AD-A032 639 PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESE--ETC F/G 9/4
APPLICATION OF STOCHASTIC APPROXIMATION TO FREQUENCY DOMAIN ADA--ETC(U)
OCT 75 F J EVERLY N00017-73-C-1418
UNCLASSIFIED TM-75-248 NL

1 OF 2
AD A032639

AD A032639

APPLICATION OF STOCHASTIC APPROXIMATION TO FREQUENCY DOMAIN ADAPTIVE FILTERS

Frederick J. Everly

Technical Memorandum
File No. TM 75-248
October 10, 1975
Contract No. N00017-73-C-1418

Copy No. 5

The Pennsylvania State University
Institute for Science and Engineering
APPLIED RESEARCH LABORATORY
Post Office Box 30
State College, PA 16801

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

DDC
RECEIVED
NOV 30 1976
B

NAVY DEPARTMENT

NAVAL SEA SYSTEMS COMMAND

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: TM-75-248	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): APPLICATION OF STOCHASTIC APPROXIMATION TO FREQUENCY DOMAIN ADAPTIVE FILTERS	5. TYPE OF REPORT & PERIOD COVERED: Ph.D. Thesis
	6. PERFORMING ORG. REPORT NUMBER: TM 75-248
7. AUTHOR(s): Frederick J. Everly	8. CONTRACT OR GRANT NUMBER(s): N00017-73-C-1418
9. PERFORMING ORGANIZATION NAME AND ADDRESS: The Pennsylvania State University, Applied Research Laboratory, P. O. Box 30, State College, PA 16801	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 186p.
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Sea Systems Command, Department of the Navy, Washington, D. C. 20362	12. REPORT DATE: Oct [redacted] 75
	13. NUMBER OF PAGES: 183 pages & figures
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Doctoral thesis.	15. SECURITY CLASS. (of this report): Unclassified, Unlimited
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited, per NSSC (Naval Sea Systems Command) 11/10/75

Technical memo.

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)

Beam Decouplers
Signal Processors
Stochastic Adaptive Filters

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research described is concerned with developing a system for adaptively processing the outputs of an array of sensing elements subject to certain optimality and processing time constraints. The adaptive part of the processor is designed to be a stochastic adaptive filter so that the processor can handle the inputs which are random processes. The processor is also recursive so as to be easily updated as the environment changes.

In the development of the theory of the stochastic adaptive filter, the orthogonal projection lemma is used to perform the minimization of the

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

391 007

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (continued)

complex variable problem resulting from these frequency domain considerations. Using the results of the minimization, a special recursive algorithm used to obtain the matrix filter coefficients used in the adaptive processor is derived from the theory of stochastic approximation.

The special recursive stochastic algorithm is shown to be a frequency domain multi-input, multi-output adaptive realization of the Wiener filter, and has as its goals, operating as an adaptive processor, the ability to gain fast increase in output signal-to-noise (S/N) and yet maintain statistical smoothing characteristics necessary for practical real time use of an adaptive processor.

ACCESSION for
NTIS White Section ☑
DDC Buff Section ☐
UNANNOUNCED ☐
JUSTIFICATION .....
.....
BY .....
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

Abstract

The research described here is concerned with developing a system for adaptively processing the outputs of an array of sensing elements subject to certain optimality and processing time constraints. Before the sensor outputs are passed to the adaptive processor, the sensing elements are combined to form a set of fixed beams. Each of the preformed beams is pointed in a different desired look direction, and it is on these preformed beams, after spectral analysis by the Fast Fourier Transform (FFT), that the adaptive part of the processor will operate. The adaptive part of the processor is designed to be a stochastic adaptive filter to be able to handle the inputs to the processor which are random processes. The processor is also recursive so as to be easily updated as the environment changes.

In the development of the theory of the stochastic adaptive filter, the orthogonal projection lemma is used to perform the minimization of the complex variable problem resulting from these frequency domain considerations. Using the results of the minimization, a special recursive algorithm, used to obtain the matrix filter coefficients used in the adaptive processor, is derived from the theory of stochastic approximation.

From the convergence proofs of the stochastic adaptive algorithm comes the fact that the sequence that describes the recursive filter is a martingale. This fact is of general use because it cannot only be used to prove convergence of a recursive stochastic algorithm, but also to show stability in the sense of a stochastic control system.

The special recursive stochastic algorithm is shown to be a frequency domain multi-input, multi-output adaptive realization of the Wiener filter, and has as its goals, operating as an adaptive processor, the ability to gain fast increase in output signal-to-noise (S/N) and yet maintain statistical

smoothing characteristics necessary for practical real time use of an adaptive processor.

The key phrase in the above expression is that the adaptive processor is a specially derived stochastic adaptive processor. The stochastic nature of the underlying processes takes the central role in the derivation of the stochastic adaptive filter.

The concept used in this research can be viewed as stochastic decoupled beam data processor. The complete data processing system can be viewed as an adaptive system which forms N beams and decouples these beams by putting them through an N input, N output linear system. This linear system contains a decoupling matrix (adaptive filter) derived from a recursive stochastic approximation algorithm. The linear system is adaptive and it decouples the beams in the statistical sense of decorrelation by removing the unwanted effect of interference signals from one beam to another.

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to Dr. Leon H. Sibul for his encouragement, direction and assistance throughout this investigation. The author would like to thank Dr. George J. McMurtry for his direction and help during the course of this study.

The author would like to thank Dr. John B. Lewis, Dr. John R. Mentzer and Dr. Anthony J. Ferraro for their reading of the manuscript.

The author also wishes to acknowledge the support of The Applied Research Laboratory of The Pennsylvania State University which is operated under contract to the Naval Sea Systems Command for providing the facilities necessary to the completion of this study.

TABLE OF CONTENTS

	Page
Acknowledgments	ii
List of Tables	vi
List of Figures	vii
Glossary of Symbols	viii
I. INTRODUCTION AND PROBLEM STATEMENT	
1.0 Introduction	1
1.1 Problem Statement	6
1.2 Outline	12
II. BACKGROUND THEORY	
2.0 Introduction	16
2.1 Array Theory	17
2.2 Detection Theory	23
2.3 Estimation: Linear Filters	26
2.3.1 Wiener Filter	26
2.3.2 Kalman-Bucy Filter	29
2.3.3 Implementation of Multichannel Wiener Filters	34
2.4 Tapped Delay Line Filters	35
2.4.1 Optimum Filters	35
2.4.2 Discrete Approximations	39
2.5 Adaptive Algorithms	40
III. STOCHASTIC APPROXIMATION AND PROBABILISTIC PROOFS	
3.0 Introduction	48
3.1 History of Stochastic Approximation Techniques	49
3.1.1 Robbins-Monro Procedure	50
3.1.2 Kiefer-Wolfowitz	52
3.2 Mathematical Preliminaries	58
3.2.1 Minimum Mean Square Error Estimation	58
3.2.2 Regression Function and Probabilistic Convergence	64
3.3 Convergence Proof of Stochastic Approximation	68
3.4 Statistical and Geometrical Significance of the Convergence Criterion	75

TABLE OF CONTENTS

		Page
IV.	ADAPTIVE FILTER DERIVATION	
4.0	Introduction	77
4.1	Derivation of Complex Matrix Filter	78
4.2	Decoupling Concept of Adaptive Filter	82
4.3	Proof of Convergence of Decoupling Matrix Filter	83
4.4	Mean Square Error	93
V.	DYNAMIC PROPERTIES OF STOCHASTIC FILTER	
5.0	Introduction	95
5.1	Dynamic Properties of Stochastic Adaptive Filter	96
5.1.1	Optimum Gain Calculation	96
5.2	Stochastic Stability Considerations	103
5.3	Unification of Previous Adaptive Systems Using Stochastic Approximation	107
VI.	COMPUTER SIMULATION RESULTS	
6.0	Introduction	112
6.1	Beam Construction and Signal Generation	113
6.2	Algorithm Parameter Determination	118
6.3	Adaptive Algorithm Applications	121
6.3.1	Effects of Special Gain Sequence	127
6.4	Generic Case Results	137
6.5	Stopping Rule	145
VII.	CONCLUSIONS AND EXTENSIONS	
7.0	Problem Statement	148
7.1	Results and Future Work	148
	Bibliography	158
	Appendix A: S/N Calculation	167
	Appendix B: Linear System Relationships	171

TABLE OF CONTENTS

		Page
Appendix C:	Probability Theory Definitions	173
	C.1 Conditional Expectations	173
	C.2 Martingales	174

LIST OF TABLES

Table		Page
I.	Coupling Coefficients	116
II.	Signal Variance	116

LIST OF FIGURES

Figure		Page
1.	System Model	2
2.	Stochastic Adaptive System	4
3.	Array Geometry	18
4.	Field Strength at Different Angles	21
5.	Optimum Likelihood Receiver	25
6.	State Variable Model and Optimum State Estimator	30
7.	Tapped Delay Line Filter	36
8.	Stochastic Approximation Algorithm	81
9.	Matrix Filter Decoupler	84
10.	Implementation of Adaptive Filter Decoupling Network	123
11.	Exponent Variation	128
12.	Gain Constant Variation	130
13.	Long Term Variation	134
14.	Beam Power Effects	135
15.	MSE Versus S/N	139
16.	Generic Case One	140
17.	Generic Case Two	142
18.	Generic Case Two	143
19.	Worst Case: Beam Intersection	144
20.	Generic Case Three	146

GLOSSARY OF SYMBOLS

H	Capital letters represent matrices except in probability definitions
$\underline{h}$	Small letters underlined represent vectors
h	Small letters represent scalars
h^i	A scalar which is a specific component of a vector
$\underline{h}^i$	A vector which is a specific column in a matrix
$h^{i,j}$	A scalar term in a matrix
$\underline{h}_n$	A vector at the n^{th} iteration
∇	Del operator
J	Jacobian
S/N	Signal-to-noise ratio
μ_n	Infinite gain sequence
s_k	A signal source
$G_{\underline{xx}}$	Power spectral density matrix
R_{xx}	Covariance matrix

CHAPTER I

INTRODUCTION AND PROBLEM STATEMENT

1.0 Introduction

In diverse fields such as communication systems (Allen, 2; Mermoz, 87; Muellar, 89), and seismic processing systems (Burg, 22), arrays of sensors (antennas, seismometers) are used to form beam patterns and to use these beam outputs as inputs to a processor which, in general, performs further signal processing. The processing is designed to be adaptive in this research. Each input to the processor is a stochastic process consisting of a desired signal (s_k) plus interfering noise (n_k) plus independent noise (sn_k) (Figure 1).

The research described here is concerned with developing a system for adaptively processing the outputs of an array of sensing elements subject to certain optimality (mean square error) and processing time constraints. The adaptive part of the processor is designed to be a stochastic adaptive filter which must efficiently filter inputs which are complex random processes. The processor is also recursive so as to be easily updated as the environment changes.

The adaptive processing system does not have to be tied to any type of array processing system but can be applied to any root finding problem that can be put into the form of the stochastic approximation algorithm. As soon as the stochastic adaptive processor is put into the framework of stochastic approximation, it will become clear that the

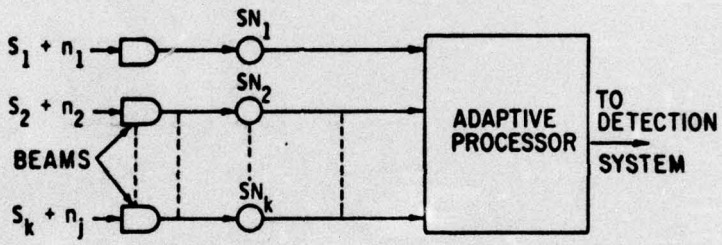


Figure 1. System Model

present algorithm has applicability to areas such as data communications (Muellar and Spaulding, 90), radio astronomy (Sakrison, 109), control system identification (Chien and Fu, 24) and biology (Cochran and Davis, 26).

In the development of the theory of the stochastic adaptive filter, the orthogonal projection lemma is used to perform the minimization of the complex variable problem resulting from the input signal considerations. Using the results of the minimization, a special recursive algorithm is derived from the theory of stochastic approximation. This recursive algorithm is used to obtain the matrix filter coefficients for the adaptive processor.

The special recursive stochastic algorithm is shown to be a frequency domain mulit-input, multi-output adaptive realization of the Wiener filter, and has as its goals, operating as an adaptive processor, the ability to gain fast increase in output signal-to-noise ratio (S/N) and yet maintain statistical smoothing characteristics necessary for practical real time use of an adaptive processor.

The key phrase in the above expression is that the adaptive processor is a specially derived stochastic adaptive processor. The stochastic nature of the underlying processes takes the central role in the derivation of the stochastic adaptive filter.

The concept used in this research can be viewed as stochastic decoupled data processor. The complete data processing system can be viewed as an adaptive system which has N inputs and decouples these inputs by putting them through an N input, N output linear system (Figure 2). This linear system contains an adaptive filter derived from a recursive stochastic approximation algorithm. The linear system is

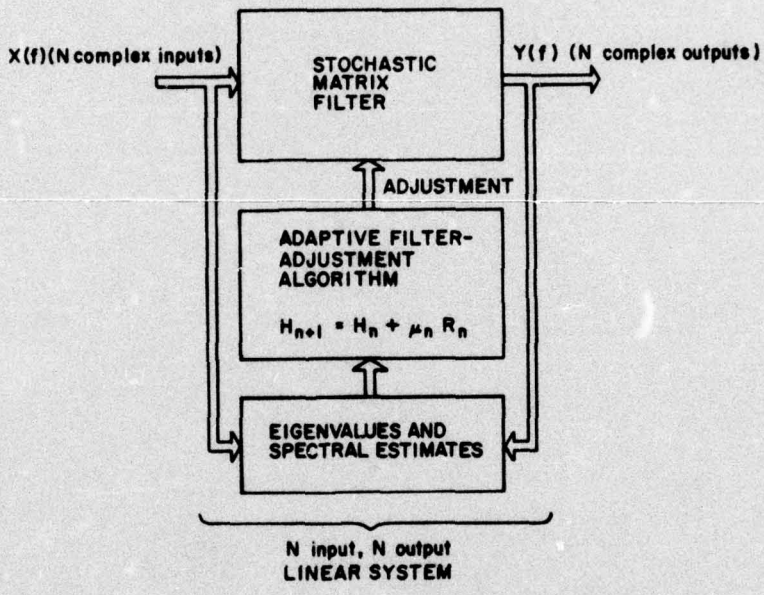


Figure 2. Stochastic Adaptive System

adaptive and it decouples input signals in the statistical sense of decorrelation by removing the unwanted effect of interference signals from one input to another.

Although no effort has been made to bridge the gap between the present algorithm and decoupling (Falb and Wolovich, 43) in multivariable control systems, there are many similarities. In decoupling of multivariable control systems, it is desired to have inputs control outputs independently, i.e., a single input influences a single output. This is, in essence, the goal of the adaptive processor. Even though the mathematical techniques involved in the two cases are different, both methods obtain a so-called decoupling matrix (or filter) which performs the stated purpose.

Schweppe (114) considered the maximum likelihood solution for a class of adaptive processors where he wished to decouple J signals from K sensors by forming J independent beams. This idea is just an application of the general ideas of statistical estimation and detection to the array processing problem and was not designed to be an adaptive processor in the same sense as is used in this research. It does, however, contain the first reference to a so-called decoupled beam data processor.

A knowledge of the statistics of the signal field as well as the noise field is necessary to describe system performance in statistical detection theory (Van Trees, 125; Whalen, 132). The adaptive processor in this research assumes certain knowledge of the statistics of the input signals and noises. The basic fact assumed is that the signal in a beam is uncorrelated with all other sources in a beam. Since optimum

detection systems require complete statistical knowledge, the adaptive processor becomes attractive in situations where complete information is not available.

It is assumed in this analysis that all signals and noises are at least wide sense stationary, and no particular statistical distributions are assumed for the signals and noises in the derivation of the stochastic adaptive filter.

The succeeding section of this chapter contains a formal statement of the problem solved and a brief outline of the properties and innovative qualities of the proposed stochastic adaptive filter. This section provides some background necessary to understand the derivation of the adaptive matrix filter from the use of the recursive algorithm which is derived from stochastic approximation techniques and its position in the literature of control systems, adaptive arrays and stochastic approximation.

1.1 Problem Statement

The problem addressed in this research is to derive a complex stochastic adaptive matrix filter as an approximation to the multi-input, multi-output Wiener filter. The adaptive filter operates on the complex stochastic inputs. The recursive algorithm used to derive the adaptive matrix filter uses a special purpose stochastic approximation technique. This special purpose stochastic approximation technique should have the properties of fast improvement in output S/N irrespective of noise statistics and statistical smoothing of the matrix adaptive filter coefficients. The recursive algorithm and the resultant

stochastic adaptive matrix filter has to be stable, and ought to be conceptually simple, and easily implemented with little computer storage in a real time environment.

The following is a summary of the features, innovations, mathematical techniques used to derive the stochastic adaptive algorithm, and the differences between the stochastic adaptive algorithm and previous deterministic adaptive algorithms. This summary lists the characteristics of the stochastic adaptive algorithm and acts as a prelude to a more detailed analysis which gives the important techniques and properties of the stochastic adaptive algorithm.

1. In previous work, most adaptive filters were synthesized by using time domain concepts such as correlations and time delays in some form of tapped delay line filter. In this research, the adaptive filter is implemented by using frequency domain concepts such as cross-spectral densities and multiplication of complex quantities.

All previous attempts at adaptive processors have been either single input, single output, or multi-input, single output systems. The present system is a first attempt at an adaptive complex multi-input, multi-output adaptive processor.

2. Since the stochastic adaptive algorithm is a frequency domain implementation of an approximation to a Wiener filter, complex variables arise naturally in the derivation of the optimum solution, and a special technique called the orthogonal projection lemma is used to derive the optimum filter.

3. Both the stochastic nature of the random processes involved and the complex variables which arise from the complex input signals

make derivation and implementation of adaptive algorithms difficult. It is shown, however, that any adaptive algorithm that does not take the stochastic nature of the underlying signals into account is not as useful as stochastic algorithms in situations where the underlying signals are random processes.

The use of a specially derived stochastic approximation procedure enables the system designer to obtain fast increase in output S/N and yet maintain the advantageous statistical smoothing properties of the stochastic adaptive algorithm not available with any deterministic gradient algorithm.

4. From the convergence proofs of the stochastic adaptive algorithm comes the fact that the sequence that describes the recursive algorithm is a martingale. This fact is of general use because it cannot only be used to prove convergence of a recursive stochastic algorithm, but also to show stability in the sense of a stochastic control system.

The following paragraphs contain a more detailed look at the possibilities of the adaptive algorithm, the mathematical techniques used in the derivation of the complex stochastic adaptive system and its potential advantages and innovations over deterministic gradient procedures.

Since the present system involves complex performance measures which do not always, in analogy with their real variable counterparts, have uniquely defined derivatives, new methods for minimization of a performance functional are necessary. In most minimization problems involving real variables, there is usually no problem in taking

derivatives of the performance functional. However, in the complex variable case, even the simple performance measure, mean square error, does not have a uniquely defined derivative. To use techniques involving complex variables which do not have a derivative, some other method for finding the optimum solution must be found.

For the complex case of the recursive algorithm, the non-analyticity of the performance functional makes it impossible to talk of differentiation in the usual sense. The technique of orthogonal projection (Halmos, 53), which does not depend on differentiation, can be used to find the optimum solution. This mathematical procedure is very general and can be applied to a wide variety of situations. Kalman (65), in one of the orthogonal projection lemmas more famous applications in the engineering literature, applied the technique to derive the minimum variance unbiased state estimator which bears his name. The orthogonal projection lemma can be applied to situations where conventional derivatives cannot be used, and it is used here to derive the optimum complex frequency domain solution to the minimization problem.

Since all the signals involved in the use of the adaptive algorithm are stochastic processes, the use of an adaptive algorithm that takes the stochastic nature of the random processes involved into account is paramount to the success of any adaptive algorithm.

A special purpose stochastic approximation method is derived to calculate the optimum filter by use of a stochastic recursive algorithm. The misadjustment or variance of the filter weights for any deterministic algorithm can be shown to be inversely proportional to the speed of the algorithm. Thus, for a fast increase in output S/N, there is a large misadjustment in the filter weights. The attainment of both small

variance in the filter weights and fast convergence (fast increase in S/N) is shown to be impossible in any algorithm which does not take into account the stochastic nature of the signals involved. It is shown that all the gradient type algorithms derived from the mean square error functional are just approximations to stochastic approximation, and since the processes involved in the adaptation are stochastic processes, stochastic approximation is a much more logical choice for the adaptive algorithm than any deterministic gradient type algorithm.

The recursive stochastic adaptive algorithm used to calculate the adaptive filter uses a special infinite sequence for the variable gain, μ_n . This special variable gain is derived from convergence and stability considerations of the idealized form of the recursive adaptive algorithm. This variable gain allows the stochastic adaptive algorithm to operate near the stability boundary of the recursive algorithm. By operating at the stability boundary, the adaptive algorithm has fast initial convergence, and yet maintains the smoothing properties one obtains from using a stochastic approximation method. It would be impossible to operate a deterministic algorithm at the stability limit of the algorithm because the deterministic algorithms possess no smoothing properties, and the variance of the calculated filter matrix would be extremely high and unusable in a practical system.

Due to properties inherent in the recursive algorithm of stochastic approximation, the method usually exhibits slow long term convergence. This attribute is mainly due to the time variable gain sequence. The real advantage of using the special time variable gain sequence derived from stability considerations of the idealized algorithm is that one obtains the maximum convergence rate obtainable

within the stability limits and yet maintain small misadjustment in the filter coefficients.

In many applications, the practical usefulness of adaptive arrays is limited by their convergence rate. The adaptively controlled filter coefficients must change at a rate equal to or greater than the rate of change of the external random inputs. The convergence rate is most severe in systems where the eigenvalues of the input covariance matrix of beams differ by several orders of magnitude. The method proposed in this research makes use of the distribution of eigenvalues to obtain fast convergence no matter how widely separated the eigenvalues are. No deterministic adaptive system (including matrix inversion) can obtain as fast convergence properties without suffering the undesirable result of extremely large filter coefficient variance.

Certain conditions are formulated which guarantee convergence of any recursive algorithm using stochastic approximation. From the proof of convergence of the form of recursive algorithm used in this research comes the fact that the sequence of solutions of the recursive algorithm is shown to be a martingale. This fact is not only used in the convergence proof of the stochastic approximation algorithm, but also to show that the adaptive algorithm, when formulated in the context of a state variable control system, is stable in terms of stochastic stability.

The convergence of the stochastic adaptive algorithm is equivalent to the stability of systems described by stochastic difference equations. The stability of these type of systems must be considered in a probabilistic sense so that stochastic analogs of Lyapunov functions can be defined. It is shown that the stochastic Lyapunov

functions are martingales, and all the useful properties of martingales (see Appendix C) accrue to these stochastic Lyapunov functions.

The purpose of an adaptive processor is that the system somehow adapts to changes in the environment and removes the effect of interfering noises from the output of the processor.

An intuitive meaning to the problem of removing any interfering signals from a given input can be gleaned from the meaning of diagonalization of the power spectral density matrix of beam outputs. When the power spectral density matrix of beam outputs is diagonalized, then the diagonalized matrix represents the covariance matrix of inputs in terms of a set of orthogonal vectors. In the statistical sense, this diagonalization decorrelates the beam outputs and makes the beams independent in the Gaussian case (Van Trees, 125).

The adaptive processing system is designed to extract a particular desired signal from a given input in the presence of many other signals which are considered to be interfering noises. We shall address the difficult problem of acquiring a weak desired signal in the presence of strong interference by the use of adaptive arrays.

1.2 Outline

Since the background, derivation and use of the stochastic adaptive processor uses concepts from a wide range of mathematical and engineering areas, a brief introduction to each of the major concepts is given in Chapter II. An introduction to array theory is presented in Section 2.1. This is presented as background to the particular type of problem solved in the experimental tests. A brief summary of linear system theory, especially the topic of Wiener filtering and the

approximations to the Wiener filter, is given in Section 2.3. The concepts of linear systems and Wiener filtering are needed to understand how the present adaptive algorithm fits into the framework of a detection system. Concepts needed from state variable control systems and probability and statistics as applied to stochastic processes appear in the appendices. Concepts from digital signal processing are used liberally in the adaptive processor solution and application.

Since the tapped delay line filters are by far the most common time domain implementation of the Wiener filter, a description of its derivation and the assumptions inherent in it are given in Section 2.3. Various adaptive algorithms considered in this research are documented in Section 2.4. This section gives insight as to how the present adaptive algorithm differs from previous ones.

Chapter III contains the probabilistic techniques necessary for the derivation and use of the stochastic adaptive filter. This chapter contains a history of the technique called stochastic approximation. The proofs deriving the sufficient conditions for minimum mean square error, and probabilistic convergence criterion are given in Section 3.2. The interpretation of mean square estimation is given in this section. Section 3.2 also contains the derivation for the regression function to be used in the recursive algorithm used to calculate the adaptive filter. The convergence proof of the stochastic approximation algorithm in Section 3.3 establishes the important result that the sequence of solutions of the adaptive filter is a martingale. The last section in this chapter contains the statistical and geometrical significance of the results of the previous sections.

Chapter IV contains the derivation and proof of convergence of the stochastic adaptive filter. Use is made of the orthogonal projection lemma to derive the optimum solution to the minimum mean square error problem. Section 4.2 establishes the meaning of the decoupling concept as it applies to the adaptive filter.

The dynamic properties of the stochastic adaptive algorithm are derived in Chapter V. The derivation of the optimum gain sequence from considerations of the idealized form of the adaptive algorithm and the extended convergence conditions are contained in Section 5.1. Section 5.2 gives the stochastic stability considerations and shows that the stochastic Lyapunov functions are martingales. The last section unifies all LMS type algorithms by using the framework of stochastic approximation but points out the fallacy of using deterministic algorithms (LMS type, etc.) when the processes involved in the adaptive processor are stochastic processes.

Chapter VI contains the construction of the computer simulation and lists the generic cases tested. The influence of the gain sequence and the optimum gain constant on convergence rates is illustrated by the computer simulation results. The physical reasoning and justification for using implied constraints in some experimental tests and the fact that these cases do not differ from the cases with no constraints for all but the long term operation of the recursive algorithm is given in Sections 6.2 and 6.3. The influence of a search strategy on long term convergence and a special stopping rule are discussed in Section 6.3.

The last chapter contains a summary of the work completed and the conclusions reached. It also contains proposals for future work and extensions into other areas.

CHAPTER II

BACKGROUND THEORY

2.0 Introduction

The present chapter gives the basic background for the detection and estimation problem, shows the connection of Wiener and Kalman filtering to the detection and estimation problem and shows how the present stochastic adaptive processor approximates the Wiener filter solution. Even though any particular type or special characteristic of an array of sensing elements is not necessary for the derivation and use of the stochastic adaptive algorithm, a brief summary of the properties of arrays is included. This summary of array theory gives a firm basis to understand the physical considerations involved in the experimental verification of the stochastic adaptive algorithm.

The following sections give the background of general detection and estimation systems and show the use of various previous optimum and adaptive methods to increase detection system performance. They show the complexity of true optimum systems and that, without some simplifying considerations, the implementation of the true detection or estimation systems in real time would not be feasible.

We illustrate the derivations for the optimum filters for both stationary (Wiener filter) environments and non-stationary (Kalman filter) environments and show various ways to implement a multichannel Wiener filter.

Almost all previous attempts at adaptive processors have been some form of time domain tapped delay line filter using deterministic minimization procedures. The tapped delay line filter is derived and its implementation is discussed. Since this filter structure is by far the most used time domain approximation to the Wiener filter, it is important to discuss its characteristics and limitations. The distinction between this time domain approach and the frequency domain approach taken here is important in understanding the differences and advantages of the frequency domain stochastic adaptive filter proposed in this research.

The last section delineates the various approximations and adaptive system approaches to the Wiener filter problem, and shows the various assumptions adaptive system designers have used to obtain realistic approximations to the optimum Wiener filter. It contains a summary of some of the most important contributions to the adaptive array literature and a brief discussion of the differences of the various adaptive techniques attempted.

2.1 Array Theory

Using results contained in Allen (2), we can make clear the use of an array of sensing elements in the recursive stochastic algorithm. Let us assume that there are N isotropic sensors at arbitrary positions specified by a set of vectors, $\underline{v}_n$ (Figure 3). The contribution of the n^{th} element to the far field at some point at a distance R is

$$f_n(e_R) = a_n e^{jk(\underline{v}_n \cdot \underline{e}_R)} ; \qquad k = \frac{2\pi}{\lambda} , \tag{2.1}$$

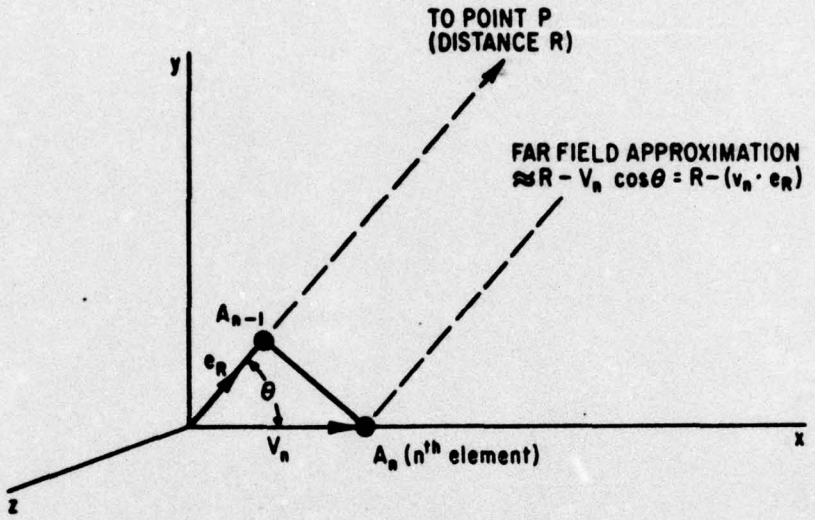


Figure 3. Array Geometry

where $(\underline{v}_n \cdot \underline{e}_R)$ is the perpendicular distance of the n^{th} element from a plane through the origin perpendicular to the direction $\underline{e}_R$, λ is the wavelength, and a_n represents the amplitude response of the n^{th} element. The product $k(\underline{v}_n \cdot \underline{e}_R)$ gives the phase shift of the signal (at distance R) relative to the reference point of the array.

The total expression for the field strength for the far field pattern is the sum of the contributions from all of the elements:

$$f(e_R) = \sum_{n=0}^{N-1} a_n e^{jk(\bar{v}_n \cdot \bar{e}_R)} \quad . \tag{2.2}$$

If we consider the simple case of an array of sensors equally spaced along a line, then Equation (2.2) becomes:

$$f(\alpha) = \sum_{n=0}^{N-1} a_n e^{jk\, n\, D \sin \alpha} \quad , \tag{2.3}$$

where $k(\underline{v}_n \cdot \underline{e}_R)$ becomes $k(n\ D \sin \alpha)$, the phase shift of the n^{th} element, α is the complement of the spherical angle θ, and D is the spacing between elements.

We find that the equation of normalized magnitude of Equation (2.3) with the amplitude factors $a_n = 1$ is of the form:

$$\frac{\sin N x}{N \sin x} \quad , \tag{2.4}$$

where $x = (\frac{\pi D}{\lambda}) \sin \alpha$.

The maximum of Equation (2.4) occurs at the origin and is referred to as a main lobe. The secondary maxima of Equation (2.4) are

called side lobes. The number of elements, the size of the aperature and the amplitude illumination factors a_n determine the characteristics of the array. One use of the proposed stochastic adaptive algorithm would be to allow relatively high side lobes in the pattern thereby reducing the cost and complexity of the array.

One of the attractive features of the array configuration is that the main beam of the array can be pointed without moving the array itself (Allen, 2). If we desire to point the main beam at an angle α_o from the perpendicular to the array, then Equation (2.3) becomes

$$f(\alpha,\alpha_o) = \sum_{n=0}^{N-1} a_n e^{jkn\, D[\sin\alpha - \sin\alpha_o]} , \qquad (2.5)$$

and the main beam is translated from the point $\alpha = 0$ to the point $\alpha = \alpha_o$.

If we have the location of point sources as in Figure 4, then the array output would be the sum of the response of the array to each source. For any given main beam, only one source could be steered on and the other sources would enter the beam as extraneous signals or noises.

