(2)

NAVAL POSTGRADUATE SCHOOL
Monterey, California

DISSERTATION

DETECTION OF ABRUPT CHANGES IN STATISTICAL MODELS

by

David Aviv

June, 1991

Dissertation Supervisors: R. Cristi
C. W. Therrien

Approved for public release; distribution is unlimited.

068

Unclassified

Security Classification of this page

REPORT DOCUMENTATION PAGE

1a Report Security Classification UNCLASSIFIED	1b Restrictive Markings
2a Security Classification Authority	3 Distribution Availability of Report: Approved for public release; distribution is unlimited.
2b Declassification/Downgrading Schedule	
4 Performing Organization Report Number(s)	5 Monitoring Organization Report Number(s)
6a Name of Performing Organization: Naval Postgraduate School; 6b Office Symbol (*If Applicable*): EC	7a Name of Monitoring Organization: Naval Postgraduate School
6c Address (*city, state, and ZIP code*): Monterey, CA 93943-5000	7b Address (*city, state, and ZIP code*): Monterey, CA 93943-5000
8a Name of Funding/Sponsoring Organization; 8b Office Symbol (*If Applicable*)	9 Procurement Instrument Identification Number
8c Address (*city, state, and ZIP code*)	10 Source of Funding Numbers: Program Element Number; Project No; Task No; Work Unit Accession No

11 Title (*Include Security Classification*) DETECTION OF ABRUPT CHANGES IN STATISTICAL MODELS

12 Personal Author(s) David Aviv

13a Type of Report	13b Time Covered	14 Date of Report (*year, month,day*)	15 Page Count
Doctoral Dissertation	From To	1991, June	251

16 Supplementary Notation The views expressed in this paper are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

17 Cosati Codes		
Field	Group	Subgroup

18 Subject Terms (*continue on reverse if necessary and identify by block number*)
Sequential tests; disorder; abrupt changes; cumulative sum tests (cumsum); Page; Lorden; Wald; diffusion processes

19 Abstract (*continue on reverse if necessary and identify by block number*

This dissertation investigates different types of disorder problems by using sequential procedures for on-line implementation. The problem is considered within the framework of detecting abrupt changes in an observed random process when the disorder can occur at unknown times. The focus of this work is on quickest detection methods for cumsum procedures implemented for different parametric and nonparametric nonlinearities and their performance evaluation. Both the non-Bayesian (Maximum-Likelihood) and the Bayesian frameworks are presented but the focus is mainly on non-Bayesian methods for which detailed analysis is provided. The use of Brownian motion approximations is also included and provides an additional viewpoint of analyzing the performance for both the non-Bayesian and Bayesian methods.

20 Distribution/Availability of Abstract	21 Abstract Security Classification
[X] unclassified/unlimited [ ] same as report [ ] DTIC users	Unclassified

22a Name of Responsible Individual	22b Telephone (*Include Area code*)	22c Office Symbol
R. C. Cristi	(408) 646-2223	EC/Cx

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

security classification of this page Unclassified

Approved for public release; distribution is unlimited

Detection of Abrupt Changes in Statistical Models

by

David Aviv
Major, Israeli Air Force
B.S., Ben Gurion University, 1981
M.S., Tel-Aviv University, 1987

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

From the

NAVAL POSTGRADUATE SCHOOL
June 1991

Author: Aviv David.
David Aviv

Approved by:

Arthur Schoenstadt, Professor
Department of Mathematics

Uno R. Kodres, Professor
Department of Computer Science

Ralph Hippenstiel, Professor
Department of Electrical and Computer Engineering

Charles W. Therrien, Professor
Department of Electrical and Computer Engineering
Dissertation Supervisor

Roberto Cristi, Professor
Department of Electrical and Computer Engineering
Dissertation Supervisor

Approved by: Michael A. Morgan
Michael A. Morgan, Chairman,
Department of Electrical and Computer Engineering

Approved by:
Richard S. Elster, Dean of Instruction

ABSTRACT

This dissertation investigates different types of disorder problems by using sequential procedures for on-line implementation. The problem is considered within the framework of detecting abrupt changes in an observed random process when the disorder can occur at unknown times. The focus of this work is on quickest detection methods for cumsum procedures implemented for different parametric and nonparametric nonlinearities and their performance evaluation. Both the non-Bayesian (Maximum-Likelihood) and the Bayesian frameworks are presented but the focus is mainly on non-Bayesian methods for which detailed analysis is provided. The use of Brownian motion approximations is also included and provides an additional viewpoint of analyzing the performance for both the non-Bayesian and Bayesian methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ABBREVIATION	EXPANSION
AR	Auto regressive
ARE	Asymptotic relative efficienc,
ARL	Average run length
$\overline{\text{ARL}}$	Stati⌣nary average run length
$\text{ARL}_\mu(x)$	Average run length when the shift in mean is μ and the initial condition is x
ARMA	Auto regressive moving average
CUMSUM	Cumulative sum
DFT	Discrete Fourier transform
FSST	Fixed sample size test
GLR	Generalized likelihood ratio
i.i.d.	Independent and identically distributed
LLR	Log likelihood ratio
LMP	Locally most powerful
LO	Locally optimum
MAP	Maximum a posteriori
ML	Maximum likelihood
MLE	Maximum likelihood rstimator
OC	Operating characteristic
PSK	Phase shift keyed
ROC	Receiver operating characteristics
SNR	Signal-to-noise ratio
SPRT	Sequential probability ratio test
ST	Sequential test
UMP	Uniformly most powerful
WGN	White Gaussian noise

LIST OF SYMBOLS

SYMBOL	MEANING
H_0	Null hypothesis
H_1	Alternative hypothesis
θ_0	Parameter under the null hypothesis
θ_1	Parameter under the alternative hypothesis
α	Probability of type I error
β	Probability of type II error
$E\{\ \}$	Expectation operator
$\Pr\{\ \}$	Probability
P_0	Probability density under the null hypothesis
P_1	Probability density under the alternative hypothesis
T	Mean time between false alarms
D	Delay for detection
Q	Operating characteristic function
$L(\theta)$	The average run length function
P_{FA}	Probability of false alarm
P_D	Probability of delay
a	The upper boundary of a sequential test
b	The lower boundary of a sequential test
N	Stopping time
N^*	Extended Stopping Time
ν	Change time
S_n	Sequential test (Wald)
$\tilde{S}_n$	Repeated sequential test (Page)
$g(\cdot)$	Nonlinearity
$g_{lo}(\ )$	The local optimum nonlinearity
h	The root of the moment generating function of $g(x)$
η	Asymptotic performance measure
$\underline{\eta}$	Lower bound for the asymptotic performance measure
$I(\theta_0, \theta_1)$	The Kullback-Liebler information number

$(a)^+$	$\max(0,a)$
μ	Mean of the Gaussian density
σ^2	Variance of the Gaussian density
B_t	Brownian motion
Z_t	General diffusion process
$s(x)$	Scale density function of a diffusion process
$S(x)$	Scale function of a diffusion process
$G(x, \xi)$	The Green function of a diffusion process
$u(x)$	The probability that a diffusion process hits a before b
$v(x)$	The mean time to reach a or b by a diffusion process
$w(x)$	The total cost up to the time when either a or b was first reached
$\alpha(y \mid x_0)$	The stationary density of an instantaneous return process
L	Number of regimes up to detection (Bayesian approach)
C	Cycle time
$\{\gamma\}$	Process of transitions
$\boldsymbol{\varepsilon}$	The vector of possible realizations of γ
$l_t(\,)$	Likelihood term calculated at time t
w_n	Waiting time of the nth customer
$\mathcal{T}_i$	Descending ladder epoch i
$\mathcal{H}_i$	Descending ladder height (corresponding to epoch i)

ACKNOWLEDGEMENTS

This work would not have been possible without the involvement of several people. I hereby express my deep appreciation and sincere thanks to the following individuals.

First and foremost, I wish to thank Professor R. Cristi for his help, guidance, support, contributions and valuable comments that he gave me during many hours of formal and informal meetings and discussion that we shared.

Next, I wish to thank Professors C. W. Therrien and R. Hippenstiel, who served in my doctoral committee, for their valuable comments, discussions and support. Gratitude is also extended to the other members of my doctoral committee, namely, Professors A. Schoenstadt and U. R. Kodres, who lent their knowledge and experience. I have also to thank Ms. Hania La Born, the typist of this dissertation, for her dedication, patience and proficiency.

A special thank you is given to Dr. Will Gersch from the University of Hawaii for his interest, knowledge, valuable comments and support during the progress of this work and to Professor M. Sherif for his friendship, his aid and assistance during my stay at the Naval Postgraduate School.

I would like to express my deepest gratitude to the people of the United States of America for honoring me with the one-in-a-lifetime privilege of studying in this country. I have made friends and met people of many different nationalities, and will treasure these memories for life.

Finally, I would like to thank the Israeli Air Force for giving me the opportunity to study at the Naval Postgraduate School.

To my mother and to the memory of my father.

I. THE DISORDER AND CHANGE DETECTION PROBLEM FORMULATION—AN OVERVIEW

A. INTRODUCTION

This dissertation presents sequential decision methods both in the non-Bayesian (maximum likelihood) framework and in the Bayesian framework. The focus is mainly on non-Bayesian methods, where the goal is to detect, as quickly as possible, changes in statistical models of a random process when these changes can occur at a random time, while the false alarm rate should be lower bounded by some given constant.

In the classical detection framework such procedures were considered by Wald (Wald, 1947), for which the binary hypothesis framework was developed under the assumption was that all the observations come from one model or from an alternative one. It was not until Page's work (Page, 1954) in the non-Bayesian framework and Shiryayev (Shiryayev, 1961, 1963, 1965) in the Bayesian framework that the problem was extended to detecting a **change** from one statistical model to a second model. Lorden (Lorden, 1971) showed that the cumulative sum tests as proposed by Page are asymptotically optimal when the mean time between false alarm tends to infinity, in the sense of minimizing the average delay time for detection. Recently, Pollak (Pollak, 1985) proved an optimality property for the Shiryayev rule.

Two types of problems depend on the time element. The first is the **disorder** problem in which the given observations correspond to one statistical model until some unknown time after which the samples

correspond to another statistical model. Hereby we will use the notations disorder and change as synonyms, even though a disorder is referred to as a general change in density which describes the change in the statistical behavior of the model, a change will refer most of the time to changes in specific parameters like mean variance, etc. The *second* problem is the **transient** problem in which the disorder decays after some time. In this dissertation we will focus only on the disorder (change) problem.

When a disorder occurs, the random variables we are concerned with are the change time and the model parameters after the change. As will be presented throughout this dissertation, the **detection** process refers to detecting the change as quickly as possible while ensuring infrequent false alarms, while the **estimation** process refers to estimating the change time and the model parameters after the change. This dissertation focuses on the detection element. The problem of *joint* estimation of the change time and the model parameters is also addressed and shown to appear in an explicit closed form in certain cases.

The question of *where do change detection problems occur* is next introduced. Three typical situations in which change detection is a critical component are considered. The *first*, in which the detection is used to produce alarms during the monitoring of dynamical systems, such as failures in sensors (Willsky, 1976,1986), detection of tsunamis and earthquake prediction (Nikiforov, 1986), and detection of production failures (Assaf and Ritov, 1988). Many more applications in industrial and military environments can be considered. Survey papers for fault detection methods are given by Isermann (Isermann, 1984), and Gertler (Gertler, 1988) The

second situation arises in the area of adaptive algorithms, were the presence of abrupt non-stationarities in the signal causes severe errors in adapting the gains of the recursive algorithms. Thus, an abrupt change detection procedure is needed to improve the tracking capability of the algorithm. For a complete survey see Ljung and Gunnarsson (Ljung and Gunnarsson, 1990).

Finally, the *third* type of application occurs when the change detection algorithm is considered as an integral part of the modeling of a signal or a system. The most popular applications are segmentation of speech signals using switching parameter methods within AR models (Andre-Obrecht, 1988) or various geophysical signals (Nikiforov, 1986). In such cases switching methods within the transition matrix of state-space models (Tugnait, 1986), or a modified Kalman filter is used to cope with changes modeled as abrupt transitions in the measurement matrices (Shumway, 1990). Also, the problem of outlier detection by modifying the Kalman filter was introduced by Pena and Guttman (Pena and Guttman, 1988).

B. THE DISORDER PROBLEM FORMULATION

1. The General Disorder Problem

The change detection problem is presented within the hypothesis testing framework, thus, requiring some statistical knowledge about the tested hypotheses which in turn are based upon statistical models of the hypotheses before and after the disorder. The model based framework is rich enough to serve as a basis for the problem formulation, resulting in parametric type tests. As it will be presented later, certain types of change detection procedures known as cumulative sum or **cumsum procedures** are able to

cope with the parametric and nonparametric forms as well. Within this framework *four* types of change detection problems will be considered.

Let H_0 and H_1 be the two (simple) hypotheses, corresponding to two possible probability distributions P_0 and P_1 on the observation space x. If a parametric notation is to be used, then the notation $P(x \mid \theta_0)$ and $P(x \mid \theta_1)$ or $P_0(x)$ and $P_1(x)$ will be used. The observations $x_1, x_2, \ldots$ are assumed to be independent random variables.

Type 1: Classical Binary Hypothesis Testing

This problem was considered by Wald (Wald, 1947) and can be written as:

$$H_0: x \sim P_0,$$

$$\text{versus} \qquad (1\text{–}1)$$

$$H_1: x \sim P_1,$$

where the notation "$x \sim P$" denotes the condition that x has distribution P. In this problem there is no time index, hence, no direct formulation of a change.

Type 2: Disorder Formulation

This problem was considered by Page (Page, 1954) and can be presented in following manner. Let ν be the unknown time when the change from P_0 to P_1 occurred. Let P_ν denote the probability when the change occurred at the ν^{th} observation. Let P_0 denote the probability there is no change, i.e., $\nu = \infty$.

The problem can be presented as

$$H_0: \quad x_1, x_2, \ldots \qquad \sim P_0 \qquad \text{no change}$$

$$\text{versus}$$

$$H_\nu: \quad x_1, x_2, \ldots, x_{\nu-1} \sim P_0. \qquad (1-2)$$

change at time ν

$$x_\nu, x_{\nu+1}, \ldots \sim P_1.$$

If the observation record is finite and equal to say s, the detection problem becomes a *multiple* hypothesis testing, since the test "looks" for at least one of the H_ν $(1 \le \nu \le s)$ to hold against H_0.

Type 3: Transient and outliers formulation:

Consider two change times ν and τ such that

$$\begin{aligned} H_0: &\ x_1, x_2, \ldots && \sim P_0 \\ &\qquad \text{versus} \\ H_1: &\ x_1, x_2, \ldots, x_{\nu-1} && \sim P_0 \\ &\ x_\nu, x_{\nu-1}, \ldots, x_{\tau-1} && \sim P_1 \\ &\ x_\tau, x_{\tau-1}, \ldots && \sim P_0. \end{aligned} \qquad (1-3)$$

The same arguments about composite testing can be applied here. Notice that this framework can be extended to the so-called **multiple disorder problem**, in which the observations $x_\tau, x_{\tau+1}, \ldots \sim P_2$ (P_2 being another probability density on the observation space).

Type 4: Initial Condition Disruption

For model based detection schemes based upon state-space, ARMA, etc., the initial condition is a part of the statistical model. Hence, besides the ordinary way to model the statistical change as a change from P_0 to P_1, a certain class of changes can be modeled as a result from changes in the initial condition. This problem is also time related since the change might occur at an unknown time.

Once H_1 is decided, i.e., disorder detected, further questions arise, such as estimating the change time ν, possibly to estimate θ_0 and θ_1, and in some cases to diagnose which type of change actually occurred. Thus, the detection and estimation following the detection problems being two separate issues can be coupled, but it is important to distinguish between them.

Both **off line** (n fixed) and **on-line** (n growing) algorithms can be designed for solving such types of problems, and as shown in the sequel differ substantially, both from the change detection formulation and from the performance evaluation point of view.

2. Solution Methods

The solution for such problems is a function of several factors.

a. Off-line versus On-line Tests

In the *off-line* formulation, a given finite record is given $x_1, x_2, ..., x_T$ and a test statistic $g_T = g(x_1, x_2, ..., x_T) \geq \lambda$ has to decide whether or not the change occurred. In the *on-line* formulation, the test statistic $g_t = g(x_1, x_2, ..., x_t) \geq \lambda$ has to reach a decision the first time when g_t exceeds a threshold λ.

b. Criterion

For the classical detection problem (1-1), the criteria in the sense of Neyman-Pearson (Ghosh, 1970), is based on a test which maximizes the **power** or the probability of detection (the probability of deciding H_1 when H_1 is actually true) subject to the constraint that the **size** or the false-alarm probability (the probability of deciding H_1 when H_0 is true) is less than or equal to a given value.

As seen from equation (1-2), in the *off-line* framework, the change detection problem involves multiple hypotheses testing, for which the Neyman-Pearson lemma is not valid (Ghosh, 1970). Therefore the test in this case cannot be defined as one of maximizing the power since H_1 is not reduced to a simple distribution but a set of distributions. In such cases, the best property for a test is said to be **Uniformly Most Powerful (UMP)**, i.e., tests which have the highest detection probability for each distribution of the alternative hypotheses H_1. Therefore no UMP tests exist for change detection problems. In this case, those UMP properties can be recovered by using asymptotic analysis (Deshayes and Picard, 1986). In order to cope with the performance analysis of test statistic functions the following definition is needed.

Definition: Stopping time. Let $x_1, x_2, \ldots$ be the sequence of independent random variables. The nonnegative integer valued random variable N is said to be a stopping time for the sequence if the event $\{N = n\}$ is independent of $x_{n+1}, x_{n+2}, \ldots$. □

Hence, the event $\{N = n\}$ corresponds to stopping after having observed $x_1, \ldots x_n$ and thus must be independent of the values of the random variables yet to come (Ross, 1989).

For *on-line* processing, the criteria is modified. Notice that by using the formulation (1-2) for a large enough number of observations, the change will be detected with probability one. Thus, a natural criterion should be the delay for detection, subject to the constraint that the size of the test is upper bounded by a given threshold (Page 1954, Shiryayev 1963). Lorden

(1971) and Nikiforov (1983) use a slightly different version of the delay for the on-line problem:

Let S_n denote the test statistic at time n. Let N be the **stopping rule**, and let λ_n be a generalized threshold. Then:

$$N = \inf\{n: S_n \geq \lambda_n\} \tag{1-4}$$

defines the stopping rule and stopping time. See Figure 1.1.

