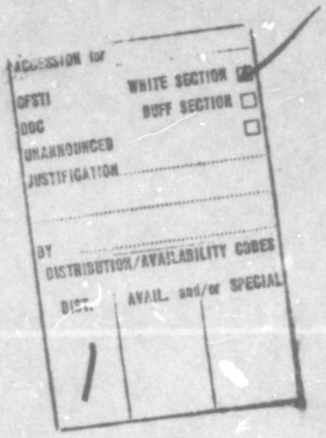
RADC-TR-69-430
Technical Report
February 1970

UNSUPERVISED ESTIMATION AND PROCESSING OF UNKNOWN SIGNALS

Purdue University

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded, by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.



Do not return this copy. Retain or destroy.

UNSUPERVISED ESTIMATION AND PROCESSING OF UNKNOWN SIGNALS

E. A. Patrick
J. P. Costello
Purdue University

FOREWORD

This interim technical report was prepared by Purdue University, School of Electrical Engineering under contract F30602-68-C-0186 sponsored by Rome Air Development Center, Griffiss Air Force Base, New York, 13440 and contract F33615-68-C-1577 sponsored by Air Force Avionics Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio; RADC project No. 5581, task 558104. Purdue's assigned report number was TR-EE-69-18. John C. Faust (EMBIS) was the RADC project engineer.

This technical report has been reviewed by the Office of Information (EMLS) and is releasable to the Clearinghouse for Federal Scientific and Technical Information.

This technical report has been reviewed and is approved.

John C. Faust

Approved: JOHN C. FAUST
Project Engineer

Frank J. Morreale

Approved: CARLO P. CROCETTI
Chief, Info Processing Branch

ABSTRACT

Many communications systems, sonar and radar systems, control systems, and pattern recognition systems such as biomedical signal processing systems must partition a multidimensional sample space so that decisions can be made on the underlying active source events. Unfortunately, the a priori information necessary to construct an acceptable partition is not always available. For many problems estimation using samples of unknown classification is the only source of additional knowledge on the sample space statistical structure. Since the sample classifications are unknown, these estimators are called unsupervised estimation algorithms.

This research is concerned with investigating practical approaches to the unsupervised estimation problem which are in some sense optimum. The emphasis is on recursive estimation algorithms having fixed storage requirements and on sequential sample processing. A Bayesian framework is utilized as a guide towards "optimality", and to provide a unifying relationship for the approaches of the report. The relationship between Bayes a posteriori, stochastic approximation, and decision directed approaches is determined. It is shown, for example, that an optimization criterion derived from the Bayes approach can be used to relate maximum

likelihood-related stochastic approximation algorithms with decision directed estimators. The application of unsupervised estimation algorithms to a practical problem is illustrated using the problem of intersymbol interference.

A direct implementation of an optimum a posteriori approach is to approximate the parameter space with a finite set of vector points. The Bayes estimator on such a discretized parameter space is proven to converge to an asymptotic vector with probability one and in mean square. The asymptotic estimator and asymptotic rate of convergence are also found. These results on a discretized parameter space lead to a new continuous parameter space criterion. The maximum of this criterion minimizes average risk against the a priori assumption of a particular parametric density function family. The properties of the criterion surface are largely unknown, but contours evaluated for a two class Gaussian problem show unimodality for this problem. For the case where the criterion surface is unimodal, stochastic approximation algorithms which seek the maximum are defined. Also, a criterion form resulting from a separable Gaussian assumption allows the definition of a simple clustering technique for maximizing the criterion.

A class of decision directed algorithms are defined which minimize a criterion derived from the separable Gaussian criterion. This class of algorithms unifies several previous papers on decision directed estimators. The decision directed estimators are given the interpretation of stochastic approximation algorithms with random weights. This allows a comparison of properties found for the decision directed

BLANK PAGE

algorithm class with results in the literature on conventional stochastic approximation algorithms. Theoretical asymptotic probability of error curves and experimental dynamic convergence results are presented.

The problem of intersymbol interference occurs when a channel smears energy from one signal baud onto following ones. The mode or cluster structure of the multibaud sample space is discussed and used to relate the approaches of previous papers. Two decision directed estimators suitable for the interference problem are defined. Experimental and asymptotic performance curves are presented. Extensions and related problems are also discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF SYMBOLS (CONTINUED)

Pe_{CB}	The conditional average probability of error given the last decision on m_{n-2} was correct
Pe_{WB}	The conditional average probability of error given the last decision on m_{n-2} was incorrect
$d(\underline{X}_n \mid \underline{u}^k)$	The suboptimum decision equation
$2r$	The number of bauds skipped so that samples with intersymbol interference are independent
$\underline{u}_N^k$	The estimate of the k^{th} mode at stage N
$\underline{s}_N$	The estimate of the channel gain vector using one dimensional binary antipodal message signals at stage N
$\underline{u}_\infty^k$	The asymptotic estimate of the k^{th} mode
$\underline{s}_\infty$	The asymptotic estimate of the channel gain vector
$\delta(\cdot)$	The indicator function which is none zero and equal to one only if $(\cdot)=0$
$Q_n(m^k)$	A $(v \times c)_k$ matrix of zeros and ones corresponding to message m^k active
$\underline{s}_N^k$	The k^{th} message signal mapping of the channel gain vector for arbitrary message signals
A_k	The k^{th} $(\ell \times \ell)$ channel gain matrix
T	The time duration of one signal baud

LIST OF FIGURES (CONTINUED)

LIST OF FIGURES (CONTINUED)

LIST OF SYMBOLS

SYMBOL	DESCRIPTION
$\mathfrak{F}$	Family of distribution functions
$F(x\|\alpha)$	A distribution function indexed by α
ℓ	Dimensionality of the sample space
R^k	k dimensional Euclidean space
x	A point in the sample space
α	Functional family vector index
$\mathfrak{A}$	Set of vector indices indexing $\mathfrak{F}$
X	Cartesian product
ν	Lebesgue measure
$f(x\|\alpha)$	A density function indexed by α
c	A constant
c'	A constant
$\mathfrak{F}'$	Family of densities corresponding to $\mathfrak{F}$
L_p	Function space of Lesbeque p integrable functions
$H(x\|B)$	Parameter conditional mixture distribution
$h(x\|B)$	Parameter conditional mixture density
$P(\alpha)$	Mixing parameter
$\mathfrak{P}$	Assumed set of all possible mixing parameter values
M^k	The number of non zero mixing parameters in the k^{th} parameter conditional mixture

LIST OF SYMBOLS (CONTINUED)

M	The actual number of non zero mixing parameters in the true mixture
M'	Assumed upper bound on M^k
B	Mixture vector index
$\mathcal{B}^{M'}$	The set of unique mixture vector indices having $M^k \leq M'$
$\underline{\alpha}(\cdot)$	An operator on B
$\underline{\sigma}(\cdot)$	A permutation operator
$(G \times P)^{M'}$	Product parameter space
$p_o(B)$	A priori probability density on $(G \times P)^{M'}$
$p(B\|Y_n)$	Posterior density on $\mathcal{B}^{M'}$ given Y_n
Y_n	$X_1, X_2, \ldots, X_n$
n	An integer count
V	The number of points in $\mathcal{B}^{M'}$
$\hat{B}(Y_n)$	Bayes estimator given Y_n
$\eta(B)$	A measure of information or distance; also a regression function
$E[\cdot]$	The expected value of $[\cdot]$
i	Integer
j	Integer
k	Integer
s	Integer
$\|(\cdot)\|^k$	The k^{th} power of the absolute value of $(\cdot)$
$h(x)$	The true mixture density

LIST OF SYMBOLS (CONTINUED)

X_k	k^{th} sample from h(x)
B^*	An index in $\mathcal{B}^{M'}$ maximizing $\eta(B)$
$\{B^{i*}\}$	The set of all $B^* \in \mathcal{B}^{M'}$
V^*	The number of points in $\{B^{i*}\}$
$\bar{B}^*$	An average value of $\{B^{i*}\}$
$\|\cdot\|$	Norm
$\lim_{n \to \infty} [\cdot]$	The limit of $[\cdot]$ as n tends to infinity
$\min[\cdot,\cdot]$	The smaller value in $[\cdot,\cdot]$
$z^k(Y_n)$	A random variable used to bound $p(B^k \mid Y_n)$
$r^k(Y_n)$	A random variable used to bound $p(B^k \mid Y_n)$
(d^k)	$\eta(B^*) - \eta(B^k)$ a measure of distance
U	An arbitrary random variable
$(\cdot)^t$	The transpose of $(\cdot)$
D	Support of the true mixture density h(x)
ρ	A non-negative constant less than one
$(\sigma)^2$	A variance
ϵ	Small positive number
$\hat{B}_n$	The n^{th} hillclimbing algorithm estimate of the mixture index B maximizing $\eta(B)$
D_n	The n^{th} update vector of the $\eta(B)$ hillclimbing algorithm
b_n	Weighting sequence

LIST OF SYMBOLS (CONTINUED)

c_n	Weighting sequence
e^j	The j^{th} column of the appropriate identity matrix I
γ	Mean vector
Φ	Covariance matrix
Ψ	Lagrange multiplier
$L^i(x)$	The i^{th} likelihood function
α_k	A weighting sequence
$\eta(B;\hat{B})$	Parameter conditional regression surface
M*	The number of non zero mixing parameters in a parameter conditional mixture
$\{S^i\}_{i=1}^{M*}$	A partition of R^ℓ into M* disjoint regions
γ^{i_o}	The true mean vector of the i^{th} class
SNR	Signal to Noise Ratio
B_{max_n}	Maximum Likelihood estimator
$C(B^*)$	Fischer information matrix
$c^{ij}(B^*)$	The i^{th} row, j^{th} column element of the Fischer information matrix
ω^i	The i^{th} message source or class
$\underline{\gamma}$	$(\gamma^1,\gamma^2,\ldots,\gamma^{M'})$
$\underline{P}(\underline{\gamma})$	$(P(\gamma^1),P(\gamma^2),\ldots,P(\gamma^{M'}))$
A_N^i	The N^{th} update region of the i^{th} estimator
$\underline{A}_N$	$(A_N^1,\ldots,A_N^M)$

LIST OF SYMBOLS (CONTINUED)

K	The number of samples during which the $\underline{\gamma}_N$ are held fixed
ξ_k^i	The k^{th} conditional sample mean of the i^{th} estimator
β_k^i	An update weight
ρ_N^i	The N^{th} update weight of the i^{th} class
w_N^i	The N^{th} count on the i^{th} class
$W(\underline{\gamma})$	The average second moment about $\underline{\gamma}$
$V(\underline{\gamma})$	The average variance given $\underline{\gamma}$
μ_A^i	$E[x \mid x \in A]$
χ	A subset of R^{ℓ}
χ^c	The complement of χ
W_∞	Limit of $W(\underline{\gamma}_N)$ as $N \to \infty$
$\underline{S}_N$	$(S_N^1, S_N^2, \ldots, S_N^M)$
I_N^i	The characteristic function of an event
μ	Sample mean
$U(x)$	Denotes a hyperplane in the sample space
$\mathcal{I}(A)$	The interior of the set A
$\mathcal{Y}$	The convex closure of the support of $h(x)$
$\Gamma(\gamma^1, \gamma^2)$	A mapping of $(R^{\ell})^2$ onto itself
τ	Truncation
$\{a^k\}_{k=1}^{\ell}$	A set of basis vectors
ζ^i	$E[x \mid \omega^i]$

LIST OF SYMBOLS (CONTINUED)

λ	A proportionality constant
ϕ	A one dimension Gaussian distribution with mean zero variance one
ϕ'	The corresponding density
t^i	The mean vector on q^1
Pe	Probability of error
Pe_{min}	Minimum probability of error with all parameters known
ΔPe	$Pe - Pe_{min}$
m_n	The n^{th} transmitted message signal
u_n	The n^{th} linear channel output signal
a_k	The k^{th} channel gain
c	The largest index of a non zero channel gain
n_n	Additive white Gaussian noise
v	The number of bauds used in the multibaud sample space
v*	The message signal index before n + 1 for which a decision is made at time n
$\underline{X}_n$	A v dimensional sequence vector
$\underline{u}_n$	A v dimensional sequence vector
Q^k	A (v x c) matrix of possible messages
$f(\underline{X}^n \mid \underline{u}^k)$	A v dimensional Gaussian density with mean vector $\underline{u}^k$, covariance matrix $(\sigma)^2 I$
$P(\underline{u}^k)$	The mixing parameter associated with $f(\underline{X}_n \mid \underline{u}^k)$
$Pe[m_n]$	The average probability of error for a decision on m_n

I. INTRODUCTION

Many communications systems, sonar and radar systems, control systems, and pattern recognition systems such as biomedical signal processing systems must operate in environments which are unknown to the system designer. A classical design approach, which is still in widespread use, is to assume a particular environment and then maximize a measure of performance for operation in this environment. While the resulting system is optimum in the assumed environment, it may be almost useless in another. A more conservative approach is to establish a worst case environment and optimize performance for this worst case. However, this criterion may be too conservative for the actual operating conditions, and the performance of such a system may be considerably poorer than the maximum attainable performance. Classical design approaches such as these usually result in systems with parameters that are fixed rather than dependent on the actual statistics. The ability of such fixed parameter systems to perform well in a wide variety of environments is fairly limited.

A "learning" system which can determine current operating conditions offers a considerable increase in system flexibility and perhaps a significant improvement in performance over a fixed system. The learning system takes advantage of favorable environments while still maintaining an ability to perform as well as can be expected

under unfavorable conditions. Also, less conservative design criteria can be used because more accurate information on the actual mode of operation is available to the system. For example, if a receiver is to process signals transmitted through an unknown time varying channel, a design based on estimates of the current received signal set has obvious advantages over a system design that minimizes average risk over the entire ensemble of possible received signal sets. The use of learning systems to improve performance is still in its infancy, and major results on practical problems are very limited at this time.

The principal topic of interest here is the development of practical solutions to problems characterized by a finite number of sources[+] and by the requirement that a multidimensional sample space must be partitioned so that decisions can be made on the underlying active source events. If enough prior knowledge to construct an acceptable partition is not available, it is necessary to estimate (i.e., to "learn") the statistics corresponding to each set of active source events. The unknown statistics can be estimated using samples of known classification as <u>training samples</u> if such samples are available. This estimation approach is called <u>supervised estimation</u> and is particularly valuable where the statistics are stationary and and enough samples are available. However, for many problems it is

---

[+]The terms sources, classes, and patterns will be used almost interchangeably throughout the report.

either impractical or impossible to obtain enough fully classified[+] samples to define an acceptable partition of the sample space. For these problems the statistical structure of the sample space must be estimated using samples of unknown classification. Since the classifications of the samples are unknown, such a procedure is called unsupervised estimation.

This report is concerned with finding practical approaches to the unsupervised estimation problem which are in some sense optimum. The emphasis is on recursive estimation algorithms having fixed storage requirements and on sequential sample processing. The application of unsupervised estimation algorithms to a practical problem is illustrated using the problem of intersymbol interference.

In Section 1.1 the unsupervised estimation problem is examined further, and previous results are discussed. A literature survey for the intersymbol interference problem is presented in Section 1.2. Finally, the chapter concludes with a discussion of the approaches and contributions of this report.

1.1 The Unsupervised Estimation Problem

In unsupervised estimation problems the "structure" of the sample space is inadequately known and it is necessary to "learn" this structure using samples $X_1, X_2, \ldots, X_n$ whose classifications are unknown. If the density function of X_k can be expressed as a linear combination of density functions from a known family whose members

---

[+]A sample is called fully classified if all the active source events are known. For example, it is difficult to diagnose all of a patients physical aliments; on the other hand, in a particular communications system the totality of message events might be either a zero or a one.

are indexed by a parameter vector, this is called parametric structure. An often used parametric family is the family of Gaussian density functions which is indexed by a mean vector and a covariance matrix. Another type of structure of importance is mode or cluster structure in which groups of samples that are close in some sense are defined to be clusters.

There are several excellent tutorial papers available on clustering algorithms and unsupervised estimation algorithms [1], [2], and [3]. Thus, in this section only results of historical importance or particularly relevant to the work in this report will be described. As discussed previously, the report is principally concerned with recursive estimation algorithms that process sample vectors sequentially. This objective can be contrasted with the approach used by most of the clustering algorithms described in [1]-[3] where it is assumed that a fixed finite data set is to be processed (usually repeatedly) to determine the clusters.

The earliest unsupervised problem to receive considerable interest was one where a single unknown signal waveform was aperiodically transmitted through an unknown noisy stationary channel. Hence, the problem was to estimate the unknown waveform and to determine when it was present. An energy detector which evolved into a matched filter was formulated by Glaser [4]. Each time a signal was decided to be present, the current waveform was averaged into the estimate of the unknown signal waveform and the match filter part of the detector was weighted more heavily. The sampling times were determined by the zero crossings of the likelihood

function derivative. A simulation was presented to indicate convergence. Jakowitz, Shuey, and White [5] presented a digital system which used a "window" on samples of the received waveform. The vector representing the last k time samples of the received waveform were correlated with the current estimated signal vector. If the correlation coefficient exceeded a variable threshold, the estimated signal vector was updated. There was no apparent statistical basis for the definition of the threshold variation. This system was slightly modified, and the mean and variance of the threshold analyzed theoretically by Hinick [6].

In decision directed estimators, decisions made on past samples are utilized to make a decision on a current sample. These three papers are examples of better, early uses of the decision directed concept. In each of the papers an arbitrarily defined threshold determines whether or not the estimated statistics are updated. There have been several other papers using the decision directed concept which were heuristic to the extreme, and some of them show how a basically good approach can be abused.

In other early work, a minimum conditional risk solution to the unsupervised estimation problem was found by Daly [7] when he formulated the problem in terms of the Bayes algorithm. Since it partitioned the parameter space, the complexity of Daly's approach grew exponentially with the number of samples. Fralick [8] under an assumption that the posterior density of parameters characterizing M classes factors into the product of the posterior densities of the parameters charactering each class, found an iterative form of

the algorithm. This independence assumption resulted in some odd behavior of his estimator. Finally, Patrick and Hancock [13] found the general iterative form. The equivalence between Daly's work [7] and the general solution is shown in [54].

In [9], Cooper and Cooper showed how simple, easily calculated statistics such as a sample mean and the eigenvector corresponding to the largest eigenvalue of the sample space could be used to define an optimum two class decision boundary for a wide variety of statistics. Moment estimators for a two class Gaussian case were presented in [10]. The complexity of these estimators serves as an indication that the use of moment estimators in a complex multimodal sample space appears limited.

As approaches to unsupervised estimation became more statistical and less heuristic, the mathematical definition of the problem was emphasized more. Statistically, the samples are from several active source events, and hence, are from what is called a mixture.

Properties of mixtures were first considered in the statistical literature [11], [55] and applied to the unsupervised learning problem by engineers [13]. The problem of unsupervised estimation then is to resolve an unknown mixture into the underlying active source events, or equivalently, to find the indices (parameter vectors) and weights (mixing parameters) that express the unknown mixture density as a linear combination of density functions.

Implicit in the solution of unsupervised estimation problems is the concept of identifiability or that there should be a 1-1 mapping or relationship between a set of mixing parameters and the

resulting mixtures. Teicher's work on finite mixtures [11] was reduced by Yakowitz to a sufficiency theorem that a necessary and sufficient condition for identifiability of a class of finite mixtures is linear independence of the density functions in each finite mixture [12]. Yakowitz also showed that a large number of parametric families (including Gaussian) are identifiable.

The unsupervised estimation problem involves a class of problems including nonstationary class probabilities, statistically dependent observations, and unknown synchronization. The general problems were formulated in a Bayesian minimum sample conditional risk framework by Patrick in [13],[14] with the mixture concept emphasized. Combined with similar work by Lainiotis [15], this provides a precise formal definition of the problem.

More recently, maximum likelihood equations were used to obtain a maximum likelihood estimator in [16] and [17]. Under a Gaussian mixture assumption, numerical methods for finding a solution for a fixed data set were presented in [17], and a similar approach was used repeatedly on a growing data set in [16]. Stochastic approximation algorithms which seek the maximum of the average likelihood function were defined in [64] for a nonmixture problem. In [18]-[23] similar decision directed approaches were examined. The algorithms partition the sample space into M regions and adjust the boundaries according to where the samples fall. MacQueen [18] proved convergence with probability one for an M-ary case algorithm. The asymptotic probability of error for a two class Gaussian problem was evaluated in [19] and [21] for one and two unknown means respectively. Both of these results assumed

the correct values of the mixing parameters were known. A moment estimator for the two unknown mean, unknown mixing parameter case was presented in [20] and [51]. The updating in [18]-[21] was done using a uniform weighting sequence (i.e. 1, $\frac{1}{2}$, $\frac{1}{3}$,...). A nonuniform weighting sequence was derived in [22] and resulted from an attempt to find an "optimum" weighting sequence. A hardware implementation of the two class decision directed estimation algorithm was described in [23].

Other approaches to unsupervised estimation were examined in [37], [53],[65]. One group of algorithms selected a finite subset of a parametric family in order to obtain digitally implementable forms. The Bayes estimator was defined by computing the posterior density on the discretized parameter space and taking the average. The storage requirement for the algorithm was found to be impractically high for more than three or four unknown parameters. Since a direct implementation of the Bayesian results of [13]-[15] requires such a discretized parameter space[+], another method of seeking the minimum risk solution is needed. A similar conclusion on the inadequacy of a discretized parameter space approach can be reached for the minimum integral square error algorithms discussed in [37] and [53]. In an effort to reduce the storage requirements, a $\frac{1}{r}$ net was constructed on the parameter space. For the $\frac{1}{r}$ net variations of these algorithms, the parameter space is searched to find the optimum parameter vector. However, the entire sample sequence $X_1, X_2, \ldots, X_n$

[+]If there are only one or two unknown parameters, an analog technique such as described in [8] for example, does not require a discretized parameter space.

must be stored. Hence, $\frac{1}{r}$ net algorithms exchange a storage limitation for a time limitation and a growing storage requirement. Other algorithms discussed in these two papers had similar complexity difficulties.

The performance of decision procedures that use consistent unsupervised estimation algorithms to improve knowledge on the underlying statistics has received considerable study ([56]-[59], for example). Decision procedures that make decision on $X_1, X_2, \ldots, X_K$, only after their unsupervised estimation algorithm has processed an entire finite data set $\{X_k\}_{k=1}^{K}$ are called <u>compound</u> decision procedures. However, if $\{X_k\}_{k=1}^{n}$ is used to make a decision on X_n, this is called a <u>sequential compound</u> decision procedure. Most results assume that the samples corresponding to each active source come from a known parametric family of density functions. The algorithms used in the decision procedures estimate the parameters characterizing the mixture using $\{X_k\}_{k=1}^{K}$ and $\{X_k\}_{k=1}^{n}$ respectively. The decision procedures then make a Bayes decision against these estimates (i.e. assuming them to be correct). The measure of performance used in analyzing compound and sequential compound decision procedures is the <u>regret function</u>. This is defined as the difference between the average risk of the procedure using the estimator and the minimum average risk given the number of samples from each member of the parametric family. There is a good brief tutorial on compound and sequential compound decision procedures in [59].

Assuming a known fixed finite parametric family, and estimating the unknown mixing parameters, Van Ryzin found uniform bounds on the

rate of convergence of the regret function to zero for both compound [56] and sequential compound [57] decision procedures. This concluded over a decade of work on the problem defined by the fixed finite family assumption. Van Ryzin's unsupervised estimation algorithm used Robbins' functions [61] which minimize integral square error between the empirical density function and the mixture density resulting from the estimated mixing parameters [37]. Alens [58] assumed an infinite known parametric family and showed that the respective regret functions for sequential compound decision procedures using moment estimators and using maximum likelihood estimators converge to zero.

These results can be summarized as showing that if an unsupervised estimation algorithm converges, a Bayes decision procedure based on the algorithm converges. This is a worthwhile finding, but the algorithm used in [56]-[59] are impractical for most problems of interest.

1.2 Intersymbol Interference Literature Survey

The problem of intersymbol interference arises when using signaling rates at which some of the energy from one baud is smeared onto the following bauds. Methods to reduce the effect of intersymbol interference include signal design, receiver design, and joint transmitter-receiver design. In [24], Schwartzlander showed that it is possible to design band limited transmitted signals so that the received signals are confined to one sample baud. However, using a standard one sample correlation receiver[+] the resulting probability of error is larger

[+]Sample bauds are linearly operated on and then compared with a threshold in a correlation receiver.

than if a slight overlap is allowed. Aein and Hancock [25] showed that the use of past samples can improve the performance over a one sample optimum correlator approach. Aaron and Tufts [26] used both minimum average probability of error and minimum mean square error criteria to specify linear receiving filters for digital data transmission with intersymbol interference. For signal to noise ratios of practical interest, the optimum filters can be represented by matched filters followed by tapped delay lines. In [27], Tufts used a joint transmitter-receiver approach where the receiver was constrained to be linear. A minimum mean square error criterion was used on a pulse modulation system. Quincy [28] used a joint transmitter-receiver design approach in which the receiver was allowed to be nonlinear, and the performance criterion was minimum average probability of error. All of these results for reducing the effect of intersymbol interference require complete knowledge of the statistics of the problem.

Other results are available which bound the correlation receiver's error if the channel statistics are known [29] and [30]. These results can be used to evaluate a system's performance under wider conditions than assumed in [24]-[28].

Lucky [31] et. al. used a supervised estimation technique to implicitly estimate the channel gains. The equalization system was implemented using a tapped delay line with an adjustable gain on each tapped point. If there are $k + 1$ taps, mathematically the system seeks the best hyperplane projection of the k dimensional space corresponding to the last k samples. Moment estimators were

used by Chang [32] for unsupervised estimation on a one baud sample space with binary antipodal signals. However, moment estimators become cumbersome rapidly and it would be quite difficult to extend Chang's work to sample spaces of more than one baud. Also, moment estimators are not really suited for multimodal problems [2]. In Chapter 4, it will be shown that the use of a multiple baud sample space can improve performance, and estimation algorithms suitable for the intersymbol interference problem will be defined. A preliminary version of these results was presented in [33].

1.3 Approaches and Contributions

In this report a Bayesian framework is utilized as a guide towards "optimality", and to provide a unifying relationship for the approaches of the report. A Bayesian approach has the advantages of requiring a list of the a priori assumptions used to solve a problem, and of minimizing conditional risk against these assumptions.

In most unsupervised estimation problems, a direct implementation of an a posteriori approach requires the approximation of the parameter space with a finite set of vector points. In Chapter 2, the Bayes estimator on such a discretized parameter space is proven to converge to an asymptotic vector with probability one and in mean square. The asymptotic estimator and asymptotic rate of convergence are also found. These results on a discretized parameter space lead to a new continuous parameter space criterion. The maximum of this criterion minimizes average risk against the a priori assumption of a particular parametric family. For most parametric density function families the criterion is difficult to evaluate without resorting to

numerical methods; however, stochastic approximation algorithms which seek the maximum of the criterion surface are readily defined. The stochastic approximation algorithms defined in Chapter 2 are recursive and have fixed storage requirements. Since the algorithms converge under reasonable conditions on the criterion surface, they are asymptotically optimum.

The properties of the criterion surface are unknown in general. In Section 2.6 the contours of the criterion surface are found numerically. The contours show that for this problem the criterion has a unimodal surface. For other problems however, particularly for incorrect a priori knowledge (e.g., assuming the mixture is Gaussian when it is not), the criterion surface may be multimodal. Also, the contours show the fallacy of the assumption (such as in [8]) that the parameters are conditionally independent.

The assumption of a separable Gaussian family results in an easily evaluated form of the criterion. This criterion form allows the definition of a simple clustering technique in Subsection 2.5.2. The algorithm is similar to others in the literature except that it uses a criterion derived from an optimum approach. The criterion resulting from a separable Gaussian assumption has the disadvantage of not being able to resolve some mixtures which violate the separability assumption. A mixture resolution limit is investigated in Section 2.6.

A class of decision directed algorithms is defined in Chapter 3 which minimize a criterion derived from the separable Gaussian family criterion. The class of algorithms contains the estimators from

several previous papers. The decision directed estimators are given the interpretation of stochastic approximation algorithms with random weights. Results in the literature on conventional stochastic approximation algorithms are then compared with the properties of the decision directed estimator class found here. In Subsection 3.1.2, the relationships between different decision directed estimators are discussed, and it is shown the scattered results on particular algorithms actually apply to the entire algorithm class. The M-ary class of estimators is proven to converge with probability one using an extension of a proof due to MacQueen [18]. The remaining results of this chapter are for the M = 2 class case. Three sets of assumptions on the mixing parameters lead to the definition of three subclasses of algorithms. For each of these subclasses, all of the algorithms in the subclass are proven to converge to a common vector point. Theoretical asymptotic probability of error curves and experimental dynamic convergence results for these M = 2 subclasses are presented.

In Chapter 4, decision directed estimators are applied to the problem of intersymbol interference. The mode or cluster structure of the multibaud sample space is discussed and used to relate the approaches of previous papers. The decision directed estimator defined in this chapter as Algorithm 1 is a direct application of the M-ary algorithm class of Chapter 3. Experimental dynamic convergence curves are presented and it is noted that this algorithm converges to the asymptotic vector very slowly. The second algorithm defined in Chapter 4 uses the a priori knowledge on the mutual constraints on

the locations of modes in the multibaud sample space. The experimental results converged extremely rapidly for a multimodal problem such as this. Since basically the algorithm estimates the unknown channel gains, a "tracking mode" form for nonstationary statistics (i.e., slowly varying channel gains) is very reasonable. Extensions and related problems such as differential phase shift keying (DPSK) communications systems are also discussed.

II. MINIMUM RISK SOLUTIONS

Except for a few degenerate cases, an unsupervised estimation algorithm which minimizes sample conditional risk must compute the posterior density on the parameters that characterize the problem. In this chapter attention is restricted to the case of a stationary random process with sample conditional independence and only one of M classes ω^i, $i = 1,2,\ldots,M$ active for each ℓ dimensional vector sample X_k. It is assumed that the class distributions belong to a known functional family $\mathfrak{F}$ in which every member is indexable by a parameter vector. In general, such a family has an infinite number of members and the indexing set forms a continuous parameter space. In order to actually construct the posterior density, it is necessary to approximate the family with a finite subset. In Section 2.1, the Bayes estimator on the finite subset is proven to converge with probability one to the parameter index maximizing a measure of information. Rates of convergence in mean square are then evaluated for some families of interest. The rest of the chapter examines the problem of asymptotically maximizing the measure of information (thus minimizing the risk) on a continuous parameter space with algorithms constrained to have fixed storage and to be recursive.

2.1 Preliminaries

Let $\mathfrak{F} = \{F(x|\alpha);\ \alpha \in G,\ x \in R^{\ell}\}$ constitute a family of ℓ dimensional cumulative distribution functions indexed by a point α in a Borel subset G of Euclidean m-space R^m such that $F(x|\alpha)$ is measurable in $R^{\ell} \times G$, and $\mathfrak{F}$ is dominated by Lebesgue measure [33] (denoted by ν) with

$$f(x|\alpha) = (dF(x|\alpha)/d\nu)(x) \leq c$$

for some $c < \infty$. Denote this corresponding family of densities by $\mathfrak{F}' \subset L_1 \cap L_2$ where L_1 and L_2 are function spaces of Lebesque integrable and square integrable functions, respectively. Additional assumptions on $\mathfrak{F}$ will be made for some of the results obtained later.

A finite <u>mixture</u> from $\mathfrak{F}$ is defined [11]

$$H\left(x \middle| \left\{\alpha^{k_i}, P\left(\alpha^{k_i}\right)\right\}_{i=1}^{M^k}\right) = \sum_{i=1}^{M^k} F\left(x|\alpha^{k_i}\right) P\left(\alpha^{k_i}\right) \qquad (2.1)$$

where $P(\alpha^{k_i}) > 0$, $i = 1,2,\ldots,M^k$, $\sum_{i=1}^{M^k} P(\alpha^{k_i}) = 1$, and $M^k < \infty$. The weights $\{P(\alpha^{k_i})\}_{i=1}^{M^k}$ are called <u>mixing parameters</u>. Let ρ (a Borel subset of R) be a set of mixing parameter values, and M' an a priori upper bound on the number of non zero mixing parameters. The class of all finite mixture of $\mathfrak{F}$ having $M^k \leq M'$ and $P(\alpha^{k_i}) \in \rho$ is then,

$$\{H(x|B^k)\} = \left\{\sum_{i=1}^{M'} F\left(x|\alpha^{k_i}\right) P\left(\alpha^{k_i}\right) : B^k \in \mathfrak{B}^{M'}\right\} \qquad (2.2)$$

where

$$B^k \triangleq \left\{\alpha^{k_i}, P\left(\alpha^{k_i}\right)\right\}_{i=1}^{M'}$$

and

$$\mathfrak{B}^{M'} = \left\{ B^k : B^k \in (\Omega \times \mathcal{P})^{M'}, \sum_{i=1}^{M'} P\left(\alpha^{k_i}\right) = 1, \alpha^{k_i} \neq \alpha^{k_j}\ i \neq j \right.$$
$$\left. \text{for } 1 \leq i,j \leq M'; \sigma[\underline{\alpha}(B^k)] \neq \sigma[\underline{\alpha}(B^j)], k \neq j \right\} \tag{2.3}$$

The notation X denotes cartesian product, $\underline{\alpha}(\cdot)$ is an operator such that $\underline{\alpha}(B^k) = \{\alpha^{k_i} P(\alpha^{k_i}), P(\alpha^{k_i})\}_{i=1}^{M'}$, and $\sigma[\underline{\alpha}(B^k)]$ denotes any permutation of the component pairs $(\alpha^{k_i} P(\alpha^{k_i}), P(\alpha^{k_i}))$ in $\underline{\alpha}(B^k) = \{\alpha^{k_i} P(\alpha^{k_i}), P(\alpha^{k_i})\}_{i=1}^{M'}$.

Further restrictions on the α^{k_i} and $P(\alpha^{k_i})$ that can be combined in a vector B^k may reduce the number of points in $\mathfrak{B}^{M'}$ from that of (2.3); however, for notational simplicity such additional assumptions are not considered here. The set $(\Omega \times \mathcal{P})^{M'}$ is called the product parameter space.

2.2 Bayes Minimum Conditional Risk Solution for Finite $\mathfrak{B}^{M'}$

The Bayes solution which minimizes the conditional risk given $\mathfrak{B}^{M'}$ and the a priori p.d.f. on the product parameter space $p_o(B^k)$, computes the posterior probability density function $p(B^k|Y_n)$ on all $B^k \in \mathfrak{B}^{M'}$,

$$p(B^k|X_1) = \frac{\left[\sum_{i=1}^{M'} f\left(X_1|\alpha^{k_i}\right)P\left(\alpha^{k_i}\right)\right]p_o(B^k)}{\sum_{B^k \in \mathcal{B}^{M'}} [\text{numerator}]}$$

$$p(B^k|Y_n) = \frac{\left[\sum_{i=1}^{M'} f\left(X_n|\alpha^{k_i}\right)P\left(\alpha^{k_i}\right)\right]p(B^k|Y_{n-1})}{\sum_{B^k \in \mathcal{B}^{M'}} [\text{numerator}]} \quad n > 1 \tag{2.4}$$

where $Y_n \triangleq \{X_1, X_2, \ldots, X_n\}$, [14]. For (2.4) to actually be implemented, it is necessary that $\mathcal{B}^{M'}$ be a finite set of V vector points $\{B^k\}_{k=1}^{V}$. The results of this and the next section assume such a finite set.

The Bayes estimator for a quadratic loss function on the discretized parameter space $\mathcal{B}^{M'}$ is

$$\hat{B}(Y_n) = \sum_{k=1}^{V} B^k \, p(B^k|Y_n) \tag{2.5}$$

where $p(B^k|Y_n)$ is calculated from (2.4). Denoting the true mixture by $h(x)$, define the quantities

$$\eta(B^k) \triangleq E[\ln h(x|B^k)]$$

$$= \int [\ln h(x|B^k)]h(x)dx \qquad B^k \in \mathcal{B}^{M'} \tag{2.6}$$

The convergence properties of the posterior density presented here

strongly depend on $\eta(B^k)$ which is a measure of distance between $h(x|B^k)$ and $h(x)$.

Convergence properties of the Bayes estimator defined on the finite set $\mathcal{B}^{M'}$ will be established in Theorem 1. In this theorem it is shown that under certain conditions the Bayes estimator on a finite parameter space $\mathcal{B}^{M'}$ converges in mean square and with probability one to an asymptotic vector point. Rates of convergence in mean square and the asymptotic vector are also evaluated. For the case where there is a unique vector point (denoted by B^*) in $\mathcal{B}^{M'}$ that maximizes the $\eta(B)$ measure of (2.6), the Bayes estimator is proven to be a superefficient estimator (i.e., converges in mean square faster than $1/n$) if $E[|\ln h(x|B^*)|^s] < \infty$ for some $s > 3$. The asymptotic solution for this case is B^*. If there is more than one vector point maximizing $\eta(B)$, a bound on the mean square probability mass associated with the rest of the points in $\mathcal{B}^{M'}$ is shown to go to zero as a power of $1/n$.

If the set of mixtures defined by the finite $\mathcal{B}^{M'}$ contains the true mixture (i.e., $h(x) \in \{h(x|B^k)\}$), it is shown in the theorem that under certain conditions only those $h(x|B^k) \equiv h(x)$ will maximize $\eta(B)$. However, if the restriction on $\{h(x|B^k)\}$ is added that $h(x|B^i) \neq h(x|B^j)$ on a set of measure greater than zero for all $B^i, B^j \in \mathcal{B}^{M'}$ with $i \neq j$, then only one $h(x|B^k)$ is in the set maximizing $\eta(B)$.

It should be noted that since $\eta(B)$ may be multimodal, there is no guarantee that the vector points maximizing $\eta(B)$ will be close

to each other in the Euclidian distance sense if $h(x) \notin \{h(x|B^k)\}$. Exceptions may occur in cases where the discretization can be taken "fine enough" as in some one parameter problems.

The following list of assumptions will be used as needed in the different parts of Theorem 1.

(1) $h(X_n|X_1,X_2,\ldots,X_{n-1}) = h(X_n)$

(2) There exists a positive integer $s > 1$ such that $E[|\ln h(x|B^k)|^s] < \infty$ for all $B^k \in \mathcal{B}^{M'}$. Denote by s^* the largest even integer for which this absolute moment exists.

(3) The probability measures corresponding to $\{h(x|B^k)\}$ are absolutely continuous with respect to Lebesque measure ν.

(4) $\nu[x : |h(x|B^k) - h(x|B^j)| > 0] > 0$ for all B^k, $B^j \in \mathcal{B}^{M'}$, $k \neq j$.

These last two conditions require no diracs and that pairs of mixture densities be different on an open set. They are satisfied for such functional families used in practice as the Gaussian, exponential, and uniform density families.

(5) $\{h(x|B^k)\}$ contains the true mixture $h(x)$.

In the proof of the theorem, the sample conditional independence and absolute moment existence assumptions ((1) and (2)) will allow application of the strong law of large numbers. If assumptions (3) and (4) are not satisfied, the distance measures between the $\{h(x|B^k)\}$ and $h(x)$ defined in (2.6) are not different for the respective $h(x|B^k)$. These two conditions together with (5) will be used to establish a uniqueness property of the Bayes estimator.

THEOREM 1

If assumptions (1) and (2) are satisfied, and for $\bar{B}^*$ defined

$$\bar{B}^* = \frac{\sum_{B^k \in \{B^{i*}\}} B^k \, p_o(B^k)}{\sum_{B^k \in \{B^{i*}\}} p_o(B^k)} \tag{2.7}$$

where

$$\{B^{i*}\} = \{\arg[\max_{B^k \in \mathcal{B}^M} \eta(B^k)]\}^+, \tag{2.8}$$

then the Bayes estimator $\hat{B}(Y_n)$ defined by (2.5) has the properties

(a) $P[\lim_{n \to \infty} \hat{B}(Y_n) = \bar{B}^*] = 1$

(b) There exists a positive number $c < \infty$ such that for n large enough,

$$E\left[\left\| \sum_{B^k \notin \{B^{i*}\}} B^k \, p(B^k|Y_n) \right\|^2\right] \leq cn^{-s^*/2}$$

and if there is only one point[++] in $\{B^{i*}\}$ this implies

$$E[\|\hat{B}(Y_n) - \bar{B}^*\|^2] \leq cn^{-s^*/2}$$

[+] For convenience in following the proof it can be assumed that there is only one B^k which maximizes $\eta(B^k)$, although for $h(x) \notin \{h(x|B^k)\}$ this is not necessarily the case.

[++] If $\eta(B)$ has no flat regions and is reasonably smooth, the probability that two parameter space points selected at random have the same $\eta(B)$ value is essentially zero.

(c) $\lim_{n \to \infty} E[\|\hat{B}(Y_n) - B^*\|^2] = 0$

If in addition (3), (4), and (5) are satisfied

(d) B^* is unique and $h(x|B^*) \equiv h(x)$.

Proof: From (2.4) and (2.5) the Bayes estimator can be expressed

$$\hat{B}(Y_n) = \sum_{k=1}^{V} B^k \frac{\prod_{j=1}^{n} h(X_j|B^k)\, p_o(B^k)}{\sum_{i=1}^{V} \prod_{j=1}^{n} h(X_j|B^i) p_o(B^i)}$$

$$= \sum_{k=1}^{V} B^k \frac{\left[e^{\frac{1}{n}\sum_{j=1}^{n} \ln h(X_j|B^k)}\right]^n p_o(B^k)}{\sum_{i=1}^{V} \left[e^{\frac{1}{n}\sum_{j=1}^{n} \ln h(X_j|B^i)}\right]^n p_o(B^i)} \qquad (2.9)$$

The strong law of large numbers can be applied because of conditions (1) and (2) so that

$$P\left[\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \ln h(X_j|B^k) = \eta(B^k)\right] = 1$$

Thus,

$$P\left[\lim_{n \to \infty} e^{\frac{1}{n}\sum_{j=1}^{n} \ln h(X_j|B^k)} = e^{\eta(B^k)}\right] = 1 \qquad (2.10)$$

and with probability one for any $B^* \in \{B^i*\}$,

$$\lim_{n \to \infty} \left\{ \frac{p(B^k|Y_n)}{p(B^*|Y_n)} \right\}^{1/n} = \frac{e^{\eta(B^k)}}{e^{\eta(B^*)}} \qquad (2.11)$$

From (2.11), if $B^k \notin \{B^i*\}$, $P[\lim_{n \to \infty} \frac{p(B^k|Y_n)}{p(B^*|Y_n)} = 0] = 1$ so as a consequence of (2.7) and (2.9) it follows that $P[\lim_{n \to \infty} \hat{B}(Y_n) = \bar{B}^*] = 1$ proving the first assertion of the theorem.

The mean square error expression can be expanded

$$E[\|\hat{B}(Y_n) - \bar{B}^*\|^2] = E\Big[\sum_{k=1}^{V} \sum_{j=1}^{V} (B^k - \bar{B}^*)^t (B^j - \bar{B}^*) p(B^k|Y_n) p(B^j|Y_n)\Big]$$

$$= \sum_{k=1}^{V} \sum_{j=1}^{V} (B^k - \bar{B}^*)^t (B^j - \bar{B}^*) E[p(B^k|Y_n) p(B^j|Y_n)] \qquad (2.12)$$

Denote the number of vectors in $\{B^i*\}$ by V^* and for notational convenience assume the vectors $\{B^k\}_{k=1}^{V}$ are ordered so that the first V^* are in $\{B^i*\}$. The mean square error can then be bounded,

$$E[\|\hat{B}(Y_n) - \bar{B}^*\|^2] \leq \sum_{k=1}^{V^*} \sum_{j=1}^{V^*} (B^k - \bar{B}^*)^t (B^j - \bar{B}^*) E[p(B^k|Y_n) p(B^j|Y_n)]$$

$$+ 2 \sum_{k=1}^{V^*} \sum_{j=V^*+1}^{V} (B^k - \bar{B}^*)^t (B^j - \bar{B}^*) \min\{E[p(B^k|Y_n)], E[p(B^j|Y_n)]\}$$

$$+ \sum_{k=V^*+1}^{V} \sum_{j=V^*+1}^{V} (B^k - \bar{B}^*)^t (B^j - \bar{B}^*) \min\{E[p(B^k|Y_n)], E[p(B^j|Y_n)]\} \tag{2.13}$$

The next several steps will be used to show that for $B^k \notin \{B^i*\}$, $E[p(B^k|Y_n)] \to O(n^{-s^*/2})$ which along with (2.13) will allow us to conclude (b).