Mathematically this would be

$$\begin{aligned} f_{\text{TOTAL BEAM 1}} &= \text{Array output for beam 1} \\ &= c_{1,1}s_1 + c_{1,2}s_2 + c_{1,3}s_3 + \ldots + c_{1,n}s_n \end{aligned} \qquad (2.6)$$

where $c_{1,j}$ represents the proportion of source s_j that appears in beam 1; $c_{n,n}$ is always taken to be one because it is the constant

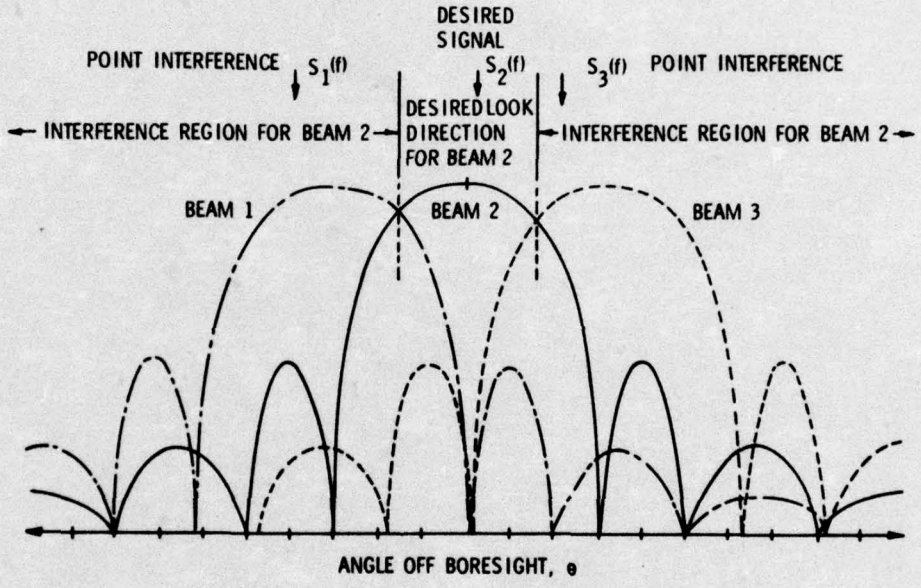


Figure 4. Field Strength at Different Angles

representing the position of the signal s_n in the beam b_n and this signal is considered the signal steered on in that beam and not a noise signal.

The output of the array comes not only from the signal that the particular beam is steered on, i.e the signal direction is known in each beam, but also from any signals in the far field of that particular beam. However, since all other signals except the one steered on are not at the center of the main lobe, they contain less power in this beam than in the beam which steers on them. The same can be said of any steered beam. We can summarize by stating that the output power of the processor due to a directional noise depends on the direction of propagation of the noise and its power spectrum, as well as the geometry of the array and each beams particular steering direction.

We call a process with certain spatial characteristics a signal if it impinges on the sensors in a main beam, but we call the same spatial process a noise in any other beam. In this research, it is assumed that there are multiple main beams. Other spatial processes such as self-noise and environment noise are considered noises in all beams.

We would like to design a processor which would detect signals anywhere inside the 3 dB points of the main beam but reject anything outside these points (Figure 4). In designing the beam outputs of an array, the designer is limited by the size (aperature) of the array and the number of sensors. These considerations place limits on the width of the main beam and the amplitude of the side lobes. The purpose of any processing system is to improve the detection system performance

using a given array. With the stochastic adaptive technique being proposed, the side lobe level of a beam can be relatively high as long as the directionality of the main beams is maintained and the adaptive processor will reject unwanted signals from the main beam (outside the 3 dB points) and the side lobes so as to improve the systems detection performance. These facts reduce the constraints on the array designer and reduce the complexity of the array design while improving detection system performance by reducing the false alarm rate.

2.2 Detection Theory

It is a fact from statistical detection theory that both the Neyman-Pearson and Bayes detection criteria lead to a likelihood-ratio detector (Van Trees, 125). The objective of any filter used before detection is to improve the probability of detection by increasing the signal-to-noise ratio (S/N) at the output of the filter which is the input to the likelihood-ratio detector.

When the statistics of the signal are known exactly and the noise is white Gaussian noise, the optimum detector is a matched filter (Whalen, 132). This is a relatively simple system and much knowledge has been accumulated on matched filter systems. However, the matched filter assumes complete statistical information about the input processes.

The problem of detecting Gaussian signals in additive Gaussian noise fields was studied by Bryn (20), who showed that, assuming K antenna elements in the array, the Bayes optimum detector could be implemented by the measurement and inversion of a 2K by 2K correlation matrix. Mermoz (87) proposed a similar scheme for known narrowband

signals, using the signal-to-noise ratio (S/N) as a performance criterion. It is obvious that for any reasonable size array (50 sensors or more) the optimum Bayes detector would be almost impossible to implement in real time.

Let us now discuss the detection of an unknown signal (Goode, 50). If we assume Gaussian statistics and make the following assumptions

$$\underline{x}(t) = \underline{s}(t) + \underline{n}(t) \quad , \tag{2.7}$$

$$E\{\underline{x}(t)\} = 0 = E\{\underline{n}(t)\} \quad , \tag{2.8}$$

$$E\{\underline{x}(t)\ \underline{x}^T(u)\} = R_{\underline{xx}}(t,u) \tag{2.9}$$

and

$$E\{\underline{x}(t)\ \underline{n}^T(u)\} = 0 \quad , \tag{2.10}$$

then the stochastic signal is completely described in the statistical sense. The optimum detection system (Figure 5) for the preceding problem comes from the solution of the following integral equation (Middleton and Groginsky, 88)

$$\int_0^T \int_0^T R_{\underline{nn}}(t,u)\ K(u,v)[R_{\underline{nn}}(v,\tau) + R_{\underline{ss}}(v,\tau)]du\ dv = R_{\underline{ss}}(t,\tau) \quad . \tag{2.11}$$

The likelihood ratio detector which results from the solution of this integral equation is quite complicated and, for practical reasons, almost impossible to implement. The array processor design engineer must simplify the structure of the likelihood ratio detector in order for it to operate in a real time environment. Any a priori knowledge

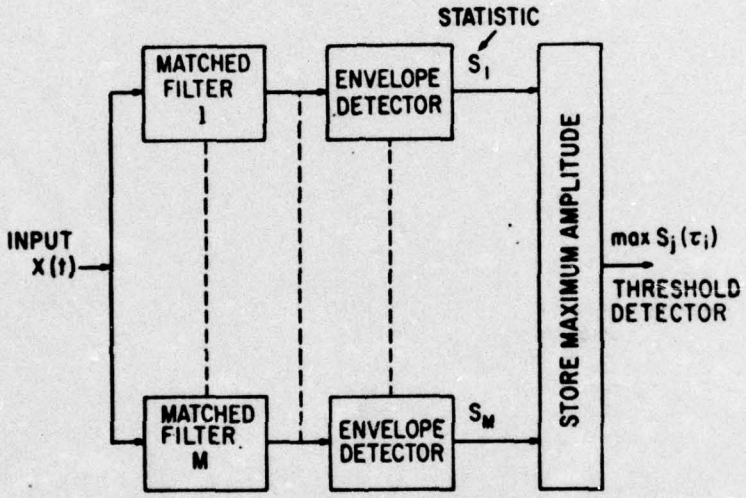


Figure 5. Optimum Likelihood Receiver

of signal or noise statistics, short of complete knowledge in the optimum case, allows the designer the opportunity to use some adaptive structure.

The standard engineering parameter, signal-to-noise ratio (S/N), which is defined as the ratio of the average signal power to the average noise power is the basic measure of good performance of a detection system. The family of curves plotted in detection theory to predict the likelihood ratio detectors performance are referred to as receiver operating characteristics (ROC). A study of these curves will show immediately that the most important variable in those curves is the signal-to-noise ratio. As the signal-to-noise ratio goes up, the receiver performance goes up even faster than a linear function (for a fixed false alarm rate). Any system which increases the signal-to-noise ratio is invaluable in a detection system.

The adaptive processor described in this research increases the signal-to-noise ratio at the output of the adaptive filter so as to increase the probability of success in the likelihood ratio detector.

2.3 Estimation: Linear Filters

2.3.1 Wiener Filter If we assume that the input process is corrupted by noise and we want to extract the signal from the noise, then the problem is one of filtering. The performance measure used to calculate the Wiener filter is the mean square error between the signal and a linear estimate of the signal. We have assumed (Section 2.2) that we know the desired signal $d(t)$, as well as the input signal correlation matrix, $R_{ss}(t,u)$. Let us also assume that we know a correlation

vector $\underline{r}_{d\underline{x}}(\tau,u)$ defined by:

$$\underline{r}_{d\underline{x}}(\tau,u) = E[d(\tau)\ \underline{x}(u)] \quad . \tag{2.12}$$

The solution of the following integral equation gives the filter vector, $h_o(w,v)$, which minimizes the mean square error.

$$\underline{r}_{d\underline{x}}(t,\sigma) = \int_{T_i}^{T_f} [R_{\underline{nn}}(t,u) + R_{\underline{ss}}(t,u)]\ \underline{h}_o(u,\sigma)\ du \quad . \tag{2.13}$$

If we assume all processes are stationary and we use a change of variables

$$\tau = t - \sigma \qquad \text{and} \qquad v = t - u \tag{2.14}$$

then we can write Equation (2.13) as:

$$\underline{r}_{d\underline{x}}(\tau) = \int_o^{\infty} R_{\underline{xx}}(v)\ \underline{h}(\tau - v)\ dv \quad . \tag{2.15}$$

The solution $\underline{h}_o(t)$ is called a Wiener filter (Wiener, 134).

The analytical solution of the Wiener-Hopf integral Equation (2.15) requires spectral factorization which is difficult to implement as a simple iterative procedure.

If certain restrictions (including those made above) are put on the input signal and the optimum filter, then certain simplifications can be made.

The specific restrictions on the input $\underline{x}(t) = \underline{s}(t) + \underline{n}(t)$ and on the optimum filter $\underline{h}_o(t)$ are:

(1) Both the signal, $\underline{s}(t)$, and the noise, $\underline{n}(t)$, sample functions from random processes that are at least wide sense stationary.

(2) The filter, $\underline{h}_o(t)$, is to be physically realizable.

If we require that $\underline{h}_o(t)$ be physically realizable, then we must guarantee

$$\underline{h}_o(t) = 0 \quad \text{for} \quad t \leq 0 \ . \tag{2.16}$$

Under the assumptions of stationarity and physical realizability, we can change the lower limit on the integral in Equation (2.15) to minus infinity and then take the Fourier transform of the resulting equation. The result is well known and it is

$$\underline{h}_o^*(f) = G_{\underline{xx}}^{-1}(f)\, g_{d\underline{x}}(f) \ . \tag{2.17}$$

What is assumed, in most of adaptive filter literature, by Equation (2.17) is that in doing the spectral factorization the correct singularities are selected so as to make the filter realizable. While both the unrealizable and realizable filters have the same generic form, they are in fact two different filters because of the types of singularities contained in each. The fact that $\underline{h}_o^*(f)$ is optimum but

unrealizable since, in general, it possesses singularities in the lower half plane and hence its Fourier transform does not vanish for negative time, is sometimes overlooked.

Spectral factorization is a time consuming process and difficult to implement in real time. Anderson et al. (3) discuss a recursive algorithm for spectral factorization but gives no estimate of the time required to do the factorization. Since the method requires the recursive solution of a Riccati difference equation, it involves considerable computational effort.

The purpose of the proposed adaptive algorithm is to obtain an adaptive approximation to the optimum Wiener filter in a computationally fast and efficient manner and to use the resulting filter in a real time system.

2.3.2 Kalman-Bucy Filter For optimum state estimation in non-stationary environments, the Kalman filter gives the best estimate in the minimum mean-square error sense (Sage, 107). Kalman (63) changed the classical formulation of the minimum mean square error problem by using a state variable model for the state estimator. The linear system, whose state we want to estimate, is assumed driven by white noise (Figure 6). We wish to find a state estimator that gives the best estimate $\hat{\underline{x}}(t)$ of $\underline{x}(t)$ in the minimum mean-square error sense. Thus we want to find an estimator such that

$$J = E\{[\underline{x}(t) - \hat{\underline{x}}(t)]^T[\underline{x}(t) - \hat{\underline{x}}(t)]\} = E\{\varepsilon^T\varepsilon\} \qquad (2.18)$$

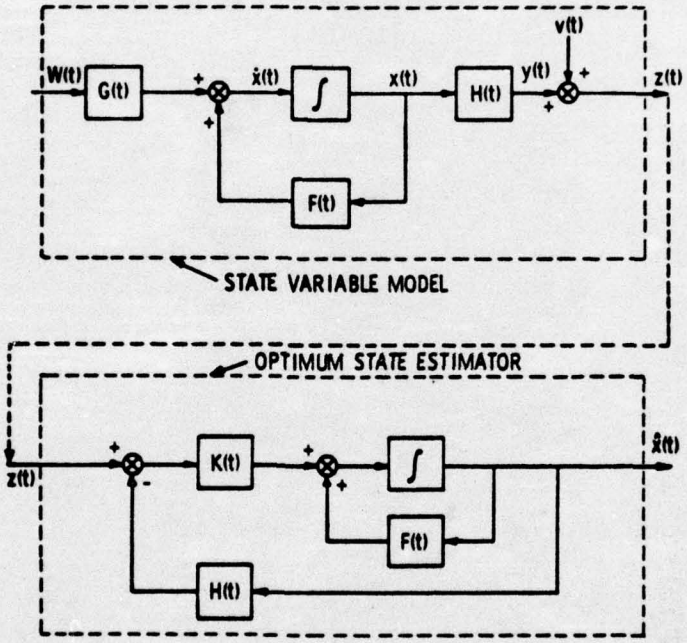


Figure 6. State Variable Model and Optimum State Estimator

is a minimum. The equation describing the system model is

$$\dot{\underline{x}} = F(t)\,\underline{x}(t) + G(t)\,\underline{w}(t) \quad . \tag{2.19}$$

The observation is

$$\underline{z}(t) = H(t)\,\underline{x}(t) + \underline{v}(t) \quad . \tag{2.20}$$

It is assumed that the input $\underline{w}(t)$ and the measurement noise $\underline{v}(t)$ are zero mean and white, and that they are uncorrelated such that we have

$$E\{\underline{w}(t)\,\underline{w}^T(\tau)\} = Q(t)\,\delta(t-\tau) \tag{2.21}$$

$$E\{\underline{v}(t)\,\underline{v}^T(\tau)\} = R(t)\,\delta(t-\tau) \tag{2.22}$$

and

$$E\{\underline{w}(t)\,\underline{v}^T(\tau)\} = E\{\underline{v}^T(t)\,\underline{w}^T(\tau)\} = 0 \quad . \tag{2.23}$$

We wish to find the estimator that gives the linear minimum mean square error and is unbiased such that

$$E\{\hat{\underline{x}}(t)\} = E\{\underline{x}(t)\}$$

and

$$E\{\tilde{\underline{x}}^T(t)\,\tilde{\underline{x}}(t)\} = \text{minimum} \tag{2.24}$$

where $\tilde{x}(t)$ is the estimation error given by:

$$\tilde{\underline{x}}(t) = \underline{x}(t) - \hat{\underline{x}}(t) \quad . \tag{2.25}$$

The solution to this linear minimum mean square error problem still results, as in Wiener filtering, in an integral equation. Rather than solving the integral equation directly, Kalman converts the integral equation to a nonlinear Riccati type differential equation. The solution of this Ricatti type nonlinear differential equation is the covariance matrix of the minimum mean square error. Kalman originally performed this linear mean square error minimization by using the orthogonal projection lemma as formulated from Hilbert space considerations.

The orthogonal projection lemma (Halmos, 53) states that the minimum linear mean square error estimate is orthogonal to the estimation error. This means that the covariance of the estimate and the estimation error is zero,

$$\text{cov}[\underline{\hat{x}}(t), \underline{\tilde{x}}(t)] = 0 \quad . \tag{2.26}$$

It can be shown, under the assumptions made, that the minimum mean square error is obtained if we minimize each element of the error covariance matrix

$$P(t) = \text{cov}[\underline{\tilde{x}}(t), \underline{\tilde{x}}(t)] = E\{\underline{\tilde{x}}(t)\, \underline{\tilde{x}}^{T}(t)\} \tag{2.27}$$

separately.

The following equations are the result of the minimization procedure using the orthogonal projection lemma. The derivation (Sage, 107) appears in many papers and books and only the result is reproduced here. The Kalman filter is described by the matrix filter

$$K(t) = P(t)\, H^T(t)\, R^{-1}(t) \quad , \tag{2.28}$$

where $P(t)$ is the solution of the following matrix differential equation:

$$\dot{P}(t) = F(t)\, P(t) + P(t)\, F^T(t) - P(t)\, H^T(t)\, R^{-1}(t)\, H(t)\, P(t) + G(t)\, Q(t)\, G^T(t) \tag{2.29}$$

with the initial conditions

$$P(t_o) = E[\underline{\tilde{x}}(t_o)\, \underline{\tilde{x}}^T(t_o)] = \mathrm{cov}[\underline{\tilde{x}}(t_o), \underline{\tilde{x}}(t_o)] \quad . \tag{2.30}$$

Figure 6 is the complete model of the optimum filter and the message model. Equation (2.29) is a matrix Riccati differential equation and is very difficult to solve.

Since the Kalman filter is the minimum variance unbiased estimator of the signal, $\underline{x}(t)$, given only the observation $\underline{z}(t)$, then anything less than complete a priori information [Equations (2.21), (2.22), (2.23)] quickly deteriorates system performance. Errors in Kalman filtering can arise from many sources. These errors include an incorrect model for the system dynamics and incorrect statistics describing the time history of the covariance of the plant noise, $Q(t)$, and an incorrect covariance matrix, $R(t)$, for the measurement noise. Any of the errors listed, plus any other statistical inaccuracies, result in a suboptimal Kalman gain. These errors not only produce an incorrect Kalman gain but they make it difficult to solve the matrix

Riccati differential equations which are sensitive to errors. The results of the Kalman filter will not be used directly but the mathematical minimization will be employed in the derivation of the proposed stochastic adaptive filter.

2.3.3 Implementation of Multichannel Wiener Filters The most obvious implementation of the Wiener filter uses the inversion of the $R_{\underline{xx}}$ matrix. This idea can be dismissed because if there are many sensors, a real time computation of the matrix inverse is not practical. Iterative techniques have distinct advantages over the direct matrix inversion in that storage and time delay for computation of $R_{\underline{xx}}^{-1}$ is not required, and the iterative algorithms result in systems where the optimum filter can be updated as the environment is sampled.

Baird (12) has implemented a matrix inversion routine that is part of an adaptive array processor. This matrix inversion lemma approximates the matrix inversion by a recursive algorithm. While this approximate matrix inversion takes longer than the basic gradient procedures, it is much faster than true matrix inversion and offers possibilities for real time use.

One could also implement a multi-input, single output Wiener filter in the frequency domain (Goode, 50). One could take the fast Fourier transform (FFT) of the input vector $\underline{x}(t)$ and multiply the result by the optimum filter vector $\underline{h}(\omega)$ defined by Equation (2.17). The Fourier coefficients obtained from the FFT are multipled by $[h_j(\omega)]_r$ and $[h_j(\omega)]_i$, the real and imaginary parts of the j-th weight, and the corresponding products for all the sensors are summed to form the beam output at frequency ω. In a real system there would

be a discrete number of values, N , for the frequency ω , and the total number of weights required to form a beam over the entire bandwidth of interest is 2KN . For a multiple output system, this approach requires a large amount of hardware.

The approach taken here is similar to the frequency domain approach described above in that it is also a frequency domain approach. However, there are many differences. One difference is the fact that the proposed system does not use the sensor outputs directly but instead uses preformed beams. The system is not only a multi-input but also a multi-output system.

Another approach is to adjust the weights of a tapped delay line multi-channel filter to minimize mean-square error. This time doamin approach of tapped delay lines is the application of the LMS algorithm (Widrow et al., 133). A tapped delay line filter (Figure 7) is used to approximate the general continuous time filter, with the approximation improving as the time delays become smaller and the number of taps increases. This system can increase in complexity very rapidly if it has to operate over a wide frequency range. Since the tapped delay line filter structure is the most common implementation of the Wiener filter in the adaptive literature, the following section will show the derivation of the optimum tapped delay filter and the approximations used to implement the filter structure in discrete form.

2.4 Tapped Delay Line Filters

2.4.1 Optimum Filters Using the criterion for a minimum from the regular calculus, we can optimize our performance criterion and find the optimum filter coefficients, $h_i (i = 1, 2, \ldots, KL)$. To find the

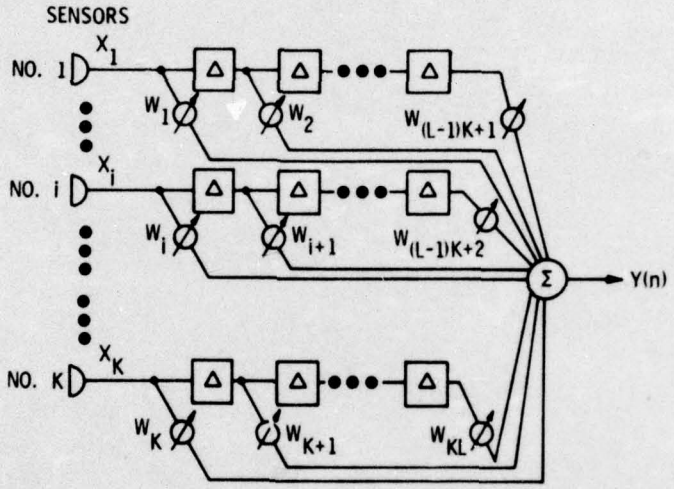


Figure 7. Tapped Delay Line Filter

optimum filter, the mean square error (MSE) performance criterion

$$E\{[d_n - y_n]^2\} = E\{\varepsilon^2\} \tag{2.31}$$

is used because of its mathematical tractability and its physical meaning in terms of power. If we take the expected value of the MSE [Equation (2.31)], differentiate the resulting expression with respect to the weights, and set the resulting expression to zero, we can obtain the optimum tapped delay line filter. Since

$$E\{\varepsilon^2\} = E\{d_n^2 - 2d_n y_n + y_n y_n^T\} \quad , \tag{2.32}$$

then, with $y = \underline{h}^T\underline{x}$, and $\underline{x} = \underline{s} + \underline{n}$, we get (dropping the sample number n)

$$E\{\varepsilon^2\} = E\{d^2 - 2d\,\underline{h}^T\underline{x} + (\underline{h}^T\underline{x})(\underline{h}^T\underline{x})^T\} \quad . \tag{2.33}$$

Taking the indicated expected values gives:

$$E\{\varepsilon^2\} = E\{d^2\} + \underline{h}^T R_{\underline{x}\underline{x}}\underline{h} - 2\underline{h}^T R_{\underline{x}d} \quad . \tag{2.34}$$

If we differentiate the above expression with respect to $\underline{h}$, we get

$$\nabla\{E(\varepsilon^2)\} = 2R_{\underline{x}\underline{x}}\underline{h} - 2r_{\underline{x}d} \quad . \tag{2.35}$$

The optimal solution occurs when the gradient is zero. Thus, we get

$$\underline{h}_{OPT} = R_{\underline{xx}}^{-1} \underline{r}_{\underline{x}d} \quad , \tag{2.36}$$

which is, in fact, the form of solution obtained from the Wiener-Hopf equation.

Let us recall that the optimum linear filter developed by Wiener has from Equation (2.17) the form

$$\underline{h}_O^*(\omega) = G_{\underline{xx}}^{-1}(\omega)\, \underline{g}_{\underline{x}d}(\omega) \quad . \tag{2.37}$$

A large body of mathematics is devoted to approximating a function by using a series of orthogonal functions. We can represent the desired optimum linear filter as a sum of the following form of orthogonal functions:

$$H(\omega) = \sum_{n=0}^{N} A_n \phi_n(x) \quad . \tag{2.38}$$

If we use a weighted mean square deviation as our closeness criterion and if we use an infinite number of terms in our approximation, then the desired filter can be represented with arbitrary accuracy. The polynomial approximation to the desired transfer function of the filter can then be converted into a rational fraction (for system use) using Padé approximants. This method is very sophisticated and extremely difficult to implement. In practice it is easier to find simple approximations to the desired transfer function.

2.4.2 Discrete Approximations Recall that one particular implementation of the Wiener filter is the tapped delay line filter. This filter consists of a tapped delay line with adjustable weights at each tap. For this implementation, the filter functions, ϕ_n , from Equation (2.38) then represent delayed versions of the input signal. In this case,

$$\underline{h}(\omega) = \sum_{k=0}^{\infty} h_k \exp(-jk\Delta) \quad , \tag{2.39}$$

where Δ is the time increment between delay line taps, and h_k is the weight at the kth tap.

A particular implementation of the tapped delay line filter is shown in Figure 7. For this system, the input signals are in discrete sampled data form.

The preceding section showed how the tapped delay line filter could be obtained as an approximation to the optimum Wiener filter. If we knew the statistical properties of both the desired signal and the noises, then it might be possible to determine the optimum linear Wiener filter. Usually only part of these statistics are known and the other statistics are unavailable. If we assume knowledge of either signals or noise, then it would be possible to develop an optimum system using these statistics. In most adaptive systems, either the statistics of the desired signal or the correlation functions of the desired signal are assumed known. Since part of the complete statistical information is unknown, adaptation is performed by adjusting the filter coefficients according to some performance measure. In all that follows, it is

assumed that the random functions involved in the adaptive processor are stationary random processes. The most common performance measure used is the minimum mean square error between the desired signal and the output of the filter. This performance criterion is used because it represents a tractable mathematical function.

The most common adaptive implementation of the tapped delay line filter is the Widrow LMS algorithm. This algorithm will be described in Section 2.5. The accuracy of the tapped delay line filter increases as the number of weights becomes large and the time difference between the taps becomes small.

The adaptive algorithm proposed in this research is the adaptive realization of the frequency domain version of the optimum linear Wiener filter. It is a much more practical implementation of the Wiener filter when the number of sensors (not beams) becomes large.

2.5 Adaptive Algorithms

The following section contains a summary of the adaptive processors reported in the literature. It contains a short account of the approximations, techniques of implementation, and varied use of adaptive processors.

This section has been written to put into perspective the relations between the complex frequency domain stochastic adaptive processor and the previous attempts at adaptive processors which were almost all time domain deterministic methods.

Urkowitz (124) considers the detection and estimation of a signal field in the presence of a noise field. He uses a Karhunen-Loève expansion, a generalization of Fourier Series, to obtain a series

representation with uncorrelated coefficients, of the random process of transducer outputs. If one looks at the complexity in the generation of the test statistic (likelihood ratio) used in this detection system, one can quickly grasp the fact that an implementation of the complete optimal system is extremely complicated.

One of the difficulties in implementing the general optimum array processor is the inversion of the $\underline{R_{xx}}$ matrix. For any size system, an on-line implementation of straightforward matrix inversion is impractical. Processors based on this direct computation procedure are poorly suited for on-line calculations, since they require tedious computations which must be repeated if the environment changes. To avoid the computation problems associated with the direct calculation of the optimal filter parameters, several adaptive procedures have been developed. These adaptive algorithm have the advantages that the required computations are generally much simpler and that they can be continually updated, thereby accounting for time variations in the environment.

Many authors have attempted to apply deterministic gradient minimization to derive the matrix filter that realized an adaptive antenna system in the time domain. Widrow's LMS algorithm adaptively formed patterns that placed nulls at the spatial location in the beam pattern of the noise sources. This adaptive processor performs filtering in space. By far, the largest class of adaptive array processors are the various forms (Daniell, 30; Frost, 47; Goode, 50; Griffiths, 52) of the Widrow LMS algorithm which uses the tapped delay line as their basic filter structure. All of these LMS algorithms, except Frost's,

use steepest descent to calculate the adaptive weight vector. There are many variations of the basic LMS algorithm and they only differ in their assumptions of known statistics. For the basic LMS algorithm of the form

$$\underline{w}_{n+1} = \underline{w}_n + \mu[d_n \underline{x}_n - \underline{x}_n^T \underline{x}_n \underline{w}_n] \quad , \tag{2.40}$$

it is assumed that the desired signal or pilot signal, $d(n)$, is known and an input cross correlation must be calculated. Griffiths (51) assumes knowledge of cross correlations between the observed signal vector, $\underline{x}_n$, and the target signal, d_n .

The gain constant μ not only determines the speed of convergence, but also the misadjustment or noisiness of the estimation process. All of the LMS type algorithms have commonality in the facts that they use the tapped delay line as the filter model and deterministic gradient type steepest descent to calculate the adaptive filter weights. The LMS type algorithms assume that the direction of the desired signal is known a priori and they use this knowledge to put specified gains on both the desired signal and the noises by adjusting the antenna or directivity pattern.

Frost's algorithm (47) belongs to the class of LMS algorithms but Frost uses a projected steepest descent algorithm to find the optimum filter. His algorithm requires knowledge of input cross correlation matrix and detailed knowledge of signal and noise geometry used in the formulation of the constraint which maintains a chosen frequency characteristic for the array in the direction of interest.

These characteristics, which are formulated in the form of a constraint, are very difficult to design for general classes of array problems.

R. L. Riegler and R. T. Compton (101) also used a steepest descent minimization of mean square error. In this system, no a priori information about the angles of arrival of signals was required. However, certain statistical characteristics of the desired signal had to be known. A reference signal replica of the desired signal must be available and the input cross correlation matrix is also needed to calculate the experimental mean square error. They found, as was found in this research, that minimizing mean square error is equivalent to maximizing S/N for all but very low input S/N. Their filter structure was a form of the tapped delay line filter used by all the Widrow LMS algorithms.

Zahm (139) extended the technique of power equalization developed by Riegler and Compton for a wider band of signals. The power equalization technique is based on proportional feedback control which equalized the power out of the array of sensing elements due to the desired signal and an interference source.

Schwartz and Winkler (113) use Rosen's projected gradient algorithm to design an iterative algorithm to minimize the mean square error or maximize the S/N subject to certain linear constraints. Their algorithm is very similar to Frost's algorithm in that both use projected gradient algorithms. The Winkler and Schwartz algorithm assumes knowledge of both a desired signal and the cross spectrum between the desired and received signals.

The side lobe canceller (Howells, 61) is a system which has a main beam and forms an auxiliary beam from elements distributed over the face of the array. Auxiliary beam signals are subtracted with right amplitude and phase from the main beam to cancel any interference not in the main lobe. It is known that in the case of non-zero bandwidth interference, the side lobe canceller does not give perfect cancellation because the auxiliary beam does not contain the exact replica of the main beam signal. This effect is more pronounced the larger the cancellation system's fractional bandwidth. To minimize these bandwidth effects, one must use multiple auxiliary beams and cancellation loops. The objective is to distribute auxiliary beam elements over the face of the array as much as possible in the same way as the elements of the main beam are distributed. Thus, the side lobe canceller cannot be used for cancelling wideband interference unless the number of cancellation loops is made very large.

Claerbout (25) designed a processor in which signal information is given in the form of various constraints on the filter coefficients rather than being given as a signal correlation function in the design of least-squares filters.

Kobayashi (70) used both the methods of steepest descent and conjugate gradients to design an adaptive filter based on the data taken over some fitting interval that minimizes the output noise power without distorting the signal. His algorithm requires knowledge of the input cross correlation matrix or cross power spectral matrix. He applied his algorithm to seismic processing system used to detect earthquakes.

Another method of calculating a set of weight values for a tapped-delay-line multichannel filter is called maximum-likelihood-ratio (MLR) processing. This method has been applied by Capon et al. (23) to the processing of large aperture seismic array (LASA) data. Maximum-likelihood-ratio processing assumes that the velocity and directional properties of the target signal are known a priori. From a knowledge of the spatial characteristics of the target, one is able to design spatial correction filters. The outputs of the spatial correction filters are then used as inputs to the tapped-delay line processor which determines the filter weights.

Brennan and Reed (19) developed a theory for an adaptive processor which maximizes the probability of detection for a fixed false alarm rate. Their derivation, however, was incorrect because they incorrectly took derivatives of non-analytic complex functions.

Muellar and Spaulding (90) used the gradient steepest descent algorithm with minimum mean square error as the performance measure to determine a technique for start-up of adaptive transversal filter (tapped delay lines) equalizers used in high speed synchronous data communications. They use a specially derived training sequence for the required pilot signal to obtain the data line synchronization. Their technique is called cyclic equalization.

Sondhi (117) used a technique similar to Muellar and Spaulding to obtain echo cancellation in long distance telephone communications. Both these systems used in communications problems use the tapped delay line filter as their filter structure.