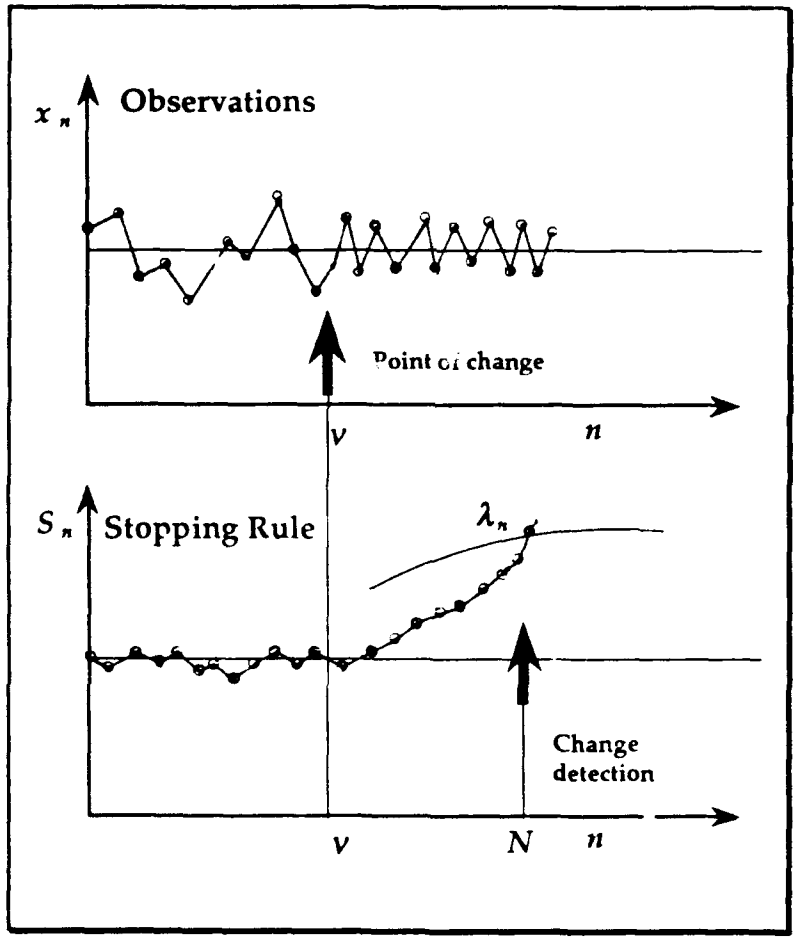


Figure 1.1. General Characteristics of the Detection Model. The observation sequence $\{x_n\}$ is transformed into a sequence $\{S_n\}$. A change in the model structure of $\{x_n\}$ results in a cumulative departure of $\{S_n\}$. The change is detected by comparison of $\{S_n\}$ with a generalized threshold $\{\lambda_n\}$ (from Segen and Sanderson, 1980).

The worst case average delay D (Lorden, 1971) is defined by

$$D = \bar{E}\{N \mid H_1\} = \sup_{\nu \geq 1} \operatorname{esssup} E_\nu\{(N - \nu + 1)^+ | x_1, x_2, \ldots, x_{\nu-1}\} \qquad (1-5)$$

$$\text{where } (a)^+ = \max(0, a),$$

where E_ν denotes the expectation of the change time under the probability law P_ν, where P_ν denotes the distribution of the sequence $x_1, x_2, \ldots$, under which x_ν is the first term with distribution P_1. In other words, D is the smallest value such that for any $\nu = 1, 2, \ldots$

$$E_\nu\{(N - \nu + 1)^+ | x_1, x_2, \ldots, x_{\nu-1}\} \leq D$$

meaning that this "minimax" type criterion defines the best worst case for delay.

Thus, the criteria is defined in terms of the quickest detection of a change subject to the constraint that the size of the test is upper bounded, i.e., the desire for large mean time between false alarms T, where T is also defined in terms of the stopping time

$$T = F\{N | H_0\} \qquad (1-6)$$

which denotes the expectation under the no-change hypothesis H_0. The pair (T,D) will specify the performance of a given algorithm.

Notice that in the transient or multiple disorder setting of the equation (1-3), a fast detection is necessary since if $\tau - \nu + 1 < D$ the transient cannot be detected.

Thus, for the *on-line* framework, this natural criterion should lead to the optimal stopping rule, and the question that arises: are there test statistics

which are optimal in that sense? A positive answer will be presented in the sequel.

Different types of criteria can be used for deriving optimal stopping times for change detection, see Bojdecki and Hosza (Bojdecki and Hosza, 1984) and Pelkowitz (Pelkowitz, 1987).

For the *off-line* problem, this question is more difficult, because as was shown in equation (1.2), change detection problems are multiple hypotheses problems for which there exists no optimum test in the classical sense of power, (Neyman-Pearson lemma), hence, no UMP tests exist. In such situations, an **asymptotic analysis** for which UMP tests can be recovered is of interest. Deshayes and Picard (Deshayes and Picard, 1986) showed that UMP tests exist for likelihood-oriented methods in the sense of large deviation asymptotic analysis. (Sample size goes to infinity.)

c. Optimal Stopping Rules

The *off-line* point of view was addressed in the last section where it was shown that optimality exists only in the sense of asymptotic analysis. For the *on-line* point of view, in the **non-Bayesian framework**, the only optimality results are given by Shiryayev and Lorden. Lorden (Lorden, 1971) showed that for some constant γ, the stopping rule N must satisfy:

$$E_0\{N\} = E\{N|\nu = \infty\} \geq \gamma.$$

The speed in which a stopping rule detects a (true) change of distribution is evaluated by (1-5)

$$\operatorname{supess\,sup} E_\nu\left\{(N-\nu+1)^+ | x_1, x_2, \ldots, x_{\nu-1}\right\}.$$

Lorden showed that a certain class of stopping rules is asymptotically ($\gamma \to \infty$) optimal, and that the cumsum procedure (Page's test which can be described as repeated sequential tests) belongs to this class.

In the **Bayesian framework**, Shiryayev (Shiryayev, 1968, 1978) solved the problem. He considered a cost function whereby one loses one unit if $N<\nu$, and loses c units for each observation taken after ν if $N \geq \nu$. The prior on ν is assumed to be geometric. Shiryayev showed that the stopping rule prescribes stopping as soon as the posterior probability of the change having occurred exceeds a fixed level.

d. Use of Prior Knowledge

For change detection problems, prior knowledge can be useful in two cases:

The *first* case is related to the problem of estimating the change time after detection. From the **Bayesian** point of view, the knowledge of the statistical nature of **change time** makes up the prior needed for such a test. Such knowledge on the distribution of the change time (or initial conditions) will assist in the quickest delay detection, i.e., estimation of the time change. In the **non-Bayesian** approach this is equivalent to assuming a uniform prior distribution over the observation set, resulting in a detector which computes the likelihood function for all possible disorder times.

The *second* case is the estimation following detection of the statistical model after the change of the parameter set θ_1. For test procedures implemented on line, the use of prior knowledge on the parameters set θ_0, θ_1 $\in \Theta$ improves the quickest detection since in such situations, only a short

sample is available from the true change time to the detection time, thus, it is difficult to identify θ_1.

In this context, we shall consider two different forms of the prior on the distribution after the change. The first form of prior uses the **composite hypothesis** testing framework. As an example, the Darmois-Koopman family of distributions (Govindarajulu, 1975, Siegmund, 1985) which is presented in the sequel, allows suitable parametric tests, using the assumption that the statistics after the change have a form of a one parameter exponential distribution. The second form of prior uses the popular method of **multiple models** whenever the set of parameters θ_1 is finite. Such methods can be found in the literature (Anderson and Moore, 1979).

The problem of detecting the change time and estimating the statistical model after the change is a difficult task because of the reasons given. Except for cases where the solution to the detection-estimation can be made explicit, like estimation of the jump amplitude in the case of additive changes in Gaussian linear models (Willsky and Jones, 1976), the combined detection-estimation solution cannot be shown in a closed form. This point is further discussed in Chapter III when the generalized likelihood ratio algorithm (GLR) is applied to linear models.

This dissertation focuses on the methods of the quickest detection problem which provides in the case of detection of jumps in the mean, a convenient way to estimate the unknown jump. However, it will be shown that a lot of complicated problems like changes in spectral properties or eigenstructure (changes in State Space models, AR models or ARMA models) can be transformed to changes in the mean of a statistic function g_n, enabling

the use of quite easy detection schemes to detect rapid changes in the dynamics of the signal model. As shown in the sequel, such detection algorithms are based on the cumsum procedure which provides a tradeoff of computation efficiency and complexity.

3. Performance Evaluation

In the *off-line* processing, the process is observed only over a finite interval, hence only a finite number of samples is used. The problem is then considered as that of classical hypothesis testing (1–1). In this case the performance is measured in terms of probability of detection versus the probability of false alarm.

In the *on-line* processing, the approach of "quickest detection" is adapted as the performance criterion used in sequential analysis. This approach is used by Nikiforov (Nikiforov, 1979, 1980). For this setting, the terms **run length** and **average run length** (**ARL**) will be used in order to determine the number of observations needed to reach a detection decision. This function will be shown to be the main tool in the performance evaluation of the test procedures. The first time the test statistic, i.e. the **stopping rule** (statistic used to determine the change) crosses the pre-determined threshold according to desired performance, is called the **stopping time** or sometimes also the **Markov time** (Shiryayev, 1978).

C. MODEL BASED METHODS

In designing the change detection/estimation algorithms, the philosophy developed in Chow and Willsky (Chow and Willsky, 1986) distinguishes two tasks which are depicted in Figure 1.2.

The *first* task is the generation of change indication signals (residuals) sometimes also called error signals. These signals are designed to reflect the possible changes in the measurements or models and to make a subsequent detection possible. These signals are designed to have a certain mean (usually zero) and a white noise correlation signature when no change occurs. This is referred to as the "white noise" hypothesis. In general the mean value or spectral properties change under a disorder.

The *second* task is design of the stopping rules (or decision rules) based upon these residuals.

Sometimes an additional task diagnostics is added. This is the problem of estimating the origin of the change (for example: which pole location changed). A broad class of change detection methods makes explicit use of a mathematical model of the observed system or signal. For example, the setting of the system or signal in a state-space form enables the use of Kalman filtering methods to generate the residuals (innovations in this case). This twofold problem will be presented next.

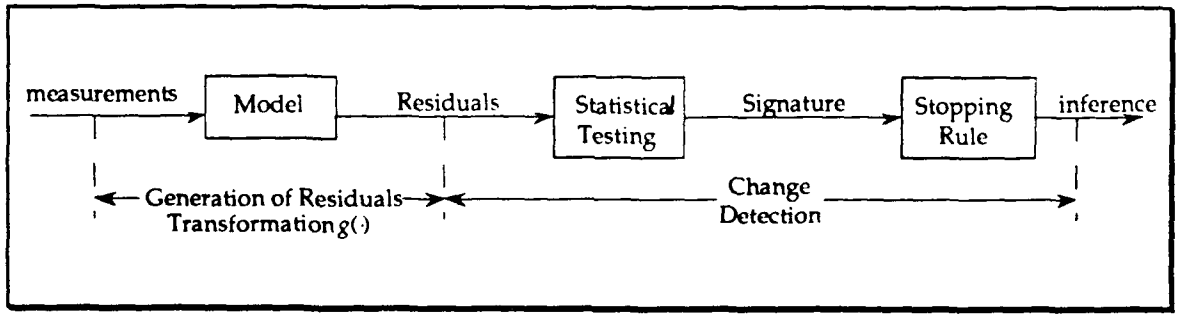


Figure 1.2. Model Based Change Detection Scheme as a Twofold Problem (from Gertler, 1988)

1. Generating the Change Indication Signals (Residuals)

As shown in Figure 1.2, modeling is an integral part of the change detection process, usually for creating "white" residuals under the "no change" hypothesis. Using the state-space setting, residuals may be generated in a number of different ways, which will be presented briefly.

a. Straight Input-Output Residuals (Gertler, 1988)

Given the state-space model

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{B}\mathbf{u}(n)$$

$$y(n) = \mathbf{C}\mathbf{x}(n)$$

an equivalent input-output model can be presented by using the shift operator with matrices G(z) and H(z), z being the shift operator and H being a diagonal matrix:

$$\mathbf{H}(z) \cdot y(n) = \mathbf{G}(z) \cdot u\,(n)$$

where

$$\mathbf{G}(z) = \mathbf{C}\cdot[\mathrm{adj}(\mathbf{I}z-\mathbf{A})\mathbf{B}]$$

$$\mathbf{H}(z) = \det\,(\mathbf{I}z-\mathbf{A})\mathbf{I}.$$

Defining

$$\mathbf{q}(n)^{\mathrm{T}} = [\mathbf{u}(n),y(n)]^{\mathrm{T}}$$

$$\mathbf{F}(z) = [\mathbf{G}(z),-\mathbf{H}(z)]$$

then the input-output equation can be written as:

$$\mathbf{F}(z) \cdot \mathbf{q}(n) = 0.$$

Consider now the model matrix $\hat{\mathbf{F}}(z)$ which represents the discrepancies between the input-output models $\hat{\mathbf{G}}(z)$ and $\hat{\mathbf{H}}(z)$, and the true system $\mathbf{G}(z)$ and $\mathbf{H}(z)$:

$$\hat{\mathbf{F}}(z) = [\hat{\mathbf{G}}(z), -\hat{\mathbf{H}}(z)]$$

where

$$\hat{\mathbf{G}}(z) = \mathbf{G}(z) + \Delta\mathbf{G}(z,t); \quad \hat{\mathbf{H}}(z) = \mathbf{H}(z) + \Delta\mathbf{H}(z,t).$$

Such discrepancies may account for plant faults or changes. Applying this equation to the measurements $\tilde{\mathbf{q}}(n)$ with the model matrix $\hat{\mathbf{F}}(z)$ yields the residuals vector $\mathbf{e}(n)$:

$$\hat{\mathbf{F}}(z) \cdot \tilde{\mathbf{q}}(n) = \mathbf{e}(n).$$

b. Filtering and Parameter Identification Methods

A popular solution (Willsky, 1976) consists of monitoring the innovations or the prediction errors, using estimation filters or parameter identification methods. Using the optimal state estimator, the Kalman filter is designed according to the "normal mode" or no change situation. If prior knowledge is known about the change or if a diagnosis is required in addition to detection, a possible solution consists of using a bank of Kalman filters designed according to all the possible models for each hypothesis (see Figure 1.3). Notice, that the Kalman filter produces under the null hypothesis zero mean and independent residuals. Consequently deviations from this behavior are indicators of change. However, in some practical problems, it may be necessary to monitor a function of the innovations rather than the innovations themselves (Basseville, 1988).

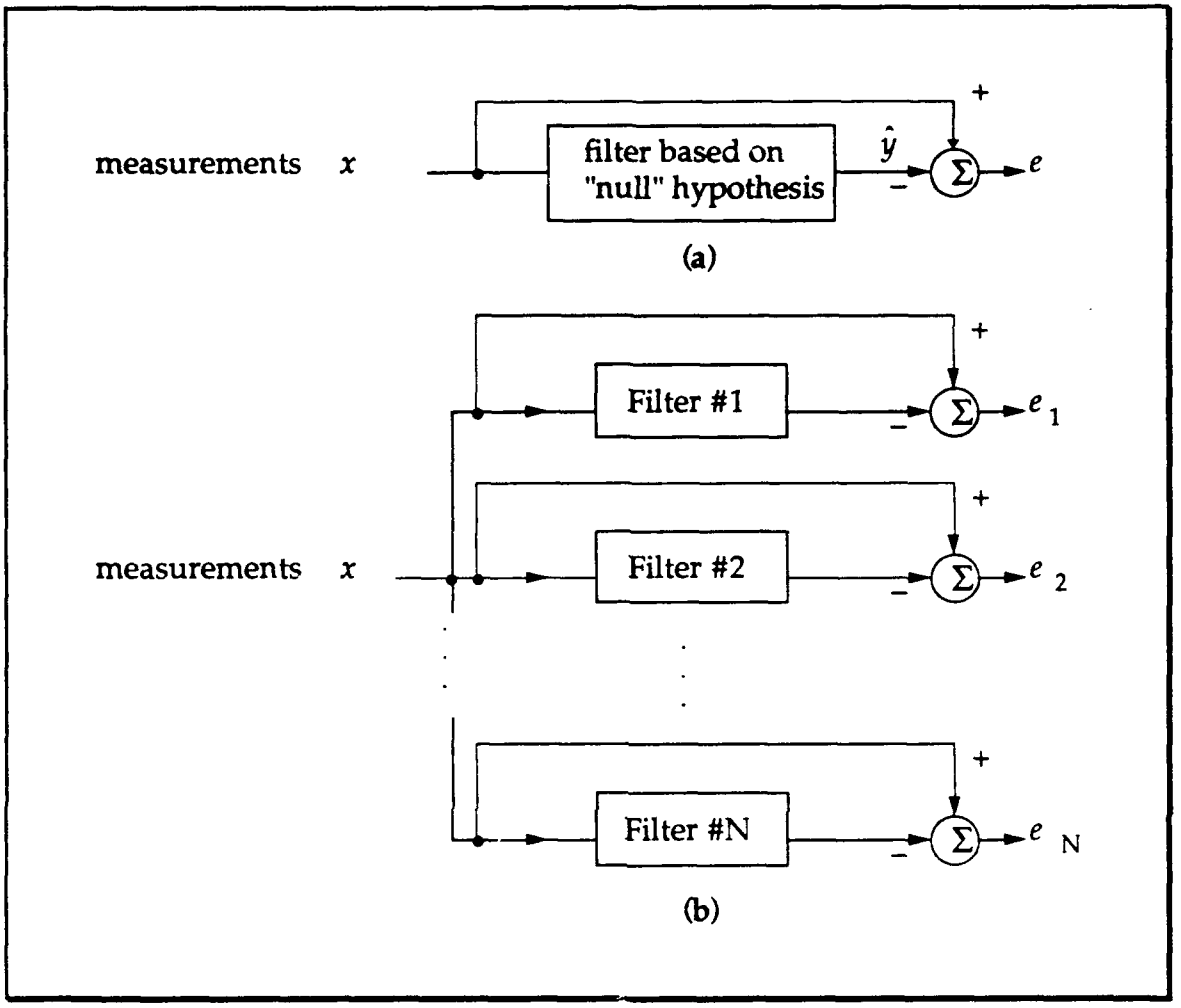


Figure 1.3. Filtering Methods for Generating the Residuals.
(a) "normal mode" filter
(b) generating error signatures due to possible change hypotheses

In identification-based methods, a residual quantity is defined in relation to the plant parameters. The plant is identified in a fault-free reference situation, then repeatedly on line. The results of the latter, are compared to the reference values and a parameter error (residual) is formed. **Remark**: these model-based methods do not include explicit **model switching**.

In Chapter VI such methods will be described, thus enabling us to modify the Kalman filter to detect the change.

c. Redundancy Methods

These techniques are used primarily for failure detection (sensor failures). Two classes can be distinguished. The *first* class is **direct or physical redundancy**. Using several identical sensors measuring the same quantities, the differences between each possible pair may reflect a change. These residuals are processed using voting methods (Willsky, 1976). Another approach consists of searching subsets of measurements for inconsistency, thus indicating changes.

The *second* class is indirect or analytical redundancy. This method monitors of all the existing relationships between the inputs and the outputs that are zero under the hypothesis of no change exists. These techniques were studied by Deckert (Deckert et al. 1977), Chow and Willsky (Chow and Willsky 1980, 1984), and others.