For $B^k \notin \{B^i*\}$, the random variable $p(B^k|Y_n)$ can be bounded by the random variable[+]

$$z^k(Y_n) \triangleq \begin{cases} \dfrac{p(B^k|Y_n)}{p(B^*|Y_n)} : \dfrac{p(B^k|Y_n)}{p(B^*|Y_n)} < e^{-n(d^k)/2} \\ \\ 1 \quad : \text{elsewise} \end{cases} \tag{2.14}$$

where B^* is any vector in $\{B^i*\}$ defined by (2.8) and $(d^k) \triangleq \eta(B^*) - \eta(B^k)$. Then the expectation of the random variable $z^k(Y_n)$ bounds the expectation of the random variable $p(B^k|Y_n)$,

$$E[p(B^k|Y_n)] \leq E[z^k(Y_n)]. \tag{2.15}$$

Defining

$$r^k(Y_n) = \frac{p(B^k|Y_n)}{p(B^*|Y_n)}$$

[+]This avoids difficulties in bounding (2.13) caused by the normalization term $\sum_{B^k \in \mathcal{B}^{M'}} p(B^k|Y_n)$ in the denominator of (2.3).

the expectation of $z^k(Y_n)$ can be expanded,

$$E[z^k(Y_n)] = \int_0^\infty z^k(Y_n) p(z^k(Y_n)) d\, z^k(Y_n)$$

$$= \int_0^{e^{-n(d^k)/2}} r^k\, p(r^k) dr^k + \int_{e^{-n(d^k)/2}}^\infty p(r^k) dr^k \qquad (2.16)$$

The second integral in (2.16) is the probability $r^k(Y_n)$ is greater than or equal to $e^{-n(d^k)/2}$. Re-expressing this integral as a probability,

$$P\left[\frac{p(B^k|Y_n)}{p(B^*|Y_n)} \geq e^{-n(d^k)/2}\right]$$

$$= P\left[\frac{1}{n} \sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) \geq -(d^k)/2 - \frac{1}{n} \ln \frac{p_o(B^k)}{p_o(B^*)}\right]$$

$$= P\left[\frac{1}{n} \sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) + d^k \geq \frac{d^k}{2} - \frac{1}{n} \ln \frac{p_o(B^k)}{p_o(B^*)}\right]$$

$$\leq P\left[\left|\frac{1}{n} \sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) + d^k\right| \geq \frac{d^k}{3}\right] \qquad (2.17)$$

for $n > 6\left|\ln \left(\frac{p_o(B^k)}{p_o(B^*)}\right)\right|/(d^k)$. Using the Markov inequality (see Loeve [35] p. 158) equation (2.17) can be bounded giving

$$\int_{e^{-n(d^k)/2}}^{\infty} p(r^k)dr^k \leq \left(\frac{3}{d^k}\right)^s$$

$$\cdot E\left[\left|\frac{1}{n}\sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) + d^k\right|^s\right] \qquad (2.18)$$

Using the property of L_p spaces that for any random variable U,

$$(E[|U|^k])^{1/k} \leq (E[|U|^r])^{1/r} \qquad 1 \leq k \leq r \leq \infty \qquad (2.19)$$

and a straightforward application of Minkowski's inequality (see Rudin [34], p. 62), it is shown in Appendix A that after some simplification (2.18) can be bounded for s even and $n \geq s$ by

$$\int_{e^{-n(d^k)/2}}^{\infty} p(r^k)dr^k \leq n^{-s/2}\left[\frac{s+2^{s-1}-1}{s/2!}\right]\left\{\sum_{i=0}^{s}\sum_{j=0}^{i} (d^k)^{s-i}\binom{s}{i}\binom{i}{j}\right.$$

$$\left.\cdot(E[|\ln h(x|B^k)|^i])^{i-j}(E[|\ln h(x|B^*)|^i])^i\right\}\left(\frac{3}{d^k}\right)^s \qquad (2.20)$$

From (2.19) this exists for any even $s \leq s^*$ and s^* gives the fastest rate. Substituting (2.20) into (2.16) and bounding the first integral we finally obtain,

$$E[z^k(Y_n)] \leq e^{-n(d^k)/2} + n^{-s^*/2}\left(\frac{3}{d^k}\right)^{s^*}\left[\frac{s^*+2^{s^*-1}-1}{s^*/2!}\right]\left\{\sum_{i=0}^{s^*}\sum_{j=0}^{i} (d^k)^{s^*-i}\right.$$

$$\left.\binom{s^*}{i}\binom{i}{j}(E[|\ln h(x|B^k)|^i])^{i-j}(E[|\ln h(x|B^*)|^i])^j\right\} \qquad (2.21)$$

As a consequence of (2.15) this proves the first part of (b).

Also this shows that the second and third summands in (2.13) go to zero at $O(n^{-s^*/2})$. If $V^* = 1$ (i.e., there is only one point in $\{B^k\}$ that maximizes $\eta(B^k)$) then the first summand in (2.13) is identically zero and $E[\|\hat{B}(Y_n) - \bar{B}^*\|^2] \to O(n^{-s^*/2})$ proving the second part of (b).

To establish (c), a proof similar to the one used in (a) shows that

$$P\left[\lim_{n \to \infty} p(B^k|Y_n)p(B^j|Y_n) = \begin{cases} \dfrac{p_o(B^k)p_o(B^j)}{\left(\sum_{B^t \in \{B^{i*}\}} p_o(B^t)\right)^2} & : B^k, B^j \text{ both } \in \{B^{i*}\} \\ 0 & : \text{elsewise} \end{cases}\right]$$

$$= 1 \qquad (2.22)$$

Since the random variables $[p(B^k|Y_n)p(B^j|Y_n)]$ are bounded, (2.22) gives that

$$\lim_{n \to \infty} E[p(B^k|Y_n)p(B^j|Y_n)] = \begin{cases} \dfrac{p_o(B^k)p_o(B^j)}{\left(\sum_{B^t \in \{B^{i*}\}} p_o(B^t)\right)^2} & : B^k, B^j \text{ both } \in \{B^{i*}\} \\ 0 & : \text{elsewise} \end{cases}$$

$$(2.23)$$

Using this result in (2.12) and rearranging $\lim_{n \to \infty} E[\|\hat{B}(Y_n) - \bar{B}^*\|^2] = 0$ proving (c).

To finish the proof, it is well known (see Kullback [36], p. 14, for example) that under conditions (3) and (4), if $h(x) \in \{h(x|B^k)\}$, only those $h(x|B^i{}^*) \equiv h(x)$ maximize $E[\ln h(x|B^k)]$. Also (4) implies uniqueness of the mixture densities (such a unique mapping between the $\{B^k\}$ and the $\{h(x|B^k)\}$ is called identifiability). Hence, there is only one B^* such that $h(x|B^*) = h(x)$ proving (d).

Comment: The definition of $\eta(B^k)$ in (2.8) is a measure of information. Hence asymptotically, the Bayes algorithm maximizes this measure of information.

2.3 Convergence Rates for Some Functional Families

If some restrictions are applied to the density functions of the finite family or the true mixture, a stronger mean square convergence result can be proven. In Proposition 1, a Gaussian assumption is made establishing "almost exponential" convergence, and Proposition 2 requires that $0 < h(x|B_k) < \infty$, $B_k \in \mathcal{B}_M$, for all points $x \in R^\ell$ for which the true mixture $h(x)$ is non zero--a reasonable assumption for many practical problems (for example, $\{h(x|B_k)\}$ Gaussian mixtures and $h(x)$ a truncated Gaussian mixture).

Definition: If a sequence $\{T_k\}$ is defined such that for any positive finite integer s, $\lim_{k \to \infty} k^s T_k = 0$, then $\{T_k\}$ is said to converge to zero almost exponentially.

Proposition 1. If assumption (1) of Theorem 1 holds and in addition, if

(2) the family of densities $\{f(x|\alpha^k)\}$ is Gaussian

(3) the true mixture $h(x)$ is a finite mixture of Gaussian densities

(4) only one parameter vector $B^* \in \{B^k\}_{k=1}^{V}$ maximizes $\eta(B^k)$

then

$$E[\|\hat{B}(Y_n) - B^*\|^2] \to 0 \text{ almost exponentially}$$

Proof: If it can be established that assumptions (1) and (2) of Theorem 1 are satisfied, then from the proof of that theorem it is sufficient to show $E[|\ln h(x|B^k) - \ln h(x|B^*)|^s]$ exists for all positive finite integers s. Because $\{f(x|\alpha^k)\}$ are Gaussian from (2) above, then as $x \to \pm \infty$, $\ln h(x|B^k) \to 0\ (-x^t x)$ and there exist constants $c, c' < \infty$ such that $c + c'|x^t x|^s > |\ln h(x|B^k) - \ln h(x|B^*)|^s$

$$\int c'|x^t x| h(x)dx \geq \int |\ln h(x|B^k) - \ln h(x|B^*)|^s h(x)dx - c \tag{2.24}$$

From assumption (3) above, $h(x)$ is a Gaussian mixture and the left integral is just an absolute moment of $h(x)$. Since all finite moments of a Gaussian mixture exist, the integral of (2.24) exists for all finite s. Hence (1) and (2) of Theorem 1 are satisfied for all positive s and the proposition proven.

Proposition 2. If assumption (1) of Theorem 1 is satisfied and in addition, if

(2) $\sup_{x \in D} |\ln h(x|B^k)| \leq c$ for some $c < \infty$ for all $B^k \in \mathcal{B}^{M'}$

where the set D is the support of the true mixture $h(x)$ (i.e., D equals the closure of the set $\{x : h(x) > 0, x \in R^{\ell}\}$).

(3) only one parameter vector $B^k \in \mathcal{B}^{M'}$ maximizes $\eta(B^k)$

then

$$E[\|B(Y_n) - B^*\|^2] < \rho^{n^{\frac{1}{2}}} \qquad 0 \leq \rho < 1 \text{ for } n \text{ large enough.}$$

Proof: From the proof of Theorem 1, it is sufficient to find an exponential bound on $P[\sum_{j=1}^{n} h(X_j|B^k) - \ln h(X_j|B^*) + n(d^k) > n(d^k)/3]$. Since the random variable $[\ln h(X_j|B^k) - \ln h(X_j|B^*)]$ is effectively bounded because of assumption (2) above, then $(\sigma^k)^2 \triangleq E[|\ln h(x|B^k) - \ln h(x|B^*)|^2]$ is finite. Also, letting

$$c' \triangleq \sup_{\{x : h(x) > 0\}} |\ln h(x|B^k) - \ln h(x|B^*) + (d^k)|$$

then $c' < \infty$. As a consequence of Kolmogorov's inequalities (see Loeve [5], p. 254), for arbitrary $\epsilon > 0$ and $(\frac{c'}{\sigma^k})\epsilon > 1$,

$$P\left\{\frac{1}{n^{\frac{1}{2}}\sigma^k}\left[\sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) + n(d^k)\right] > \epsilon\right\} < \exp\left[-\frac{\epsilon(\sigma^k)}{4c'}\right] \tag{2.25}$$

Letting $\epsilon = n^{\frac{1}{2}}(\frac{d^k}{3\sigma^k})$, (2.25) is bounded for $n > (\frac{3(\sigma^k)^2}{c'(d^k)})^2$,

$$P\left\{\left[\sum_{j=1}^{n} \ln h(X_j|B^k) - \ln h(X_j|B^*) + n(d^k)\right] > n(d^k)/3\right\} < e^{-n^{\frac{1}{2}}(d^k)/12c'}$$

proving the proposition.

2.4 Some Approaches to Asymptotic Minimum Risk Solutions

The previous two sections showed that the maximum of the measure of information $\eta(B)$ on a finite set of vectors $\mathcal{B}^{M'}$ defines the asymptotic minimum risk solution relative to $\mathcal{B}^{M'}$. Because of the Bayes algorithm definition in (2.4), the posterior density on a continuous parameter space is proportional to the posterior density on a discrete parameter space at the discrete parameter vector points. Hence, for <u>any</u> set of vectors $\mathcal{B}^{M'}$, the asymptotic minimum risk solution relative to $\mathcal{B}^{M'}$ is defined by the parameter vector $B \in \mathcal{B}^{M'}$ that maximizes $\eta(B)$. Storage requirement calculations in [37] show that for most practical problems[+], storage limitations prohibit taking the vector points in the discretized parameter space close enough for the Bayes estimator to be useful. The rest of this chapter investigates asymptotically optimum algorithms with fixed storage requirements which seek the maximum of the regression surface $\eta(B)$ on a continuous parameter space. Of course for finite sample size, the algorithms are suboptimum with respect to a non implementable posterior density approach on such a continuous space. The approaches discussed in this section are:

[+]Exceptions include problems characterized by one parameter such as target bearing in sonar or signal frequency [8], [60].

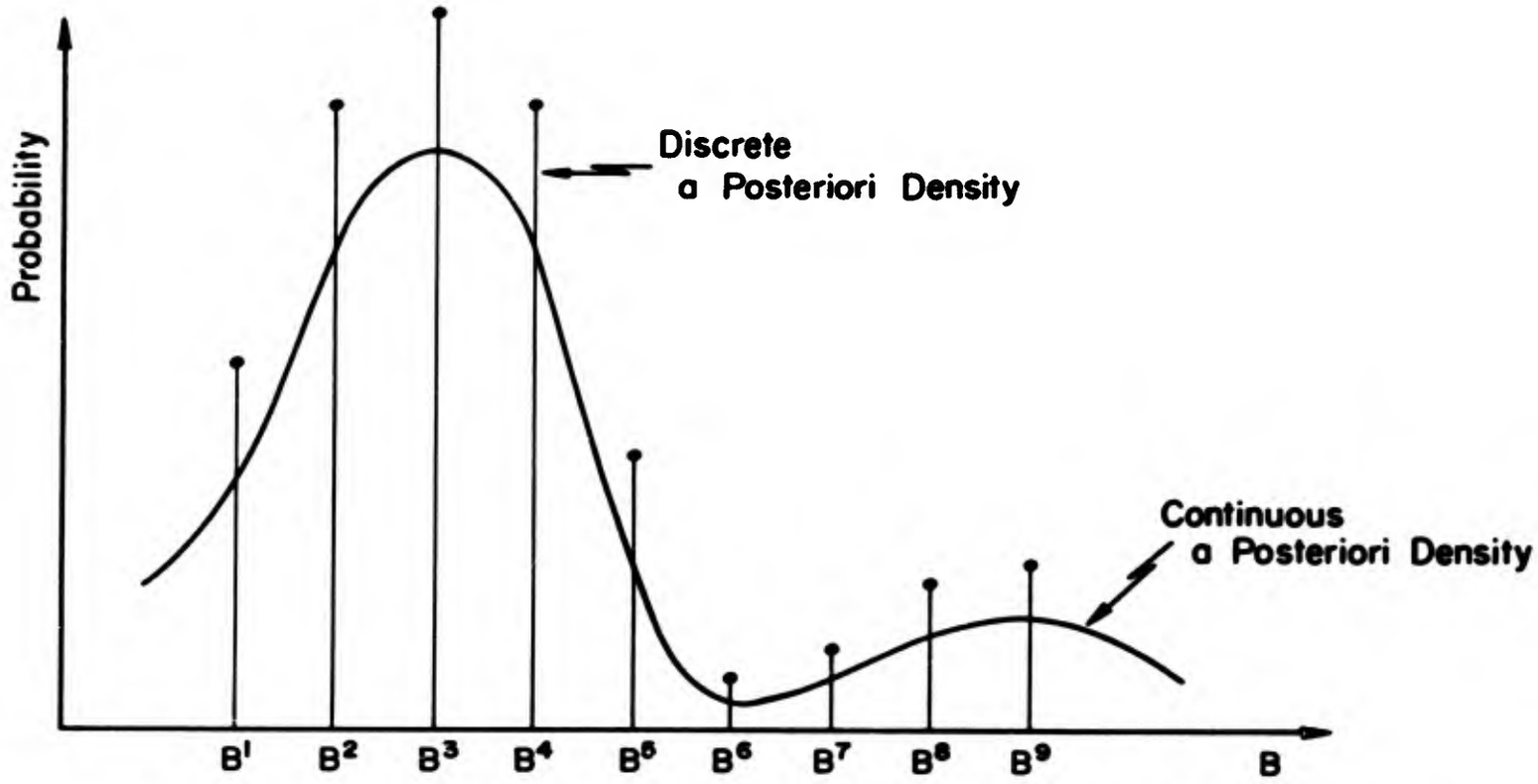


Figure 1. Relationship Between the Continuous and Discrete A Posteriori Densities Computed from the Same Set of Samples $\{X_k\}_{k=1}^{n}$.

1. Stochastic hillclimb on the regression surface $\eta(B)$

2. Stochastic solution of maximum likelihood related equations.

3. Nonlinear regression on estimates of $\eta(B)$ on a finite set of vectors $\mathcal{B}^{M'}$.

The stochastic approximation algorithm of 2.4.1 searches for the maximum of the regression surface $\eta(B)$ by estimating the slope and hillclimbing. The algorithm of 2.4.2 seeks the zero of the regression surface $\frac{d}{dB}[\eta(B)]$. The last approach discussed in this section is an unsuccessful attempt to utilize knowledge on the structure of $\eta(B)$ to interpolate between points of the natural log transformed posterior density on a finite set $\{B^k\}_{k=1}^{V}$.

2.4.1 Stochastic Hillclimb on the Regression Surface $\eta(B)$.

This stochastic approximation algorithm is a modification of the Keifer Wolfowitz procedure [38] to maximizing $E[\ln h(x|B)]$ subject to the constraints that $\sum_{i=1}^{M'} P(\alpha^i) = 1$ and $P(\alpha^i) \geq 0 \quad i = 1,2,\ldots,M'$. Let $\mathcal{B}^{M'}$ be a known closed bounded convex set (see equation (2.3)) containing the solution vector, and for notational convenience, order the components of B such that the first M' are the mixing parameters. The maximization problem is then to maximize the expectation of

$$\ln h(x|B) - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right] \tag{2.26}$$

(which is $\eta(B)$) where Ψ is a unknown Lagrange multiplier. Taking the partial derivative with respect to $P(\alpha^i)$,

$$\frac{f(x|\alpha^i)}{h(x|B)} - \Psi = 0 \qquad i = 1,2,\ldots,M'$$

multiplying by $P(\alpha^i)$ and summing over i shows that $\Psi = 1$. To approximate the slope of (2.26) in the direction of the j^{th} component of B,

$$D^j = \left\{ \begin{array}{l} \ln h(x|B + ce^j) - \left[\sum_{i=1}^{M'} (P(\alpha^i) + ce^j \delta^{ij}) - 1 \right] \\ - \ln h(x|B - ce^j) + \left[\sum_{i=1}^{M'} (P(\alpha^i) - ce^j \delta^{ij}) - 1 \right] \end{array} \right\} /2c \qquad (2.27)$$

where e^j is the j^{th} column of an $(m + 1)M'$ dimensional identity matrix, c is a positive number, and δ^{ij} is the Kronecker delta function. Evaluating (2.27),

$$D^j = \begin{cases} \dfrac{\ln h(x|B + ce^j) - \ln h(x|B - ce^j)}{2c} - 1 & 1 \le j \le M' \\ \dfrac{\ln h(x|B + ce^j) - \ln h(x|B - ce^j)}{2c} & M'+1 \le j < (m+1)M' \end{cases} \qquad (2.28)$$

Substituting $X_{n+1} = x$, $c_{n+1} = 2c$, and $B_n = B$ in (2.28), the stochastic approximation algorithm for maximizing $\eta(B)$ is defined,

$$B^j_{*n+1} = \begin{cases} [B^j_n + b_{n+1} D^j_{n+1}] / \left[1 + b_{n+1} \sum_{i=1}^{M'} D^i_{n+1}\right] & 1 \le j \le M' \\ [B^j_n + b_{n+1} D^j_{n+1}] & M'+1 \le j \le (m+1)M' \end{cases}$$

and

$$B_{n+1} = \begin{cases} B_{*n+1} & : B_{*n+1} \in \mathcal{B}^{M'} \\ B_n & : \text{otherwise} \end{cases}$$

where the vector $D_{n+1} \triangleq (D^1_{n+1}\ D^2_{n+1}, \ldots, D^{(m+1)M'}_{n+1})^t$, with

$$D^j_{n+1} = \begin{cases} \dfrac{\ln h(X_{n+1}|B_n + c_{n+1}\, e^j) - \ln h(X_{n+1}|B_n - c_{n+1}\, e^j)}{c_{n+1}} - 1 & 1 \le j \le M' \\ \dfrac{\ln h(X_{n+1}|B_n + c_{n+1}\, e^j) - \ln h(X_{n+1}|B_n - c_{n+1}\, e^j)}{c_{n+1}} & M'+1 \le j \le (m+1)M' \end{cases}$$

and the non-negative sequences satisfy

$$\sum_{n=2}^{\infty} b_n = \infty, \quad \lim_{n \to \infty} c_n = 0, \quad \sum_{n=2}^{\infty} \left(\frac{b_n}{c_n}\right)^2 < \infty.$$

While being of extreme importance in the performance of the algorithm, the selection of the initial vector B_1 is arbitrary.

If the first order partials of $\eta(B)$ exist and are bounded, $E[|\ln h(x|B) - \eta(B)|^2] < c$ for some $c < \infty$ for $B \in \mathcal{B}^{M'}$, and if for arbitrary $\epsilon > 0$,

$$\epsilon < \sup_{B \in \mathcal{B}^{M'}} \|B - B^*\| < \epsilon^{-1} \qquad (B^* - B)^t \left(\frac{\partial}{\partial B}\, \eta(B)\right) > 0$$

a slight extension of Venter's Theorem 3 [39] can be used to prove convergence with probability one to B^*, the vector that uniquely maximizes $\eta(B)$. The unimodality of $\eta(B)$ will be examined further in Section 2.6.

2.4.2 Stochastic Solution of Maximum Likelihood-Related Equations. In this section an algorithm utilizing the Robbins-Munro procedure [40] obtains the solution to

$$\frac{\partial}{\partial B}\int\left\{\ln[h(x|B)] - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right]\right\}h(x)dx = 0 \qquad (2.30)$$

for $f(x|\alpha^i)$ multivariate Gaussian with mean vector γ^i, covariance matrix Φ^i. It is assumed that $h(x)$ is bounded and has compact support [34]. Under these conditions, the integral and partial differentation operations may be interchanged in (2.30),

$$\int\frac{\partial}{\partial B}\left\{\ln h(x|B) - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right]\right\}h(x)dx = 0 \qquad (2.31)$$

Equation (2.31) is a maximum likelihood-related equation having the constraint that $\sum P(\alpha^i) = 1$.

Define the likelihood

$$L^i(x) = \frac{f(x|\alpha^i)}{\sum_{i=1}^{M'} f(x|\alpha^i)P(\alpha^i)} \qquad (2.32)$$

Evaluating the partials in (2.31) with respect to the mixing parameters,

$$\frac{\partial}{\partial P(\alpha)^k}\left\{\ln h(x|B) - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right]\right\} = L^k(x) - \Psi$$

$$k = 1,2,\ldots,M' \qquad (2.33)$$

To obtain the value of the Lagrange multiplier Ψ, multiply (2.33)

for each k by the corresponding mixing parameter $P(\alpha^k)$ and sum,

$$\sum_{k=1}^{M'} L^k(x)P(\alpha^k) - \Psi P(\alpha^k) = 0 \qquad (2.34)$$

Using the definition of $L^k(x)$ from (2.32), then $\Psi = 1$. Evaluating the other partial derivatives in (2.31),

$$\frac{\partial}{\partial \gamma^k}\left\{\ln h(x|B) - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right]\right\}$$

$$= P(\alpha^k)L^k(x)(\phi^k)^{-1}(x - \gamma^k) = 0$$

$$\frac{\partial}{\partial(\phi^k)^{-1}}\left\{\ln h(x|B) - \Psi\left[\sum_{i=1}^{M'} P(\alpha^i) - 1\right]\right\}$$

$$= \frac{1}{2} P(\alpha^k)L^k(x)(\phi^k - (x - \gamma^k)(x - \gamma^k)^t) = 0 \qquad (2.35)$$

where 0 is the appropriate ℓ-dimensional vector or $\ell x \ell$ matrix of zeros. Removing the non zero factors in (2.35) which are not functions of x, then together with (2.33) we obtain a set of stationary equations which are equivalent to (2.31)

$$\int [L^k(x) - 1]h(x)dx = 0$$

$$\int (x - \gamma^k)L^k(x)h(x)dx = 0$$

$$\int [(x - \gamma^k)(x - \gamma^k)^t - \phi^k]L^k(x)h(x)dx = 0 \qquad (2.36)$$

Given a sequence of conditionally independent samples $X_1, X_2, \ldots$, a modification of the Robbins-Munroe procedure for solving (2.36) yields

the following set of estimators[+] to be updated simultaneously

$$P_{n+1}(\alpha^k) = [P_n(\alpha^k) + \alpha_{n+1}[L_n^k(X_{n+1}) - 1]]/\left[1 + \alpha_{n+1}\sum_{j=1}^{M'}[L_n^j(X_{n+1}) - 1]\right]$$

$$\gamma_{n+1}^k = \gamma_n^k + \alpha_{n+1}[X_n - \gamma_n^k]L_n^k(X_{n+1})$$

$$\phi_{n+1}^k = \phi_n^k + \alpha_{n+1}[(X_n - \gamma_n^k)(X_n - \gamma_n^k)^t - \phi_n^k]L_n^k(X_{n+1})$$

$$k = 1,2,\ldots,M' \qquad (2.37)$$

if $B_{n+1} \in \mathcal{B}^{M'}$; otherwise, $P_{n+1}(\alpha^k) = P_n(\alpha^k)$, $\gamma_{n+1}^k = \gamma_n^k$, and $\phi_{n+1}^k = \phi_n^k$.

Again, if $\eta(B)$ is unimodal on $\mathcal{B}^{M'}$, it may be possible to use results in [39] to prove convergence with probability one and in mean square. The approach leading to the recursive algorithm of (2.37) can be contrasted with approaches that end up effectively assuming that n samples are stored and the maximum likelihood estimator for these n samples is to be found. For example, in [13] implicit equations involving $X_1, X_2, \ldots, X_n$ are presented for the M′ class case; however, the constraint that $\sum P(\alpha^i) = 1$ is not imposed and when the constraint is violated, the resulting estimators do not converge. The two class case presented in [13] does include this constraint and an iterative method to obtain the solutions is presented in [16]. For the M′ class case with the constraint that $\sum P(\alpha^i) = 1$, Wolfe [17] found simplified implicit stationary equations for the maximum likelihood estimator. He also presented

---

[+]Slightly more complex estimators are obtained by replacing the terms multiplying α_n on the r.h.s. of (2.37) with the respective derivatives in (2.33) and (2.35).

an algorithm based on the Newton-Raphson technique, and an iterative algorithm for finding a solution. The algorithms, of course, are not recursive requiring the entire data set to be repeatedly reprocessed. In [64], Sakrison defined a recursive algorithm form for finding the zeros of the derivative regression surface in a nonmixture problem (i.e. $\frac{d}{d\alpha} E[\ln f(x|\alpha)]$). A discussion is presented in Subsection 2.6.3 comparing assumptions such as Sakrison's that $\{f(x|\alpha):\alpha\in G\}$ contain the true statistics with the minimum risk approach used here.

In comparing the algorithm of (2.37) with the hillclimbing algorithm of the last section, it might be noted that users of stochastic approximation algorithms generally have observed better performance of algorithms based on the Robbins-Munro procedure than those based on the Keifer-Wolfowitz procedure. Intuitively, this is because estimation of a noisy slope is far more difficult than estimation of the value of a noisy surface. Also, for the R-M procedure, if a priori bounds on the magnitude of the derivative regression surface can be obtained, Dvoretzky [42] has found weighting sequences that can be used to accelerate convergence over that for $\frac{1}{n}$ weighting.

The assumption that allowed the simplification of (2.35) and lead to the recursive algorithms of (2.37)--that the $P(\alpha^i)$ $i = 1,2,\ldots,M'$ are nonzero--effectively requires that all values of M up to M' be considered as hypotheses. This can be partially accomplished by parallel processing the samples with algorithms from (2.37) having different hypothesized values of M. Unfortunately, since $\eta(B)$ is never actually calculated, it is difficult to establish which hypotheses value of M is correct without resorting to a rule.

It should be emphasized that given an unlimited number of samples, the number M of classes can be estimated by the number of non zero estimated mixing parameters, assuming M' is larger than or equal to M. The parallel processing discussed above is an attempt to experimentally compensate for a limited number of observations.

2.4.3. Nonlinear Regression on $\{\hat{\eta}(B_t)\}_{t=1}^{V}$. This section presents an attempt to utilize knowledge that the class densities belong to a known parametric family to interpolate between points of the posterior density on a finite vector set $\{B^k\}_{k=1}^{V}$. Thus, the samples have the mixture density $h(x|B^*)$. Define the regression surface

$$\eta(B;B^*) = \int \ln [h(x|B)]h(x|B^*)dx$$

and on $\{B^k\}_{k=1}^{V}$ construct the estimates

$$\hat{\eta}(B^k)_n = \frac{n-1}{n}\,\hat{\eta}(B^k)_{n-1} + \frac{1}{n}\ln[h(X_n|B^k)] \quad k = 1,2,\ldots,V$$

The approach is then to least squares fit the $\{\hat{\eta}(B^k)_n\}_{k=1}^{V}$ with a regression surface $\eta(B;\hat{B})$ using nonlinear regression. The resulting vector $\hat{B}_n$ is then defined as the n^{th} estimate of the vector that maximizes $\eta(B)$.

Inherent in the use of the method is the requirement that $\eta(B;\hat{B})$ be a convenient function of B and $\hat{B}$. An attempt is made in Appendix B to evaluate $\eta(B;\hat{B})$ for a mixture of two, one dimensional Gaussian probability densities. The difficulties encountered in evaluating such an extremely simple example

suggest that the potential of this approach for practical problems is limited.

2.5 Asymptotic Minimum Risk Solutions Under a Separable Assumption

For an asymptotic minimum risk solution it is necessary to find the parameter vector $B \in \mathcal{B}^{M'}$ with M^* non zero mixing parameters $1 \leq M^* \leq M'$ that maximizes

$$\eta(B) = \int \ln h(x|B)h(x)dx$$

$$= \int \ln\left[\sum_{k=1}^{M^*} f(x|\alpha^k)P(\alpha^k)\right]h(x)dx \qquad (2.38)$$

In this section the sample space is partitioned into M^* disjoint regions where the regions are defined

$$S^k \triangleq \{x : f(x|\alpha^k)P(\alpha^k) > f(x|\alpha^j)P(\alpha^j) \quad \text{all } j \neq k\}$$
$$k = 1,2,\ldots,M^* \qquad (2.39)$$

given a parameter vector B. It is assumed that over each partitioned set the class density is Gaussian having mean vector γ^k, covariance matrix Φ^k, with the density truncated at the partition boundary[+]. The true mixture h(x) is assumed bounded.

Under these assumptions $\eta(B)$ in (2.38) can be expanded,

[+]Separability results from this truncation assumption.

$$\eta(B) = \sum_{k=1}^{M^*} \int_{S^k} \ln \left[f(x|\alpha^k) P(\alpha^k) \right] h(x)dx$$

$$= \sum_{k=1}^{M^*} \left[\int_{S^k} h(x)dx \right] \left\{ \ln P(\alpha^k) + \ln \left[\frac{1}{(2\pi)^{\ell/2} |\phi^k|^{1/2}} \right] \right.$$

$$\left. - 1/2 \frac{\int_{S^k} (x - \gamma^k)^t (\phi^k)^{-1} (x - \gamma^k) h(x)dx}{\int_{S^k} h(x)dx} \right\} \quad (2.40)$$

2.5.1. General Case Maximization of Separable $\eta(B)$. Ignoring for now the definition of the $\{S^k\}_{k=1}^{M^*}$ regions in (2.39), take a fixed set of regions which are independent of the B vector components. It can be shown by taking partial derivatives with respect to each parameter under the constraint $\sum_{k=1}^{M^*} P(\alpha^k) = 1$ that for this fixed partition, (2.40) is maximized if the parameters are defined,

$$P(\alpha^k) = \int_{S^k} h(x)dx$$

$$\gamma^k = \int_{S^k} x\, h(x)dx / \int_{S^k} h(x)dx$$

$$\phi^k = \int_{S^k} (x - \gamma^k)(x - \gamma^k)^t\, h(x)dx / \int_{S^k} h(x)dx$$

$$k = 1,2,\ldots,M^* \quad (2.41)$$

Since these definitions maximize (2.40) for any partition into M* regions, they maximize $\eta(B)$ for the regions satisfying (2.39). If the parameters are defined as in (2.41), then maximizing $\eta(B)$ is equivalent to finding the partition and the value of M* which maximizes

$$\eta(B) = \sum_{k=1}^{M^*} P(\alpha^k) \ln\left[\frac{P(\alpha^k)}{(2\pi)^{\ell/2}|\Phi^k|^{1/2}}\right] - \frac{\ell}{2} \qquad (2.42)$$

This equation illustrates the advantage of the separable approach. Since $\eta(B)$ can be expressed explicitly as a function of the parameter vector B, unsupervised estimation algorithms can be defined (as is done in Section 2.5.2 below) the ability to increase or decrease M* according to the current estimate of $\eta(B)$. Equation (2.42) for $\eta(B)$ can be used to determine the quality of M* or the significance of relatively small mixing parameter estimates. There is a cost associated with the separable Gaussian assumption--the inability to resolve mixtures having low signal to noise ratios except in some cases where M is known. One aspect of the resolution problem for M unknown is investigated in 2.6.2.

If the complexity reducing assumption that $\Phi^k = (\sigma^k)^2 I$ is made, it can be shown by taking partial derivatives as discussed previously, that $(\sigma^k)^2$ should be defined

$$(\sigma^k)^2 = \frac{\int_{S^k} \|x - \gamma^k\|^2 h(x)dx}{\int_{S^k} h(x)dx}, \quad k = 1,2,\ldots,M^* \qquad (2.43)$$

The "metric" of (2.42) remains the same.

This subsection has been concerned with determining a criteron for evaluating the quality of estimated parameters and the corresponding partition (defined by (2.39)). The following subsection presents an approach for maximizing $\eta(B)$ using estimators corresponding to (2.41).

2.5.2. An Unsupervised Estimation Algorithm for Maximizing $\eta(B)$ Under a Separable Gaussian Assumption. This unsupervised estimation algorithm is designed to maximize the $\eta(B)$ criterion (2.42) given an upper bound M' on the number of classes. Since the algorithm operates without knowledge of the number of classes M^* which maximizes (2.42), it is possible that M' is not large enough. In contrast to the previous algorithms, this algorithm will provide an indication when M' should be increased so that the maximum of $\eta(B)$ can be reached. The algorithm utilizes $\eta(B)$ to determine whether $\eta(B)$ is maximized by assigning a sample to one of M' classes, or maximized by combining two classes and defining the sample as a new class. No convergence proof is presented.

A distinction is maintained in the algorithm between two types of statistical classes: isolated points and clusters. A cluster is a collection of samples, while an isolated point is a single

sample which potentially generates a new cluster. The number of clusters which the algorithm determines as maximizing $\eta(B)$ is M*. The M'-M* isolated points have no effect on $\eta(B)$. If the number of isolated points becomes too small, M' can be increased to obtain more isolated points provided there is sufficient storage available. An isolated point at x is defined to have the statistics

1) zero probability mass
2) mean vector $\gamma = x$
3) correlation matrix $C = XX^t$

and the convention used when calculating $\eta(B)$ is that $[0 \ln(\frac{0}{0})] \triangleq 0$. The k^{th} cluster has the statistics

1) $n^k \triangleq$ total number of samples assigned to this class and $P(\alpha^k) = \frac{n^k}{n}$
2) mean vector γ^k
3) correlation matrix C^k

The following rules define the manner in which the statistics of two combined classes are updated. It is assumed that classes r and s are combined into class j.

Case A. Two Clusters

1) $n^j = n^r + n^s$
2) $\gamma^j = \frac{n^r \gamma^n + \alpha^s \gamma^s}{n^r + n^s}$
3) $C^j = \frac{n^r C^r + n^s C^s}{n^r + n^s}$

Case B. A cluster (class r) and an isolated point (class s)

1) $n^j = n^r + 1$

2) $\gamma^j = \frac{n^r \gamma^r + \gamma^s}{n^r + 1}$

3) $C^j = \frac{n^r C^r + C^s}{n^r + 1}$

Case C. Two isolated points.

1) $n^j = 2$

2) $\gamma^j = \frac{\gamma^r + \gamma^s}{2}$

3) $C^j = \frac{C^r + C^s}{2}$

The rule for updating the statistics of the r^{th} class with a sample are Cases D and E.

Case D. A cluster (class r) and the n^{th} sample.

1) $n^r = n^r + 1$

2) $\gamma^r = \frac{n^r \gamma^r + X_n}{n^r + 1}$

3) $C^r = \frac{n^r C^r + X_n X_n^t}{n^r + 1}$

Case E. An isolated point (class r) and the n^{th} sample.

1) $n^r = 2$

2) $\gamma^r = \frac{\gamma^r + X_n}{2}$

3) $C^r = \frac{C^r + X_n X_n^t}{2}$

Figure 2 contains a flow chart of the algorithm. The algorithm is started by using the first M′ samples as M′ isolated points. Then, given sample X_n, there are two[+] types

---

[+]A third possible action is to make X_n a new class (isolated point) and thus increase M′ by 1.

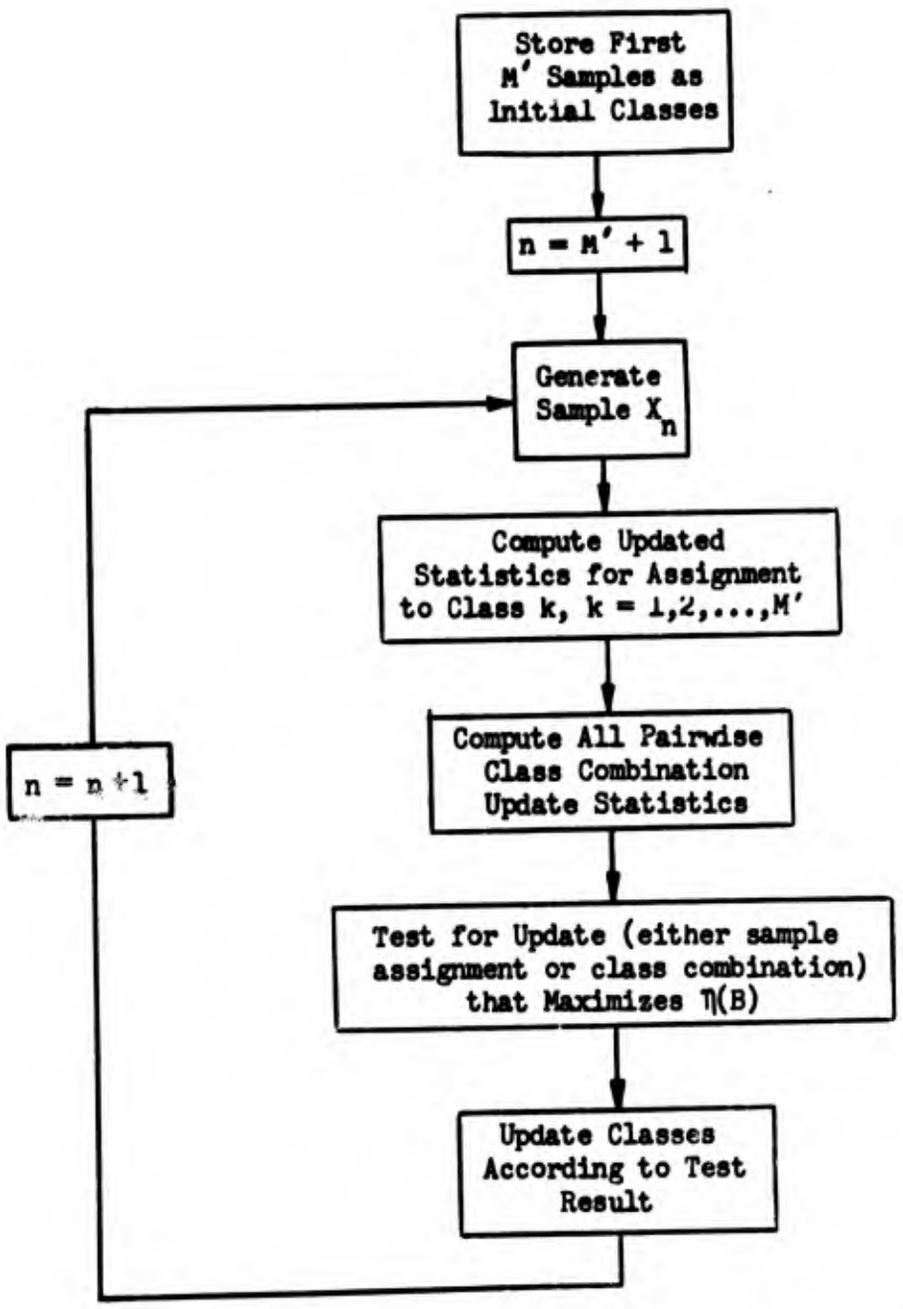


Figure 2. Flow Chart of Algorithm to Maximize $\eta(B)$ Under a Separable Gaussian Assumption.

of possible actions:

1) Assign X_n to one of the M′ classes (M′ possible actions).

2) Combine two of the M′ classes and make X_n a new class (isolated point)($\binom{M'}{2}$ possible actions).

The $\eta(B)$ criterion of (2.42) is calculated for each of the $M' + \binom{M'}{2}$ possible actions. The update corresponding to the action which maximizes $\eta(B)$ is then performed. Effectively, the favorable action produces the largest estimated $\eta(B)$ by accepting classes having large estimated values of $P(\alpha^k)\ln[P(\alpha^k)/|\Phi^k|^{\frac{1}{2}}]$.

This procedure is somewhat similar to such clustering techniques as adaptive sample set construction [48] and Isodata [49]. However, these other algorithms do not utilize a criterion such as $\eta(B)$ derived from an optimum approach.

2.5.3. Special Case: $\Phi^k = (\sigma)^2 I$ and $P(\alpha^k) = \frac{1}{M}$ Where M is Known. Under these conditions the $\eta(B)$ expression in (2.40) becomes

$$\eta(B) = \ln\left[\frac{1}{M(2\pi)^{\ell/2}(\sigma)^{\ell}}\right] - \frac{1}{2(\sigma)^2}\sum_{k=1}^{M}\int_{S^k}\|x - \gamma^k\|^2 h(x)dx \qquad (2.44)$$

and maximizing $\eta(B)$ is equivalent to finding

$$\min_{\{\gamma^k\}_{k=1}^{M}} \sum_{k=1}^{M}\int_{S^k}\|x - \gamma^k\|^2 h(x)dx \qquad (2.45)$$

This criterion, or the slight generalization for $P(\alpha^k)$ not identical,

is asymptotically minimized by the class of decision directed algorithms defined in the next chapter. As will be seen, it is extremely easy to implement an algorithm to asymptotically minimize risk under the a priori assumptions that lead to (2.45). One disadvantage of the criterion of (2.45) is that it cannot be used to determine M if this knowledge is not available. This drawback occurs because (2.45) does not incorporate a cost for adding additional classes as does the criterion of (2.44). Parallel processing as suggested by MacQueen [18] for $M^* = 1,2,\ldots,M'$ can produce more classes than there really are. For example, suppose h(x) is composed of a single one dimensional Gaussian density function with mean zero, variance $(\sigma)^2$. If it is assumed $M^* = 2$, the solution of (2.45) is a partition through $x = 0$. Denoting the variances on either side of this solution partition by $(\sigma^i)^2$, $i = 1,2$ the strict inequality $(\sigma^1)^2 + (\sigma^2)^2 < (\sigma)^2$ holds. This shows that even though there was only one class, (2.45) is less for $M^* = 2$ then for $M^* = 1$ and the use of (2.45) to determine M failed. For many applications, knowledge of M is not an unreasonable assumption, and the criterion of (2.45) motivates one of the simplest unsupervised estimation algorithms currently in existence. This class of algorithms will be discussed in the next chapter.

2.6 Topological Properties of the $\eta(B)$ Measure of Information

Because all of the algorithms discussed in this chapter deal with the measure of information $\eta(B)$ under various assumptions, it is of interest to establish some properties of the regression surface defined by $\eta(B)$. While the subject is not examined extensively, in this section contour

plots of constant $\eta(B)$ for a Gaussian mixture are presented, and a mixture resolution limit under the separable Gaussian mixture assumption is determined.

2.6.1. $\eta(B)$ Contour Plots for a Gaussian Mixture. Suppose the samples are from a mixture of two, one dimensional Gaussian distributed random variables. Assume that the density family, the equilikely mixing parameters, and the common variance $(\sigma)^2 = 1$ are known, and that the true mean values γ^{1_0} and γ^{2_0} are unknown. From (2.6) we want to plot contours of constant $\eta(B)$ where $\eta(B) = \eta(\gamma^1,\gamma^2)$ is defined

$$\eta(\gamma^1,\gamma^2) = \int_{-\infty}^{\infty} \left[\ln \sum_{k=1}^{2} 1/2\ f(x|\gamma^k,(\sigma)^2)\right] \cdot \sum_{j=1}^{2} 1/2\ f(x|\gamma^{j_0},(\sigma)^2)dx \qquad (2.46)$$

In the above equation, $f(x|\gamma,(\sigma)^2)$ is a one dimensional Gaussian density with mean γ, variance $(\sigma)^2$. This equation was evaluated numerically since, as noted in Section 2.4.3, explicit evaluation of (2.46) is unattainable at this time. The contours were found for parameter values $\gamma^{1_0} = 5$, $\gamma^{2_0} = 2$ at signal to noise ratios $(\text{SNR} \triangleq (\frac{\gamma^{1_0} - \gamma^{2_0}}{\sigma})^2)$ of 4.0 and 9.0, and plotted in Figures 3 and 4 respectively. At SNR = 4.0, the mixture density is unimodal, while at SNR = 9.0 the class densities are beginning to become separable (although there is still considerable overlap).