It is true with all on-line implementations of estimation algorithms whether they are optimal implementations or adaptive realizations that each has its limitations. It has been shown by Comer (27) that the algorithms with constant gain, μ , have comparatively little noise resistance. Furthermore, in the presence of measuring error with variance σ^2 , convergence in the usual sense does not occur but instead we get;

$$\lim_{n\to\infty} E\{||\underline{w}_n - \underline{w}_o||^2\} < \mathrm{Var}(\mu_o, \sigma^2) \quad , \tag{2.41}$$

where

$$\mathrm{Var}(\mu_o, \sigma^2) \to 0 \quad \text{as} \quad \mu_o \to 0 \quad . \tag{2.42}$$

The only way to reduce the variance of the weights is to make the gain constant very small. This technique results in an inordinate amount of time to get convergence and makes the gradient techniques inapplicable for on-line use. The precise manner in which the gain in stochastic approximation is varied, which is exactly that suggested above, reduces the effect of additive measurement noise to the point where the variance of the process goes to zero.

As was noted before, the tapped delay line filter only approximates the continuous time filter. In a broadband system, there would have to be a large number of taps with small time delays between successive taps to get a good representation. The stochastic approximation algorithm in this research is potentially more powerful

than non-stochastic type algorithms because it has the statistical nature of the system built into the adaptive minimization problem and since it is specifically a frequency domain technique, it is designed to operate over a wide range of frequencies.

There are many other authors who have tried to realize adaptive filters. To try and give a description of each one's work would be impractical. A partial listing of the more important methods has been given here. The references contain a large survey of both the methods described and many others.

CHAPTER III

STOCHASTIC APPROXIMATION AND PROBABILISTIC PROOFS

3.0 Introduction

In this chapter, a brief summary of theorems and convergence criterion of both the Robbins-Monro and the Kiefer-Wolfowitz stochastic approximation algorithms is given. We show the evolution and weakening of the convergence conditions of stochastic approximation and how these techniques have been applied to many diverse areas. The importance of the variable gain sequence, μ_n , for both rapid convergence and statistical smoothness of the stochastic algorithm is emphasized. These are important reasons why the stochastic approximation algorithm is more useful than deterministic algorithms. The last part of Section 3.1 lists some of the many areas in which the classical stochastic approximation techniques have been applied.

We formally state the classical projection theorem and use a theorem formulated by Karlin and Taylor (66) to establish the necessary and sufficient conditions for both minimum mean square error and linear minimum mean square error. It is shown that linear estimators are just subsets of general minimum mean square error estimators and that the optimum linear estimator has a simple geometrical interpretation.

The analysis in Section 3.2.1 shows that the optimum linear solution to the minimum mean square error minimization can be derived from the orthogonal projection lemma. The non-analytic performance

measure (in the complex case) makes it impossible to derive the optimum filter by ordinary calculus. This fact gives the impetus to use some other means for the minimization and the orthogonal projection lemma is used for the minimization. The optimum matrix filter for minimum mean square error is derived from considerations of the necessary and sufficient conditions for minimization obtained from Theorem 3.6. This theorem is just the stochastic version of the classical (non-stochastic) orthogonal projection theorem (Theorem 3.3).

A derivation is done to obtain the regression function for the recursive stochastic approximation algorithm which is used to derive the matrix adaptive filter for the experimental tests. The various modes of probabilistic convergence are established and their interrelations are compared. It is shown why non-probabilistic convergence criteria make no sense when the underlying processes are stochastic processes.

We prove convergence of the stochastic adaptive algorithm under the convergence criterion of stochastic approximation and establish the important result that the sequence of solutions of the adaptive filter is a martingale. The fact that the sequence of solutions is a martingale not only establishes the convergence of the stochastic algorithm but also is used to show the stability of the stochastic adaptive algorithm in terms of a stochastic control system.

3.1 History of Stochastic Approximation Techniques

While the concepts involved in the use of stochastic approximation are relatively simple to understand, they are mathematically difficult. The convergence conditions of stochastic approximation are difficult to establish on a global basis and many authors have dealt with extensions

and modifications of the original Robbins-Monro (102) and Kiefer-Wolfwitz (68) processes and they have devised ways to adapt stochastic approximation techniques to diverse applications. The following section traces the history of the Robbins-Monro and Kiefer-Wolfwitz methods and shows the evolution of the convergence criterion and their extensions to a wider range of minimization or root finding problems than their original derivations.

3.1.1 Robbins-Monro Procedure Stochastic approximation had its beginnings in Robbins and Monro's classic paper. Robbins and Monro formulated their minimization problem and their convergence conditions by presenting the following theorem. The sequence μ_n is the same type of sequence used in this research.

Theorem 3.1 (Robbins and Monro, 102)

Let $r(x)$ be a given function and α a given constant such that the equation $r(x) = \alpha$ has a uniquely defined root $x = \theta$. Let $y(x)$ be a realization of a measurement. Assume $y(x)$ has distribution $P[y(x) \leq y] = g(y|x)$ such that $r(x) = \int_{-\infty}^{\infty} y\,dg(y|x)$ (i.e., $r(x) = E\{(y|x)\}$. Choose x_1 arbitrary and define the recursive relation: $x_{n+1} = x_n + \mu_n\{\alpha - y_n(x_n)\}$.

Let the sequence μ_n be of the form $1/n$ and assume there exists some constant $c > 0$ such that

$$P\{|y(x)| \leq c\} = 1 , \tag{3.1}$$

and that the conditions

A. $r(x)$ is a nondecreasing function ,

B. $r(\theta) = \alpha$,

C. $r'(\theta) > 0$; $r'(\theta) = \frac{d(r)}{d\theta}$

are satisfied. Then, the recursive relation,

$$x_{n+1} = x_n + \mu_n\{\alpha - y_n(x_n)\} \tag{3.2}$$

and the preceding assumptions, imply the result

$$\lim_{n\to\infty} E\{(x_n - \theta)^2\} = 0 \quad . \tag{3.3}$$

Wolfowitz (136) weakened the convergence conditions of the Robbins-Monro theorem so that if the regression function satisfied

$$|r(x)| \leq c \quad , \tag{3.4}$$

then the Robbins-Monro process x_n converges in probability to θ .

Blum (16) weakened the convergence conditions and required that the regression function lie between two lines.

Friedman (46) further weakened the convergence conditions. Friedman's theorem enables one to construct a convergence process when $|r(x)|$ and $\sigma^2(x)$ are bounded by known functions $f_1(x)$ and $f_2(x)$. One then takes

$$f(x) = \text{Max}[f_1(x) , f_2(x)]^{1/2} \tag{3.5}$$

and $r(x)$, the regression function, has only to satisfy

$$r(x) \leq [\ell|x| + k] \tag{3.6}$$

for positive constants ℓ and k . The Robbins-Monro process then becomes

$$x_{n+1} = x_n - \mu_n[\alpha - y_n(x_n)]/f(x_n) \quad . \tag{3.7}$$

Gladyshev (47) simplified the conditions for convergence and established proofs whereby the algorithm converged with probability one. It is on his theorem that most of the convergence proofs of stochastic approximation are based.

Blum (16) generalized the Robbins-Monro procedure to the multi-dimensional case. It is on a matrix version of Blum's multidimensional procedure that the method in this research is based.

3.1.2 Kiefer-Wolfowitz Kiefer-Wolfowitz suggested a method to estimate the maximum of a regression function where the regression function is not directly available. Since more measurements are necessary at each step to form some approximation to the regression function, the Kiefer-Wolfowitz process exhibits slower convergence properties than the Robbins-Monro process. They proposed the following theorem and the resultant convergence conditions. It is to be noted that the gain sequence, a_n/c_n , used in the method is analogous to μ_n used in the Robbins-Monro method.

Theorem 3.2 (Kiefer and Wolfowitz, 68)

Let $r(x)$ be a regression function and $F(y|x)$ a family of distribution functions and assume that the following conditions are satisfied:

$$\int_{-\infty}^{\infty} [y(x) - r(x)]^2 \, dF(y|x) \leq \sigma^2 < +\infty \quad . \tag{3.8}$$

Assume that $r(x)$ is strictly increasing for $x < \theta$, and that $r(x)$ is strictly decreasing for $x > \theta$.

Let $\{a_n\}$ and $\{c_n\}$ be infinite sequences of positive real numbers such that

$$\sum_{n=1}^{\infty} a_n = \infty \, , \quad \sum_{n=1}^{\infty} a_n c_n < \infty \, , \quad \sum_{n=1}^{\infty} a_n^2 c_n^{-2} < \infty \quad . \tag{3.9}$$

Then, the recursive scheme defined by

$$x_{n+1} = x_n + \frac{a_n}{c_n} [y(x_n + c_n) - y(x_n - c_n)] \tag{3.10}$$

converges in probability to the maximum, θ, of the regression function $r(x)$ if certain regularity conditions are satisfied.

Many authors, notably Venter (130), have weakened the convergence conditions for the Kiefer-Wolfowitz process.

One can construct a multidimensional Kiefer-Wolfowitz process by forming $2N$ observations of the random vector

$$\underline{y}_k \pm c_k \underline{e}_j \, , \qquad k = j = 1 \, , \ldots , n \tag{3.11}$$

where the e_j are unit vectors. This observation vector is used to evaluate the regression function in a recursive algorithm of the form

$$\underline{z}_{n+1} = \underline{z}_n + \frac{a_n}{c_n} \underline{r}(y_n) \quad . \tag{3.12}$$

The $\{a_n\}$ and $\{c_n\}$ are sequences of positive real numbers that satisfy

$$\sum_{n=1}^{\infty} a_n = \infty \ , \quad \sum_{n=1}^{\infty} a_n c_n < \infty \ , \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{a_n^2}{c_n^2} < \infty \quad . \tag{3.13}$$

Kiefer-Wolfowitz (68) made some highly restrictive assumptions from which they proved that Equation (3.12) converges with probability one. Venter (130) relaxed the original convergence criterion of Kiefer-Wolfowitz and made the technique applicable to a wider variety of cases.

Most experimenters in stochastic approximation have concluded that the method be carried out in two stages. The first stage would take large step sizes to estimate the most likely region of search, and the second stage would take smaller step sizes to fine tune the search. Since the general stochastic approximation algorithm is a local technique rather than a global one, this segregation of search stages is a logical procedure. There does not exist, however, either in the global case or the local case any general method for carrying out the search nor for adjusting the search steps in an optimum manner while still satisfying the convergence conditions.

Many authors have proposed specific ways to accelerate convergence. Kesten (65) proposed keeping the gain, μ_n, constant if the

difference $(x_n - x_{n-1})$ had the same sign as $(x_{n-1} - x_{n-2})$ and decreasing the step size otherwise. Fabian (43) proposed a method where one would make observations on the regression function in random directions and pick the direction where the sign of the regression function was the negative of the others. This method is very similar to the deterministic method of steepest descent.

The choice of the sequence μ_n is of major importance for the speed of convergence of the stochastic algorithm and for the statistical smoothness of the process. No general procedure exists for choosing this sequence. The rapid speed of convergence for the proposed adaptive filter has been obtained by considering the least mean square error as a control system, and then deriving the sequence which allows the algorithm to operate on the stability limit of the system. This starting value gives rapid improvement at the expense of initial noisy results. However, the further application of the infinite sequence smooths the resulting matrix filter coefficients because of the built in statistical smoothing function of the stochastic approximation algorithm. This smoothing is not possible with deterministic systems.

Kushner (76) has recently attempted to formulate a general theory for optimum search strategies for stochastic approximation. He has also tried to formally incorporate linear constraints in the solution of the general optimization problem solved by stochastic approximation and in the search strategy.

Mendel and Fu (84) discuss a method to accelerate convergence whereby they use the orthogonal projection of the current estimate of the random variable to force the estimate to be within a cube. They use

knowledge of the convergence region of some of the random variables being estimated to speed the convergence of the total procedure. This method of orthogonal projection discussed by Mendel and Fu is based on the same physical considerations as the method used in the matrix adaptive filter in this research.

Stochastic approximation has been used in a wide variety of applications. Many authors in pattern recognition (Fu, 46), adaptive automatic control systems (Holmes, 59), operations research (Albert and Gardner, 1), and biological research (Cochron and Davis, 26) have applied the methods of stochastic approximation. Problems in adaptive control systems fall into the stochastic approximation framework and yield computationally simple algorithms which require little storage and can be performed in real time. Stochastic approximation has been used to estimate probability density functions for application not only in statistics but also in communication theory.

Chien and Fu (24) have used stochastic approximation to deal with state estimation in stationary dynamical systems where only some state components are accessible for measurement.

Dupac (40) has extended the basic Robbins-Monro procedure so that it can be used to estimate parameters which are slowly time varying, and are varying in either a deterministic manner or in a random fashion with their variances (time varying) going to zero as time tends to infinity. This extension has use in auto regressive and moving average processes not only in statistics but also in any system that needs reliable estimates of the mean of a process. One fact that can be gleaned from this short history of stochastic approximation is that the

applications of the technique have been varied. In the sections that follow, stochastic approximation will be used to derive a recursive relation which will give a stochastic adaptive realization of multi-input, multi-output Wiener filter to be used as the matrix filter in an adaptive array processor.

The following section formulates the general minimum mean square estimation problem, and establishes a theorem whereby the linear minimum mean square minimization can be solved. We formally state the definition of a regression function to be used in all the succeeding analysis and compare the relationship of the various modes of probabilistic convergence.

Theorem 3.3 is a formal statement of the classical projection theorem. It should be contrasted with Theorem 3.4 which is the stochastic version of the classical theorem. The stochastic version provides the basis for finding the minimum mean square error and the analytic method whereby the optimum matrix adaptive filter and the regression function are derived.

Theorem 3.3 (Luenberger, 79)

Let S be a complete metric space. Let T be a closed vector space which is a subspace of S. Let $\underline{x}$ be a vector that is in T but not necessarily in S. If we let $\underline{u}$ be any vector in T, then there exists a unique $\underline{u}_o$ such that $||\underline{x} - \underline{u}_o|| \leq ||\underline{x} - \underline{u}||$ for all $\underline{u}$ in T. From geometric considerations, the necessary and sufficient condition that u_o must satisfy to fulfill the above conditions is that the vector $(\underline{x} - \underline{u}_o)$ be orthogonal to the subspace T.

3.2 Mathematical Preliminaries

3.2.1 Minimum Mean Square Error Estimation Since we are interested in estimating the value of a random variable resulting from the observation of an experiment, we now formulate the mean square error estimation problem (Karlin and Taylor, 66). Since stationary random processes are assumed in this research, the estimation problem is considered in this context. From this formulation will evolve a theorem which gives the necessary and sufficient conditions for linear minimum mean square error which is used in the derivation of the stochastic adaptive filter.

If we are concerned with the problem of estimating the random variable x from an observation of past and present values of the random variable y, then the problem is called a prediction problem. The prediction problem involves construction of certain types of estimators. The physical problem that concerns us in this research is more precisely called a filtering problem because it consists of estimating a random variable based on a observation of a process containing both signal and noise. However, the estimator designed uses both past and present values and is used to predict future values. It is for these reasons that the terms describing the estimator will be used interchangeably.

We want to estimate the best value for x, the result of some observation. We call $\hat{x}$ our estimation and we desire to make our estimation error $(x - \hat{x})$ as small as possible in some mathematical sense. Since the experimental observations are random processes, a good probabilistic measure of performance for our estimator is the mean

squared error

$$E\{(x - \hat{x})^2\} = E\{\varepsilon^2\} \quad . \tag{3.14}$$

It is to be observed that this performance measure is an average measure and sometimes we will be correct, and sometimes we will be wrong in our estimations. We will minimize our error on the average.

Thus, we want to find an estimator that minimizes over all estimators, $\hat{x}$, (both linear and nonlinear), the mean square error $E\{(x - \hat{x})^2\}$. In most systems, and adaptive systems in particular, the most important design criterion is the mathematical tractability of the performance criterion that is chosen. The mean square error criterion is not only mathematically tractable but it possesses the realistic physical design criterion that the error is minimized in the power sense. Wiener (134) made a major contribution to the engineering literature when he used the minimum mean square error criterion to find his optimal estimator, the Wiener filter. The ultimate result of this research is a recursive algorithm to find the Wiener filter and the results of Theorem 3.6 are the cornerstones of the derivation of the adaptive algorithm.

The mean square error is defined as the error over a band of frequencies. Since each term of the mean square error at every discrete frequency in the band is positive or zero, then each term can be minimized separately. In this research the minimum mean square error is found at each frequency in the band of interest. Since each term is

positive or zero, we can guarantee that the mean square error is minimum over the entire band of interest.

The following analysis (Karlin and Taylor, 66) illustrates the fact that if linear estimators, as realistic approximations to true minimum mean square error estimators, are desired, then some degradation, except for Gaussian statistics, of performance from the true minimum mean square error estimation must be expected.

If we assume that y is the result of an experiment from which we wish to estimate a value of x, then for the jointly distributed random variables x and y we allow any estimator having finite variance.

In the case where the outcome of an experiment, x, has a known mean, μ, the best predictor for x in the mean square error sense is $\hat{x} = \mu$.

We could now argue that in the general case the appropriate distribution for our estimator is the mean of x computed from the conditional distribution, $\mu_{x|y} = E\{x|y\}$. This can be shown to be true by the following example (Karlin and Taylor, 66). We compute the mean square error as

$$E\{(x - \hat{x})^2\} = E\{(x - \mu_{x|y})^2\} + 2E\{(x - \mu_{x|y})(\mu_{x|y} - \hat{x})\} + E\{(\mu_{x|y} - \hat{x})^2\} . \quad (3.15)$$

The second term in Equation (3.15) can be shown to be zero by evaluating the conditional expectations indicated. Expanding the second term by use of identities from Appendix C, we get

$$E\{(x - \mu_{x|y})(\mu_{x|y} - \hat{x}) = E\{E[(x - \mu_{x|y})(\mu_{x|y} - \hat{x})|y]\}$$

$$= E\{(\mu_{x|y} - \hat{x})\, E[(x - \mu_{x|y})|y]\}$$

$$= 0 \quad . \tag{3.16}$$

Thus, Equation (3.15) becomes

$$E\{(x - \hat{x})^2\} = E\{x - \mu_{x|y}\}^2 + E\{\mu_{x|y} - \hat{x}\}^2 \quad , \tag{3.17}$$

and the right-hand side is minimized by setting $\hat{x} = E\{x|y\}$.

In practical situations, the optimal estimator, $\hat{x} = E\{x|y\}$, is almost never known. Linear estimators, while only approximations to true minimum mean square error estimators, lead to easily implemented estimators. Since mathematical tractability and ease of implementation are the major concerns in this research, all the succeeding analysis concerns linear estimators. We allow as estimators only those estimators that are linear functions of the random variable, y . Since this class of allowable estimators is smaller than in the general minimum mean square estimation, the resulting minimum mean square estimation error is at best equal to but generally greater than true minimum mean square error.

A useful result of minimum mean square error estimation (Rhodes, 100) is that the estimation error $\tilde{x} \triangleq x - \hat{x}$ in the least mean square estimator $\hat{x} = E\{x|y\}$ is uncorrelated with any function g of the random vector y , i.e.,

$$E\{g(y)\tilde{x}\} = 0 \tag{3.18}$$

and

$$E\{g(y)\tilde{x}|y\} = 0 \quad . \tag{3.19}$$

The following theorem provides both the necessary and sufficient conditions for the linear minimum mean square error and the theorem shows that the linear mean square error estimator is unique. Proofs of each of the assumptions and assertions of this theorem appear in Karlin and Taylor (66), Van Trees (124), and Feller (44).

Theorem 3.4 (Karlin and Taylor, 66)

Let x satisfy the relation $E\{x^2\} < \infty$ and assume that all estimators $\hat{x}$ satisfy the same relation. We will also allow as estimators any linear combination of estimators, i.e. if $\hat{x}_1$ and $\hat{x}_2$ are estimators then $a\hat{x}_1 + b\hat{x}_2$ is also a estimator for any real a and b .

A. A estimator $\hat{x}$ has minimum mean square error if and only if $E\{(x - \hat{x})u\} = 0$ for every estimator u .

B. If we assume the existence of minimum mean square estimators $\hat{x}_1$ and $\hat{x}_2$, then the minimum mean square error estimator is unique in the sense that $E\{(\hat{x}_1 - \hat{x}_2)^2\} = 0$.

The following definitions and theorems contain mathematical concepts necessary in the proofs of convergence of the stochastic approximation algorithm. Other results from conditional probability

theory used in succeeding proofs appear in Appendix C.

Definition 3.1

If we assume that $\{x_n , n = 1,2, \ldots\}$ and $\{y_n , n = 1,2, \ldots\}$ are stochastic processes and that $\{x_n\}$ satisfies

$$E\{|x_n|\} < \infty \tag{3.20}$$

and

$$E\{x_{n+1} \mid y_1 , \ldots y_n\} = x_n \tag{3.21}$$

then $\{x_n\}$ is a martingale with respect to $\{y_n\}$.

From a practical point of view, the stochastic process $\{y_n\}$ can be considered as the available information up to time n. From real analysis $\{x_n\}$ can be considered a function of $\{y_n\}$.

Theorem 3.5

Let $\{x_n\}$ be a martingale with respect to $\{y_n\}$ satisfying

$$E\{x_n^2\} < \infty . \tag{3.22}$$

Then, $\{x_n\}$ converges to a limit both with probability one and in the mean square, i.e.,

$$Pr\{\lim_{n\to\infty} [x_n = x_\infty]\} = 1 \tag{3.23}$$

and

$$\lim_{n\to\infty} E\{\left|x_n - x_\infty\right|^2\} = 0 \quad . \tag{3.24}$$

Definition 3.2

A wide-sense stationary process is a stochastic process $\{x_n\ ,\ n\epsilon N\}$ having finite second moments, $E\{x_n^2\} < \infty$, a constant mean, $m = E\{x_n\}$, and a covariance function, $E\{[x_{n_1} - m][x_{n_2} - m]\}$, that depends only on the time difference $\left|n_1 - n_2\right|$.

The preceding definitions and theorems along with the following property of conditional expectation

$$E\{f(y_1, \ldots y_n)\left|y_1, \ldots y_n\right.\} = f(y_1, \ldots y_n) \tag{3.25}$$

are the cornerstones for the proofs involving the stochastic approximation algorithm.

3.2.2 Regression Function and Probabilistic Convergence We want to estimate the random variable, y, by a suitable function $g(x)$ of x so that the mean square estimation error

$$E\{[y - r(x)]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - r(x)]^2 f(x,y)\, dx\, dy \tag{3.26}$$

is minimum. The function $g(x)$ that minimizes the above expression is, from the previous section, the conditional expected value of y

assuming x i.e.,

$$r(x) = E\{y|x\} \quad . \tag{3.27}$$

The function

$$r(x) = E\{y|x\} \tag{3.28}$$

is known as a regression function. This definition of a regression function is used in all the analysis that follows and specifically in the stochastic approximation algorithm.

Since the value of the gradient or its counterpart, the regression function, is not known exactly but depends on some stochastic process, we cannot use the ordinary concepts of convergence for non-random processes. The concepts of convergence must be redefined to encompass the stochastic nature of the recursive algorithm.

There are three types of stochastic convergence (Cramer and Ledbetter, 29): convergence in probability, mean square convergence, and convergence with probability one. A vector $\underline{h}_n$ converges in probability to $\underline{h}_{OPT}$ as $n \to \infty$ if for any $\varepsilon > 0$, the probability that the norm $||\underline{h}_n - \underline{h}_{OPT}||$ does not exceed ε converges to zero, i.e.

$$\lim_{n\to\infty} \text{PROB}\{||\underline{h}_n - \underline{h}_{OPT}|| > \varepsilon\} = 0 \quad . \tag{3.29}$$

Convergence in probability does not imply that every sequence $\underline{h}_n$ of

random vectors converges to $\underline{h}_{OPT}$ in the ordinary sense. A random vector $\underline{h}_n$ converges to $\underline{h}_{OPT}$ in the mean square sense as $n \to \infty$ if the mathematical expectation of the square of the norm $||\underline{h}_n - \underline{h}_0||$ converges to zero

$$\lim_{n\to\infty} E\{||\underline{h}_n - \underline{h}_{OPT}||^2\} = 0 \quad . \tag{3.30}$$

Convergence in the mean square implies convergence in probability but it does not imply ordinary convergence for any random vector $\underline{h}_n$. Convergence in the mean-square is related to the investigations of the moments of the second order. The above two types of convergence may not be satisfactory since in both types of convergence, the probability that a given vector $\underline{h}_n$ converges to $\underline{h}_{OPT}$ in an ordinary sense is zero. Since, in the recursive algorithm we only have available a sample or noisy regression function (or gradient), it is desirable that the limit exists for that particular sequence of random vectors $\underline{h}_n$ which is actually observed, and not for a family of random sequences which may never be observed. This type of convergence can be assured if we introduce the concept of convergence almost everywhere (a.e.) or convergence with probability one. Since $\underline{h}_n$ is a stochastic process, we can consider the convergence of a sequence $\underline{h}_n$ to $\underline{h}_{OPT}$ as a random event. The sequence of random vectors $\underline{h}_n$ converges to $\underline{h}_{OPT}$ as $n \to \infty$ almost certainly or with probability one, if the probability of ordinary convergence of $\underline{h}_n$ to $\underline{h}_{OPT}$ is equal to one

$$\text{PROB}\{\lim_{n\to\infty} ||\underline{h}_n - \underline{h}_{OPT}||^2 = 0\} = 1 \quad . \tag{3.31}$$

Thus, by neglecting the set of sequences of random vectors with total probability equal to zero, we have an ordinary convergence.

The great power of the method of stochastic approximation lies in the fact that, if the conditions of convergence are satisfied, then the random vector $\underline{h}_n$ converges with probability one to $\underline{h}_{OPT}$.

The following section contains the general convergence proof for a recursive algorithm (vector case) using stochastic approximation and it establishes that the sequence of solutions is a martingale. It will be seen throughout this analysis that the fact that the recursive algorithm is a martingale leads to many advantageous properties. The matrix case of the recursive algorithm is a simple extension of the vector case. We can define a single output as

$$y^i = (\underline{h}^i)^T \underline{x} \quad , \tag{3.32}$$

where $\underline{x}$ is the input vector and the $\underline{h}^i$ are the columns of the matrix filter. For the multi-output case, we get

$$\underline{y} = [\underline{h}^1 \; \underline{h}^2 \; . \; . \; . \; \underline{h}^n]^T \; \underline{x}$$

$$\underline{y} = H^T \underline{x} \tag{3.33}$$

where the columns of the matrix filter are the results of the vector case.

After the proof of convergence of the general recursive algorithm has been established, we show the statistical and geometrical interpretations involved in this convergence theorem. These geometrical

interpretations are important in understanding the intuitive arguments employed in the analysis of the complex matrix adaptive filter.

3.3 Convergence Proof of Stochastic Approximation

Theorem 3.6 and the succeeding proof apply to any recursive algorithm that satisfies the conditions of this theorm.

It is proposed that the matrix algorithm of the form

$$H_{n+1} = H_n - \mu_n R(\underline{x}_n | H_n) \quad , \tag{3.34}$$

where $R(x|H)$ is the regression function and μ_n is the special gain sequence, or the vector form of the algorithm

$$\underline{h}_{n+1} = \underline{h}_n - \mu_n \underline{r}(\underline{x}_n | \underline{h}_n) \tag{3.35}$$

be used as the recursive algorithm to find the matrix filter weights. It is shown in this section that if any algorithm of the form of Equation (3.35) satisfies the convergence conditions of stochastic approximation theory, then the recursive algorithm will converge to $\underline{h}_{OPT}$ with probability one. The matrix form [Equation (3.34)] follows from the vector form [Equation (3.35)]. In Equation (3.34), $R(\underline{x}|H)$ can either be the regression function derived from the orthogonal projection lemma or the gradient of the performance measure. Certain restrictions will be placed on the regression function to guarantee convergence and it will be shown that the adaptive algorithm (vector case) satisfies these conditions.

Theorem 3.6

Let μ_n be a sequence of numbers such that,

$$\text{A1.} \quad \mu_n > 0 \tag{3.36}$$

$$\text{A2.} \quad \sum_{n=1}^{\infty} \mu_n = \infty \tag{3.37}$$

$$\text{A3.} \quad \sum_{n=1}^{\infty} \mu_n^2 < \infty \ . \tag{3.38}$$

Let the following conditions be satisfied,

$$\text{B.} \quad \inf_{\varepsilon < ||\underline{h} - \underline{h}_{OPT}|| \leq \frac{1}{\varepsilon}} E\{(\underline{h} - \underline{h}_{OPT})^T \underline{r}(\underline{x}|\underline{h})\} \geq 0 \qquad \varepsilon > 0 \tag{3.39}$$

$$\text{C.} \quad E\{\underline{r}(\underline{x}|\underline{h})^T \underline{r}(\underline{x}|\underline{h})\} \leq d(1 + ||\underline{h} - \underline{h}_0||^2) \tag{3.40}$$

for all $\underline{h}$ in a bounded set and $d>0$. If the preceding conditions (A1, A2, A3, B, C) are satisfied then the sequence $\underline{h}_n$ defined by Equation (3.35) converges with probability one and in the mean square to the root of the regression function.

It is to be noted that where derivatives exist, the regression function can be replaced by the gradient. It will be simpler in this proof if the existence of the gradient is assumed. In any proofs that follow, the gradient ∇G will be equivalent to the regression function $\underline{r}(\underline{x}|\underline{h})$ and they will be used interchangeably.

The preceding theorem contains the sufficient conditions for convergence of any stochastic algorithm. If these convergence

conditions are satisfied, then the recursive algorithm [Equation (3.35)] can be guaranteed to converge to the optimum solution with probability one (Gladyshev, 49). The mean square convergence is not formally established but it can be gleaned from the analysis.

The following proof, patterned after that of Gladyshev (49), shows that the stochastic algorithm (vector case) converges with probability one and in the mean square to the optimal solution if it satisfies the previous conditions (A1, A2, A3, B, and C). This proof applies to both the real and complex variable cases [if the hermitian transpose $(\cdot)^H$ is used instead of the regular transpose $(\cdot)^T$ and the regression function is used instead of the gradient].