2. Statistical Testing (stopping rules)

The resulting residual vector contains the combined effects of the changes and the noise (as well as the modeling errors). Two approaches can be considered.

The *first* consists of the deterministic modeling of the changes. (It is important not to confuse the random nature of the change time and (usually) the change magnitude with the deterministic modeling of the change). For example, consider the case:

$$X_n = \theta_n + n_n \quad n_n \sim N(0,\sigma^2)$$

where

$$\theta_n = \begin{cases} \theta_0 & 1 \le n \le \nu-1 \\ \theta_1 & n \ge \nu \end{cases}$$

ν being the change time (random). Therefore, the effect of changes on the residuals has to be separated from that of the noise. This is done by statistical testing, making use of the assumption of the the non-changing statistical structure of the noise, versus the changing statistical nature of the observations (change in mean, variance, etc.).

In the *second* approach the observed changes in the time series are modeled in a statistical manner. Therefore, the noise is part of the modeling. Hence, the statistical nature of the changes can be modeled as changes in the noise characteristics.

Several testing methods will be described briefly, while the main part of the dissertation will focus on a subset of them.

a. Compound Scalar Testing (χ^2-type off-line test)

Consider a single scalar test statistic

$$\mathbf{e}^T(n) \cdot \mathbf{S}_e^{-1} \cdot \mathbf{e}(n) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda.$$

where $\mathbf{e}(n)$ is the residual vector and $\mathbf{S}_e$ is the covariance matrix of the vector $\mathbf{e}$. Then, under the no change hypothesis, the residual vector $\mathbf{e}(n)$ consists of normal i.i.d. components. Hence, the threshold λ follows a chi-square distribution with ρ degrees of freedom (ρ being the vector size or number of residuals). Recursive chi-square tests are also available.

b. *Likelihood-oriented Methods*

The likelihood ratio approach is a general tool for change detection. Different methods can be considered (Ghosh, 1970; Willsky, 1980). For example, consider a test which compares the hypothesis H_1 of nonzero residual mean to the null hypothesis H_0 of zero mean. The decision is based on the likelihood ratio between the joint distributions of the residuals

$$S_1^n = \log \frac{P\{e(1), e(2), \ldots, e(n) | H_1\}}{P\{e(1), e(2), \ldots, e(n) | H_0\}} \qquad (1-7)$$

The numerator and the denominator, respectively, are the Probability densities of the observed time series under the two hypotheses. If the residuals are independent, then (1–7) is easy to compute. Under the hypotheses testing given by (1–2):

$$S_1^n(\mathbf{e}) = \log \prod_{i=1}^{n} \frac{P_1(e_i)}{P_0(e_i)} = \sum_{i=1}^{n} \log \frac{P_1(e_i)}{P_0(e_i)}.$$

If the residuals monitored are the innovations of a Kalman filter, then it can be shown (Anderson and Moore, 1979) that the distribution of these innovations is given by the conditional distribution of the observations x_i (conditioned by their past values), hence, (1–7) can be written in the general form

$$S_1^n = \sum_{i=1}^{n} \log \frac{P_1(x_i | x_{i-1}, \ldots, x_0)}{P_0(x_i | x_{i-1}, \ldots, x_0)}. \qquad (1-8)$$

This kind of test is called **cumulative sum** test (or cumsum test) and can be written as

$$S_1^n = \sum_{i=k}^{n} g(x_i). \tag{1-9}$$

where

$$g(x_i) = \log \frac{P_1(x_i|x_{i-1},\ldots,x_0)}{P_0(x_i|x_{i-1},\ldots,x_0)}.$$

Notice that in this case the computation of S_1^n is recursive.

The tests based on (1-8), (1-9) are **stopping rules** (i.e., tests which enable us to estimate the change time ν), based upon the knowledge of the parameterized densities before and after the change. In this case the estimated stopping time can be found by using the **maximum likelihood estimate (MI** under H_1, namely

$$\hat{\nu} = \arg\max_{1\le \nu \le n} S_\nu^n. \tag{1-10}$$

In general, the statistical properties after the change (i.e., using the parameterized format of P_1 as θ_1) are not known. Hence, the cumsum test (1–9) can be used to reach the change decision

$$\max_{1\le \nu \le n} \max_{\theta_1} S_\nu^n(\theta_0, \theta_1) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{1-11}$$

This test is called the **generalized likelihood ratio (GLR) test** (Willsky and Jones, 1976) and involves a double maximization of high computation cost. Only in special cases like additive changes in linear systems modeled in the state-space representation, it can be shown (Willsky and Jones, 1976) that the effect of the resulting changes in the innovation vector $\mathbf{e}_n$ are also additive. Therefore, in the case of Gaussian state and observation noises, there are cases for which explicit solutions for θ_1 exist. For example, if θ_1 represents the

mean after the change. Then, the maximization over θ_1 is explicit (Basseville, 1988), resulting in *joint* estimation of the vector (ν, θ_1) by *recursive* computation of S_1^n and $\hat{v}_n$.

The theoretical optimality of the GLR has been investigated recently (Deshayes and Picard, 1986) from the off-line point of view. They show that under asymptotic exponential decay rates of the error probabilities α,β (where α is the Type 1 error probability or the false alarm probability and similarly β is the Type 2 probability or the probability of detection) and for specific families of distributions, the GLR tests are UMP.

Remarks

- The stopping rule based upon a cumsum statistic can use any general nonlinearity $g(\cdot)$. For example, instead of the probability ratio of conditioned observations as in (1-9), a probability ratio of the observations x_i can be used. In this case

$$g(x_i) = \log \frac{P_1(x_i)}{P_0(x_i)}.$$

- Both off-line and on-line implementations (using "sliding" windows) can be used. Examples for using this method for ARMA and AR models can be found in the literature (Segen and Sanderson, 1980, Basseville, 1986, and Basseville and Benveniste, 1983).

c. The Statistical Local Approach

This approach is used in order to overcome the main drawback of the GLR test, namely its computation cost due to the double maximization. This approach was introduced by Nikiforov (Nikiforov, 1986) for on-line

detection of changes in spectral characteristics of ARMA models. The original idea consists of looking for small changes in models and using a special type of Taylor's expansion of the log-likelihood function. Thus, the nonlinearity $g(\cdot)$ becomes

$$g(x_n) = \frac{d}{d\theta} P_\theta(x_n | x_{n-1}, \ldots)\Big|_{\theta = \theta_0}. \qquad (1-12)$$

Deshayes and Picard (Deshayes and Picard, 1986), showed that for the statistic $g(x_n)$ there exists a central limit theorem. Any change in θ is reflected as a change in $g(x_n)$, for which stopping rules based on cumsum tests can be designed.

d. Bayesian Oriented Methods

Bayesian oriented methods are based upon some prior statistical knowledge on the change time, or uses some knowledge on the switching model used to describe the statistical behavior of the changes. The use of hidden Markov models to describe the changes in state-space models (Shumway, 1990) is very popular, and leads to some change detection algorithms. However, in the Bayesian framework, it is very difficult to find a general solution because of the use of different cost functions or different prior assumptions. As mentioned in Section B.2 of this chapter, Shiryayev (Shiryayev, 1977) introduced a Bayesian competitor as an alternative to the Page cumsum test. Recently, Pollak (Pollak, 1985), proved an optimality property for the Shiryayev-Roberts rule.

e. Heuristics Associated with a Two-model Approach (Basseville, 1986)

This method called the "two models approach" is in fact a simplification of the GLR test. Implementation of GLR tests leads to "boundary" problems, because models are not very reliable when identified on short segments. In order to overcome this problem, the two model approach was introduced. These algorithms are less efficient than likelihood ratio methods but more efficient than the tests based upon the local approach.

D. ORGANIZATION OF THE DISSERTATION

This dissertation focuses on the on-line analysis of detection algorithms, hence, the quickest detection methods are explored. Both the non-Bayesian and Bayesian points of view are investigated but the focus is on non-Bayesian (maximum likelihood) methods. In this context, sequential analysis and a certain type of cumulative sum procedures which form a generalization of a test first studied by Page (Page, 1954) to detect a change in the distribution of random variables observed at random times are investigated. The Bayesian point of view is also included. Shiryayev (Shiryayev, 1978) results are shown to play a key role in any Bayesian approximation.

Different disorder types (Type, 2, 3, and 4) are investigated throughout the dissertation in the sequential (on-line) detection framework.

The body of this dissertation is divided into four groups as shown in Figure 1.4. Chapters II and III form the maximum likelihood solution of the detection problem while Chapter V presents the Bayesian approach. Chapter IV provides additional tools to analyze the performance of both the

non-Bayesian and Bayesian methods by using diffusion type approximations. Finally, Chapter VI presents a MAP estimator to a Type 4 problem, namely, discontinuity type disorder.

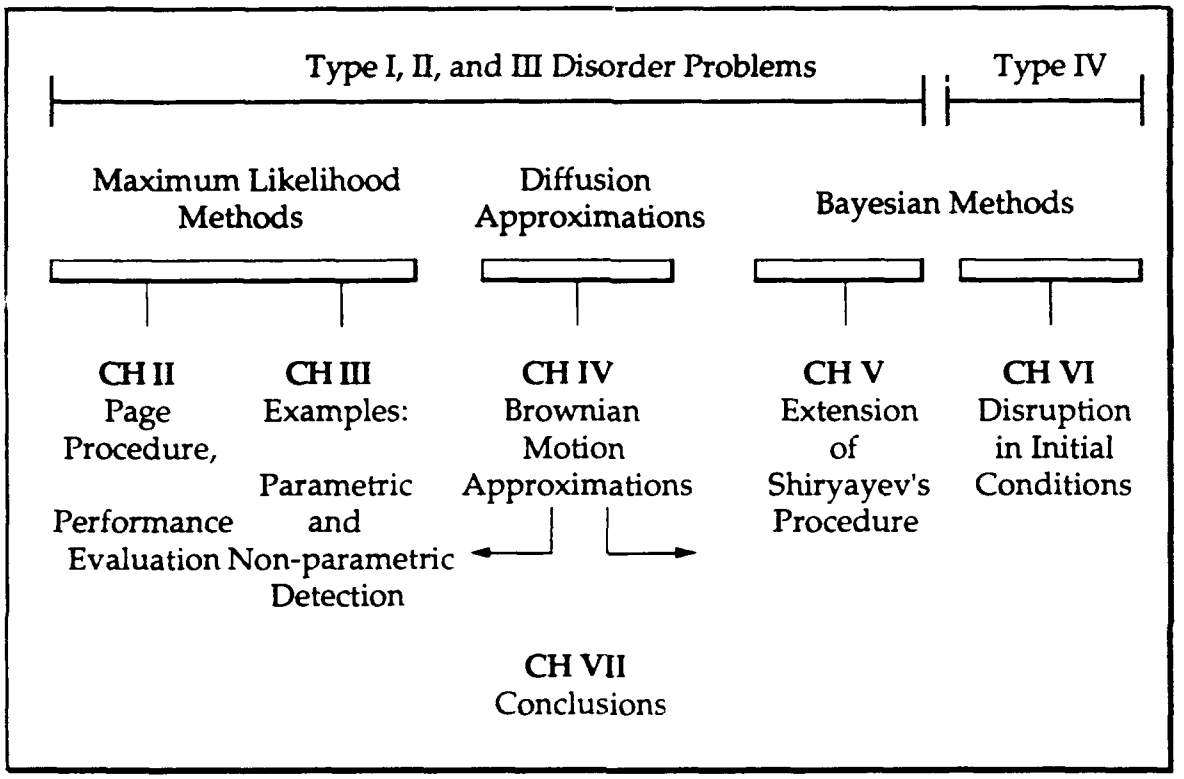


Figure 1.4. Sequential Methods for Quickest Disorder Detection

Each chapter includes an introduction and a summary section which will assist in relating all the topics presented throughout this dissertation. An appendix which summarizes the basic concepts of hypothesis testing and detection theory is also provided.

II. SEQUENTIAL METHODS FOR QUICKEST DETECTION OF CHANGES IN PROBABILITY: THE NON-BAYESIAN FRAMEWORK

A. INTRODUCTION

Consider the observation process $\{x_n\}$ with probability density $P_\theta(x_n)$ or conditional probability density $P_\theta(x_n \mid x_{n-1}, ..., x_0)$ depending upon an unknown parameter θ. This parameter can describe two different situations: In the first situation, θ can be for example, the mean or variance of the density of the time series, and will reflect directly the statistical properties of the time series. In the second situation, using some convenient parameterization of a system or signal denoted by θ, i.e. the state-space representation or ARMA modeling, θ describes the dynamics of a system or signal.

In the context of detecting jumps (sudden changes) in the parameter set θ, we are interested in detecting changes in the dynamics, or in the statistical properties of complicated structures.

Since the jump time is unknown, the problem is twofold: detection of the change, and estimation of the change time. In this chapter we will focus on the detection problem only. As shown in the last chapter there are *different issues that must be addressed:* on-line versus off-line implementation, parametric versus non-parametric methods, etcetera. These issues were briefly presented in Chapter I, and are investigated in more detail in the context of change detection in this chapter. In particular we will be addressing the following points:

1. Off-line versus On-line Viewpoints

Even though the final goal is to implement on-line (sequential) procedures, the off-line viewpoint is significant, since it can be used to derive on-line tests. This point will be clarified in this chapter. These two viewpoints differ in: (a) problem formulation and (b) performance evaluation as related to different criteria.

In the *off-line* formulation, the change detection problem is implemented as multiple hypotheses testing, for which the Neyman-Pearson lemma is not valid so that no UMP tests exists. Thus, the criterion from this viewpoint is that of classical detection problems, namely: size and power of the test.

In the *on-line* formulation, the criteria is modified to detect a change in the parameter θ as quickly as possible. In the on-line point of view the detection is performed by a **stopping rule** of the general form

$$N = \inf\{n: S_n \geq \lambda\}$$

S_n being an appropriate test statistic (see Chapter I).

The performance of a stopping rule is evaluated by T the *mean time between false alarms* (1– 6), and by D *the delay for detection* (1– 5) as proposed by Lorden (Lorden, 1971). This is a "minimax" type of average delay referred to as the best least favorable change time.

The difference between the off-line and the on-line viewpoints is significant: whereas no optimal test does exist in the off-line framework, optimal stopping rules do exist in the on-line framework for independent identically distributed (i.i.d.) sequences with known distributions before and

after the change. Moustakides (Moustakides, 1986) extended this result to the non i.i.d. case. Since in this chapter we take the non-Bayesian approach, another difference is viewed: In the *off-line* processing we assume a uniform prior distribution over all the observation set, resulting in a likelihood detector which computes the likelihood for all possible disorder times, whereas in the *on-line* approach, the disorder time is assumed to be an unknown parameter. Lorden showed (Lorden, 1971) that a certain class of stopping rules called cumulative sum tests (cumsum) are optimal in the sense of his criteria. The cumsum tests form a rich enough family of tests, and are the focus of investigation of this chapter. In particular, the test called the **Page-Hinkley** stopping rule is investigated in depth.

2. Composite Testing

As mentioned earlier, optimal stopping rules do exist in the case of i.i.d. sequences with known distributions before and after the change. When the distribution after the change is not known, a composite framework needs to be used. This issue is addressed by using the Darmois-Koopman Distribution for a one parameter exponential family.

3. Parametric versus Non-parametric Methods

The nonlinearities or transformations $g(\cdot)$ used for the cumsum detection procedures (1–9) can have a parametric or non-parametric form (sign, rank tests, etc.). The analyses will provide a general framework which can be used for either type.

B. ORGANIZATION OF THIS CHAPTER

The main goal of this chapter is the analysis of sequential methods for change detection, namely, the cumsum procedures and in particular the Page-Hinkley stopping rule. The purpose is to set a general framework in which the transformation (nonlinearity) used can be of a general form (different parametric and non-parametric forms). Thus, the following two sections (C and D) can be considered as a "guided tour" through theorems and results needed to understand and analyze cumsum procedures and their performance (presented in Sections E and F).

In Section C, sequential tests known as one-sided and two-sided Wald tests are presented in the classical detection formulation. Some basic theorems (Wald identity) which are shown to be important for the general disorder or change detection are presented.

In Section D, the sequential tests implemented with the log-likelihood function known as the Sequential Probability Ratio Tests (SPRT) are presented. Optimal properties of these tests are shown. Performance evaluation of the one and two sided SPRT, known as Wald approximation are analyzed. Within this framework, composite SPRTs using the Koopman-Darmois family of distributions are presented. Basic performance measures in the presence of strong and weak changes are shown.

In Section E, we introduce the cumsum stopping rules in the on-line framework (using Lorden's criterion). The Page test is presented and shown to be as a repeated one-sided Wald's test. Both the off- and on-line viewpoints are presented. Observing the renewal property of cumsum tests,

using Ladder variables and results from queueing theory, new aspects of cumsum tests are addressed. The Page test is also shown to be a maximum likelihood detector. Finally, optimal properties of the Page tests are presented; this test is shown to be optimal in the sense of Lorden criteria.

Section F presents the performance evaluation of Page's test. The run length function is shown to be the primary tool needed for the analysis of delay and average false alarm rate of the test. Using the results in Sections C and D we derive two results known as Lorden's and Wald's approximations. Finally, the asymptotic performance framework is introduced and used for two important results: first, the asymptotic approximation of the run length function, and second, the generalization of Lorden's results to general nonlinearities, other than the log-likelihood transformation used in the Page stopping rule. A general framework of asymptotic performance evaluation of Page's test is provided. The resulting measure is shown to be used for any nonlinearity in the presence of various noise distributions.

Section G presents a short summary of the main results of this chapter.

C. SEQUENTIAL TESTS

An alternative approach to the fixed size tests is to fix the desired performance and allow the number of measurements to vary in order to achieve this performance.

To formulate the problem, suppose that the observations $\{x_k: k = 1, 2, ...\}$ are i.i.d. and distributed according to

$$H_0: X_k \sim P_0, \quad k = 1, 2, ...$$

versus (2-1)

$$H_1: X_k \sim P_1, \quad k = 1, 2, ...$$

where P_0 and P_1 are two possible distributions. A sequential test is defined by the pair of indicator sequences (ϕ,d) where:

$\phi = \{\phi_k: k = 0, 1, 2, ...\}$ is the **stopping rule indicator,** $(\phi: \Re^n \rightarrow \{0,1\})$,

$d = \{d_k: k = 0, 1, 2, ...\}$ is called the **terminal decision rule.**

For an observation sequence $\{x_k: k = 0, 1, 2, ...\}$ the rule (ϕ,d) makes a decision $d(x_1, x_2, ..., x_n)$ whether or not any change occurred. In particular, sequential tests can be described as follows: Continue sampling as long as $\phi(x_1, x_2, ..., x_n) = 0$, and stop when $\phi(x_1, x_2, ..., x_N) = 1$. We define two kinds of tests: two-sided and one-sided.