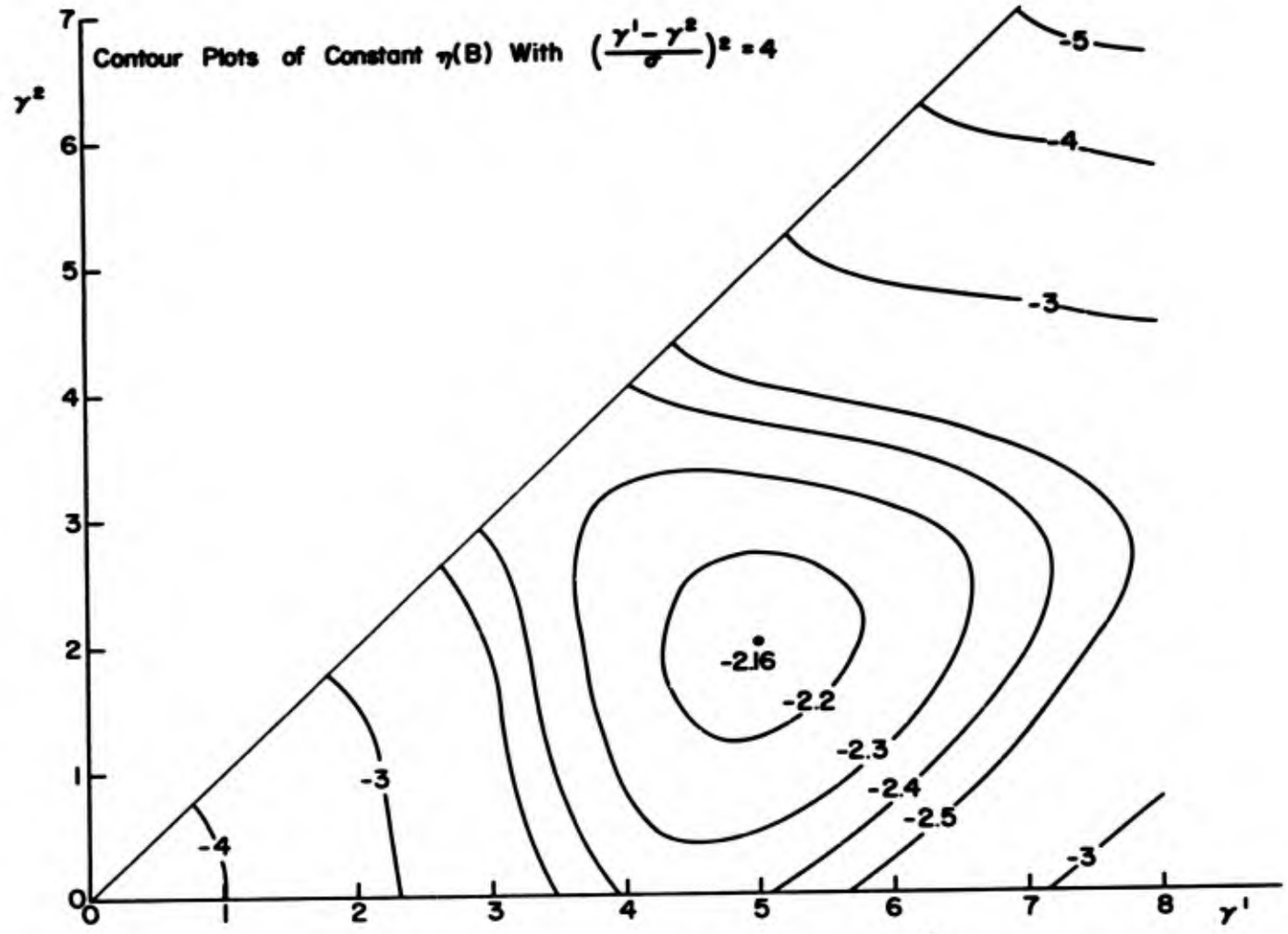


Figure 3. Contour Plot of Constant η(B) for $\gamma^1_0 = 5.0$, $\gamma^2_0 = 2.0$, and SNR = 4.0.

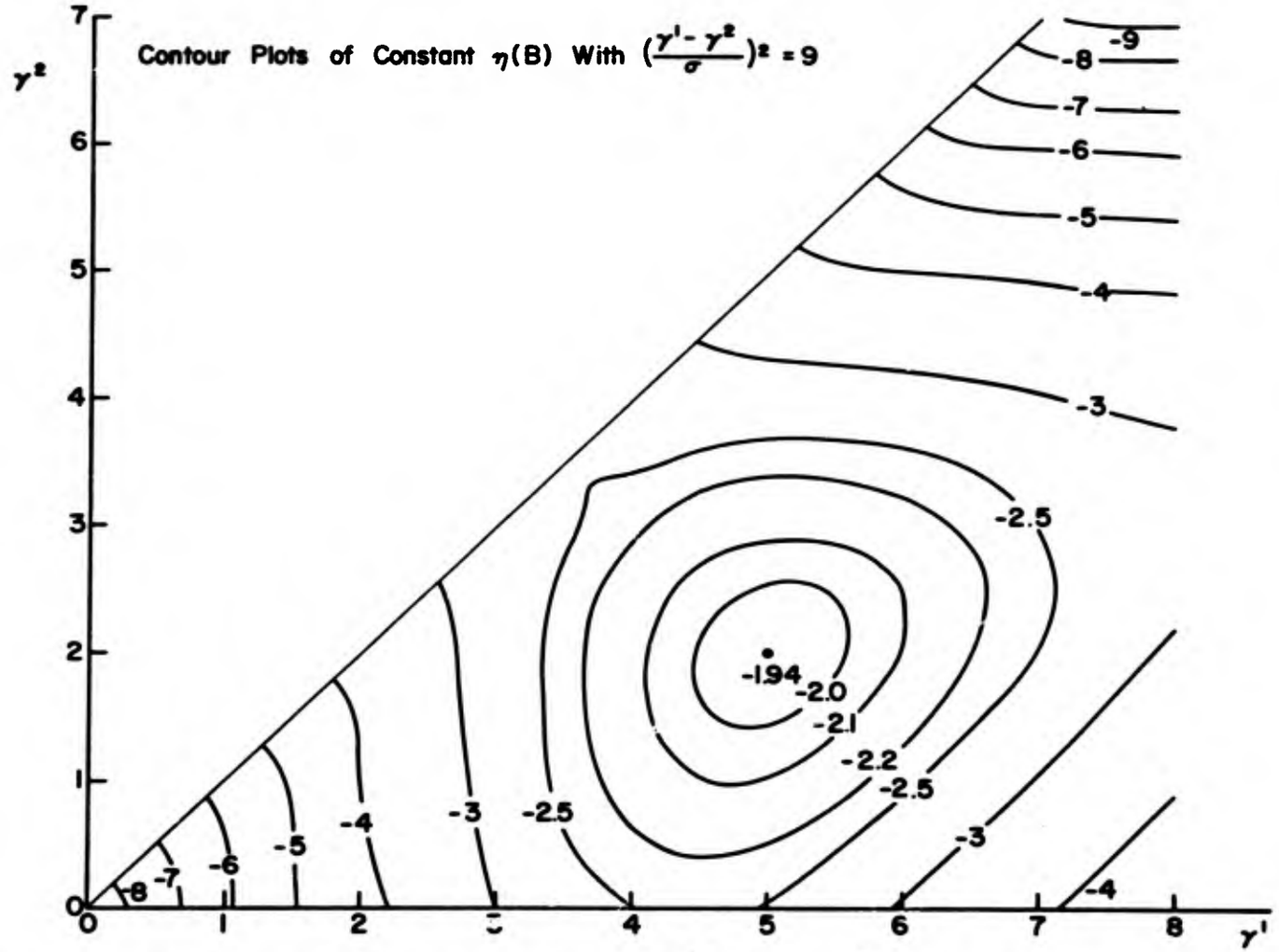


Figure 4. Contour Plot of Constant $\eta(B)$ for $\gamma^{1}_{0} = 5.0$, $\gamma^{2}_{0} = 2.0$, and SNR = 9.0.

From the complexity of the contour shapes, it is not surprising that the attempt to explicitly evaluate (2.46) was unsuccessful. Also, the contours in Figures 3 and 4 do not exhibit marginal density function properties that are characteristic of the case where the parameters were conditionally independent (i.e. if $p(\{\alpha^i, P(\alpha^i)\}_{i=1}^{M'} | Y_n) = \prod_{i=1}^{M'} p(\alpha^i|Y_n)p(P(\alpha^i)|Y_n)$). This shows the fallacy of the assumption (Fralick [8]) that the parameters are conditionally independent for the mixture case.

Results on the asymptotic posterior density and maximum likelihood estimator can be used to give additional information about the regression surface $\eta(B)$. LeCam [43] found conditions which were applied in [37] to the mixture case under which the posterior density was shown to be asymptotically normal with mean vector the maximum likelihood estimator and inverse covariance matrix the Fischer information matrix multiplied by n. Also, under these conditions the maximum likelihood estimator converges with probability one. The $(i,j)^{th}$ term in the Fischer information matrix $C(B^*)$ is defined,

$$c^{ij}(B^*) = \int \left\{ \frac{\partial}{\partial \theta^i} \frac{\partial}{\partial \theta^j} \ln h(x|B) \right\} h(x)dx \Big|_{B = B^*} \qquad (2.47)$$

where θ^i denotes the i^{th} component of B. Expanding the posterior density,

$$\frac{\prod_{k=1}^{n} h(X_k|B)p_o(B)}{\int_{\mathcal{B}^{M'}} \prod_{k=1}^{n} h(X_k|B)p_o(B)dB}$$

$$= \exp\left\{ \sum_{k=1}^{n} \ln h(X_k|B) + \ln p_o(B) - \ln[\text{norm}] \right\} \tag{2.48}$$

then under LeCam's conditions,

$$p(B|Y_n) \rightarrow \frac{|C(B^*)|^{1/2} n^{1/2}}{(2\pi)^{\ell/2}}$$

$$\cdot \exp\left\{ -\frac{n}{2}\left(B - B_{\max_n}\right)^t C(B^*)\left(B - B_{\max_n}\right) \right\} \tag{2.49}$$

as n tends to infinity where $B_{\max_n}$ is the maximum likelihood estimator given $Y_n = \{X_1, X_2, \ldots, X_n\}$. Since $B_{\max_n}$ converges with probability one to B*, equations (2.48) and (2.49) imply that for B arbitrarily close to B*,

$$P\left[\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \ln h(X_k|B) \right.$$

$$\left. = -\frac{1}{2}(B - B^*)^t C(B^*)(B - B^*) + c \right] = 1 \tag{2.50}$$

where c comes from the normalization term and is finite. Applying the definition of $\eta(B)$ from (2.6), for B in an arbitrarily small neighborhood of B*,

$$\eta(B) \overset{\Delta}{=} -1/2(B - B^*)^t C(B^*)(B - B^*) + c \tag{2.51}$$

This result combined with the definition of the matrix $C(B^*)$ shows that at higher signal to noise ratios where the densities are

essentially separable, the contours near the solution tend to ellipses with axes parallel to the coordinate axes. This information about the quadratic contours "near" the solution could possibly be used to accelerate the convergence of the likelihood weighted estimators of Subsection 2.4.2.

2.6.2. A Mixture Resolution Limit Resulting from the Separable Gaussian Assumption. The criterion of (2.42) was derived under a separability assumption. The following is an investigation of the limiting case for which the criterion can resolve a mixture of two classes when the separable Gaussian assumption is violated.

Suppose the true mixture density $h(x)$ is the sum of two one dimensional Gaussian densities $f(x|(-1)^k \gamma, (\sigma)^2)$, $k = 1,2$ having means at $-\gamma$ and $+\gamma$, variances $(\sigma)^2$, and mixing parameters $1/2$. Evaluating (2.42) under these assumptions for one active class $M^* = 1$,

$$\max_{M^*} \sum_{i=1}^{M^*} P(\alpha^i) \ln\left(\frac{P(\alpha^i)}{\sqrt{2\pi}(\sigma^i)}\right) - \frac{1}{2} = \ln\left(\frac{1}{\sqrt{2\pi}(\sigma^1)}\right) - \frac{1}{2} \qquad (2.52)$$

where $((\sigma^1)^2$ is the variance of x assuming there is one class when in fact there are two)

$$(\sigma^1)^2 \triangleq \int_{-\infty}^{\infty} x^2 \, h(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 \sum_{k=1}^{2} f(x|(-1)^k \gamma, (\sigma)^2)dx$$

$$= (\sigma)^2 + (\gamma)^2 \tag{2.53}$$

Substituting (2.53) into (2.52),

$$\ln\left(\frac{1}{\sqrt{2\pi}(\sigma^1)}\right) - \tfrac{1}{2} = \tfrac{1}{2}\ln\left(\frac{1}{2\pi[(\sigma)^2 + (\gamma)^2]}\right) - \tfrac{1}{2} \tag{2.54}$$

Similarly evaluating (2.42) for two active classes $M^* = 2$,

$$\max \sum_{i=1}^{M^*} P(\sigma^i)\ln\left(\frac{P(\sigma^i)}{\sqrt{2\pi}(\sigma^i)}\right) - \tfrac{1}{2} = \ln\left(\frac{1}{2\sqrt{2\pi}(\sigma^2)}\right) - \tfrac{1}{2} \tag{2.55}$$

where $((\sigma^2)^2$ is the variance of x assuming there are two classes when in fact there are two)

$$(\sigma^2)^2 \triangleq 2\int_0^\infty x^2\, h(x)dx - \left[2\int_0^\infty x\, h(x)dx\right]^2 \tag{2.56}$$

Evaluating the terms in (2.56),

$$\int_0^\infty x^2\, h(x)dx = \frac{(\sigma^1)^2}{2}$$

$$= \frac{(\sigma)^2 + (\gamma)^2}{2} \tag{2.57}$$

and letting $\Phi(\cdot)$ denote a normalized zero mean Gaussian distribution, from Cramer [44],

$$\int_0^\infty x\, h(x)dx = \tfrac{1}{2}\int_0^\infty x\, f(x|\gamma,(\sigma)^2)dx + \tfrac{1}{2}\int_0^\infty x\, f(x|-\gamma,(\sigma)^2)dx$$

$$= \frac{1}{2}\{\gamma + \lambda^1 \sigma\}(1 - \phi(\frac{-\gamma}{\sigma})) + \frac{1}{2}\{-\gamma + \lambda^2 \sigma\}(1 - \phi(\frac{\gamma}{\sigma})) \tag{2.58}$$

where

$$\lambda^1 = \frac{\phi'(\frac{-\gamma}{\sigma})}{1 - \phi(-\frac{\gamma}{\sigma})} \qquad \lambda^2 = \frac{\phi'(\frac{\gamma}{\sigma})}{1 - \phi(\frac{\gamma}{\sigma})} \tag{2.59}$$

Combining (2.58) and (2.59),

$$\int_0^\infty x\, h(x)dx = \gamma(\phi(\frac{\gamma}{\sigma}) - \frac{1}{2}) + \sigma\phi'(\frac{\gamma}{\sigma}) \tag{2.60}$$

Substituting (2.57), (2.60) into (2.56), then (2.55) becomes

$$\ln\left(\frac{1}{2\sqrt{2\pi(\sigma^2)}}\right) - \frac{1}{2}$$

$$= \frac{1}{2}\ln\left(\frac{1}{8\pi\{(\sigma)^2 + (\gamma^2) - [\gamma(2\phi(\frac{\gamma}{\sigma}) - 1) + 2\sigma\phi'(\frac{\gamma}{\sigma})]^2\}}\right) - \frac{1}{2} \tag{2.61}$$

In Figure 5, $\eta(B)$ is plotted for $M^* = 1$ (one assumed active sample class) and $M^* = 2$ (two assumed active classes) as a function of the signal to noise ratio $4(\frac{\gamma}{\sigma})^2$. The intersection of the two curves gives the dividing line between the SNR values where $M^* = 1$ maximizes $\eta(B)$ and the SNR where $M^* = 2$ maximizes $\eta(B)$. The figure shows that below a signal to noise ratio ~ 9.0 (the two class means are 3σ apart), $\eta(B)$ evaluated under a separable Gaussian mixture assumption cannot resolve two classes for the case considered in this section.

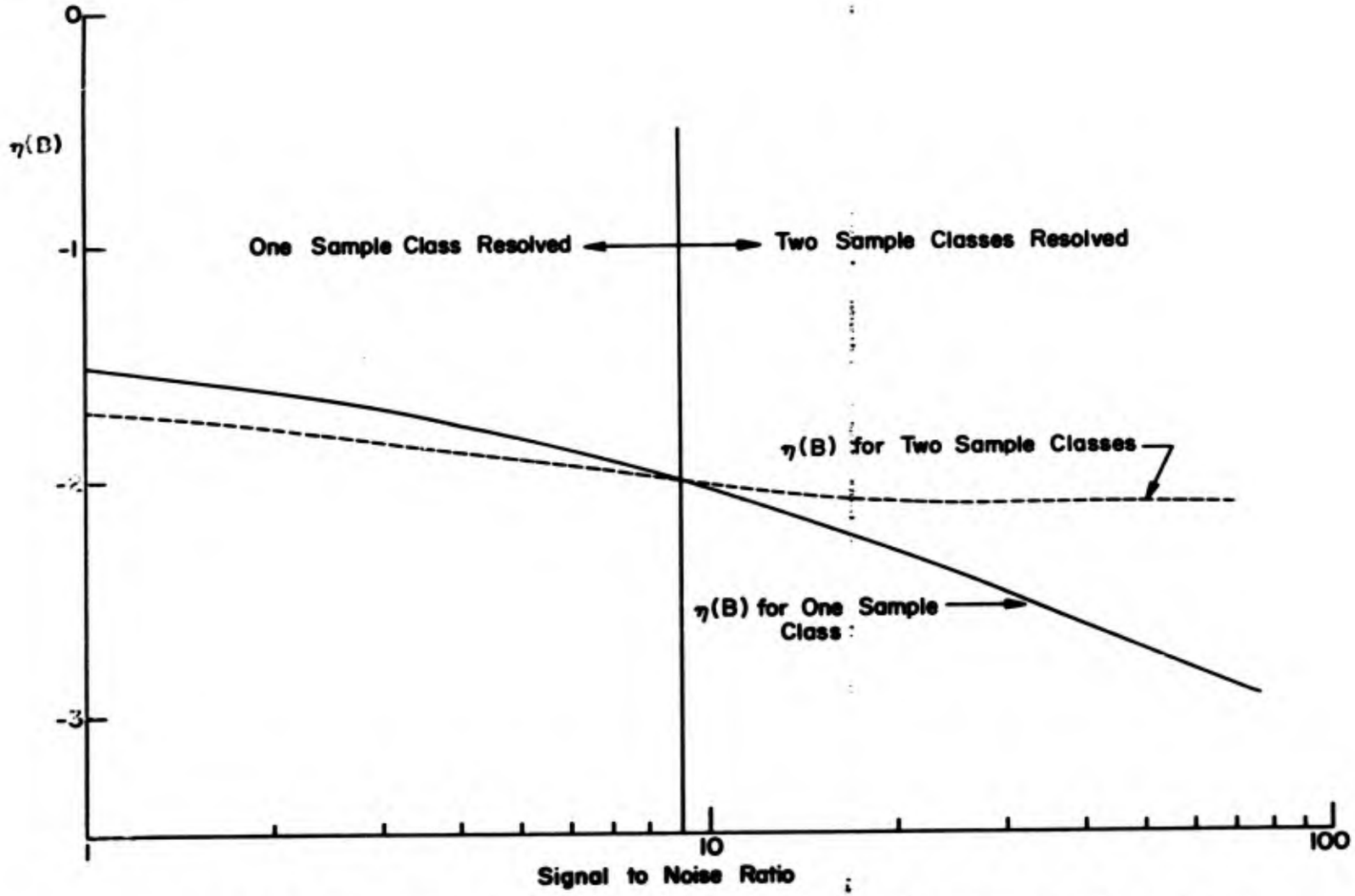


Figure 5. Resolution of Two Sample Classes Using the Separable Gaussian Assumption.

2.6.3. Comments on Classical Maximum Likelihood Results and the $\eta(B)$ Approach. Classical results on maximum likelihood estimators [50] assume that the true mixture $h(x)$ can be expressed as a linear combination of a finite number of functions from a known parametric family $\{f(x|\alpha) : \alpha \in G\}$. Estimators are constructed using this known family and if certain conditions are satisfied the estimators are proven to converge. However, in practice, the choice of parametric family is not only dictated by an intuative feeling for the statistics of the problem, but also by the convenience or simplicity of the resulting description. For example, a Gaussian assumption requires only a mean vector and a covariance matrix to completely describe each density function. The approach of this report has been to assume a parametric family (such as Gaussian), and then to determine the parameter vector that best describes an unknown $h(x)$ using the $\eta(B)$ criterion. There is obviously a close duality between the classical maximum likelihood approach and the $\eta(B)$ regression surface approaches and, in fact, if $h(x)$ is a finite mixture of functions from the assumed family, they are solutions to the same problem. Since it can be stated almost categorically that actual sample distributions never belong to a known parametric family, the $\eta(B)$ approach used here seems to be a more realistic model of the physical world.

III. A CLASS OF DECISION DIRECTED ESTIMATION ALGORITHMS

In unsupervised estimation problems the samples are from a mixture of several classes and the statistics required to classify a sample X_n are incompletely known. These unknown statistics must be "learned" from a sample sequence $\{X_k\}_{k=1}^{n-1}$ whose true classifications are unknown. The problem of concern here is illustrated in Figure 6. One of M classes ω^i, $i = 1,2,\ldots,M$ is active with probability $P(\omega^i)$ to cause the ℓ dimensional vector sample X_n. It is assumed that the unknown mixture density $h(x)$ is to be represented by a sum of separable Gaussian densities having mean vectors γ^i, mixing parameters $P(\gamma^i)$, and a common covariance matrix $(\sigma)^2 I$. Letting $f(x|\gamma^i,(\sigma)^2 I)$ denote a Gaussian density with mean γ^i, covariance matrix $(\sigma)^2 I$ and defining

$$\underline{\gamma} \triangleq (\gamma^1,\gamma^2,\ldots,\gamma^{M'})$$

$$\underline{P}(\underline{\gamma}) \triangleq (P(\gamma^1),P(\gamma^2),\ldots,P(\gamma^{M'})) \qquad (3.1)$$

then under such a separable Gaussian assumption, it was found in Section 2.5.1 that the average risk is minimized by maximizing over $\underline{\gamma}$ and $\underline{P}(\underline{\gamma})$

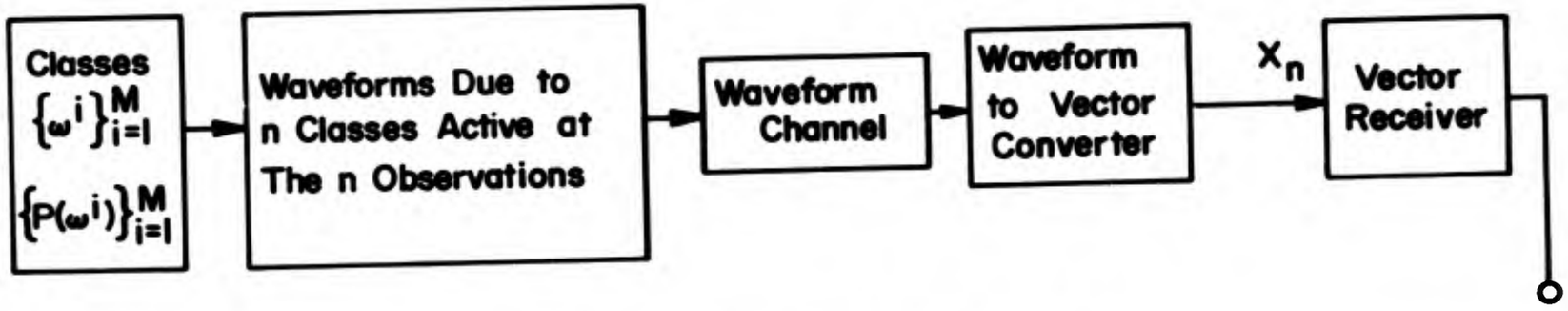


Figure 6. Problem Model for Chapter 3.

$$\eta(\underline{\gamma},\underline{P}(\underline{\gamma})) = \sum_{i=1}^{M'} \int_{S^i(\underline{\gamma},\underline{P}(\underline{\gamma}))} h(x)dx \ln P(\gamma^i)$$

$$- \frac{1}{2(\sigma)^2} \sum_{i=1}^{M'} \int_{S^i(\underline{\gamma},\underline{P}(\underline{\gamma}))} \|x - \gamma^i\|^2 h(x)dx \qquad (3.2)$$

where

$$S^i(\underline{\gamma},\underline{P}(\underline{\gamma})) \triangleq \{x : f(x|\gamma^i,(\sigma)^2 I)P(\gamma^i) > f(x|\gamma^j,(\sigma)^2 I)P(\gamma^j)$$

$$1 \le j \le M'\}$$

$$j \ne i \qquad (3.3)$$

and tied points are assigned arbitrarily to the lowest indexed set. In this chapter a class of algorithms is defined that minimize

$$\sum_{i=1}^{M'} \int_{S^i(\underline{\gamma},\underline{P}(\underline{\gamma}))} \|x - \gamma^i\|^2 h(x)dx \qquad (3.4)$$

and are thus suboptimum with respect to the assumptions unless the $P(\gamma^i)$, $i = 1,2,\ldots,M'$ are assumed equal. As discussed in Section 2.5.3, the "metric" of (3.4) cannot be used directly to determine the number of classes. Hence, it will be assumed throughout this chapter that M is known (although perhaps incorrectly). Methods of handling the mixing parameters include:

1. $P(\gamma^i)$ unknown and to be estimated.
2. $P(\gamma^i)$ unknown but assumed equal to $1/M$ or $P(\omega^i)$ known equal to $1/M, 1 \le i \le M$, implying $P(\gamma^i) = \frac{1}{M}$, $1 \le i \le M$.

3. $P(\nu^i)$ unknown but $P(\omega^i)$ known ($\neq \frac{1}{M}$) and an ordering known on the component vectors in $\underline{\nu} = (\nu^1,\ldots,\nu^M)$ for every $\underline{\nu}\in(R^\ell)^M$.

This last condition reflects the difficulty that knowledge of the class probabilities is not useful unless there is a way to associate them with correct estimates of the components in $\underline{\nu}_N = (\nu_N^1,\ldots,\nu_N^M)$ (such as an ordering on the component vectors of $\underline{\nu}$ for all $\underline{\nu}\in(R^\ell)^M$) is known.

Several previous papers have considered either the two class problem [19]--[23] or the more general M class problem [2],[18], under assumptions equivalent to a separable Gaussian assumption. In this chapter a class of estimators is presented which includes, among others, the convergent algorithms (and with minor modifications of a weighting sequence the nonconvergent algorithms) from these papers. The algorithms take a weighted sum of the samples falling into a subset of each $S^i(\underline{\nu}_N,\underline{P}_N)$ region to estimate the mean vectors. Estimation of the mixing parameters is assumed done using a function of the estimated mean vectors and other recursive statistics.

In Section 3.1 the class of M-ary algorithms is presented and related to some previous results. These estimators will be used in the next chapter on intersymbol interference. An extension of a convergence with probability one proof due to MacQueen [18] for the M-ary case is presented in 3.2. This is a generalized proof in the sense that general conditions on an estimator are assumed. It may be nontrival to show that a particular estimator satisfies the required conditions.

Most of the results of this chapter are concerned with the special case of M = 2. In Section 3.3 the specialization of the M-ary case estimator is discussed. For M = 2, considerably weaker assumptions on the mixture density h(x) than those of Section 3.2 are sufficient to prove convergence. In Section 3.4 a set of nonparametric conditions on h(x) is presented for each of the three M = 2 subclasses defined by the three methods of handling the $P(\gamma^i)$ discussed above. These conditions are shown to be sufficient to prove convergence with probability one and in mean square for all the estimators in each subclass in Propositions 3-5. Also under these conditions, all the estimators in a particular subclass converge to a common asymptotic vector point which is unique on $\mathcal{B}^{M'}$. Analytical evaluation of the M = 2 asymptotic vector point in Section 3.5 leads to a set of implicit equations which can be evaluated using a numerical technique. Curves are presented showing the algorithm subclass's increase in asymptotic probability of error over the optimum system with all parameters known for a Gaussian mixture. In Section 3.6 experimental convergence results are presented and used in Section 3.7 in a discussion of the tradeoffs involved between different members of the algorithm class.

3.1 The M-ary Class of Decision Directed Algorithms

The decision directed algorithm class is defined as an algorithmic form with several parameters unspecified. Specification of these parameters gives particular members of the algorithm class. The algorithmic form is defined by updating the current

estimates of the mean vectors γ_N^i $i = 1,2,\ldots,M$ with samples falling into subsets of the $\{S^i(\underline{\gamma}_N, \underline{P}_N)\}_{i=1}^M$ regions, denoted

$$A_N^i \subset S^i(\underline{\gamma}_N, \underline{P}_N) \qquad i = 1,2,\ldots,M \tag{3.5}$$

In this and the following section's general formulation, the method of obtaining the A_N^i subsets is not specified nor is the method of obtaining the mixing parameter estimates $P(\gamma_N^i)$, $i = 1,2,\ldots,M$. However, in Section 3.3, a particular method of obtaining the A_N^i subsets and of estimating the mixing parameters are each defined forming some algorithm subclasses of interest. Also, the case where there are fewer than M mean vectors unknown follows as a special case and will not be discussed.

3.1.1 The Algorithm Class. In the algorithms, the mean vectors $\underline{\gamma}_N = (\gamma_N^1,\ldots,\gamma_N^M)$ and the corresponding update regions $\underline{A}_N = (A_N^1,\ldots,A_N^M)$ are held fixed for K samples $\{X_n\}_{n=NK+1}^{(N+1)K}$. While the $\underline{\gamma}_N$ are being held fixed, conditional sample mean vectors ξ_k^i $k = 1,2,\ldots,K$ are calculated from the samples falling into each A_N^i region. The $\{\xi_K^i\}_{i=1}^M$ will be used to update[+] the $\underline{\gamma}_N$. Thus, for $NK + 1 \le n \le (N + 1)K$

$$\xi_k^i = (1 - \beta_k^i)\xi_{k-1}^i + \beta_k^i X_{NK+k} \tag{3.6}$$

where

[+]This introduces a "smoothing effect" for $K > 1$.

$$\beta_k^i = \begin{cases} \frac{1}{K^i} : X_{NK+k} \in A_N^i \\ \text{and } K^i = K^i + 1 \qquad i = 1,2,\ldots,M \\ 0 : \text{otherwise} \end{cases} \tag{3.7}$$

and $\{\xi_o^i\}_{i=1}^M \overset{\Delta}{=} 0$. After $X_{(N+1)K}$ has been classified and $\{\xi_K^i\}_{i=1}^M$ calculated, the $\{\gamma_N^i\}_{i=1}^M$ and the $\{S^i(\underline{\gamma}_N, \underline{P}_N)\}_{i=1}^M$ (along with the corresponding A_N^i update regions) are updated using the ξ_K^i, $i = 1,2,\ldots,M$ and a weighting sequence $\{\alpha_r\}_{r=1}^\infty$. Hence, at $\underline{n = (N+1)K}$

$$\gamma_{N+1}^i = (1 - \alpha_{N+1}^i)\gamma_N^i + \alpha_{N+1}^i \xi_K^i \qquad i = 1,2,\ldots,M \tag{3.8}$$

where

$$\alpha_{N+1}^i = \begin{cases} \alpha_{w_N^i} : K^i > 0 \\ \text{and } w_{N+1}^i = w_N^i + 1 \qquad i = 1,2,\ldots,M \\ 0 : K^i = 0 \\ \text{and } w_{N+1}^i = w_N^i \end{cases} \tag{3.9}$$

and if $K^i \equiv 0$ for all $i = 1,2,\ldots,M$, $w_{N+1}^i = w_N^i + 1$, $i = 1,2,\ldots,M$. After the updating is completed, the K^i, $i = 1,2,\ldots,M$ are reinitialized to zero. The selection of the initial vectors γ_1^i and counts w_1^i, while important to the algorithms' performance, is arbitrary[+]. If $\{\alpha_r\}_{r=1}^\infty$ is a non-negative sequence satisfying

---

[+]Of course, w_1^i is an integer greater than zero, $i = 1,2,\ldots,M$.

$$\sum_{r=1}^{\infty} \alpha_r = \infty \qquad \sum_{r=1}^{\infty} (\alpha_r)^2 < \infty \tag{3.10}$$

Then it may be possible to establish convergence of the algorithms.

The update sets $\{A_N^i\}_{i=1}^M$ have important functions for $P(\gamma_N^i) \neq \frac{1}{M}$ any $i = 1,2,\ldots,M$. If the mixing parameters are being estimated and A_N^i is replaced by $S_N^i \triangleq S^i(\underline{\gamma}_N, \underline{P}_N)$ in (3.9), it is possible that $\gamma_{N+1}^i \notin S_{N+1}^i$. Since this leads to undesirable trap states[+], the A_N^i subsets can be used to allow only those updates with $\gamma_{N+1}^i \in S_{N+1}^i$ $i = 1,2,\ldots,M$. Similarly, if an ordering on the components of $\underline{\gamma} \in (R^{\ell})^M$ is known, $\underline{\gamma}_N$ is updated only if the updated vector components satisfy the required ordering relationship.

Like most minimum seeking stochastic approximation algorithms, the decision directed algorithms are sensitive to sample correlation (i.e. if the correlation is severe enough, the algorithms do not converge). By computing sample means on each partitioned region with $K > 1$, some of the convergence power of a sample mean is combined with partition updatability in an algorithm more widely applicable than the $K = 1$ estimator.

3.1.2 Relationship with Other Estimators. Equation (3.8) has the form of a stochastic approximation algorithm [42] with random weights ρ_N^i, $i = 1,2,\ldots,M$. In conventionally weighted stochastic approximation algorithms, the non-negative ρ_N^i sequences

---

[+]A trap state is defined as a state at which no further updating is possible for one or more of the $\{\gamma_N^i\}_{i=1}^M$.

essentially are non zero for n large enough. However, in the decision directed estimators, any of the weights ρ_N^i, $i = 1,2,\ldots,M$ may be zero for <u>any</u> value of N, depending on the current set of K samples.

There are two parameter sets besides M which are undefined in (3.6)-(3.10): K and the weighting sequence $\{\alpha_r\}_{r=1}^{\infty}$. We now look at some members of the algorithm class defined by (3.6)-(3.10).

<u>Case 1</u>. $K = 1$, $\alpha_r = 1/r$

For $M = 2$ this is the "classical" uniformly weighted decision directed estimator. Under assumptions of a two class Gaussian mixture and the $P(\gamma^i)$ $i = 1,2$ known along with a corresponding ordering on $(\gamma^1,\gamma^2)\in(R^\ell)^2$, the method of evaluating the asymptotic probability of error was found by Scudder [19] for one unknown mean vector. This was extended to two unknown mean vectors in [21] by Patrick and Costello. These analytical methods will be shown to be applicable for a far wider set of values for K and α_r. Both results were found under an assumption that the estimators converge. A moment estimator for $P(\gamma^1)$ (and thus for $P(\gamma^2) = 1 - P(\gamma^1)$) was presented in [20], [51] for the two unknown mean vector, unknown mixing parameter case. For the M-ary case with the $P(\gamma^i)$ $i = 1,2,\ldots,M$ set equal to $1/M$, MacQueen [18] established conditions for convergence with probability one to a vector point which is a local minimum of the "metric" of (3.4).

<u>Case 2</u>. $K = 1$, $\alpha_r = \frac{1}{r+c}$

In a recent paper Gregg and Hancock [22] attempted to "optimize" the α_r weighting coefficients in a different

(but equivalent) estimator form for the two class case. The criterion of optimality was the minimization of a functional containing the parameter mean square error, the error convergence rate, and the average estimate dispersion. Their result was dependent on the unknown parameters; however, by taking the low signal to noise ratio limit of the weights and applying a normalization, the weights reduced to a function of an unknown Lagrange multiplier. Under these conditions, the "optimum" weights were found to be $\alpha_r = \frac{1}{r+c}$ where c is an unknown constant, to be determined by experiment. It is interesting to note that Dvoretzky [42] has established conditions for conventionally weighted stochastic approximation algorithms under which the sequence $\frac{1}{r+c}$ minimizes the parameter mean square error with the value of c an explicit function of the conditions. The use of this result is wide spread ([45], [46], for example). The appearance of this minimizing weight sequence in "optimized" decision directed estimators is thus not surprising.

Case 3. $1 < K < \infty, \alpha_r = 1$

While the algorithms defined by these parameters are not in the class defined by (3.6)-(3.10)((3.10) is not satisfied), they require only a minor modification of the α_r weighting sequence. These algorithms are the so called "batch processing" algorithms [2] or "tracking mode" algorithms [23] and are the easiest $K > 1$ algorithms to implement. Although in general there is no mean square convergence property for algorithms whose weighting sequence do not satisfy (3.10), for reasonably

large K relative to the inverse signal to noise ratio, computer simulation results presented in 3.6 have moderately worse small sample performance than Case 1 algorithms for multidimensional samples. Algorithms of this type are particularly useful for tracking slowly varying class statistics where "convergence" has vague meaning anyway.

3.2 Generalized Convergence for the M-Ary Algorithm Class

The mixture statistical structure is in general not known exactly. Hence, it is of interest to determine broad conditions on the mixture density h(x) under which an algorithm converges to a solution, and to define what is meant by a "solution". The following theorem is an extension of MacQueen's result [18] to cover the M-ary Algorithm Class of (3.6)-(3.10) and follows his proof closely. We first list conditions used in the theorem.

(1) $h(X_n|X_1,X_2,\ldots,X_{n-1}) = h(X_n)$

(2) h(x) is absolutely continuous with respect to Lesbegue measure.

(3) There exists a closed and bounded convex set $\chi \subset R^{\ell}$ such that $\int_{\chi} h(x)dx = 1$ and $\int_{c\chi} h(x)dx = 0$.

(4) For any open set $A \in \chi$, $\int_A h(x)dx > 0$

(5) $P[\lim_{n \to \infty} \frac{p(A_N^i)}{p(S_N^i)} = 1] = 1, i = 1,2,\ldots,M$ where $p(A) \triangleq \int_A h(x)dx$.

Condition (5) is the only one dealing with the generation of the $\{A_N^i\}_{i=1}^M$ regions and, indirectly, the unspecified estimation of the mixing parameters. This condition is very general and may

be nontrival to prove for a specific algorithm.

THEOREM 2.

If conditions (1)-(5) hold, and for $W(\underline{\gamma})$ and $V(\underline{\gamma})$ defined,

$$W(\underline{\gamma}_N) \triangleq \sum_{i=1}^{M} \int_{S_N^i} \|x - \gamma_N^i\|^2 h(x)dx \tag{3.11}$$

$$V(\underline{\gamma}_N) \triangleq \sum_{i=1}^{M} \int_{S_N^i} \|x - \mu_{S_N^i}\|^2 h(x)dx \tag{3.12}$$

where

$$\mu_A \triangleq E[x|A] \tag{3.13}$$

then the sequence of random variables $W(\underline{\gamma}_1)$, $W(\underline{\gamma}_2)$,... generated by any member of the class of estimators defined by (3.6)-(3.10) with $\gamma_1^i \in \chi$, $i = 1,2,\ldots,M$ converges with probability one. Also with probability one $W_\infty \triangleq \lim_{N \to \infty} W(\underline{\gamma}_N)$ is equal to $V(\underline{\gamma})$ for some $\underline{\gamma}$ in the set of points satisfying

$$\gamma^i = E[x|S^i(\underline{\gamma},\underline{P}(\underline{\gamma}))] \qquad i = 1,2,\ldots,M \tag{3.14}$$

having the property that $\gamma^i \neq \gamma^j$ if $i \neq j$.

Proof: Assume that the initial vectors $\{\gamma_1^i\}_{i=1}^M$ have been determined, the initial counts $\{w_1^i\}_{i=1}^M$ set to positive integers, and updating starts with sample X_2.