PROOF

Subtracting both sides of Equation (3.35) by $\underline{h}_{OPT}$, gives

$$\underline{h}_{n+1} - \underline{h}_{OPT} = \underline{h}_n - h_{OPT} - \mu_n \nabla G_n \quad . \tag{3.41}$$

where $\underline{r}(\underline{x}|\underline{h}) = \nabla G_n$ and $\underline{h}_{OPT} = \underline{h}_0$. Squaring Equation (3.41), gives

$$(\underline{h}_{n+1} - \underline{h}_0)^T(\underline{h}_{n+1} - \underline{h}_0) = (\underline{h}_n - \underline{h}_0)^T(\underline{h}_n - \underline{h}_0) - 2\mu_n(\underline{h}_n - \underline{h}_0)^T\nabla G_n + \mu_n^2 \nabla G_n^T \nabla G_n \quad . \tag{3.42}$$

Taking the mathematical expectation for a given $\underline{h}_1, \underline{h}_2 \ldots, \underline{h}_n$, we get:

$$E\{||\underline{h}_{n+1} - h_0||^2 \mid \underline{h}_1, \underline{h}_2 \ldots \underline{h}_n\} = ||\underline{h}_n - \underline{h}_0||^2 - 2\mu_n E\{(\underline{h}_n - \underline{h}_0)^T \nabla G_n\} + \mu_n^2 E\{\nabla G_n^T \nabla G_n\} \quad . \tag{3.43}$$

where $||\underline{h}_{n+1} - \underline{h}_0||^2 = (\underline{h}_{n+1} - \underline{h}_0)^T(\underline{h}_{n+1} - \underline{h}_0)$. From Condition C, Equation (3.43) becomes:

$$E\{||\underline{h}_{n+1} - \underline{h}_0||^2 \mid \underline{h}_1, \underline{h}_2 \cdots \underline{h}_n\} \leq$$

$$||\underline{h}_n - \underline{h}_0||^2 - 2\mu_n E\{(\underline{h}_n - \underline{h}_0)^T \nabla G_n + \mu_n^2 d[1 + ||H - H_0||^2] \quad (3.44)$$

Using Condition B, Equation (3.44) is reduced to

$$E\{||\underline{h}_{n+1} - \underline{h}_0||^2 \mid \underline{h}_1, \underline{h}_2 \cdots \underline{h}_n\} \leq$$

$$||\underline{h}_n - \underline{h}_0||^2 (1 + \mu_n^2 d) + d\mu_n^2 \quad . \quad (3.45)$$

If we define

$$\beta_n =$$

$$||\underline{h}_n - \underline{h}_0||^2 \prod_{k=n}^{\infty} (1 + \mu_k^2 d) + \sum_{k=n}^{\infty} 2d\mu_k^2 \prod_{m=k+1}^{\infty} (1 + \mu_m d) \quad (3.46)$$

then

$$\beta_{n+1} =$$

$$||\underline{h}_{n+1} - \underline{h}_0||^2 \prod_{k=n+1}^{\infty} (1 + \mu_k^2 d) + \sum_{k=n+1}^{\infty} 2d\mu_k^2 \prod_{m=k+1}^{\infty} (1 + \mu_m^2 d) \quad (3.47)$$

Taking the conditional expectation for given $\underline{h}_1, \underline{h}_2 \cdots \underline{h}_n$, and using Equations (3.45), (3.46), (3.47), we get:

$$E\{\beta_{n+1} \mid \underline{h}_1, \underline{h}_2 \cdots \underline{h}_n\} = E\{||\underline{h}_{n+1} - \underline{h}_0||^2 \mid \underline{h}_1, \underline{h}_2 \cdots \underline{h}_n\}$$

$$\prod_{k=n+1}^{\infty} (1 + \mu_k^2 d) + \sum_{k=n+1}^{\infty} 2d\mu_k^2 \prod_{m=k+1}^{\infty} (1 + \mu_k^2 d) \leq$$

$$[||\underline{h}_n - \underline{h}_0||^2 (1 + d\mu_n^2) + 2\mu_n^2 d] \prod_{k=n+1}^{\infty} (1 + \mu_k^2 d)$$

$$+ \sum_{k=n+1}^{\infty} 2d\mu_k^2 \prod_{m=k+1}^{\infty} (1 + \mu_m^2 d) = \beta_n \qquad (3.48)$$

or, more compactly,

$$E\{\beta_{n+1} \mid \underline{h}_1, \underline{h}_2 \cdots \underline{h}_n\} \leq \beta_n \ . \qquad (3.49)$$

Since $\beta_n = f(\underline{h}_1, \underline{h}_2 \cdots \underline{h}_n)$, then if we take conditional expectation for given $\beta_1, \beta_2 \cdots \beta_n$ of both sides of Equation (3.49), and using definitions from the conditional expectation identities in Appendix C, we obtain

$$E\{\beta_{n+1} \mid \beta_1, \beta_2 \cdots \beta_n\} \leq \beta_n \ . \qquad (3.50)$$

The inequality in Equation (3.49) shows that the β_n's are a semimartingale, where:

$$E\{\beta_{n+1}\} \leq E\{\beta_n\} \leq \cdots \leq E\{\beta_1\} < \infty \ , \qquad (3.51)$$

so that according to the theory of martingales (Doob, 38) the sequence β_n converges with probability one, and by virtue of Equations (3.45)

and (3.46), the sequence $(\underline{h}_n - \underline{h}_0)$ also converges with probability one to some random number ξ. The fact that the sequence of filters is a martingale will be used not only to prove convergence of the recursive algorithm [Equation (3.34)] which is used to calculate the matrix filter but also to show that the stochastic control system, represented by the recursive algorithm, is stable in the sense of stochastic stability. A list of martingale properties appears in Appendix C.

To complete the proof of the theorem, it remains to be shown that PROB $(\xi = 0) = 1$.

It can be seen from Equations (3.51) and (3.45) and the fact that $\mu_n > 0$ that the sequence $E\{||\underline{h}_n - \underline{h}_0||^2\}$ is bounded. Let us take the mathematical expectation on both sides of inequality Equation (3.44):

$$E\{||\underline{h}_{n+1} - \underline{h}_0||^2\} \leq E\{||\underline{h}_n - \underline{h}_0||^2\} - 2\mu_n E(\underline{h}_n - \underline{h}_0)^T \nabla G_n + \mu_n^2 d[1 + E||\underline{h} - \underline{h}_0||^2] \quad . \tag{3.52}$$

Adding the first n inequalities together, we have by deduction:

$$E\{||\underline{h}_{n+1} - \underline{h}_0||^2\} \leq E\{||\underline{h}_1 - \underline{h}_0||^2\} + \sum_{k=1}^{n} d\mu_k^2 (1 + E||\underline{h}_k - \underline{h}_0||^2) - \sum_{k=1}^{n} 2\mu_k E\{(\underline{h}_k - \underline{h}_0)^T \nabla G_n\} \quad . \tag{3.53}$$

Since $E\{||\underline{h}_n - \underline{h}_0||^2\}$ is bounded and Condition A3 is fulfilled, using Equation (3.52), it follows that:

$$\sum_{k=1}^{\infty} \mu_k E\{(\underline{h}_k - \underline{h}_0)^T \nabla G_n\} < \infty \quad . \tag{3.54}$$

Using Condition A2 and noting that from Condition B,

$$\inf_{\varepsilon<||\underline{h} - \underline{h}_0||\leq\frac{1}{\varepsilon}} E\{(\underline{h} - \underline{h}_0)^T \nabla G\} \geq 0 \quad , \tag{3.55}$$

we deduce from Equation (3.54) that

$$\{(\underline{h}_n - \underline{h}_0)^T \nabla G_n\} \rightarrow 0 \quad \text{with probability one for some sequence } n_N \ .$$

Using the fact that $(\underline{h}_n - \underline{h}_0)^T \nabla G_n \rightarrow 0$ with probability one, the fact that the regression function has a unique root and the fact that $(\underline{h}_n - \underline{h}_0) \rightarrow \xi$ with probability one from the first half of the proof, we conclude that $\xi = 0$ with probability one. Therefore, the algorithm

$$\underline{h}_{n+1} = \underline{h}_n - \mu_n \underline{r}_n(\underline{x}|\underline{h}) \quad \forall \quad \underline{r}_n(\underline{x}|\underline{h}) = \nabla G_n \tag{3.56}$$

converges with probability one,

$$\text{PROB}\{\lim_{n\rightarrow\infty} (\underline{h}_n - \underline{h}_0) = 0\} = 1 \tag{3.57}$$

as well as in the mean square sense,

$$\lim_{n\rightarrow\infty} E\{||\underline{h}_n - \underline{h}_0||^2\} = 0 \ . \tag{3.58}$$

3.4 Statistical and Geometrical Significance of the Convergence Criterion

Recall from the proof of the recursion relation for the adaptive filter the conditions (A1, A2, A3, B, C) imposed on the properties of the gain sequence, μ_n, as well as on the behavior of the regression function $\underline{r}(\underline{x}|\underline{h})$. These conditions not only guarantee convergence of the recursive algorithm with both probability one and in the mean square sense but boasts important statistical and geometrical concepts.

The condition that $\mu_n > 0$ is to assure that the corrections, on the average, are to be made in the direction toward the minimum.

The condition that $\sum_{n=1}^{\infty} \mu_n^2 < \infty$ is to account for the accumulative effect of the error in measurement. If random noise enters the measurement at each iteration step, then this condition assures that the random measurement error approaches zero as the number of iterations becomes large. This condition also implies the condition that

$$\lim_{n\to\infty} \mu_n \to 0 \quad . \tag{3.59}$$

We can see that if we let μ_n approach zero in this manner then we can attain arbitrary accuracy in our movement to the minimum of the regression function.

The above conditions assure that $\underline{h}_n$ converges on some value $\underline{h}_\infty$. The condition $\sum_{n=1}^{\infty} \mu_n \to \infty$ assures that value, $\underline{h}_\infty$, converges to $\underline{h}_{OPT}$. If $\underline{h}_n$ approaches any value other than $\underline{h}_{OPT}$, the total correction effect $\sum_{n=1}^{\infty} \mu_n \underline{r}_n(\underline{x}|\underline{h})$ is infinite.

The above conditions guarantee that if the sequence μ_n satisfies conditions A1, A2 and A3, the total correction effect of the

sequence μ_n allows the recursive algorithm to approach $\underline{h}_{OPT}$ with infinite correction, and as μ_n gets smaller we can approach the minimum with arbitrary accuracy.

The condition that

$$\inf_{\varepsilon<||\underline{h} - \underline{h}_0||<\frac{1}{\varepsilon}} E\{(\underline{h} - \underline{h}_0)^T \underline{r}(\underline{x}|\underline{h})\} \geq 0 \tag{3.60}$$

for $\varepsilon > 0$ determines the behavior of the surface $E_{\underline{x}}\{\underline{r}(\underline{x}|\underline{h})\} = 0$ close to the zero of the regression function. If the error criterion does have a unique minimum, the above condition is satisfied.

The condition that

$$E\{\underline{r}^T(\underline{x}|\underline{h})\ \underline{r}(\underline{x}|\underline{h})\} \leq d(1 + ||\underline{h} - \underline{h}_0||^2) \tag{3.61}$$

for $d > 0$ requires that the mathematical expectation of the quadratic forms

$$E\{\underline{r}^T(\underline{x}|\underline{h})\ \underline{r}(\underline{x}|\underline{h})\} \tag{3.62}$$

increase, as $\underline{h}$ increases, no faster than a quadratic function of the weights. This condition guarantees that the variance of the regression function is bounded.

CHAPTER IV

ADAPTIVE FILTER DERIVATION

4.0 Introduction

Using the results from the previous chapter, the optimum matrix filter is derived and its convergence established. The sequence of proofs that show the adaptive algorithm satisfies the convergence conditions of stochastic approximation appear in Section 4.3. While the satisfaction of the convergence conditions guarantees convergence in the mathematical sense, the main thrust of this research has been to devise a system that obtains fast initial (in less than 50 iterations of the recursive algorithm) increase in S/N (fast initial convergence). The satisfaction of the convergence conditions of stochastic approximation then provides the basis to expect ultimate convergence if the recursive algorithm were run long enough. This fact, although mathematically elegant, is not very useful for practical applications of the adaptive processor.

We show that under the weak signal strong interference assumption, the present adaptive algorithm solution corresponds to the diagonalization of the cross spectral density matrix. The proof of convergence of the adaptive algorithm is done for both the real and complex forms. It is shown rigorously that for the real variable case, the convergence conditions for stochastic approximation are satisfied. In the complex case, the five convergence conditions can be shown to be satisfied by

mathematical analysis. The satisfaction of the fifth condition is also clear from geometrical and intuitive reasoning.

In the final section, an expression is derived for the minimum mean square error and it is shown that in the case of the matrix filter, a decomposition of the mean square error into the contributions from each beam is more meaningful than considering the total mean square error due to all beams.

4.1 Derivation of Complex Matrix Filter

In order to avoid embarrassing difficulties in differentiation of complex functions, the orthogonal projection lemma is used for obtaining expressions for linear approximation to minimum mean square estimation of random processes. The justification for the use of the orthogonal projection lemma comes from Theorem 3.4. In this theorem the basic fact used is that the necessary and sufficient condition for the optimum linear filter for minimum mean square error is the satisfaction of the orthogonal projection lemma.

The following development obtains the frequency domain matrix Wiener-Hopf equation for a multi-input, multi-output system by use of the orthogonality conditions for stochastic processes. From the orthogonality condition of Theorem 3.4, the optimum linear matrix filter can be derived. From considerations of the optimum matrix filter and the orthogonality conditions, we can formulate an equation for the regression function to be used in the stochastic approximation algorithm. This regression function is used in the complex stochastic approximation algorithm to perform the same function as the gradient in determining the adaptive matrix filter. The following theorem provides

the regression function needed in the stochastic approximation algorithm which calculates the matrix filter.

Theorem 4.1

A necessary and sufficient condition for the n x n matrix filter, H(f) , to minimize the mean square error

$$E(\underline{\varepsilon}^H\underline{\varepsilon}) = E[(\underline{d} - H^T\underline{x})^H(\underline{d} - H^T\underline{x})] \tag{4.1}$$

is that the orthogonality conditions

$$E\{(\underline{d} - H^T\underline{x})\underline{x}^H\} = 0 \tag{4.2}$$

or, equivalently,

$$E\{\underline{x}(\underline{d} - H^T\underline{x})^H\} = 0 \tag{4.3}$$

are satisfied.

It is only needed to state the results from Theorem 3.4 to prove the above theorem. The optimum filter can now be derived from the orthogonality conditions of Theorem 4.1. From Equation (4.2), we have:

$$E\{\underline{\varepsilon x}^H\} = E\{(\underline{d} - H^T\underline{x})\ \underline{x}^H\} = 0 \ ,$$

$$= G_{\underline{dx}} - G_{\underline{yx}} = 0 \ ,$$

$$G_{\underline{dx}} = G_{\underline{yx}} \ . \tag{4.4}$$

From Equation (4.3), we have:

$$E\{\underline{x}\varepsilon^H\} = E\{\underline{x}(d - H^T\underline{x})^H\}$$

$$= E(\underline{x}d^H) - E(\underline{x}\underline{x}^H)H^*$$

$$= G_{\underline{x}d} - G_{\underline{x}\underline{x}}H^* = 0 \quad ,$$

and

$$H^* = [G_{\underline{x}\underline{x}}]^{-1} G_{\underline{x}d} \quad . \tag{4.5}$$

Cases where the gradient of the mean square error exists, the orthogonality condition is equivalent to the gradient. Hence, the estimated orthogonality condition can be used instead of the estimated gradient in the adaptive algorithm. The orthogonal projection lemma suggests a stochastic approximation algorithm to find the adaptive matrix filter of the form

$$H^*_{n+1} = H^*_n + \mu_n[G_{\underline{x}d} - G_{\underline{x}\underline{x}}H^*_n] = H^*_n + \mu_n R(\underline{x}|H) \quad . \tag{4.6}$$

Equation (4.6) is in state variable form and this implementation appears in Figure 8. It is to be noted that the optimum filter depends only on the power spectral densities of the input signal and noise. Due to the mean square error criterion, any two realizations of input random processes with the same power spectral density will require the same optimum filters.

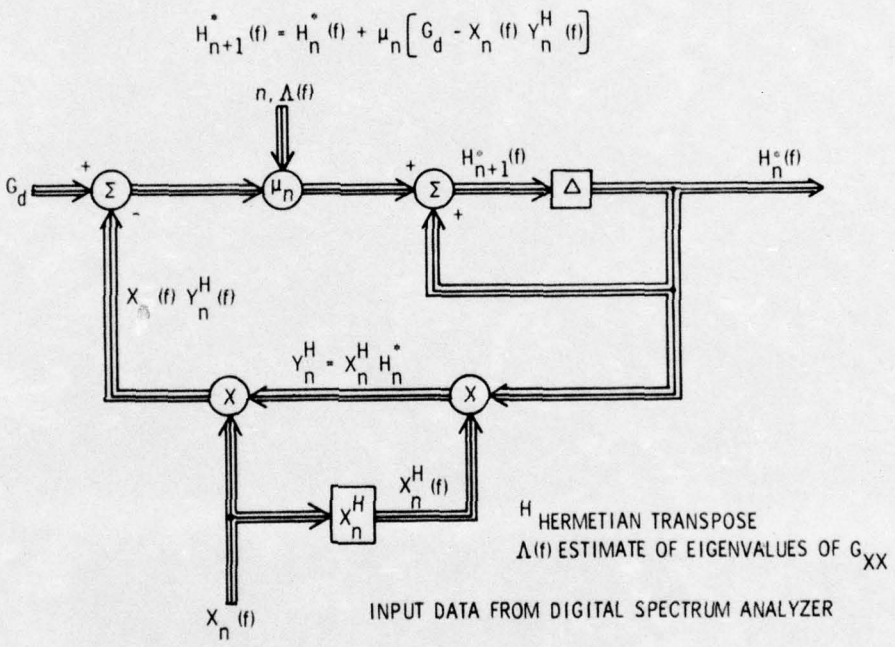


Figure 8. Stochastic Approximation Algorithm

It is true that all processes, Gaussian or non-Gaussian, with the same auto-power spectral density and cross-power spectral density matrices lead to the same processor and the same mean square error if the processing is required to be linear (Rhodes, 100; Van Trees, 125).

In the statistical case, while the performance measure has a unique minimum, the random process that attains this minimum may be different for each realization of the random process represented by the filter weights and only the long term statistical properties determine the optimum filter when the filter is defined to be a linear function of the inputs. It is in the case of Gaussian processes that no nonlinear filter is superior to the best linear filter in the mean square error sense.

4.2 Decoupling Concept of Adaptive Filter

According to the orthogonality condition, the sufficient condition for minimization of the mean square error is

$$G_{\underline{dx}}(f) - H(f)^T G_{\underline{xx}}(f) = 0 \tag{4.7}$$

which implies

$$G_{\underline{dx}}(f) = G_{\underline{yx}}(f) = H(f)^T G_{\underline{xx}}(f) \quad . \tag{4.8}$$

Now, if one assumes that the cross-spectrum matrix between desired signal, $\underline{d}$, and input vector, $\underline{x}$, is

$$G_{\underline{dx}}(f) = D_{iag}[G_{d_i x_i}] \quad , \tag{4.9}$$

then a typical off-diagonal term is:

$$G_{y_i x_j} = 0 \quad \text{for} \quad i \neq j \ . \tag{4.10}$$

This means that under so-called weak signal strong interference assumption, minimization of mean square error is essentially the same as diagonalization of cross-spectral density matrix. It is under this assumption that the performance of the proposed adaptive algorithm is to be judged. The matrix representation for the decoupling concept appears in Figure 9.

4.3 Proof of Convergence of Decoupling Matrix Filter

This section shows that the adaptive decoupler algorithm (the vector case) satisfies the convergence conditions required by stochastic approximation and therefore converges as the number of iterations become large, to the optimum decoupling matrix filter with probability one and in the mean square.

The algorithm

$$\underline{h}^*_{n+1} = \underline{h}^*_n - \mu_n \, \underline{r}(\underline{x}_n | \underline{h}_n) \tag{4.11}$$

converges if the following conditions are satisfied:

$$\text{A1. } \mu_n > 0 \ ; \quad \text{A2. } \sum_{n=1}^{\infty} \mu_n = \infty \ ; \quad \text{A3. } \sum_{n=1}^{\infty} \mu_n^2 < \infty \tag{4.12}$$

$$\text{B.} \quad \inf_{\varepsilon < ||\underline{h} - \underline{h}_0|| < \frac{1}{\varepsilon}, \ \varepsilon > 0} E\{(\underline{h} - \underline{h}_0)^T \, \underline{r}(\underline{x}|\underline{h})\} > 0 \quad \text{in the neighborhood of } \underline{h}_0 \ . \tag{4.13}$$

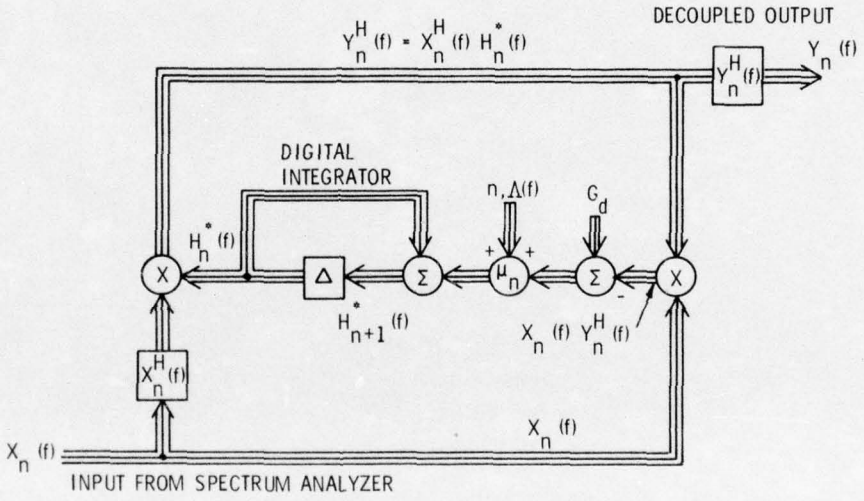
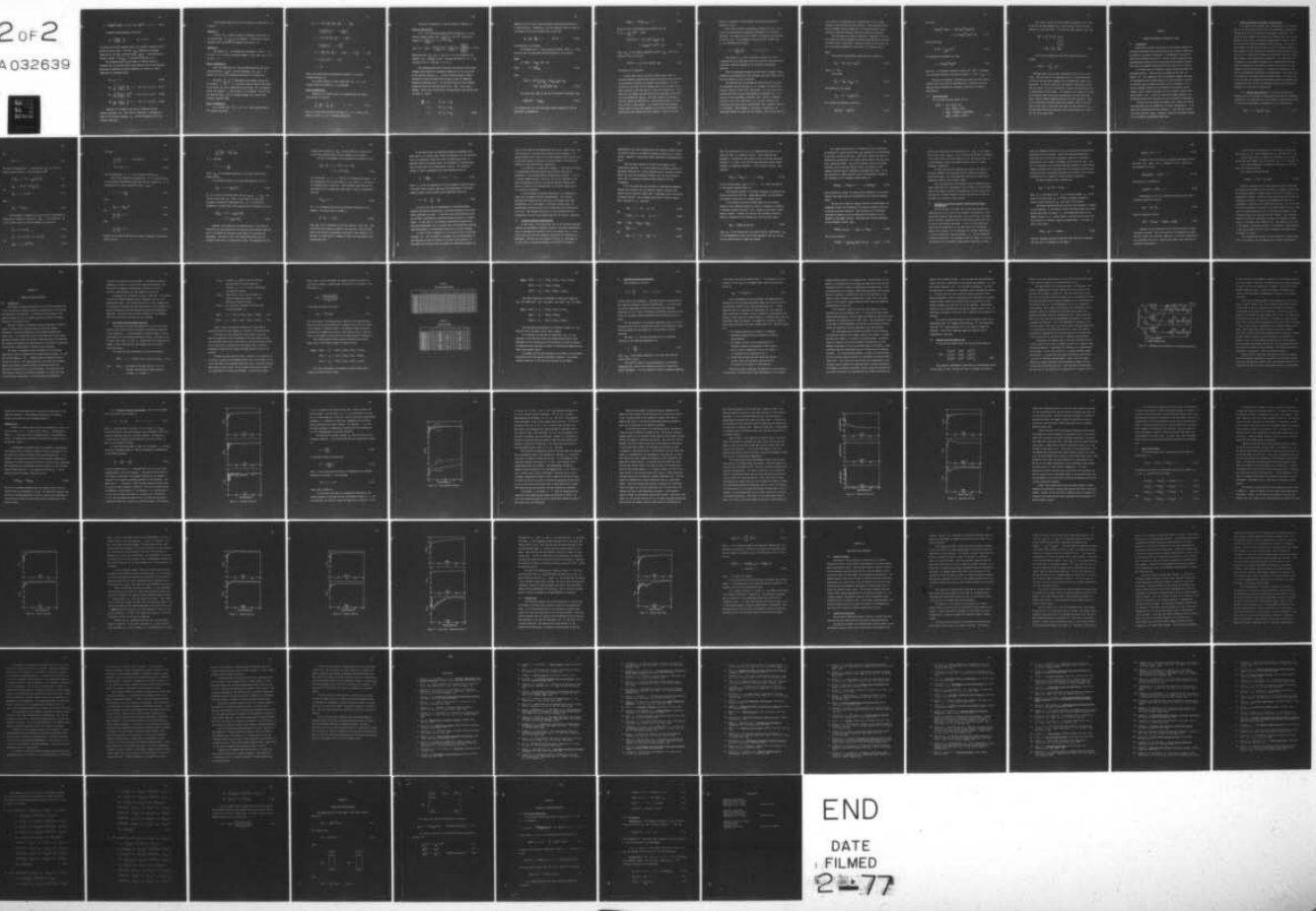


Figure 9. Matrix Filter Decoupler

AD-A032 639 PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESE--ETC F/G 9/4
APPLICATION OF STOCHASTIC APPROXIMATION TO FREQUENCY DOMAIN ADA--ETC(U)
OCT 75 F J EVERLY N00017-73-C-1418
UNCLASSIFIED TM-75-248 NL

2 OF 2
AD A032639



END
DATE FILMED
2-77

$$C.\quad E\{\underline{r}(\underline{x}|\underline{h})^T\,\underline{r}(\underline{x}|\underline{h})\} \le d(1 + ||\underline{h} - \underline{h}_0||^2)\ ,\quad d > 0\ . \qquad (4.14)$$

A special infinite sequence of the form

$$\mu_n = \left(\frac{2.0}{\lambda_{max}}\right)\ \frac{1}{n^i}\ ; \qquad 0.5 < i \le 1.0 \qquad (4.15)$$

is proposed as the gain sequence used in the recursive algorithm used to obtain the matrix filter. The term λ_{max} represents the maximum eigenvalue of the input covariance matrix $\underline{G_{xx}}(f)$. The derivation of the gain constant $(2.0/\lambda_{max})$ is contained in Section 5.1.1.

The following proofs use the theory of infinite series to establish that the special infinite gain sequence used in the stochastic approximation algorithm satisfies Conditions A1, A2 and A3. These conditions are reproduced below.

$$A1.\quad \mu_n > 0 \qquad (4.16)$$

$$A2.\quad \sum_{n=1}^{\infty}\ \left(\frac{2}{\lambda_{max}}\right)\ \frac{1}{n^i} = \infty \quad \text{for}\ \ 0.5 < i \le 1.0 \qquad (4.17)$$

$$A3.\quad \sum_{n=1}^{\infty}\ \left[\left(\frac{2}{\lambda_{max}}\right)\ \frac{1}{n^i}\right]^2 < \infty \quad \text{for}\ \ 0.5 < i \le 1.0 \qquad (4.18)$$

$$A3'.\quad \lim_{n\to\infty}\ \left(\frac{2}{\lambda_{max}}\right)\ \frac{1}{n^i} \to 0 \quad \text{for}\ \ 0.5 < i \le 1.0 \qquad (4.19)$$

Condition A3' is added to the list to illustrate how the sequence approaches zero. This result is important in assessing the effect of the infinite sequence, μ_n , on the convergence rate of the recursive algorithm.

The following results are used in the proofs of Conditions A1, A2, A3 and A3'.

Theorem 4.2

A series Σa_ν of positive terms is convergent if and only if, its partial sums, $S_\nu = \sum_{k=0}^{\nu} a_k$, are bounded. If the partial sums are unbounded, then the series is divergent to the value $+\infty$.

Theorem 4.3

The sequence a_n is convergent and converges to a limit L if, for each $\varepsilon > 0$, there is a positive integer N such that $|a_n - L| < \varepsilon$ for all $n > 1$.

Proof of Condition A1

Since the eigenvalues of a complex Hermetian matrix [the input covariance matrix, $G_{\underline{xx}}(f)$] are real (Wilkinson, 134), and n is greater than zero, we can guarantee satisfaction of Condition A1.

The series $\sum_{n=1}^{\infty} \frac{1}{n^i}$ is designated as the harmonic series with the exponent i. For $i < 1$, its partial sums are greater than those of the series $\Sigma\frac{1}{n}$ and by comparison with the series $\Sigma\frac{1}{n}$, the harmonic series with exponent i, will be shown to be divergent. The fact that the harmonic series $\sum_n \frac{1}{n}$ is divergent (Knopp, 69) is used in the following comparison test.

Proof of Condition A2

Let us observe that $n^k \leq n$ for $k \leq 1$. Then, considering the $2^{n'}$th partial sum we get

$$S_{2^n} = 1 + \frac{1}{2^k} + \left(\frac{1}{3^k} + \frac{1}{4^k}\right) + \left(\frac{1}{5^k} + \ldots + \frac{1}{8^k}\right)$$

$$+ \left(\frac{1}{(2^{n-1} + 1)^k} + \ldots + \frac{1}{(2^n)^k} \right)$$

$$\geq 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \ldots + \frac{1}{8}\right) + \ldots$$

$$+ \left(\frac{1}{(2^{n-1} + 1} + \ldots + \frac{1}{2^n} \right)$$

$$\geq 1 + \frac{1}{2}\left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \ldots + \frac{1}{8}\right) + \ldots + \left(\frac{1}{2^n} + \ldots + \frac{1}{2^n}\right)$$

$$= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots + \frac{1}{2}$$

$$= 1 + \frac{n}{2} \quad . \tag{4.20}$$

Hence, the partial sums are unbounded and therefore, the series diverges (Theorem 4.2).

By a similar method, it can be shown that, for $i > 1$, the harmonic series with exponent i is convergent.

Proof of Condition A3

Condition A3 is obvious once it is recognized that the series can be written in the form

$$\sum_{n=1}^{\infty} \frac{1}{n^{2i}} = \sum_{n=1}^{\infty} \frac{1}{n^j} \quad , \qquad j = 2i \quad . \tag{4.21}$$

Since it is known that the series converges for $j > 1$, then it only remains to bound $j > 0.5$ to satisfy Condition A3.

The proof of Condition A3' uses the results of Theorem 4.3.

Proof of Condition A3'

If ε is an arbitrarily small positive number and N is any positive integer greater than $\left(\frac{2}{\lambda_{max}}\right)\frac{1}{n}$, the m^{th} term of the sequence is $\left(\frac{\hat{\alpha}}{\lambda_{max}}\right)\frac{1}{m}$. If $m > N$, then,

$$|a_m - 0| = |a_m| = \left|\left(\frac{2}{\lambda_{max}}\right)\frac{1}{m}\right| = \left(\frac{2}{\lambda_{max}}\right)\frac{1}{m} < \frac{\left(\frac{2}{\lambda_{max}}\right)}{\left(\frac{2}{\lambda_{max}}\right)/\varepsilon} = \varepsilon \quad (4.22)$$

which shows that $|a_m - 0| < \varepsilon$ for all $m > N$ and, therefore, the sequence $\{a_n\}$ converges to zero. The proof has been for $i = 1$ but it holds for all i such that $0.5 < i \leq 1$.

The preceding proofs have shown that the specially derived gain sequence does satisfy the convergence Conditions (A1, A2, A3 and A3') on the infinite sequence μ_n. In the following analysis, results obtained in Tuteur and Chang (123) are used to show the algorithm satisfies Conditions B and C. We know from results of the orthogonal projection lemma that the mean square error, $\varepsilon(\underline{h})$, has a unique minimum. We can take a derivative of the mean square error and see that for every i we get

$$\begin{aligned} \frac{\partial \varepsilon}{\partial h_i} &> 0 \qquad \text{for } h_i > h_{i0} \\ &= 0 \qquad \text{for } h_i = h_{i0} \\ &< 0 \qquad \text{for } h_i < h_{i0} \,. \end{aligned} \quad (4.23)$$

Equation (4.14) is just a formal statement indicating the behavior of a convex function. Consequently, with the regression function equal to the gradient in the real variable case, we see that

$$(\underline{h}_1 - \underline{h}_{10}) \frac{\partial \varepsilon}{\partial h_i} \geq 0 \qquad \text{for all } i \tag{4.24}$$

and Condition B is satisfied.

If we assume each x (input signal) satisfies $E\{x^2\} < \infty$, then the proof that the algorithm satisfies Condition C is as follows:

Proof

$$\text{If } \nabla\varepsilon(\underline{h}) = \nabla\varepsilon\Big|_{\underline{h}_0} + J(\underline{h} - h_0)$$

$$\text{where} \qquad J = 2\underline{xx}^T\Big|_{h_0} \tag{4.25}$$

then,

$$\nabla\varepsilon^T\nabla\varepsilon = (\nabla\varepsilon)^T\Big|_{\underline{h}_0}(\nabla\varepsilon\Big|_{\underline{h}_0}) + 4\nabla\varepsilon\Big|_{\underline{h}_0} G_{\underline{xx}}(\underline{h} - \underline{h}_0)$$

$$+ 4(\underline{h} - \underline{h}_0)\, \underline{xx}^T\, \underline{xx}^T(\underline{h} - \underline{h}_0) \quad . \tag{4.26}$$

If we note that $\underline{x}^T\underline{x}$ is the sum of the power in the input, then

$$E\{\underline{xx}^T\, \underline{xx}^T\} = P_{IN}G_{\underline{xx}} \quad . \tag{4.27}$$

It is obvious that the first and second terms in Equation (4.25) are zero since, by definition,

$$\underline{r}(\underline{x}|\underline{h}_0) = \nabla\varepsilon(\underline{h})\Big|_{\underline{h} = \underline{h}_0} = \underline{0} \quad . \tag{4.28}$$

Using the result from Beckenbach and Bellman (13) that $\underline{x}^T A \underline{x} \leq \lambda_{\substack{\max \\ \text{of A}}} ||\underline{x}||^2$, gives

$$E\{(\nabla\varepsilon)^T(\nabla\varepsilon)\} = 4(\underline{h} - \underline{h}_0)^T P_{IN} G_{\underline{xx}} (\underline{h} - \underline{h}_0)$$

$$\leq 4P_{IN}\lambda_{max}(\underline{h} - \underline{h}_0)^T(\underline{h} - \underline{h}_0) \tag{4.29}$$

where λ_{max} is the largest eigenvalue of matrix $G_{\underline{xx}}$. Then from Equation (4.29), we see that

$$E\{\nabla\varepsilon^T\nabla\varepsilon\} \leq c_1[1 + (\underline{h} - \underline{h}_0)^T(\underline{h} - \underline{h}_0)] \tag{4.30}$$

and Condition C is satisfied.