The **two-sided** sequential test is based on the definition of the cumulative sum:

$$S_n = \sum_{i=1}^{n} g(x_i) \qquad (2-2)$$

$$S_0 = s$$

where $g: \Re \rightarrow \Re$ is a memoryless function of the observations, and s is called the **initial score**.

We detect a change according to the following stopping rule:

$$\phi_n(x_1, x_2, ...x_n) = \begin{cases} 0 & \text{if } S_n \in (a,b) \quad \text{continue} \\ 1 & \text{if } S_n \notin (a,b) \quad \text{stop} \end{cases} \qquad (2-3)$$

where a, b are the stopping thresholds; $b < 0 < a$.

Also, the terminal decision indicator is given by:

$$d_n(x_1, x_2, \ldots x_n) = \begin{cases} 0 & \text{if } S_n \le b \quad \text{no disorder} \\ 1 & \text{if } S_n \ge a \quad \text{disorder.} \end{cases}$$

The stopping time N (sometimes called the **sample size** or the **run length** of the test) is defined as:

$$N = \inf\{n: \ S_n \notin (a,b)\}$$

and the **exit times** are defined by:

$$N_a = \inf\{n: \ S_n \ge a\} \tag{2–4}$$

$$N_b = \inf\{n: \ S_n \le b\}$$

The error probabilities for the two-sided tests are defined as:

$$\alpha = \Pr\{S_N \ge a | H = H_0\}$$

$$\beta = \Pr\{S_N \le b | H = H_1\}.$$

In classical detection theory, α is defined as the probability of false alarm, and β is the probability of miss. In terms of hypothesis testing the **acceptance zone** w_a is defined as the zone where $x_k \in w_a$ or $x_k \sim P_0$. The **rejection zone** w_r is defined by $x_k \in w_r$ or $x_k \sim P_1$ (disorder zone). The **indifference zone** w_i is defined by $x_k \in \Theta - w_a - w_r$ (See Figure 2.1).

The **one-sided test** is defined by letting: $b \to -\infty$

$$\phi_n(x_1, x_2, \ldots x_n) = \begin{cases} 0 & \text{if } S_n < a \quad \text{continue} \\ 1 & \text{if } S_n \ge a \quad \text{stop} \end{cases}$$

and (2 – 5)

$$d_n(x_1, x_2, \ldots x_n) = \begin{cases} 0 & \text{if } S_n < a \\ 1 & \text{if } S_n \ge a \quad \text{disorder} \end{cases}$$

and the stopping time is given by

$$N = \inf\{n: S_n \geq a\}$$
$$= \infty \qquad S_n < a.$$

The error probabilities are defined in this case as:

$$\alpha = \Pr\{S_N \geq a | H_0\}$$

$$\beta = \Pr\{S_N \leq a | H_1\}.$$

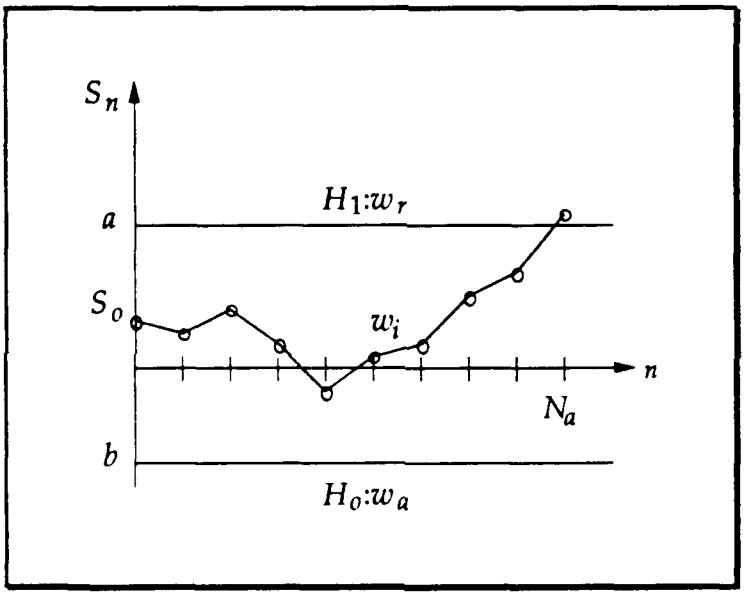


Figure 2.1. Two-sided Sequential Test

1. The Fundamental Identity (Wald's Identity) of the Sequential Analysis

This identity forms the basis of subsequent analysis for the Operating Characteristics (OC) and ARL functions of a Sequential Test (ST). It gives a convenient way to derive the moment of the sample size required to terminate the ST.

Theorem (Wald, 1947):

Let $x_1, x_2, \ldots,$ be independent random variables and let $S_n = \sum_{k=1}^{n} g(x_k)$. Let ST($a$, b, S_n) be any sequential test of H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$ based on i.i.d. $\{g(x_n)\}$, and let N be the stopping time for this test.

Let $\psi_i(h)$ denote the moment generating function of the random variable $g(x)$ under the hypothesis H_i:

$$\psi_i(h) = E\{\exp(g(x)\cdot h)|H_i\} \qquad i = 0,1$$

for every real h for which $\psi_i(h)$ is bounded. Then, if $P(g(x) = 0 \,|\, H_i) < 1$ and $P(|g(x)|<\infty \,|\, H_i) = 1$, we have:

$$E\left\{\exp(S_N \cdot h)[\psi_i(h)]^{-N} | H_i\right\} = 1 \qquad i = 0,1 \qquad (2-6)$$

The proof can be found in Ghosh (Ghosh, 1970, p. 208) or Feller (Feller, 1971, p. 603).

2. Applications of Wald's Identity

As a direct corollary to Wald's identity, immediate results for the ARL function can be obtained:

$$\text{define} \quad z = g(x),$$

then, the **average run length** (ARL) is given by Govindarajulu (Govindarajulu, 1975):

$$\begin{aligned} E\{N|\theta_i\} &= \frac{E\{S_N|\theta_i\}}{E\{z|\theta_i\}} \qquad \text{if } E\{z|\theta_i\} = \psi_i'(0) \neq 0 \\ E\{N|\theta_i\} &= \frac{E\{S_N^2|\theta_i\}}{E\{z^2|\theta_i\}} \qquad \text{if } E\{z|\theta_i\} = \psi_i'(0) = 0 \end{aligned} \qquad (2-7)$$

Bounds on the stopping thresholds can be associated with the ST(a, b, g)

$$a \le \log\frac{1-\beta}{\alpha}$$
$$b \ge \log\frac{\beta}{1-\alpha}. \qquad (2-8)$$

The strict equalities hold if and only if $b \le 0 < a$, and in terms of the error probabilities for $|a, b| > 0$:

$$\alpha \le (1-\beta)e^{-a}$$
$$\beta \le (1-\alpha)e^{b}. \qquad (2-9)$$

These approximation are known as **Wald approximations** and were derived by ignoring the "excess over the boundaries" (Siegmund, 1985). Notice that we can get yet cruder inequalities when we consider the asymptotic case where $\alpha \downarrow 0$, $\beta \downarrow 0$. Then:

$$\alpha \le e^{-a} \quad \beta \le e^{b}.$$

3. Comparison of Sequential Tests (ST) and Fixed Sample Size Tests (FSST)

Our object is to investigate the number of samples saved by an ST(a,b,g) over the corresponding optimum FSST, both designed to achieve the same performance (α,β).

The **relative efficiency** of ST(a,b,g) at θ is defined (Ghosh, 1970) by

$$RE(\theta) = \frac{n(\alpha,\beta)}{E\{N|\theta\}}$$

where $n(\alpha,\beta)$ is the sample size required by FSST test and $E\{N|\theta\}$ is the ARL function of the ST test, both designed to achieve the same performance (α,β). It can be shown (Poor, 1988) that for the case of a simple sequential detection of a constant signal in the presence of white Gaussian noise, using a likelihood ratio detector for both the ST and the FSST, the limiting RE is given by

$$\lim_{\alpha=\beta\to 0} RE = 4.$$

Thus for vanishingly small error probabilities (with $\alpha = \beta$) the SPRT requires on the average only one-fourth as many samples as does the FSST test. Further discussion can found in Ghosh (Ghosh, 1970).

D. SPRT TESTS

When the test procedure given by (2-2 and 2-3) uses the the log-likelihood ratio as the nonlinearity $g(x)$

$$g(x) = \log \frac{dP(x|\theta_1)}{dP(x|\theta_0)}$$

the sequential test is called the **sequential probability ratio test** (SPRT). The relation between any ST(a,b,g) to SPRT (A,B) is given by

$$A = e^a \quad B = e^b$$

where

$$b < 0 < a \quad \text{and} \quad 0 < B \le 1 \le A.$$

The bound approximations (2-8), (2-9) can be converted to SPRT test terms by placing $e^a = A$ and $e^b = B$.

The SPRT test has a fundamental property which is extremely important and will be used in the sequel (Therrien, 1989):

$$\begin{aligned} &\text{under disorder:} && E\{g(x_i) \mid \theta_1\} \geq 0 \\ &\text{under no disorder:} && E\{g(x_i) \mid \theta_0\} \leq 0. \end{aligned} \tag{2-10}$$

In the following sections, several properties of SPRT test will be presented, and the two- and one-sided SPRT tests will be analyzed, followed by the composite hypothesis framework for SPRT.

1. Optimal Properties of SPRT

For testing a simple hypothesis against a simple alternative with i.i.d. observations, the SPRT test is optimal among all sequential and fixed sample size tests in the sense of minimizing the expected run length both under H_0 and under H_1 among all the tests having no large error probabilities. The following theorem establishes this result.

The Wald-Wolfowitz Theorem (1948):

Among all tests (FSST and ST) for which $\Pr\{\text{accept } H_1 \mid H_0\} \leq \alpha$, and $\Pr\{\text{accept } H_0 \mid H_1\} \leq \beta$ and for which $E\{N \mid \theta_i\} \leq \infty \quad i = 0,1$; the SPRT with error probabilities α and β minimizes both $E\{N \mid \theta_0\}$ and $E\{N \mid \theta_1\}$. □

The proof can be found in (Ghosh, 1970). The optimal property of the SPRT test can be viewed as analogous to the Neyman-Pearson lemma.

Definition (Wijsman, 1960):

A SPRT is said to have a **monotonicity** property if when the upper stopping bound of the SPRT is increased and the lower bound is decreased, then at least one of the error probabilities decreases.

Theorem (Lehmann, 1959)

Let $x_1, x_2, \ldots$ be independent random variables having probability density $P(x;\theta)$ which has monotone likelihood ratio. Then any SPRT for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$ $(\theta_0 \leq \theta_1)$ has a nondecreasing power function. □

The proof can be found in Lehmann (Lehmann, 1986).

2. The Termination Property of SPRT

The SPRT test is a **closed** test if and only if, the termination property holds for every $\theta \in \Theta$. When $g(x_i)$ are i.i.d. any SPRT is closed under the *following mild restriction* (Poor, 1988):

suppose that for any $\theta \in \Theta$, $g(x_i)$ are i.i.d. random variables and $P(g(x_i) \mid \theta) < 1$, then:

- $\lim_{n\to\infty} \Pr\{n > N|\theta\} = 0.$
- there exists a $t_0>0$ such that the *moment generating function* $E\{e^{nt} \mid \theta\}$ exists for all real $t< t_0$.

This means that the entire statistics of n can be found. The result is that the SPRT or the $ST([\beta/(1-\alpha)],[(1-\beta/\alpha])$ based on Wald's approximations are always closed. Ghosh (Ghosh, 1970) extended the result to the situation $g(x_i)$ are not i.i.d.

Another optimal property of the SPRT was shown by Wald which established a lower bound on the ARL of competitors of the SPRT. This result will be presented in the sequel (2–24) when presenting the problem of composite hypothesis testing.

3. The Operational Characteristics (OC) and AKL Functions of Two-sided SPRT Tests

The use of Wald's identity (2–6) forms the basis of certain bounds for the OC function $Q(\theta)$ and ARL functions for the SPRT.

Wald's Approximations

Wald's approximations are based on the use of the moment generating function of $g(x)$ (2–6) provided that we can find two nonzero real numbers h_0 and h_1 such that

$$\psi_i(h_i) = E\{\exp(h \cdot g(x))|\theta_i\} = 1 \quad i = 0,1. \tag{2-11}$$

Existence and uniqueness of such roots are guaranteed when $g(x)$ has a nonzero mean and satisfies certain other conditions (Feller, 1971).

The key results for our purposes is that $\psi_i(h) = 1$ has:

- one and only one nonzero root

$$-\infty < h(\theta) < 0 \quad \text{if} \quad E\{g(x)|\theta\} = E_\theta\{g(x)\} > 0$$

$$0 < h(\theta) < \infty \quad \text{if} \quad E\{g(x)|\theta\} = E_\theta\{g(x)\} < 0. \tag{2-12}$$

- No non zero real root if $E\{g(x) \mid \theta\} = 0$.

When we try to detect a change from a negative trend $E\{g(x) \mid \theta_0\} < 0$ to a positive trend $E\{g(x) \mid \theta_1\} > 0$ then, it implies that $h(\theta_1) < 0 < h(\theta_0)$.

Notice that the roots are functions of two parameters: the probability density of the observations $P(x)$ and the nonlinearity $g(x)$. The approximate

formulas of SPRT test ST(a,b) for the OC and ARL are derived when $g(x_i)$ are i.i.d. and $b < 0 < a$, using the assumption of no excess of S_n over a and b.

Lower bounds for the ARL function of two-sided SPRT test under hypotheses H_0 and H_1 in terms of the error probabilities are given by Wald (Wald, 1947):

$$L(\theta_0) = E_0\{N\} \geq \frac{(1-\alpha)\log(\beta/1-\alpha)+\alpha\log(1-\beta/\alpha)}{E_0\{g(x_1)\}} \qquad \text{if } \theta = \theta_0.$$

$$L(\theta_1) = E_1\{N\} \geq \frac{\beta\log(\beta/1-\alpha)+(1-\beta)\log(1-\beta/\alpha)}{E_1\{g(x_1)\}} \qquad \text{if } \theta = \theta_1. \tag{2-13}$$

Bounds for the operational characteristic function $Q(\theta)$ are given by Ghosh (Ghosh, 1970):

- For detecting a change of positive trend (upward change) $h(\theta_0) > 0$:

$$\frac{\exp\{h(\theta)\cdot a\}-1}{\exp\{h(\theta)\cdot a\}-\eta(\theta)\exp\{h(\theta)\cdot b\}} \leq Q(\theta) \leq \frac{\delta(\theta)\exp\{h(\theta)\cdot a\}-1}{\delta(\theta)\exp\{h(\theta)\cdot a\}-\exp\{h(\theta)\cdot b\}}$$

- For detecting a change of negative trend (downward change) $h(\theta_0) < 0$:

$$\frac{\exp\{h(\theta)\cdot a\}-1}{\exp\{h(\theta)\cdot a\}-\delta(\theta)\exp\{h(\theta)\cdot b\}} \leq Q(\theta) \leq \frac{\eta(\theta)\exp\{h(\theta)\cdot a\}-1}{\eta(\theta)\exp\{h(\theta)\cdot a\}-\exp\{h(\theta)\cdot b\}}$$

where (2–14)

$$\eta(\theta) = \inf_{1<\xi<\infty} \xi E\left\{\exp\{h(\theta)\cdot g(x)\}\middle|\exp\{h(\theta)\cdot g(x)\} \leq \frac{1}{\xi};\theta\right\} \leq 1$$

$$\delta(\theta) = \inf_{0<\xi<1} \xi E\left\{\exp\{h(\theta)\cdot g(x)\}\middle|\exp\{h(\theta)\cdot g(x)\} \leq \frac{1}{\xi};\theta\right\} \geq 1.$$

Recall that for any test, the OC should result in $Q(\theta_0) \geq 1-\alpha$, and $Q(\theta_1) \leq \beta$ (see Appendix). Thus, the motivation is to find bounds for the ARL in terms of the OC function and the stopping thresholds a,b. These upper and lower bounds for the ARL are given by Ghosh (Ghosh, 1970):

$$L(\theta) = E_\theta\{N\} \begin{cases} \leq \dfrac{[a+\gamma(\theta)][1-Q(\theta)]+bQ(\theta)}{E\{g(x)|\theta\}} & E\{g(x)|\theta\} > 0 \\ \geq \dfrac{a^2[1-Q(\theta)]+b^2Q(\theta)}{E\{g^2(x)|\theta\}} & E\{g(x)|\theta\} = 0 \\ \geq \dfrac{[a+\gamma(\theta)][1-Q(\theta)]+bQ(\theta)}{E\{g(x)|\theta\}} & E\{g(x)|\theta\} < 0 \end{cases} \quad (2-15)$$

where $\gamma(\theta) = \sup_{r>0} E\{g(x_1) - r | g(x) \geq r > 0; \theta\}$ is the "excess over the boundary." The mean time between false alarms T is given by $L(\theta_0)$ while the delay for detection is given by $L(\theta_1)$.

Detecting a change from a negative trend $E\{g(x) \mid \theta_0\} < 0$ to a positive trend $E\{g(x) \mid \theta_1\} > 0$ (upward change) can result in effective bounds for $L(\theta)$. Notice that $Q(\theta_0) \geq 1-\alpha$ and $Q(\theta_1) \leq \beta$, result in consistent inequality directions in (2-15). Thus, upper bounds can be evaluated in the case of disorder detection. Similarly, bounds for $L(\theta)$ in the case of detecting a change from a positive trend to a negative one (downward change) can be found by reversing the inequalities in (2–15).

4. The Operational Characteristic OC and ARL Functions for One-sided SPRT Tests

For detecting a change from a negative to positive trend, the probability that the one-sided test does not stop under θ_0, can be found by using the limit of the two-sided OC function, as b tends to negative infinity. Thus, this probability is lower bounded by:

$$\begin{aligned}\Pr\{L(\theta_o)=\infty\} &= \lim_{b\to-\infty} Q(\theta_0) \\ &\geq \lim_{b\to-\infty} \frac{\exp\{h(\theta_0)\cdot a\}-1}{\exp\{h(\theta_0)\cdot a\}-\eta(\theta_0)\exp\{h(\theta_0)\cdot b\}} \\ &= \frac{\exp\{h(\theta_0)\cdot a\}-1}{\exp\{h(\theta_0)\cdot a\}}.\end{aligned}$$

Notice that the obtained lower bound avoids the use of the functions $\delta(\theta)$ and $\eta(\theta)$ which are difficult to generalize.

The probability that the one-sided test terminates under θ_0 which is the size (α) of the one-sided test is upper bounded by

$$\begin{aligned}\alpha = \Pr\{S_N \geq a | H_0\} \quad &= \Pr\{L(\theta_0)<\infty\} \\ &= 1-\Pr\{L(\theta_0)=\infty\} \qquad (2-16) \\ &\leq \exp\{-h(\theta_0)\cdot a\}.\end{aligned}$$

This result is very important and will be used in the sequel when analyzing the cumsum procedures due to Lorden's criteria.