Since[+] S_{N+1} is the minimum distance partition relative to Y_{N+1},

$$E[W(Y_{N+1})|Y_{NK}] = E\left[\sum_{i=1}^{M} \int_{S_{N+1}^i} \|x - y_{N+1}^i\|^2 h(x)dx \middle| Y_{NK}\right]$$

$$\leq E\left[\sum_{i=1}^{M} \int_{S_N^i} \|x - y_{N+1}^i\|^2 h(x)dx \middle| Y_{NK}\right] \tag{3.15}$$

where

$$Y_{NK} \triangleq \{X_n\}_{n=1}^{NK} \tag{3.16}$$

Let r^i denote the number of samples from $\{X_n\}_{n=NK+1}^{(N+1)K}$ that fall in A_N^i and note that if $r^i = 0$, $y_{N+1}^i = y_N^i$. Then

$$E[W(Y_{N+1})|Y_{NK}] \leq \sum_{i=1}^{M} \sum_{r=0}^{K} E\left[\int_{S_N^i} \|x - y_{N+1}^i\|^2 h(x)dx \middle| r^i = r, Y_{NK}\right]$$

$$\cdot \binom{K}{r} (1 - p(A_N^i))^{K-r} p(A_N^i)^r \tag{3.17}$$

$$= \sum_{i=1}^{M} \int_{S_N^i} \|x - y_N^i\|^2 h(x)dx(1 - p(A_N^i))^K$$

[+]Inequality (3.15) is due to MacQueen [18].

$$+ \sum_{i=1}^{M} \sum_{r=1}^{K} \int\cdots\int_{A_N^i} \int_{A_N^i} \int_{S_N^i} \|x - (1 - \rho_{N+1}^i)\gamma_N^i$$

$$- \rho_{N+1}^i \frac{1}{r}\left(\sum_{j=1}^{r} y_j\right)\|^2$$

$$\cdot h(x)dx \prod_{j=1}^{r} h(y_j)dy_j \binom{K}{r}(1 - p(A_N^i))^{K-r}\, p(A_N^i)^r \tag{3.18}$$

Applying the relation $\int_A \|x - y\|^2 h(x)dx = \int_A \|x - \mu'\|^2 h(x)dx$ $+ p(A)\|y - \mu'\|^2$ where $\int_A (x - \mu')h(x)dx = 0$, two times and expanding enables us to write the second term in (3.18) as,

$$\sum_{i=1}^{M} \sum_{r=1}^{K} \int\cdots\int_{A_N^i} \left[\left(\int_{S_N^i} \|x - \mu_{S_N^i}\|^2 h(x)dx\right)\right.$$

$$+ p(S_N^i)\|(1 - \rho_{N+1}^i)\gamma_N^i$$

$$\left. - \mu_{S_N^i} + \rho_{N+1}^i \left(\frac{1}{r}\sum_{j=1}^{r} y_j\right)\|^2\right] \prod_{j=1}^{r} h(y_j)dy_j\Big\}$$

$$\cdot \binom{K}{r}(1 - p(A_N^i))^{K-r}\, p(A_N^i)^r$$

$$= \sum_{i=1}^{M} \sum_{r=1}^{K} \left\{p(A_N^i)^r \left(\int_{S_N^i} \|x - \gamma_N^i\|^2 h(x)dx\right)\right.$$

$$\left. - p(S_N^i)p(A_N^i)^r \|\gamma_N^i - \mu_{S_N^i}\|^2\right.$$

$$+ p(S_N^i)p(A_N^i)^r \left\|(1 - \rho_{N+1}^i)\left(\gamma_N^i - \mu_{S_N^i}\right)\right\|^2$$

$$+ p(S_N^i)\int_{A_N^i} \cdots \int_{A_N^i} \left\|\rho_{N+1}^i \left(\mu_{S_N^i} - \frac{1}{r}\sum_{j=1}^{r} y_j\right)\right\|^2 \prod_{j=1}^{r} h(y_j)dy_j$$

$$+ 2p(S_N^i)\rho_{N+1}^i (1 - \rho_{N+1}^i)\left(\gamma_N^i - \mu_{S_N^i}\right)^t \left(p(A_N^i)^{r-1} \int_{A_N^i} xh(x)dx\right.$$

$$\left. - p(A_N^i)^r \mu_{S_N^i}\right)\binom{K}{r}(1 - p(A_N^i))^{K-r} p(A_N^i)^r$$

$$= \sum_{i=1}^{M}\sum_{r=1}^{K} \left\{p(A_N^i)^r \left(\int_{S_N^i} \|x - \gamma_N^i\|^2 h(x)dx\right)\right.$$

$$- p(S_N^i)p(A_N^i)^r \|\gamma_N^i - \mu_{S_N^i}\|^2$$

$$+ p(S_N^i)p(A_N^i)^r \left\|(1 - \rho_{N+1}^i)\left(\gamma_N^i - \mu_{S_N^i}\right)\right\|^2$$

$$+ p(S_N^i)(\rho_{N+1}^i)^2 \frac{1}{r}\int_{A_N^i} \|\mu_{S_N^i} - x\|^2 h(x)dx$$

$$\left. + 2p(S_N^i)p(A_N^i)^r \rho_{N+1}^i (1 - \rho_{N+1}^i)\left(\gamma_N^i - \mu_{S_N^i}\right)^t \left(\mu_{A_N^i} - \mu_{S_N^i}\right)\right\}$$

$$\cdot\binom{K}{r}(1 - p(A_N^i))^{K-r} p(A_N^i)^r \qquad (3.19)$$

where the last two terms reduced because of the sample conditional independence assumption. Combining equations (3.19) and (3.18),

$$E[W(\underline{Y}_{N+1})|Y_{NK}] \le W(\underline{Y}_N) + \sum_{i=1}^{M} \left(\int_{S_N^i} \|x - Y_N^i\|^2 h(x)dx \right)$$

$$\cdot \left[-1 + (1 - p(A_N^i))^K + \sum_{r=1}^{K} \binom{K}{r} (1 - p(A_N^i))^{K-r} p(A_N^i)^{2r} \right]$$

$$- \sum_{i=1}^{M} p(S_N^i) \|Y_N^i - \mu_{S_N^i}\|^2 (2\rho_{N+1}^i - (\rho_{N+1}^i)^2)$$

$$\cdot \sum_{r=1}^{K} \binom{K}{r} (1 - p(A_N^i))^{K-r} p(A_N^i)^{2r}$$

$$+ \sum_{i=1}^{M} p(S_N^i)(\rho_{N+1}^i)^2 \sum_{r=1}^{K} \frac{1}{r}(\sigma_N^i)^2 \binom{K}{r} (1 - p(A_N^i))^{K-r} p(A_N^i)^r$$

$$+ 2 \sum_{i=1}^{M} p(S_N^i)(\rho_{N+1}^i + (\rho_{N+1}^i)^2)(Y_N^i - \mu_{S_N^i})^t (\mu_{A_N^i} - \mu_{S_N^i})$$

$$\sum_{r=1}^{K} \binom{K}{r} (1 - p(A_N^i))^{K-r} p(A_N^i)^{2r} \qquad (3.20)$$

where

$$(\sigma_N^i)^2 \triangleq \frac{\int_{A_N^i} \|x - \mu_{S_N^i}\|^2 h(x)dx}{p(A_N^i)} \qquad i = 1,2,\ldots,M \qquad (3.21)$$

Simplifying (3.20),

$$E[W(\underline{Y}_{N+1}) | \underline{Y}_{NK}] \leq W(\underline{Y}_N)$$

$$- 2 \sum_{i=1}^{M} p(S_N^i) p(A_N^i)^K \left[1 - (1 - p(A_N^i))^K\right] \rho_{N+1}^i \, \| Y_N^i - \mu_{S_N^i} \|^2$$

$$+ \sum_{i=1}^{M} p(S_N^i) \left[1 - (1 - p(A_N^i))^K\right] \left[(\sigma_N^i)^2 + \| Y_N^i - \mu_{S_N^i} \|^2\right.$$

$$\left. + 2 \left| \left(Y_N^i - \mu_{S_N^i}\right)^t \left(\mu_{A_N^i} - \mu_{S_N^i}\right) \right| \right] (\rho_{N+1}^i)^2$$

$$+ 2 \sum_{i=1}^{M} p(S_N^i) \left[1 - (1 - p(A_N^i))^K\right] \rho_{N+1}^i \left| \left(Y_N^i - \mu_{S_N^i}\right)^t \right.$$

$$\left. \left(\mu_{A_N^i} - \mu_{S_N^i}\right) \right| \qquad (3.22)$$

Because of the boundedness assumption, $W(\underline{Y}_N)$ is bounded with probability one, as is $(\sigma_N^i)^2$. We now prove that

$$\sum_{N=1}^{\infty} (\rho_{N+1}^i)^2 \left[1 - (1 - p(A_N^i))^K\right] \qquad (3.23)$$

and

$$\sum_{N=1}^{\infty} \rho_{N+1}^i \left| \left(Y_N^i - \mu_{S_N^i}\right)^t \left(\mu_{A_N^i} - \mu_{S_N^i}\right) \right| \left[1 - (1 - p(A_N^i))^K\right] \qquad (3.24)$$

converge with probability one for each i = 1,2,...,M, thus showing that

$$\sum_{N=1}^{\infty} \sum_{i=1}^{M} p(S_N^i)[1 - (1 - p(A_N^i))^K]\left[(\sigma_N^i)^2 + \|Y_N^i - \mu_{S_N^i}\|^2 + 2\left|\left(Y_N^i - \mu_{S_N^i}\right)^t \left(\mu_{A_N^i} - \mu_{S_N^i}\right)\right|\right](\rho_{N+1}^i)^2 \qquad (3.25)$$

and

$$2 \sum_{N=1}^{\infty} \sum_{i=1}^{M} p(S_N^i)[1 - (1 - p(A_N^i))^K]\rho_{N+1}^i \left(Y_N^i - \mu_{S_N^i}\right)^t \left(\mu_{S_N^i} - \mu_{A_N^i}\right) \qquad (3.26)$$

converge with probability one. Then Lemma 3 in Appendix C can be applied with $r_N = W(\underline{Y}_N)$ and $t_N = \sum_{i=1}^{M} p(S_N^i)[1 - (1 - p(A_N^i))^K]$ $\{[(\sigma_N^i)^2 + \|Y_N^i - \mu_{S_N^i}\|^2 + 2|(Y_N^i - \mu_{S_N^i})^t (\mu_{A_N^i} - \mu_{S_N^i})|\cdot(\rho_{N+1}^i)^2 + |(Y_N^i - \mu_{S_N^i})^t (\mu_{A_N^i} - \mu_{S_N^i})|\rho_{N+1}^i\}$ to establish that $E[W(\underline{Y}_{N+1})|Y_{NK}] - W(\underline{Y}_N)$ tends to zero with probability one.

It might be commented upon here that the notation of (3.23) is slightly misleading. What is actually being considered is

$$\sum_{N=1}^{\infty} \left[(\rho_{N+1}^i)^2 [1 - (1 - p(A_N^i))^K]|Y_{NK}\right]$$

Due to the definition of the $\{\alpha_r\}_{r=1}^{\infty}$ sequence, it is sufficient to consider convergence of

$$\sum_{N=1}^{\infty} [1 - (1 - p(A_N^i))^K]/[\epsilon + 1 + w_N^i]^2 \qquad (3.27)$$

for $\epsilon > 0$, since this implies convergence of (3.23). Letting I_N^i denote the characteristic function of the event [at least one of $\{X_n\}_{n=KN+1}^{K(N+1)} \in A_N^i$]$\cup$[none of $\{X_n\}_{n=KN+1}^{K(N+1)} \in \underline{A}_N$], then $E[I_N^i \mid Y_{NK}] = (1 - (1 - p(A_N^i))^K + (1 - p(\underline{A}_N))^K)$ since they are mutually exclusive. Noting that $w_N^i = 1 + \sum_{k=1}^{N-1} I_k^i$, an application of Theorem 1 of [47], p. 274 (see Appendix C for its statement) says that for any positive numbers c, ϵ,

$$P\Big\{\epsilon + w_N^i \geq 1 + \sum_{k=1}^{N-1} [1 - (1 - p(A_k^i))^K + (1 - p(\underline{A}_k))^K]$$

$$- c \sum_{k=1}^{N-1} [1 - (1 - p(A_k^i))^K + (1 - p(\underline{A}_k))^K]$$

$$- [1 - (1 - p(A_k^i))^K + (1 - p(\underline{A}_k))^K]^2\Big\} \geq 1 - \frac{1}{1+c\epsilon} \qquad (3.28)$$

or letting $c = 1$,

$$P\Big\{\epsilon + w_N^i \geq 1 + \sum_{k=1}^{N-1} [1 - (1 - p(A_k^i))^K + (1 - p(\underline{A}_k))^K]^2\Big\}$$

$$\geq 1 - (1 + \epsilon)^{-1} \qquad (3.29)$$

Thus, with probability at least $1 - (1 + \epsilon)^{-1}$ the series of (3.27) is less than

$$\sum_{N=2}^{\infty} \frac{1 - (1 - p(A_N^i))^K + (1 - p(\underline{A}_N))^K}{\left[1 + \sum_{k=1}^{N-1} \left[1 - (1 - p(A_k^i))^K + (1 - p(\underline{A}_k))^K\right]^2\right]^2} \tag{3.30}$$

which is finite for any $\{p(A_N^i)\}_{N=1}^{\infty}$ sequence. Since the choice of ϵ is arbitrary, this establishes that (3.25) converges with probability one. A similar approach which also uses the assumption that $p(A_N^i) \to p(S_N^i)$ with probability one giving $\mu_{A_N^i} \to \mu_{S_N^i}$, $i = 1,2,\ldots,M$, shows (3.24) converges with probability one, and thus (3.26).

To determine the limit W_∞, Lemma 3 as applied above entails convergence with probability one of $\sum_{N=1}^{\infty} \{W(\underline{Y}_N) - E[W(\underline{Y}_{N+1}) | Y_{NK}]\}$. Hence, (3.22) implies convergence with probability one of

$$\sum_{N=1}^{\infty} \left\{ \sum_{i=1}^{M} p(S_N^i) \| Y_N^i - \mu_{S_N^i} \|^2 \, 2p(A_N^i)^K \, \rho_{N+1}^i \right.$$

$$\left. \cdot [1 - (1 - p(A_N^i))^K] \right\} \tag{3.31}$$

Since ρ_N^i goes to zero no faster than $O(\frac{1}{n})$, and from assumption (5) $p(S_N^i)/p(A_N^i)$ tends to a finite limit with probability one, then convergence of (3.31) implies that

$$\sum_{i=1}^{M} p(S_N^i) \| Y_N^i - \mu_{S_N^i} \|^2 \tag{3.32}$$

converges to zero on a subsequence $\{\underline{Y}_{N_s}\}$ and this subsequence itself has a convergent subsequence, say $\{\underline{Y}_{N_t}\}$. Letting

$$\underline{y} = (y^1, y^2, \ldots, y^M)$$

$$= \lim_{t \to \infty} \underline{y}_{N_t}$$

then because of the continuity of W(y) (which does not depend on continuity of $p(S_N^i)$, i = 1,2,...,M),

$$\lim_{t \to \infty} W\left(\underline{y}_{N_t}\right) = W_\infty = W(\underline{y}) \tag{3.33}$$

Since h(x) was assumed absolutely continuous with respect to Lebesque measure, $p(S_{N_t}^i)$ i = 1,2,...,M are continuous functions of $\underline{y}_{N_t}$. This continuity gives

$$\lim_{t \to \infty} \sum_{i=1}^{M} p\left(S_{N_t}^i\right) \| y_{N_t}^i - \mu_{S_{N_t}^i} \|^2 = 0$$

$$= \sum_{i=1}^{M} p(S^i) \| y^i - \mu_{S^i} \|^2 \tag{3.34}$$

From assumption (4) that the integral of h(x) is non zero on any open set contained in X, and the definition of the A_N^i subsets, then the $y_{N_t}^i$ are in the interior of the $S_{N_t}^i$ i = 1,2,...,M. As a consequence of this and the absolute continuity of h(x), $p(S^i(\underline{y})) > 0$, i = 1,2,...,M. Thus, it follows from (3.34) that

$$y^i = E[x | S^i(\underline{y})] \qquad i = 1,2,\ldots,M$$

and since $y_{N_t}^i \neq y_{N_t}^j$ if $i \neq j$, then $y^i \neq y^j$ $i \neq j$. Finally,

$$W(\underline{\gamma}_{N_t}) = V(\underline{\gamma}_{N_t}) + \sum_{i=1}^{M} p(S_{N_t}^i) \| \gamma_{N_t}^i - \mu_{S_{N_t}^i} \|^2 \qquad (3.35)$$

and from (3.32)-(3.34) this proves the theorem.

Comment: One of the more interesting results of this proof is contained in (3.31). The conclusion can be drawn from this equation that either the estimate of γ^i is updated an infinite number of times, or there is a trap state after a finite number of samples and $p(S_N^i) \to 0$. An example of how the A_N^i subsets can be used to avoid such trap states is shown in Section 3.3.

3.3 The Class of Algorithms for M = 2

The decision boundary between the two $\{S^i(\underline{\gamma}_N, \underline{P}_N)\}_{i=1}^2$ regions is defined for M = 2 by

$$(\gamma_N^1 - \gamma_N^2)^t \left(x - \frac{\gamma_N^1 + \gamma_N^2}{2}\right) \geq (\sigma)^2 \ln\left(\frac{P(\gamma_N^2)}{P(\gamma_N^1)}\right) \quad : \quad S_N^1$$

$$\text{otherwise} \quad : \quad S_N^2 \qquad (3.36)$$

Denote the threshold $(\sigma)^2 \ln\left(\frac{P(\gamma_N^2)}{P(\gamma_N^1)}\right)$ by T_N. Three subclasses of the M = 2 class of algorithms defined by the manner in which T_N is determined will be considered in the remainder of this chapter.

Subclass A. $\{P(\gamma^i)\}_{i=1}^2$ unknown and to be estimated.

A moment estimator for $P(\gamma_N^1)$ can be defined as a function of $\{\gamma_N^i\}_{i=1}^2$ and the sample mean

$$\mu_N \triangleq \frac{1}{NK} \sum_{k=1}^{NK} X_k \tag{3.37}$$

Letting $\mu = E[x]$, then

$$\mu = P(\gamma^1)\gamma^1 + (1 - P(\gamma^1))\gamma^2$$

or solving for the mixing parameter,

$$P(\gamma^1) = \frac{(\gamma^1 - \gamma^2)^t\ (\mu - \gamma^2)}{(\gamma^1 - \gamma^2)^t\ (\gamma^1 - \gamma^2)} \tag{3.38}$$

Using the estimates for γ^1, γ^2, and μ in (3.38) and noting that $P(\gamma^1) = 1 - P(\gamma^2)$, the expression for the threshold becomes,

$$T_N \triangleq T_N\ (\gamma_N^1, \gamma_N^2)$$

$$= (\sigma)^2 \ln\left[\frac{(\gamma_N^1 - \gamma_N^2)^t\ (\gamma_N^1 - \mu_N)}{(\gamma_N^1 - \gamma_N^2)^t\ (\mu_N - \gamma_N^2)}\right]. \tag{3.39}$$

An easy way to implement the constraint that $\gamma_{N+1}^i \in \mathcal{I}(S_{N+1}^i)^+$ $i = 1,2$ is to replace the $\{A_N^i\}_{i=1}^2$ sets of (3.7) with $\{S_N^i\}_{i=1}^2$,

---

$^+\mathcal{I}(A)$ denotes the interior of the set A.

calculate a $\underline{Y}^{*}_{N+1}$, and then test whether all the component vectors $\underline{Y}^{i*}_{N+1}$ fall into the correct regions S^{i*}_{N+1}. Thus,

$$\underline{Y}_{N+1} = \begin{cases} \underline{Y}^{*}_{N+1} & : \quad \underline{Y}^{i*}_{N+1} \in \mathcal{J}(S^{i*}_{N+1}) \text{ all } i = 1,2 \\ \\ \underline{Y}_{N} & : \quad \text{otherwise} \end{cases} \tag{3.40}$$

through this procedure of updating $\underline{Y}_{N+1}$, $\{A^{i}_{N}\}^{2}_{i=1}$ regions are implicitly defined and trap states are avoided.

Subclass B. $\{P(\underline{Y}^{i})\}^{2}_{i=1}$ assumed equal to 1/2.

For this subclass

$$T_{N} \triangleq 0 \tag{3.41}$$

and

$$A^{i}_{N} \triangleq S^{i}_{N} \qquad i = 1,2 \tag{3.42}$$

Subclass C. $\{P(\omega^{i})\}^{2}_{i=1}$ known along with a corresponding ordering on the components of $\underline{Y} = (\underline{Y}^{1}, \underline{Y}^{2})$ for all $\underline{Y}^{i} \in R^{\ell}$.

For this subclass

$$T_{N} \triangleq (\sigma)^{2} \ln\left(\frac{P(\omega^{2})}{P(\omega^{1})}\right) \tag{3.43}$$

In order that $\underline{Y}^{i}_{N+1} \in \mathcal{J}(S^{i}_{N+1})$ $i = 1,2$ for $K = 1$ the samples must lie in the sets[+],

---

[+]For $K > 1$, $\underline{Y}^{*}_{N+1}$ is computed and tested for both $\mathcal{J}(S^{i}_{N+1})$ and the ordering.

$$S_N^i \cap \left\{x : \alpha_{w_N^i}(x - \gamma_N^i) - (-1)^i (\gamma_N^1 - \gamma_N^2)\|^2 > 2T_N\right\}$$

$$i = 1,2 \qquad (3.44)$$

where the sets $\{x : \|\alpha_{w_N^i}(x - \gamma_N^i) - (-1)^i (\gamma_N^1 - \gamma_N^2)\|^2 > 2T_N\}$ were obtained for K = 1 by substituting γ_{N+1}^i in the decision equation of (3.36).

In addition, the updated component vectors of $\gamma_{N+1} = (\gamma_{N+1}^1, \gamma_{N+1}^2)$ must satisfy the known ordering relationship. Examples of possible ordering relationships include an energy ordering such as occurs in an ON-OFF problem, and a value ordering of a particular component of the γ_{N+1}^i, i = 1,2. The value ordering is meaningful, for instance, in a communications problem where known signals (e.g. binary antipodal) are sent through an unknown channel whose gain corresponding to the first component of the γ^i, i = 1,2, is known to be positive. This assumption is used, in fact, in the intersymbol interference simulations of Chapter 4.

To fulfill the requirement that the updated components γ_{N+1}^i i = 1,2 satisfy the ordering, implicit A_N^i, i = 1,2, regions are obtained using a procedure similar to that for Subclass A. Replace the A_N^i sets used in (3.7) by the sets defined in (3.44). At each N calculate a γ_{N+1}^* with the modified equations and test it. Then, γ_{N+1} is updated,

$$\underline{Y}_{N+1} = \begin{cases} \underline{Y}^{*}_{N+1} & : \ (Y^{1*}_{N+1}, Y^{2*}_{N+1}) \text{ satisfy the ordering} \\ \\ \underline{Y}_{N} & : \ \text{otherwise} \end{cases} \tag{3.45}$$

3.4 Algorithm Subclass Convergence for M = 2.

For the M = 2 subclasses defined in the last section, the boundary between the $\{S^i\}_{i=1}^2$ regions is a hyperplane. The boundary locations are thus considerably more limited than the more general M-ary case considered in Section 3.2. The M-ary convergence result of that section required that there exists a convex set χ such that $p(\chi) = 1$, $p(R^\ell - \chi) = 0$, and for any open set $A \in \chi$, $p(A) > 0$ where $p(A) \triangleq \int_A h(x)dx$. These assumptions are necessarily restrictive. However, for the M = 2 case, it is sufficient to define a set $\mathcal{Y} \in R^\ell$ as the convex closure of the support of h(x).[+] This relationship is illustrated in Figure 7. For any hyperplane denoted U(x) crossing $\mathcal{Y}(\mathcal{Y})$,

$$\int_{\{x \,:\, U(x)>0\}} h(x)dx > 0$$

and (3.46)

$$\int_{\{x \,:\, U(x)<0\}} h(x)dx > 0$$

Hence, for any pair $\underline{Y}^i_N \in \mathcal{Y}(S^i \ (\underline{Y}_N, \underline{P}(\underline{Y}_N)) \cap \mathcal{Y})$ i = 1,2, then $p(S^i_N) > 0$, i = 1,2.

[+]The support of a function h(x) is defined as the closure of the set $\{x : h(x) > 0\}$ [34].

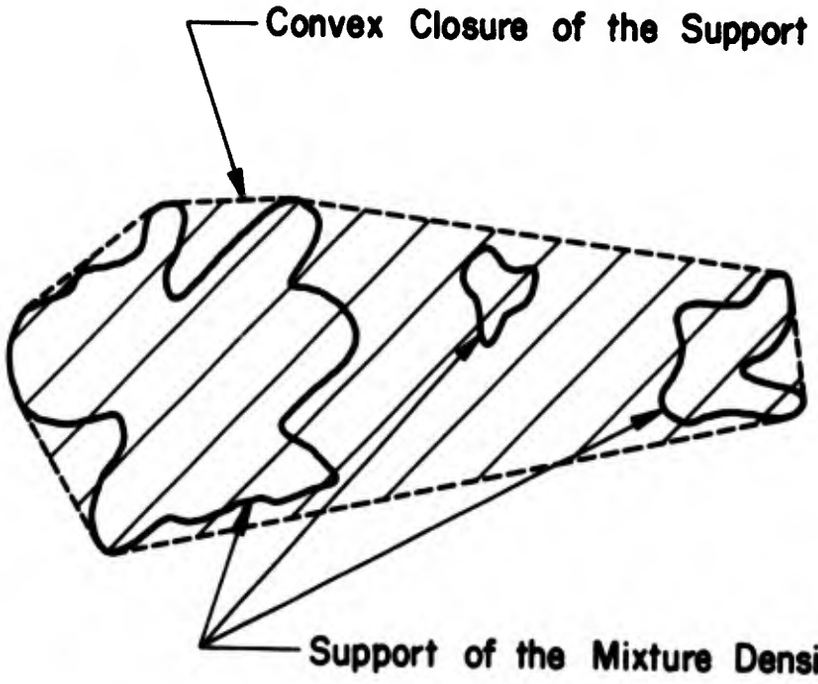


Figure 7. The Relationship Between the Support of the Mixture Density h(x) and Its Convex Closure $\mathcal{Y}$.

An additional property of the $M = 2$ case is that for Subclasses B and C it is not necessary that the starting vectors be in $\mathcal{Y}$, but rather that they be close enough to be updated eventually (this corresponds to assumption (5b) in the list below).

The following is a list of statements which will be used as needed in proving convergence of the three algorithm subclasses:

(1) $h(X_n|X_1,X_2,\ldots,X_{n-1}) = h(X_n)$

(2) $h(x)$ is absolutely continuous with respect to Lebesque measure.

(3) The support of $h(x)$ is a bounded set.

(4) Except for a permutation of the component vectors there exists only one solution of

$$\gamma^i = E[x|S^i(\underline{\gamma},\underline{P}(\underline{\gamma}))] \qquad i = 1,2$$

where

$$\gamma^i \in \mathcal{B}(\mathcal{Y}) \qquad i = 1,2$$

for each of the algorithm subclasses defined in Section 3.3.

(5) For all pairs of starting vectors $(\gamma_1^1,\gamma_1^2) = \underline{\gamma}_1$

(a) $\gamma_1^i \in \mathcal{B}(S^i(\underline{\gamma}_1,\underline{P}(\underline{\gamma}))) \quad i = 1,2$

(b) the integral inequalities

$$\int_{\{x : (\gamma_1^1 - \mu)^t (x - \frac{\gamma_1^1 + \mu}{2}) > T\}} h(x)dx > 0$$

$$\int_{\{x : (\mu - \gamma_1^2)^t (x - \frac{\mu + \gamma_1^2}{2}) < T\}} h(x)dx > 0$$

are satisfied for Subclasses B and C where T has the respective subclass definitions:

B) $T = 0$

C) $T = (\sigma)^2 \ln\left[\frac{P(\omega^2)}{P(\omega^1)}\right]$

Proposition 3. If (1)--(4), (5a) are satisfied and if $\gamma_1^i \in \mathcal{Y}$ $i = 1,2$ for all allowable starting vectors with probability one, then all the estimation algorithms in Subclass A converge with probability one and in mean square to a unique solution (except for a permutation of the component vectors) $\underline{\gamma} = (\gamma^1, \gamma^2)$, $\gamma^i \in \mathcal{Y}$.

Proof: From (4) and (5a), the restriction of the starting vectors, and the definition of the implicit $\{A_N^i\}_{i=1}^2$ region in Section 3.3, there are no trap states. Hence, there are an infinite number of updates for each mean vector estimate, and w_N^i tends to infinity (implying $\alpha_{w_N^i} \to 0$) $i = 1,2$. Together with the boundedness assumption (3), this shows that $A_N^i \to S_N^i$ with probability one $i = 1,2$. Also, for all N, $\gamma_N^i \in \mathcal{Y}(S_N^i)$, $i = 1,2$. A slight modification of the last part of the proof for Theorem 2 to take advantage of the hyperplane integral properties of $\mathcal{Y}$ ((3.46)) shows that $\underline{\gamma}_N$ converges to a unique solution (except for a permutation of vector components) with probability one. Call this solution $\underline{\gamma}^* = (\gamma^{1*}, \gamma^{2*})$ (the other solution is $\underline{\gamma}^{**} = (\gamma^{1**}, \gamma^{2**}) = (\gamma^{2*}, \gamma^{1*})$, i.e. a permutation of the vectors).

To prove convergence in mean square, suppose a mapping $\Gamma(\underline{Y}_N, \underline{Y}^*)$ of $\underline{Y}^*$ is defined such that given $\underline{Y}_N$,

$$\Gamma(\underline{Y}_N, \underline{Y}^*) = \begin{cases} (Y^{1*}, Y^{2*}) \\ \text{or} \\ (Y^{2*}, Y^{1*}) \end{cases} \tag{3.47}$$

depending on which ordering minimizes $\sum_{i=1}^{2} \|Y_N^i - \Gamma^i(\underline{Y}_N, \underline{Y}^*)\|^2$. Since $Y^{1*} \neq Y^{2*}$ there exist open neighborhoods of $\underline{Y}^*$ and $\underline{Y}^{**}$ on $(\mathcal{Y})^2$ such that Γ is continuous on each. Since $\{Y_N^i\}_{i=1}^2$ are bounded random variables, Γ is continuous at $\underline{Y}^*$ and $\underline{Y}^{**}$, and $p(Y_N^1, Y_N^2)$ converges to diracs at $\underline{Y}^*$ and $\underline{Y}^{**}$,

$$\lim_{N \to \infty} E\left[\sum_{i=1}^{2} \|Y_N^i - \Gamma^i(\underline{Y}_N, \underline{Y}^*)\|^2\right] = 0 \tag{3.48}$$

which completes the proof.

Proposition 4. If (1)--(4), (5b) are satisfied with probability one, then all the estimation algorithms in Subclass B converge with probability one and in mean square to a unique (except for a reordering of the component vectors) solution $\underline{Y} = (Y^1, Y^2), Y^i \in \mathcal{Y}$ i = 1,2.

Proof: From condition (5b), and the fact the mixture mean value converges with probability one, then for almost every Y_N^i sequence, $Y_N^i \in \mathcal{Y}$ for N large enough. The rest of the proof is similar to that for Proposition 3 and is thus omitted.

The observation is made that for this subclass, assumption (5a) implies (5b) is satisfied. Furthermore, if the initial vectors

$\gamma_1^i \in \mathcal{Y}$ i = 1,2 with probability one and $\gamma_1^1 \neq \gamma_1^2$, assumption (5) can be dropped since it is a consequence of assumption (2) and (3.46).

For the algorithms in Subclass C, not every possible ordering relationship is useful, and in fact, pathological ordering relationships are easily exhibited. Rather than to exhaustively examine each individual ordering relationship, assumptions (6) and (7) in Proposition 5 restrict the ordering relationships to those satisfying certain reasonable conditions.

Proposition 5. If (1)--(4),(5b) are satisfied and in addition, if

(6) For any $\underline{\gamma}_N = (\gamma_N^1, \gamma_N^2)$ with $\gamma_N^i \in \mathcal{I}(S_N^i \cap \mathcal{Y})$ i = 1,2 the ordering relation is such that $\prod_{i=1}^{2} p(A_N^i) > 0$ for some α_k with k finite.

(7) The ordering relationship is such that as $\alpha_{w_N^i} \to 0$, $p(A_N^i) \to p(S_N^i)$, i = 1,2.

Then the algorithms in Subclass C converge with probability one and in mean square to a unique solution $\underline{\gamma} = (\gamma^1, \gamma^2)$, $\gamma^i \in \mathcal{Y}$ i = 1,2.

Proof: Assumption (6) in addition to (4) and (5b) gives that the ordering relation does not introduce any trap states. Hence, there are an infinite number of updates for each vector and $\alpha_{w_N^i} \to 0$ i = 1,2 with probability one. From (7) this implies that $p(A_N^i) \to p(S_N^i)$ i = 1,2. The rest of the proof follows that of Proposition 3 and is thus omitted.

3.5 Asymptotic Probability of Error for Gaussian ON-OFF and Binary Cases

In this section the asymptotic probability of error of each of the M = 2 estimator subclasses defined in Section 3.3 is determined under a Gaussian mixture assumption. The method of error analysis was originally developed in [19] for the ON-OFF case and extended to the binary case in [21]. The approach of this section will be to find a set of implicit equations for the binary case asymptotic vector, with the implicit equations for the ON-OFF case obtained as a degenerate case of the binary equations.

Given a sequence of samples $\{X_k\}_{k=1}^{\infty}$ for which the estimators converge, the asymptotic vectors of the binary decision directed estimators can be expressed,

$$\lim_{N \to \infty} \nu_N^1 = E[x|\tau_\ell]$$

$$\lim_{N \to \infty} \nu_N^2 = E[x|\tau_u] \tag{3.49}$$

where τ_ℓ and τ_u refer to lower and upper truncation, respectively, such that samples following on a particular side of the ultimate decision boundary are classified as belonging to that particular class.

Define the basis vectors $\{q^k\}_{k=1}^{\ell}$ of R^ℓ such that $q^1 \triangleq [\frac{\nu^1 - \nu^2}{\|\nu^1 - \nu^2\|}]^t x$ and $\{q^k\}_{k=2}^{\ell}$ are perpendicular to q^1. Because of the spherical Gaussian assumption, the $\{q^k\}_{k=2}^{\ell}$ are uncorrelated with q^1 and

untruncated by the decision hyperplane. Hence, the analysis of the probability of error is reduced to that of the one dimensional subspace of q^1. Denote $\zeta^i = E[q^1 | \omega^i]$ $i = 1,2$. From Cramer [44], the mean of a one dimensional Gaussian random variable q^1 with mean ζ^i and variance $(\sigma)^2$ but then truncated at $q^1 = q$ is

$$t^i = \zeta^i + \lambda\sigma \qquad i = 1,2$$

where

$$\lambda = \lambda_\ell (r^i) = \frac{\phi' (r^i)}{1 - \phi(r^i)} \text{ for lower truncation}$$

$$\lambda = \lambda_u (r^i) = - \frac{\phi' (r^i)}{\phi(r^i)} \text{ for upper truncation} \qquad (3.50)$$

and

$$r^i = \frac{q - \zeta^i}{\sigma} \qquad i = 1,2 \qquad (3.51)$$

with $\phi(x)$ a Gaussian distribution function with zero mean, variance one. Evaluating the terms in (3.51),

$$r^i = \frac{\sigma}{t^1 - t^2} \ln\left[\frac{P(\gamma^2)}{P(\gamma^1)}\right] + \frac{t^1 + t^2 - 2\zeta^i}{2\sigma} \qquad (3.52)$$

The distribution function of X_n is a mixture and is upper and lower truncated by the decision hyperplane. Expanding (3.49) into the expected value of the respective truncated distributions and evaluating the factors gives,

$$t^1 = \frac{P(\omega^1)(1 - \Phi(r^i))}{D_1} \{\zeta^1 - \zeta^2) + (\lambda_\ell (r^2) - \lambda_\ell (r^2))\sigma\}$$

$$+ \{\zeta^2 + \lambda_\ell (r^2)\sigma\}$$

$$t^2 = \frac{P(\omega^1)\Phi(r^1)}{D_2} \{\zeta^1 - \zeta^2) + (\lambda_u (r^1) - \lambda_u (r^2))\sigma\}$$

$$+ \{\zeta^2 + \lambda_u (r^2)\sigma\} \qquad (3.53)$$

where

$$D^1 = P(\omega^1)(1 - \Phi(r^1)) + P(\omega^2)(1 - \Phi(r^2))$$

$$D^2 = P(\omega^1)\Phi(r^1) + P(\omega^2)\Phi(r^2)$$

The implicit equations (3.53) were solved numerically using an iterative technique on a digital computer as a function of the true class probabilities $\{P(\omega^i)\}$, the mixing parameters $\{P(\gamma^i)\}$, and the signal to noise ratio ($SNR \triangleq (\frac{1}{(\sigma)^2} (\gamma^{1_0} - \gamma^{2_0})^t (\gamma^{1_0} - \gamma^{2_0}))$ of the samples X_n. These solutions determine the r^i expressions of (3.52), and the estimation system's asymptotic probability of error is found by substituting these values of r^i in

$$P_e = P(\omega^1)\Phi(r^1) + P(\omega^2)\Phi(r^2) \qquad (3.54)$$

The minimum probability of error $P_{e_{min}}$ can be found by substituting ζ^i for t^i in (3.52) and evaluating (3.54).

The implicit equations for the ON-OFF decision directed estimators can be found by substituting zero for ζ^2 and t^2 in (3.52) and (3.53). Similar substitutions in (3.54) allows evaluation of the respective optimum and suboptimum ON-OFF system's probability of error.

The difference between the estimation system's probability of error and the minimum is defined as ΔP_e. Figure 8 contains a plot of the minimum probability of error with all parameters known. Curves are presented in Figures 9--14 showing ΔP_e vs $P(\omega^1)$ with SNR as a parameter for the ON-OFF and binary cases of each of the three algorithm subclasses defined in Section 3.3.

Examination of the results showed that the deflections in the value of ΔP_e in Figures 9--14 were due to the relative movements of the optimum and estimation system decision boundaries as the parameters varied. At certain parameter set values the boundaries cross and the respective probabilities of error are the same. The ΔP_e curves for each of the two unknown mean case algorithm subclasses given in Figures 12--14 are symmetric about $P(\omega^1) = 1/2$.

The algorithms from Subclass A estimate the mixing parameter $P(\gamma^1)$ in addition to the one or two unknown mean vectors. The asymptotic ΔP_e curves for these algorithms are given in Figures 9 and 12 for the ON-OFF and binary cases respectively. The estimators for the unknown means are biased, becoming unbiased as the signal to noise ratio increases (the densities are becoming separable). Below a signal to noise ratio of about 5, this bias severely reduces the effectiveness of the mixing parameter estimator, and

in fact, negative estimates can be produced. The class probability values $P(\omega^1)$ for which the asymptotic estimator system is nearest in performance to the optimum system are also effected by the biases. For the two unknown mean case, the biases tend to cancel near $P(\omega^1) = 1/2$ and the asymptotic estimator system performs best near $P(\omega^1) = 1/2$; however, only one mean is biased for the ON-OFF case and the best region is off-set from $P(\omega^1) = 1/2$. In both cases at resonable signal to noise ratios, the rate of degradation in moving from these "best" regions is not high.

The effect of arbitrarily assuming $P(\gamma^1) = 1/2$ as done for algorithm Subclass B is shown in Figures 10 and 13 for the ON-OFF and two unknown mean cases respectively. The curves of the ON-OFF system's increase in probability of error are nonsymmetric. The rate of degradation in performance for both ON-OFF and binary systems is extremely rapid as the differences between the assumed and actual class probabilities increases. However, for higher signal to noise ratios, these systems are adequate if $P(\omega^1)$ is not "too close" to the extreme values of zero and one. For $P(\omega^1) = 1/2$ the asumptotic probability of error of the two unknown mean system is the same as the system with all parameters known. From Figure 11, this optimality property is generally not true for the corresponding ON-OFF system. A comparison of the Subclass B curves with those discussed above for Subclass A shows that at reasonable signal to noise ratios, the Subclass A algorithms are considerably better asymptotically for most values of $P(\omega^1)$.

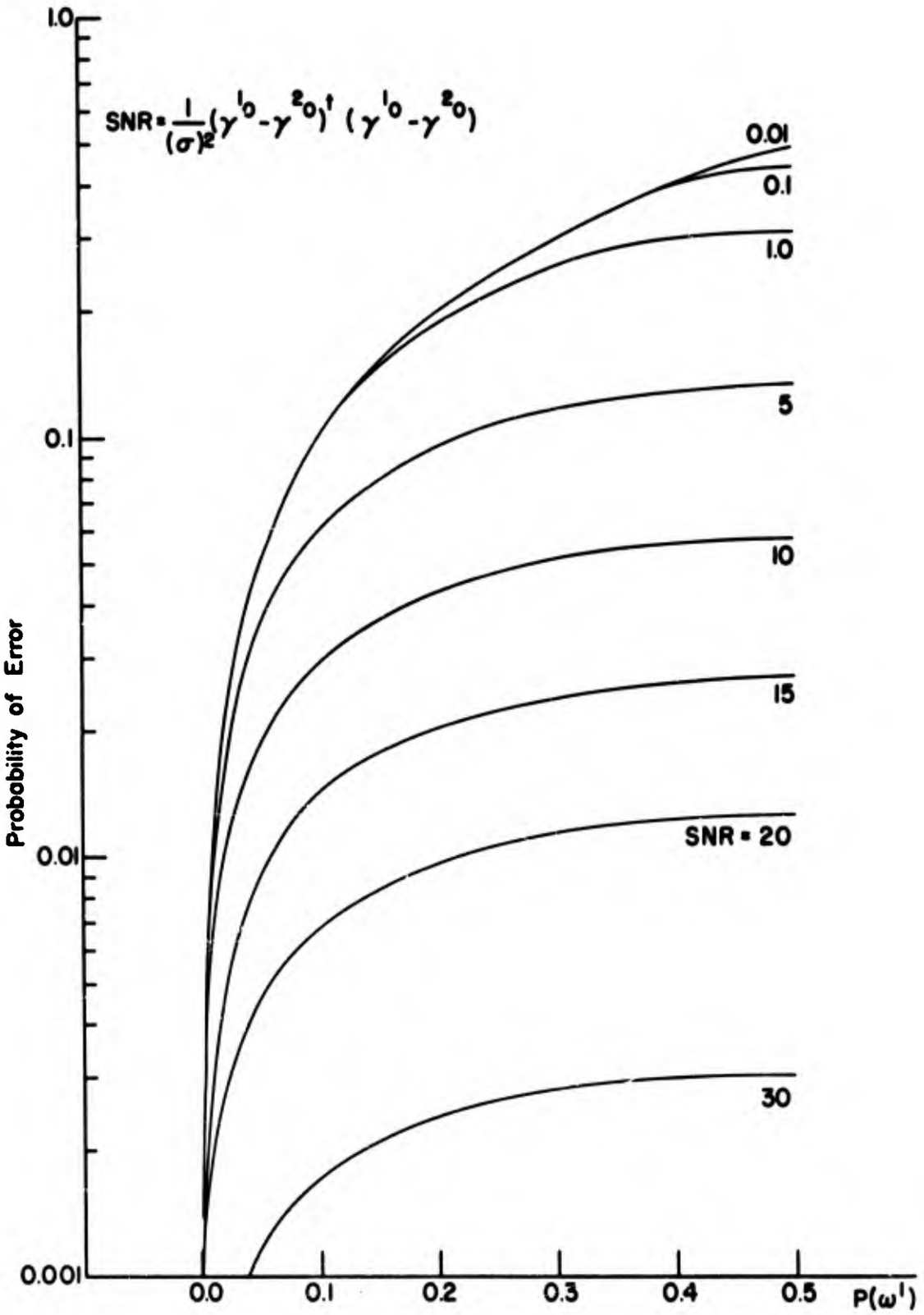


Figure 8. Minimum Probability of Error vs. $P(\omega^1)$ for Several SNR Values.

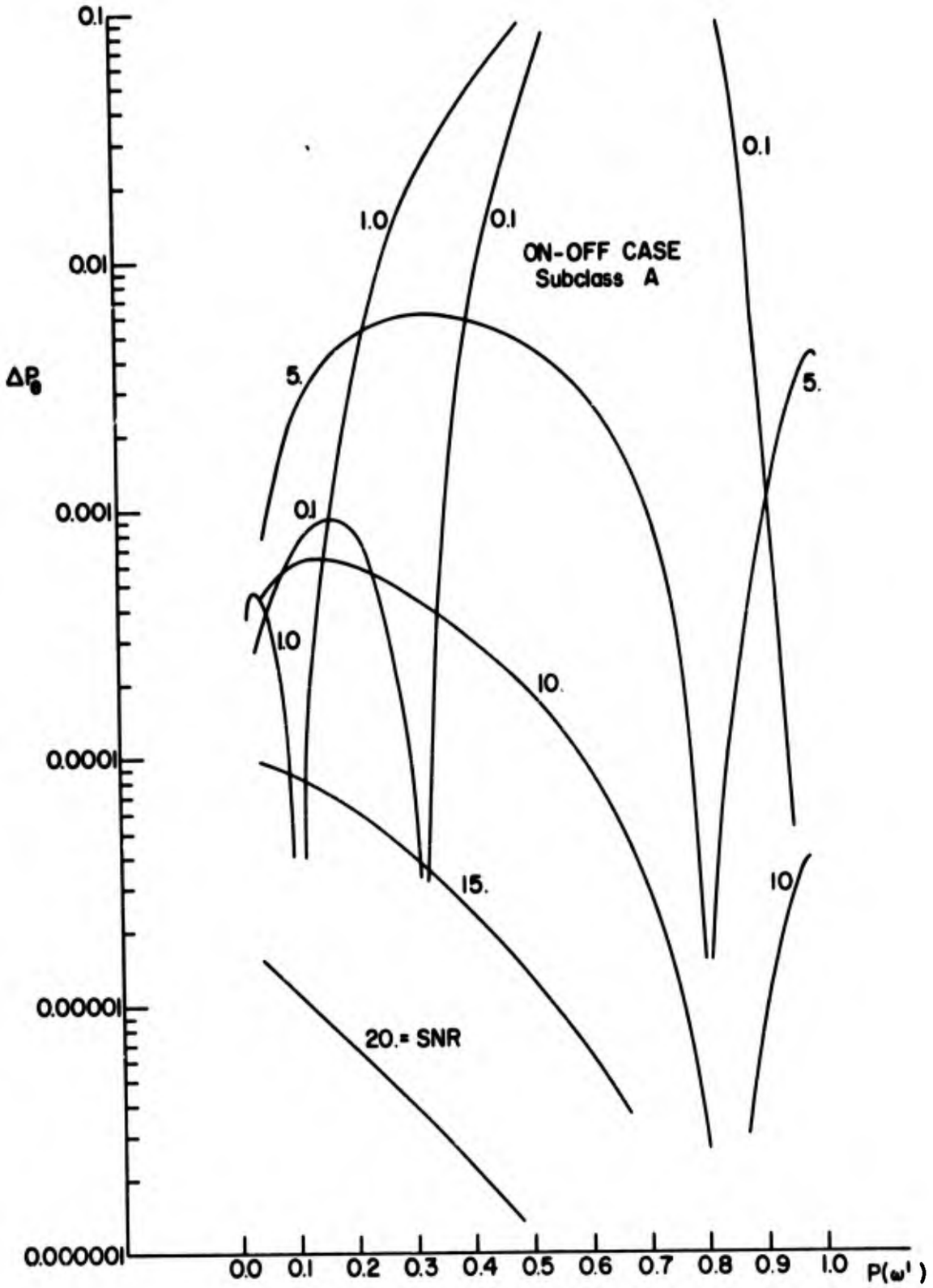


Figure 9. ON-OFF ΔP_e for $P(\nu^1)$ Estimated vs $P(\omega^1)$.