We have shown that for the real variable vector case, the proposed decoupling algorithm which derives the matrix filter satisfies the convergence conditions for stochastic approximation. The only difference in the proof for the complex case is that we cannot talk about gradients in the same sense as in the real variable case. It is shown that the complex case satisfies the five convergence conditions in strict mathematical fashion. The fifth condition is imbedded in classical analysis's definition of a convex function and it is difficult to prove in this context. It is obvious that the performance measure used is no more than quadratic in the filter coefficients, H , and therefore must satisfy the final condition. Since we can from

Theorem 3.4 guarantee a unique minimum, the reasoning behind this condition is valid.

Recall from stochastic approximation [Equations (4.12), (4.13), (4.14)] the conditions which guarantee convergence of the iterative algorithm. The complex case can be shown to converge in the same sense as the real case. The convergence conditions on the variable gain, μ_n , are satisfied by the choice

$$\mu_n = \left(\frac{2}{\lambda_{max}}\right)\frac{1}{n^i}; \qquad 0.5 < i \leq 1.0 \quad . \tag{4.31}$$

This choice of μ_n will be derived in Section 5.1.1 from considerations of the idealized case of the recursive algorithm and the convergence conditions imbedded in the stochastic approximation algorithm.

Since the performance measure has been shown to possess a unique minimum from considerations of the orthogonal projection lemma, then we know that Equation (4.13) is satisfied.

The condition that the mathematical expectation of the norm of the regression function increases, as $\underline{h}$ increases, no faster than a quadratic function of the filter coefficients, $\underline{h}$, can be seen to be satisfied from both intuitive and geometrical considerations. It is obvious from the performance functional itself, which is a quadratic function of the filter coefficients, $\underline{h}$, that the surface of the performance functional cannot increase faster than a quadratic function of the weights. This statement is equivalent to saying that the performance measure is convex near the minimum. Since we know that the

error surface is bounded above by a quadratic and it has a unique minimum, it must satisfy the final condition. This reasoning follows that of Friedman (46) who showed that, if the given function can be bounded by some known function, then this condition is satisfied.

The following is the proof of Condition C [Equation (4.14)] for the complex form of the stochastic recursive algorithm which calculates the matrix filter coefficients. The proof is done for the vector case but can be easily extended for each vector in the matrix filter.

Proof

The stochastic approximation algorithm can be written as

$$\underline{H}^*_{n+1} = \underline{H}^*_n + \mu_n [\underline{G}_d - \underline{x}_n \underline{y}_n^H] \tag{4.32}$$

where $\underline{y} = \underline{H}^T \underline{x}$.

If we define

$$\underline{Z}^*_{n+1} = \underline{H}^*_{n+1} - \underline{G}_{\underline{x}_n \underline{x}_n}^{-1} \underline{G}_d \quad , \tag{4.33}$$

then Equation (4.32) becomes

$$\underline{Z}^*_{n+1} = [\underline{I} - \mu_n \underline{x}_n \underline{x}_n^H] \underline{Z}^*_n \quad . \tag{4.34}$$

If we define the regression function as

$$\underline{r}_n(\underline{x}_n | \underline{z}_n) = \underline{x}_n \underline{x}_n^H \underline{z}^*_n \quad , \tag{4.35}$$

then we get

$$E\{\underline{r}(\underline{x}|\underline{z})^H \, r(\underline{x}|\underline{z})\} = E\{\underline{z}_n^T(\underline{x}_n\underline{x}_n^H)^H(\underline{x}_n\underline{x}_n^H)\underline{z}_n^*\}$$

$$= \underline{z}_n^T \, E\{(\underline{x}_n\underline{x}_n^H)^H(\underline{x}_n\underline{x}_n^H)\} \, \underline{z}_n^* \quad . \tag{4.36}$$

Since we know that

$$\underline{z}^T A\underline{z}^* \leq \lambda_{\substack{\max \\ \text{of } A}} ||\underline{z}||^2 \tag{4.37}$$

from Beckenbach and Bellman (13) where

$$A = E\{(\underline{x}_n\underline{x}_n^H)^H(\underline{x}_n\underline{x}_n^H)\} \quad , \tag{4.38}$$

and since A is bounded (a physical process) and $||\underline{z}||^2$ is bounded, then to satisfy Condition C we need only choose d to be $\leq \lambda_{\max}$ of A.

The proof of Condition C completes the proofs that show that the complex form of the recursive algorithm used to calculate the matrix adaptive filter satisfies the convergence conditions of stochastic approximation.

4.4 Mean Square Error

We can compute the mean square error as

$$E\{\underline{\varepsilon}^H\underline{\varepsilon}\} = E\{(\underline{d} - \underline{y})^H(\underline{d} - \underline{y})\}$$

$$= E\{(\underline{d} - H^T\underline{x})^H(\underline{d} - H^T\underline{x})\}$$

$$= E\{\underline{d}^H\underline{d} - 2(H^T\underline{x})^H\underline{d} + (H^T\underline{x})^H(H^T\underline{x})\}$$

$$= E\{\underline{d}^H\underline{d} - 2(\underline{y})^H\underline{d} + (\underline{y})^H(\underline{y})\} \quad . \tag{4.39}$$

If we look at any of the three products in Equation (4.29), then we can see that each expression is a sum of terms in which each term depends on a particular beam. If we take the $\underline{d}^H\underline{d}$ product, we see that

$$\underline{d}^H\underline{d} = [d_1^* \quad d_2^* \quad d_3^*] \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$$

$$= d_1^* d_1 + d_2^* d_2 + d_3^* d_3 \quad . \tag{4.40}$$

We can see from Equation (4.39) that the mean square error becomes

$$E\{\varepsilon^H\varepsilon\} = E\{\sum_{i=1}^{N} |\varepsilon_i|^2\} \quad . \tag{4.41}$$

The mean square error is seen from Equation (4.31) to be a sum of terms. With this form of the mean square error, it becomes desirable to look at each individual component of the minimum mean square error separately. In the case derived in this research, where there is more than one output, it is very instructive to see how the mean square error is being reduced in each output. For instance, if the power in one beam is one hundred times the power in another beam, then the interfering noise could be removed completely from the low power beam and it would not be detectable in the mean square error because a small 1% or less error in the filter coefficients that affect the mean square error in the high power beam would completely mask the mean square error reduction in the low power beam.

CHAPTER V

DYNAMIC PROPERTIES OF STOCHASTIC FILTER

5.0 Introduction

This chapter contains the analysis of the dynamic properties of the matrix decoupling filter and an analysis of stochastic stability as it applies to the adaptive algorithm. An optimum gain is derived to obtain fast convergence properties. The use of this optimum gain has been mentioned previously and its properties are discussed in greater detail when the computer simulation results of Chapter VI are given.

It is shown that the stochastic Lyapunov functions possess the martingale property. This was first recognized by Bucy (21) and Kushner (73). The recognition that the stochastic Lyapunov functions are martingales should provide impetus for future results in the area of automatic control systems and adaptive array processing. With a readily available analysis for the stability of stochastic control systems, control systems designers will find stochastic systems techniques easier to use. A more firm connection between adaptive array processing and the vast literature of control systems is advantageous.

The last section contains the analysis that unifies all the LMS type algorithms into the framework of stochastic approximation. It is shown that algorithms which use a constant μ when the input signals are random processes cannot, in general, obtain the advantages inherent in the stochastic approximation algorithm.

5.1 Dynamic Properties of Stochastic Adaptive Filter

It is shown in this chapter, from considerations of the idealized case for the decoupling algorithm, that the maximum gain constant can be determined for use in the stochastic approximation algorithm. This particular gain constant can be used as the starting value for a special sequence in the context of stochastic approximation to greatly increase the rate at which the matrix adaptive filter converges. The purpose of this analysis, stated in the introduction, was to design an algorithm which operates in real time and yet increases the output signal-to-noise ratio so as to increase the probability of detection. With this special gain sequence, these objectives are satisfied. It is shown that this optimum gain is the largest allowable in the idealized case and represents the borderline of the stability region in the context of state variable control system analysis. If, however, this gain were to be used in a deterministic algorithm, the misadjustment of the weights and the noisiness of the system would be extremely large. Of course, this is not the case with stochastic approximation because this mathematical technique takes into account the stochastic nature of the system and is insensitive to this type of misadjustment.

5.1.1 Optimum Gain Calculation To derive the special gain sequence used in the stochastic approximation algorithm, the convergence properties of the idealized form of the stochastic algorithm are explored. If the recursive algorithm is rearranged into the following form:

$$H^*_{n+1} = [I - \mu G_{\underline{xx}}] H^*_n + \mu G_{\underline{xd}} \quad , \tag{5.1}$$

then a substitution of the form

$$Z^*_{n+1} = H^*_{n+1} - G_{\underline{xx}}^{-1} G_{\underline{xd}} \tag{5.2}$$

reduces Equation (5.1) to

$$\begin{aligned} H^*_{n+1} &= Z^*_{n+1} + G_{\underline{xx}}^{-1} G_{\underline{xd}} \\ &= [I - \mu G_{\underline{xx}}][Z^*_n + G_{\underline{xx}}^{-1} G_{\underline{xd}}] + \mu G_{\underline{xd}} \\ &= [I - \mu G_{\underline{xx}}] Z^*_n + [I - \mu G_{\underline{xx}}] G_{\underline{xx}}^{-1} G_{\underline{xd}} + \mu G_{\underline{xd}} \quad . \end{aligned} \tag{5.3}$$

Simplifying, we have:

$$Z^*_{n+1} = [I - \mu G_{\underline{xx}}] Z^*_n \quad . \tag{5.4}$$

Equation (5.1) neglects the stochastic nature of the algorithm because it assumes that the cross spectral density matrix $G_{\underline{xx}}$ is either given or can be computed exactly from input data. This assumption assumes importance when the positive definiteness of the cross spectral density matrix $G_{\underline{xx}}$ is required. The above equation also assumes a constant gain instead of a variable gain.

Let U be an unitary transformation which diagonalizes the power spectral density matrix $G_{\underline{xx}}$, i.e.,

$$U^H G_{\underline{xx}} U = \Lambda \quad , \tag{5.5}$$

and

$$U^H U = I \quad . \tag{5.6}$$

The unitary transformation U exists because $G_{\underline{xx}}$ is a positive definite Hermitian matrix. The algorithm becomes:

$$U^H Z^*_{n+1} U = U^H[I - \mu G_{\underline{xx}}] UU^H Z^*_n U \quad , \tag{5.7}$$

$$Q^*_{n+i} = [U^H IU - \mu U^H G_{\underline{xx}} U] Q^*_n \tag{5.8}$$

and

$$Q^*_{n+1} = [I - \mu\Lambda] Q^*_n \quad , \tag{5.9}$$

where

$$Q^*_{n+1} = U^H Z^*_{n+1} U \quad . \tag{5.10}$$

The convergence of Equation (5.9) can be easily established by examining the sequence of solutions of Q^*_{n+1}. If we assume that Q^*_0 is the initial condition, then the sequence of solutions is

$$Q^*_1 = [I - \mu\Lambda] Q^*_0 \quad , \tag{5.11}$$

$$Q^*_2 = [I - \mu\Lambda] Q^*_1 = [I - \mu\Lambda]^2 Q^*_0 \tag{5.12}$$

and

$$Q^*_{n+1} = [I - \mu\Lambda]^{n+1} Q^*_0 \quad . \tag{5.13}$$

The limit

$$\lim_{n+1 \to \infty} Q^*_{n+1} = 0 \quad \text{if and only if} \tag{5.14}$$

$$||[I - \mu\Lambda]|| < 1 \ , \tag{5.15}$$

where the expression $||\cdot||$ is any suitably defined norm.

Since it has already been assumed that $\mu > 0$ in the derivation of the stochastic approximation algorithm, the only condition on μ for satisfaction of the norm inequality is that $\mu \lambda_{max} < 2$,

$$0 < \mu < \frac{2}{\lambda_{max}} \ . \tag{5.16}$$

Since

$$Z^*_{n+1} = U Q^*_{n+1} U^H \tag{5.17}$$

and

$$\lim_{n+1 \to \infty} Q^*_{n+1} = 0 \ , \tag{5.18}$$

then

$$\lim_{n+1 \to \infty} Z^*_{n+1} = 0 \ . \tag{5.19}$$

Equation (5.19) implies that the actual filter converges to the optimum filter, that is,

$$\lim_{n+1 \to \infty} H^*_{n+1} = G_{\underline{xx}}^{-1} G_{\underline{xd}} \tag{5.20}$$

if μ satisfies

$$0 < \mu < \frac{2}{\lambda_{max}} , \tag{5.21}$$

where λ_{max} is the maximum eigenvalue of the input power spectral density matrix.

The stochastic version of the algorithm [Equation (5.4)] is:

$$Z^*_{n+1} = [I - \mu \underline{x}_n \underline{x}_n^H] Z^*_n . \tag{5.22}$$

Z^*_n is a function of previous input data $\underline{x}_0, \underline{x}_1, \underline{x}_2, \ldots \underline{x}_{n-1}$, but not the current input $\underline{x}_n$, hence, if each new set of input data represents statistically independent data, Z^*_n is statistically independent of present data and the expected value of Equation (5.22) is:

$$\begin{aligned} E\{Z^*_{n+1}\} &= E[I - \mu \underline{x}_n \underline{x}_n^H] E\{Z^*_n\} \\ &= [I - \mu G_{\underline{xx}}] E\{Z^*_n\} . \end{aligned} \tag{5.23}$$

Equation (5.23) shows that the expected value of the filter is governed by the same algorithm as the idealized non-stochastic algorithm when it is assumed that successive input samples are statistically independent. The same is true if the weakened hypothesis that the successive input data are uncorrelated is used. The expression for the

second order statistics of Z^*_{n+1} is more difficult to obtain and it has not been computed since it is not needed in this analysis.

The rate of convergence can be determined from Equation (5.13):

$$Q^*_{n+1} - Q^*_n = [I - \mu\Lambda]^n[I - \mu\Lambda - I]\, Q^*_0$$

$$= - [I - \mu\Lambda]^n\, \mu\Lambda\, Q^*_0 \quad . \tag{5.24}$$

In the case when $\mu\, \lambda_i << 1$, the rate of convergence for small n in each independent loop is proportional to $\mu\, \lambda_i$, where λ_i is equal to the decoupled power in that loop. This statement comes from the fact that Equation (5.5) assumed that there exists a unitary transformation such that

$$U^H\, G_{xx}\, U = \Lambda \tag{5.25}$$

and Λ is a diagonal matrix with the decoupled eigenvalues on the diagonal. The initial rate of change is:

$$Q^*_1 - Q^*_0 = - \mu\Lambda\, Q^*_0 \quad , \tag{5.26}$$

where $\frac{1}{\mu\lambda_i}$ can be interpreted as the time constant of each loop. This relation has been observed in Widrow's LMS algorithm. Since the gain used in the stochastic approximation algorithm is time variable, it is difficult to assess speed of response in exactly the same sense as the constant gain case.

We can observe that the underlying relaxation phenomenon which takes place in the weight values (filter coefficients) for the idealized case is of exponential nature and, since the mean square error is a quadratic form in the weight values, the transients in the mean-square error function must also be exponential in nature. The transients consist of sums of exponentials with time constants given by:

$$T_p = \frac{1}{(\mu)\lambda_i} , \qquad i = 1, 2, \ldots, n , \tag{5.27}$$

where λ_i is the ith eigenvalue of the input signal correlation matrix. It has been shown by Widrow (133) that the misadjustment of the basic adaptive element using the pilot vector algorithm is given by

$$M = \frac{1}{2} \sum_{p=1}^{n} \frac{1}{T_p} . \tag{5.28}$$

It can be seen that the misadjustment of the deterministic gradient algorithms is proportional to the time constants of the filter weight adjustment procedure. It has been observed experimentally that the stochastic approximation algorithm performs initially in a similar manner. As time goes on, the stochastic approximation algorithm has much less misadjustment than is possible with the deterministic gradient algorithms because of the smoothing properties of the gain sequence.

It will be seen in Chapter VI, in the plots of S/N for different starting μ's , that the speed of convergence and the misadjustment of the algorithm are directly related to the size of the starting μ's . The smaller the starting value, the slower the algorithm converges, but

there is less noise in the algorithm and the S/N is a smooth curve. The price one pays for this less noisy system is slow convergence. The greatest benefit that accrues to the system designer from the use of the special gain sequence is the fast initial increase in S/N without the drawback of noisy matrix filter coefficients.

It is difficult to compare the constant μ case with the stochastic approximation algorithm because the stochastic algorithm is essentially a time varying feedback system and the constant μ case is a constant feedback system. Initially, the two are comparable if the same starting gain is used but the time varying gain possesses the ability to smooth out the misadjustment of the filter weights due to its statistical properties.

In the deterministic gradient algorithms, it is impossible to obtain fast convergence and small misadjustment. The stochastic approximation technique allows fast initial convergence with large misadjustment, but as the number of iterations becomes large, the misadjustment (variance of the filter coefficients) will automatically go to zero. This variance reduction is a very valuable result for systems which can allow long running times for the recursive algorithm.

5.2 Stochastic Stability Considerations

At this point, it has been proven that if the recursive algorithm satisfied the convergence conditions imposed by stochastic approximation, then it converged with probability one to the optimum filter. It was noted in the proof that the sequence of matrix filter coefficients is a martingale. The fact that the sequence of filters is a martingale is not only useful in proving convergence in the sense of stochastic

approximation, but also in showing that the stochastic adaptive filter algorithm is stable in the sense of stochastic stability of a control system. Appendix C contains some useful definitions concerning martingales.

There are many examples of the use of martingales in the engineering literature. It is well known that the Wiener process is a martingale (McGarty, 83), certain integrals used in stochastic differential equations are martingales (McShane, 85), and the likelihood ratios used in statistical detection theory (McGarty, 83) are martingales.

It is well known that the stability of deterministic dynamical systems can be proven if we can construct Lyapunov functions that satisfy certain conditions (Sage, 108). Using Ogata (91), we can define a Lyapunov function. Let us suppose there exists a scalar function $V(\underline{x})$ continuous in $\underline{x}$, such that:

$$V(\underline{x}) > 0 \quad \text{for} \quad \underline{x} \geq \underline{0} \tag{5.29}$$

and

$$\Delta V(\underline{x}) < 0 \quad \text{for} \quad \underline{x} \geq 0 \ , \tag{5.30}$$

where $\Delta V(\underline{x}) = V(\underline{x}_{k+1}) - V(\underline{x}_k)$,

$$V(\underline{0}) = 0 \tag{5.31}$$

and

$$V(\underline{x}) \to \infty \quad \text{as} \quad ||\underline{x}|| \to \infty \ . \tag{5.32}$$

Then, the equilibrium state $\underline{x} = \underline{0}$ is asymptotically stable in the large and $V(\underline{x})$ is a Lyapunov function. These conditions are equivalent to saying that there exists a positive definite continuous function of $\underline{x}$, $V(\underline{x})$, such that its time derivative along a system trajectory is nonpositive definite. If the Lyapunov function, $V(\underline{x})$, satisfies

$$V(\underline{x}_{n_1}) \geq V(\underline{x}_{n_2}) \geq \ldots \geq V(\underline{x}_{n_i}) \tag{5.33}$$

for the discrete system $\{\underline{x}_n, \; n = 0, 1, \ldots i\}$, then the stability of the origin is implied (Aoki, 8).

The following analysis of stochastic stability uses analysis and results contained in Aoki (8). Using these results, we can establish the stability conditions for a dynamic system.

If we construct the state variable model for the recursive adaptive filter algorithm (Figure 9), then it is possible to construct a stochastic Lyapunov function, $V(\underline{h})$, similar to that used in deterministic systems. Consider the discrete time stochastic dynamical model of the adaptive filter described by (for the vector case):

$$\underline{h}_{k+1} = \underline{f}_k(\underline{h}_k, \underline{x}_k, \underline{y}_k, \underline{\mu}_k) \tag{5.34}$$

where $\underline{h}_k$ is the n-dimensional state vector (filter coefficients), $\underline{\mu}_k$ is the m-dimensional control vector (gain sequence), and $\underline{x}_k$ and $\underline{y}_k$ are the random vectors of inputs and outputs.

For a given control policy, a collection of state trajectories is possible for a given initial state $\underline{h}_0$, depending on the realizations of the input stochastic processes. This result differs from that of a deterministic system where, for a given control policy, only one unique state trajectory results. Due to the stochastic nature of the input processes and the resulting adaptive filter, the behavior of the Lyapunov functions must be considered in probabilistic terms. It is thus indicated to replace Equation (5.33) on deterministic systems by the following one for stochastic systems (Aoki, 8):

$$E\{V(\underline{x}_1)\} \geq E\{V(\underline{x}_2)\} \geq \ldots \geq E\{[V(\underline{x}_n)]\} \quad . \tag{5.35}$$

This condition has exactly the same intuitive meaning as in the deterministic case except that now the definition is in terms of expected values.

The idea that stochastic Lyapunov functions are martingales was recognized by Bucy (21) and Kushner (73) and use is made of this fact in proving stochastic stability. Bucy arrived at the relationship described by Equation (5.35) by noting that the Lyapunov function satisfies a martingale property. Bucy noted that for any realization of the stochastic process $\{\underline{x}_n\}$,

$$E\{V(\underline{x}_n) \mid \underline{x}_0, \underline{x}_1 \ldots, \underline{x}_{n-1}\} \leq V(\underline{x}_{n-1}) \quad . \tag{5.36}$$

If we use the equality

$$E\{V(\underline{x}_n)\} = E_{\underline{x}_n}\{E_{V|\underline{x}_n}[V(\underline{x}_n) \mid \underline{x}_0, \underline{x}_1 \ldots \underline{x}_{n-1}]\} \quad , \tag{5.37}$$

then we can conclude that the expected value of $V(\underline{x}_n)$ is not greater than the last value $V(\underline{x}_{n-1})$. It is precisely the property in Equation (5.36) which allows us to write Equation (5.35) and conclude that, for a stable dynamical system, the Lyapunov function must be a supermartingale (Appendix C).

By direct analogy with deterministic systems, this behavior of stochastic Lyapunov functions means that the conditional expected generalized energy is not increasing with time and the system can be regarded as stable in the stochastic sense.

Thus, by proving that the stochastic process representing the filter coefficients is a martingale not only allows us to prove convergence in the sense of stochastic approximation, but also allows us to show that the stochastic process of filter coefficients is stable in the sense of a stochastic control system.

5.3 Unification of Previous Adaptive Systems Using Stochastic Approximation

In this section, it is shown that the LMS type deterministic algorithms can be put in the form of stochastic approximation. Other algorithms, beside the LMS algorithms, can be unified into the framework of stochastic approximation but the unification method will only be done for the LMS type algorithms. In fact, all algorithms that result from the minimization of the mean square error performance measure can be shown to fit the framework of stochastic approximation if certain conditions on the gain sequence, μ_n, and the regression function are satisfied. The orthogonal projection lemma, as it was in the derivation of the stochastic approximation algorithm, is used to

derive the regression function for use in the recursive algorithm which determines the filter weights. Since these LMS type deterministic algorithms do not satisfy the convergence conditions of stochastic approximation, only a weaker convergence can be proved for them. This lack of convergence is an extremely important weakness of these adaptive techniques because these algorithms work with stochastic processes and strong statistical convergence properties are necessary if the recursive algorithms are to be applied in a wide variety of situations.

The Widrow adaptive filter algorithm is:

$$\underline{h}_{n+1} = \underline{h}_n + \mu[d_n - y_n]\,\underline{x}_n \quad , \tag{5.38}$$

where $\underline{h}_n$ is the weight vector, d_n is the pilot signal, y_n is the output of the filter and $\underline{x}_n$ is a vector of signals composed of various delayed versions of $\underline{x}_n$, the output of the sensors.

The Widrow algorithm can now be put into the framework of stochastic approximation. It is desired to obtain the tapped delay line filter weights which make the output, y_n, the minimum mean squared error (MMSE) estimate of d_n, a pilot signal. The direction of arrival and the statistics of the pilot signal are assumed known. If we want to find a filter where

$$E\{[d_n - y_n]^2\} = \text{minimum} \quad , \tag{5.39}$$

then the orthogonal projection lemma says this objective is achieved when the error is orthogonal to the signal

$$E\{\underline{x}_n(d_n - y_n)\} = 0 \quad . \tag{5.40}$$

An identity from the theory of conditional expectation (Karlin and Taylor, 66), $E\{g(x)\, f(y)\} = E\{E[g(x)|y]\, f(y)\}$, allows us to write Equation (5.40) as:

$$E_{\underline{h}}\{E_{y/\underline{h}}[\underline{x}_n(d_n - y_n)/\underline{h}_n]\} = 0 \quad . \tag{5.41}$$

This equation is satisfied for

$$E_{y/\underline{h}}\{\underline{x}_n[d_n - y_n]/\underline{h}_n\} = 0 \quad . \tag{5.42}$$

Using Equation (5.42) as a regression function we can write a recursive algorithm of the form

$$\underline{h}_{n+1} = \underline{h}_n + \mu \underline{r}_n \tag{5.43}$$

with the regression function

$$\underline{r}(\underline{h}) = E\{\underline{r}_n/\underline{h}_n\} = E\{\underline{x}_n[d_n - y_n]/\underline{h}_n\} \quad . \tag{5.44}$$

Equation (5.43) and Equation (5.44) constitute Widrow's weight adjustment algorithm. Since this algorithm, as implemented by Widrow, does not satisfy the convergence conditions of stochastic approximation, any convergence must be in a statistically weaker sense than that for stochastic approximation.

To derive Griffith's algorithm, it is necessary to define the input vector $\underline{x}_n$ as a sum of signal vector $\underline{s}_n$ and noise vector $\underline{n}_n$.

We also assume that the signal to be estimated d_n is uncorrelated with every signal in the noise vector $\underline{n}_n$, i.e., $E\{d_n \underline{n}_n^i\} = 0$.

With these assumptions, the regression function can be derived using the preceding analysis.

The regression function becomes

$$E\{\underline{r}_n/h_n\} = E\{[d_n - y_n]\, \underline{x}_n/h_n\} \quad . \tag{5.45}$$

Griffith's algorithm has the same drawback as Widrow's algorithm. Even though the regression function represents a stochastic process and the resultant sequence of weight vectors, the use of a constant μ is suggested by Griffiths. Obviously, this algorithm does not satisfy the convergence conditions of stochastic approximation and thus only a convergence of the mean of the weights can be proven. It is to be noted, that while convergence of the mean of the filter weights is valid, large errors or misadjustments can occur for different realizations of the stochastic process that describe the filter weights. The convergence results for Griffith's algorithm are weaker than convergence in mean square or convergence with probability one. The weakness of this form of convergence is intimately related to the usefulness of the algorithm. In many applications with this type of convergence, the algorithm is, in general, not useful.

All algorithms that result from the minimization of mean square error can be shown to fit the framework of stochastic approximation.

However, the application of adaptive techniques to stochastic processes without the benefit of stochastic approximation techniques results in a reduced statistical performance of the adaptive processor.

CHAPTER VI

COMPUTER SIMULATION RESULTS

6.0 Introduction

This chapter contains a description of the three beam system used to test the stochastic adaptive filter and a discussion of the different tests applied to the algorithm. A discussion of the results of the computer simulation tests for normal operation is given for a sample from each of the generic cases.

The use of implicit constraints, derived from physical considerations, is shown to give the same results as the cases without constraints. The only difference is in the minimum allowable mean square error. It is shown that both the constrained and unconstrained modes give the same result for the short term operation of the algorithm. The physical and mathematical interpretations of the implicit constraints are given and a justification for their use is established.

The speed of convergence and the noisiness of the recursive algorithm are shown to depend on both the special gain constant S_c and the infinite sequence $1/n^i$. Changing these two parameters gives a spectrum of results. The results show that the special gain sequence gives a fast initial increase in output S/N and thus allows the algorithm to operate in an on-line environment. The ease with which the parameters needed in the adaptive algorithm, in a practical case, can be obtained is described. A fast eigenvalue and eigenvector

calculation is described in this chapter. The analytical method of eigenvalue calculation is called the QL algorithm (Wilkinson, 135). This method of eigenvalue determination, basically a series of similarity transformations, is fast enough for on-line use.

A stopping rule is proposed for use in future work. The stopping rule uses an averaged Frobenius norm to determine accuracy in the probabilistic sense. From a theoretical point of view, stopping rules are necessary to determine the long term accuracy of the technique. No general stopping rule has been devised for stochastic approximation algorithms. The matrix filter is only calculated for one frequency since the calculation is the same for each frequency.

6.1 Beam Construction and Signal Generation

A three-beam system has been simulated to be used to test the adaptive filter derived from the stochastic approximation algorithm. In this simulation, s_1 will represent the signal in Beam 1, and all other signals will be considered interfering noises. The same is true of Beams 2 and 3 where s_2 and s_3, respectively, will represent the signals in their beams and all other signals will be considered interfering noises.

The beams have been constructed in the following manner:

$$\text{Beam 1} = x_1 = (sl)s_1 + (a)s_2 + (b)s_3 + (cc)sn_1 \quad , \quad (6.1)$$

where $(sl)s_1$ = the signal in the main lobe one. sl is a constant representing the angular position of signal s_1 in Beam 1.

$(a)s_2$ = a signal s_2 mainly in Beam 2, entering the side lobes of the main beam that contains s_1, a is a constant representing its position in Beam 1.

$(b)s_3$ = same form as $(a)s_2$ but with signal s_3 mainly in Beam 3 and constant b representing its position in Beam 1.

$(cc)sn_1$ = any self noise contained in the same lobe as the main signal s_1.

$$\text{Beam 2} = x_2 = (f)s_1 + (s2)s_2 + (g)s_3 + (dd)sn_2 \quad . \qquad (6.2)$$

$$\text{Beam 3} = x_3 = (h)s_1 + (p)s_2 + (s3)s_3 + (ee)sn_3 \quad . \qquad (6.3)$$

Figure 4 shows the physical description of signal spatial position relative to other signals as a function of the beam pattern.

The beams have been constructed in this manner so that an analytical expression could be obtained for both the signal-to-noise ratio and mean square error at the output of the matrix adaptive filter. If the beams are not constructed in this manner, then a tractable mathematical expression for either performance measure would not be possible.

Knowing the input statistics (mean, variance), it is possible to calculate an estimate of the signal-to-noise ratio at the output of the filter and compare the estimated theoretical value to the value obtained when using the matrix filter. We use signal-to-noise ratio because it is a good measure of detection performance. If we call the system

output $\underline{y}(f)$, then we can define the signal-to-noise ratio (S/N) to be the ratio of variance of signal power to the sum of the variances of the noise power, i.e.,

$$S/N = \frac{H^T(f)G_{ss}(f)H^*(f)}{H^T(f)G_{nn}(f)H^*(f)} \quad . \tag{6.4}$$

If we assume an output of the form

$$\underline{y}(f) = H^T(f)\,\underline{x}(f) \tag{6.5}$$

and if we construct the antenna pattern outputs in the preceding form, then we can use the above definition of signal-to-noise ratio to observe, on an iteration to iteration basis of the adaptive weight calculation, how the S/N is changed at the output of the adaptive filter. The derivation of the signal-to-noise ratio for each beam appears in Appendix A.

The following generic cases have been simulated and the parameters used in the simulation appear in Tables I and II.

<u>Case 1</u>. Beam 1 = $x_1 = (s1)s_1 + (a)s_2 + (b)s_3 + (cc)sn_1$

Beam 2 = $x_2 = (f)s_1 + (s2)s_2 + (g)s_3 + (dd)sn_2$

Beam 3 = $x_3 = (h)s_1 + (p)s_2 + (s3)s_3 + (ee)sn_3$.

This case corresponds to the physical situation where there is coupling of signals into all beams.