An upper bound for the ARL function of the one-sided test under θ_1 (Delay for detection) is obtained by using the upper bound for the ARL of the two-sided test (2-15):

$$L(\theta_1) = E_1\{N\} \le \lim_{b \to -\infty} \frac{[1+\gamma(\theta)][1-Q(\theta_1)] + bQ(\theta_1)}{E\{g(x)|\theta_1\}}.$$

Since $b \to -\infty$, the OC function $Q(\theta_1)$ is a decreasing function (monotonicity property of the SPRT function), $E\{g(x) \mid \theta_1\} > 0$ (detecting a positive trend), hence, the right-hand side is a decreasing function and the inequality is preserved as $Q(\theta_1) \to 0$, resulting in:

$$L(\theta_1) = E_1\{N\} = E\{N|\theta_1\} \le \frac{a+\gamma(\theta_1)}{E\{g(x)|\theta_1\}} \qquad E\{g(x)|\theta_1\} > 0. \qquad (2-17)$$

5. SPRT for Composite Hypotheses

Although the SPRT was derived from a test of a simple hypothesis against a simple hypothesis, it was shown that from the on-line point of view of detecting abrupt changes, optimal stopping rules do exist in the case of i.i.d. observations with known distributions before and after the change. When the distribution after the change is not known, some other hypotheses can be considered. Thus, it is natural to consider to test for example H_0: $\theta < \theta^*$ against H_1: $\theta > \theta^*$.

Wald (Wald, 1947) considered the method of weight functions in order to deal with unknown composite alternatives where the alternative may be a parameter within a surface (Rejection Region). If the method of weight function is not feasible, so-called **open-ended (one-sided) likelihood ratio test** procedures can be considered. Lorden (Lorden, 1971) investigated that approach for the problem of open ended (one-sided) tests for the one parameter exponential **Darmois-Koopman families of distributions**. This approach leads to easily computed procedures to obtain approximations to the

detection probability and ARL functions of the SPRT of composite hypotheses using only the theory developed for simple hypotheses. The following shows that this is generally possible in the context of a one parameter exponential family, and will form the base for Lorden's cumsum procedure.

a. Composite Testing for Darmois-Koopman Distribution Families (Siegmund, 1985)

Consider a general SPRT test defined by (2-2) and (2-3) with the additional assumption that $x_1, x_2, \ldots$ are i.i.d., so that

$$S_1^n = \log \prod_{i=1}^{n} \frac{P_1(x_i)}{P_0(x_i)}.$$

Next we follow Siegmund's analysis (Siegmund, 1985) to derive a new observation. Let P, P^* be third and fourth density functions, such that the original test of P_0 against P_1 is equivalent to a test of P against P^* with new stopping boundary values, such that

$$\frac{P^*(x)}{P(x)} = \left[\frac{P_1(x)}{P_0(x)}\right]^{\theta_1} \qquad \theta_1 \neq 0. \qquad (2-18)$$

Note that $P^*(x)$ must satisfy

$$\int_{-\infty}^{\infty} \left[\frac{P_1(x)}{P_0(x)}\right]^{\theta_1} P(x)dx = 1.$$

Define

$$z(x) = \log \frac{P_1(x)}{P_0(x)}$$

hence

$$\int_{-\infty}^{\infty} e^{z(x)\theta_1} P(x)dx = 1.$$

Now we define a function $b(\theta)$ such that

$$\int_{-\infty}^{\infty} e^{z(x)\theta} P(x)dx = e^{b(\theta)} \neq 1 \qquad (2-19)$$

where θ represents the statistical "distance" between the null and the alternative hypotheses. Notice that $b(\theta_1) = b(0) = 0$. If the last integral converges then

$$\int_{-\infty}^{\infty} e^{z(x)\theta - b(\theta)} P(x)dx = 1$$

and

$$P_\theta(x) = e^{z(x)\theta - b(\theta)} P(x)$$

and

$$\frac{P_\theta(x)}{P(x)} = e^{z(x)\theta - b(\theta)} \qquad (2-20)$$

represents the new test since $\frac{P_\theta(x)}{P(x)} = \left[\frac{P_1(x)}{P_0(x)}\right]^\theta$. The resulting test defines a one-parameter exponential family of distributions under which composite tests can be evaluated easily.

Differentiation of (2-19) w.r.t θ gives:

$$b'(\theta) = \int_{-\infty}^{\infty} z(x) \cdot P_\theta(x)dx = E_x\{z(x)\}$$

and

$$b''(\theta) = \int_{-\infty}^{\infty} [z(x)]^2 P_\theta(x)dx - [b'(\theta)]^2 = \mathrm{var}_x(z(x)) \geq 0 \qquad (2-21)$$

so $b(\theta)$ is convex. The desired $\theta_1 \neq 0$ satisfying (2-18) exists if and only if

$$b'(0) = \int_{-\infty}^{\infty} z(x)P(x)dx = E_x\{z(x)\} \neq 0$$

since $b(\theta)$ is convex and $b(\theta_1) = b(0) = 0$ (see Figure 2.2).

The original test of P_0 against P_1

$$N = \inf\left\{n: \quad \prod_{k=1}^{n} \frac{P_1(x_k)}{P_0(x_k)} \notin (a,b)\right\} \tag{2-22}$$

is equivalent to a test of

$$N = \inf\left\{n: \quad \prod_{k=1}^{n} \frac{P^*(x_k)}{P(x_k)} \notin \left(a^{\theta_1}, b^{\theta_1}\right)\right\}.$$

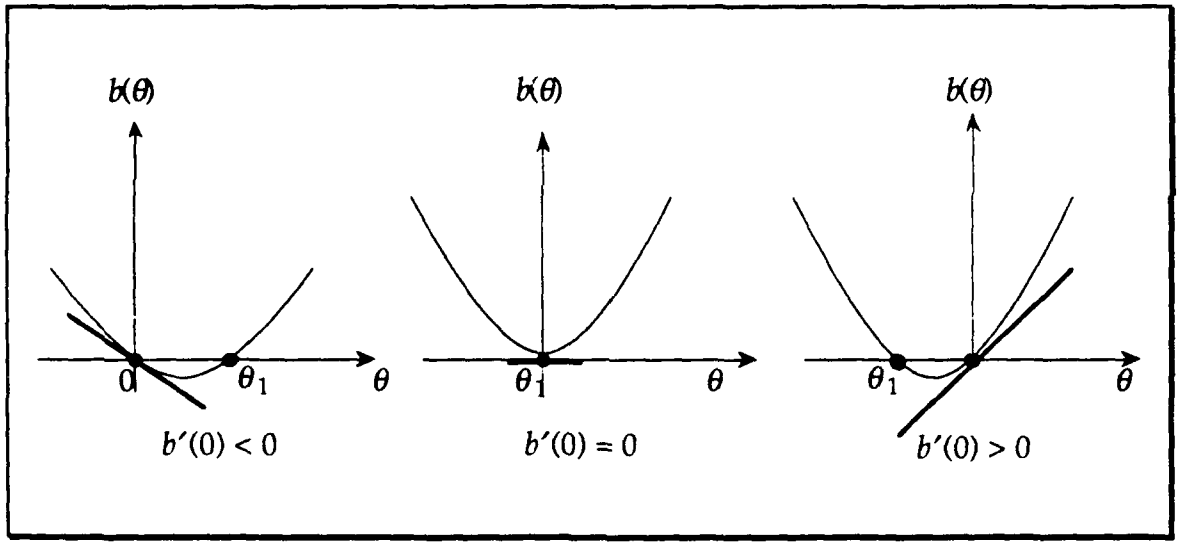


Figure 2.2. The $b(\theta)$ Function (from Siegmund, 1985)

Since $P(x)$ represents the null hypothesis under (2-18), it is clear that for composite testing, $b'(0) = E\{z(x_1)\} = E\{z(x)|H_0\} \neq 0$ implies that under the null hypothesis (no disorder) the test should give a negative trend $b'(0) < 0$ when detecting a change from a negative to positive trend (see Figure 2.2). This result is consistent with another one which is presented in the sequel,

namely, that it is worthwhile to bias the detector if it is known that before the disorder occurs, the test will have zero mean. This gives some degree of robustness to the test under composite hypothesis testing.

b. Performance Evaluation

The following proposition establishes an important result about the performance of the SPRT within the composite framework. This result will be shown to play a key role in Lorden's work about the optimality of Page's test in the on-line framework (minimum average delay for detection), by assigning a lower bound on the ARL for competitors of Page's test.

Proposition (Wald, 1947):

Given a two-sided sequential test of H_0:$X \in \theta_0$ against H_1:$X \in \theta$, suppose N_1 and N_2 are stopping times for $x_1, x_2, \ldots \in X$ such that:

$$P_{\theta_0}(N_1 < \infty) \le \alpha < 1 \quad \text{and} \quad P_{\theta}(N_2 < \infty) \le \beta < 1$$

where α and β are the false alarm and miss probabilities (respectively).

Define:

$$I(\theta, \theta_0) = E_\theta\left\{\log\left[\frac{f_\theta(x)}{f_{\theta_0}(x)}\right]\right\} = \int_{-\infty}^{\infty} \log\left[\frac{f_\theta(x)}{f_{\theta_0}(x)}\right] dP_\theta \qquad (2-23)$$

this is the **information number** or the **Kullback-Liebler number**. Then:

- $I(\theta, \theta_0) \cdot E_\theta\{\min(N_1\ N_2)\} \ge (1-\beta)\ln\alpha^{-1} - \ln 2$
- and for $N_2 \to +\infty$(one - sided test) and $\beta \downarrow 0$, (2–24)

$$L(\theta) = E_\theta\{N_1\} \ge \frac{\ln\alpha^{-1}}{I(\theta, \theta_0)} \qquad \square$$

The proof can be found in Wald's book (Wald, 1947, p. 197)

Remark: The last proposition gives a lower bound on the average delay for detection $D = E\{N \mid H_1\}$ for *any* stopping rule for which $\alpha = P_0(N<\infty) < 1$ (see Figure 2.3). For the SPRT we have the approximate relations (2-13) between the ARL and the error probabilities. The last proposition generalizes (2-17) for composite tests, in asserting that the ARL function in (2-13) is approximately minimal.

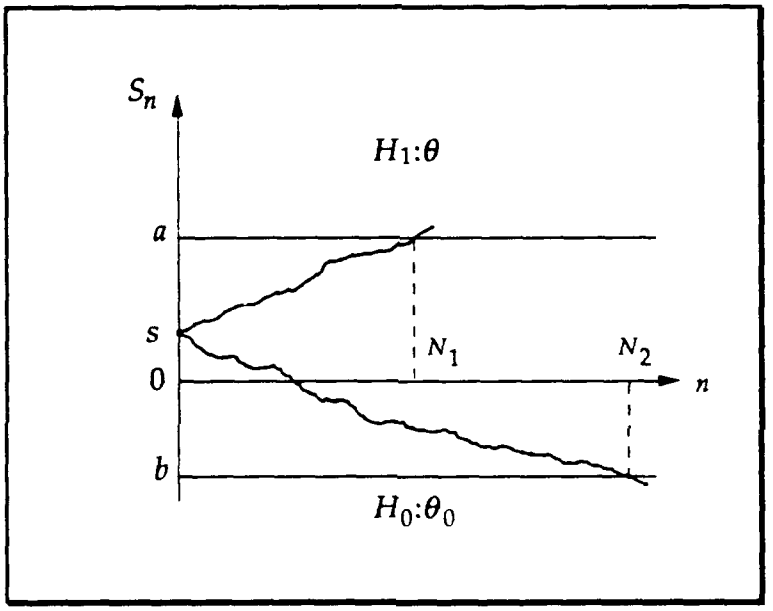


Figure 2.3. Sequential Test Exit Times

c. *Performance of Sequential Composite Tests in the Presence of a Weak Signal*

In classical detection theory, the **locally optimum** detector maximizes the slope of the power curve with respect to signal strength (evaluated at zero signal strength or at the presence of a weak signal) for a fixed false alarm (Neyman-Pearson locally optimum procedure). In this composite alternative hypotheses approach, the alternatives θ are close in the

sense of metric or distance to the null hypothesis θ_0. It can be shown (Kassam, 1987) (Poor, 1988) (Kazakos, 1977) that the locally optimum detector in the classical detection problem of H_0: $x_i \sim P(x \mid \theta_0)$ versus H_1: $x_i \sim P_1(x) = P(x \mid \theta)$ for $\theta > \theta_0$ is given by:

$$g_{\ell o}(x) = -\frac{\frac{d}{d\theta}P(x|\theta)\Big|_{\theta\to\theta_0}}{P(x|\theta_0)} = -\frac{P_0'(x)}{P_0(x)}$$

$$= -\frac{d}{d\theta}\ln\{P(x|\theta)\}\Big|_{\theta=\theta_0} \qquad (2-25)$$

where $\theta - \theta_0$ indicates the "distance" between H_1 and H_0, and $\theta \to \theta_0$ indicates a weak signal situation, resulting in the locally most powerful (LMP) nonlinearity $g_{\ell o}$. For the ST defined by (2-2) we can define the Signal-to-Noise Ratio (SNR) (Kassam, 1988):

$$\text{SNR} \triangleq \frac{\left(E\{S_n|\theta\}\right)^2}{\text{var}\{S_n|\theta_0\}}.$$

We seek to maximize the SNR when $\theta \to \theta_0$.

The **efficacy** ε of a test is defined (Kassam, 1988) as the limiting incremental signal to noise ratio:

$$\varepsilon(g) = \lim_{\substack{n\to\infty \\ \theta\to\theta_0}} \frac{\left(\frac{d}{d\theta}E\{S_n|\theta\}\Big|_{\theta=\theta_0}\right)^2}{n\cdot\text{var}\{S_n|\theta_0\}}$$

$$= \frac{\left(E\{g'(x)|\theta_0\}\right)^2}{\text{var}\{g(x)|\theta_0\}}.$$

The nonlinearity $g(x)$ that maximizes the efficacy is the local optimum nonlinearity $g_{\ell o}(x) = -P_0'(x) / P_0(x)$ which is also the Neyman-Pearson locally optimum procedure. In this case the efficacy is equal to Fisher's information for a location shift test, namely: H_0: $x_i \sim f(x)$ versus H_1: $x_i \sim f(x-\theta)$, and is given by (Kassam, 1987):

$$\varepsilon(g) = \int_{-\infty}^{\infty} \left[\frac{P_0'(x)}{P_0(x)} \right]^2 P_0(x) dx.$$

6. Practical Criticism of the SPRT and Truncated Tests

The optimality property of the SPRT is a remarkably strong property but it applies only to simple hypotheses. Even for the simple case of constant signal detectors as shown in (Poor, 1988), it is necessary to know the signal value in order to implement the test. This is in contrast to the Fixed Sample Size tests which are UMP for $\theta \geq 0$. For applications involving composite testing, the open continuation region can lead to very large sample sizes, especially when $E\{\log[f_1(x)/f_0(x)]\} \cong 0$. Thus, although the ARL of the SPRT is finite with probability 1, it is not bounded. This difficulty can be overcome by modifying the SPRT to stop sampling and make a hard (single-tnreshold) decision after the ARL has reached some maximum number of samples. This type of test is known as the **truncated test** and can be described as follows: The sequential test is defined by

$$N = \inf\{n:\ S_n \notin (a,b)\}.$$

In the absence of a definite upper bound on N, we define an upper bound M. Hence, the new (truncated) stopping rule is given by

$$\min(N,M).$$

Another problem associated with the SPRT is the estimation following detection. If we want to stop sampling as soon as it is possible to tell in which of two subsets of the parameter space a parameter lies, then usually, the estimation procedure will require an adequate number of samples which is larger than the sample size. A possible solution is to artificially enforce a larger sample size. However, sequentially stopped versions of the estimators are biased, while we would like to consider unbiased estimators. The problem of estimation following sequential tests is not a part of this work. However, in the disorder detection framework, we are interested in randomly stopped averages $\sum_{i=1}^{m} g(x_i)$ where m is a random variable. The Anscombe-Doeblin theorem (Siegmund, 1985) shows that such averages are asymptotically normal under quite general conditions.

E. CUMSUM PROCEDURES

Assuming that a given process has i.i.d. observations $x_1, x_2, \ldots$, whose distribution possibly changes from P_0 to P_1 at an unknown point in time ν, then, in the hypothesis testing framework the problem can be presented as:

$$\begin{aligned} H_0: &\quad x_1, x_2, \ldots && \sim P_0 \\ &\qquad \text{versus} \\ H_\nu: &\quad x_1, x_2, \ldots, x_{\nu-1} && \sim P_0 \qquad \nu \geq 1 \\ &\quad x_\nu, x_{\nu+1}, \ldots && \sim P_1 \end{aligned} \qquad (2-26)$$

Let P_ν and E_ν denote the probability measure and the expectation under P_ν respectively, when the change from P_0 to P_1 occurs at the ν^{th} sample (ν=1, 2, ...). Let P_0 denote the probability that there is no change, i.e., $\nu = \infty$.

Using Lorden's definition (Lorden, 1971):

$$D = \bar{E}_1\{N\} = \sup_{\nu \geq 1} \operatorname{ess\,sup} E_\nu\left\{(N-\nu+1)^+ \middle| x_1, x_2, , x_{\nu-1}\right\}$$

$$\text{where: } (a)^+ = \max(0,a) \qquad (2-27a)$$

which is the worst possible or the least favorable conditional, expected detection delay or quickness of reaction to a change (disorder). Thus, a "minimax" type of criterion is defined for which the delay D is the smallest value such that for every $\nu \geq 1$

$$E_\nu\left\{(N-\nu+1)^+ \middle| x_1, x_2, \ldots, x_{\nu-1}\right\} \leq \bar{E}_1\{N\}$$

almost surely under F_0.

The goal is to find the stopping time N which allows the quickest detection of the change, subject to:

$$E_0\{N\} \geq \gamma. \qquad (2\text{–}27b)$$

The constraint implies that if the change does not occur, then the expected time for false alarm is no less than the threshold of γ (where $\gamma \to \infty$ asymptotically).

Several ad hoc proposals to solve this **multiple hypothesis** problem that at least one of the H_ν hold ($1 \leq \nu \leq n$) against H_0. The most well known procedures are the Page-Hinkley and Shiryayev-Roberts tests, and will be presented in the sequel. Both are based on the probability ratio, hence, presuming the properties presented in the last section.