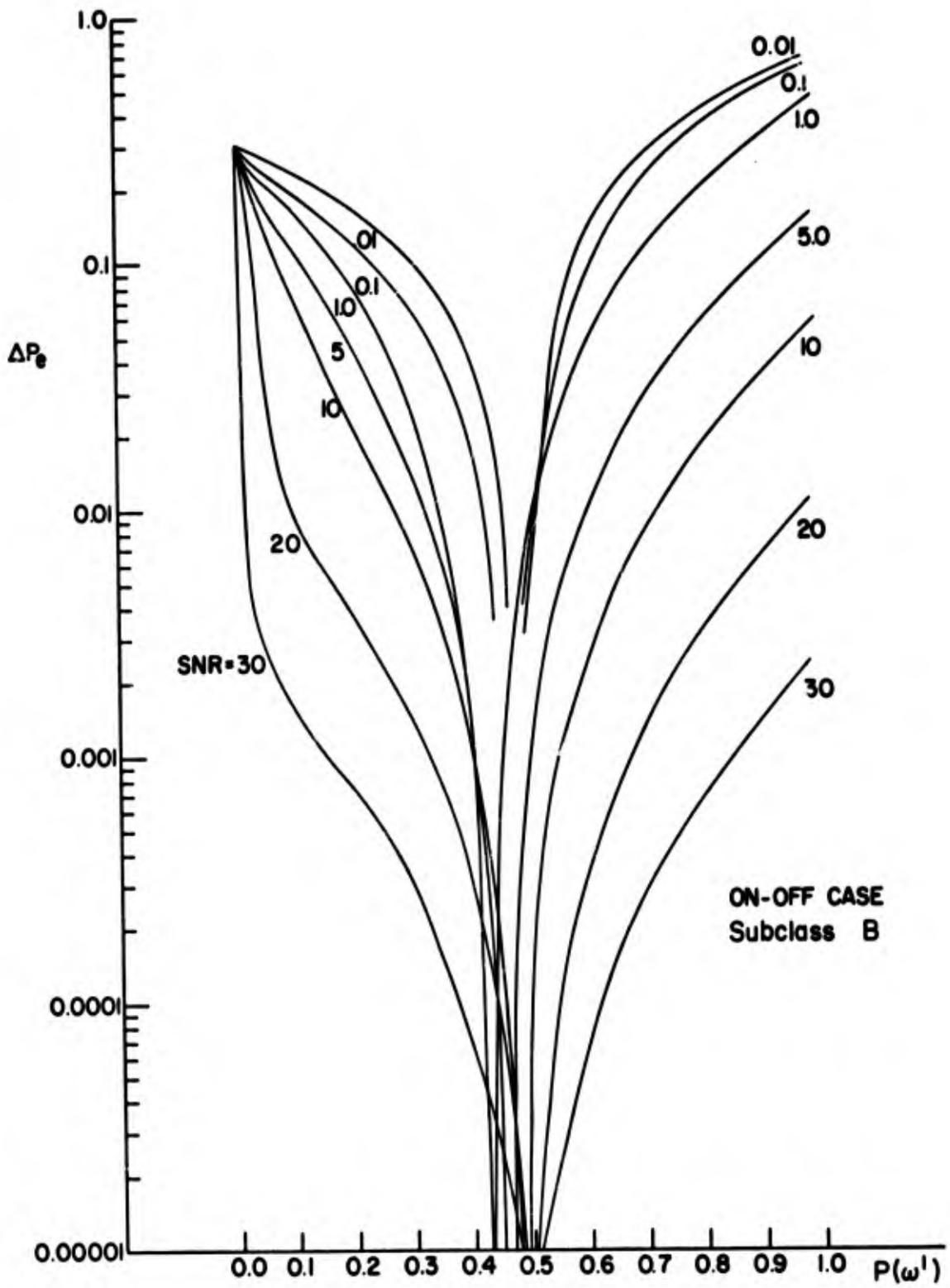


Figure 10. On-Off ΔP_e for $P(\gamma^1)$ Assumed $\frac{1}{2}$ vs. Actual Value of $P(\omega^1)$.

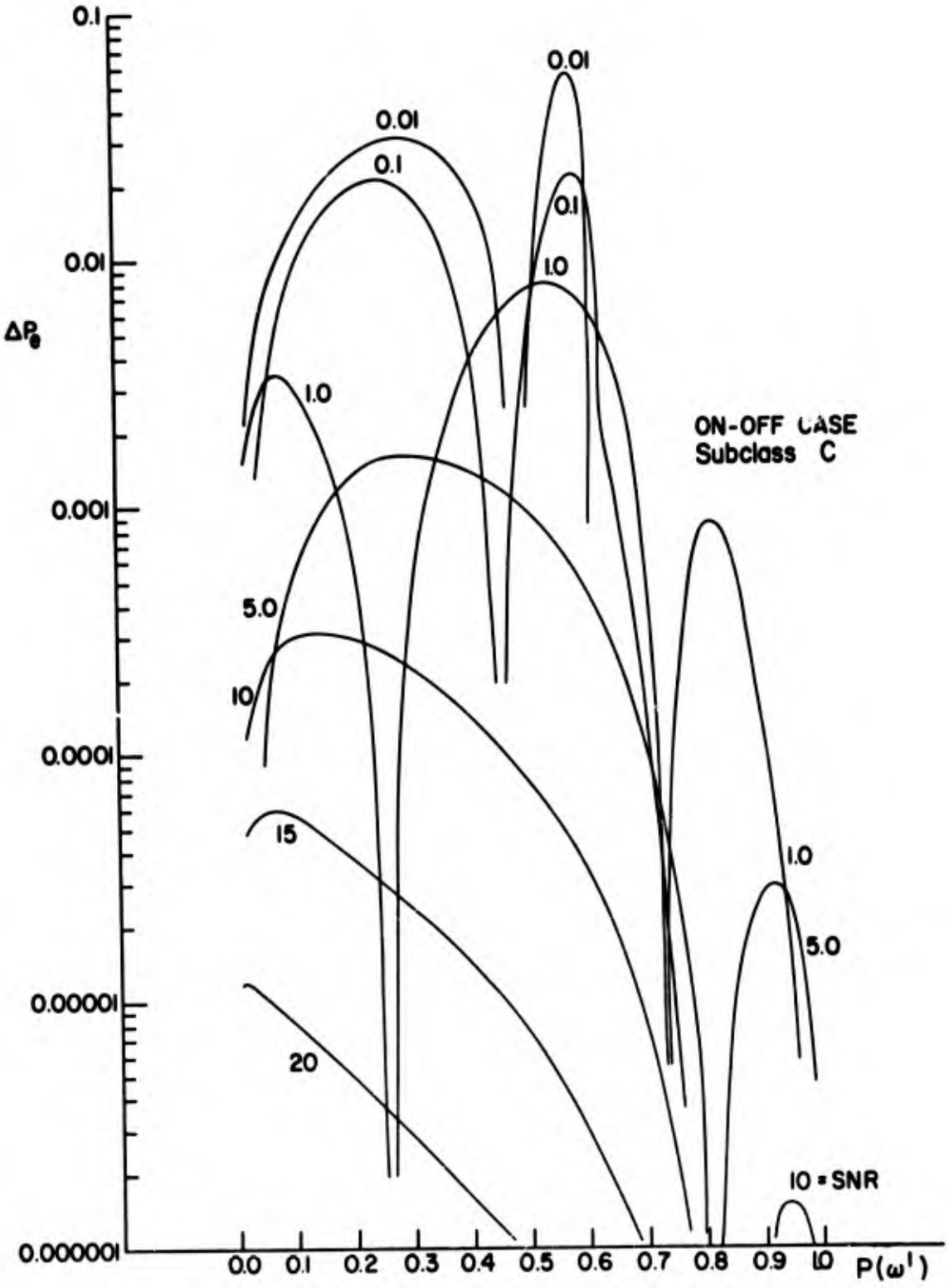


Figure 11. On-Off ΔP_e vs. $P(\omega^1)$ for $P(\omega^1)$ and an Ordering on $\{Y^i\}_{i=1}^2$ Known.

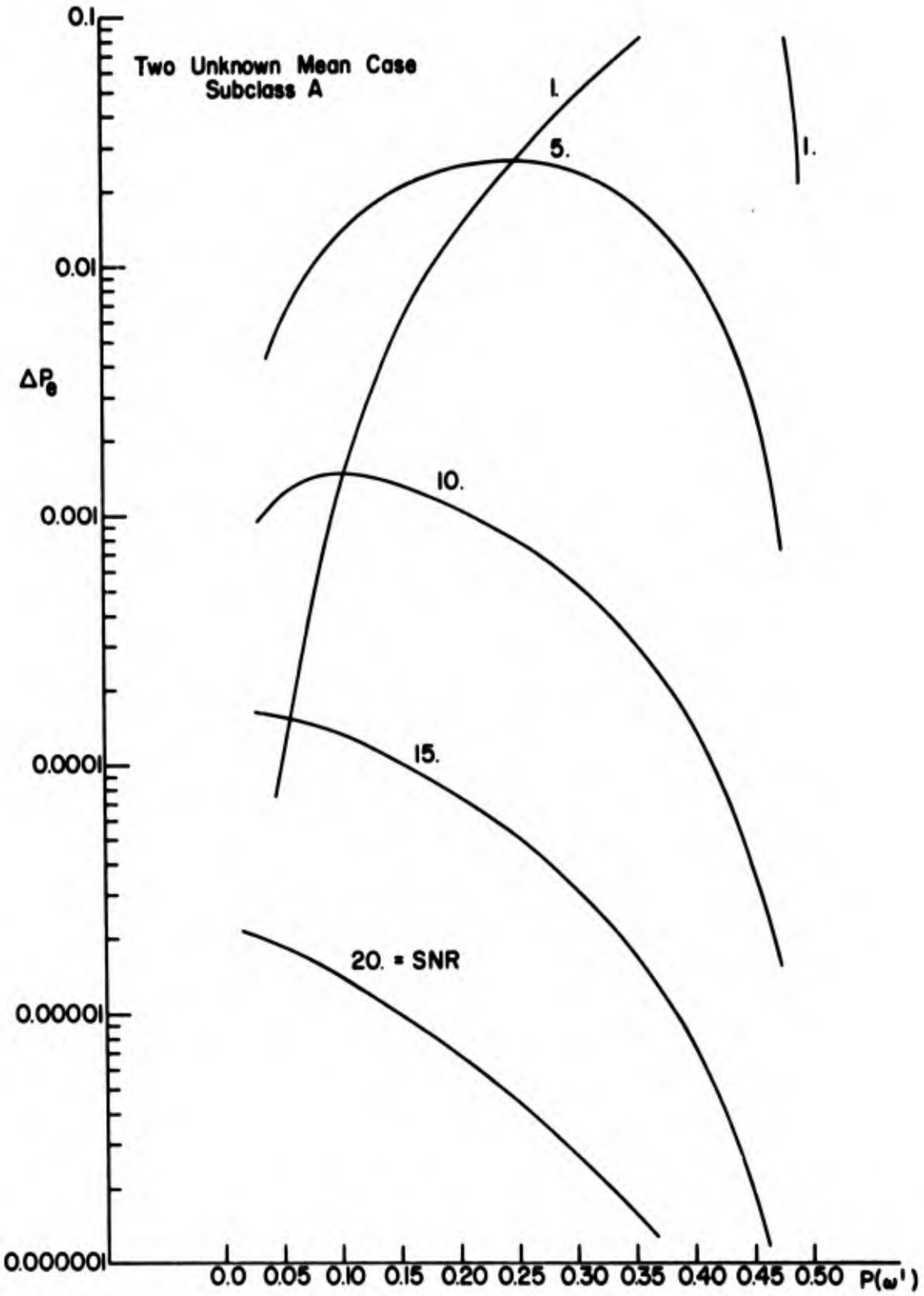


Figure 12. Two Unknown Mean ΔP_e for $P(\gamma^1)$ Estimated vs $P(\omega^1)$.

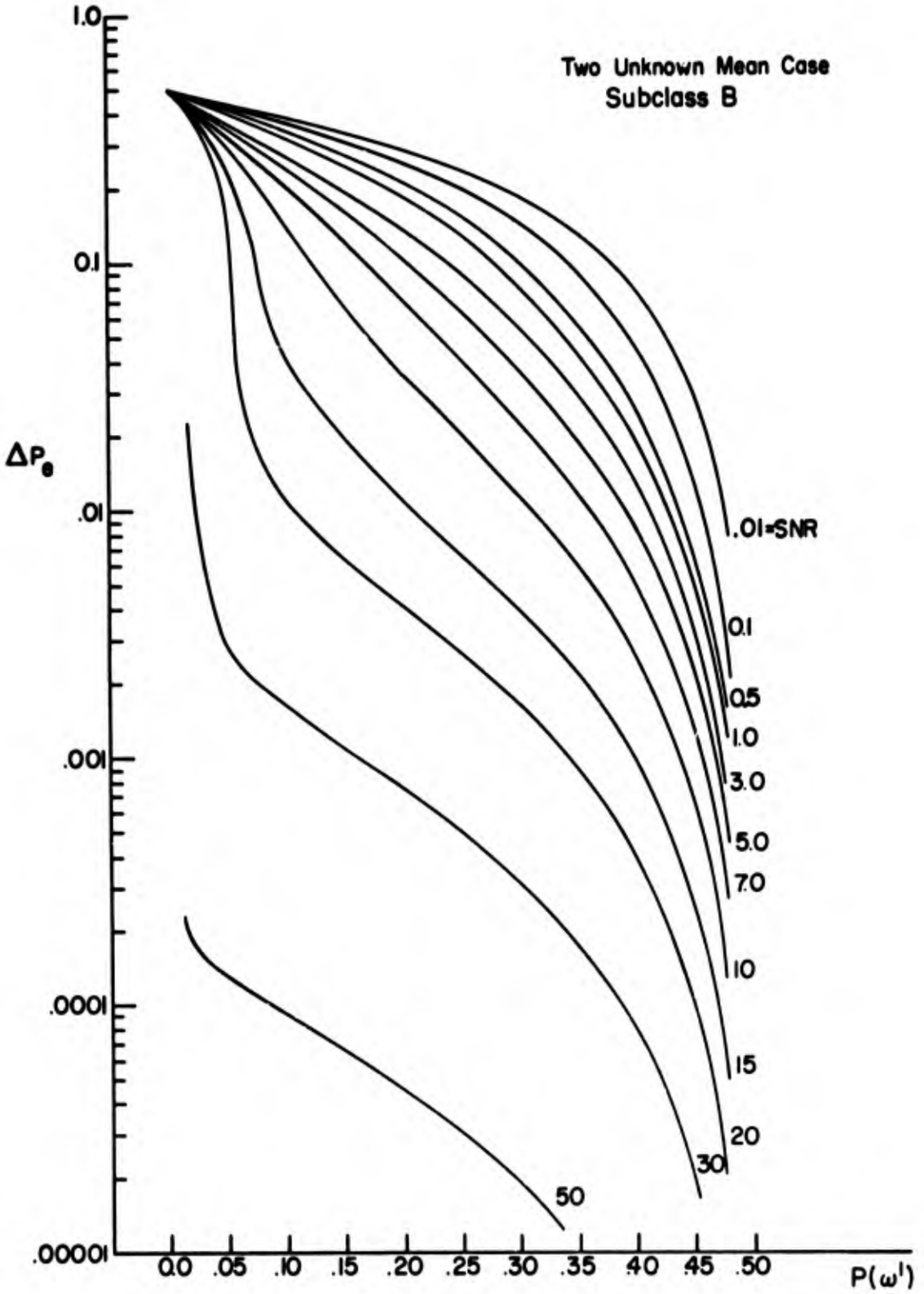


Figure 13. Two Unknown Mean ΔP_e for $P(\gamma^1)$ Assumed $\frac{1}{2}$ vs. Actual Value of $P(\omega^1)$.

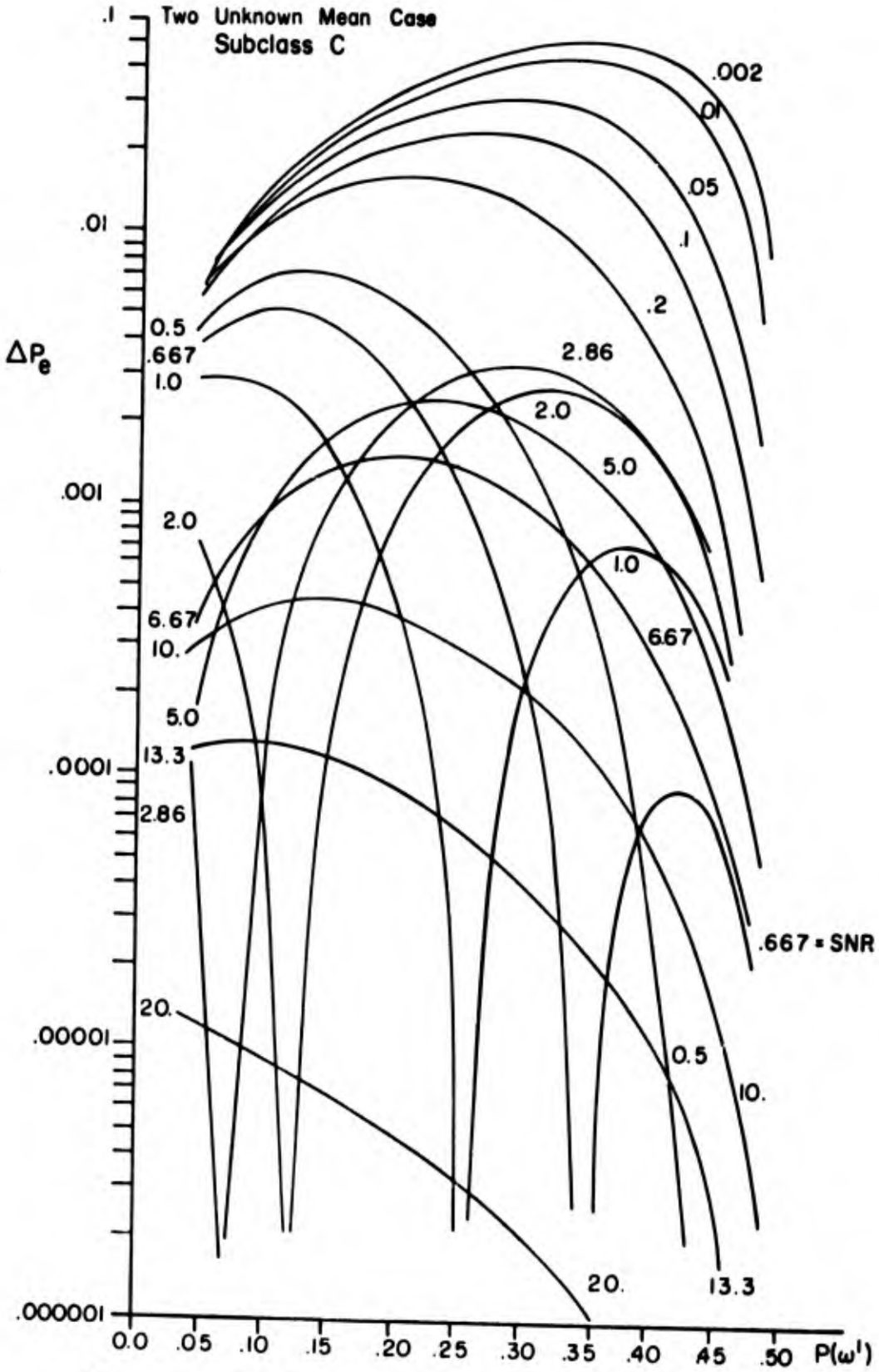


Figure 14. Two Unknown Mean ΔP_e vs $P(\omega^1)$ for $P(\omega^1)$ and an Ordering on $\{\gamma^i\}_{i=1}^2$ Known.

Figures 11 and 14 show that for the known $P(\omega^1)$ and known ordering on $\{\gamma^i\}_{i=1}^2$ assumptions made for Subclass C, the asymptotic performance of the ON-OFF and binary systems are not much worse than the system with all parameters known. The higher signal to noise ratio curves for the ON-OFF case exhibit the same off-set to the right noted on the Subclass A curves.

3.6 Experimental Algorothm Performance

Experimental dynamic performance curves for several of the $M = 2$ algorithms discussed in this chapter were obtained using a computer simulation. All of the experiments were for a four dimensional sample space ($\ell = 4$) with additive white Gaussian noise. Also, the $P(\omega^i)$, $i = 1,2$ class probabilities were defined 1/2, and the initial vectors γ_1^1 or $\{\gamma_1^i\}_{i=1}^2$ were defined as the first or first two samples for the ON-OFF and binary cases respectively. Since these parameters were common to all the simulation results of this chapter, they are not specifically labeled on the plots. Five sets of fifty experiments were performed and the experimental average probability of error for each set of experiments was calculated. The appropriate median experimental average probability of error was plotted on each of the graphs.

Curves showing the experimental convergence of the Subclass B [$K = 1$, $\alpha_k = 1/k$] algorithms for the ON-OFF and binary cases are presented in Figures 15 and 16. Similar curves for the binary case algorithm where the $P(\gamma^i)$ are estimated (Subclass A, [$K = 1$, $\alpha_k = 1/k$]) are given in Figure 17. The figures show

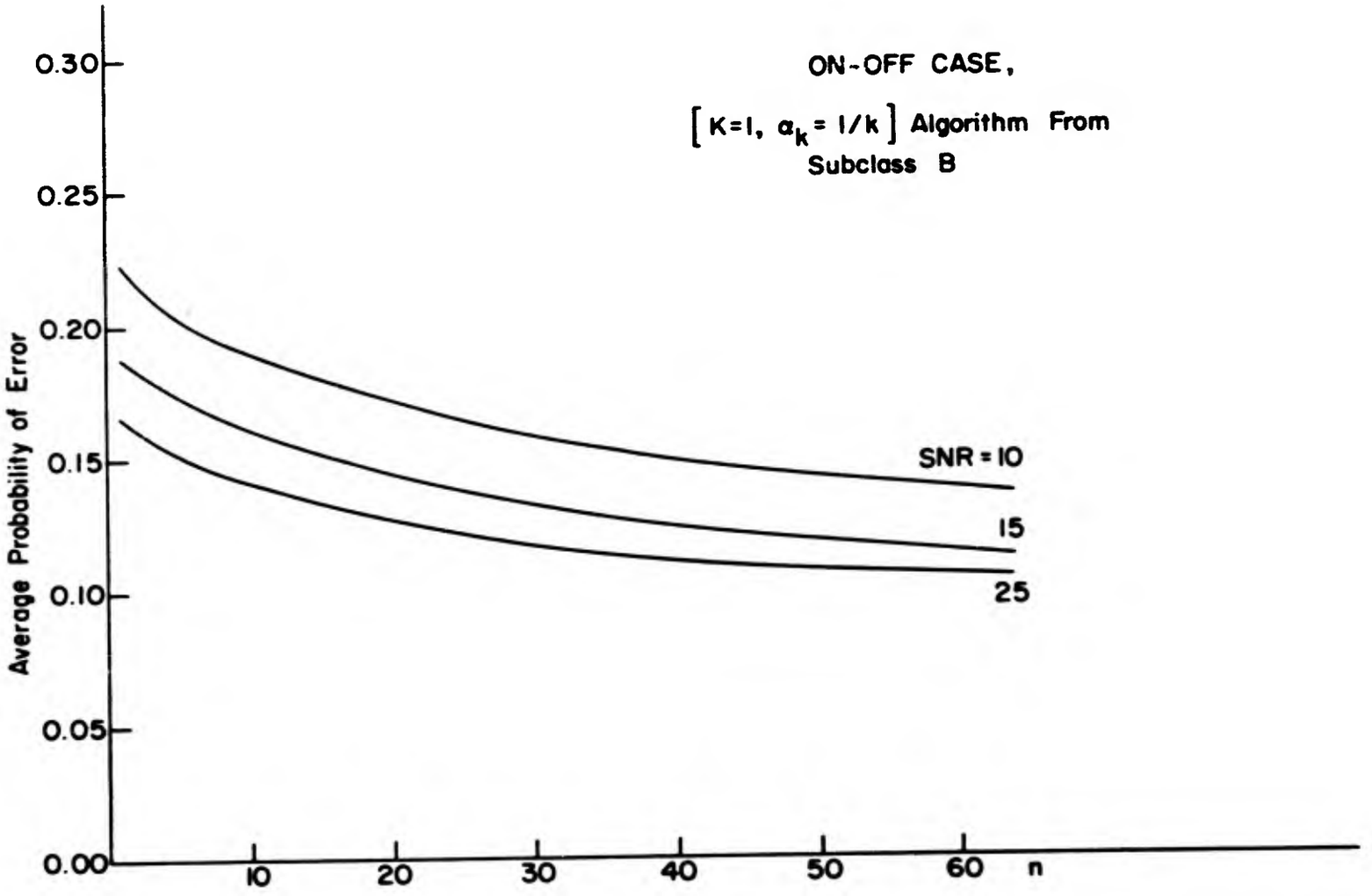


Figure 15. ON-OFF Case Experimental Average Probability of Error vs n for Several SNR Values.

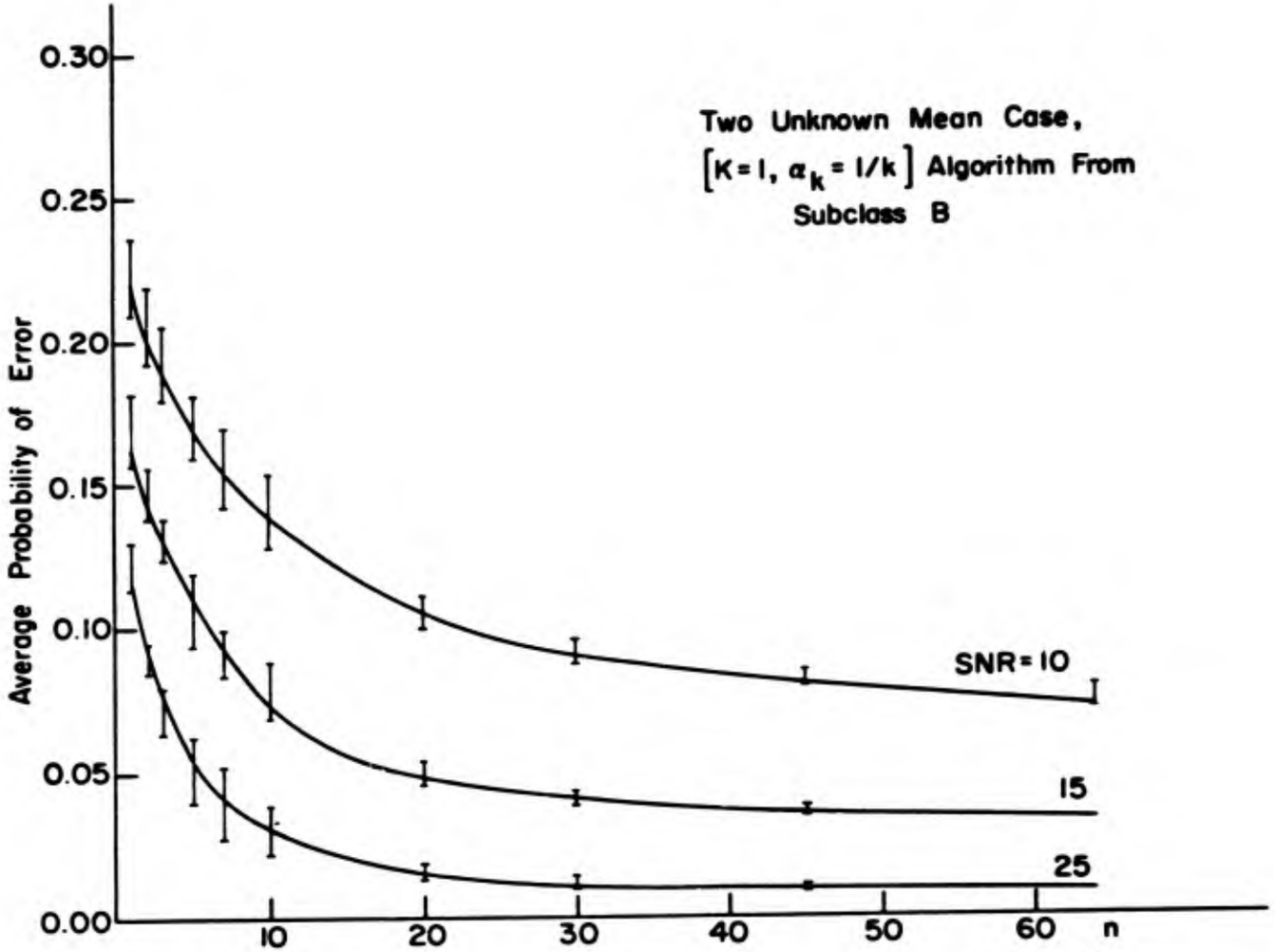


Figure 16. Two Unknown Mean Case Experimental Average Probability of Error vs n for Several SNR Values.

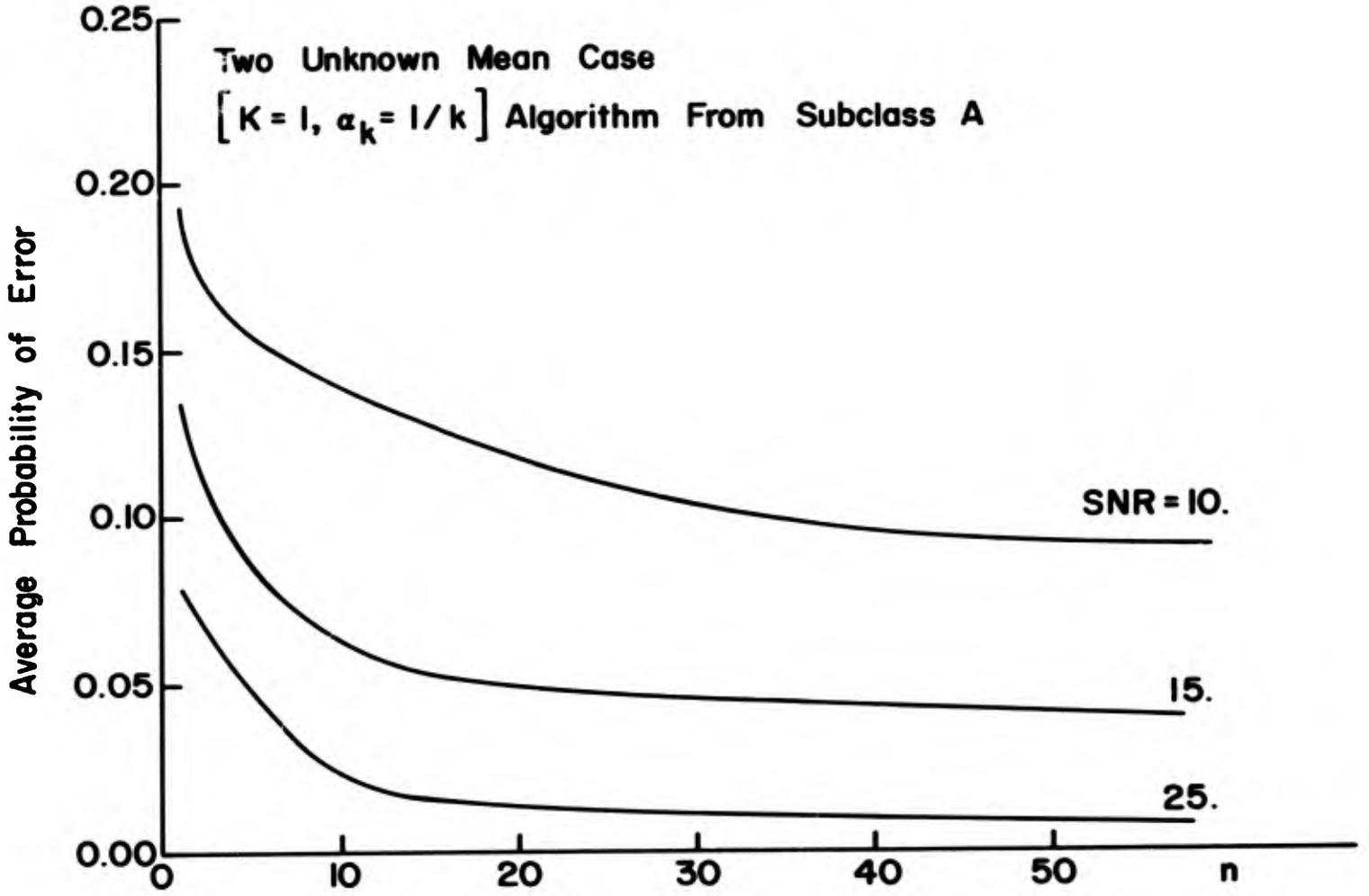


Figure 17. Experimental Average Probability of Error vs. n at Several SNR Values for Estimation of Two Unknown Mean Vectors and an Unknown Mixing Parameter.

that the ON-OFF algorithm performed considerably poorer than either of the binary algorithms, and although not indicated on the graph, the standard deviation per experiment was about 0.2. These results were expected since the binary algorithms learn only where the sample space clusters are located, while the ON-OFF algorithms also associate the $\{\omega^i\}_{i=1}^2$ with the clusters. As a result of this class association, ON-OFF algorithm starting vectors γ_1^1 located on the opposite side of the known mean γ^{2_0} from the asumptotic solution require γ_N^1 to rotate around γ^{2_0}--a time consuming task. In contrast, for starting vectors γ_1^1 located near the asymptotic solution, convergence is extremely rapid. Since at the beginning of each experiment γ_1^1 was selected as the first sample, either event can occur resulting in the large experimental variance that was observed.

As shown in Figure 17, the algorithm which estimates the $\{P(\gamma^i)\}_{i=1}^2$ in addition to the two unknown means converges slightly more slowly than the algorithm which assumes $P(\gamma^i) = 1/2$ $i = 1,2$. Also, the experimental variance was several times larger. This shows the effect of the $\{A_N^i\}_{i=1}^2$ regions which accept only "allowable" samples and throw the rest away. For some sample sequences the estimates of $P(\gamma^i)$, $i = 1,2$ differ considerably from their true value of 1/2 resulting in a large number of rejected samples.

The effect of the parameter K (the number of samples processed between estimate updates) on the convergence of the tracking mode algorithm obtained from Subclass B (i.e. $\alpha_k = 1$) is shown in Figure 18. Initial convergence is faster for smaller K values

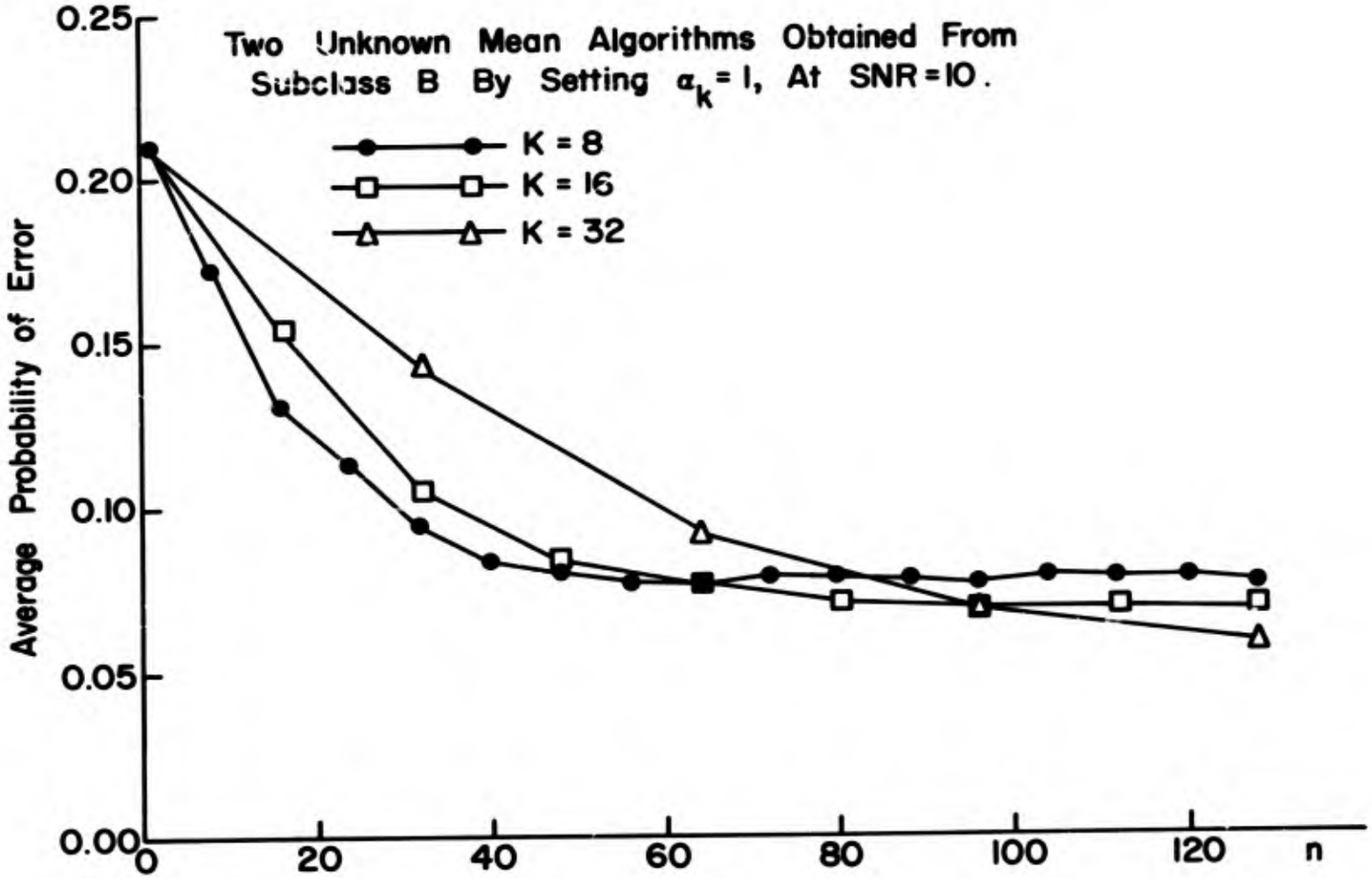


Figure 18. Two Unknown Mean Case Experimental Average Probability of Error vs. n for Algorithms with $\alpha_k = 1$ and Several Values of K.

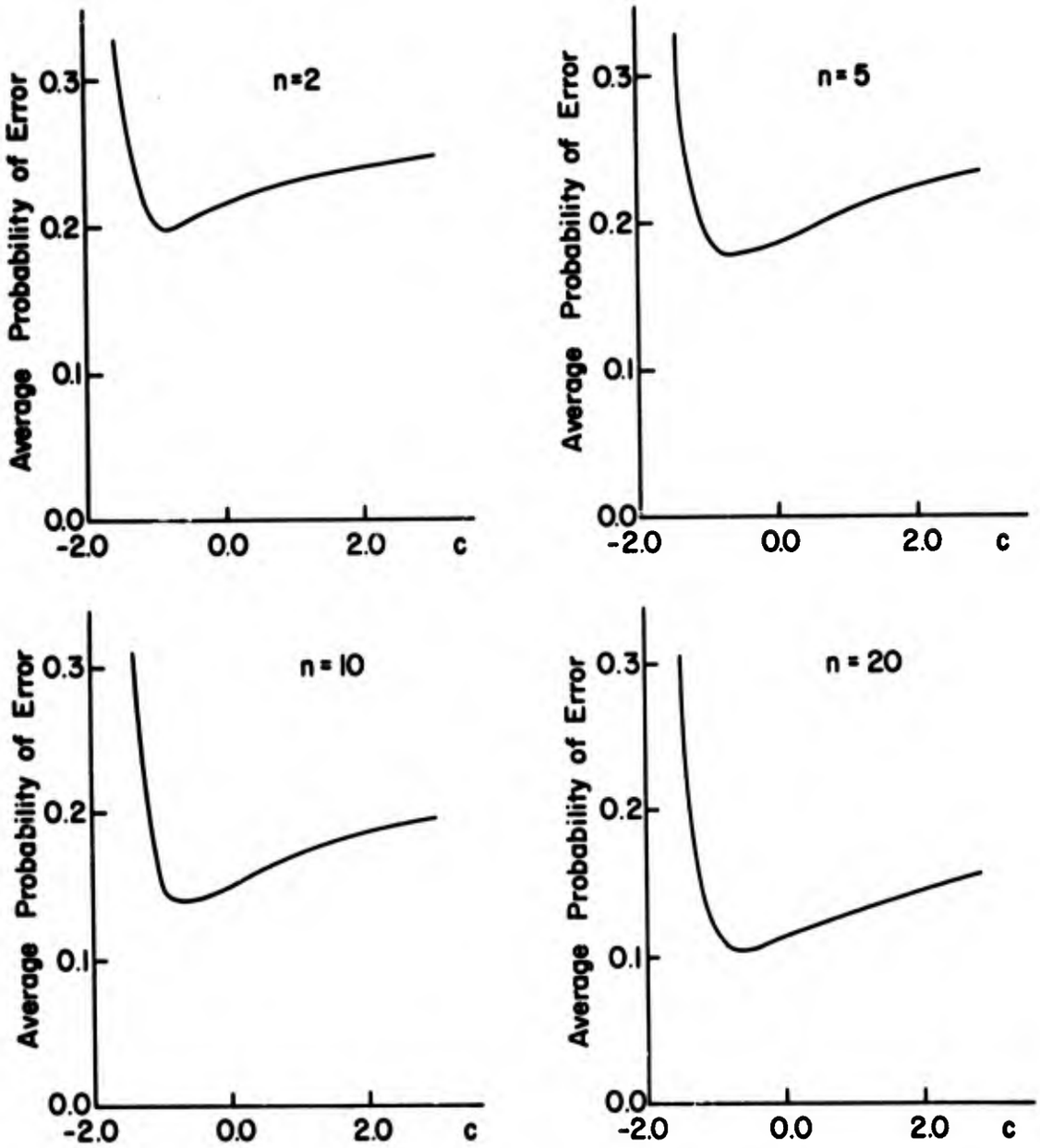


Figure 19. Experimental Average Probability of Error of Two Unknown Mean Case Algorithms from Subclass B with $[K = 1, \alpha_k = \frac{1}{k+c}]$ vs c at SNR = 10.

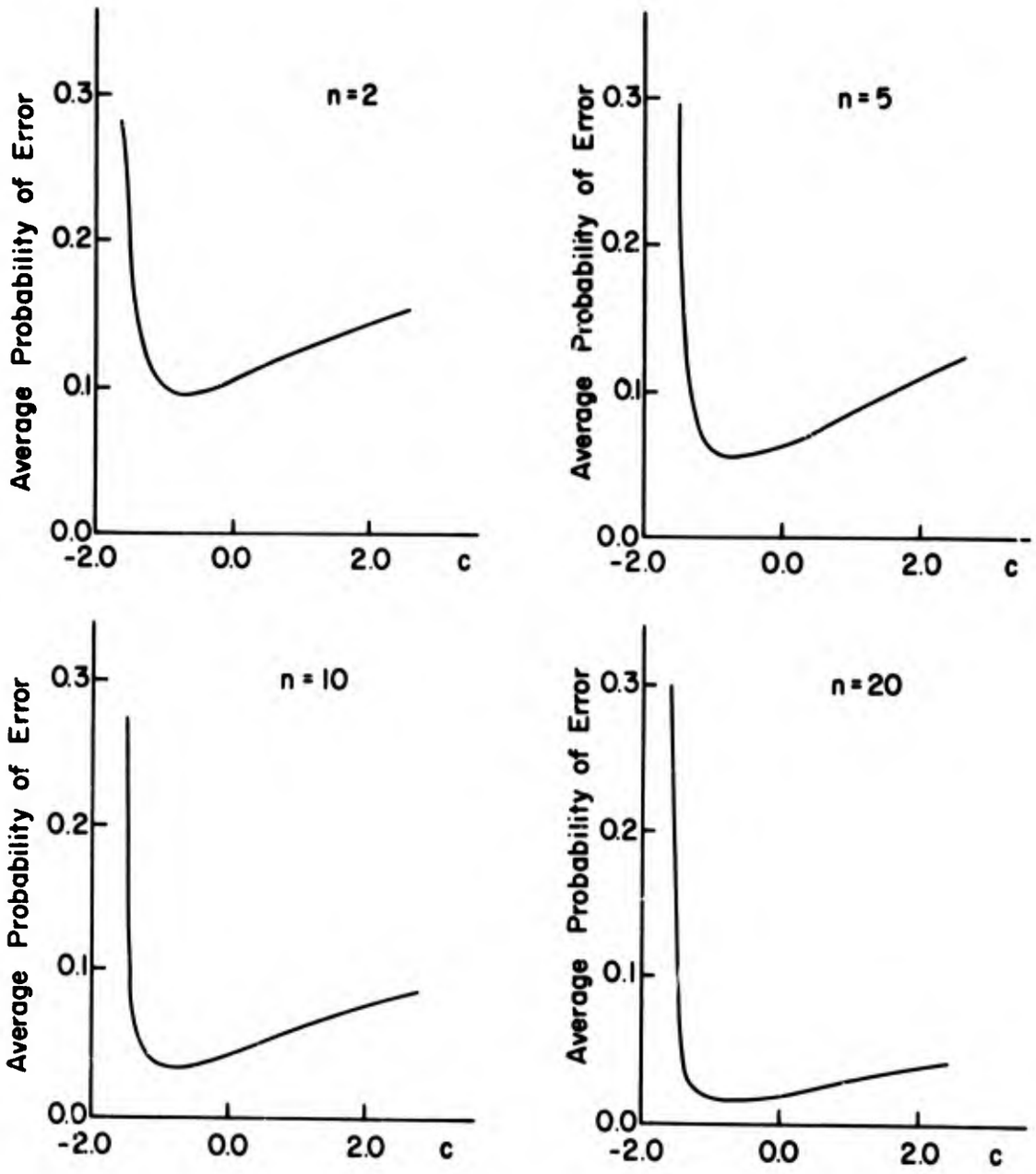


Figure 20. Experimental Average Probability of Error of Two Unknown Mean Case Algorithms from Subclass B with $[K = 1, \alpha_k = \frac{1}{k+c}]$ vs c at SNR = 25.

while the algorithms are more stable for larger values of K. Algorithms using the larger K values thus have performance close to the asymptotic error for a convergent Subclass B algorithm if the statistics are stationary.