TABLE I

COUPLING COEFFICIENTS

	s1	a	b	cc	f	s2	c	dd	h	p	s3	ee
Figure 11	1.0	0.4	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.4	1.0	0.1
Figure 12	1.0	0.4	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.4	1.0	0.1
Figure 13	1.0	0.1	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.0	1.0	0.1
Figure 14	1.0	0.1	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.1	1.0	0.1
Figure 15	1.0	0.4	0.4	0.1	0.0	1.0	0.1	0.1	0.1	0.4	1.0	0.1
Figure 16	1.0	0.4	0.1	0.1	0.4	1.0	0.4	0.1	0.1	0.4	1.0	0.1
Figure 17	1.0	0.2	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.2	1.0	0.1
Figure 18	1.0	0.4	0.4	0.1	0.0	1.0	0.0	0.1	0.0	0.4	1.0	0.1
Figure 19	1.0	0.707	0.1	0.1	0.0	1.0	0.0	0.1	0.0	0.0	1.0	0.1
Figure 20	1.0	0.1	0.0	0.1	0.0	1.0	0.0	0.1	0.0	0.1	1.0	0.1

TABLE II

SIGNAL VARIANCE

	s_1	s_2	s_3	sn_1	sn_2	sn_3
Figure 11	1	100	1	1	100	1
Figure 12	1	100	1	1	100	1
Figure 13	1	100	10	1	1	1
Figure 14	1	100	100	1	100	1
Figure 15	1	100	1	1	100	1
Figure 16	1	100	1	1	1	1
Figure 17	1	1	1	1	1	1
Figure 18	1	100	1	1	100	1
Figure 19	1	100	10	1	1	1
Figure 20	1	100	10	1	100	10

Case 2. Beam 1 = x_1 = $(s1)s_1 + (a)s_2 + (b)s_3 + (cc)sn_1$

Beam 2 = x_2 = $(s2)s_2 + (dd)sn_2$

Beam 3 = x_3 = $(p)s_2 + (s3)s_3 + (ee)sn_3$.

This beam configuration corresponds to coupling of signal two (s_2) and signal three (s_3) into Beam 1, and signal (s_2) into Beam 3.

Case 3. Beam 1 = x_1 = $(s1)s_1 + (a)s_2 + (cc)sn_1$

Beam 2 = x_2 = $(s2)s_2 + (dd)sn_2$

Beam 3 = x_3 = $(p)s_2 + (s3)s_3 + (ee)sn_3$.

This beam pattern corresponds to a coupling of signal two (s_2) into both Beam 1 and Beam 3 and no other coupling.

It is important to note that the self noise (sn_1 to sn_3) contained in each beam only limits the maximum signal-to-noise ratio obtainable and the effectiveness of the algorithm can best be measured by comparing percent improved or percent of theoretical maximum attained rather than absolute levels.

All signals used in these simulations are assumed to have Gaussian distributions with zero mean and adjustable variances. The variable variance allows one to vary the power contained in any signal.

6.2 Algorithm Parameter Determination

A gain sequence of the form

$$\mu_n = \frac{S_c}{n^i} \qquad 0.5 < i \leq 1.0 \qquad (6.6)$$

has been used in the experiments. This gain sequence was derived from both speed and stochastic stability considerations. It not only gives fast initial improvement and has built-in statistical smoothing properties, but it is easy to determine the parameters necessary for its use in the stochastic adaptive algorithm which determines the matrix filter.

From the analysis of the idealized algorithm, a gain, S_c, can be determined which lies on the stability boundary of the dynamical control system representation (see Figure 8) of the stochastic recursive algorithm.

This gain, S_c, has been determined from the convergence conditions of the idealized algorithm to be:

$$S_c = \frac{2.0}{\lambda_{max}} , \qquad (6.7)$$

where λ_{max} is the largest eigenvalue of the input power spectral density matrix $G_{xx}(f)$.

The power in a beam is a good approximation to the desired eigenvalues and a circuit that measures the power in a beam can be easily implemented. It is also possible to bound the maximum eigenvalue

of the input power spectral density matrix. If we define the trace of a matrix to be the sum of the diagonal terms, then the bound can be written as

$$\lambda_{max} < \text{trace}[G_{xx}(f)] \quad . \tag{6.8}$$

In the experiments that were performed, the eigenvalues were obtained both from the power measurement method and from the more accurate analytical method, the QL algorithm (Wilkinson, 134). The QL algorithm finds the eigenvalues and eigenvectors by three stage processes consisting of finding a series of similarity transformations (QL matrices), which reduce the original matrix to a diagonal matrix with the eigenvalues along the diagonal. The eigenvalue algorithm is based on three simple principles from linear algebra (Hohn, 58). These principles are:

1. Any Hermitian matrix may be reduced to tridiagonal (Hessenberg generally) form by elementary similarity transformations.
2. Any symmetric matrix A can be diagonalized by an orthogonal similarity transformation and the matrix which is used to diagonalize A has as its columns an orthonormal set of eigenvectors for A.
3. Similar matrices have the same eigenvalues (characteristic roots) but different eigenvectors which are related by the matrix of transformations.

The special purpose eigenvalue and eigenvector routine consists of three parts. The first part is the transformation of the original

complex Hermitian matrix to tridiagonal form. The second part of the method is the applicaton of the implicit QL algorithm to the resultant tridiagonal matrix to obtain its eigenvalues and eigenvectors. Since the eigenvalues of the original complex Hermitian matrix are preserved under the similarity transformations used in the algorithm, it only remains to back transform the eigenvectors of part two to the eigenvectors of the original complex Hermitian matrix using the similarity transformations of part one.

If we consider a unit of time as the time required to do one multiplication and one addition, and N is the order of the original complex Hermitian matrix, then the total time needed to obtain all of the eigenvalues and eigenvectors is $4/3N^2$ units. For nultiplication times on the order of 300 nanoseconds, this method of calculating the eigenvalues and eigenvectors can be considered a real time performer.

One could employ many different strategies to both accelerate the search and obtain filter weights closer to the optimum. No effort has been made, however, to optimize the search strategy. The motive behind the adaptive filter was to develop a method which gave the greatest improvement in S/N in the fewest number of iterations. These attributes make the system applicable for on line use in an adaptive processor. It is theoretically possible to develop a search strategy for the recursive algorithm whereby one approaches the optimum within a small error. To do this would require many more computations at each iteration and the speed of the algorithm would be slowed considerably. All methods to accelerate convergence require taking more observations at a given iteration of the algorithm and this clearly slows down the

speed of the estimation process. It can be seen from the experimental results that both a modification of the special gain constant, S_c, and the infinite sequence, $1/n^i$, can accelerate convergence. No effort has been made to accelerate the convergence rate by varying the exponent i. The main reason for this is that one design objective of this adaptive algorithm is that the algorithm must be obtained in as computationally simple a manner as possible within the constraint of rapid enough convergence for real time use of the algorithm. These design objectives can be obtained with the simplest exponent which is one. If even faster convergence is desired, then a variation of the exponent is required.

The recursive algorithm used to derive the matrix filter can be implemented using $2N^2$ complex multiplications, N^2 real multiplications and $2N^2$ complex additions (N is the number of beams) per iteration. With present computing speeds the algorithm can be implemented in real time.

6.3 Adaptive Algorithm Applications

We can write the matrix filter for the three beam system as

$$H(f) = \begin{bmatrix} h_{11}(f) & h_{12}(f) & h_{13}(f) \\ h_{21}(f) & h_{22}(f) & h_{23}(f) \\ h_{31}(f) & h_{32}(f) & h_{33}(f) \end{bmatrix} . \qquad (6.9)$$

The stochastic approximation algorithm will simultaneously derive all the terms in $H(f)$ and then the filter is applied in recursive

fashion to the three beam system to see how well the interfering noises in each beam are removed as the system evolves in time.

Two applications of the decoupling matrix are considered. The physical implementation in control system form of either case of the adaptive filter appears in Figure 10. The first application is when the filter coefficients along the diagonal are not constrained to be one, and the second application is when the diagonal filter coefficients are constrained to be one. If the experiment is set up in a physically correct manner, then it makes no difference in the performance of the algorithm, except for differences in attainable mean square error, whether the diagonal terms are constrained or not. Sample cases have been run for each case and the experimental verification of the above statements are discussed. One of the advantages of operating in the constraint mode is that there is an upper bound on the mean square error or a lower bound on the signal-to-noise ratio. While this might be considered an advantage for the long term operation of the algorithm, it has little effect on the time frame of interest considered here.

Constraining the diagonal terms of the matrix decoupling filter to have a gain of one is what is commonly referred to in adaptive array processing (Angerson, 5; Applebaum, 10; Claerbout, 25) as maintaining the main beam. This is done in practice to prevent boresight signal cancellation. To prevent main lobe signal cancellation, the artifice of a pilot or reference steering signal is used. The pilot signal inhibits the adaptive processor from responding to signals with prescribed directional characteristics. This technique for maintaining the main lobe is the equivalent of incorporation of a spatial filter to

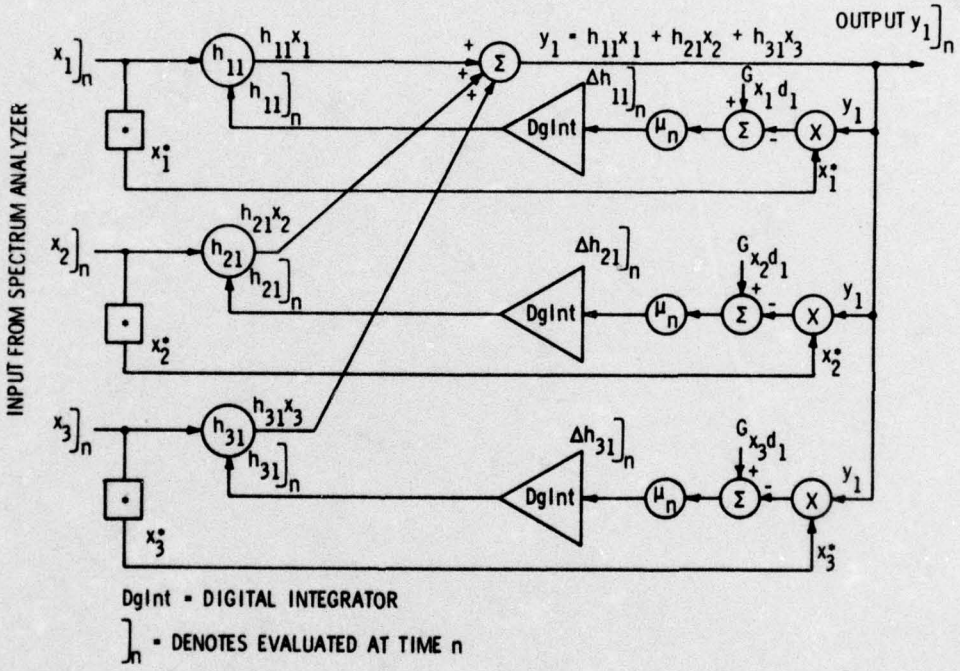


Figure 10. Implementation of Adaptive Filter Decoupling Network

restrict the gain of the adaptive processor in the boresight direction, and is well known in practice. By making the gain one along the diagonal, it insures that no more signal than is already in that beam is transmitted through the filter. A perusal of Figure 4 illustrates the physical reasoning behind the mathematical technique of the diagonal constraint.

This idea is also equivalent to the mathematical idea of a projection operator (Halmos, 53). We are forcing (constraining) certain filter values to be on a certain part of the minimization surface corresponding to the rules of the constraint. The idea of projection operator is well known in linear algebra (Stoll, 120). The method used here produces exactly the same result. In essence, it limits certain matrix filter coefficients by projecting a particular filter coefficient to have a certain value corresponding to the constraint boundary and no other. Many authors have used the idea of a projection operator. Rosen (103) used the idea in basic gradient search techniques. However, he had no means to keep the projection exactly on the constraint boundary. Frost (47) has applied the same idea to array processing. With the projection operator used by Frost, one can guarantee that the projection will always be along a boundary. The technique used here is equivalent, in operation, to Frost's projection operator.

Mendel and Fu's (86) first order method of projection operator forces the adaptive algorithm to use values of the estimated variable only if the variables are within a cube. They know from physical considerations the limits on some of the random variables to be estimated.

For the case of the beam patterns considered here, the use of this technique is indicated from the consideration of the weak signal strong interference assumption. For, if the experiment is constructed in a physically realizable manner, then there cannot possibly be a situation where a signal has more power in the beam where it is considered a noise than in the beam where it is considered a signal. Since it is unreasonable to allow the filter to feed more signal into a beam than is already there, the technique of constraining the filter estimate, by use of the projection operator technique, is dictated by physical considerations.

In stochastic approximation, sufficient conditions for mean square and probability-one convergence are satisfied within some unknown bounded convex set. If a convergence region were known, a reflecting barrier at the boundary would solve the global convergence conditions and the estimate would converge in mean square and with probability one. Davisson (33) has proven that if we let A be the event where the estimate sequence remains within some given convergence region, then convergence conditioned on A occurs in mean square and with probability one because the sequence of estimates is the same as if a reflecting barrier were placed at the boundary. Davisson showed that the unconditional probability of convergence is bounded below by the probability of the event A. The point is that with a reflecting barrier, convergence is guaranteed. Without a reflecting barrier, global conditions can be bounded by the probability of event A. When there are equality constraints imposed on the solution, the reflecting barriers can be considered, by analogy with Markov processes, absorbing

barriers and the absorbing barriers then give the same result as the projection operators. The mathematical meaning of an absorbing barrier can be given by the following definition.

Definition 6.1

A state in a Markov chain is said to be absorbing if $P(j,i) = 0$ for all states $i \neq j$. In other words a state j is absorbing if $P(j,j) = 1$. One cannot leave an absorbing state. The probability, $P(i,j)$, is defined as the transition probability of being in state i and moving to state j.

For physically realizable cases, the technique of constraining the diagonal terms of the matrix adaptive filter is analogous to both the projection operator technique and the reflecting barrier. It is similar to the projection operator in that the linear transformation necessary to convert the old matrix filter to the new filter (satisfying the constraint) can be determined from the solution of a set of simultaneous linear equations. If we assume that the matrix A is the desired projection operator, then we can write:

$$A\, H(f)_{OLD} = H(f)_{NEW} \quad . \tag{6.10}$$

This set of equations guarantees that the diagonal terms of the new matrix filter are constrained to be one. The constraint technique uses the idea of an absorbing barrier in that it uses a priori knowledge of the solution to constrain certain matrix filter coefficients to have certain values.

6.3.1 Effects of Special Gain Sequence Recall that the generic form of the special gain sequence is

$$\mu_n = (S_c)\ \frac{1}{n^i} \qquad 0.5 < i \leq 1.0 \qquad (6.11)$$

where S_c has been shown in Section 5.1.1 to be equal to $2/\lambda_{max}$. To show the effect of this gain sequence on the convergence rate of the recursive algorithm, there are two tests conducted. The exponent i is varied over its range and the proportion of the gain constant S_c used in the algorithm is varied.

It can be seen from Figure 11 that, as the exponent i is varied, the rate of convergence goes up. From the mathematical considerations of the infinite sequence

$$\frac{1}{n^i} < \frac{1}{n^{i-1}} < \frac{1}{n^{i-2}}\ , \qquad (6.12)$$

it can be observed that as i increases from 0.51 to 1.0, the corresponding terms in each series decrease. This means that the effect on the stochastic algorithm by the sequence increases as the exponent i decreases if we consider succeeding iterations of the algorithm. This means that as i decreases, it takes the gain sequence a shorter time to effect the same increase in output S/N. The price one pays for this acceleration is noisy results. As the exponent i is varied from 0.6 to 1.0, one can observe the effect on the output S/N. The exponent $i = 0.6$ has the fastest increase but it is so noisy that the filter coefficients would not be usable in a practical system. The case where

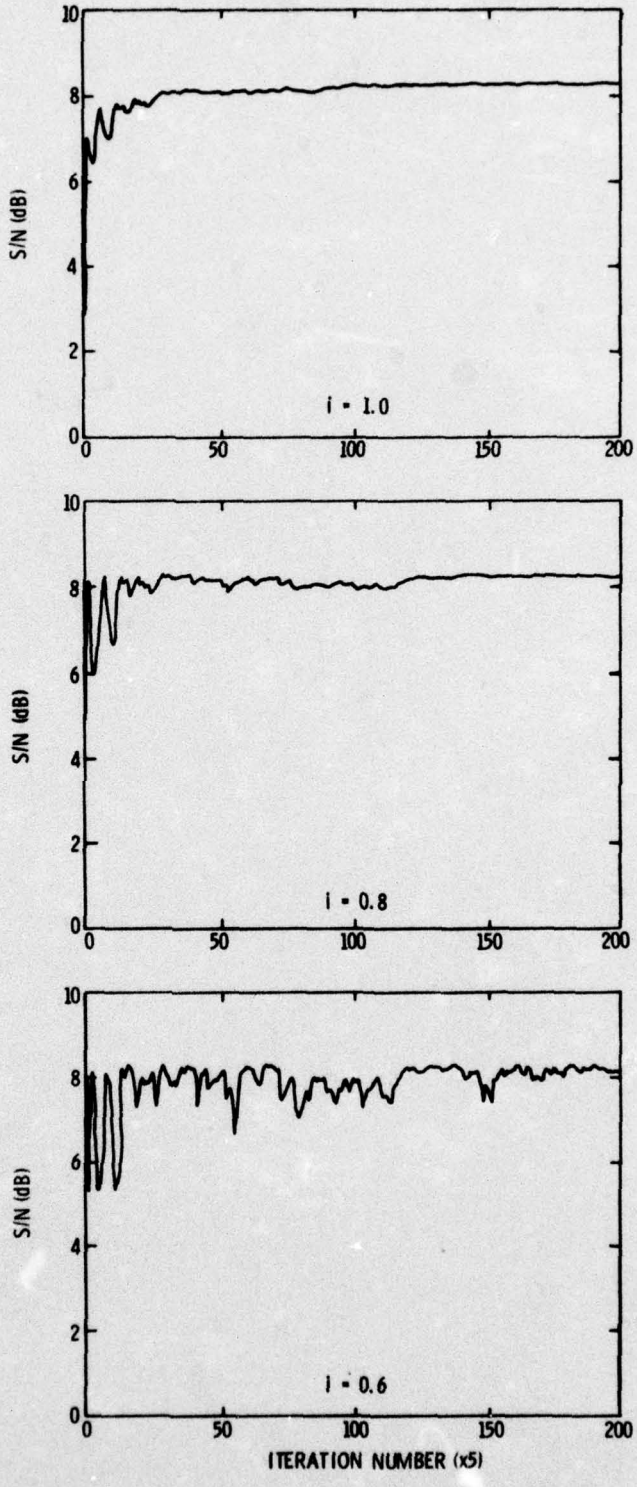


Figure 11. Exponent Variation

$i = 0.8$ is slower but the filter values have a smaller variance and could be usable. The case where $i = 1.0$ is the smoothest of all and yet it is fast enough for on-line use. With the stochastic approximation technique, it is possible to obtain fast convergence and yet obtain filter coefficients with small variance. The exponent $i = 1.0$ has the added advantage of computational simplicity and this advantage is not to be underestimated in practical system use.

The variation of the gain constant S_c also has effects on the stochastic algorithm. It has been derived that the optimum gain constant is

$$S_c = \frac{2.0}{\lambda_{max}} \quad . \tag{6.13}$$

If this gain constant is rewritten as:

$$S_c^{\,1} = \left(\frac{2.0}{\lambda_{max}}\right) G \quad , \tag{6.14}$$

then G can be varied and its effect on convergence can be observed. The results of varying G over the range

$$0.07 \leq G \leq 0.9 \tag{6.15}$$

can be seen in Figure 12.

It can be seen that there is a significant difference in the values obtained for the output S/N for the different values of G. For 20 iterations, the values of output S/N varied from -10 dB for $G = 0.07$

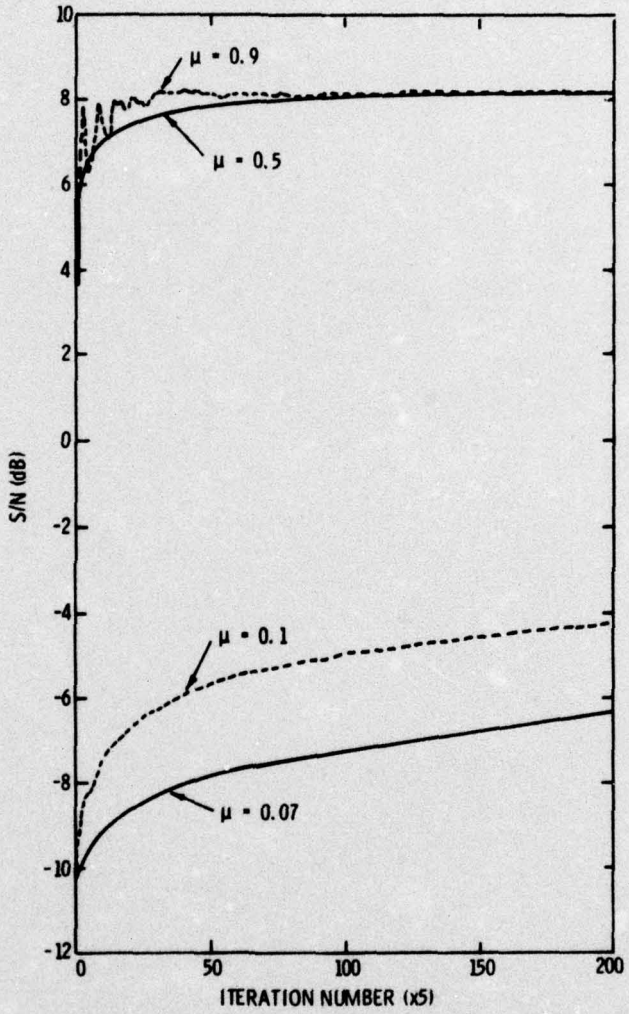


Figure 12. Gain Constant Variation

to +6 dB for G = 0.9 . For G = 0.9 , the algorithm attains a S/N of +6 dB in less than ten iterations. For G = 0.7 , it takes approximately 25 iterations. For G = 0.1 and 0.07 , the algorithm never does reach +6 dB in the running time of the test. It can also be observed that the G = 0.9 case is noisier than any other case. This result can be expected since the algorithm is operating close to its stability limit. The case where G = 0.5 is a very smooth curve and would be desirable for practical use. However, it is possible to operate the algorithm near its stability limit and still obtain small variance in the filter coefficients. For all of the cases tested in the experimental setup, a gain G of 0.7 was used.

These results are significant because they show that the algorithm must be operated as near to the maximum gain constant S_c as possible to be able to operate the algorithm on-line. If a deterministic algorithm were operated with values of G = 0.01 , the resultant filter coefficients would not be usable. The deterministic attempts at adaptive filters (Griffiths, 51) have suggested that G be 0.001 or less to obtain usable filter coefficients. The great power of this stochastic approximation method is that it allows the fastest increase in output S/N but yet its built-in statistical smoothing qualities permit use of the filter coefficients even at these high levels of gains. This result has not been obtained with any deterministic algorithm.

A case where G was equal to 1.5 times the maximum gain was tried and it did indeed become unstable as predicted by theory. No plots were obtained because the filter coefficients became too large to plot after only a few iterations.

Under the weak signal, strong interference assumption, the adaptive filter performs its function exactly as predicted by theory. In all the cases tested in the simulation, greater than 50% [on a decibel (dB) basis] of the strong interfering signal was removed in less than 20 iterations of the adaptive algorithm.

In the case where there is no interfering noise, the adaptive filter initially exhibits a decrease in S/N. The amount of decrease depends directly on the strength of the signal in the no-noise beam and whether the implicit constraints are applied. The reason for this is twofold. The first reason is that the filter values are initialized to an identity matrix which, for the beams with no noise in them, corresponds to the optimum filter. Any deviation from this point will result in a degradation in the performance of the beam with no noise. Thus, while the filter weights are adapting in other dimensions, the beam with no noise suffers. If the first big step takes the search far from the optimum, then if the power is large, the variable gain does not have enough length, i.e., it needs a larger step size to progress rapidly in the no-noise beam search surface. Theory guarantees eventual success but the process can become inexorably slow. From the theory of minimization or search techniques, this is a predictable result. Since the variable gain sequence determines the convergence rate, the longer the recursive algorithm is run, the less effect the variable gain has on the minimization procedure. A more powerful search strategy can considerably enhance this process. This result only applies to the no-noise beam and not to any beams containing interfering noise because the largest steps are made toward the optimum and not

away from the optimum in the no-noise beam. Figure 13 shows a case where the simulation was run to over 5000 iterations of the stochastic algorithm. It can be discerned from this plot that the filter is approaching the optimum with a small but finite slope. To obtain the true optimum with this particular search strategy was not possible due to limitations on computer time. Running the algorithm, a large number of iterations was not desirable because the long term results were not of primary interest.

When the power of the signal in the beam is equal to the other beams or the implicit constraints are used, then it can be seen from Figure 14 that the S/N does not decrease as much as when these conditions are not fulfilled. It can be seen from Figure 14 that the S/N only goes down to about 12 dB (from a theoretical maximum of 20 dB) then starts to increase almost immediately.

It has been shown from the experimental results that the more noise there is in adjacent beams, the longer it takes the beam with no noise to start increasing its S/N. This fact can also be gleaned from intuitive considerations. The filter is derived so as to reduce the effect of interfering noises, and the filter performs this function before it attacks any other inaccuracies. Thus, when there are large interferences in a given beam, the filter weights must adjust to remove both large interferences and large power, and it is slow in attacking the inaccuracies of the no-noise beam because the mean square error is smaller, on a percentage basis, in the no-noise beam than in the beam with large interferences. When there are few interferences and the power is not large, the filter weights do not have to adapt greatly to

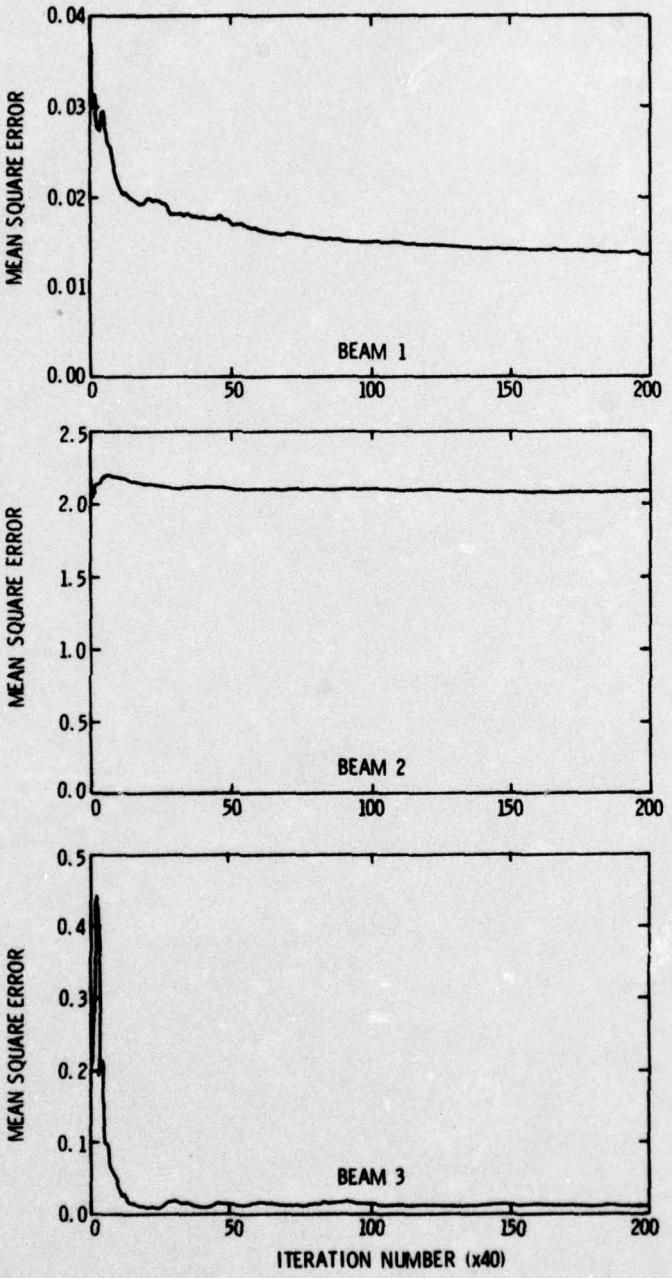


Figure 13. Long Term Variation

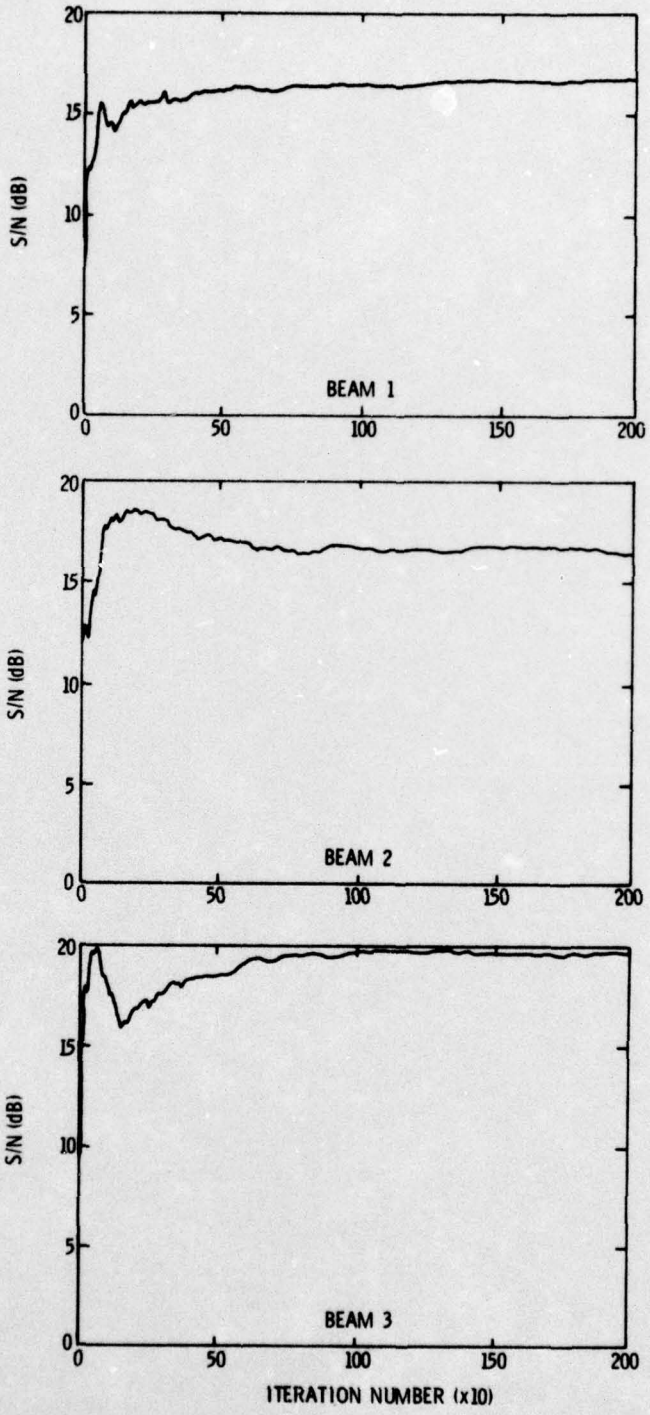


Figure 14. Beam Power Effects

remove the interfering noises, and thus the filter weights can adjust out the interfering noises and then work on the beams where there are no interfering noises. The more power and the greater the number of interfering noises, the longer it takes the beam with no noise to approach its own theoretical minimum mean square error or maximum signal-to-noise ratio.

Singer and Frost (115) have derived an expression which bounds the steady state error covariance of the Wiener filter. The relation shows that the greater the interfering noise intensities, the greater the steady state filtering error. This means that with high interfering noise powers, a small error in filter coefficients results in a large increase in mean square error. This result is deceiving because if all the signals were normalized, this effect would not appear and a small error in filter weights would result in a small mean square error. This attribute of the filter was also observed in the experimental tests. The use of the largest eigenvalue to speed convergence helps considerably when there is a large spread in eigenvalues. If a more sophisticated search procedure were devised where each eigenvalue was used in sequence after a percentage change in the Frobenius norm, then this phenomenon of the matrix filter could be alleviated even more and possibly eliminated altogether.

Using a more sophisticated search procedure similar to those employed in deterministic systems would decrease the time to reach the optimum. Kushner (74,75) has tried to formulate search strategies for stochastic type algorithms but these techniques have been proven for a limited number of cases.

If the filter were started at some other point, then depending on where it was started, the S/N could either increase or decrease. If one was far from the optimum, then the S/N would increase, but in this simulation all filters are initialized to the same identity matrix; i.e., optimum filter for no noise, and thus, in the case of the beam with no noise, its S/N ratio goes down for a time and then starts to increase after a number of iterations of the algorithm. Since we assume no a priori knowledge as to where to start the recursive algorithm, this initialization is as good as any. Of course, any a priori knowledge could be used to alleviate this problem and increase the convergence rate.