1. The Page Cumsum Test (Page, 1954)

Page's procedure has two equivalent implementations: Recursive test which can be considered as a repeated modified one-sided SPRT test with

constant stopping limits, and a Repeated SPRT with a moving indifference zone.

a. Repeated SPRT with Moving Indifference Zone

Consider the test:

$$g_n = \max_{0\le k\le n}\left(\tilde{S}_n - \tilde{S}_k\right) = \tilde{S}_n - \min_{0\le k\le n}\tilde{S}_k$$

where

$$\tilde{S}_n = \sum_{i=1}^{n}\log\frac{P_1(x_i)}{P_0(x_i)} \qquad (2-28)$$

$$\tilde{S}_0 = 0$$

The indifference interval equals to $(0,a)$. The stopping rule based upon (2-28) is defined as

$$N^* = \inf\left\{n:\ \tilde{S}_n - \min_{0\le k\le n}\tilde{S}_k \ge a\right\} \qquad (2-29)$$

Note that $g_n = \tilde{S}_n - \min_{0\le k\le n}\tilde{S}_k$ measures the current height of the random walk $\tilde{S}_k$, $k = 0, 1, 2, \ldots$ above its minimum value. Whenever the random walk establishes a new minimum, the process forgets its past and starts again in the sense of a **renewal process** (see Figure 2.4):

$$\tilde{S}_{n+j} = \min_{0\le k\le n+j}\tilde{S}_k = \tilde{S}_{n+j} - \tilde{S}_n - \min_{0\le k\le j}\left(\tilde{S}_{n+k} - \tilde{S}_n\right) \qquad (2-30)$$

This renewal property has important consequences. It means that N^* can be defined in terms of a sequence of one-sided SPRTs as follows:

$$N_1 = \inf\{n: \ \tilde{S}_n \notin (0,a)\}$$

if $S_{N_1} \geq a$, then: $N^* = N_1$.

otherwise: $\tilde{S}_{N_1} = \min_{0 \leq k \leq N_1} S_k$ and define:

$$N_2 = \inf\{n: \ \tilde{S}_{N_1+n} - \tilde{S}_{N_1} \notin (0,a)\} \tag{2-31}$$

if $\tilde{S}_{N_1+N_2} - \tilde{S}_{N_1} \geq a$, then: $N^* = N_1 + N_2$

otherwise: $\tilde{S}_{N_1+N_2} = \min_{0 \leq k \leq N_1+N_2} \tilde{S}_k$

and in general,

$$N_k = \inf\{n: \ \tilde{S}_{N_1+\ldots+N_{k-1}+n} - \tilde{S}_{N_1+\ldots+N_{k-1}} \notin (0,a)\}$$

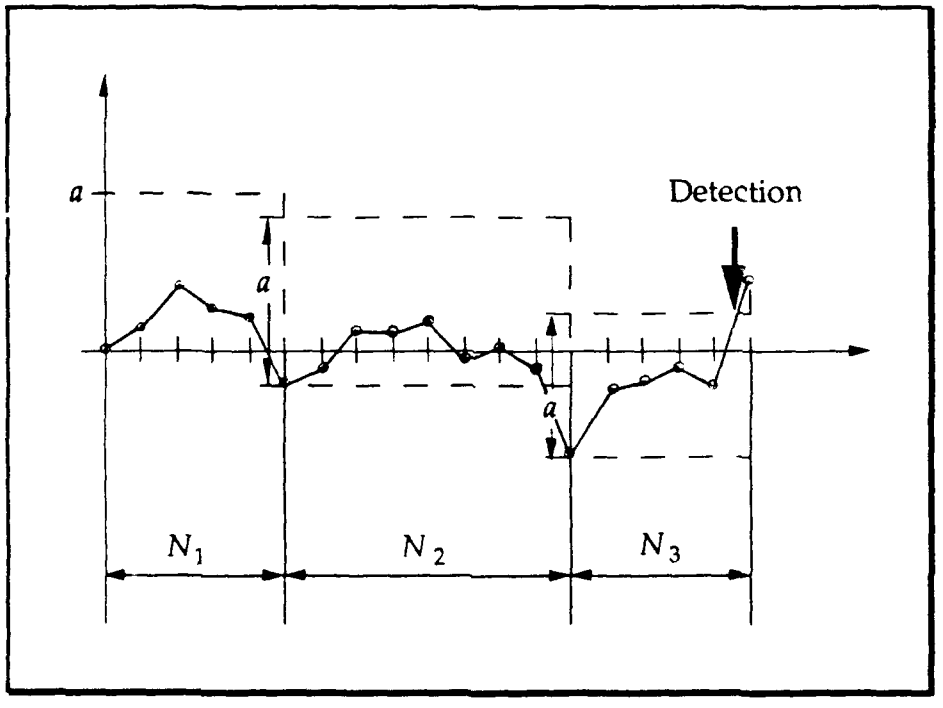


Figure 2.4. Page Test (Moving Indifference Zone)

The overall stopping time is given by:

$$N^* = N_1 + N_2 + \ldots + N_M \tag{2–32}$$

and is called the **extended stopping time**, since it consists of sum of single SPRT stopping times, where:

$$M = \inf\left\{m: \ \tilde{S}_{N_1+\ldots+N_m} - \tilde{S}_{N_1+\ldots+N_{m-1}} \geq a\right\} \tag{2-33}$$

is the number of the repetitions (renewals). By (2-33), M is geometrically distributed (see also (Siegmund, 1985)) with:

$$E\{M\} = 1 / \Pr\left\{\tilde{S}_{N_1} \geq a\right\}.$$

Define: $N^* = \sum_{i=1}^{M} N_i$, and using Wald's identity (2 – 7) we obtain

$$E\{N^*\} = E\{N_1\} \cdot E\{M\} = \frac{E\{N_1\}}{\Pr\left\{\tilde{S}_{N_1} \geq a\right\}} \tag{2-34}$$

which expresses the extended stopping time in terms of the expected stopping size and error probability of a *single SPRT*.

b. Recursive Implementation

As will be shown, the recursive implementation has two interpretations.

The *first* is the relation to the repeated one-sided Wald sequential test with boundaries 0 and a, which forms a renewal process whenever the random walk $\tilde{S}_n$ hits the lower boundary 0 (see Figure 2.5). The renewal process is repeated until such time that a Wald test exceeds the threshold a. Thus, the process $\tilde{S}_n$ can be described as

$$\tilde{S}_n = \max\left\{0, \tilde{S}_{n-1} + g(x_n)\right\}$$
$$S_0 = s \qquad (3-35)$$

while the extended stopping time is given by:

$$N^* = \inf\left\{n : \tilde{S}_n \geq a\right\}$$

This representation is equivalent to the original Wald test stopping rule:

$$N = \inf\left\{n : S_n < 0 \text{ or } S_n > a\right\}$$

which is repeated from the initial score S_0 each time $S_n < 0$ (zero being the renewal boundary, hence, the name: repeated one-sided Wald Test), and so on, until such time that a Wald test exceeds the threshold a.

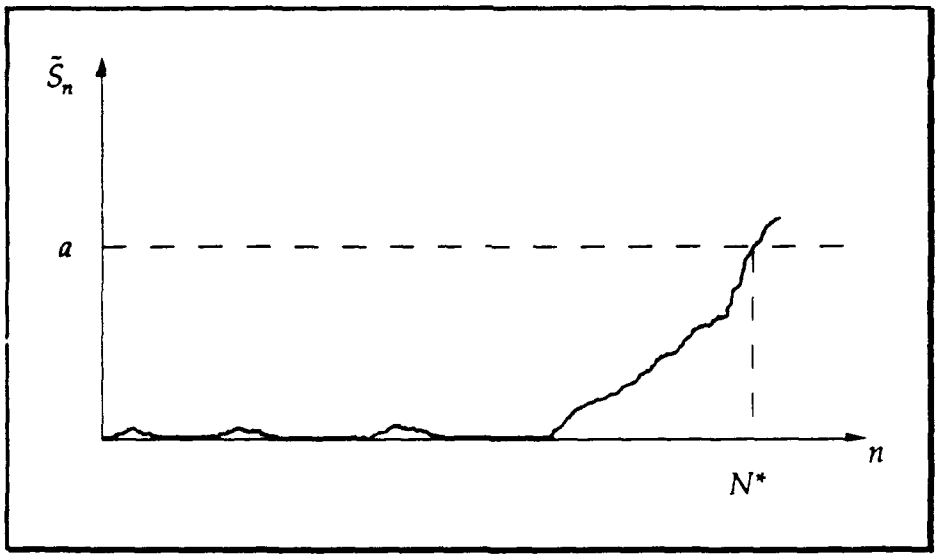


Figure 2.5. Recursive Implementation of Page's Test

The *second* interpretation of the recursive algorithm is related to the connection of the one-sided **first passage problem** with a single server queue. It will be shown that a queueing process w_n can be described in terms

of a random walk S_n. This fact forms the basis of the asymptotic distribution analysis of w_n as $n \to \infty$.

Assume the customers arrive at single server at time $a_1 < a_1 + a_2 < a_1 + a_2 + a_3$. These arrival times are the **arrival epochs** which form a **renewal process**. Assume, $a_1, a_2, \ldots$ are i.i.d. and let b_n ($n = 1, 2, \ldots$) denote the service time and w_n the waiting time of the n^{th} customer. Suppose that the $(n\text{-}1)^{th}$ customer arrives at epoch t. His service time starts at epoch $t + w_{n-1}$ and terminates at $t + w_{n-1} + b_{n-1}$ (See Figure 2.6). The next customer arrives at time $t + a_n$. He finds the server free if $w_{n-1} + b_{n-1} < a_n$, but has a waiting time (server busy) $w_n = w_{n-1} + b_{n-1} - a_n$ if this quantity is greater or equal to 0. Denote the queueing process by $x_n = b_{n-1} - a_n$. In short:

$$w_n = \begin{cases} w_{n-1} + x_n & \text{if } w_{n-1} + x_n \geq 0 \\ 0 & \text{if } w_{n-1} + x_n < 0 \end{cases}$$

$$\text{or:} \quad w_n = \max\{0, w_{n-1} + x_n\}$$

$$w_0 = 0$$

This result shows that if the service times $b_1, b_2, \ldots$ are i.i.d., then the x_n's are also i.i.d., hence the process w_n is a **random walk** which resets to 0 whenever it enters $(-\infty, 0)$. In order to describe the random walk w_n in terms of the random walk generated by the random variables x_n, define:

$$S_n = x_1 + x_2 + \ldots + x_n$$

and adhere to the notation for **ladder variables**. Define ν as the subscript for which $S_1 \geq 0$, $S_2 \geq 0$, ..., $S_{\nu-1} \geq 0$, but $S_\nu < 0$. By definition, ν is the *first*

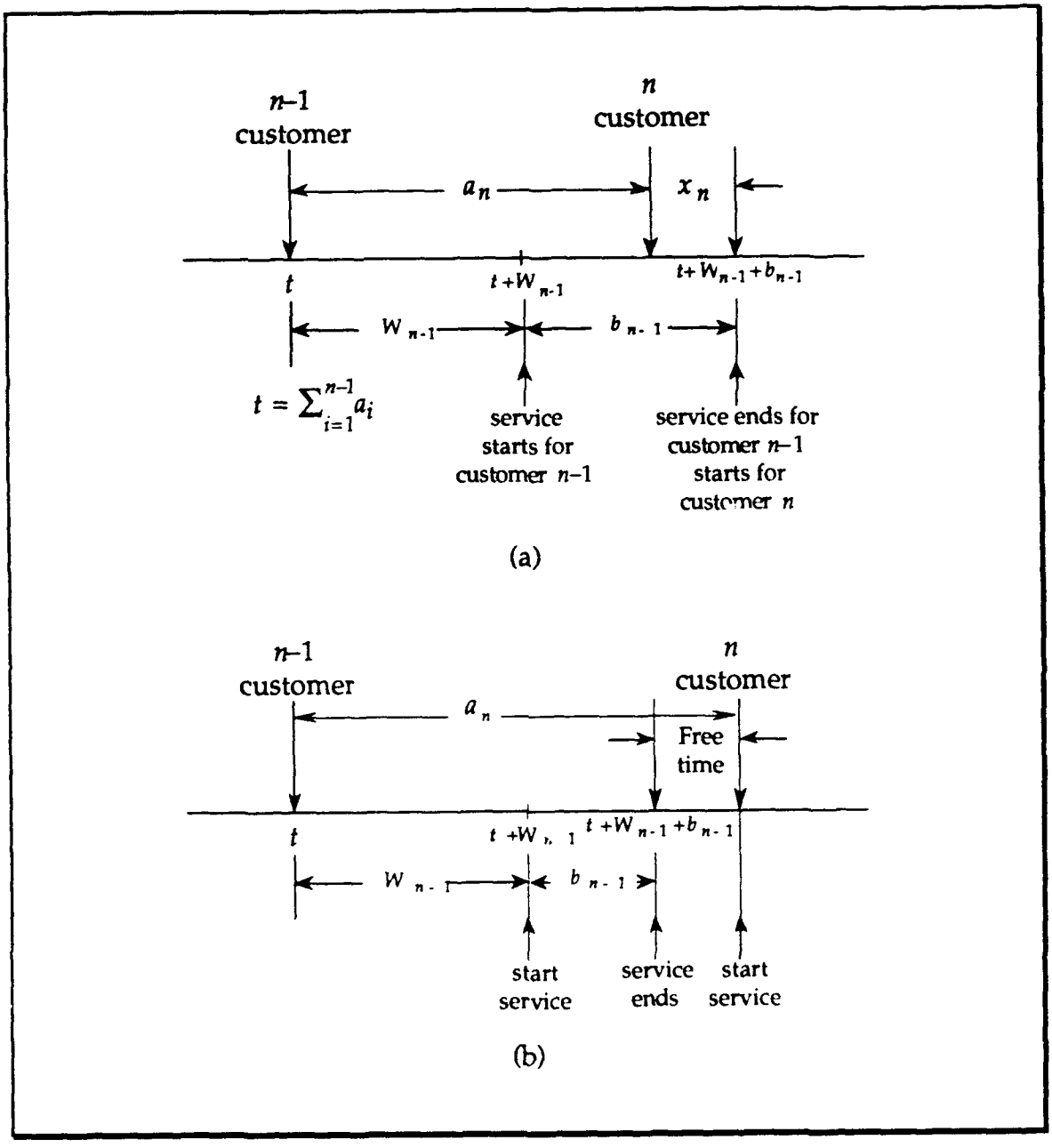


Figure 2.6. Two Situations of Server
(a) server busy
(b) server free

descending ladder epoch denoted by $\mathcal{F}_1^-$. Up to this epoch, the customers 1, 2, ..., ν–1 had positive waiting times $w_1 = s_1, w_2 = s_2, \ldots, w_{\nu-1} = s_{\nu-1}$. The ν customer is the first one to find the server free. The first conclusion is (Feller, 1971): *The descending ladder epochs correspond to the customers who find the server free, (i.e. $w_k = 0$) and constitute a renewal process with recurrence times distributed as $\mathcal{F}_1^-$.* (Since the continuation of the random walk beyond epoch $\mathcal{F}_1^-$ is a probabilistic replica of the entire random walk).

Suppose now that customer ν–1 arrived at epoch τ. His waiting time was $w_{\nu-1} = S_{\nu-1}$, the epoch of his departure is $\tau + w_{\nu-1} + b_{\nu-1}$ (see Figure 2.6). The customer ν arrived at epoch $\tau + a_\nu$, when the server is free. Thus, the time for which the server is free is given by

$$\text{free time} = \tau + a_\nu - (\tau + w_{\nu-1} + b_{\nu-1}) = a_\nu - w_{\nu-1} - b_{\nu-1}$$

$$= -x_\nu - w_{\nu-1} = -x_\nu - S_{\nu-1} = -S_\nu.$$

But by definition S_ν is the *first descending ladder height* $\mathcal{H}_1^-$. Thus, as the second conclusion we have: *The duration of free periods are i.i.d. random variables which constitute a renewal process with recurrence time distributed as $-\mathcal{H}_1^-$.*

To summarize, customer number k which arrives at epoch $\mathcal{F}_1^- + \ldots + \mathcal{F}_k^-$, is the customer that finds the server free. At the epoch of his arrival the server has been free for $-\mathcal{H}_k^-$ time units, at the same time the k^{th} descending ladder height is given by $S_k = -\sum_{i=1}^{k} \mathcal{H}_i^-$.

The remarkable statistical property of the random walk as of containing two imbedded renewal process: the ladder epochs and ladder heights, and the fact that the random walk is a probabilistic replica of the

entire random walk after the first ladder epoch (and each other ladder epoch) enables important results to be found about the distribution of the ladder variables in terms of the first ladder variables. It is easy to follow the next analysis of Page's cumsum tests (2-29) and (2-35) in terms of ladder variables.

c. Page Procedure Revisited

The first Page's version presented by (2-28) and (2-29), implies that the time k for which $\min_{0 \le k \le n} S_k$ gets its minima, is a *descending* ladder epoch. Hence, at that time k the test is renewed. The descending ladder epochs indicate the time where the change is more likely to happen. A change is declared when the test is terminated, i.e., the "distance" from the last descending epoch is at least a. For the repeated one-sided Wald's SPRT version (2-35), the descending ladder epochs are defined at the times where the random walk hits the lower boundary 0 (see Figure 2.7). At that ladder epochs the test is renewed. Once again, the test measures the "distance" between the current value of the random walk from the last ladder epoch. This distance is equivalent to the "statistical distance" between P_0 and P_1 as defined by (2-26) or the disorder distance. Notice that this analysis was done for detecting upward changes. Similarly, for detecting downward changes we will use *ascending* ladder epochs and the test terminates when the test reaches a "distance" a below the last ascending ladder variable.

Three important observations can be made: The *first*, as pointed out before in the analysis of Page's test in a composite hypothesis testing, is that it is worthwhile to bias the detector if it is known that before the disorder occurs, the test will have zero mean. It can be shown from Figure 2.8 that a random walk with negative drift will improve the chance of rapid disorder

detection under the restriction of low false alarm rate, since the number of ladder epochs will be larger, resulting in more renewals, thus having the effect of "forgetting" the irrelevant past observations. This result is supported analytically in the sequel when it is shown that the expected delay for detection is reduced by biasing the test.