Dynamic convergence curves of the binary algorithm from Subclass B, $[K = 1; \alpha_k = \frac{1}{k+c}, k = 2,3,\ldots]$ with c as a parameter are presented in Figures 19 and 20 for signal to noise ratios $(\text{SNR} \triangleq \frac{1}{(\sigma)^2}(\gamma^{1_0} - \gamma^{2_0})^t (\gamma^{1_0} - \gamma^{2_0}))$ of 10 and 25 respectively. At both signal to noise ratios the optimum value of c was approximately - 0.8 and resulted in a moderately better small sample performance than that for c = 0. This weight shows that a de-emphasis of the $\{\gamma_1^i\}_{i=1}^2$ initial vectors when compared to later updates[+] results in the best performance with the method of selecting the $\{\gamma_1^i\}_{i=1}^2$ used here.

3.7 Discussion

The purpose of this section is to make additional comments on the M = 2 algorithms described in this chapter. Results for the M > 2 algorithms will be reserved for the next chapter. The experimental results showed that for the initial vector selection method used, the binary Subclass B algorithms are superior to any of the others. In particular, the experimental variance per

---

[+] The parameter value c = 0 corresponds to an equal weighting and c > 0 emphasizes the initial vectors. In contrast to the result above, Dvoretzky's $\alpha_k = \frac{1}{k+c}$ error minimizing sequence for convential stochastic approximation algorithms [42] always has $c \geq 0$.

experiment was at least several times smaller than that of the other algorithms tested, and it converged faster. However, the asymptotic probability of error curves showed a sharp deterioration in the performance of the binary Subclass B algorithms for $P(\omega^i) \neq 1/2$. This suggests that a potentially desirable approach might be a two stage algorithm--a binary Subclass B algorithm is used to obtain good mean vector estimates that are then used as initial vectors in an estimator that matches the a priori knowledge. For an ON-OFF case for example, the nearest mean after a certain number of samples is associated with γ^{2_0} and the other called γ_1^1.

Another possibility not considered in the experimental results is that of supervised starting. Providing one or more samples of known classification to obtain γ_1^1 would almost eliminate the orientational difficulties of the ON-OFF algorithms that was noted in Figure 15. It would also change the result on the optimum $\alpha_k = \frac{1}{k + c}$ sequences from a negative value of c reflecting a lack of confidence in $\{\gamma_1^i\}_{i=1}^2$, to a positive value indicating the relative accuracy of initial vectors obtained with supervised samples.

The results seem to indicate that the moment estimator for the unknown mixing parameter is of questionable value. The estimator does not work at lower signal to noise ratios because of bias in the mean vector estimates. At higher signal to noise ratios the $P(\gamma^i) = 1/2$ i = 1,2 assumption gives adequate performance with faster convergence and greater simplicity.

Use of $K > 1$ parameter values will obviously slow down experimental convergence for the independent samples used in the computer simulation. It remains to be shown for dependent samples what degree of stability $K > 1$ adds to the classical $K = 1$ decision directed estimator.

IV. UNSUPERVISED ESTIMATION OF SIGNALS WITH INTERSYMBOL INTERFERENCE

In an effort to cope with the information requirements of an expanding technology, information transmission system designers have often desired to increase signaling rates through the usual bandlimited channels. These channels exhibit the deleterious effects of "memory" (actually energy storage) at higher signaling rates when part of the energy from one transmitted signal band is smeared onto the next band. Since the smeared transmitted bands overlap, this condition is called intersymbol interference. The seriousness of the problem has motivated a considerable amount of work on techniques that reduce the effect of the interference. For the case where the channel and noise statistics are known, the probability of error [24], [25], [26] and [28], or mean square error [26] and [27] has been minimized through proper signal design, receiver design, or joint transmitter-receiver design. These results have been obtained by using almost classical optimization procedures. Also, bounds on the probability of error of conventional correlator receivers [29] and [30] have been obtained. For many practical problems however, the effect of the channel may be unknown, or if initially known, the channel may be time varing. In either case estimation is

required. Rather than sending known signals through the channel (implying the message source is turned off during this period) and either implicitly [31] or explicitly estimating the channel statistics[+], the approach used here is to estimate the necessary statistics without interfering with the transmitter's primary function--information transmission.

This chapter investigates the application of an unsupervised estimation and processing receiver to intersymbol interference in the class of linear, time invariant channels.[++] The unsupervised estimation algorithms used are similar to the class of decision directed algorithms defined in the last chapter. In contrast to previous papers, the underlying statistical structure of the intersymbol interference sample space is emphasized. This provides a unifying framework for the problem within which different approaches can be compared. For convenience and simplicity in notation, we restrict attention to the case of one dimensional binary antipodal transmitted signals ($\ell = 1$, $M = 2$). A more general formulation is not presented until Section 4.6 since the resulting complex notation obscures the basic concepts. However, the multidimensional sample space generalizations of the unsupervised estimation algorithms are immediate, as is the

---

[+]This is called supervised estimation using sounding signals.

[++]Practically speaking, this also includes those channels which vary slowly compared to the convergence of the estimators.

interpretation of the $\ell > 2$ statistical structure of the problem. More general signal classes require only a minor (although cumbersome) modification of the defining equations.

4.1 The Intersymbol Interference Problem

A block diagram of the intersymbol interference problem is given in Figure 21. One of two sources ω^i, $i = 1,2$, with corresponding one dimensional message waveforms $\{m^i\}_{i=1}^{2}$ produces the transmitted waveform m_n at the n^{th} time. The message sequence is assumed statistically independent. This waveform is propagated through a linear time invariant channel which can be characterized by a linear channel transformation followed by additive noise. The linear channel transformation is defined

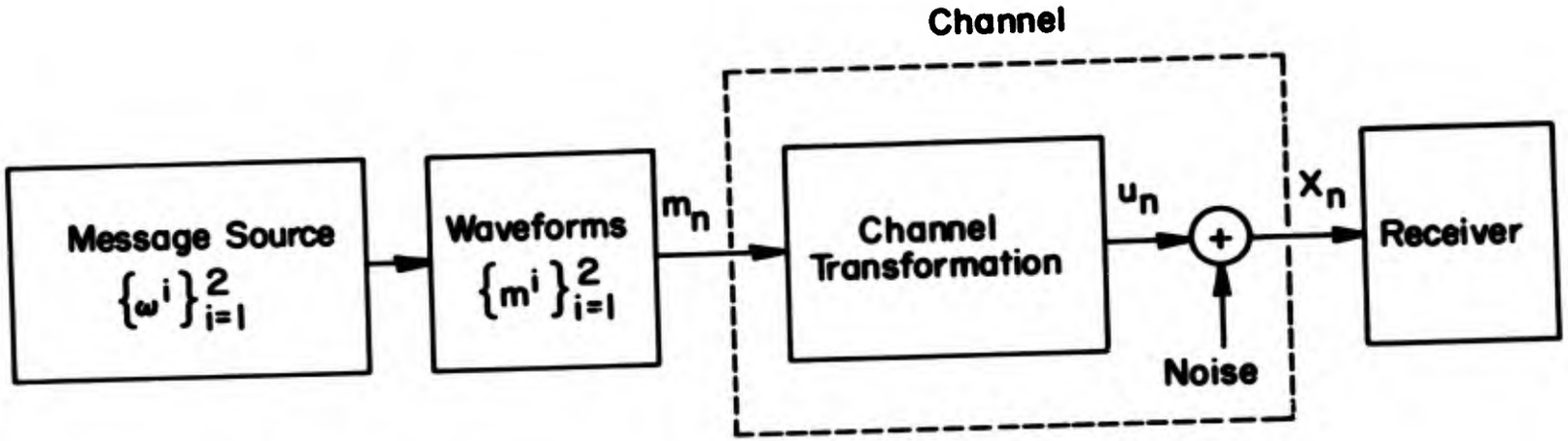
$$u_n = \sum_{k=1}^{n} a_{n-k+1}\, m_k \tag{4.1}$$

where the $\{a_k\}_{k=1}^{\infty}$ are unknown channel gains with a_c the last non zero gain ($c < \infty$), and u_n is the n^{th} signal resulting from the channel transformation. The index on the last non zero channel gain defines the channel memory as $c - 1$ bauds. The received signal X_n is given by

$$X_n = u_n + h_n$$

$$= \sum_{k=1}^{n} a_{n-k+1}\, m_k + h_n \tag{4.2}$$

where h_n is assumed to be additive white Gaussian noise.



Channel Transformation

$$u_n = \sum_{k=1}^{n} a_{n-k+1}\, m_k$$

Figure 21. Problem Model

From the channel model given in (4.2), if $c > 1$ the sample sequence $\{X_k\}_{k=1}^{n}$ is <u>not</u> statistically independent even though the original message sequence and additive noise were assumed to have this property. Because of the statistical dependence of the samples, the optimum decision procedure to determine which source ω^i, $i = 1,2$, produced m_n differs from approaches which assume statistical independence. For such an optimum procedure, all the energy from m_n must be transmitted through the channel <u>before</u> a decision on m_n is made. As a consequence, there is a <u>delay</u> of at least $c - 1$ samples between when m_n is transmitted, and when a decision is made on the source active at the n^{th} time. Hence, an optimum decision procedure requires $\{X_k\}_{k=1}^{n+c-1}$ be available for processing.

An optimum decision procedure thus has the liabilities of being an exponentially growing decision problem with an unlimited storage requirement. In this chapter we are interested in a more practical class of decision procedures which assume that at time j only the <u>last v samples</u> $\{X_k\}_{k=j-v+1}^{j}$ are stored. The decision procedures being considered are those that when given $\{X_k\}_{k=n-v+1}^{n}$, decide which class ω^i was active at time $n - v^* + 1$ where $1 \leq v^* \leq v$. Decision procedures used in several previous papers [24]-[26], [28], [32] and [33] are in this class, and it will be possible to compare their assumptions on v and v^*.

The unsupervised estimation algorithms defined in Section 4.3 assume that the signal class (binary antipodal) and the value of c are known at the receiver. However, these parameters could

also be estimated using, for example, one of the unsupervised estimation algorithms defined in Section 2.4.

4.2 The Statistical Structure of the Sequence Sample Space of $\{X_k\}_{k=n-v+1}^{n}$

As described in the previous section, the sequence of the last v samples $\{X_k\}_{k=n-v+1}^{n}$ is to be used to discriminate between the events ω^1 active at time $n - v^* + 1$, and ω^2 active at $n - v^* + 1$. In this section the structure of the sequence sample space will be examined. The number and relative locations of modes will give the statistical objectives that estimation algorithms should be designed to achieve. An understanding of the mutual constraints on the locations of the modes will indicate an estimation strategy utilizing this a priori knowledge, for which a decision directed type of estimator is particularly well suited. The discussion also includes the relative advantages of different decision procedures, and the definition of the suboptimum procedure used here.

Define the v dimensional sequence vectors,

$$\underline{X}_n = \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-v+1} \end{pmatrix} \qquad \underline{u}_n = \begin{pmatrix} u_n \\ u_{n-1} \\ \vdots \\ u_{n-v+1} \end{pmatrix} \tag{4.3}$$

From the definition of (4.2), $\underline{X}_n$ is a v dimensional Gaussian random variable with mean vector $\underline{u}_n$, and covariance matrix

$(\sigma)^2$ I. Because of the finite channel memory of c - 1 bauds, there are only a finite number of unique vector values that $\underline{u}_n$ can attain. To be more precise, given any sequence of message events $\{m_k\}_{k=1}^{n}$, the channel transformation of $\{m_k\}_{k=1}^{n}$ into $\underline{u}_n$ ((4.1)) can be expressed in matrix form,

$$\begin{pmatrix} u_n \\ u_{n-1} \\ \vdots \\ u_{n-v+1} \end{pmatrix} = \begin{pmatrix} m_n & m_{n-1} & \cdots & m_{n-c+1} \\ m_{n-1} & m_{n-2} & & m_{n-c} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-v+1} & m_{n-v} & & m_{n-v-c+2} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{pmatrix} \tag{4.4}$$

Defining a (vxc) matrix Q with general term $q^{ij} = \{m_{n+2-i-j}\}$ $1 \le i \le v$, $1 \le j \le c$, and a vector $\underline{a} \triangleq (a_1, a_2, \ldots, a_c)^t$, equation (4.4) can be written,

$$\underline{u}_n = Q\,\underline{a} \tag{4.5}$$

There are exactly 2^{c+v-1} unique Q^k matrices generated by the totality of message event sequences, and this number is not a function of n for $n > c + v - 1$. Hence, there are at most $J \le 2^{c+v-1}$ unique $\underline{u}_n$ vectors denoted $\{\underline{u}^k\}_{k=1}^{J}$. The assumption is made in this chapter that there are no zero elements in $\underline{a}$. Under these conditions, all 2^{c+v-1} $\underline{u}^k$ vectors are unique.[+] Since all the $\underline{u}^k$ can be generated using the known $\{Q^k\}_{k=1}^{J}$ matrices and

---

[+]To show this, assume $\underline{u}^1 = \underline{u}^2$. This implies that $Q^1\,\underline{a} = Q^2\,\underline{a}$ or $(Q^1 - Q^2)\underline{a} = 0$. Hence, $\underline{a}$ lies in the null space of $(Q^1 - Q^2)$. Since $Q^1 \neq Q^2$ (by definition), the rank of $(Q^1 - Q^2)$ is at least one. This implies that $\underline{a}$ has at least one zero, a contradiction.

the vector $\underline{a}$, these matrices can be interpreted as containing the mutual constraints on the mode locations. The $\{Q^k\}_{k=1}^{J}$ matrices will be used in an unsupervised estimation algorithm that was found to perform very well for the intersymbol interference problem.

4.2.1 Figures Illustrating the Mode Structure. Because of the zero mean additive noise assumed in the channel model, around each $\underline{u}^k$ there is a "cluster" or concentration of density mass. Figures 22-24 are presented to show the mode structure of the sequence sample space. These figures will be used in comparing different decision procedures in Subsection 4.2.2. All figures are for one dimensional equilikely antipodal transmitted signals, and the channel gains of $a_1 = .95$ and $a_2 = .3$ are typical for an RC channel with $c = 2$.

As shown in Figure 22, the mixture density for $\underline{v = 1}$ has four modes corresponding to the four possible message events at times $(n - 1, n)$ that produce X_n. The message events corresponding to each mode are labeled on the figure.

The case where both X_{n-1} and X_n are available $\underline{(v = 2)}$ for decision making is illustrated in Figure 23. There are 8 modes and the message event sequence at times $(n - 2, n - 1, n)$ corresponding to each mode is enclosed in parenthesis. The center of each mode is one of the $2^{c+v-1} = 8$ unique $\underline{u}^k$ sequence vectors discussed previously.

Finally, the case where $\underline{v = 3}$ and $\{X_k\}_{k=n-2}^{n}$ are available for use in a decision procedure has the mode structure illustrated

in Figure 24. Again each mode is labeled with the message event sequence at times (n - 3, n - 2, n - 1, n) corresponding to the particular mode.

The underlying relationship between these plots is that Figures 22 and 23 could be generated from Figure 24 by taking the respective v < 3 dimensional projection onto (X_n) and (X_{n-1}, X_n) respectively. Similarly, a projection of Figure 23 onto (X_n) gives Figure 22. As the parameter v is taken smaller, the message event sequences corresponding to the mode vectors in the original space are truncated. Clusters which were distinct in the original space, but whose truncated message event sequences are identical for the smaller v value, "collapse" into each other. Conversely, if v is increased by one, each mode in the original space forms two new modes in the higher dimensional space.

Let v_1 and v_2 denote two sample sequence lengths with $v_1 < v_2$ and $v^* \leq v_1$. Also, let $\{m'_k\}_{k=n-v_1-c+2}^{n}$ and $\{m''_k\}_{k=n-v_2-c+2}^{n}$ denote the message event sequences corresponding to the closest v_1 dimensional modes having $m'_{n-v^*+1} = 1$ and $m''_{n-v^*+1} = -1$. It should be noted from Figures 22-24 that the $2^{v_2-v_1}, v_2$ dimensional modes that can be generated from each of $\{m'_k\}_{k=n-v_1-c+2}^{n}$ and $\{m''_k\}_{k=n-v_1-c+2}^{n}$ do not necessarily contain the closest pairs of v_2 dimensional mode vectors with one having $m_{n-v^*+1} = 1$ and the other $m_{n-v^*+1} = -1$. This implies that the interclass distance increases with increasing v, or equivalently, the probability of error is a monotone decreasing function of increasing v.

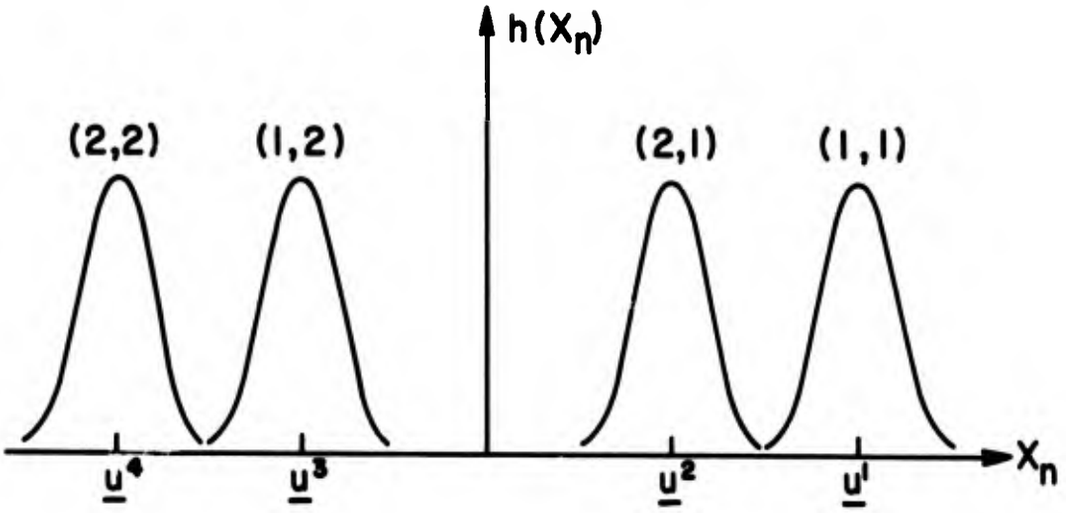


Figure 22. Mode Structure for v = 1, c = 2.

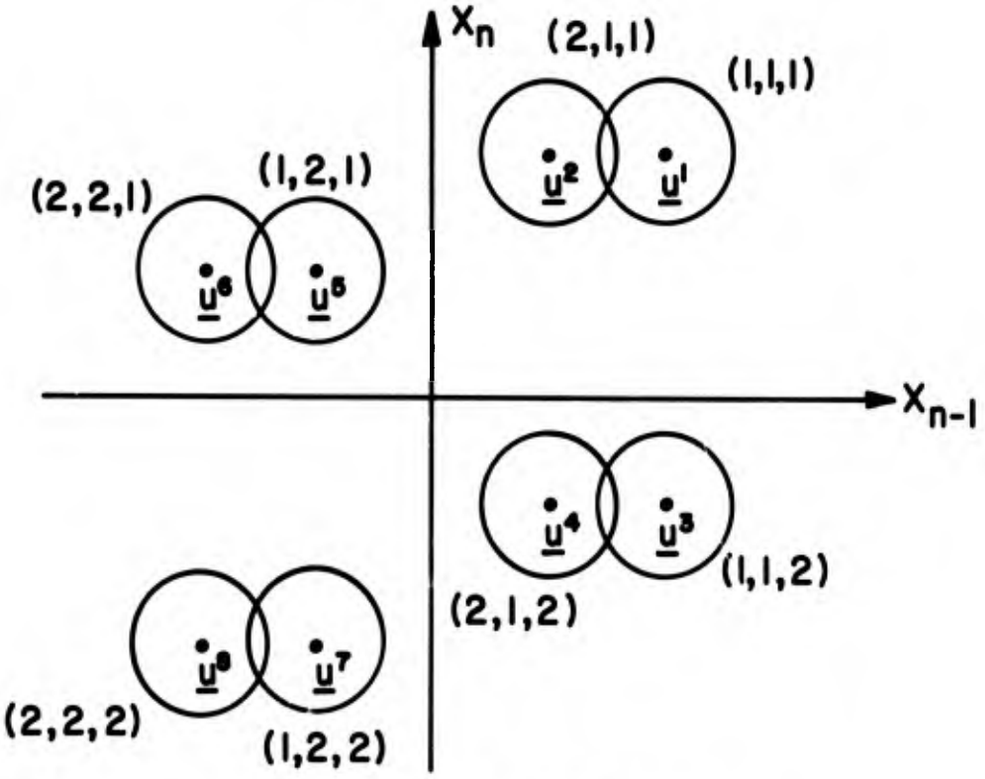


Figure 23. Mode Structure for v = 2, c = 2.

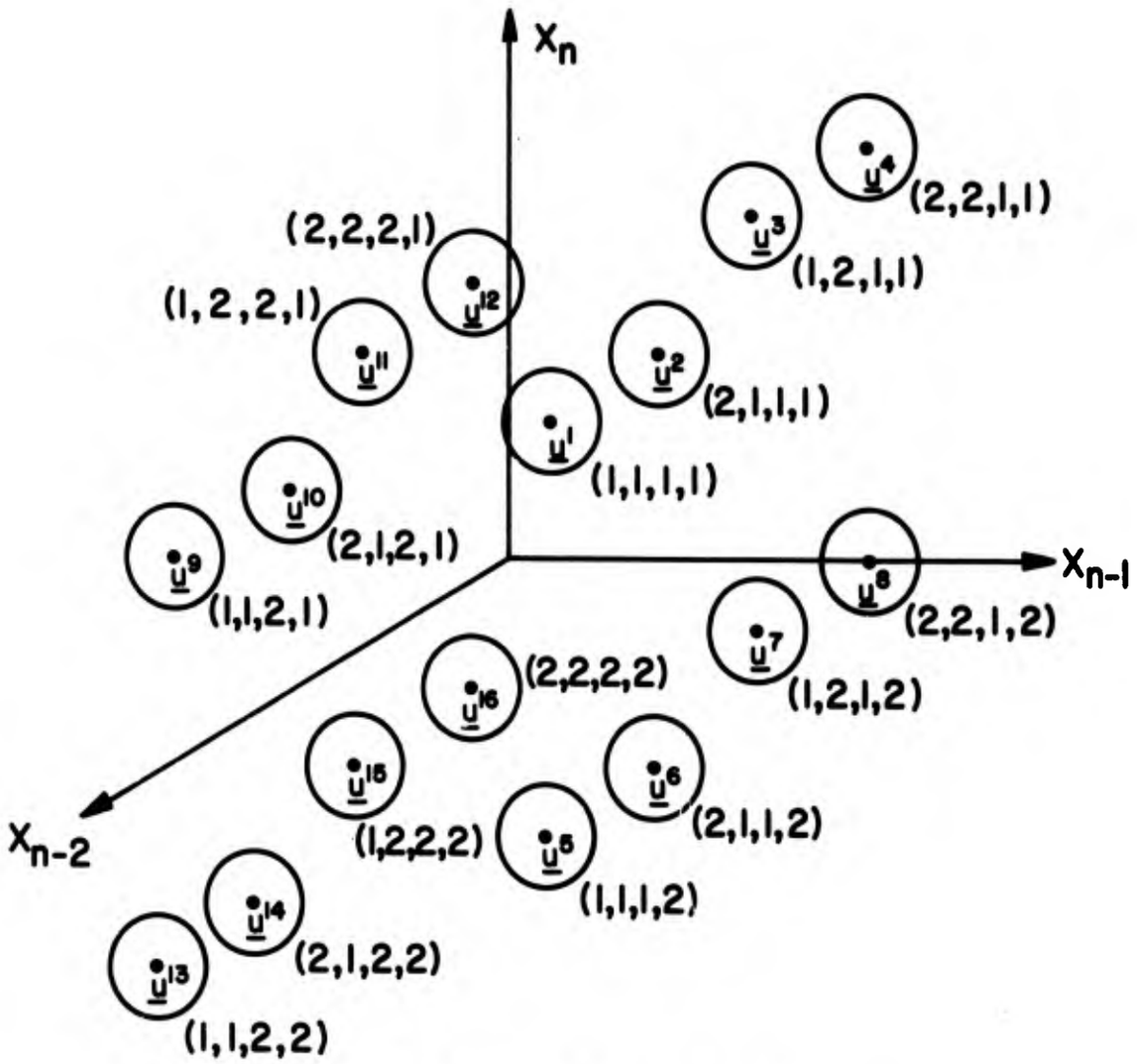


Figure 24. Mode Structure for v = 3, c = 2.

4.2.2 Decision Procedures and the Estimation Problem. The decision procedures discussed in this subsection will be compared using the v = 2, c = 2 mode structure plot of Figure 23. This is the simplest case which shows the distinctions between the various procedures.

An optimum correlator approach[+] assuming v = v* = 1 is a standard decision procedure used in communications. This widespread usage includes the intersymbol interference problem [24], [26], [28]-[30], [32]. The projection property of v = 2 modes onto v = 1 modes discussed in the previous subsection can be used to relate an optimum correlator approach to the v = 2, c = 2 plots. The optimum correlator decision boundary for binary antipodal signals is thus the line perpendicular to $X_{n-1} = 0$ (i.e the X_n axis as shown on Figure 25) for a decision on X_{n-1}, or the line perpendicular to $X_n = 0$ (i.e. the X_{n-1} axis) for a decision on X_n. Both, of course, are completely equivalent because of the symmetry of the v = 1 projections.

The optimum correlator decision boundary shown in Figure 25 goes unnecessarily close to the modes in a v = 2, c = 2 space. Intuitively, a better decision boundary should avoid the modes as much as possible. The following expresses this intuitive concept mathematically. Denote the subset of $\{\underline{u}^k\}_{k=1}^{J}$ having source ω^i active at time n - v* + 1 by $\{\underline{u}^{i_k}\}_{k=1}^{J/2}$. The optimum decision procedure against the v and v* constraints is then to determine which index i maximizes

[+]The optimum linear receiver for v = v* = 1 is refered to here as the optimum correlator.

$$\sum_{k=1}^{J/2} f(\underline{X}_n \mid \underline{u}^{i_k}) P(\underline{u}^{i_k}) \qquad (4.6)$$

where from the equilikely message assumption $P(\underline{u}^k) = \frac{1}{J}$, $k = 1,2,\ldots,J$, and $f(\underline{X}_n \mid \underline{u}^{i_k})$ is a v dimensional Gaussian density having mean vector $\underline{u}^{i_k}$ and covariance matrix $(\sigma)^2$ I. This decision boundary is plotted along the X_n axis in Figure 25 for the case ($v = 2$, $v^* = 2$) where a decision on m_{n-1} is to be made given (X_{n-1}, X_n). The $v^* = 1$ decision boundary has a shape similar to the $v^* = 2$ boundary, but is oriented along the X_{n-1} axis. By observing in Figure 25 that the modes corresponding to ω^1 active at $n - v^* + 1$ are considerably farther away for $v^* = 2$ than for $v^* = 1$ from the modes corresponding to ω^2 active at $n - v^* + 1$, the probability of error for $v^* = 2$ is less than or equal to the probability of error for $v^* = 1$. Since $c = 2$, the conclusion can be made that the better decision procedure ($v^* = 2$) waits until all the energy is transmitted through the channel before making a decision. This conclusion parallels the unconstrained optimum decision procedure discussed in Section 4.1. However, the decision procedure for $v = 2$, $v^* = 1$ still performs better than the optimum correlator ($v = 1$, $v^* = 1$) because of the decision boundary's avoidance of the modes.

A decision procedure described by Aein and Hancock [25] uses a decision feedback scheme which combines some of the discrimination power of a higher dimensional sequence sample space, with the practical advantage of processing sample sequences in a lower

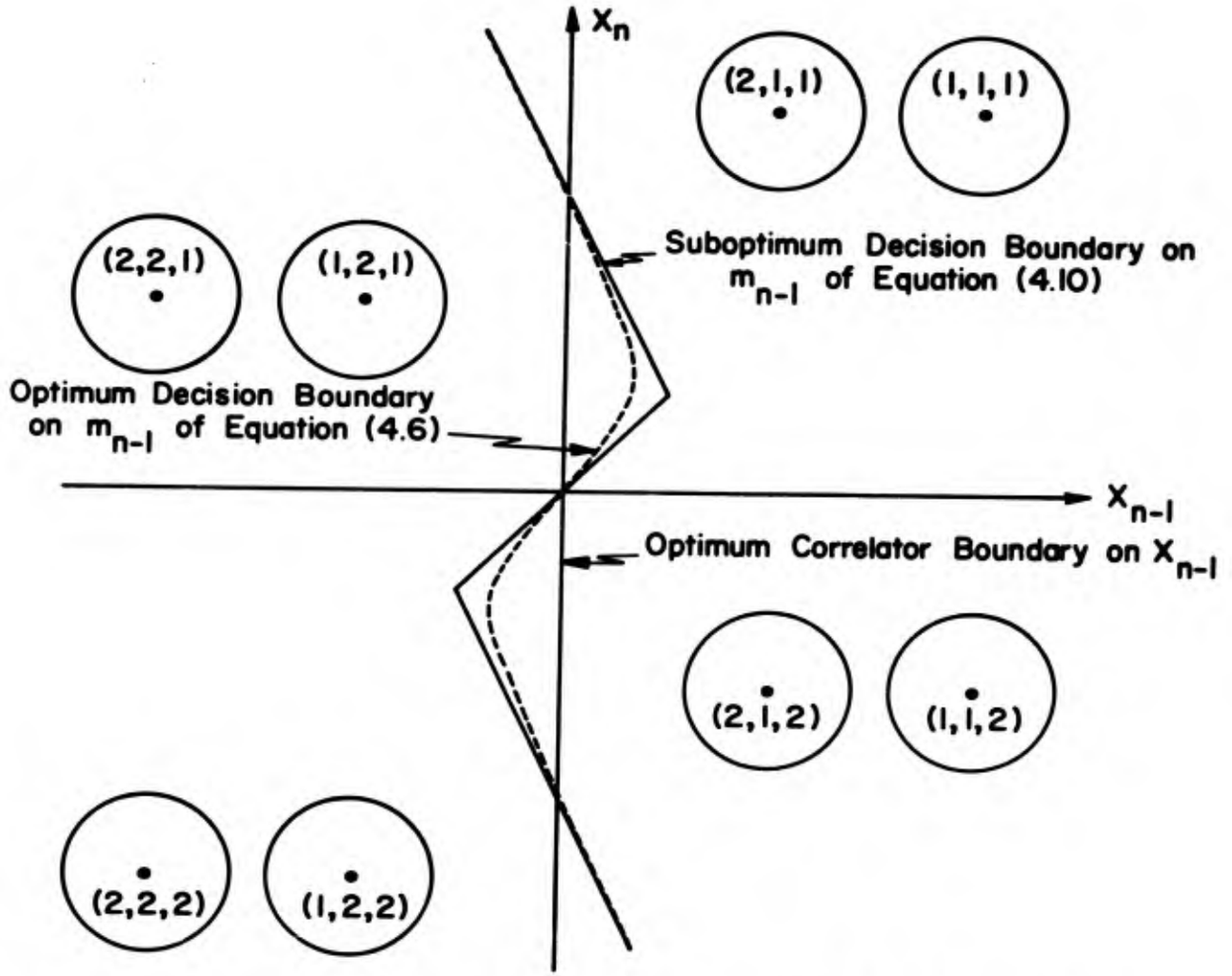


Figure 25. Decision Boundaries and Mode Structure for $v = v^* = c = 2$.

dimensional space. As discussed in Section 4.2, to each mode vector $\underline{u}^k$ there corresponds a c + v - 1 dimensional message event vector which generates it ((4.4)), or corresponding to each possible $\underline{u}_n$ is a message event sequence $(m_{n-v-c+2},\ldots,m_{n-v^*}, m_{n-v^*+1},m_{n-v^*+2},\ldots,m_n)$. At time n, decisions have already been made on $\{m_k\}_{k=1}^{n-v^*}$, and in particular, on $\{m_k\}_{k=n-v-c+2}^{n-v^*}$. The decision feedback scheme makes a decision on m_{n-v^*+1} using only those $\underline{u}^k$ whose corresponding message event sequence matches the decisions made on $\{m_k\}_{k=n-v-c+2}^{n-v^*}$. Decision feedback then uses information on the last v + c - 1 samples in a sequence sample space whose dimensionality is only v.

Figure 26 shows the mode structure for the case $v = v^* = c = 2$ when $m_{n-2} = 1$. The decision boundary for m_{n-1} where the decision on m_{n-2} is correct is shown in Figure 26.a, and the decision boundary resulting from an incorrect decision on m_{n-2} is shown in Figure 26.b. The mode structure and decision boundaries are similar for $m_{n-2} = -1$. Given the a priori knowledge of the value of m_{n-2}, the interclass distances in the sample sequence space (see Figure 26.b) are considerably greater than the interclass distances shown in Figure 25 without this knowledge. Of course the decision on m_{n-2} is not always correct, but if the effect of an incorrect decision is not too great, the decision feedback scheme offers a considerable improvement in performance.

The effect of decision errors made on m_{n-2} can be easily computed for this particular case of $v = v^* = c = 2$. Let $P_e[m_k]$ denote the probability of an incorrect decision on m_k, $P_{e_{CB}}$ the

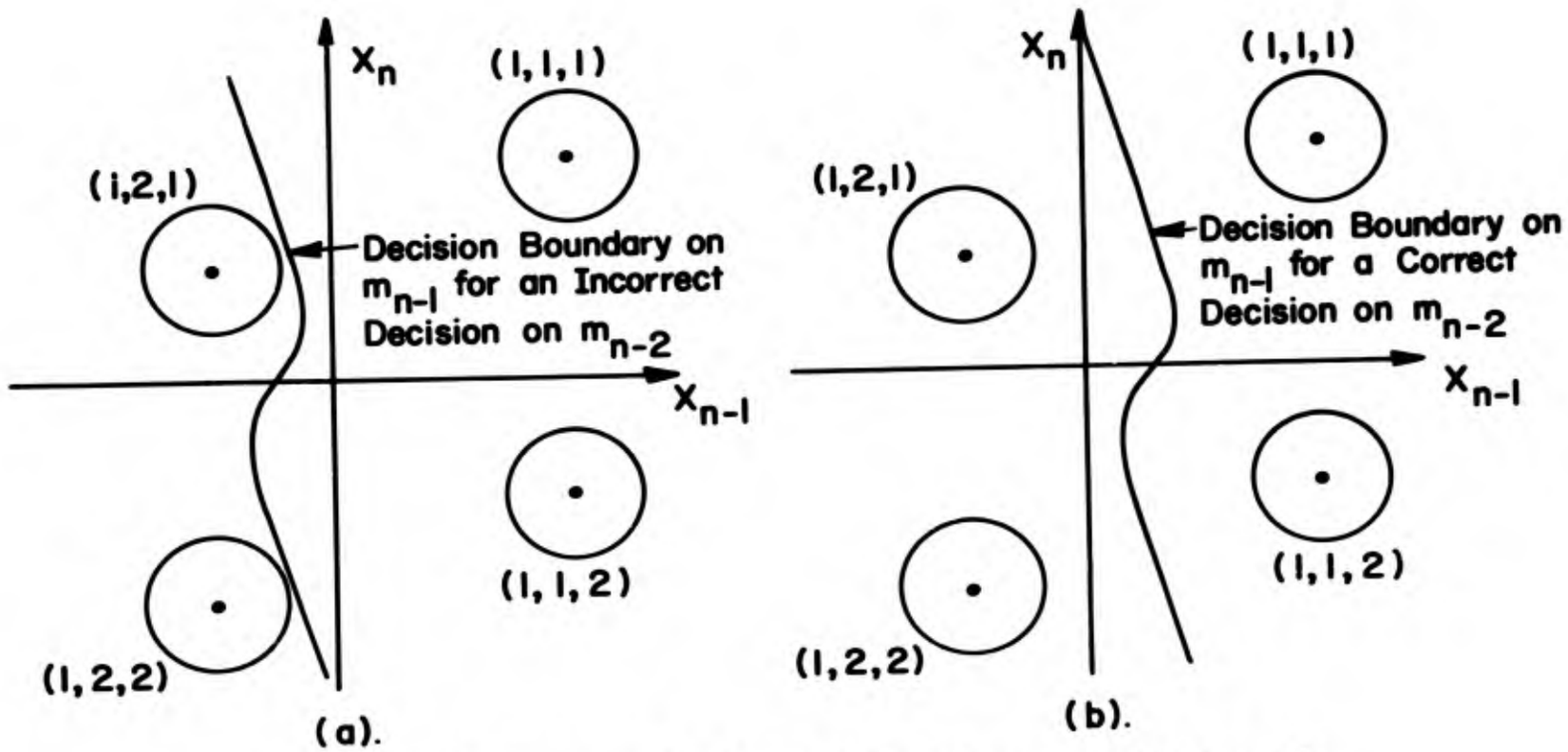


Figure 26. Mode Structure for the Case $v = v^* = c = 2$ when $m_{n-2} = 1$.

probability of error given a correct decision on m_{n-2}, and $P_{e_{WB}}$ the probability of error given a wrong decision on m_{n-2}. These last two quantities are the masses on the wrong side of the respective decision boundaries in Figure 26.a and 26.b.

$$P_e[m_{n-1}] = (1 - P_e[m_{n-2}])P_{e_{CB}} + P_e[m_{n-2}]P_{e_{WB}} \qquad (4.7)$$

Since $P_e[m_{n-1}] = P_e[m_{n-2}]$, then solving for $P_e[m_{n-1}]$,

$$P_e[m_{n-1}] = \frac{P_{e_{CB}}}{1 + P_{e_{CB}} - P_{e_{WB}}} \qquad (4.8)$$

Since $P_{e_{CB}}$ may be a factor of 100 or more smaller than an equivalent decision procedure without decision feedback, $P_{e_{WB}}$ can be relatively large (i.e. almost one) and still have a significant improvement in performance. The disadvantage of decision feedback is that errors will tend to run in batches.

For larger values of $v + c - 1 - v^*$ (the number of decisions being fed back), the errors on $\{m_k\}_{k=n-v-c+2}^{n-v^*}$ will eventually cancel the performance gained by taking subsets of the $\{\underline{u}^k\}$. Perhaps the most important use of the decision feedback approach may be as a complexity reducing method allowing signal estimation and processing at higher values of c than currently practical.

This discussion has indicated that a wide variety of decision procedures is available with varing degrees of complexity

and performance. For most problems of interest[+], the underlying estimation problem is to determine the locations of modes, and enough of the mapping between the modes and the message event sequences to be able to use the chosen decision procedure.

Procedures using a criterion such as (4.6) need knowledge only of the mapping between the estimated $\underline{u}^k$ and the class ω^i active at $n - v^* + 1$ (i.e. one element of the $v + c - 1$ element message event sequence). For example, given a set of estimated mode locations for a $v = v^* = c = 2$ case with $a_1 > |a_2|$, modes having a positive component[++] corresponding to time $n - v^* + 1$ might be assigned to class ω^1 and the rest to ω^2. This, in fact, was done for the simulation of Algorithm 1 in Section 4.5.

Other procedures, such as those using decision feedback, require knowledge of the entire message event sequence that corresponds to the estimated $\underline{u}^k$. Letting $\hat{\underline{u}}^k$ denote for now an estimate of $\underline{u}^k$, one impractical way to establish this mapping is to use the $\{Q^k\}$ matrices defined in Section 4.2. The method is to search through all possible assignments of the $\{Q^k\}$ matrices to the $\{\hat{\underline{u}}^k\}$ vectors until the least squares criterion

---

[+]This is a reference to more general practical problems involving multidimensions or multiclasses which are an extension of these concepts and results.

[++]Refering back to Figure 25, for $a_1 > |a_2|$ modes corresponding to ω^1 active at $n - v^* + 1$ always stays on the right side of the X_n axis, and those corresponding to ω^2 active are always on the left side.

$$\sum_{J=1}^{J'} \left\{ \left[\left(Q^{r_j} \right)^t \left(Q^{r_j} \right) \right]^{-1} Q^{r_j} \hat{\underline{u}}^j - \frac{1}{J'} \sum_{s=1}^{J'} \left[\left(Q^{r_s} \right)^t \left(Q^{r_s} \right) \right]^{-1} Q^{r_s} \hat{\underline{u}}^s \right\}^2 \tag{4.9}$$

is minimized where r_j denotes the j^{th} count of the r^{th} ordering, and only $J' < J$ of the $\{Q^k\}$ matrices having linearly independent columns are used. If $v < c$, $J' = 0$ of course. The message event sequences corresponding to the minimizing assignment of the $\{Q^k\}$ are then assigned to the respective $\underline{u}^k$ vectors. The method is impractical because of the large number of orderings; however, a practical variation is to take two $\hat{\underline{u}}^k$ and apply the same approach. The right hand term in (4.9) is an estimate of $\underline{a}$. Using the $\{Q^k\}$ matrices and (4.5) each mode is regenerated and the closest $\hat{\underline{u}}^k$ is given the sequence assignment belonging to the generating Q^{r_k} matrix. This concept is closely related to Algorithm 2 defined below in Subsection 4.3.2.

Decision directed estimators will be defined in Section 4.3 which will estimate the $\{\underline{u}^k\}$ mode vectors. Hence, a decision equation is needed to relate the sample sequences $\underline{X}_n$ to the modes. The optimum constrained decision equation of (4.6) is inappropriate for use since it gives a decision on ω^i directly. To provide the necessary sample assignments, the suboptimum decision equation

$$d(\underline{X}_n \mid \underline{u}^k) = \begin{cases} 1 : f(\underline{X}_n \mid \underline{u}^k) P(\underline{u}^k) > f(\underline{X}_n \mid \underline{u}^j) P(\underline{u}^j) \text{ all } j \neq k \\ 0 : \text{elsewise} \end{cases} \tag{4.10}$$

is used. The decision on which ω^i was active is determined by which subset $\{\underline{u}^{i_k}\}_{k=1}^{J/2}$ contains the vector maximizing (4.10). Because most errors at reasonably high signal to noise ratios come from the closest $\underline{u}^k$ from each class, the use of this suboptimum decision procedure does not degrade performance very severely in the present application. This decision boundary is illustrated in Figure 25. Curves will be presented in Section 4.4 showing that for $v = v^* = 2$, this criterion gives much better performance than an optimum correlator.

4.3 The Unsupervised Estimation Algorithms

Algorithms from the general decision directed estimator class of Chapter 3 must be modified slightly when applied to the problem of intersymbol interference. The convergence proof of that chapter required statistical independence of the update vectors. This condition can be satisfied in the present application by updating only after every other $r \geq \max(v,c)$ sample bauds. Two algorithms suitable for the intersymbol interference problem are defined in this section. The first algorithm estimates the sequence sample space mode vectors $\{\underline{u}^k\}$ directly using the decision equation of (4.10) to assign sample sequences to the modes. The mapping between the estimated mode vectors and the message event sequences is not specified, but two methods of obtaining the mapping were discussed in the last section. The second algorithm uses the a priori knowledge of the signal set to estimate the channel gain vector $\underline{a}$. Estimates of the $\{\underline{u}^k\}$

mode vectors are constructed using the $\{Q^k\}$ matrices and equation (4.5). Because of this method of generating the mode vector estimates from the $\{Q^k\}$ matrices, the mapping between the estimated mode vectors and the message event sequences is known completely. In both algorithms, estimates of the mode vectors $\{\underline{u}^k\}$ will be denoted $\{\underline{u}_N^k\}$, and in Algorithm 2, the estimate of the channel gain vector $\underline{a}$ will be denoted by $\underline{s}_N$ for reasons that will be discussed later.+

4.3.1 Direct Estimation of the Mode Vectors $\{\underline{u}^k\}$. Since only every $2r^{th}$ sample baud sequence is used in the algorithm, the estimation problem for samples with intersymbol interference is equivalent to determining the mode locations in a conditionally independent, multimodal, multidimensional (if $v > 1$) sample space. This is the estimation problem that the algorithm class defined in the last chapter is supposed to be capable of solving. Using notation similar to that in Chapter 3, the estimator for the v-dimensional vectors $\{\underline{u}^j\}$ is defined:

ALGORITHM 1.

For $2rKN + 1 \leq n \leq 2rK(N + 1)$,

$$\underline{\xi}_k^j = (1 - \beta_k^j)\underline{\xi}_{k-1}^j + \beta_k^j \underline{X}_{2r(KN+k)} \quad k = 1,2,\ldots,K$$
$$1 \leq j \leq J \qquad (4.11)$$

+The motivation for using this definition will be the extension to unknown message signal sets.

where

$$\beta_k^j = \begin{cases} \dfrac{1}{K^j} : d(\underline{X}_{2r(KN+k)} | \underline{u}_N^j) = 1 \\ \qquad \text{and } K^j = K^j + 1 \qquad 1 \le j \le J \\ 0 \quad : \text{elsewise} \end{cases} \tag{4.12}$$

and $\{\underline{\xi}_0^j\}_{j=1}^J \triangleq 0$. The decision procedure $d(\cdot|\cdot)$ used in (4.12) was defined previously in (4.10). Then, at $\underline{n = 2rK(N+1)}$,

$$\underline{u}_{N+1}^j = (1 - \rho_{N+1}^j)\underline{u}_N^j + \rho_{N+1}^j \underline{\xi}_K^j \qquad 1 \le j \le J \tag{4.13}$$

where

$$\rho_{N+1}^j = \begin{cases} \alpha_{w_N^j} : K^j > 0 \\ \qquad \text{and } w_{N+1}^j = w_N^j + 1 \qquad 1 \le j \le J \\ 0 \quad : K^j = 0 \\ \qquad \text{and } w_{N+1}^j = w_N^j \end{cases} \tag{4.14}$$

After the updating is completed, the K^j are reinitialized to zero.