6.4 Generic Case Results

It has been assumed in the simulations that all signals are uncorrelated i.e.,

$$E\{s_1 s_2\} = E\{s_1 s_3\} = E\{s_2 s_3\} = 0 \tag{6.16}$$

and that all self noises are uncorrelated both with the signals s_1, s_2 and s_3 and with other self noises

$$E\{sn_1 sn_2\} = E\{sn_1 sn_3\} = E\{sn_2 sn_3\} = 0 \ , \tag{6.17}$$

$$E\{sn_1 s_1\} = E\{sn_1 s_2\} = E\{sn_1 s_2\} = 0 \ , \tag{6.18}$$

$$E\{sn_2 s_1\} = E\{sn_2 s_2\} = E\{sn_2 s_3\} = 0 \tag{6.19}$$

and

$$E\{sn_3 s_1\} = E\{sn_3 s_2\} = E\{sn_3 s_3\} = 0 \ . \tag{6.20}$$

The position of a signal or noise in any beam is adjusted by varying the coupling coefficients (Table I). The power of the signals or noises is adjusted by changing the variances of the signals. The power contained in each case tested appears in Table II. The coupling coefficients give the amount of signal which is contained in another beam as noise. The actual form of the individual cases can be gleaned from a perusal of Table II and Equations (6.1), (6.2) and (6.3). A graphic picture of the significance of these coefficients can be seen in Figure 4. Unless otherwise indicated, only the curves of S/N for beams with interfering signals are included.

Figure 15 shows that the S/N ratio goes up in direct relation to the way the mean square error goes down. From Figure 15, we can see that the mean square error decreases initially and the S/N goes up in the same number of iterations. This is not a surprising result and it is not only intuitively pleasing but also verified by experimental tests. The fact that the mean square error and the signal-to-noise ratio are related in an inverse manner is not a new result. Riegler and Compton (100) also verified in their experiments that the MSE goes up in the same way that the S/N goes down. For the remainder of the tests, the signal-to-noise ratio is used as the criterion of system performance. Mean square error is used only to illustrate certain results.

The results of the tests for the first generic case can be seen in Figure 16. Figure 16 contains the test results when there were no constraints. Signal s_2 had 100 times the power of the other two signals and it was coupled as an interfering signal in both of the other

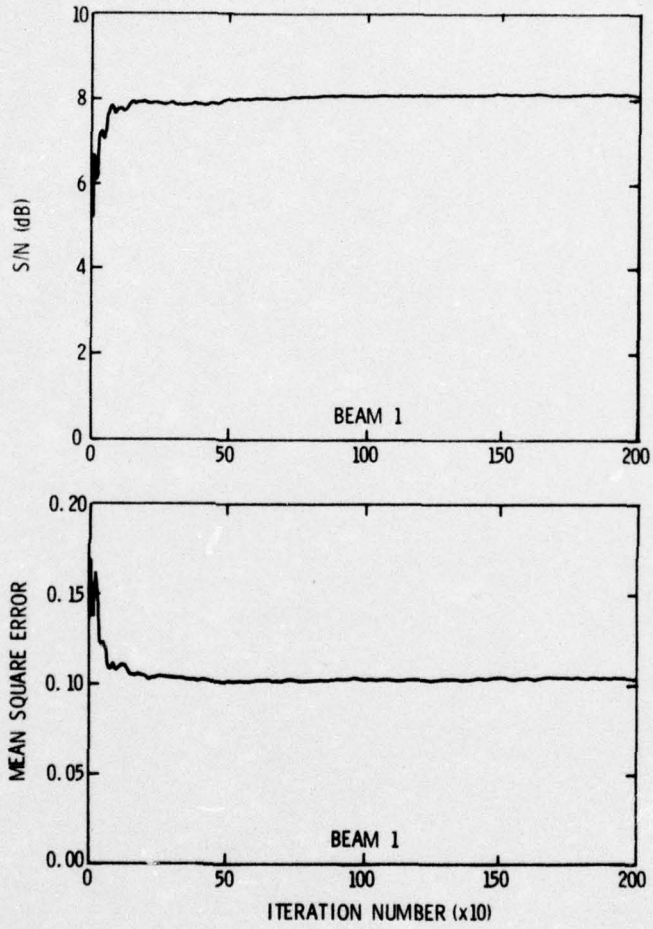


Figure 15. MSE Versus S/N

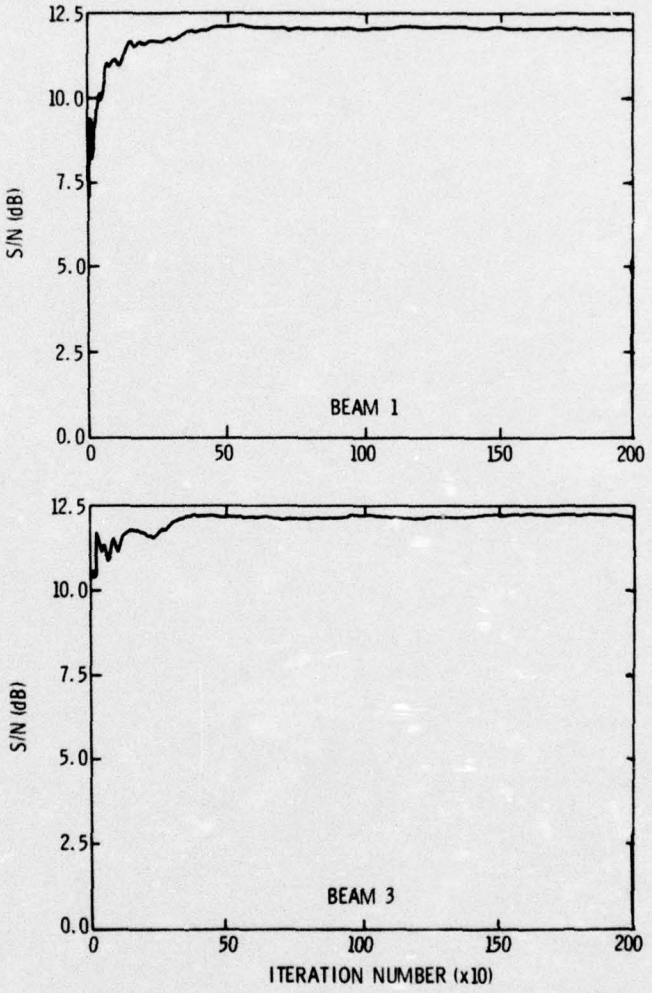


Figure 16. Generic Case One

beams. For this case, there were two interfering signals, s_2 and s_3, in Beam 1 and two interfering signals, s_1 and s_2 in Beam 3. Even with these large interfering signals, the interference effect on the desired signals was removed in less than 20 iterations of the algorithm. The output S/N ratio obtained greater than 50% of its theoretical maximum value in less than 10 iterations. The percentage is calculated by taking the ratio of the increase in S/N to the maximum attainable S/N. The output S/N started at -12.5 and increased to +8 in approximately 20 iterations.

It can be seen from Figure 17 that the constraint mode operated in exactly the same manner as the no constraint mode. It can be stated that the only difference discovered in the operation of the stochastic algorithm between the two cases (constraint and no constraint) was in beams which had no noise. *This result was also discussed in Section 6.2.*

Figure 18 shows the case where one signal s_2, has one hundred times the power of the other two. This large signal was injected into the other two beams by the coupling coefficients which appear in Table II. Another signal, s_3, was also injected into Beam 1. Thus, Beam 1 had two interfering signals and Beam 3 had one interfering signal. The interfering signal in each beam was decreased in the output S/N by over 50%. The output S/N started at -12.8 and increased to +10 dB in less than 20 iterations. In Beam 3, the output S/N started at -12.4 dB and increased to +10 dB in less than 20 iterations.

Another test was conducted for generic case 2 and the results appear in Figure 19. In this test, the power of s_2 was still 100 times the power of s_1 but the power of s_3 was increased to 10 times

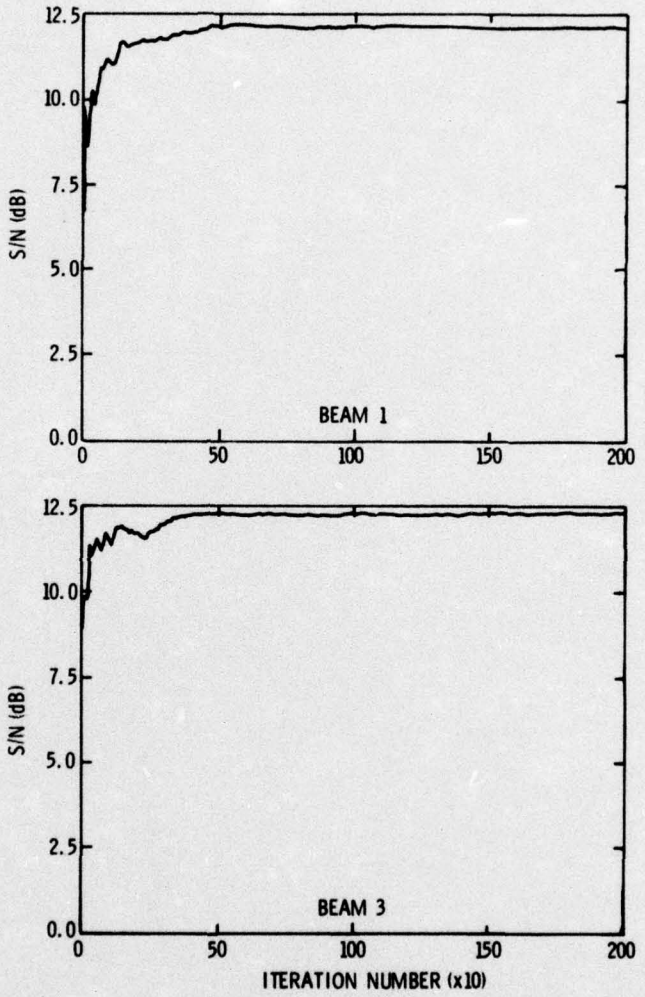


Figure 17. Generic Case Two

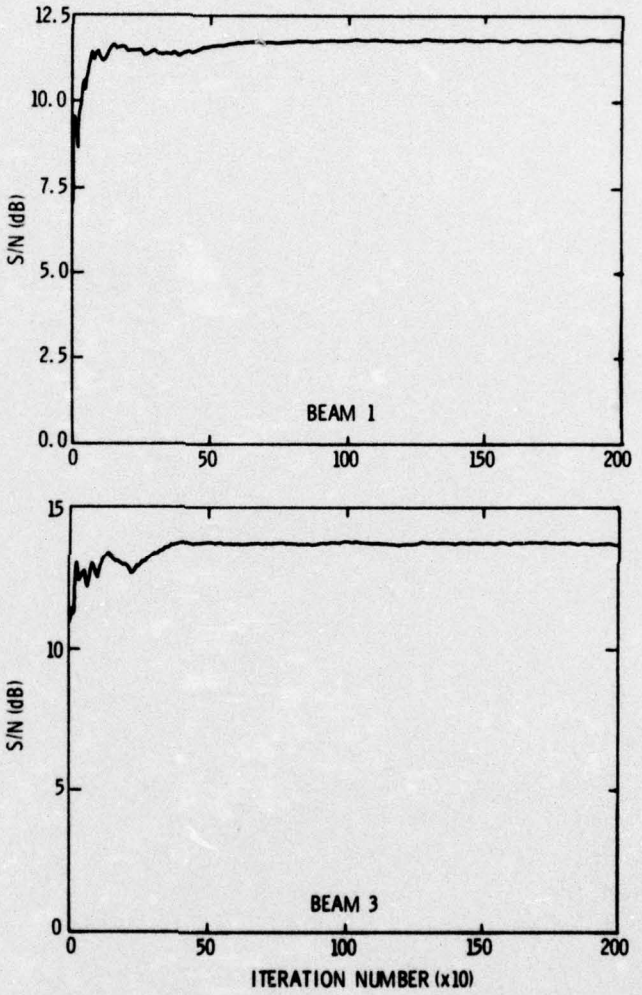


Figure 18. Generic Case Two

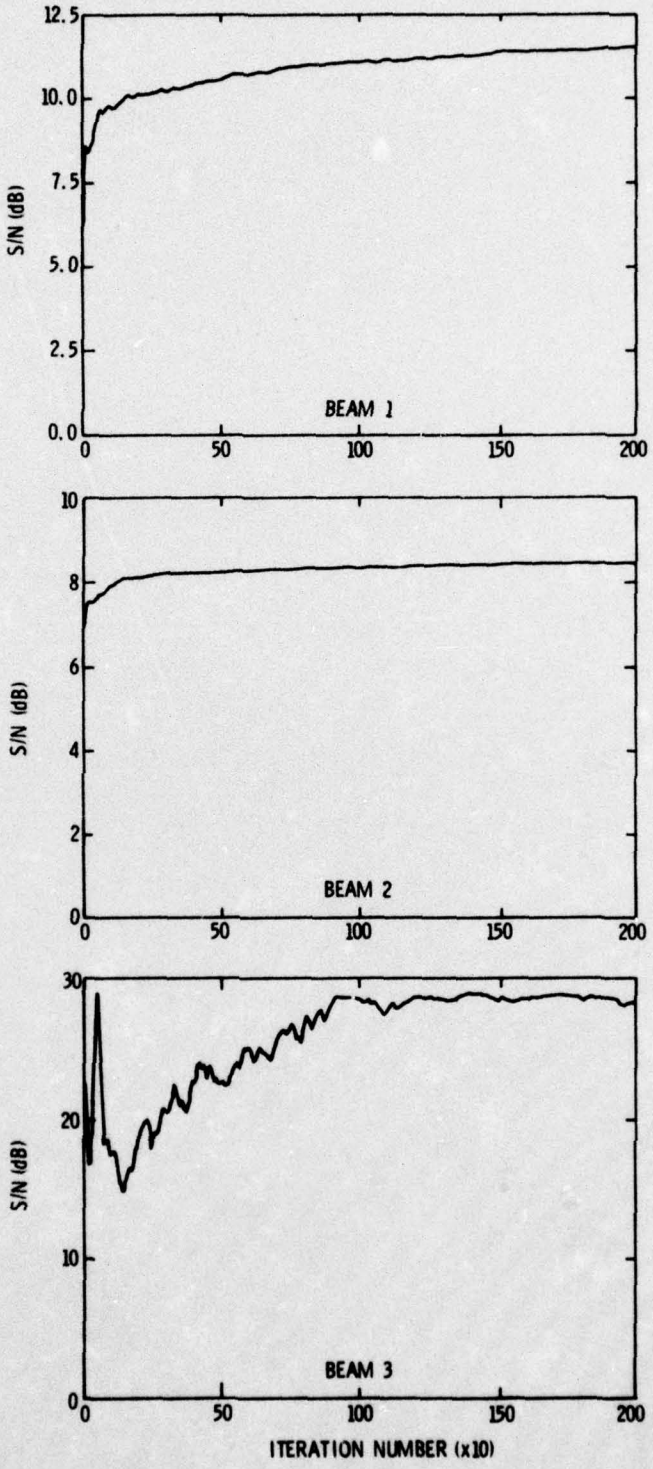


Figure 19. Worst Case: Beam Intersection

the power of s_1. Both s_2 and s_3 were injected into s_1 as noises. The signal s_1 was completely masked and Beam 1 had an extremely poor initial S/N of -18 dB. This case also had the added hindrance in that the interfering signal s_2 was put near the crossover point of the beams. This should be the most difficult position to remove an interfering noise. In spite of these worst case conditions, the algorithm increased the output S/N to +8 dB in less than 20 iterations. This was an increase of 20 dB on a theoretical perfect solution of 38 dB. Surely, an impressive result.

The results for generic case 3 appear in Figure 20. This case was constructed with s_2 having 100 times the power in s_1 and s_3 had 10 times the power in s_1. Signal s_2 was coupled into both Beams 1 and 3 and there was no other coupling. As can be seen from Figure 20, both Beam 1 and Beam 3 attained greater than 50% of their theoretical maximum S/N in approximately 25 iterations. Beam 1 started at approximately -12 dB and increased +10 in approximately 25 iterations.

6.5 Stopping Rule

The following stopping rule is being proposed to determine when the optimum is attained with a certain accuracy in the probabilistic sense. It is being proposed only as an indication of needed future results for long term properties of the recursive algorithm. There are special stopping rules for regular iterative methods which are based on the comparison of the last two iterations $H(n-1)$ and $H(n)$ of the iterative algorithm. The stopping rule being proposed for the probabilistic algorithm is to compute a running average of the form

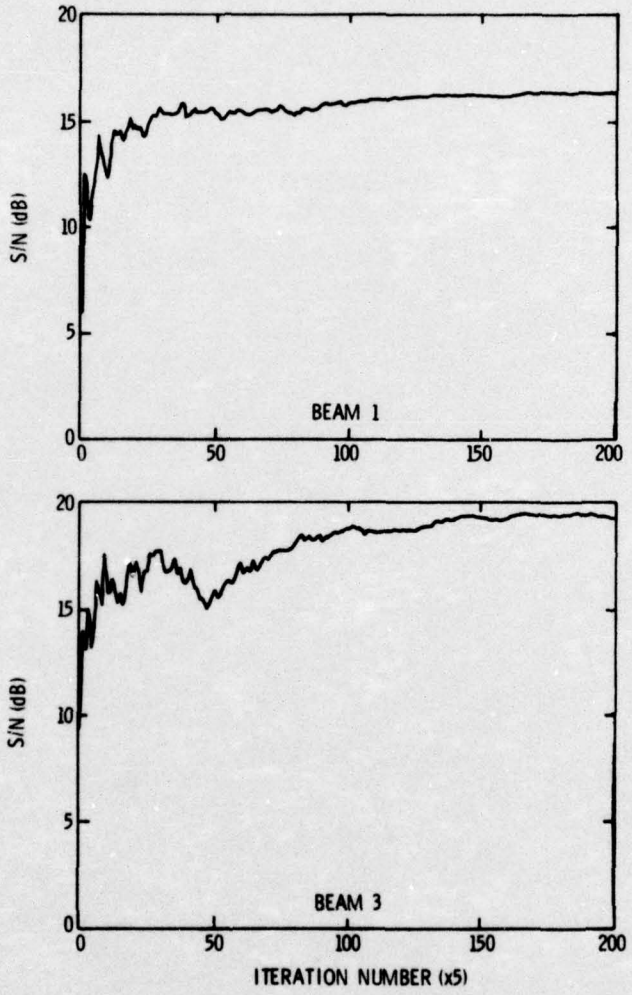


Figure 20. Generic Case Three

$$RAV_k^{i,j}(f) = \frac{1}{N} \sum_{k=n}^{n+N} H_k^{i,j}(f) \quad , \tag{6.21}$$

where k is the iteration number of the recursive algorithm and (i,j) represents the position of the filter coefficient in the matrix filter, and then compute the Frobenius norm of the difference of any two iterations of

$$\begin{aligned} RAV_N^{i,j}(f) &= ||RAV_{[(k+1)N]}^{i,j}(f) - RAV_{[kN]}^{i,j}(f)|| \\ &= ||DIFF(f)|| \leq \varepsilon \end{aligned} \tag{6.22}$$

where ε is a small real number.

If the Frobenius norm of the difference is smaller than a certain number, ε, the algorithm can be stopped and one can be assured of the required accuracy in the probabilistic sense.

It is, however, difficult to choose, N, the number of iterations of the adaptive algorithm to average over. This quantity is dependent upon the statistics of the underlying process and is difficult to determine a priori. Of course, this stopping rule does not apply to the on line system because it requires a great amount of computation. This stopping rule has not been used in this analysis because the long term characteristics of the stochastic adaptive filter were not desired.

CHAPTER VII

CONCLUSIONS AND EXTENSIONS

7.0 Problem Statement

The purpose of this research is a derivation of a complex stochastic processor as the adaptive implementation of the multi-input, multi-output Wiener filter. The stochastic adaptive processor operates in an environment where the noise statistics were not known a priori. It operates under the weak signal, strong interference assumption. The adaptive matrix filter was to be able to remove the strong interference and increase the output S/N in a few iterations of the recursive algorithm (derived from stochastic approximation considerations) used to derive the adaptive matrix filter. The stochastic adaptive processor requires only complex multiplications and additions (Section 6.2). On a digital computer the adaptive processor can be implemented simply through the use of arithmetic units. With present computing speeds, the recursive algorithm has as its goal the ability to be implemented in real time and to obtain the adaptive matrix filter on-line.

7.1 Results and Future Work

The stochastic adaptive processor's ability to perform the above functions has been demonstrated by the computer simulation results.

The adaptive processor was designed under the weak signal, strong interference assumption and, in all cases tested, this objective was

attained. One can, as a consequence of the results obtained, expect a significant improvement in signal-to-noise ratio at the output of the adaptive processor.

The stochastic processor removed the strong interfering signals in all cases tested in a small enough number of iterations to make the algorithm an on-line performer. The adaptive processor demonstrated its ability to rapidly increase the output S/N and make it a desirable input to a detection system. Since the stochastic adaptive algorithm takes into account the fact that its inputs are random processes, it is better able to handle the processing requirements than any deterministic algorithm. Its greatest assets would seem to be the ability of fast increase in S/N and the built in ability to handle stochastic processes. One important extension of the work in this research would be the use of stochastic approximation when the input random processes are non-stationary.

The extension of adaptive array processors to the multi-input, multi-output case gives impetus for new types of processing systems (including the likelihood-ratio detector part of the receiver) beyond those possible with single output array processors.

The extension of the adaptive array processors to complex signals offers new possibilities not available in theory based on real signals. Further theoretical development is facilitated by analytic representation of band pass signals which naturally leads to complex signals.

The most attractive method for obtaining the optimum solution would seem, at first glance, to be matrix inversion. An iterative

scheme for matrix inversion would seem a most promising method. But even when $\underline{R_{xx}}^{-1}(t)$ or $\underline{G_{xx}}^{-1}(f)$ is obtained, there still must be a complex matrix multiplication to determine the optimum filter $[\underline{h}^{*}_{OPT}(f) = \underline{G_{xx}}^{-1}(f)\ \underline{g_{xd}}(f)]$. Since the adaptive algorithm calculates the optimum filter directly and does not need a succeeding matrix multiplication, a savings accrues to the adaptive filter. A more important consideration is how many iterations and how much arithmetic is involved in the iterative matrix inverse. Since $\underline{R_{xx}}^{-1}(t)$ is not known exactly, the inverse must be continually updated. It is not clear that any type of matrix inverse that must be updated has any advantage over the stochastic adaptive algorithm described here. In fact, the simplicity of calculation and the ease of updating seems to indicate that, at present, the stochastic adaptive algorithm is superior. It is not proven, at this time, that any iterative matrix inverse has a smaller computational load than the adaptive filter.

With only a few exceptions, all previous adaptive processors have been based on real signals and finite impulse response filters with real weights. The present adaptive matrix filter operates completely on complex input signals. In many applications this approach leads to a simpler overall system.

The first objective, to derive the optimum filter, was obtained through the use of the orthogonal projection lemma. This general purpose technique allows derivation of the minimum mean square error when other more standard minimization procedures fail. The idea of a projection operator is widely used in the abstract theory of vector spaces (Halmos, 53) and topology (Kolomogorov and Fomin, 71). The idea of a projection

operator is very general and has found and will continue to find diverse applications in the engineering literature. Its potential applications to engineering problems have not yet been exploited in any depth. Application of projection operators have crept into the engineering literature from many diverse problem areas. Kalman (65) used the idea to derive the minimum variance unbiased estimator which is used so often in both theoretical control system work and practical control systems. Frost (47) uses the idea of a projection operator in his adaptive processor design. Rosen (103) has incorporated the projection operator into a deterministic search or minimization procedure which is coming into widespread use.

The special purpose variable gain sequence μ_n permits the stochastic adaptive processor to obtain fast increase in output S/N, thereby enabling the adaptive processor to operate in a real time environment. Using the largest possible stable gain sequence, the adaptive algorithm obtains fast initial convergence but retains the advantages of statistical smoothing inherent in the stochastic approximation technique. This fast increase in output S/N is important in any practical application of the adaptive processor to communication systems or control systems. The long term statistical considerations are important if one needs an accurate answer as to how close to the optimum solution one can get with either a naive or sophisticated search strategy. It was seen that drastic changes in both the convergence rate and the statistical smoothness of the matrix filter coefficients can be affected by small changes in either the gain constant, or the variable gain sequence. The variable gain sequence

variation affected primarily the smoothness of the coefficients and the gain constant affected primarily the initial convergence rate. The variable gain sequence also has a marked effect on initial convergence rate but its effect is not as adjustable as the effect of the gain constant. This fact makes variation of the exponent of the gain sequence less valuable for a real time application. Better mathematical strategies for variation of both the gain constant and the gain sequence can result in both better initial speed and long term smoothness. This area of research could produce the most fruitful results in the application of stochastic approximation techniques to engineering problems. The stochastic adaptive processor successfully combines both stochastic principles and complex variables in an adaptive real time processing environment. The ability to use the stochastic nature of the input signals without slowing down its performance is a major attribute of this adaptive matrix filter.

The following method is proposed as a means of accelerating convergence of the stochastic recursive algorithm. The derivation of the optimum gain constant from considerations of the idealized form of the recursive algorithm shows that the knowledge and use of the eigenvalues of the power spectral density matrix $G_{\underline{xx}}(f)$ gives a large increase in the convergence rate. The mathematical operation of the recursive algorithm used to derive the matrix filter coefficients is not a scalar process but a vector process. Since each dimension of the vector space of the recursive algorithm has a different convergence rate, it is necessary to take this fact into account to be able to approach the fastest convergence rate. If one uses a matrix gain of

the form (for a three-beam system),

$$G = \begin{bmatrix} \frac{2}{\lambda_1} & 0 & 0 \\ 0 & \frac{2}{\lambda_2} & 0 \\ 0 & 0 & \frac{2}{\lambda_3} \end{bmatrix}$$

instead of a scalar gain

$$g = \frac{2}{\lambda_{max}} \qquad \text{or} \qquad \frac{1}{\sum_i \lambda_i} ,$$

then an overall increase in the convergence rate can be expected. If we consider each dimension of the recursive algorithm separately, then there are two possibilities for the search convergence. Either the gain in that dimension is too small and the convergence is slow or the gain is too large and there is overshoot of the minimum. These effects have been illustrated by the experimental results described earlier. In the no-noise case, the recursive filter was at the optimum but the gain was large and the first move was far from the optimum. The gain constant then decreased in size and the convergence was slow but in the correct direction. Using the knowledge of the eigenvalues in the matrix gain, it is possible to both increase the convergence rate in the noisy signals and decrease the initial movement from the optimum in the non-noisy signals. While this technique requires more arithmetic than the scalar gain, an increase in convergence rate would justify its use.

The movement in any dimension is governed by how far or how close one is to the optimum and the size of the gain. Using the knowledge of the results in this research, it is possible to adjust each dimension of the recursive algorithm separately. The most desirable situation is to be able to separately adjust the gain in each dimension depending on its position relative to the optimum and not on the general consideration of all the dimensions. With the results derived here, increased convergence can be expected and it is possible to develop a search strategy closer to the optimum one than is possible with any scalar gain. Kushner (75) has attempted to find the optimum step size in each dimension by a minimization in each dimension. The technique proposed here would take fewer calculations and it is computationally simpler but it is, of course, not the optimum movement at each step of the recursive algorithm.

The implementation of the adaptive algorithm is extremely simple. It involves only complex multiplications and additions. The time considerations in Section 6.2 show how fast the algorithm can operate with available technology. It has also been shown that it is computationally simple to obtain the parameters used in the recursive algorithm.

Contained in the convergence proofs of the stochastic recursive algorithm used to derive the adaptive filter was the important result that the filter coefficients resulting from the successive operations of the recursive algorithm were martingales. As was pointed out, this fact is not only valuable in proving convergence, but is also useful in stochastic stability considerations.

It has been shown that stochastic Lyapunov functions satisfy the martingale property. This result has many possibilities to be exploited

in the analysis of stochastic control systems. It is possible to develop a stochastic analog of deterministic Lyapunov functions. This would give the control system designer a framework in which to analyze stochastic control system which are becoming more prevalent in engineering applications. Without this framework for stochastic stability, it is very difficult to design reliable stochastic control systems when the order of the system or the number of state variables is large.

The stochastic adaptive processor has application to a wide range of different problems. It could be used as an adaptive filter for a digital communication system. It could be used in biological research for predicting a level of response in an experiment. It could be used by statisticians to estimate probability distribution functions. Control system engineers could use this technique for system identification and process control. Workers in the field of pattern recognition and machine learning could use these stochastic approximation principles to great advantage. The stochastic technique has wide applicability in seismic signal processing. In the area of array processing and digital signal processing (communication systems and seismic signal processing), the advent of the fast Fourier transform has made frequency domain methods more attractive for real time use. The present technique of adaptive processing gives a viable means for handling the complex variables that arise naturally from frequency domain problems. A much needed extension to the present work would be to build a mathematical framework in which to handle a wider range of complex variable minimization problems. Not all of the theory of real variables can be used in complex variables. A simple extension of real analysis to complex

analysis is not possible in some important problems and, since most frequency domain techniques involve complex variables, a wider framework to handle these problems is important.

A unification with the techniques used in the decoupling theory in multivariable control systems would be valuable because it would give some insight into stochastic approximation theory from the point of view of stability and observability in modern control theory. Since stochastic approximation and stability theory are related through the martingale property, it might be possible to simplify the convergence proofs and provide a tehcnique for a more optimum search strategy.

The present stochastic adaptive filter is a valuable addition to any detection or parameter estimation system. From the receiver operation curves (ROC), it can be gleaned that the higher the input S/N is to any detection system, the greater the probability of detection at a given false alarm rate. Since the present adaptive system increases the S/N to over half of its theoretical maximum in a few iterations, it can be used as an effective input to any optimum or suboptimum detector.

While the present processing system has not been considered in the framework of a complete detection system but only as an input to the detection part of any receiver, it is instructive to state the following results about the space, time, and detection factorability of optimum processors. These results become important when a general form for the various optimum array processing schemes is desired.

Middleton and Groginsky (88) were the first to consider the space-time factorability of optimum processors. They showed that factorization is not, in general, possible in optimum passive array processing systems.

The space-time structure of optimum active array processors and, in particular, the factorability of such processors into spatial and temporal operations was studied by Pasupathy and Venetsanopoulos (95). They studied the problem for linear continuous array in a reverberation-limited environment and derived the conditions on signal, reverberation, and array parameters under which such a factorization is possible. They found that the factorability was limited, in general, to narrowband systems.

Van Trees (125) has shown that for a certain class of optimal criteria and signal models, all optimum space-time processing schemes are composed of a beamformer or spatial processor followed by a scalar filter and, for plane wave signals, the spatial processor is common to all. Only the filter reflects the particular criterion of optimality selected.

The processing system considered here, where beamforming is performed before any adaptive filtering function, does not fit exactly into the preceding categories of optimum processors. One major difference is that the final filtering function is a matrix filter rather than a scalar one. While the optimality criterion are in some cases the same, it is in the particular implementation of the spatial and timporal parts of the processor that the various schemes differ.

BIBLIOGRAPHY

1. Albert, A. E., and Gardner, L. A. Jr., Stochastic Approximation and Nonlinear Regression, Riverside Press, Cambridge, Massachusetts, 1967.

2. Allen, J. L., "Array Antennas: New Applications for an Old Technique," IEEE Spectrum, Vol. 35, pp. 115-130, November, 1964.

3. Anderson, B. D. O., Hitz, K. L., and Diem, N. D., "Recursive Algorithm for Spectral Factorization," IEEE Trans. Circuits Syst., Vol. CAS-21, pp. 742-750, November, 1974.

4. Anderson, T. W., Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York, 1958.

5. Anderson, V. C., "Digital Array Phasing," J. Acoust. Soc. Am., Vol. 32, pp. 867-870, July, 1960.

6. Anderson, V. C., "DICANNE, A Realizable Adaptive Process," J. Acoust. Soc. Am., Vol. 45, pp. 393-405, March, 1969.

7. Anderson, V. C., and Rudnick, P., "Rejection of a Coherent Arrival at an Array," J. Acoust. Soc. Am., Vol. 45, pp. 406-410, April, 1969.