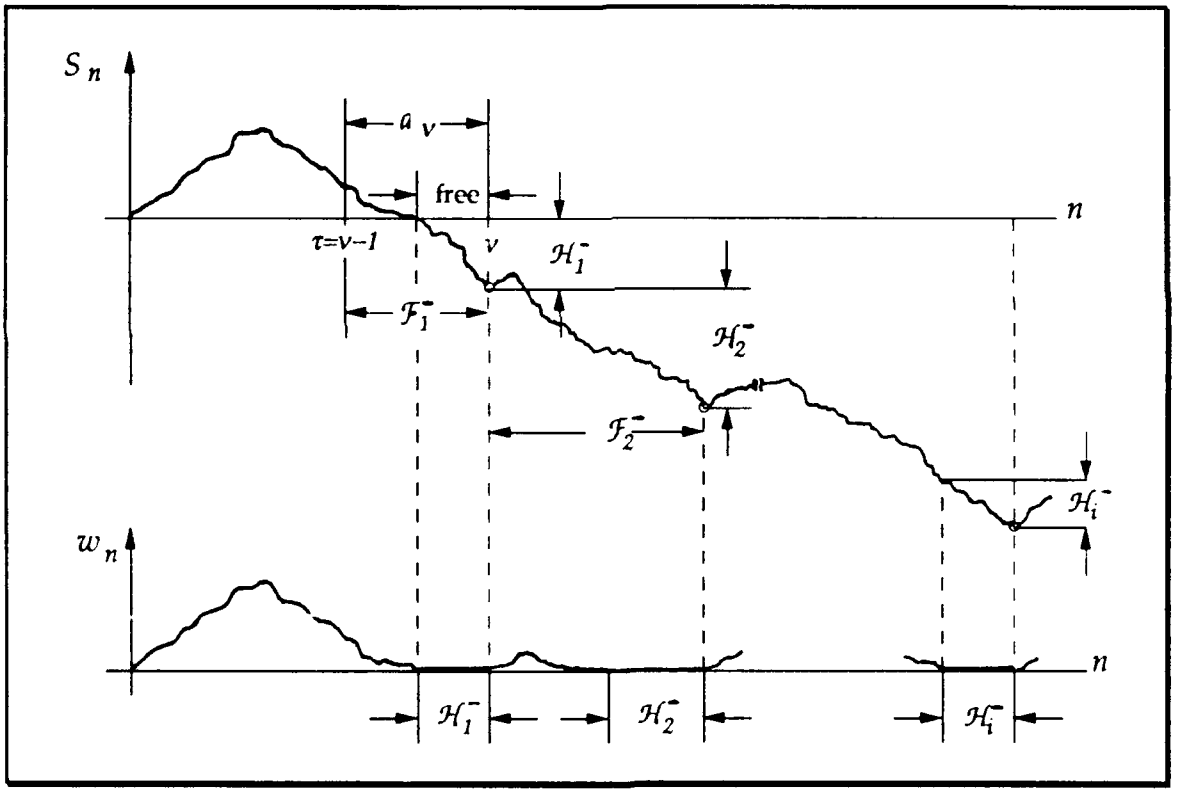


Figure 2.7. Random Walks w_n and S_n Containing two Renewal Processes: $\mathcal{F}_i^-$, $\mathcal{H}_i^-$

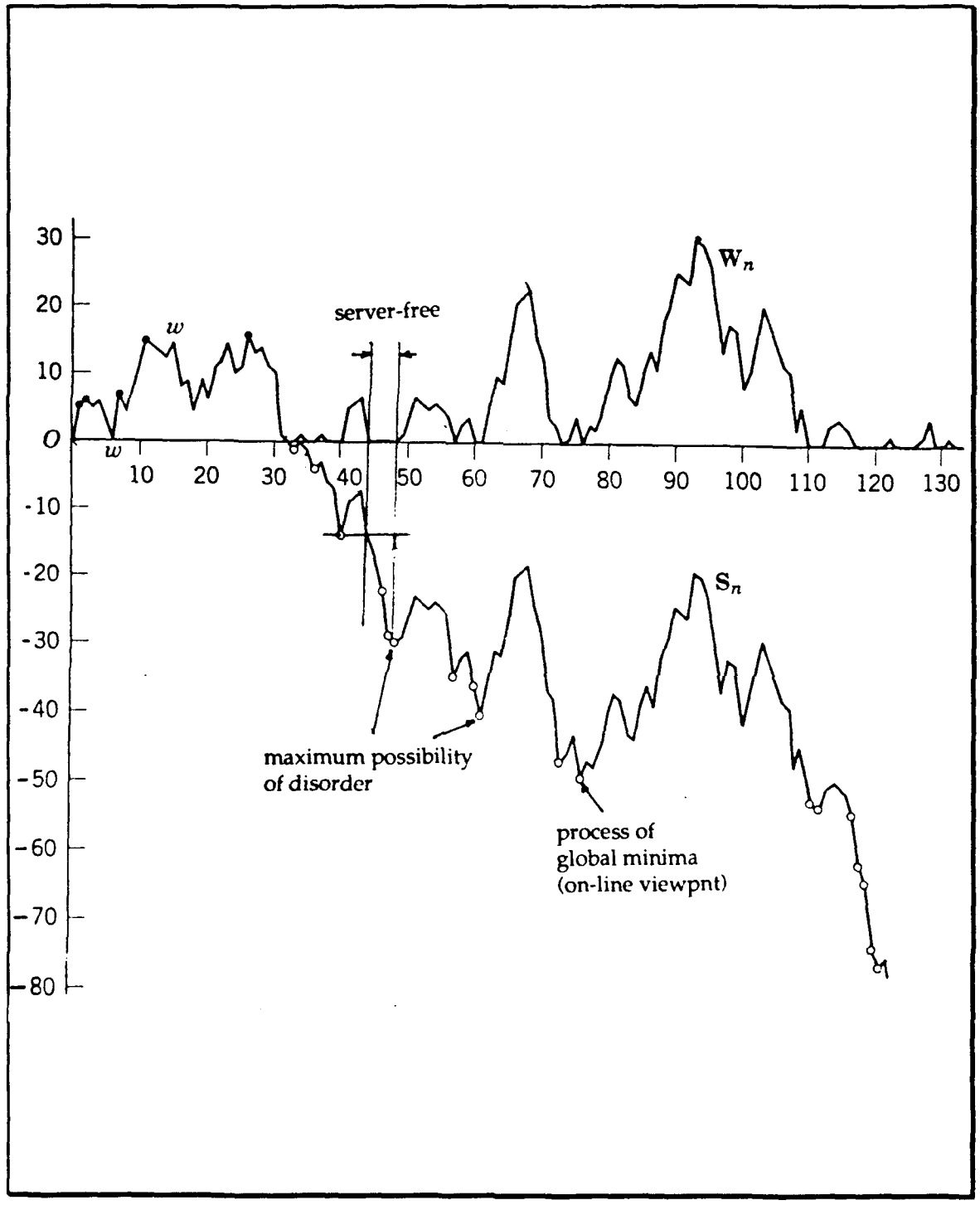


Figure 2.8. The Relationship between the Recursive Implementation W_n and the Random Walk Process S_n (From Feller, 1971)

The *second* observation is related to the first by realizing that each time the test is renewed (having the effect of resetting the test, hence "forgetting" the past), the test or detector behaves like an **adaptive detector,** since when reaching the descending ladder epoch, the past noisy observations containing no data about a possible change can be ignored. The fact that at each ladder point the likelihood of a change is the greatest implies that since the disorder is a local phenomena, the detection will occur if the signal-to-noise ratio and/or disorder duration is large enough, resulting in a threshold passage. Thus, this adaptive detector acts like a low-pass filter which filters the incoming signal except the changes.

The *third* observation leads to the analytical equivalency between the Page tests (2-27,2-29) and (2-35) and is found in Siegmund (Siegmund, 1985).

Let $S_n = x_1 + \ldots + x_n$. By backward recursion,

$$
\begin{aligned}
w_n &= \max(0,\ w_{n-1} + x_n) = \max\left(0,\ (w_{n-2} + x_{n-1})^+ + x_n\right) \\
&= \max(0,\ w_{n-2} + x_{n-1} + x_n,\ x_n) \\
&= \max\left(0,\ (w_{n-3} + x_{n-2})^+ + x_{n-1} + x_n,\ x_n\right) \qquad (2-36) \\
&= \max(0,\ w_{n-3} + x_{n-2} + x_{n-1} + x_n,\ x_{n-1} + x_n,\ x_n) \\
&= \ldots \max(0,\ S_n,\ S_n - S_1,\ S_n - S_2, \ldots,\ S_n - S_{n-1}) \\
&= \max_{0 \le k \le n} (S_n - S_k) = S_n - \min_{0 \le k \le n} S_k
\end{aligned}
$$

which shows the equivalent interpretation of the two procedures. Namely, the queueing variable w_n measures the departure of the process S_n from its last maxima.

This shows also that the distribution of w_n and its asymptotic behavior can be studied.

Theorem (Feller, 1971): The distribution of the queueing variable w_n is identical to the distribution of the random variable M_n, where:

$$M_n = \max_{0 \le k \le n} [0, S_1, \ldots, S_n]$$

where:
$$S_n = \sum_{i=1}^{n} x_i. \qquad (2-37)$$

□

Hence,

$$\Pr\{w_n \ge a\} = \Pr\{N^*(a) \le n\}$$

where
$$N^*(a) = \inf\{n: \ S_n \ge a\} \qquad (2-38)$$

and
$$\lim_{a \to \infty} \Pr\{w_n \ge a\} = \Pr\{N^*(a) < \infty\}$$

This relation is used in the sequel to derive an expression for the probability of the ARL function of the test (see 2-52).

The Wiener-Hopf integral equation (Feller, 1971) can be used to find an explicit solution to the probability distribution $m(x) = \Pr\{M_n \le x\} = \Pr\{w_n \le x\}$.

d. The Page Test as a Sequential Maximum Likelihood Detector

Let the problem be specified as in (2-26). When Page's test is implemented with the log-likelihood ratio (repeated SPRT Test), it is equivalent to a sequential implementation of the maximum likelihood detector. The log-likelihood function $l(x_1, \ldots, x_n)$ is given by:

$$l(x_1,\ldots,x_n) = \sum_{i=1}^{\nu-1}\log\left[P_0(x_i)\right] + \sum_{i=\nu}^{n}\log\left[P_1(x_i)\right]$$
$$= \sum_{i=1}^{n}\log\left[\frac{P_1(x_i)}{P_0(x_i)}\right] + \sum_{i=1}^{n}\log\left[P_0(x_i)\right]$$

Notice that the last term does not depend on the disorder time ν and can be neglected.

Define: $$S_\nu^n = \sum_{i=\nu}^{n}\log\frac{P_1(x_i)}{P_0(x_i)}$$

and replacing the unknown jump time ν by its maximum likelihood estimate under H_ν, we get the following change detector:

$$g_n(\hat{\nu}) = \max_{\nu} S_\nu^n \quad S_1^n - \min_{1\le k\le n} S_1^k$$
$$= S_n - \min_{1\le k\le n} S_k \qquad (2-39)$$

resulting in the same detection test as in (2-28,2-36).

2. Optimal Properties of the Page Test

In this section we will review two important results due to Lorden (Lorden, 1971). The first result (the following theorem) will enable the use of Wald approximations (2-16,2-17) to find an efficient way of calculating the performance measure for Page's test. The second result establishes the asymptotic optimality of Page's test in the sense of Lorden's criterion.

a. Bounds on the Performance of Quickest Detection for Repeated One-sided Tests

Let N be the stopping variable of a one-sided Wald test:

$$N = \inf\{n:\ S_n > a\}$$

for some statistics $\{S_n\}$ defined as functions of the i.i.d. observations $x_1, x_2, \cdots$. Let N_k be the stopping variable of the same test applied to $x_k, x_{k+1}, \ldots$, for $k = 1, 2, \ldots$, and define

$$N^* = \min_{k \geq 1}\{N_k + k - 1\}$$

N^* is the extended stopping variable of the one-sided test which stops when one of the sequence of tests $\{N_k\}$ applied to $x_k, x_{k+1}, \ldots$, stops the first time.

Theorem (Lorden, 1971): Let N be a one-sided stopping variable with respect to $x_1, x_2, \ldots$, such that

$$\Pr\{N < \infty | H_0\} \leq \alpha.$$

Let N_k denote the one-sided stopping variable obtained by applying N to $x_k, x_{k+1}, \ldots$, and define

$$N^* = \min_{k \geq 1}\{N_k + k - 1\}.$$

Then,

$$E_0\{N^*\} \geq \frac{1}{\alpha} = \gamma \qquad (2-40)$$

and for *any* alternative distribution P_1,

$$E_1\{N^*\} \leq E_1\{N\}$$ □

Notice that the one-sided Wald test is applied to $x_k, x_{k+1}, \ldots$, stopping the first time one of these tests stops. This result can now be viewed by using the renewal argument: Each time the test statistic $\tilde{S}_k$ (2-35) falls below zero, $\tilde{S}_k$ is

reset to zero indicating that a new test is starting from $k+1$ and so on until the first test reaches the stopping boundary.

This theorem establishes the optimality of Page's test (N^*) versus unrepeated one-sided tests (N). This result will be used in the next section to derive a performance measure for Page's test.

b. Asymptotic Optimality of Page's Test

Recall Lorden's criterion definition for the performance of cumsum procedures. The stopping rules N must satisfy

$$E\{N \mid \nu = \infty\} = E_0\{N\} \geq \gamma.$$

The quickness for which the stopping rule detects a true change in distribution is evaluated by $E_1\{N\}$ given by (2-27a).

The problem of minimizing $\overline{E}_1\{N\}$ subject to the constraint $E_0\{N\} \geq \gamma$ becomes more interesting if we replace the distribution P_1 by the Darmois-Koopman family of distributions $\{P_\theta, \theta \in \Theta\}$ with θ unknown, and try to achieve small $\overline{E}_\theta\{N\}$ (defined as $\overline{E}_1\{N\}$) for each θ subject to $E_0\{N\} \geq \gamma$. To handle composite $\{P_\theta\}$, one-sided sequential tests of P_0 vs. $\{P_\theta\}$ are applied to $x_k, x_{k+1}, \ldots$, for $k = 1, 2, \ldots$, stopping the first time one of these tests stops.

Lorden showed that we can simultaneously minimize $\overline{E}_\theta\{N\}$ for each θ asymptotically as $\gamma \rightarrow \infty$ for a wide class of tests. Lorden's main result was that Page's test (N^*) implemented with the log-likelihood function and a zero score with a stopping boundary γ belongs to this class. The following result will show that Page test achieves the lowest possible $\overline{E}_1\{N^*\}$, resulting as an optimal test both when P_1 is known, and when P_1 is unknown (composite testing case).

Let N and N^* be defined as in the last section. If N is the stopping variable of a one-sided SPRT of P_0 vs. P_1 with likelihood-ratio boundary $1/\alpha$, then by using Wald's approximation (2-7) we have that

$$E_1\{N\} \sim |\log \alpha| / I(\theta_1, \theta_0) \text{ as } \alpha \to 0$$

where $I(\theta_1, \theta_0)$ is the information number as defined by (2-23). Applying the last theorem, we obtain that N^* (Page's procedure) satisfying $E_0\{N^*\} \geq \alpha^{-1}$ and $E_1\{N\}$ is asymptotically *at most* $|\log \alpha| / I(\theta_1, \theta_0)$ as $\alpha \to 0$, and this is asymptotically the best we can do. In other words:

$$\inf \bar{E}_1\{N\} \geq \bar{E}_1\{N^*\} \sim \frac{\log \gamma}{I(\theta_1, \theta_0)} \qquad \text{as } \gamma = \alpha^{-1} \to \infty. \qquad (2-41)$$

Lorden also showed that we can simultaneously minimize $\bar{E}_\theta\{N\}$ for each θ asymptotically as $\gamma \to \infty$ for a wide class of tests:

$$\bar{E}_\theta\{N^*\} \sim \frac{\log \gamma}{I(\theta, \theta_0)} \qquad \text{as } \gamma \to \infty.$$

Moustakides (Moustakides, 1986), extended these results to the non-asymptotic case where γ is finite.

F. PERFORMANCE ANALYSIS OF THE PAGE TEST

In 1954, Page (Page, 1954) introduced a control chart procedure based on the repeated one-sided Wald-SPRT test with boundaries $(0,a)$, zero being the renewal boundary and a the stopping boundary.

Let the problem formulation be according to (2-26), that is,

$$N^* = \inf\{n: \; \tilde{S}_n > a\}$$

and let $L(s,\theta)$ be the ARL of this test with initial score $\tilde{S}_0 = s$ when $\{x_i\}$ are i.i.d. $F(x \mid \theta)$ distributed, i.e.,

$$L(s,\theta) = E\left\{ N^* \middle| \tilde{S}_0 = s,\theta \right\}.$$

Consider now Wald's test

$$N = \inf\{n:\ S_n < 0 \ \text{ or } \ S_n > a\}.$$

Similarly, let $L_w(s,\theta)$ denote the ARL of Wald's test,

$$L_w(s,\theta) = E\{N \mid S_0 = s,\theta\}$$

and let $Q_w(s,\theta)$ be the Operating Characteristic (OC) of the same Wald test, that is

$$Q_w(s,\theta) = P(S_n < 0 \mid S_0 = s,\theta).$$

Then, the ARL of Page test $L(s,\theta)$ is given (Page, 1954)

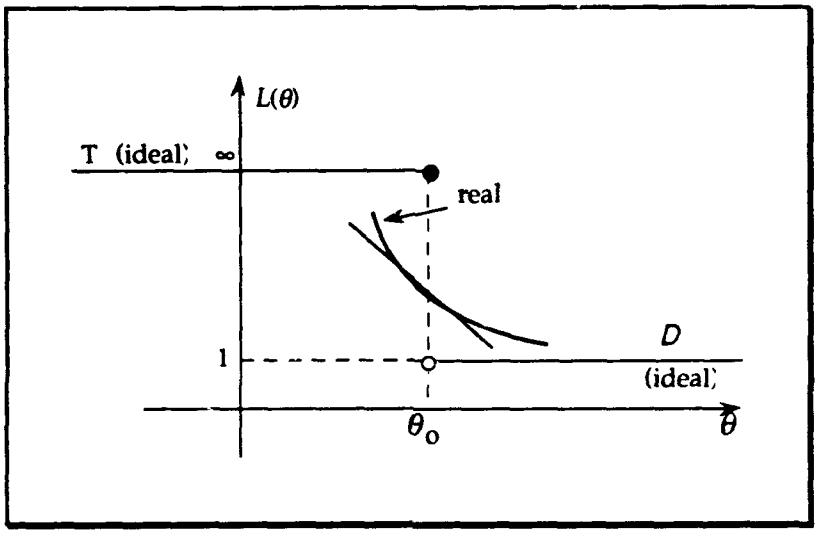
$$L(s,\theta) = \frac{Q_w(s,\theta)}{1-Q_w(0,\theta)} L_w(0,\theta) + L_w(s,\theta). \qquad (2-42)$$

Lorden (Lorden, 1971) and Benveniste (Benveniste, Ed., 1986) showed that the least favorable delay for detection, D, as defined by Lorden for Page's test occurs when the test statistic is zero when the change or the disorder occurs, i.e., $S_{v-1} = 0$, since the test statistic has the longest path to travel towards the stopping boundary. Thus, the ARL function with initial score $S_0 = 0$, determines *both* the false alarm rate T and the delay for detection D as given by

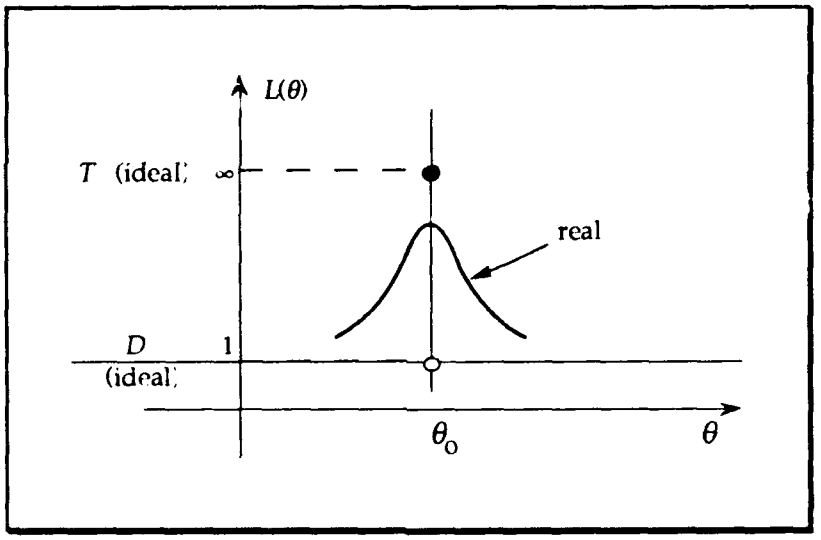
$$T = L(0,\theta_0) = E_0(N^*) = \frac{L_w(0,\theta_0)}{1-Q_w(0,\theta_0)} \qquad (2-43)$$

$$D = L(0,\theta_1) = E_1(N^*) = \frac{L_w(0,\theta_1)}{1-Q_w(0,\theta_1)}. \qquad (2-44)$$

Possible ideal and real ARL functions are presented in Figure 2.9.