The method of obtaining the starting values is arbitrary, but obviously the more clusters found initially, the easier the algorithm's job. Unfortunately, the computer simulation results discussed in Section 4.5 show that this algorithm's performance is highly dependent on the starting values. In other words, its mode seeking ability in a multimodal space is poor unless the algorithm is started "near" the asymptotic solution.

4.3.2 Implicit Estimation of the Mode Vectors $\{\underline{u}^j\}$ Using the $\{Q^j\}$ Matrices. As discussed in Section 4.2, the $\{Q^j\}$ matrices relate the $\{\underline{u}^j\}$ mode vectors and the channel gain vector $\underline{a}$,

$$\underline{u}^j = Q^j \underline{a} \qquad j = 1,2,\ldots,J \tag{4.15}$$

In order to map back to $\underline{a}$ from the $\{\underline{u}^j\}$, it is necessary that $[(Q^j)^t\ (Q^j)]^{-1}$ exist for at least one value of j, $1 \le j \le J$. This inverse exists only when Q^j has linearly independent columns. Hence, a necessary condition for an inverse to exist is $v \ge c$. In Algorithm 2, the estimated mode vectors $\{\underline{u}_N^j\}$ are generated from the $\{Q^j\}$ matrices and the estimated channel gain vector $\underline{a}_N$ using (4.15). The decision equation defined in (4.10) assigns the sample sequence $\underline{X}_{2r(KN+k)}$ to one of the $\{\underline{u}_N^j\}$ (denote the index by j_k). The corresponding $\{Q^j\}$ matrix with index j_k is then used to map the sample sequence into the channel gain space, and $\underline{a}_N$ is updated if the mapping exists (i.e. if $[(Q^{j_k})^t\ (Q^{j_k})]^{-1}$ exists). A decision on which ω^i caused m_{n-v^*+1} can be made without further computation since the entire $c + v - 1$ message sequence is chosen along with Q^{j_k}.

ALGORITHM 2.

For $2rKN + 1 \le n \le 2rK(N + 1)$,

$$\underline{\varepsilon}_k = (1 - \beta_k)\underline{\varepsilon}_{k-1} + \beta_k \left[\left(Q^{j_k}\right)^t \left(Q^{j_k}\right)\right]^{-1} Q^{j_k} \underline{X}_{2r(KN+k)}$$

$$k = 1,2,\ldots,K \tag{4.16}$$

where

$$\beta_k = \begin{cases} \frac{1}{K^1} : \left[\left(Q^{j_k}\right)^t \left(Q^{j_k}\right)\right]^{-1} < \infty \\ \qquad \text{and } K^1 = K^1 + 1 \\ 0 \ : \text{elsewise} \end{cases} \tag{4.17}$$

$$j_k = \text{index}\left\{ \max_{1 \le j \le J} d(\underline{X}_{2r(KN+k)} | \underline{u}_N^j) \right\} \tag{4.18}$$

and $\xi_0 \triangleq 0$. The decision procedure $d(\cdot|\cdot)$ used in (4.18) was defined previously in (4.10). Then, at $\underline{n = 2rK(N + 1)}$,

$$\underline{s}_{N+1} = (1 - \rho_{N+1})\underline{s}_N + \rho_{N+1}\, \xi_K \tag{4.19}$$

where

$$\rho_{N+1} = \begin{cases} \alpha_{w_N} \ : K^1 > 0 \\ \qquad \text{and } w_{N+1} = w_N + 1 \\ 0 \quad : K^1 = 0 \\ \qquad \text{and } w_{N+1} = w_N \end{cases} \tag{4.20}$$

After updating is completed, K^1 is reinitialized to zero.

The easiest way to obtain a starting value $\underline{s}_1$ for Algorithm 2 (although the method is not used in the computer simulation) is to take the first two samples and apply equation (4.9) as discussed in Subsection 4.2.2. Algorithm 2 discards a sizable fraction of the sample sequences because they fall into regions whose

corresponding $[(Q^{j_k})^t\ (Q^{j_k})]$ inverse matrix does not exist. However, samples used in updating give information about the mode structure of the entire sequence sample space.

4.4 Asymptotic Probability of Error

As discussed in Subsection 4.3.1, Algorithm 1 is equivalent to an algorithm from the class of decision directed estimators proven to converge in Theorem 2. Hence, under the conditions of this theorem, Algorithm 1 converges with probability one and in mean square to a solution of

$$\underline{u}_\infty^k = E[\underline{x} \mid \underline{x} \in S_\infty^k] \tag{4.21}$$

where

$$S_\infty^k \triangleq S^k(\{\underline{u}_\infty^j\}_{j=1}^J)$$

$$= \{\underline{x} : \|\underline{x} - \underline{u}_\infty^k\|^2 < \|\underline{x} - \underline{u}_\infty^j\|^2, \text{ all } j \neq k\} \tag{4.22}$$

Also, each set of solution vectors $\{\underline{u}_\infty^k\}_{k=1}^J$ to (4.21) gives a local minimum of

$$\sum_{k=1}^{J} \int_{S_\infty^k} \|\underline{x} - \underline{u}_\infty^k\|^2 \, h(\underline{x}) d\underline{x} \tag{4.23}$$

While Algorithm 2 is not a member of the class of estimators proven to converge in Theorem 2, a minor modification of the

proof can be made so that it applies to Algorithm 2. Thus, under the conditions stated in Theorem 2, Algorithm 2 converges with probability one and in mean square to a solution of

$$\underline{s}_\infty = E\left[\sum_{j=1}^{J'} \left\{\left[\left(Q^{k_j}\right)^t\left(Q^{k_j}\right)\right]^{-1}\left(Q^{k_j}\right)^t \underline{x}|\underline{x}\in S_\infty^{k_j}\right\}\right] \qquad (4.24)$$

where only those S_∞^k whose corresponding Q^k matrix has $[(Q^k)^t\ (Q^k)]^{-1} < \infty$ are used. Expanding (4.24),

$$\underline{s}_\infty = \frac{\sum_{j=1}^{J'} \left[\left(Q^{k_j}\right)^t\left(Q^{k_j}\right)\right]^{-1}\left(Q^{k_j}\right)^t E\left[\underline{x}|\underline{x}\in S_\infty^{k_j}\right]P\left(S_\infty^{k_j}\right)}{\sum_{j=1}^{J'} P\left(S_\infty^{k_j}\right)} \qquad (4.25)$$

Also, the components in the solution $\underline{s}_\infty$ give a local minimum of

$$\sum_{j=1}^{J'} \int_{S_\infty^{k_j}} \left\|\left[\left(Q^{k_j}\right)^t\left(Q^{k_j}\right)\right]^{-1}\left(Q^{k_j}\right)^t\left(\underline{x} - E\left[\underline{x}|\underline{x}\in S_\infty^{k_j}\right]\right)\right\|^2 h(x)dx \qquad (4.26)$$

The asymptotic probability of error for the case where $v = v^* = c = 2$ was determined for Algorithm 1. The solution to (4.21) was found using an iterative technique [41]. Starting with an initial partition, the average value in each two dimensional region was calculated numerically. This defines a new partition and again the average values were calculated. The procedure is repeated until judged to have sufficently converged. This iterative method is the same as that used in Section 3.5 to

obtain the solution of the M = 2 implicit asymtotic equations. However, in contrast to the M = 2 case which required about ten iterations, the rate of convergence for this eight mode case was extremely slow. To speed up convergence, considerable computer interaction was necessary (i.e. guesses on $\{\underline{u}_N^k\}$ were entered by keyboard). The effect of inaccuracies due to numerical integration is impossible to determine.

Comment: Evaluation of the average vector on each region of the partition as discussed above, is equivalent to letting $K \to \infty$ in Algorithm 1. Since Algorithm 1 is essentially a stochastic approach to the iterative technique of finding a solution, if it takes several hundred iterations (each of which is equivalent to letting $K = \infty$ in the algorithm) to obtain a solution, Algorithm 1 with $K < \infty$ may be expected to have difficulty converging. Similar slow convergence has been noted by Casey and Nagy [52] for their "batch processing" algorithm (where K = size of sample set, $\alpha_k = 1$ in (3.6)-(3.10)).

The definition of signal to noise ratio used for the M = 2 case in Chapter 3 was an extremely meaningful measure because it directly related the interclass distance and the noise in the probability of error expression. Unfortunately, nothing as meaningful is available for a multimodal sample space. The signal to noise ratio used here for the multimodal case is defined,

$$SNR|m^i \triangleq \frac{\text{Total energy from the channel} | m^i \text{ was sent}}{\text{Average noise power}} \qquad i = 1,2 \qquad (4.27)$$

For this chapter's assumptions of binary antipodal signals with values of ±1, and zero mean additive noise, equation (4.27) reduces to

$$SNR = \left(\frac{\|a\|}{\sigma}\right)^2 \tag{4.28}$$

The $v = v^* = c = 2$ asymptotic probability of error using (4.10) of Algorithm 1 vs the signal to noise ratio is presented in Figures 27 and 28 for channel gains $a_1 = .95$, $a_2 = .31$ and $a_1 = .87$, $a_2 = .5$ (i.e. energy overlaps of 10 and 25 percent) respectively. Also, plotted on these figures are the probabilities of error of the optimum correlator and the suboptimum decision equation of (4.10). All results were obtained using numerical integration on the $v = 2$ dimensional sequence sample space. The figures show that at higher SNR values where the modes tend to become separable, the asymptotic estimator is almost indistinguishable from the suboptimum decision equation. Also, as $|\frac{a_2}{a_1}| \to 1$, the optimum correlator performs worse and worse compared to the suboptimum decision equation. At lower signal to noise ratios however, the mode estimators are highly biased. The relative locations of the asymptotic estimators $\{\underline{u}_\infty^k\}$ and the mode vectors $\{\underline{u}^k\}$ are presented in Figure 29 for a signal to noise ratio of 6 and $a_1 = .95$, $a_2 = .31$ (i.e. an energy overlap of 10 percent). The bias is very apparent in this figure.

The asymptotic results also show that at low SNR values, an optimum correlator approach using the a priori knowledge about the signal set performs better than the algorithm. However, communications

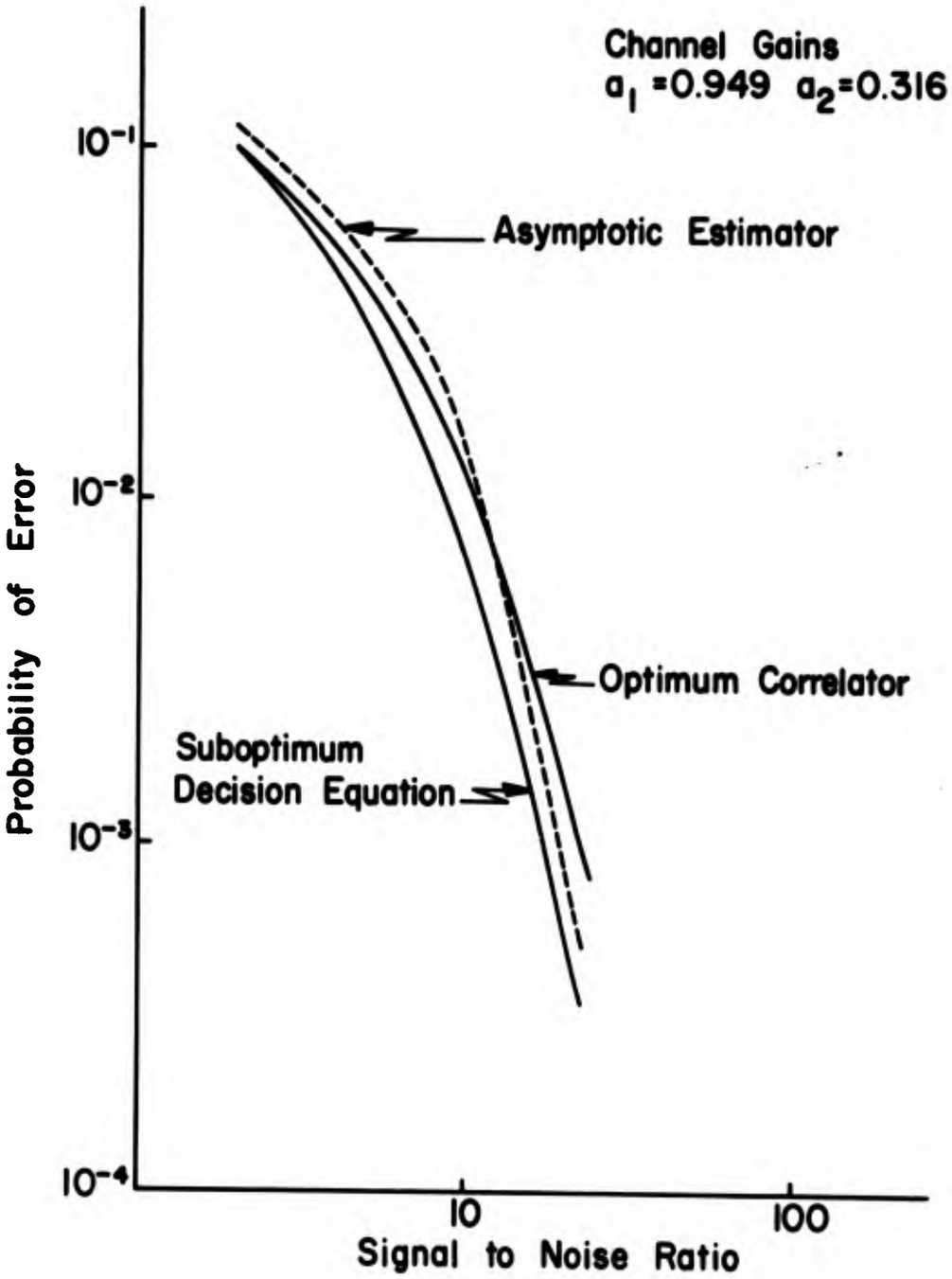


Figure 27. Respective Probabilities of Error vs SNR for Binary Antipodal Signals With v = 2, c = 2, and Channel Gains a_1 = .949, a_2 = .316.

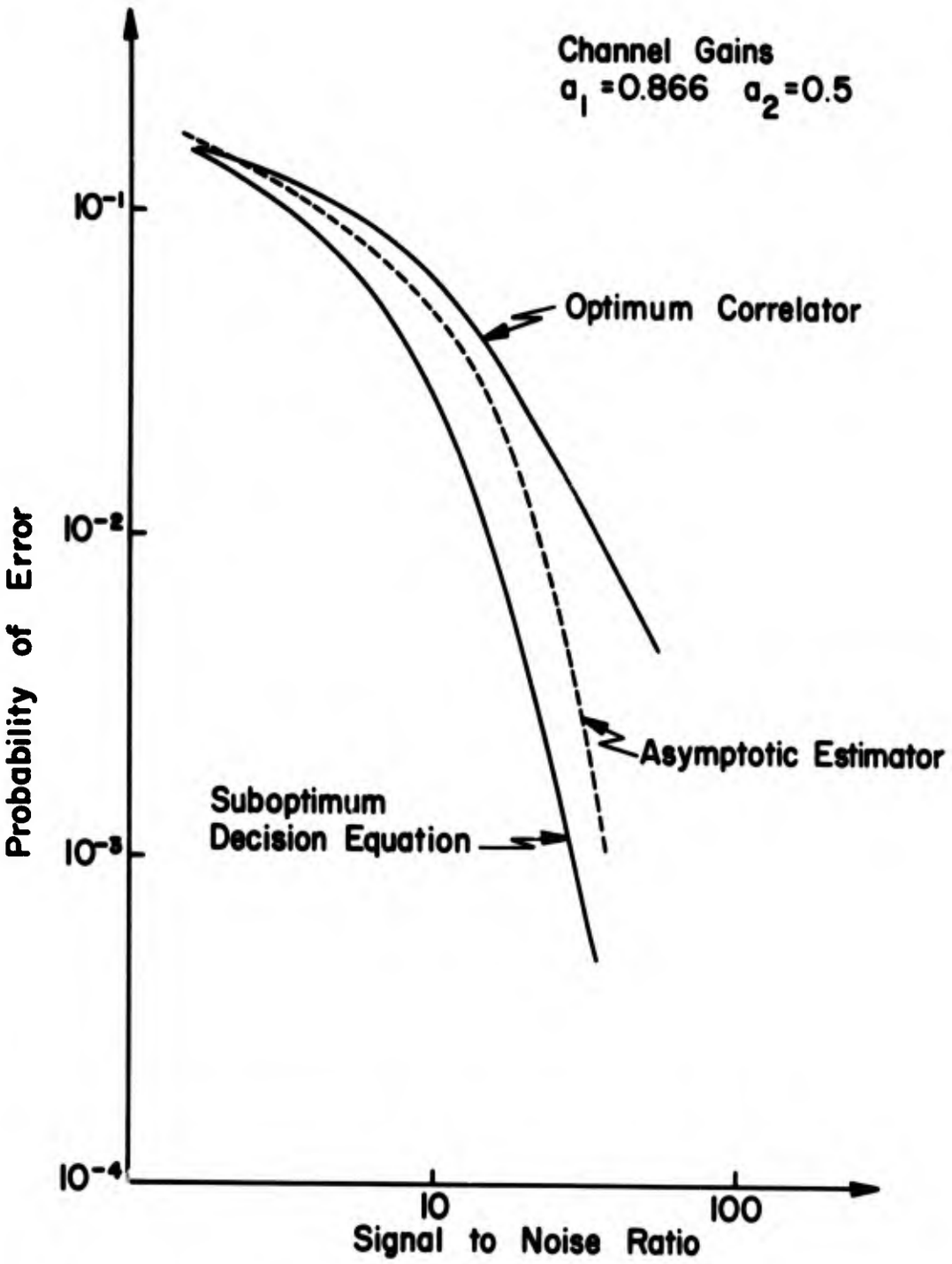


Figure 28. Respective Probabilities of Error vs SNR for Binary Antipodal Signals With v = 2, c = 2, and Channel Gains a_1 = .866, a_2 = .5.

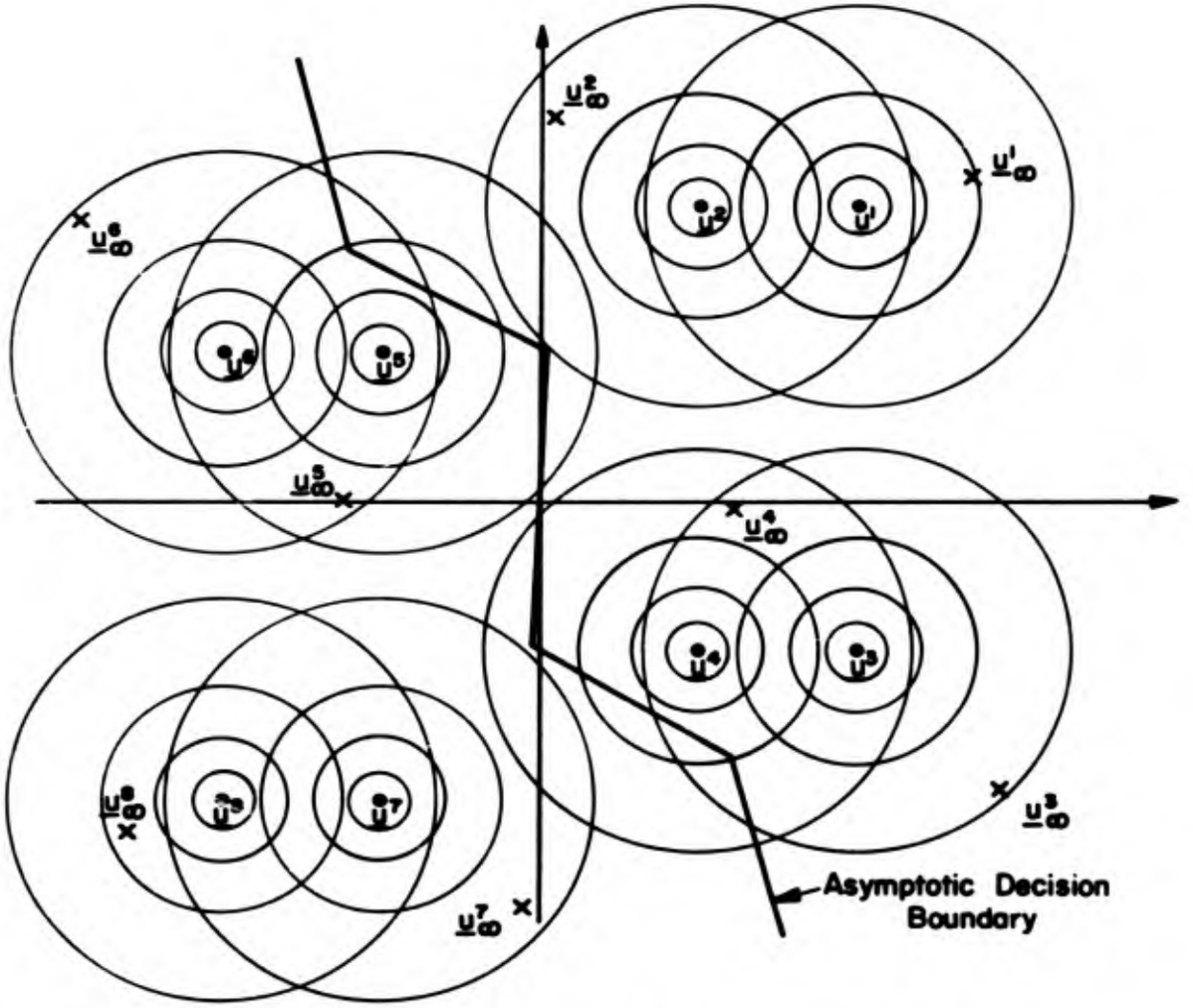


Figure 29. Relative Locations of Asymptotic Estimators $\{\underline{u}_\infty^k\}$ and Mode Vectors $\{\underline{u}^k\}$ for Channel Gains $a_1 = .95$, $a_2 = .3$ at SNR = 6.

systems are usually operated at significantly higher signal to noise ratios, and probabilities of error of 10^{-4} or 10^{-5} may be considered practically minimal performance. At these higher signal to noise ratios, the asymptotic estimator is almost as good as the suboptimum decision equation and offers a significant improvement in performance over an optimum correlator approach.

4.5 Experimental Performance Results

While Algorithms 1 and 2 have very strong asymptotic convergence properties, their actual use will involve processing a finite number of samples. Obviously, it is desirable that the algorithms converge rapidly for small sample sizes. To experimentally establish the two algorithms' dynamic convergence properties, computer simulations were performed for a case where (X_{n-1},X_n) was available to make a decision on m_{n-1}, and the channel had a memory of one baud (i.e. $v = v^* = c = 2$). The channel gains were $a_1 = .95$, $a_2 = .31$ corresponding to a ten percent energy overlap.

The ultimate measure of performance of estimation systems such as these is the average probability of error. Although it was extremely time consuming, the probability of error was determined by numerically integrating the sequence sample space during the computer simulations. In the simulations of the two algorithms, twenty-five experiments were performed at each of the selected signal to noise ratios. For each of the twenty-five experiments the probability of error was calculated at several chosen values of N. The median of the twenty-five experiments is plotted on Figures 30 and 31.

Algorithm 1 was started for each experiment by taking the first eight independent sample sequences and selecting a subset. Note in Figure 23 that the four modes corresponding to ω^2 active at n - 1 can be mapped onto the four modes for ω^1 active at n - 1 by multiplying the ω^2 active mode vectors by (-1). This idea is used in the determination of starting values by testing the second component of each of the eight sample sequences, and if it is negative (i.e. falls into the ω^2 active at n - 1 half of the plane)[+] both components are multiplied by (-1). This procedure has the effect of mapping the the sample sequences into the ω^1 half of the plane. The four farthest apart were selected as the starting values $\{\underline{u}_1^k\}_{k=1}^4$ and the other four estimated mode vectors $\{\underline{u}_1^k\}_{k=5}^8$ were defined $\underline{u}_1^k \triangleq -\underline{u}_1^{8-k+1}$, k = 5,6,7,8. On the average, this method of obtaining vectors appeared to find about five of the eight modes.

The experimental average probability of error found for Algorithm 1 with K = 1 and $\alpha_k = 1/k$ is shown in Figure 30. Not indicated on the figure is the fact that after 250 sample sequence updates Algorithm 1 rarely found modes that it was not started near, and this problem did <u>not</u> improve as the signal to noise ratio <u>was increased</u>. Hence Algorithm 1 is only as good as its starting values, and its use is restricted to such estimation problems as tracking slowly varing statistics which have been initially found by some other means (e.g. supervised starting).

[+]This is, obviously utilizing a correlation decision procedure on which ω^1 was active at time n - 1.

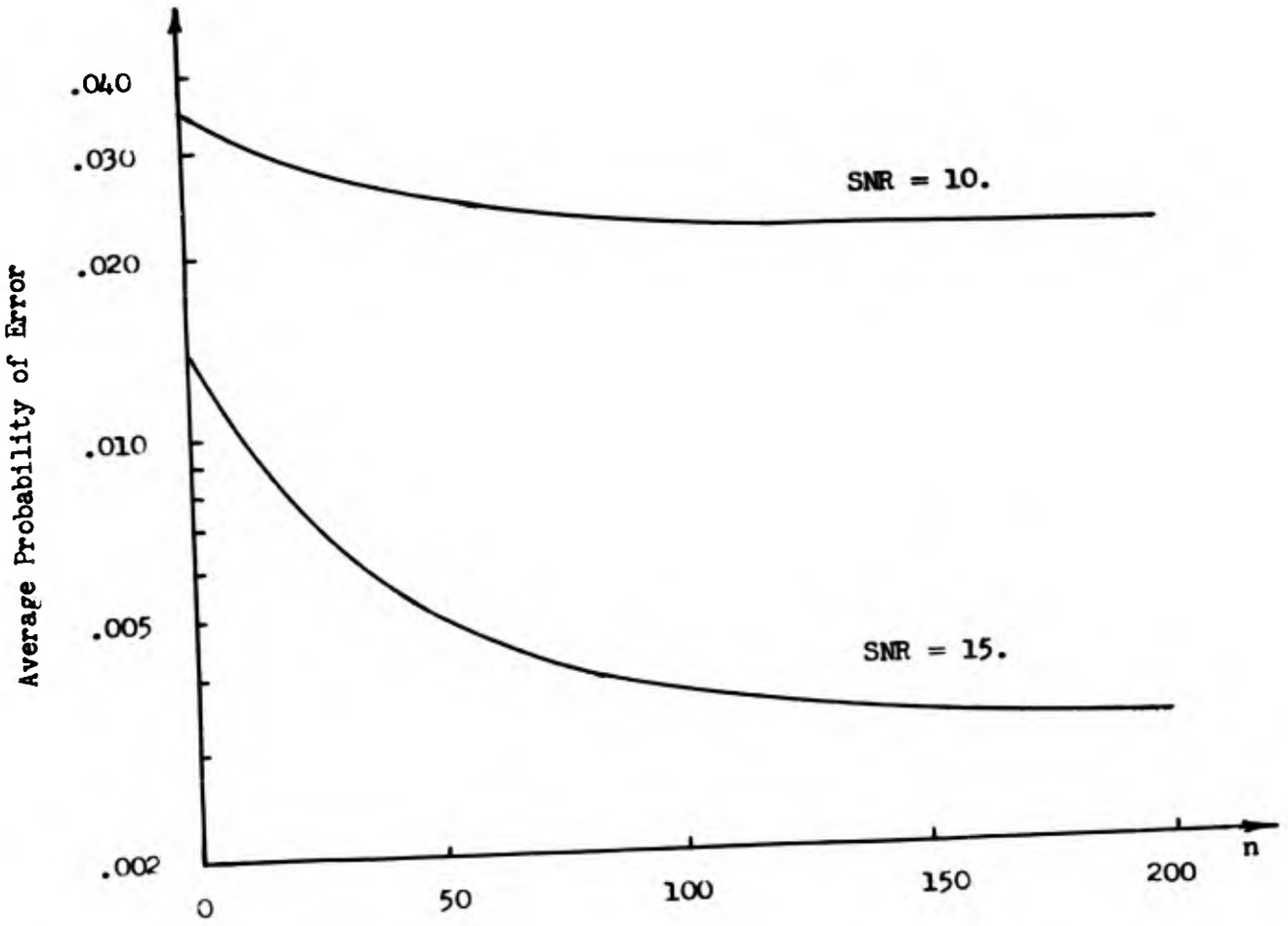


Figure 30. Average Probability of Error vs n for Algorithm 1.

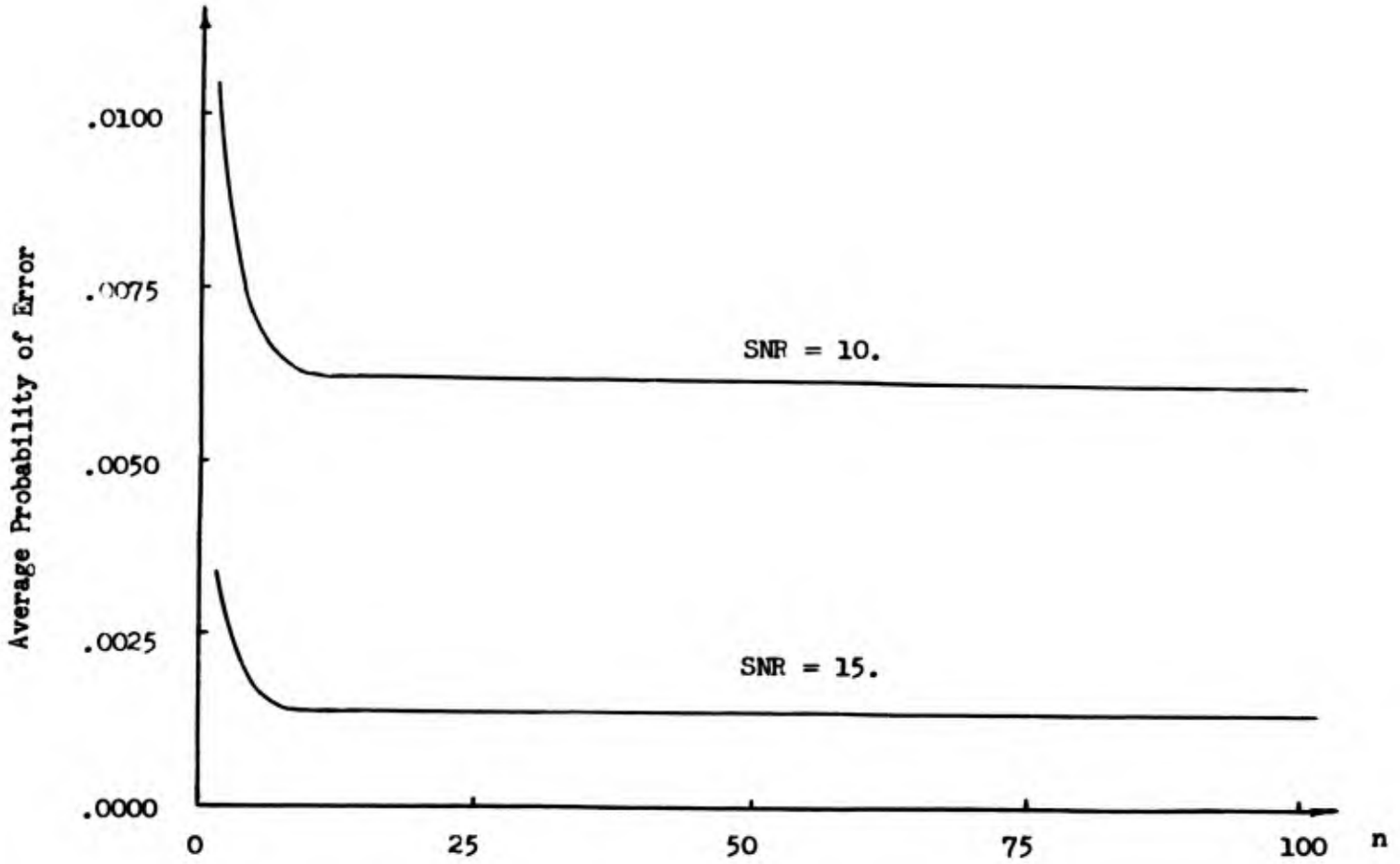


Figure 31. Average Probability of Error vs n for Algorithm 2.

Algorithm 2 was started using the first sample with $\underline{s}_1 \triangleq (\|\underline{x}_1\|, 0)^t$. This initialization is equivalent to assuming the initial estimated decision boundary is that of an optimum correlator. The experimental average probability of error curves will then show how performance improves as the receiver evolves into its asymptotic form.

The experimental average probability of error for Algorithm 2 with $K = 1$ and $\alpha_k = 1/k$ is shown in Figure 31. The algorithm convergence was extremely rapid and most of the twenty-five experiments were almost the same after 100 samples.

Algorithm 2 can be interpreted as "learning" a mapping from the sequence sample space to the channel gain vector space. Every acceptable updating sample sequence from one particular cluster gives information about <u>all</u> the other clusters. Thus, the rapid convergence. The only major weakness of the algorithm is that it does not accept a significant portion of the sequence sample space for updating purposes, and hence, does not utilize all the information that is available in the sample sequences.

4.6 Extensions and Related Problems

The formulation of the intersymbol interference problem in terms of a one dimensional binary antipodal message set which was assumed known to the receiver, greatly simplified the presentation of this chapter. In this section, extensions to other message signal set assumptions and more general linear channel models are discussed. Due to the excellent performance of Algorithm 2, the emphasis of the discussion is on obtaining mappings from the $\{u^k\}$ into either the

channel gain vector space or a related space. The application of Algorithm 2 is then immediate and will not be discussed further. Also presented in this section is a discussion of two problems whose statistical structure is closely related to that of the intersymbol interference problem. These two problems provide additional examples of how knowledge of the sample space mode structure can be used in obtaining a solution to an unsupervised estimation problem.

4.6.1 Extensions. Processing in the presence of non-white noise and providing for delay in the channel are two obvious extensions to the problem formulated in Section 4.1. The intersymbol interference receiver processes sample sequences in a $v\ell$ dimensional sequence sample space. Suppose the conditional density of $\underline{X}_n$ given $\underline{u}^k$ is active has mean vector γ^k and covariance matrix Φ^k. The decision equation of (4.10) is then modified to

$$d(\underline{X}_n|\underline{u}^k) = \begin{cases} 1 : (\underline{X}_n - \gamma^k)^t (\Phi^k)^{-1} (\underline{X}_n - \gamma^k) \\ \quad < (\underline{X}_n - \gamma^j)^t (\Phi^j)^{-1} (\underline{X}_n - \gamma^j) \text{ all } j \neq k \\ 0 : \text{elsewise} \end{cases} \tag{4.29}$$

If $\gamma^k \neq \underline{u}^k$ for any k, the $\{Q^k\}$ matrices no longer contain the mutual constraints on the mode locations and Algorithm 2 no longer applies to the problem. A delay of j bauds can be included in the channel model by changing the channel transformation of (4.1) to

$$u_n = \sum_{k=1}^{n-j} a_{n-j-k+1} m_k \tag{4.30}$$

A similar change in interpretation of the times of the message sequence events must be made, but the basic mode structure of the problem remains the same. It might be noted here that generally, if baud synchronization is maintained, the relative u_j and m_k count subscripts are irrelavent to the processing of the samples.

Methods of obtaining a mapping from the $\{u^k\}$ into the channel gain vector space or a similar space for different linear channel models and message signal assumptions are considered below. The results are for the multidimensional signal vector case ($\ell > 1$) with the message signal basis set of ℓ functions assumed known to the receiver[+]. Denote the ℓ dimensional message signal active at time n by

$$m_n \triangleq (m_{n,1}, m_{n,2},\ldots,m_{n,\ell})^t \tag{4.31}$$

and use similar notation for the ℓ dimensional vectors u_n, X_n, and n_n. Also, define the ℓ dimensional message signal vector corresponding to source ω^i as

$$m^i \triangleq (m^{i,1}, m^{i,2},..,m^{i,\ell})^t \tag{4.32}$$

for i = 1,2,...,M, the k^{th} indexed ℓ dimensional channel output vector

---

[+]In this discussion all signals whose actual representations in the system are waveforms (e.g. m_n,u_n) will be expressed in terms of a vector of coefficients of the known basis set.

$$u^k \triangleq (u^{k,1}, u^{k,2}, \ldots, u^{k,\ell})^t \tag{4.33}$$

and finally the $v\ell$ dimensional vectors

$$\underline{u}_n = \begin{pmatrix} \underline{u_n} \\ \underline{u_{n-1}} \\ \vdots \\ \underline{u_{n-v+1}} \end{pmatrix} \qquad \underline{X}_n = \begin{pmatrix} \underline{X_n} \\ \underline{X_{n-1}} \\ \vdots \\ \underline{X_{n-v+1}} \end{pmatrix} \tag{4.34}$$

This notation is a slight modification to multidimensional vectors of the notation defined in Section 4.1 and 4.2 for the one dimensional case.

Case 1. A Convolutional Channel and Pulse Coded Modulation.

The ℓ dimensional v-baud channel transformation model under these assumptions can be expressed in matrix form,

$$\begin{pmatrix} u_{n,1} \\ \vdots \\ u_{n,\ell} \\ u_{n-1,1} \\ \vdots \\ u_{n-v+1,\ell} \end{pmatrix} = \begin{pmatrix} m_{n,1} & m_{n,2} & \cdots & m_{n-c+1,\ell} \\ m_{n,2} & m_{n,3} & & m_{n-c,1} \\ \vdots & & \ddots & \vdots \\ m_{n-v+1,\ell} & m_{n-v,1} & & m_{n-v-c+1,\ell-1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{c\ell} \end{pmatrix} \tag{4.35}$$

or defining the $(v\ell \times c\ell)$ matrix of message signal components as Q_n and the $c\ell$ dimensional vector of channel gains as $\underline{a}$,

$$\underline{u}_n = Q_n \underline{a} \tag{4.36}$$

This is the same form as (4.5) which relates the $\underline{u}$ vectors to the vector of channel gains for the one dimensional message signal case. If there are M message vectors, there are exactly M^{c+v} unique Q^k matrices generated by the totality of message event sequences for $\ell > 1$, and M^{c+v-1} matrices for $\ell = 1$.

If the message signal vectors $\{m^i\}_{i=1}^{M}$ are <u>known</u> to the receiver the values can be substituted into the different Q^k matrices. This means the Q^k are known and Algorithm 2 can be applied with no changes other than to the ℓ dimensional baud space. The only additional difficulty of $\ell > 1$ is that the computation of the $(v\ell \times v\ell)$ $[(Q^k)^t\ (Q^k)]^{-1}\ Q^k$ matrices will probably have roundoff error problems as ℓ becomes large.

If the message signal vectors are <u>unknown</u> to the receiver, a slightly different approach can be used to obtain the inverse mapping. Equation (4.35) can be re-expressed ($\delta(\cdot)$ is the indicator function),

$$\begin{pmatrix} u_{n,1} \\ \vdots \\ u_{n-v+1,\ell} \end{pmatrix} =$$

$$\left[\begin{pmatrix} m_{n,1}\,\delta(\|m_n - m^1\|) & \cdots & m_{n-c+1,\ell}\,\delta(\|m_{n-c+1} - m^1\|) \\ \vdots & \ddots & \vdots \\ m_{n-v+1,\ell}\,\delta(\|m_{n-v+1} - m^1\|) & \cdots & m_{n-v-c+1,\ell-1}\,\delta(\|m_{n-v-c+1} - m^1\|) \end{pmatrix}\right.$$

$$\left. + \begin{pmatrix} m_{n,1}\,\delta(\|m_n - m^2\|) & \cdots & m_{n-c+1,\ell}\,\delta(\|m_{n-c+1} - m^2\|) \\ \vdots & \ddots & \vdots \\ m_{n-v+1,\ell}\,\delta(\|m_{n-v+1} - m^2\|) & \cdots & m_{n-v-c+1,\ell-1}\,\delta(\|m_{n-v-c+1} - m^2\|) \end{pmatrix}\right.$$

$$+\ldots+\begin{pmatrix} m_{n,1}\,\delta(\|m_n - m^M\|) & \ldots m_{n-c+1,\ell}\,\delta(\|m_{n-c+1} - m^M\|) \\ \vdots & \ddots \\ m_{n-v+1,\ell}\,\delta(\|m_{n-v+1} - m^M\|) & \ldots m_{n-v-c+1,\ell-1}\,\delta(\|m_{n-v-c+1} - m^M\|) \end{pmatrix}\Bigg]$$

$$\cdot\begin{pmatrix} a_1 \\ \vdots \\ a_{c\ell} \end{pmatrix} \tag{4.37}$$

Expanding the k^{th} matrix in the above equation,

$$\begin{pmatrix} m_{n,1}\,\delta(\|m_n - m^k\|) & \ldots m_{n-c+1,\ell}\,\delta(\|m_{n-c+1} - m^k\|) \\ \vdots & \ddots \\ m_{n-v+1,\ell}\,\delta(\|m_{n-v+1} - m^k\|) & \ldots m_{n-v-c+1,\ell-1}\,\delta(\|m_{n-v-c+1} - m^k\|) \end{pmatrix}\begin{pmatrix} a_1 \\ \vdots \\ a_{c\ell} \end{pmatrix}$$

$$= \begin{pmatrix} \delta(\|m_n - m^k\|) & \ldots \delta(\|m_{n-c+1} - m^k\|) \\ \vdots & \ddots \\ \delta(\|m_{n-v+1} - m^k\|) & \ldots \delta(\|m_{n-v-c+1} - m^k\|) \end{pmatrix}$$

$$\cdot\begin{pmatrix} m^{k,1} & m^{k,2} \ldots m^{k,\ell} \\ m^{k,2} & m^{k,3} \ldots m^{k,1} \\ \vdots & \vdots \\ m^{k,\ell} & m^{k,1} \ldots m^{k,\ell-1} \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{c\ell} \end{pmatrix} \tag{4.38}$$

$$= Q_n(m^k)\underline{s}^k \tag{4.39}$$

where $Q_n(m^k)$ defined as the first matrix on the right side of the equal sign in (4.38) is a ($v\ell$ x $c\ell$) matrix of zeros and ones, and $\underline{s}^k$ is the vector resulting from the mapping of the channel gain vector by the matrix of $[m^{k,j}]$ elements. Thus, in general,

$$\underline{u}_n = Q_n(m^1)\underline{s}^1 + Q_n(m^2)\underline{s}^2 + \ldots + Q_n(m^M)\underline{s}^M \qquad (4.40)$$

As before for $\ell > 1$, there are only M^{v+c} unique $(Q_n(m^1),\ldots,Q_n(m^M))$ matrix sets that satisfy (4.40)$^+$. A mapping between the $\{\underline{u}^k\}$ and the $\{\underline{s}^i\}$ can be obtained by taking a sequence of the last j of the $\underline{u}_k$,

$$\begin{bmatrix} \underline{u}_n \\ \underline{u}_{n-1} \\ \vdots \\ \underline{u}_{n-j+1} \end{bmatrix} = \begin{bmatrix} Q_n(m^1) & Q_n(m^2) & \ldots Q_n(m^M) \\ Q_{n-1}(m^1) & Q_{n-1}(m^2) & \ldots Q_{n-1}(m^M) \\ \vdots & & \vdots \\ Q_{n-j+1}(m^1) Q_{n-j+1}(m^2) \ldots & & Q_{n-j+1}(m^M) \end{bmatrix} \begin{bmatrix} \underline{s}^1 \\ \underline{s}^2 \\ \vdots \\ \underline{s}^M \end{bmatrix} \qquad (4.41)$$

Thus, we have been able to reduce the problem again to a set of known matrices which can be used to map the $\{\underline{u}^k\}$ into the $\{\underline{s}^i\}$ vector space. Algorithm 2 will obviously have to be modified slightly to utilize the different form of the relationship between the $\{\underline{u}^k\}$ and the $\{\underline{s}^i\}$.