8. Aoki, M., Optimization of Stochastic Systems, Academic Press, New York, 1967.

9. Apostol, T. M., Mathematical Analysis, Addison-Wesley Publishing Company, Reading, Mass., 1957.

10. Applebaum, S. P., "Adaptive Arrays," Syracuse University Research Corp., Rept. SPL-769, June, 1964.

11. Baggeroer, A. B., State Variables and Communication Theory, MIT Press, Cambridge, Massachusetts, 1970.

12. Baird, C. A., "Recursive Processing for Adaptive Arrays," Proc. Adaptive Antenna Systems Workshop, Naval Research Laboratory, Washington, D. C., pp. 163-183, March 11-13, 1974.

13. Beckenbach, E. F., and Bellman, R., Inequalities, Springer-Verlag, New York, 1965.

14. Bellman, R., Introduction to Matrix Analysis, McGraw Hill Book Company, New York, 1960.

15. Birkhoff, G., and MacLane, S., Modern Algebra, MacMillan, New York, 1965.

16. Blum, J. R., "Multidimensional Stochastic Approximation Methods," Ann. Math. Stat., Vol. 25, pp. 737-744, 1954.

17. Bodewig, E., Matrix Calculus, Interscience, New York, 1959.

18. Bracewell, R., The Fourier Transform and Its Applications, McGraw Hill Book Company, New York, 1965.

19. Brennan, L. E., and Reed, I. S., "Theory of Adaptive Radar," IEEE Trans. Aerosp. Electron. Syst., Vol. AES-9, pp. 237-252, March, 1973.

20. Bryn, F., "Optimum Signal Processing of Three-Dimensional Arrays Operating on Gaussian Signals and Noise," J. Acoust. Soc. Am., Vol. 34, pp. 289-297, March, 1962.

21. Bucy, R. S., "Stability and Positive Supermartingale," J. Diff. Equations, Vol. 1, pp. 151-155, April, 1965.

22. Burg, J. P., "Three-Dimensional Filtering with an Array of Seismometers," Geophysics, Vol. 29, pp. 693-713, October, 1964.

23. Capon, J., Greenfield, R. J., and Kolker, R. K., "Multidimensional Maximum-Likelihood Processing of a Large Aperture Seismic Array," Proc. IEEE, Vol. 55, pp. 192-211, February, 1967.

24. Chien, Y. T., and Fu, K. S., "Stochastic Learning of Time Varying Parameters in Random Environment," IEEE Trans. Syst. Sci. Cybern., Vol. SSC-5, pp. 237-246, September, 1969.

25. Claerbout, J. F., "A Summary, by Illustrations, of Least-Squares Filters with Constraints," IEEE Trans. Inf. Theor., Vol. IT-14, pp. 269-272, March, 1968.

26. Cochran, W. G., and Davis, M., "The Robbins-Monroe Method for Estimating the Median Lethal Dose," J. R. Stat. Soc., Series B, Vol. 27, pp. 28-44, 1965.

27. Comer, T. R., "Some Stochastic Approximation Procedures for Use in Process Control," Math. Stat., Vol. 35, pp. 86-98, 1964.

28. Cox, H., "Optimum Arrays and the Schwarz Inequality," J. Acoust. Soc. Am., Vol. 45, pp. 228-232, March, 1969.

29. Cramer, H., and Leadbetter, M. R., Stationary and Related Stochastic Processes, John Wiley & Sons, New York, 1967.

30. Daniell, T. P., "Adaptive Estimation with Mutually Correlated Training Sequences," IEEE Trans. Syst. Sci. Cybern., Vol. SSC-6, pp. 12-19, January, 1970.

31. Darlington, S., "Linear Least Squares Smoothing and Prediction, with Applications," Bell Syst. Tech. J., Vol. 37, pp. 1221-1294, December, 1958.

32. Davenport, W. B., and Root, W. L., An Introduction to the Theory of Random Signals and Noise, McGraw Hill, New York, 1958.

33. Davisson, L. D., "A Theory of Adaptive Filtering," IEEE Trans. Inf. Theor., Vol. IT-12, pp. 97-102, April, 1966.

34. Davisson, L. D., "Steady State Error in Adaptive Mean Square Minimization," IEEE Trans. Inf. Theor., Vol. IT-16, pp. 382-385, July, 1970.

35. Davisson, L. D., "Convergence Probability Bounds for Stochastic Approximation," IEEE Trans. Inf. Theor., Vol. IT-16, pp. 680-685, November, 1970.

36. Derman, C., and Sacks, J., "On Dvoretzky's Stochastic Approximation Theorem," Ann. Math. Stat., Vol. 30, pp. 601-606, 1959.

37. DeRusso, P. M., Roy, R. J., and Close, C. M., State Variables for Engineers, John Wiley & Sons, Inc., New York, 1965.

38. Doob, J. L., Stochastic Processes, John Wiley, New York, 1953.

39. Dunford, N., and Schwartz, J. T., Linear Operators, Interscience Publishers Inc., New York, New York, 1958.

40. Dupac, V., "A Dynamic Stochastic Approximation Method," Ann. Math. Stat., Vol. 36, pp. 1695-1702, 1965.

41. Dvoretzky, A., "On Stochastic Approximation," Proc. Third Berkeley Symp. Math. Stat. Probab., Vol. I, pp. 39-55, 1956.

42. Edelblute, D. J., Fisk, J. M., and Kinnison, G. L., "Criteria for Optimum-Signal-Detection Theory for Arrays," J. Acoust. Soc. Am., Vol. 41, pp. 199-205, January, 1967.

43. Fabian, V., "On Choice of Design in Stochastic Approximation Methods," Ann. Math. Stat., Vol. 39, pp. 457-465, 1968.

44. Falb, P. L., and Wolovich, W. A., "Decoupling in the Design and Synthesis of Multivariable Control Systems," IEEE Trans. Autom. Control, Vol. AC-12, pp. 651-659, December, 1967.

45. Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley & Sons, New York, 1968.

46. Friedman, S., "On Stochastic Approximation," Ann. Math. Stat., Vol. 34, pp. 343-346, 1963.

47. Frost, O. L., "An Algorithm for Linearly Constrained Adaptive Array Processing," Proc. IEEE, Vol. 60, pp. 926-935, August, 1972.

48. Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

49. Gladyshev, E. G., "On Stochastic Approximation," Theory of Probability and Its Applications, Vol. 10, pp. 275-278, March, 1965.

50. Goode, B. B., "Adaptive Sensor Array Processing," Ph.D. Dissertation, Stanford University, Stanford, Calif., November, 1970.

51. Griffiths, L., "A Comparison of Multidimensional Wiener and Maximum-Likelihood Filters," Proc. IEEE, Vol. 55, pp. 2045-2049, November, 1967.

52. Griffiths, L. J., "A Simple Adaptive Algorithm for Real-Time Processing in Antenna Arrays," Proc. IEEE, Vol. 57, pp. 1969-1704, October, 1969.

53. Halmos, P. R., Finite Dimensional Vector Spaces, Van Nostrand, Princeton, 1958.

54. Hamming, R., Numerical Methods for Scientists and Engineers, McGraw Hill, New York, 1962.

55. Hauser, A. A., Complex Variables, Simon and Schuster, New York, 1971.

56. Helstrom, C. W., Statistical Theory of Signal Detection, Pergamon Press, Oxford, 1960.

57. Hogg, R. V., and Craig, A. T., Introduction to Mathematical Statistics, MacMillan Company, New York, 1965.

58. Hohn, F. E., Elementary Matrix Algebra, MacMillan Company, New York, 1964.

59. Holmes, J. K., "Two Stochastic Approximation Procedures for Identifying Linear Systems," IEEE Trans. Autom. Control, Vol. AC-14, pp. 292-295, May, 1969.

60. Householder, A. S., The Theory of Matrices in Numerical Analysis, Blaisdell Publishing Co., New York, 1968.

61. Howells, P. W., "Intermediate Frequency Side Lobe Canceller," U.S. Patent 3,202,990, Filed May 4, 1959.

62. Jenkins, G. M., and Watt, D. G., Spectral Analysis and Its Applications, Holden-Day, San Francisco, 1968.

63. Kadota, T. T., "Optimum Estimation of Nonstationary Gaussian Signals in Noise," IEEE Trans. Inf. Theory, Vol. IT-15, pp. 201-221, March, 1969.

64. Kalman, R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," Trans. ASME, Vol. 83, Ser. D, pp. 95-108, March, 1961.

65. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, Vol. 82, Ser. D, pp. 35-43, March, 1960.

66. Karlin, S., and Taylor, H. M., A First Course in Stochastic Processes, Academic Press, New York, 1975.

67. Kesten, H., "Accelerated Stochastic Approximation," Ann. Math. Stat., Vol. 29, pp. 41-59, 1958.

68. Kiefer, J., and Wolfowitz, J., "Stochastic Estimation of the Maximum of a Regression Function," Ann. Math. Stat., Vol. 23, pp. 462-466, 1952.

69. Knopp, K., Theory and Application of Infinite Series, Hafner Publishing Co., New York, 1947.

70. Kobayashi, H., "Iterative Synthesis Methods for a Seismic Array Processor," IEEE Trans. Geosci. Electron., Vol. GE-8, pp. 169-178, July, 1970.

71. Kolmogorov, A. N., and Fomin, S. V., Introductory Real Analysis, Prentice Hall, Englewood Cliffs, New Jersey, 1970.

72. Kraus, J. D., Antennas, McGraw Hill Book Company, New York, 1950.

73. Kushner, H. J., "New Theorems: Examples in the Liapunov Theory of Stochastic Stability," Proc. Joint Auto Control Conf., pp. 613-619, Troy, New York, 1965.

74. Kushner, H. J., "A Note on the Maximum Sample Excursions of Stochastic Approximation Processes," Ann. Math. Stat., Vol. 37, pp. 513-516, 1966.

75. Kushner, H. J., "Stochastic Approximation Algorithms for the Local Optimization of Functions with Nonunique Stationary Points," IEEE Trans. Autom. Control, Vol. AC-17, pp. 646-654, October, 1972.

76. Kushner, H. J., and Gavin, T., "Extensions of Kesten's Adaptive Stochastic Approximation Method," Ann. Stat., Vol. 1, pp. 851-862, 1973.

77. Kushner, H. J., and Gavin, T., "Stochastic Approximation Type Methods for Constrained Systems," IEEE Trans. Autom. Control, Vol. AC-19, pp. 349-357, August, 1974.

78. Lacoss, R. T., "Adaptive Combining of Wideband Array Data for Optimal Reception," IEEE Trans. Goesci. Electron., Vol. GE-6, pp. 78-86, May, 1968.

79. Lainiotis, D. G., and Sims, F. L., "Sensitivity Analysis of Discrete Kalman Filters," Int. J. Control, Vol. 12, pp. 657-669, June, 1970.

80. Lee, Y. W., Statistical Theory of Communication, John Wiley & Sons, Inc., New York, 1960.

81. Luenberger, D. G., Optimization by Vector Space Methods, New York: John Wiley & Sons, Inc., 1969.

82. Martin, R. S., and Wilkinson, J. H., "The Implicit QL Algorithm," Numeriche Mathematik, Vol. 12, pp. 377-383, 1968.

83. McGarty, T. P., Stochastic Systems and State Estimation, John Wiley and Sons, New York, 1974.

84. McGee, W. F., "Complex Gaussian Noise Moments," IEEE Trans. Inf. Theor., Vol. IT-17, pp. 149-157, March, 1971.

85. McShane, E. J., Stochastic Calculus and Stochastic Models, Academic Press, New York, 1974.

86. Mendel, J. M., and Fu, K. S., Adaptive Learning and Pattern Recognition, Academic Press, New York, 1970.

87. Mermoz, H., "Optimal and Adaptive Receiving Antennas - A Review [Essai de Synthese sur les Antennes de Detection Optimales et Adaptives]," a translation from the French by J. M. Taylor, Jr., NRL Translation 1241, January 3, 1972, Original Source: Ann. Telecommun., Vol. 24, pp. 269-280, 1970.

88. Middleton, D., and Groginsky, H. L., "Detection of Random Acoustic Signals by Receivers with Distributed Elements: Optimum Receiver Structures for Normal Signal and Noise Fields," J. Acoust. Soc. Am., Vol. 38, pp. 727-737, November, 1965.

89. Muellar, K. H., "A New Fast Converging Mean Square Algorithm for Adaptive Equalizers with Partial Response Signaling," Bell Syst. Tech. J., Vol. 54, pp. 143-153, January, 1975.

90. Mueller, K. H., and Spaulding, K. H., "Cyclic Equalization - A New Rapidly Converging Equalization Technique for Synchronous Data Communication," Bell Syst. Tech. J., Vol. 54, pp. 369-406, February, 1975.

91. Nagy, B., and Riesz, F., Functional Analysis, Ungar Company, New York, 1956.

92. Oza, K. G., and Jury, E. I., "System Identification and the Principle of Random Contraction Mapping," Siam J. Control, Vol. 6, pp. 244-257, February, 1968.

93. Papoulis, A., Probability, Random Variables and Stochastic Processes, McGraw Hill, New York, 1965.

94. Parzen, E., Stochastic Processes, Holden Day, San Francisco, 1962.

95. Pasupathy, S., and Venetsanopoulos, A. N., "Optimum Active Array Processing Structure and Space-Time Factorability," IEEE Trans. Aerosp. Elect. Syst., Vol. AES-10, pp. 770-779, November, 1974.

96. Pierre, D. A., Optimization Theory with Applications, John Wiley & Sons, Inc., New York, 1969.

97. Pratt, W. K., Generalized Wiener Filtering Computation Techniques, IEEE Trans. Comput., Vol. C-21, pp. 636-641, July, 1972.

98. Rabiner, L. R., and Gold, B., Theory and Application of Digital Signal Processing, Prentice Hall, Englewood Cliffs, New Jersey, 1975.

99. Raiffa, H., and Schlaifer, R., Applied Statistical Decision Theory, MIT Press, Cambridge, Massachusetts, 1968.

100. Rhodes, I. B., "A Tutorial Introduction to Estimation and Filtering," IEEE Trans. Autom. Control, Vol. AC-16, pp. 688-706, December, 1971.

101. Riegler, R., and Compton, R., "An Adaptive Array for Interference Rejection," Proc. IEEE, Vol. 61, pp. 748-758, June, 1973.

102. Robbins, H., and Monroe, S., "A Stochastic Approximation Method," Ann. Math. Stat., Vol. 22, pp. 400-407, 1951.

103. Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming: Part I: Linear Constraints," J. Soc. Ind. Appl. Math., pp. 181-217, 1960.

104. Royden, H. L., Real Analysis, MacMillan Company, New York, 1968.

105. Rudin, W., Real and Complex Analysis, McGraw Hill, New York, 1966.

106. Rudnick, P., "Small Signal Detection in the DIMUS Array," J. Acoust. Soc. Am., Vol. 32, pp. 871-877, July, 1960.

107. Sage, A. P., Optimum Systems Control, Prentice Hall, Inc., Englewood Cliff, New Jersey, 1968.

108. Sakrison, D. J., "Iterative Design of Optimum Filters for Non-mean Square Error Criteria," IEEE Trans. Inf. Theor., Vol. IT-3, pp. 161-167, February, 1963.

109. Sakrison, D. J., "Efficient Recursive Estimation of the Parameters of Radio Astronomy Target," IEEE Trans. Inf. Theor., Vol. IT-12, pp. 35-41, January, 1966.

110. Saridis, G. N., Nikolic, Z. Z., and Fu, K. S., "Stochastic Approximation Algorithms for System Identification, Estimation, and Decomposition of Mixtures," IEEE Trans. Syst., Sci., Cybern., Vol. SSC-5, pp. 8-15, January, 1969.

111. Sato, S. M., and Lopresti, P. V., "On the Generalization of State Feedback Decoupling Theory," IEEE Trans. Autom. Control, Vol. AC-16, pp. 133-140, April, 1971.

112. Schultheiss, P. M., "Passive Sonar Detection in the Presence of Interference," J. Acoust. Soc. Am., Vol. 43, pp. 418-425, March, 1968.

113. Schwartz, M., and Winkler, L. P., "Adaptive Nonlinear Optimization of the Signal-to-Noise Ratio of an Array Subject to a Constraint," J. Acoust. Soc. Am., Vol. 52, pp. 39-51, January, 1972.

114. Schweppe, F. C., "Sensor-Array Data Processing for Multiple-Signal Sources," IEEE Trans. Inf. Theor., Vol. IT-14, pp. 294-305, March, 1968.

115. Singer, R. A., and Frost, P. A., "On the Relative Performance of the Kalman and Wiener Filters," IEEE Trans. Autom. Control, Vol. AC-14, pp. 390-394, August, 1969.

116. Snyder, D. L., The State-Variable Approach to Continuous Estimation, MIT Press, Cambridge, Massachusetts, 1969.

117. Sondhi, M. M., "An Adaptive Echo Cancellor," Bell, Syst. Tech. J., Vol. 46, pp. 497-520, March, 1967.

118. Stakgold, I., Boundary Value Problem in Mathematical Physics, MacMillan Company, New York, 1968.

119. Stark, L., "Microwave Theory of Phased Array Antennas - A Review," Proc. IEEE, Vol. 62, pp. 1661-1701, December, 1974.

120. Stoll, R. R., Linear Algebra and Matrix Theory, McGraw Hill, New York, 1952.

121. Thrall, R. M., and Tornheim, L., Vector Spaces and Matrices, John Wiley & Sons, Inc., New York, 1957.

122. Tsypkin, Y., Adaptation and Learning in Automatic Systems, Academic Press, New York, 1971.

123. Tuteur, F. B., and Chang, J. H., "A New Class of Adaptive Array Processors," J. Acoust. Soc. Am., Vol. 49, pp. 639-649, March, 1971.

124. Urkowitz, H., "On Detection and Estimation of Wave Fields for Surveillance," IEEE Trans. Mil. Electron., Vol. ME-12, pp. 44-56, January, 1970.

125. Van Trees, H. L., Detection, Estimation and Linear Modulation Theory, John Wiley & Sons, New York, 1968.

126. Varga, R., Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, New Jersey, 1956.

127. Vaysboro, E. M., and Yudin, D. B., "Multiextremal Stochastic Approximation," Eng. Cybern., Vol. 6, pp. 1-11, January, 1968.

128. Venter, J. H., "On Dvoretzky Stochastic Approximation Theorems," Ann. Math. Stat., Vol. 37, pp. 1534-1544, 1966.

129. Venter, J. H., "An Extension of the Robbins-Monro Procedure," Ann. Math. Stat., Vol. 38, pp. 181-190, 1967.

130. Venter, J. H., "On Convergence of the Keifer-Wolfowitz Approximation Procedures," Ann. Math. Stat., Vol. 38, pp. 1031-1036, 1967.

131. Wasan, M. T., Stochastic Approximation, Cambridge University Press, Cambridge, 1969.

132. Whalen, A. D., Detection of Signals in Noise, Academic Press, New York, 1971.

133. Widrow, B., Mantey, P. E., Griffiths, L. J., and Goode, B. B., "Adaptive Antenna Systems," Proc. IEEE, Vol. 55, December, 1967.

134. Wiener, N., Extrapolation Interpolation and Smoothing of Stationary Time Series, MIT Press, Cambridge, Massachusetts, and John Wiley & Sons, Inc., New York, 1949.

135. Wilkinson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

136. Wolfowitz, J., "On the Stochastic Approximation Method of Robbins and Monro," Ann. Math. Stat., Vol. 23, pp. 457-461, 1952.

137. Wolfowitz, J., "On Stochastic Approximation Methods," Ann. Math. Stat., Vol. 27, pp. 1151-1156, 1956.

138. Zadeh, L. A., and Desoer, C. A., Linear System Theory, McGraw Hill Book Company, New York, 1963.

139. Zahm, C. L., "Application of Adaptive Arrays to Suppress Strong Jammers in the Presence of Weak Signals," IEEE Trans. Aerosp. Electron. Syst., Vol. AES-9, pp. 260-270, March, 1973.

APPENDIX A

S/N CALCULATION

Since the output of the filter has been defined as

$$\underline{y} = H^T \underline{x} \quad , \tag{A.1}$$

one can, because of the way the beams are constructed, calculate an estimate of signal-to-noise ratio at the output of the adaptive filter. Knowing the input signal statistics (zero mean, variable variance normal processes), one can calculate an approximate output signal-to-noise ratio.

If we assume the mean of all s_n are zero and that $E\{s_n s_n^*\} = \sigma_n^2$, then we can calculate the products

$$y^n (y^n)^* = [(\underline{h}^n)^T \underline{x}^n][(\underline{h}^n)^T x^n]^* \tag{A.2}$$

and collect the terms due to the signals alone and noise alone, and the ratio of these terms is the output signal-to-noise ratio. The beams (x^n) are constructed in the following manner:

$$x^1 = (s1)s_1 + (a)s_2 + (b)s_3 + (cc)sn_1 \tag{A.3}$$

$$x^2 = (f)s_1 + (s2)s_2 + (g)s_3 + (dd)sn_2 \tag{A.4}$$

$$x^3 = (h)s_1 + (p)s_2 + (s3)s_3 + (ee)sn_3 \tag{A.5}$$

Using Equations (A.3), (A.4) and (A.5), the products indicated in (A.2) can be calculated and the separation of signal terms from noise terms in the resulting expressions permits calculation of an output signal-to-noise ratio.

$$y^1 \cdot (y^1)^* = h^{11}(h^{11})^*[(s1 \cdot s1^*)s_1s_1^* + (a \cdot a^*)s_2s_2^* + (b \cdot b^*)s_3s_3^* + (cc \cdot cc^*)sn_1sn_1^*] + h^{11}(h^{12})^*[(s1 \cdot f^*)s_1s_1^* + (a \cdot s2^*)s_2s_2^* + (b \cdot g^*)s_3s_3^*] + h^{11}(h^{13})^*[(s1 \cdot h^*)s_1s_1^* + (a \cdot p^*)s_2s_2^* + (b \cdot s3^*)s_3s_3^*] + h^{12}(h^{11})^*[(f \cdot s1^*)s_1s_1^* + (s2 + a^*)s_2s_2^* + (g \cdot b^*)s_3s_3^*] + h^{12}(h^{12})^*[(f \cdot f^*)s_1s_1^* + (s2 \cdot s2^*)s_2s_2^* + (g \cdot g^*)s_3s_3^* + (dd \cdot dd^*)sn_2sn_2^*] + h^{12}(h^{13})^*[(f \cdot h^*)s_1s_1^* + (s2 \cdot p^*)s_2s_2^* + (g \cdot s3^*)s_3s_3^*] + h^{13}(h^{11})^*[(h \cdot s1^*)s_1s_1^* + (p \cdot a^*)s_2s_2^* + (s3 \cdot b^*)s_3s_3^*] + h^{13}(h^{12})^*[(h \cdot f^*)s_1s_1^* + (p \cdot s2^*)s_2s_2^* + (s3 \cdot g^*)s_3s_3^*] + h^{13}(h^{13})^*[(h \cdot h^*)s_1s_1^* + (p \cdot p^*)s_2s_2^* + (s3 \cdot s3^*)s_3s_3^* + (ee \cdot ee^*)sn_3sn_3^*] \quad (A.6)$$

$$y^2 \cdot (y^2)^* = h^{21}(h^{21})^*[(s1 \cdot s1^*)s_1s_1^* + (a \cdot a^*)s_2s_2^* + (b \cdot b^*)s_3s_3^* + (cc \cdot cc^*)sn_1sn_1^*] + h^{21}(h^{22})^*[(s1 \cdot f^*)s_1s_1^* + (a \cdot s2^*)s_2s_2^* + (b \cdot g^*)s_3s_3^*] + h^{21}(h^{23})^*[(s1 \cdot h^*)s_1s_1^* +$$

$$(a \cdot p^*)s_2s_2^* + (b \cdot s3^*)s_3s_3^*] + h^{22}(h^{21})^*[(f \cdot s1^*)s_1s_1^* +$$

$$(s2 \cdot a^*)s_2s_2^* + (g \cdot b^*)s_3s_3^*] + h^{22}(h^{22})^*[(f \cdot f^*)s_1s_1^* +$$

$$(s2 \cdot s2^*)s_2s_2^* + (g \cdot g^*)s_3s_3^* + (dd \cdot dd^*)sn_2sn_2^*] +$$

$$h^{22}(h^{23})^*[(f \cdot h^*)s_1s_1^* + (s2 \cdot p^*)s_2s_2^* + (g \cdot s3^*)s_3s_3^*] +$$

$$h^{23}(h^{21})^*[(h \cdot s1^*)s_1s_1^* + (p \cdot a^*)s_2s_2^* + (s3 \cdot b^*)s_3s_3^*] +$$

$$h^{23}(h^{22})^*[(h \cdot f^*)s_1s_1^* + (p \cdot s2^*)s_2s_2^* + (s3 \cdot g^*)s_3s_3^*] +$$

$$h^{23}(h^{23})^*[(h \cdot h^*)s_1s_1^* + (p \cdot p^*)s_2s_2^* + (s3 \cdot s3^*)s_3s_3^* +$$

$$(ee \cdot ee^*)sn_3sn_3^*] \qquad (A.7)$$

$$y^3 \cdot (y^3)^* = h^{31}(h^{31})^*[(s1 \cdot s1^*)s_1s_1^* + (a \cdot a^*)s_2s_2^* + (b \cdot b^*)s_3s_3^* +$$

$$(cc \cdot cc^*)sn_1sn_1^*] + h^{31}(h^{32})^*[(s1 \cdot F^*)s_1s_1^* +$$

$$(a \cdot s2^*)s_2s_2^* + (b \cdot g^*)s_3s_3^*] + h^{31}(h^{33})^*[(s1 \cdot h^*)s_1s_1^* +$$

$$(a \cdot p^*)s_2s_2^* + (b \cdot s3^*)s_3s_3^*] + h^{32}(h^{31})^*[(f \cdot s1^*)s_1s_1^* +$$

$$(s2 + a^*)s_2s_2^* + (g \cdot b^*)s_3s_3^*] + h^{32}(h^{32})^*[(f \cdot f^*)s_1s_1^* +$$

$$(s2 \cdot s2^*)s_2s_2^* + (g \cdot g^*)s_3s_3^* + (dd \cdot dd^*)sn_2sn_2^*] +$$

$$h^{33}(h^{31})^*[(h \cdot s1^*)s_1s_1^* + (p \cdot a^*)s_2s_2^* + (s3 \cdot b^*)s_3s_3^*] +$$

$$h^{33}(h^{32})^*[(h \cdot f^*)s_1s_1^* + (p \cdot s2^*)s_2s_2^* + (s3 \cdot g^*)s_3s_3^*] +$$

$$h^{33}(h^{33})^*[(h \cdot h^*)s_1s_1^* + (p \cdot p^*)s_2s_2^* + (s3 \cdot s3^*)s_3s_3^* +$$

$$(ee \cdot ee^{*})sn_3 sn_3^{*}] + h^{32}(h^{33})^{*}[(f \cdot h^{*})s_1 s_1^{*} +$$

$$(s2 \cdot p^{*})s_2 s_2^{*} + (g \cdot s3^{*})s_3 s_3^{*}] \quad . \tag{A.8}$$

If all the signal terms are separated from the noise terms in any given beam output, then the output signal-to-noise ratio can be calculated as the ratio of these terms. In simple terms, the output signal-to-noise ratio can be defined as:

$$S/N = SNR \triangleq \frac{H^{T}(f)\, G_{ss}(f)\, H(f)^{*}}{H^{T}(f)\, G_{nn}(f)\, H(f)^{*}} \quad . \tag{A.9}$$

APPENDIX B

LINEAR SYSTEM RELATIONSHIPS

The single output of a multi-input, multi-output system is written as

$$y^i(f) = [\underline{h}^i(f)]^T \underline{x}(f) \tag{B.1}$$

and in matrix form,

$$\underline{y}(f) = H^T(f)\, \underline{x}(f) \ , \tag{B.2}$$

where

$$\underline{y}(f) = \begin{bmatrix} y^1(f) \\ y^2(f) \\ \cdot \\ \cdot \\ \cdot \\ y^N(f) \end{bmatrix} , \qquad \underline{x}(f) = \begin{bmatrix} x^1(f) \\ x^2(f) \\ \cdot \\ \cdot \\ \cdot \\ x^N(f) \end{bmatrix}$$

and

$$H(f) = [\underline{h}^1(f)\ \underline{h}^2(f) \ . \ . \ . \ \underline{h}^N(f)] \ ,$$

$$H(f) = \begin{bmatrix} h^{11}(f) & h^{12}(f) \ldots h^{1N}(f) \\ h^{21}(f) & \ldots\ldots\ldots \\ \vdots & \vdots \\ h^{N1}(f) \ldots\ldots\ldots\ldots & h^{NN}(f) \end{bmatrix} .$$

The output power spectral density matrix is given by:

$$G_{\underline{yy}}(f) = H^T(f)G_{\underline{xx}}(f)H^*(f) = [H^T(f)\underline{x}(f)][H^T(f)\underline{x}(f)]^H \quad . \quad (B.3)$$

Some useful relations for the auto-spectrum and cross spectrum matrices are:

$$G_{\underline{xx}}{}^{-1}(f) = [G_{\underline{xx}}{}^{-1}(f)]^H \quad (B.4)$$

$$G_{\underline{xy}}{}^H(f) = G_{\underline{yx}}(f) \quad (B.5)$$

$$G_{\underline{xx}}{}^H(f) = G_{\underline{xx}}(f) \quad \text{(Hermitian property)} \quad . \quad (B.6)$$

APPENDIX C

PROBABILITY THEORY DEFINITIONS

(C.1) Conditional Expectations

The conditional distribution function $F_{x|y}(x|y)$ of x given $Y = y$ is defined by

$$F_{x|y}(x|y) = \frac{\text{PROB}\{X \leq x,\ Y = y\}}{\text{PROB}\{Y = y\}} \quad \text{for} \quad \text{PROB}\{Y = y\} > 0 \quad . \qquad (C.1)$$

For any values x and y, we can define (Karlen and Taylor, 66)

$$\text{PROB}\{X \leq x,\ Y \leq y\} = \int_{-\infty}^{z} F_{x|y}(x|z)\ dF_y(z) \quad . \qquad (C.2)$$

To define the law of total probability, we make $y = \infty$ in Equation (C.2) and get

$$\text{PROB}\{X \leq x\} = \text{PROB}\{X \leq x,\ Y \leq \infty\} = \int_{-\infty}^{+\infty} F_{x|y}(x|y)\ dF_y(y) \qquad (C.3)$$

For any bounded functions $g(x)$ and $h(y)$, Equation (C.3) becomes

$$E\{g(x)\ h(y)\} = E\{E[g(x)|y]\ h(y)\} \quad . \qquad (C.4)$$

The following equations list some conditional expectation identities.

$$E\{h(x,y) \mid Y = y\} = E\{h(x,y) \mid Y = y\} \tag{C.5}$$

$$E\{g(x)\, f(y) \mid y\} = f(y)\, E\{g(x) \mid y\} \tag{C.6}$$

$$E\{c|y\} = c \quad \text{for } c \text{ a constant} \tag{C.7}$$

$$|E\{\underline{x}^T \underline{y}\}|^2 \leq E\{||\underline{x}||^2\}\, E\{||y||^2\} \ . \tag{C.8}$$

(C.2) Martingales

Definition C.1. A martingale is defined as a real or complex stochastic process $(x_n, n\epsilon N)$ for which $E\{|x_n|\} < \infty$, $n\epsilon N$ and

$$E\{x_{n+1} \mid x_n \ldots x_1\} = x_n \tag{C.9}$$

with probability 1. If the equal sign in Equation (C.9) is replaced by $\leq$, then the process is a supermartingale.

If an x_n process is a martingale, the process defined by the real and imaginary parts of the x_n's are also martingales.

Definition C.2. Let $\{x_n, y_n, n = 1, 2, \ldots\}$ by a martingale on a probability space (Ω, β, P) with $\sup_n E\{|x_n|\} < \infty$. The following conditions are all equivalent:

$$\{x_n, y_n, n = 1, 2, \ldots, \infty\} \text{ is a martingale .} \tag{C.10}$$

$$\lim_{n\to\infty} E\{|x_n - x_\infty|\} = 0 \ . \tag{C.11}$$

$$E\{|x_\infty|\} = \lim_{n\to\infty} E\{|x_n|\} \ . \tag{C.12}$$

DISTRIBUTION

Commander (NSEA 09G32)
Naval Sea Systems Command
Department of the Navy
Washington, D. C. 20362 — Copies 1 and 2

Commander (NSEA 0342)
Naval Sea Systems Command
Department of the Navy
Washington, D. C. 20362 — Copies 3 and 4

Defense Documentation Center
5010 Duke Street
Cameron Station
Alexandria, VA 22314 — Copies 5 through 16