 585-588.

66. Trench, W.F., "An Algorithm for the Inversion of Finite Toeplitz Matrices," *J. Soc. Indust. Appl. Math., 12,* Sept. 1964, pp. 515-522.

67. Zohar, S., "Toeplitz Matrix Inversion: The Algorithm of W.F. Trench," Trench," *J. Ass. Comp. Mach., 16,* Oct. 1969, pp. 592-601.

68. Preis, D.H., "The Toeplitz Matrix: Its Occurrence in Antenna Problems and a Rapid Inversion Algorithm," *IEEE Trans. Ant. and Propagat.,* Mar. 1972, pp. 204-206.

69. Zohar, S., "The Solution of a Toeplitz Set of Linear Equations," *J. Ass. Comp. Mach., 21,* April 1974, pp. 272-276.

70. Markel, J.D., and A.H. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis," *IEEE Trans. Audio Electroacoust., AU-21,* April 1973, pp. 69-79.

71. Farden, D.C., "The Solution of a Special Set of Hermitian Toeplitz Linear Equations," *J. Ass. Comp. Mach.,* submitted Jan. 1975.

72. Farden, D.C., and L.L. Scharf, "Comments on 'The Statistical Design of Nonrecursive Digital Filters'," Authors' Reply, *IEEE Trans. Acoust., Speech, and Signal Processing,* submitted Jan. 1975.

73. Saridis, G.N., and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification," *IEEE Trans. Automat. Contr., AC-13,* Oct. 1968, pp. 515-523.

74. Graupe, D., and J. Perl, "Stochastic Approximation Algorithms for Identifying ARMA Processes," *Int. J. Systems Sci., 5,* Nov. 1974, pp. 1025-1028.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 11	2. GOVT ACCESSION NO: (14) TR-11 (ONR) ✓
	3. RECIPIENT'S CATALOG NUMBER ✓
4. TITLE (and Subtitle): (6) Stochastic Approximation with Correlated Data, ✓	5. TYPE OF REPORT & PERIOD COVERED: (9) Technical Report,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) David C. Farden	8. CONTRACT OR GRANT NUMBER(s): (15) N00014-67-A-0299-0019, N66001-72-C-0479
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Electrical Engineering, Colorado State University, Fort Collins, Colorado 80523	10. PROGRAM ELEMENT, PROJECT, AREA & WORK UNIT NUMBERS: (12) 124p.
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Code 436, Statistics and Probability Branch, Arlington, Virginia 22217 (11)	12. REPORT DATE: May 75 ✓
	13. NUMBER OF PAGES: 114
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Stochastic approximation, Robbins-Monro procedure, correlated data, time series, adaptive signal processing, moment conditions, autocovariance decay rate conditions, adaptive filters, adaptive arrays, minimum mean square error filter, FIR filter, Toeplitz matrix, block Toeplitz matrix.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

New almost sure convergence results for a special form of the multi-dimensional Robbins-Monro stochastic approximation procedure are developed. The special form treated is motivated by a consideration of several algorithms that have been proposed for discrete time adaptive signal processing applications. Most of these algorithms can also be viewed as stochastic gradient-following algorithms.

Continued on back.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

406 434 ✓ not

20. Essentially, previous convergence results contain a common "conditional expectation condition" which is extremely difficult (if not impossible) to satisfy when the "training data" is a correlated sequence. In contrast, the new convergence results developed in the present work are easily applied to cases where the "training data" is heavily correlated. In fact, the new convergence results are applicable when certain moments exist and certain "decay rates" on two autocovariance functions can be established. For example, when the data sequence is normal and (i) M-dependent, (ii) autoregressive moving average (ARMA), or (iii) can be viewed as samples of a band-limited continuous time process, the new convergence results can be applied to establish the almost sure convergence of each algorithm treated.

Several special forms of data correlation matrices that are shown to arise in descrete time signal processing are examined. New computationally efficient procedures are developed for both the inversion of a matrix having one of the treated special forms and for the solution of a corresponding set of simultaneous linear equations. The special forms treated are termed Toeplitz and block Toeplitz matrices.