Case 2. Diagonal Channel Matrices and Arbitrary Signals

The channel transformation model for these assumptions can be expressed,

$$u_n = \sum_{k=1}^{n} A_{n-k+1}\, m_k \qquad (4.42)$$

$^+$For $\ell = 1$ and binary antipodal signals, (4.40) becomes $\underline{u}_n = Q_n(m^1)m^1\underline{a} + Q_n(m^2)m^2\underline{a} = Q_n(m^1)\underline{a} - Q_n(m^2)\underline{a} = Q_n\underline{a}$. This is the same as (4.5).

where A_k is an (ℓ x ℓ) diagonal matrix of channel gains with $(i,i)^{th}$ element a_{k+j-1}. This channel transformation model implies that energy applied to one of the basis functions has no effect on the energy received on any of the others. An example of such a case is frequency shift keying (FSK) through a linear channel. Expanding (4.42),

$$\begin{pmatrix} u_{n,1} \\ u_{n,2} \\ \vdots \\ u_{n,\ell} \end{pmatrix} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & a_\ell \end{pmatrix} \begin{pmatrix} m_{n,1} \\ m_{n,2} \\ \vdots \\ m_{n,\ell} \end{pmatrix}$$

$$+ \cdots + \begin{pmatrix} a_{\ell(c-1)+1} & 0 & & \cdots 0 \\ 0 & & a_{\ell(c-1)+2} & \cdots 0 \\ \vdots & & & \ddots \vdots \\ 0 & & \cdots & a_{c\ell} \end{pmatrix} \begin{pmatrix} m_{n-c+1,1} \\ m_{n-c+1,2} \\ \vdots \\ m_{n-c+1,\ell} \end{pmatrix}$$

$$= \begin{pmatrix} m_{n,1} & & 0 & \vdots & \vdots & m_{n-c+1,1} & & 0 \\ & m_{n,2} & & \vdots & \vdots & & m_{n-c+1,2} & \\ & & \ddots & \vdots & \vdots & & & \ddots \\ 0 & & & m_{n,\ell} & 0 & & & m_{n-c+1,\ell} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{c\ell} \end{pmatrix} \tag{4.43}$$

Hence, $\underline{u}_n$ can be expressed,

$$\begin{pmatrix} u_{n,1} \\ u_{n,2} \\ \cdot \\ \cdot \\ \cdot \\ u_{n-v+1,\ell} \end{pmatrix} =$$

$$\begin{pmatrix} m_{n,1} & & 0 & & m_{n-c+1,1} & & 0 \\ & m_{n,2} & & & & m_{n-c+1,2} & \\ & & \ddots & & & & \ddots \\ 0 & & & m_{n,\ell} & 0 & & & m_{n-c+1,\ell} \\ \hline & \vdots & & & & \vdots & \\ \hline m_{n-v+1,1} & & 0 & & m_{n-v-c+2,1} & & 0 \\ & m_{n-v+1,2} & & & & m_{n-v-c+2,2} & \\ & & \ddots & & & & \ddots \\ 0 & & & m_{n-v+1,\ell} & 0 & & & m_{n-v-c+2,\ell} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_{c\ell} \end{pmatrix}$$

(4.44)

which is the same form as (4.36). Thus, the results in Case 1 can be applied to these assumptions except here there are M^{c+v-1} unique Q^k matrices for all $0 < \ell < \infty$.

Case 3. Arbitrary Channel Matrices and Signals.

The channel transformation equation utilizing these assumptions is

$$u_n = \sum_{k=1}^{n} A_{n-k+1}\, m_k \qquad (4.45)$$

where the $\{A_k\}$ are unknown $\ell \times \ell$ matrices of channel gains. The structure of this problem will not allow the definition of a set of $\{Q^k\}$ matrices so a direct approach is used. Equation (4.45) can be written,

$$(u_n \cdots u_{n-v+1}) = (A_1 \cdots A_c) \begin{pmatrix} m_n & \cdots & m_{n-v+1} \\ \vdots & & \\ m_{n-c+1} & \cdots & m_{n-v+c+2} \end{pmatrix} \quad (4.46)$$

If there are M message signals of dimension ℓ, this gives $v\ell$ equations in $c\ell^2$ unknown if the message signal set is known, and $v\ell$ equations in $c\ell^2 + M\ell$ is the signal set is unknown. In either case, the complexity of this approach indicates that this problem is not well suited to unsupervised estimation using an algorithm such as Algorithm 2.

4.6.2 Related Problems. This subsection illustrates the use of a problem's mode structure in obtaining a solution. The underlying statistical structure of the two problems discussed -- bandpass signals with unknown carrier phase, and unknown synchronization -- is similar to that of the intersymbol interference problem.

For the case where $c = 1$, (i.e. no intersymbol interference) the minimum probability of error of any binary communications system can only be attained using antipodal signals or phase shift keying (PSK). The optimum detector for a PSK system requires complete knowledge of the received bandpass signals, including the phase of the RF waveform. Unfortunately in many practical applications, the received carrier phase is not known in advance to the receiver, and

phase coherent reception is not possible. This discussion will show that a standard approach (differential phase shift keying) for handling the unknown carrier phase problem is related to a special case of Algorithm 2 applied to unknown synchronization.

One approach to obtain the unknown phase is to transmit a phase reference signal along with the message signal. However, this has the disadvantage for a fixed transmitter power output of reducing the energy in the message signal. A differential phase shift keying system (DPSK) implements the concept of a transmitted phase reference by using the last received signal as the reference [62]. For high signal to noise ratios and slow drifts in RF phase, a DPSK system performs almost as well as an optimum coherent binary communications system using the same antipodal message set. Conventional DPSK systems require supervised starting (the first transmitted signal is a phase reference) and special differential encoders and decoders.

The signal to noise ratio of the phase reference used by the receiver can be improved by utilizing a decision directed approach on past sample vectors. The phase reference is formed as a weighted sum: each term in the sum is a past sample multiplied by ± 1 depending on whether its decision is the same as the decision for the sample processed previous to the current observation. In [63] this approach with an exponential weighting on the last ten samples yielded a moderate improvement in performance over a one sample phase reference.

Suppose the duration of one sample baud is T seconds and is known to the receiver, and for convenience, denote the times at

which the received bauds are assumed to start as kT, k = 0, $\pm 1, \pm 2, \ldots$. If in addition to an unknown RF phase the exact arrival time of the received signals is not known (i.e. the bauds do not start at the times kT, k = $0, \pm 1, \pm 2, \ldots$), there are two signals active on the same [kT,(k + 1)T] times interval. Although there is no overlap of the received signals, since there are two signals active on [kT, (k + 1)T], the multibaud mode structure for this unknown synchronization problem is the same as that for intersymbol interference with c = 2. We assume the finite signal basis function set used has the property that the norm of the received signal vector representation is relatively invariant (and nonzero) under RF carrier phase shift. If the mapping between this vector representation on [k, (k + 1)T] and the received signal waveform phase and time shift is continuous, a modification of Algorithm 2 can be used to track nonstationary phase shifts in the multibaud sequence sample space. For example, letting [$0 < K < \infty$, $\alpha_k = 1$] in Algorithm 2, we obtain a "tracking mode" [23]. Another possible choice is to use [$K = 1, \alpha_k = \epsilon$ where $0 < \epsilon < 1$]; this gives an exponential weighting roughly similar to [63]. With such modifications the algorithm can follow small RF phase changes much the same way as a conventional DPSK system. However, the use of a multibaud sample space approach with $v \geq 2$ has the advantage that the cumulative effect of nonstationary changes can be determined and used to maintain synchronization. Also, if T is known well enough and the phase shifts are small enough that the mode locations in the sequence sample space are quasi stationary relative to the convergence time of Algorithm 2, the multibaud sample space can be used to initially synchronize the receiver.

V. CONCLUSIONS

5.1 Summary and Conclusions

In Chapter 2 the asymptotic properties of the Bayes estimator on a discretized parameter space $\mathcal{B}^{M'}$ were established. Such a direct application of the a posteriori approach is practical for one parameter problems such as ECM frequency detection and sonar target bearing estimation. Conditions were given for which the Bayes estimator is super efficient (i.e. mean square convergence faster than $O(\frac{1}{n})$). Asymptotically, the Bayes estimator maximizes a measure of information $\eta(B) \triangleq E[h(x|B)]$ for $B \in \mathcal{B}^{M'}$. This was a worthwhile result in itself, but by noting the relationship between $\eta(B)$ defined on a finite $\mathcal{B}^{M'}$, and $\eta(B)$ on a continuous space, a new optimization criterion was defined. The parameter vector $B \in \mathcal{B}^{M'}$ that maximizes $\eta(B)$ defines the minimum risk solution relative to the parametric family of densities assumed in $\mathcal{B}^{M'}$.

While $\eta(B)$ is difficult to evaluate in general, it defines a regression surface whose maximum can be found using stochastic approximation algorithms. It may be possible to show that under reasonable conditions on the regression surface, the stochastic approximation algorithms based on the Keifer-Wolfowitz and Robbins-Munro procedures defined in Section 2.4 converge and thus are asymptotically optimum. Unfortunately, gradient based stochastic

approximation algorithms such as these usually converge fairly slowly. In the present application, the algorithms must estimate multimodal sample space statistics. Since the parameter vectors move at only $O(\alpha_k)$, the appearance of a new class after many samples have been processed will have only a slight effect on the estimated parameter vector. The basic problem is that these two algorithms do not take advantage of any separability in the sample space.

The assumption that the true mixture can be expressed in terms of a separable Gaussian family led to an easily evaluated form of $\eta(B)$ in Section 2.5. An algorithm designed to maximize this form of $\eta(B)$ by performing the update that maximizes the estimate of $\eta(B)$ was defined in Subsection 2.5.2. It was interesting to note in Section 2.6 that the separable Gaussian $\eta(B)$ criterion could resolve mixtures with considerable overlap despite the violation of the separability assumption. Also, in this section contours of $\eta(B)$ for a two class Gaussian problem were evaluated numerically. The $\eta(B)$ contours for this problem were not ellipsoidal or any other common form indicating that an optimum recursive[+] Bayes estimator for unsupervised estimation is unlikely to exist.

A class of decision directed estimators was defined in Chapter 3 unifying several previous papers. The algorithms have the same form as conventional stochastic approximation algorithms and were given the interpretation of stochastic approximation algorithms with

---

[+]By recursive it is meant that the estimator at stage n is a known function of the estimators at stages n - 1, n - 2,...,1.

random weights. The class of algorithms seek a minimum of a criterion derived from the separable Gaussian $\eta(B)$ criterion. Except for the case where $P(\alpha^k) = \frac{1}{M}$ $k = 1,2,\ldots,M$ the criterion of this chapter yields a solution which is suboptimum. Also, as discussed in Chapter 3 the criterion can not be used to determine the value of M if this is unknown. The entire class of algorithms is proven to converge with probability one in Theorem 2.

Three subclasses of M = 2 algorithms were defined: Subclass A--$P(\gamma^i)$ $i = 1,2$ unknown to be estimated; Subclass B--$P(\gamma^i)$ $i = 1,2$ assumed 1/2; Subclass C--$P(\omega^i)$ $i = 1,2$ and on ordering on $(\gamma^1,\gamma^2)\in(R^\ell)^2$ known. These algorithms were proven to converge in mean square and with probability one under considerably weaker conditions than necessary for the general M-ary case. The theoretical asymptotic probability of error of the three M = 2 algorithm subclasses were found by iteratively solving a set of implicit equations on a computer. The mean vector estimates are biased, and this bias ruined the moment estimator used for the unknown mixing parameter in Subclass A at lower signal to noise ratios. Subclass B performed well except near the extreme values of $P(\omega^1)$ equal to zero or one. The increase in asymptotic probability of error for Subclass C algorithms over an optimum system with all parameters known is at least an order of magnitude less than the optimum probability of error. Experimental results showed that the two unknown mean cases of Subclasses A and B converged rapidly while the ON-OFF case was more sensitive to poor starting values. Convergence of Subclass C algorithms would be slow because the $\{A_N^i\}_{i=1}^2$ regions would ignore

many of the samples. The effect of using weighting sequences $\{\alpha_k = \frac{1}{k+c}\}$ was also investigated experimentally for a Subclass B algorithm. It was found that the value of c that minimized the probability of error was c = -.8 for the random starting method used. This has an interpretation of de-emphasizing the starting vectors and differs from results on conventional stochastic approximation algorithms where $c \geq 0$.

A simulation of the M > 2 case was delayed until Chapter 4. As discussed previously, Algorithm 1 is a direct application of the algorithm class of Chapter 3 to the intersymbol interference problem. The simulation of Algorithm 1 for an eight mode sample space indicated convergence was extremely slow despite a high signal to noise ratio. Again, we have an algorithm that does not take advantage of the separable mode structure of the sample space, which in this case is assumed by the criterion the algorithm asymptotically minimizes.

Unsupervised estimation algorithms were applied to the problem of intersymbol interference in Chapter 4. The chapter was developed in terms of the easily understood one dimensional binary antipodal signal set, with the signal set and the channel memory of c - 1 bauds assumed known to the receiver. The use of the concepts for more general, multidimensional signal sets was delayed until the last section of Chapter 4 so as not to obscure the basic ideas. The mode structure of the multibaud sequence sample space was illustrated for c = 2, and the assumptions on v and v* in previous papers were compared. An examination of the mode structure yielded the $\{Q^k\}$ matrices which contain the mutual mode location constraints. These

matrices were used in Algorithm 2 to map portions of the sequence sample space into the channel gain vector space.

Two algorithms were defined in Chapter 4. As mentioned above, Algorithm 1 directly estimated the 2^{v+c-1} mode locations, while Algorithm 2 used the $\{Q^k\}$ matrices to estimate the channel gain vector thus implicitly estimating the mode locations. An estimator using decision feedback was contemplated, but since errors run in batches rather than more or less independently, a decision directed algorithm using decision feedback would probably converge very slowly compared to Algorithm 2. Supervised starting may provide the good initial vectors such an approach needs, but the severe initial vector dependence makes the approach less interesting. The asymptotic probability of error of Algorithm 1 was evaluated by iteratively solving for the minimum of the decision directed estimator criterion. This required a numerical integration of the sequence sample space and a large number of iterations was necessary before the iterative technique converged. It was found that at low signal to noise ratios the bias of the estimators results in worse performance than an optimum correlator approach; however, the usual operation of such a communications system is at higher signal to noise ratios where in this case the asymptotic performance is almost the same as with all parameters known. The dynamic performance of Algorithms 1 and 2 was evaluated experimentally. The poor results for Algorithm 1 were discussed above; however, Algorithm 2 converged rapidly particularly for a multimodal sample space. In both simulation results a numerically evaluated probability of error was used as the measure of performance.

In Section 4.6 the concepts developed earlier in Chapter 4 were applied to more general message set assumptions and channel transformation models. The emphasis was on obtaining mappings from the $\{u^k\}$ to the channel gain vector space or a similar space so that the highly successful Algorithm 2 could be applied. Also, in this section are brief discussions on two related problems — unknown bandpass signal carrier phase at the receiver, and unknown synchronization.

5.2 Recommendations for Further Study

A criterion $\eta(B)$ whose maximum on $\mathcal{B}^{M'}$ defines the minimum risk solution was found in the report. Although some properties of $\eta(B)$ were found in Chapter 2, little is known about the regression surface shape. Properties of $\eta(B)$ such as the effect of incorrect or simplifying a priori assumptions should be investigated more fully. In particular, the estimates resulting from an incorrect Gaussian mixture assumption should be determined. Since $\eta(B)$ is difficult to evaluate in general, the work will probably have to be done numerically. Other assumptions on parametric families such as binomial, poisson, and exponential might also be investigated.

Two algorithms based on the Keifer-Wolfowitz and Robbins-Munro procedures were defined in Chapter 2. The R-M based algorithm was presented more or less as an example of how the maximum likelihood-related stationary equations found in this chapter might be solved stochastically. Other approaches for stochastically solving the stationary equations should be examined. Formal convergence proofs for these two algorithms need to be constructed.

Under conditions sufficient for convergence these two stochastic gradient based algorithms will converge slowly for most problems because they lack the ability to rapidly adjust to the appearance of a new class in the sample sequence. Hence, they are highly initial value dependent. Asympotically, of course, the algorithms can resolve any mixture having the same parametric form. The clustering algorithm defined in Subsection 2.5.2 has a means of determining (i.e. a test on $\eta(B)$ for a separable Gaussian family) whether a sample should be used to start the estimates of a new class. However, it has a limited ability to resolve mixtures with overlapping densities. A new unsupervised estimation algorithm combining the advantages of both approaches is needed.

The decision directed estimators defined in Chapter 3 did not perform well for larger values of M (i.e. multimodal sample spaces). The algorithm defined in Subsection 2.5.2 can be considered a generalized version of the decision directed estimators. It should be determined whether or not the algorithm converges. If the algorithm does converge, the dynamic performance should be evaluated and compared with the performance of other approaches.

If $\eta(B)$ is multimodal, an approach not discussed in the report which may be useful is a random search on the stochastic surface of $\ln h(x'B)$. The application of random search techniques to finding the maximum of $\eta(B)$ should be examined, and the dynamic performance of such a system should be evaluated.

In Chapter 4, the intersymbol interference problem was formulated in terms of a one dimensional binary antipodal signal set. The number of bauds of channel memory c - 1 and the signal set were assumed known to the receiver. For the case where $v = v^* = c = 2$, the performance of Algorithm 2 was evaluated experimentally and found to be very good for reasonable signal to noise ratios. A formal convergence proof for this algorithm should be constructed. It may be possible to prove convergence under weaker assumptions than stated in Theorem 2 because much of the sample space is mapped back to $\underline{a}$. Methods of extending the approach to $\ell > 1$, and other signal sets were discussed in Subsection 4.6.1. Performance for $\ell > 1$, other values of v, v^*, and c, and other signal sets should also be determined experimentally.

For a practical channel, an implementation of Algorithm 2 will have to deal with some nonlinearities. The effect on the sequence sample space mode structure of nonlinearities in the channel transformation should be examined. Also, an actual test over a communication channel (for example, a telephone channel or satellite relay link) would be useful in indicating what parameter values and signaling sets may be practical.

LIST OF REFERENCES

[1] G. H. Ball, "Data Analysis in the Social Sciences: What About the Details", Proceedings of the Fall Joint Computer Conference, 1965.

[2] G. Nagy, "State of the Art in Pattern Recognition", Proceedings of the IEEE, Vol. 56, No. 5, May 1968.

[3] Y. Ho and A. K. Agrawala, "On Pattern Classification Algorithms, Introduction and Survey", Proceedings IEEE, Vol. 56, No. 12, December 1968.

[4] E. M. Glaser, "Signal Detection by Adaptive Filters", IRE Transactions on Information Theory, Vol. IT-7, No. 2, April 1961.

[5] C. V. Jakowatz, R. L. Shuey, and J. M. White, "Adaptive Waveform Recognition", Information Theory, C. Cherry, Ed., Butterworths, Washington, D. C., 1961.

[6] M. J. Hinich, "A Model for a Self-Adapting Filter", Information and Control, Vol. 5, No. 3, September 1962.

[7] R. F. Daly, "The Adaptive Binary-Detection Problem on the Real Line", Stanford Technical Report TR No. 2003-3, February 1962.

[8] S. Fralick, "Learning to Recognize Patterns Without a Teacher", Stanford Technical Report TR No. 6103-3, March 1965.

[9] D. B. Cooper and P. W. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition", Information and Control, Vol. 7, No. 3, September 1964.

[10] P. W. Cooper, "Some Topics on Nonsupervised Adaptive Detection for Multivariate Normal Distributions", Computer and Information Sciences-II, Academic Press, Inc., New York 1967.

[11] H. Teicher, "Identifiability of Finite Mixtures", Annals of Mathematical Statistics, Vol. 38, No. 4, August 1967.

[12] S. Yakowitz and J. Spragins, "On the Identifiability of Finite Mixtures", Annals of Mathematical Statistics, Vol. 39, No. 1, February 1968.

[13] J. C. Hancock and E. A. Patrick, "Learning Probability Spaces for Classification and Recognition of Patterns With or Without Supervision", Purdue University Report TR-EE65-21, November 1965.

[14] E. A. Patrick, "On a Class of Unsupervised Estimation Systems", IEEE Transactions on Information Theory, Vol. IT-14, May 1968.

[15] C. G. Hilborn, Jr., and D. G. Lainiotis, "Optimal Unsupervised Learning Multicategory Dependent Hypothesis Pattern Recognition", IEEE Transactions on Information Theory, Vol. IT-14, May 1968.

[16] E. A. Patrick, "Asymptotic Distribution of Maximum Likelihood Estimators for a Nonsupervised Adaptive Receiver", IEEE International Communications Conference Record, Philadelphia, Pennsylvania, June 1966.

[17] J. H. Wolfe, "NORMIX:Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions", Research Memorandum SRM68-6, U. S. Naval Personnel Research Activity, San Diego, California, August 1967.

[18] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2, University of California Press, 1967.

[19] H. S. Scudder, "Probability of Error of Some Adaptive Pattern-Recognition Machines", IEEE Transactions on Information Theory, Vol. IT-11, No. 3, July 1965.

[20] E. A. Patrick, G. Carayannopoulos, and J. P. Costello, "Five Results on Unsupervised Learning Systems", Purdue University Report TR-EE66-21, December 1966.

[21] E. A. Patrick and J. P. Costello, "Asymptotic Probability of Error Using Two Decision Directed Estimators for Two Unknown Mean Vectors", IEEE Transactions on Information Theory, Vol. IT-14, January 1966.

[22] W. D. Gregg and J. C. Hancock, "An Optimum Decision-Directed Scheme for Gaussian Mixtures", IEEE Transactions on Information Theory, Vol. IT-14, No. 3, May 1968.

[23] E. A. Patrick, J. P. Costello, and F. C. Monds, "Unsupervised Estimation of a Two Class Decision Boundary", to appear in the Proceedings of the 1968 P. I. B. International Symposium on Computer Processing in Communications.

[24] H. Schwartzlander and J. C. Hancock, "Signal Optimization for Digital Communication Over Channels with Memory", IEEE Symposium on Signal Transmission and Processing, Columbia University, New York, New York, May 1965.

[25] J. M. Aein and J. C. Hancock, "Reducing the Effects of Intersymbol Interference", IEEE Transactions on Information Theory, Vol. IT-9, July 1963.

[26] M. R. Aaron and D. W. Tufts, "Intersymbol Interference and Error Probability", IEEE Transactions on Information Theory, Vol. IT-12, January 1966.

[27] D. W. Tufts, "Nyquist's Problem--The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation", Proceedings of the IEEE, March 1965.

[28] J. C. Hancock and E. A. Quincy, "Jointly Optimum Waveforms and Receivers for Channels with Memory", Purdue University Report TR-EE66-7, April 1966.

[29] B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling", IEEE Transactions on Information Theory, Vol. IT-14, July 1968.

[30] R. Lugannani, "Intersymbol Interference and Probability of Error in Digital Systems", presented at 1969 International Symposium on Information Theory.

[31] R. W. Lucky, "Automatic Equalization for Digital Communications", Bell System Technical Journal, Vol. 45, February 1966.

[32] R. W. Chang and J. C. Hancock, "On Receiver Structures for Channels Having Memory" IEEE Transactions on Information Theory, Vol. IT-12, October 1966.

[33] E. A. Patrick and J. P. Costello, "Unsupervised Estimation of Unknown Signals with Intersymbol Interference", presented at the 1969 Princeton Conference on Information Sciences and Systems.

[34] W. Rudin, Real and Complex Analysis, McGraw-Hill, New York, 1966.

[35] M. Loeve, Probability Theory, 3rd Edition, Divan Nostrand Company, Inc., Princeton, New Jersey, 1963.

[36] S. Kullback, Information Theory and Statistics, Wiley, New York, New York, 1959.

[37] E. A. Patrick and J. P. Costello, "On Unsupervised Estimation Algorithms", presented at the 1969 International Symposium on Information Theory; also submitted to IEEE Transactions on Information Theory.

[38] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function", Annals of Mathematical Statistics, Vol. 23, No. 3, September 1952.

[39] J. H. Venter, "On Dvoretzky's Stochastic Approximation Theorems", Annals of Mathematical Statistics, Vol. 37, No. 6, December 1966.

[40] H. Robbins and S. Munro, "A Stochastic Approximation Method", Annals of Mathematical Statistics, Vol. 22, No. 3, September 1951.

[41] P. Henrici, Elements of Numerical Analysis, Wiley, New York, New York, 1964.

[42] A. Dvoretzky, "On Stochastic Approximation", Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, 1956.

[43] L. LeCam, "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates", University of California Publications in Statistics, Vol. 1, 1953.

[44] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, 1964.

[45] Y. C. Ho and R. C. K. Lee, "Identification of Linear Dynamic Systems", Proceedings of the 3rd Symposium on Adaptive Processes, 1964.

[46] G. N. Saridis and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification", IEEE Transactions on Automatic Controls, Vol. AC-13, No. 5, October 1968.

[47] L. E. Dubins and L. J. Savage, "A Tchebysheff-Like Inequality for Stochastic Processes", Proceedings of the National Academy of Sciences, Vol. 53, 1965.

[48] G. S. Sebestyen, Decision-Making Processes in Pattern Recognition, Macmillan Company, New York, 1962.

[49] G. H. Ball and D. J. Hall, "ISODATA, an Iterative Method of Multivariate Data Analysis and Pattern Classification", 1966 International Communications Conference.

[50] S. Wilks, Mathematical Statestics, John Wiley and Sons, New York, 1962.

[51] E. A. Patrick and G. Carayannopoulos, "Codes for Unsupervised Learning of Source and Binary Channel Probabilities", Information and Control, Vol. 14, No. 3, September 1969.

[52] R. G. Casey and G. Nagy, "An Autonomous Reading Machine," IEEE Transactions on Electronic Computers, Vol. C-17, No. 5, May 1968.

[53] E. A. Patrick and J. P. Costello, "On Some Approaches to Unsupervised Estimation", Purdue University Report TR-EE68-7, August 1968.

[54] J. C. Hancock and E. A. Patrick, "Iterative Computation of a Posteriori Probability for M-ary Nonsupervised Adaptation", IEEE Transactions on Information Theory, Vol. IT-12, No. 4, October 1966.

[55] H. Robbins, "Mixtures of Distributions", Annals of Mathematical Statistics, Vol. 19, No. 3, September 1948.

[56] J. R. VanRyzin, "The Compound Decision Problem with mxn Finite Loss Matrix", Annals of Mathematical Statistics, Vol. 37, No. 2, April 1966.

[57] J. R. Van Ryzin, "The Sequential Compound Decision Problem with mxn Finite Loss Matrix", Annals of Mathematical Statistics, Vol. 37, No. 4, August 1966.

[58] N. Alens, "Compound Bayes Learning Without a Teacher", Stanford University Technical Report No. 6151-2, August 1967.

[59] K. Abend, "Compound Decision Procedures for Pattern Recognition", Proceedings of the National Electronics Conference, Vol. 23, 1967.

[60] S. Fralick, Slenkovich, and Wilson, "Design and Performance of an Adaptive Receiver for Signals of Unknown Frequency", 1966 IEEE International Communications Conference.

[61] H. Robbins, "The Empirical Bayes Approach to Statistical Decision Problems", Annals of Mathematical Statistics, Vol. 35, No. 1, March 1964.

[62] J. M. Wozencraft and I. M. Jackobs, Principles of Communication Engineering, John Wiley and Sons, Inc., New York, 1965.

[63] J. G. Proakis, P. R. Drouilhet, Jr., and R. Price, "Performance of Coherent Detection Systems Using Decision-Directed Channel Measurement", IEEE Transactions on Communications Systems, Vol. CS-12, March 1964.

[64] D. J. Sakrison, "Stochastic Approximation, a Recursive Method for Solving Regression Problems", Advances in Communication Theory, Vol. 2, A. V. Balakvishnan, Ed., Academic Press, New York, 1966.

[65] G. N. Saridis, Z. J. Nikolic, and K. S. Fu, "Stochastic Approximation Algorithms for System Identification, Estimation, and Decomposition of Mixtures", IEEE Transactions on System Science and Cybernetics, Vol. SSC-5, No. 1, January 1969.

BLANK PAGE

APPENDICES

APPENDIX A

EVALUATION OF AN ABSOLUTE MOMENT BOUND

In this appendix it will be shown that for s even

$$E\Big[\Big|\sum_{k=1}^{n}\left[\ln h(X_k|B_t) - \ln h(X_k|B^*) + d_t\right]\Big|^s\Big]$$

$$\le n^{s/2}\left\lceil\frac{s + 2^{s-1} - 1}{s/2!}\right\rceil\Big\{\sum_{i=0}^{s}\sum_{j=0}^{i}(d_t)^{s-i}\binom{s}{i}\binom{i}{j}$$

$$E[|\ln h(x|B_t)|^i]^{i-j} \cdot E[|\ln h(x|B^*)|^i]^j\Big\} \qquad (A.1)$$

This inequality is used to obtain (2.20) in the Bayes mean square convergence proof of Section 2.2. Before proving (A.1), we first present two lemmas that will be useful.

Let U be a zero mean random variable and define

$$\mu_k \triangleq E[|U|^k]$$

<u>Lemma 1</u>. If $\sum_{k=1}^{r} j_k = s$ where $0 < j_k \le s \quad k = 1,2,\ldots,r$ then

$$\mu_{j_1} \cdot \mu_{j_2} \cdot \mu_{j_3} \cdots \mu_{j_r} \le \mu_s \quad 1 \le r \le s \qquad (A.2)$$

Proof: From the property of Lp spaces

$$\mu_k^{1/k} \le \mu_t^{1/t} \qquad 1 \le k \le t \le \infty \tag{A.3}$$

or taking the k^{th} power of each side,

$$\mu_k \le \mu_t^{k/t} \tag{A.4}$$

Applying this inequality r times gives

$$\mu_{j_1} \cdot \mu_{j_2} \cdots \mu_{j_r} \le \mu_s^{\frac{1}{s}\sum_{k=1}^{r} j_k} \tag{A.5}$$

but since $\sum_{k=1}^{r} j_k = s$, (A.5) becomes

$$\mu_{j_1} \cdot \mu_{j_2} \cdots \mu_{j_r} \le \mu_s \tag{A.6}$$

which proves the lemma.

For the following lemma, let $\{U_k\}_{k=1}^{n}$ denote a sequence of statistically independent zero mean random variables.

Lemma 2. The number of products in

$$\Big(\sum_{k=1}^{n} U_k\Big)^s = \sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\cdots\sum_{t_s=1}^{n} U_{t_1} U_{t_2} \cdots U_{t_s}$$

$$= \sum_{k_1=0}^{s}\sum_{k_2=0}^{k_1}\cdots\sum_{t_{n-1}=0}^{k_{n-2}} \binom{s}{k_1}\binom{k_1}{k_2}\cdots\binom{k_{n-2}}{k_{n-1}}$$

$$\cdot U_1^{s-k_1} U_2^{k_1-k_2} \cdots U_n^{k_{n-1}} \tag{A.7}$$

for which

$$\left.\begin{matrix} s - k_1 \\ k_i - k_{i+1},\ 1 \le i \le n-2 \\ k_{n-1} \end{matrix}\right\} \neq 1 \tag{A.8}$$

is bounded by $[\frac{s + 2^{s-1} - 1}{s/2!}]n^{s/2}$ for $n \geq s$ and s even.

Proof: Since none of the exponents in (A.7) can have value one and s is even, only $1 \leq r \leq s/2$ exponents can be non zero at the same time. Denote these ordered non zero exponents by $j_1, j_2, \ldots, j_r$ where $\sum_{i=1}^{r} j_i = s$. The number of products satisfying (A.8) can be obtained by finding how many ways an ordered set of r exponents can be assigned to the ordered set of random variables $\{U_k\}_{k=1}^{n}$, and then multiplying by the number of ordered sets of r exponents for $1 \leq r \leq s/2$.

The number of ways an ordered set of r non zero exponents can be assigned to the ordered U_k is $\binom{n}{r}$. Note that for $n \geq s/2$,

$$\binom{n}{r} \leq \binom{n}{s/2} \qquad 1 \leq r \leq s/2 \tag{A.9}$$

This is the essential bound as a function of the number of samples n.

The second part of the bound is the number of ordered non zero exponents $\{j_i\}_{i=1}^{r}$ $1 \leq r \leq s/2$ where $j_i \neq 1, i = 1,2,\ldots,r$. Rather than evaluating this number, it is sufficient to obtain a bound on it by calculating all the ways the exponents $\{j_i\}_{i=1}^{r}$

satisfy $\sum_{i=1}^{r} j_i = s$. Now $\sum_{i=1}^{r} j_i = s$, $1 \le j_i \le s$, $1 \le r \le s$ defines an $s - 1$ dimensional simplex having $s + 1$ points on an edge. The number of points in the simplex [37] is

$$s + \sum_{r=2}^{s} \frac{1}{(r-1)!} \prod_{k=1}^{r-1} (s - k) = s + \sum_{r=2}^{s} \binom{s-1}{r-1}$$

$$= s + \sum_{r=1}^{s-1} \binom{s-1}{r} + \binom{s-1}{0} - 1$$

$$= s + 2^{s-1} - 1 \qquad \text{(A.10)}$$

Combining (A.9) and (A.10), the number of products satisfying (A.8) is bounded by

$$(s + 2^{s-1} - 1)\binom{n}{s/2} \le \left[\frac{s + 2^{s-1} - 1}{s/2!}\right] n^{s/2}$$

which proves the lemma.

Defining

$$U_k \triangleq \ln h(X_k|B_t) - \ln h(X_k|B^*) + d_t$$

$$U \triangleq \ln h(x|B_t) - \ln h(x|B^*) + d_t$$

Then from Lemma 2 and the fact that U_k is a zero mean random variable (implying that the expectation of all product terms having U_k^1 in them are zero),

$$E\left[\left|\sum_{k=1}^{n} U_k\right|^s\right] \le \left[\frac{s + 2^{s-1} - 1}{s/2!}\right] n^{s/2}$$

$$\sup_{\substack{j_1, j_2, \ldots, j_r \\ 1 \le j_i \le s \\ \sum_{i=1}^{r} j_i = s}} \prod_{i=1}^{r} E[|U|^{j_i}] \qquad (A.11)$$

and from Lemma 1,

$$E\left[\left|\sum_{k=1}^{n} U_k\right|^s\right] \le \left[\frac{s + 2^{s-1} - 1}{s/2!}\right] n^{s/2} E[|U|^s] \qquad (A.12)$$

All that remains to be done is the evaluation of $E[|U|^s]$.

$$E[|U|^s] = E[|\ln h(x|B_t) - \ln h(x|B^*) + d_t|^s]$$

$$= \sum_{i=0}^{s} \binom{s}{i} (d_t)^{s-i} E[|\ln h(x|B_t) - \ln h(x|B^*)|^i] \qquad (A.13)$$

Using the relation

$$|a - b| \le |a| + |b|$$

and the Minkowski inequality (see Rudin [34], p. 62)

$$E[(|a| + |b|)^k]^{1/k} \le (E[|a|^k])^{1/k} + (E[|b|^k])^{1/k}$$

in equation (A.13), it becomes

$$E[|U|^s] = \sum_{i=0}^{s} \sum_{j=0}^{i} \binom{s}{i}\binom{i}{j}(d_t)^{s-i} \left(E[|\ln h(x|B_t)|^i]\right)^{i-j}$$

$$\cdot\left(E[|\ln h(x|B^k)|^i]\right)^j \qquad (A.14)$$

Combining (A.14) and (A.12), the required result of (A.1) is obtained.

APPENDIX B

EVALUATION OF $\eta(B)$ FOR A GAUSSIAN MIXTURE USING A SERIES EXPANSION

For the nonlinear regression approach to finding the maximum of $\eta(B)$ discussed in subsection 2.4.3, it was necessary that the regression function

$$\eta(B;B^*) = \int \ln[h(x|B)]h(x|B^*)dx \qquad (B.1)$$

be an easily manipulated function of the parameter vectors B and B*. In this appendix we take the simple case where M = 2 and is known, and the parametric family $\{f(x|\alpha):\alpha\in G\}$ is composed of one dimensional Gaussian densities with mean values γ, and common covariance $(\sigma)^2 = 1$. Since integrals of the form of (B.1) with M > 1 do not appear in standard integral tables, the approach attempted here is to expand the logarithm term into an infinite series, evaluate the integral, and then simplify into a closed form expression.

Expanding the logarithm term,

$$\ln\left[\sum_{i=1}^{2} f(x|\alpha^i)P(\alpha^i)\right] = -\sum_{k=1}^{\infty} (-1)^k \left(\sum_{i=1}^{2} f(x|\alpha^i)P(\alpha^i)-1\right)^k \frac{1}{k}$$

$$= - \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left(\sum_{i=1}^{2} f(x|\alpha^i) P(\alpha^i) \right)^{k-j}$$

$$= - \sum_{k=1}^{\infty} \sum_{j=0}^{k} \sum_{s=0}^{j} (-1)^j \frac{1}{k} \binom{k}{j} \binom{j}{s} [f(x|\alpha^1) P(\alpha^1)]^{j-s}$$

$$\cdot [f(x|\alpha^2) P(\alpha^2)]^s \qquad \text{(B.2)}$$

Substituting (B.2) into (B.1), and using the Gaussian density form,

$$\eta(B; B^*) = \int \left\{ - \sum_{k=1}^{\infty} \sum_{j=0}^{k} \sum_{s=0}^{j} (-1)^j \left(\frac{1}{k}\right) \binom{k}{j} \binom{j}{s} [f(x|\alpha^1) P(\alpha^1)]^{j-s} \right.$$

$$\left. \cdot [f(x|\alpha^2) P(\alpha^2)]^s \right\} \sum_{r=1}^{2} f(x|\alpha^{r*}) P(\alpha^{r*}) dx$$

$$= - \sum_{r=1}^{2} P(\alpha^{r*}) \sum_{k=1}^{\infty} \sum_{j=0}^{k} \sum_{s=0}^{j} (-1)^j \frac{1}{k} \binom{k}{j} \binom{j}{s} \left[\frac{P(\alpha^1)}{\sqrt{2\pi}}\right]^{j-s} \left[\frac{P(\alpha^2)}{\sqrt{2\pi}}\right]^s$$

$$\cdot \frac{1}{\sqrt{2\pi}} \int \exp\left\{ - 1/2[(j - s)(x - \gamma^1)^2 + s(x - \gamma^2)^2 + (x - \gamma^{r*})^2] \right\} dx$$

(B.3)

Taking the integral term from (B.3),

$$\frac{1}{\sqrt{2\pi}} \int \exp[- 1/2[(j - s)(x - \gamma^1)^2 + s(x - \gamma^2)^2 + (x - \gamma^{r*})^2]] dx$$

$$= \frac{1}{\sqrt{2\pi}} \int \exp[- 1/2[(j + 1)x^2 - 2((j - s)\gamma^1 + s\gamma^2 + \gamma^{r*})x$$

$$+ (j - s)(\gamma^1)^2 + s(\gamma^2)^2 + (\gamma^{r*})^2]] dx$$

$$= \frac{1}{(j+1)^{1/2}} \exp\{-1/2[-\frac{1}{j+1}\left((j-s)^2\,(\gamma^1)^2 + s^2(\gamma^2)^2 + (\gamma^{r*})^2\right.$$

$$\left. + 2s(j-s)(\gamma^1)(\gamma^2) + 2(j-s)(\gamma^1)(\gamma^{r*}) + 2s(\gamma^2)(\gamma^{r*})\right)$$

$$+ (j-s)(\gamma^1)^2 + s(\gamma^2)^2 + (\gamma^{r*})^2]\} \qquad \text{(B.4)}$$

Since no significant simplification is apparent, the evaluation is terminated here. Thus, although an infinite series expansion of $\eta(B;B^k)$ was obtained, the attempt tc re-express this series as a closed form expression was unsuccessful.

APPENDIX C

TWO RESULTS USED IN THE PROOF OF THEOREM 2

The following Martingale lemma was proven by MacQueen [18].

Lemma 3. Let $r_1, r_2, \ldots$, and $t_1, t_2, \ldots$, be given sequences of random variables, and for each $N = 1,2,\ldots$, let r_N and t_N be measurable with respect to Ω_N where $\Omega_1 \subset \Omega_2 \subset \ldots$ is a monotone increasing sequence of σ - fields (belonging to the underlying probability space). Suppose each of the following conditions holds with probability one:

(1) $t_N \geq 0$

(2) $\sum t_N < \infty$

(3) $E[r_{N+1} \mid \Omega_N] \leq r_N + t_N$

Then the sequences of random variables $r_1, r_2, \ldots$, and $s_0, s_1, \ldots$, where $s_0 = 0$ and $s_N = \sum_{k=1}^{N} (r_k - E[r_{k+1} \mid \Omega_k])$, $N = 1,2,\ldots$, both converge with probability one.

The other result used in the proof of Theorem 2 is due to Dubins and Savage [47]. This result was also used by MacQueen in the converge proof of his M-ary case algorithm.

THEOREM 3

Let $U_1, U_2, U_3 \ldots$ be a real valued stochastic process. Let μ_N be the conditional expectation of U_N given the past and $(\sigma_N)^2$

the conditional variance of U_N given the past. Suppose that for any N, μ_N is finite with probability one. (No such assumption is needed for $(\sigma_N)^2$.) Let c, ϵ be positive numbers. Then the probability that there is some N for which

$$(U_1 + U_2 + \cdots U_N) \geq \epsilon + (\mu_1 + \mu_2 + \cdots + \mu_N) +$$

$$c((\sigma_1)^2 + (\sigma_2)^2 + \cdots + (\sigma_N)^2)$$

is less than $1/(c + c\epsilon)$. This bound is sharp.

The actual form of this result used in MacQueen's proof and used in the proof of Theorem 2 here is

$$P\Big[\sum_{k=1}^{N} U_k + \epsilon \geq \sum_{k=1}^{N} \mu_k - c \sum_{k=1}^{N} (\sigma_k)^2\Big] \geq 1 - \frac{1}{1+c\epsilon} \qquad (C.1)$$

In Theorem 2, $U_k \triangleq I^i_{k-1}$ for $k \neq 1$ and $U_1 = 1$. Also, $E[I^i_N \mid Y_{NK}] = E[(I^i_N)^2 \mid Y_{NK}] = 1 - (1 - p(A^i_N))^K + (1 - p(A_N))^K$. Substituting this in (C.1) gives (3.28).

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
Purdue University
School of Electrical Engineering
Lafayette, Indiana 47907

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3 REPORT TITLE
UNSUPERVISED ESTIMATION AND PROCESSING OF UNKNOWN SIGNALS

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Periodic for period February 1968 to June 1969

5 AUTHOR(S) *(First name, middle initial, last name)*
E. A. Patrick
J. P. Costello

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
February 1970	187	65

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F30602-68-C-0186	
b. PROJECT NO.	
5581	TR-EE-69-18
Task No. 558104	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	RADC-TR-69-430

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Rome Air Development Center (EMBIS) Griffiss Air Force Base, New York 13440.

13. ABSTRACT

Many communications systems, sonar and radar systems, control systems and pattern recognition systems such as biomedical signal processing systems must partition a multidimensional sample space so that decisions can be made on the underlying active source events. Unfortunately, the a priori information necessary to construct an acceptable partition is not always available. For many problems estimation using samples of unknown classification is the only source of additional knowledge on the sample space statistical structure. Since the sample classifications are unknown, these estimators are called unsupervised estimation algorithms.

This research is concerned with investigating practical approaches to the unsupervised estimation problem which are in some sense optimum. The emphasis is on recursive estimation algorithms having fixed storage requirements and on sequential sample processing. A Bayesian framework is utilized as a guide towards "optimality", and to provide a unifying relationship for the approaches of the report. The relationship between Bayes a posteriori, stochastic approximation, and decision directed approaches is determined. It is shown, for example, that an optimization criterion derived from the Bayes approach can be used to relate maximum likelihood-related stochastic approximation algorithms with decision directed estimators. The application of unsupervised estimation algorithms to a practical problem is illustrated using the problem of intersymbol interference.

A direct implementation of an optimum a posteriori approach is to approximate the

(Over)

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Pattern Recognition						
Unsupervised Learning						
Statistical Decision Theory						
Stochastic Approximation						
Unsupervised Estimation						
Intersymbol Interference						

ABSTRACT Cont'd

parameter space with a finite set of vector points. The Bayes estimator on such a discretized parameter space is proven to converge to an asymptotic vector with probability one and in mean square. The asymptotic estimator and asymptotic rate of convergence are also found. These results on a discretized parameter space lead to a new continuous parameter space criterion. The maximum of this criterion minimizes average risk against the a priori assumption of a particular parametric density function family. The properties of the criterion surface are largely unknown, but contours evaluated for a two class Gaussian problem show unimodality for this problem. For the case where the criterion surface is unimodal, stochastic approximation algorithms which seek the maximum are defined. Also, a criterion form resulting from a separable Gaussian assumption allows the definition of a simple clustering technique for maximizing the criterion.

A class of decision directed algorithms are defined which minimize a criterion derived from the separable Gaussian criterion. This class of algorithms unifies several previous papers on decision directed estimators. The decision directed estimators are given the interpretation of stochastic approximation algorithms with random weights. This allows a comparison of properties found for the decision directed algorithm class with results in the literature on conventional stochastic approximation algorithms. Theoretical asymptotic probability of error curves and experimental dynamic convergence results are presented.

The problem of intersymbol interference occurs when a channel smears energy from one signal baud onto following ones. The mode or cluster structure of the multibaud sample space is discussed and used to relate the approaches of previous papers. Two decision directed estimators suitable for the interference problem are defined. Experimental and asymptotic performance curves are presented. Extensions and related problems are also discussed.

UNCLASSIFIED

Security Classification

AFLC—Griffiss AFB NY 19 Mar 70-253