(a) Real and Ideal ARL Functions for Testing $\theta \leq \theta_0$ against $\theta > \theta_0$



(b) Real and Ideal ARL Functions for Testing $\theta = \theta_0$ against $\theta \neq \theta_0$

Figure 2.9. Possible Ideal and Real ARL Functions

Remark: The **local properties** of cumsum tests can be measured in terms of the derivative of the function $L(\theta)$ at θ_0, since the local properties of a test measures the test's performance as $\theta_1 \rightarrow \theta_0$ (when the statistical "distance"

between the two hypotheses tends to zero), meaning weak signals or a low signal to noise ratio case.

It was shown (Nikiforov, 1986) that convenient measures can be defined by using the derivative ξ of the ARL function for determining a local approximation (see Figure 2.8), where ξ is defined as

$$\xi = \frac{\partial L(\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} \qquad \text{if } \xi \neq 0$$

and if $\xi = 0$, the local approximation is given by (2 – 45)

$$\frac{\partial^2 L(\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_0} \qquad \text{if } \xi = 0.$$

This forms the basis for what is called in the sequel the **local approach** for cumsum tests, resulting in **local sequential tests**.

1. The Lorden Approximations

As shown in the last section and given by (2-40), Lorden established bounds on the delay for detection and the false alarm rate for Page's cumsum test in terms of the Wald sequential test.

Lorden's theorem (2-40) can be applied to Page's test with nonlinearity $g(x)$ and a zero score to obtain a new bound. Using Wald's lower bound (2-16) as derived for his one-sided sequential test we obtain

$$\alpha \leq \exp\{-h(\theta_0)\}$$

Thus, from (2-40), the mean time between false alarm can be lower bounded by

$$T = E_0\{N^*\} \geq \exp\{h(\theta_0)\cdot a\}, \qquad (2\text{–}46)$$

where $h(\theta_0)$ is the unique non-zero root of the moment generating function (2-11). Using (2-17) and applying Lorden's theorem, the delay for detection can be upper bounded by:

$$D = E_1[N^*] \leq \frac{a+\gamma(\theta_1)}{E\{g(x) \mid \theta_1\}}. \tag{2-47}$$

Equations (2-46) and (2-47) are known as **Lorden's bounds**.

Remark: Notice that the mean time between false alarms (2-46) is an exponential function of the stopping bound a.

2. Wald Approximations

Similar results can be obtained by using the approach proposed by Nikiforov (Nikiforov, 1986). Recall the approximation (2-14) and (2-15) obtained for the OC and ARL functions for the two-sided Wald sequential test. These approximations can be used with the modification $b \uparrow 0$ (zero renewal boundary for Page's test). Once again, a lower bound for T and an upper bound for D will be derived.

a. Lower Bound for T when $E\{g(x) \mid \theta_0\} < 0$:

Using Page's result (2-43) and the bound (2-15) yields:

$$T = L(0,\theta_0) = \lim_{b \uparrow 0} \frac{L_w(0,\theta_0)}{1-Q(0,\theta_0)}$$

$$\geq \lim_{b \uparrow 0} \frac{1}{E\{g(x)|\theta_0\}} \left[a + \gamma(\theta_0) + b \cdot \frac{Q(\theta_0)}{1-Q(\theta_0)} \right]$$

Notice that the right hand side is a decreasing function of $Q(\theta_0)$ (since $d/dQ(Q/(1-Q)) > 0$ and $b < 0$). Hence, using the upper bound for Q obtained for $h(\theta_0) > 0$ given by (2-14), we obtain:

$$T \geq \lim_{b \uparrow 0} \frac{1}{E\{g(x)|\theta_0\}} \left[a + \gamma(\theta_0) + b \cdot \frac{\delta(\theta_0)\exp\{h(\theta_0)a\} - 1}{1 - \exp\{h(\theta_0)b\}} \right]$$

The last term is a function of b, but both the numerator and denominator approaches zero when $b \uparrow 0$. Using L'Hopital's rule and using the fact that $\delta(\theta_0) \geq 1$ we obtain (Broder, 1990):

$$T \geq \frac{1}{E\{g(x)|\theta_0\}} \left[a + \gamma(\theta_0) + \frac{1 - \exp\{h(\theta_0)a\}}{h(\theta_0)} \right] \qquad (2-48)$$

where $h(\theta_0)$ and $\gamma(\theta_0)$ were defined in (2-11) and (2-15) respectively.

b. Lower Bound for T when $E\{g(x) \mid \theta_0\} = 0$

In a similar way we obtain (Broder, 1990):

$$\begin{aligned} T &= \lim_{b \uparrow 0} \frac{L_w(0, \theta_0)}{1 - Q(0, \theta_0)} \\ &\geq \lim_{b \uparrow 0} \frac{1}{1 - Q(\theta_0)} \frac{a^2[1 - Q(\theta_0)] + b^2 Q(\theta_0)}{E\{g^2(x)|\theta_0\}} \\ &= \frac{a^2}{E\{g^2(x)|\theta_0\}}. \end{aligned} \qquad (2-49)$$

Remark: By (2-49) the mean time between false alarms is a quadratic function of the stopping bound a. Recall that by (2-46) it was shown that when $E\{g(x) \mid \theta_0\} < 0$, the mean time between false alarms is an exponential function of the stopping bound. Hence, once again it is demonstrated that it is worthwhile to bias the test to have a negative drift before the change,

resulting T as an exponential function of the stopping bound instead of a quadratic one (Broder, 1990).

c. *Upper Bound for D*

Consider now the delay for detection. Since that for detecting upward change, after the disorder $E\{g(x) \mid \theta_1\} > 0$ results in $h(\theta_1) < 0$. Using (2-44), (2-14) and (2-15) in the same manner as before we obtain:

$$\begin{aligned} D &= L(0,\theta_1) \\ &= \frac{L_w(0,\theta_1)}{1-Q(0,\theta_1)} \\ &\le \lim_{b\uparrow 0} \frac{1}{1-Q(\theta_1)} \frac{[a+\gamma(\theta_1)][1-Q(\theta_1)]+bQ(\theta_1)}{E\{g(x)|\theta_1\}} \end{aligned}$$

Once again, since the right hand side is a decreasing function of $Q(\theta_1)$, and since $b < 0$ the inequality is preserved. Thus, as $Q(\theta_1)\downarrow 0$, we can replace $Q(\theta_1)$ with zero and obtain

$$D \le \frac{a+\gamma(\theta_1)}{E\{g(x) \mid \theta_1\}} \tag{2-50}$$

which is consistent with Lorden's bound (2-47).

3. Asymptotic Performance and Measures

For all disorder detection schemes the pair (T,D) determines the detector performance, just as P_D versus P_{FA} (P_D being the probability of detection and P_{FA} is the probability of false alarm) determines the performance of a classical detector.

As shown, the ARL function determines uniquely the pair (T,D). Thus, it is in our interest to examine its asymptotic performance. Hinkley (Hinkley, 1972) used an asymptotic performance measure for the nonlinearity $g(x)$ while using P_D and P_{FA} as performance criteria. This measure was derived while calculating the efficiency of the cumsum procedure. The proposed measure was

$$\log E\{\exp\{-h(\theta_0)\cdot g(x)\} \mid \theta_1\}.$$

Recently (Broder, 1990), another performance measure was proposed resulting in an alternative technique which allows recursive computation of the ARL, the stopping bound a increases, avoiding the complicated numerical integration needed to generate the performance curves (solution of Fredholm type integral equations).

a. Asymptotic Approximation of the ARL Function

Central limit theorem for renewal processes. (Ross, 1989): For large t, $N(t)$ being a renewal process is approximately normally distributed with mean $\frac{t}{\mu}$ and variance $t\sigma^2/\mu^3$, where μ ($\mu \neq 0$) and σ are respectively the mean and the variance of the interarrival distribution.

$$\lim_{t\to\infty} P\left\{\frac{N(t)-t/\mu}{\sqrt{t\sigma^2/\mu^3}} < x\right\} = \Phi(x) \qquad (2-51)$$

where $\Phi(x)$ is the Gaussian cumulative distribution function:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx.$$

Khan (Khan, 1981) used this result to show the asymptotic normality (in the sense of $a \to \infty$) of the run length of Page's test with a stopping boundary a under $P(\theta_1)$:

$$\text{define:} \qquad \mu = E\{g(x)|\theta_1\} \quad \sigma^2 = \text{var}[g(x)|\theta_1]$$

$$\text{then:} \qquad \frac{L(\theta_1) - a/\mu}{\sqrt{a\sigma^2/\mu^3}} \xrightarrow{P} N(0,1)$$

where $N(0,1)$ is a Gaussian distribution with zero mean and a unit variance. Using this asymptotic distribution and the results derived in (2-38), a new approximation can be established for the asymptotic probability that the average delay is less than a given threshold:

$$\Pr\{L(\theta_1) < x\} \approx \Phi\left(\frac{x - a/\mu}{\sqrt{a\sigma^2/\mu^3}}\right). \qquad (2-52)$$

b. *Alternative Asymptotic Performance Evaluation*

An alternative way to evaluate the Page test under different nonlinearities for various noise distributions has been shown by Broder (Broder, 1990). Define an asymptotic performance measure

$$\eta \triangleq \lim_{\text{ARL}_0 \to \infty} \frac{\log \text{ARL}(\theta_0)}{\text{ARL}(\theta_1)}$$

$$= \lim_{T \to \infty} \frac{\log(T)}{D}$$

$$= \lim_{a \to \infty} \frac{\log(T)}{D} \qquad (2-53)$$

Notice that for η to reach a finite bound, both T and D approach infinity as $a \to \infty$, thus η reflects the asymptotic performance and is the reciprocal of the slope of the (D, log$T$) performance curve. This performance measure can be interpreted in two ways: *First*, as the ratio of the run-length for the two hypotheses. Hence, the larger η indicates better asymptotic performance; *second*, as an Asymptotic Relative Efficiency (ARE) between the two tests. Recall that for a fixed mean time between false alarms (large enough):

$$\mathrm{ARE}_{1,2} = \frac{\eta_1}{\eta_2} = \lim_{\substack{T_1 \to \infty \\ T_2 \to \infty}} \frac{\dfrac{\log(T_1)}{D_1}}{\dfrac{\log(T_2)}{D_2}} = \frac{D_2}{D_1} = \frac{L_2(\theta_1)}{L_1(\theta_1)} \qquad (2-54)$$

hence, resulting in the delay ratio of the two tests.

Using Lorden's approximations (2-46) and (2-47) for the Page test, and ignoring the "excess over the boundaries" we get:

$$\log T \geq h(\theta_0)a$$

$$D \leq \frac{a}{E\{g(x)|\theta_1\}}$$

hence

$$\eta \geq h(\theta_0) \cdot E\{g(x)|\theta_1\} = \underline{\eta}. \qquad (2-55)$$

This lower bound $\underline{\eta}$ can be defined as the asymptotic performance measure, thus, being a convenient way to "measure" Page's test using different nonlinearities $g(x)$ for various noise distributions.

Notice that the lower bound $\underline{\eta}$ can be used as an alternative way to estimate the expected delay given the desired large mean time between false alarms.

Property: If g is the log-likelihood ratio, $g(x) = \log[(f(x \mid \theta_1)/f(x \mid \theta_0)]$, then under any noise distribution the bound is tight, i.e., $\eta = \underline{\eta} = I(\theta_1,\theta_0)$ where $I(\theta_1,\theta_0)$ is the **Kullback-Liebler information number** (2-23).

Proof: Since

$$E\left\{\exp\left\{\log\frac{f(x|\theta_1)}{f(x|\theta_0)}\right\}|\theta_0\right\} = \int_{-\infty}^{\infty}\frac{f(x|\theta_1)}{f(x|\theta_0)}\cdot f(x|\theta_0)dx = 1$$

using the moment generating function identity (2-11) it becomes obvious that

$$h(\theta_0) \equiv 1.$$

Hence,
$$\eta = E\{g(x)|\theta_1\} = \underline{\eta} = I(\theta_1,\theta_0).$$

$$\lim_{T\to\infty}\frac{\log T}{D} \approx I(\theta_1,\theta_0)$$

□

Recall that by definition: $T = E_0\{N^*\} \geq \alpha^{-1}$

where α is the Probability of false alarm. Hence

$$\frac{\log \alpha^{-1}}{D} \approx I(\theta_1, \theta_0) \qquad \alpha \to 0$$

or
$$D \approx \frac{\log \alpha^{-1}}{I(\theta_1, \theta_0)} \qquad \alpha \to 0. \qquad (2\text{–}56)$$

Hence, (2-55) can be seen as a generalization of Lorden's result (2-41) for any nonlinearity function. The root $h(\theta_0)$ of the moment generating function identity "scales" (2-41), thus (2-55) establishes a general bound which can be

evaluated for any nonlinearity $g(x)$ under any noise environment. We will be interested in the cases where the strict equality exists ($\eta = \underline{\eta}$), which enables us to get a precise relationship between the delay and the false alarm rate. In the next chapter we will see some examples for which $\eta = \underline{\eta}$.

G. SUMMARY

In this chapter we show that detecting a disorder presented in the multiple hypothesis framework (1-2) can be done by using cumsum type procedures. One of these procedures, called the Page test, was presented and investigated in depth. Using renewal theory and ladder variables we present a new technique to observe the properties of Page's test. Three observations are shown: first, it is worthwhile to bias the test if it is known that before the disorder the mean of the statistic is zero; second, Page's test behaves like an adaptive detector in the sense that the ladder epochs form a local minima (or maxima) process in which the past observations which do not contribute relevant information about the change are forgotten. Finally, we showed the equivalent representation of Page's test in the off-line and on-line versions.

Page's test implemented with the log-likelihood nonlinearity is shown to be the MLE of the change time (within the multiple hypotheses testing (1-2) framework). Using Lorden's results, the asymptotic optimality of Page's test is obtained in the sense that Page's test implemented with the log-likelihood nonlinearity is the optimal stopping rule, that is, the average delay for detection subject to the false alarm rate which tends to zero is minimized. Thus, the log-likelihood nonlinearity is shown to be the optimal

nonlinearity, and therefore, Page's test, which is the MLE for this case, is shown to be the quickest detector for the disorder problem.

Finally, performance evaluation of Page's test was derived. The main results are what is called the Lorden approximations for the mean time between false alarms T (2-46) and delay D(2-47) and similarly, the Wald approximations for T (2-48) and D(2-50).

In addition, using Broder's results, the asymptotic performance measure is shown to be lower bounded. The problem of how informative the bound is for different nonlinearities will be analyzed in the next chapter. Here we show that for the optimal nonlinearity the log-likelihood, the bound is tight, i.e., the bound provides all the information needed to predict Page test performance. Finally, a new simple generalization of Lorden's result was shown for any nonlinearity function in any noise environment.

III. THE APPLICATION OF PAGE'S TEST TO PARAMETRIC AND NONPARAMETRIC CHANGE DETECTION

A. INTRODUCTION

In the last chapter it is shown that implementing the Page's test with the log-likelihood ratio nonlinearity results in the Maximum Likelihood Estimator (MLE) of the change time. Furthermore, it is the quickest detector of the disorder. The problem becomes much more complicated when the model parameters after the change are not known. In this case, the unknown random variables, ν the change time, and the model parameters θ, have to be estimated. Thus the detection problem can be presented in the estimation framework. *Joint* estimation of ν and θ is a very difficult task because the disorder occurs at an unknown time and the presence of several unknown parameters forces the use of suboptimal detectors. Hereby, we present some competitive ad-hoc methods used for detection and if possible also estimation of the change time and the model's parameters.

1. Likelihood Oriented Methods

In situations such as detection of an unknown change magnitude of Gaussian signals it is possible to perform the joint estimation of ν and the unknown parameters θ (Basseville, 1988). In such cases, the detection approach consists of replacing the unknown jump magnitude of the model parameter by its MLE. The *Generalized Likelihood Ratio* (GLR) test of the *joint* estimation becomes

$$\max_{1\le v\le n}\max_{\theta_1} S_r^n(v,\theta_1) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

where $S_r^n(v,\theta_1)$ is the log-likelihood cumsum statistic. This double maximization problem of estimating both the change time and the parameters is reduced to a single maximization of the cumulative sum since the Gaussian characteristic of the signal to be detected allows an explicit solution as a function of θ_1 for the likelihood ratio test. (Basseville and Benveniste, Eds., 1986, Chapter 1). Hence the change time estimate becomes

$$\hat{v}(r) = \arg\min_{v} S_r^n\left(v,\hat{\theta}_1\right).$$

This property is still valid in a more general situation when we consider the problem of detecting additive changes in linear models described in state-space representation and leads to an *efficient change detection algorithm with* reasonable computing cost. An earlier approach consists of monitoring the innovations of a Kalman filter, because of the linear property of the system and additive effect of the change on the system, it may be shown (Willsky and Jones, 1976) that the effect on the innovation is also additive. Moreover, the Gaussian characteristic of the state and observation noises which ensures for an explicit solution in θ_1 for a likelihood ratio test, is still valid in this situation. These points were explored by Willsky and Jones (1976) who derived *a recursive* algorithm for the GLR test computed for the innovation of a Kalman filter designed under the no change hypothesis. The distribution of these innovations with respect to its past values, thus, the cumulative sum to be computed in this case is in the form of

$$S_i^j\left(P_{\theta_0}, P_{\theta_1}\right) = \sum_{k=i}^{j} \log \frac{P_{\theta_1}\left(x_k | x_{k-1}, \ldots, x_0\right)}{P_{\theta_0}\left(x_k | x_{k-1}, \ldots, x_0\right)}$$

where P_{θ_1} reflects the change in a certain parameter (change in the mean, variance, etc.). The GLR test is then

$$\max_{1 \le r \le n} \max_{\theta_1} S_r^n\left(P_{\theta_0}, P_{\hat{\theta}_1}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda.$$

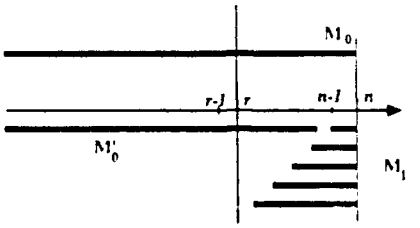
As mentioned above, the maximization over θ_1 is explicit because of the Gaussian assumptions of white noise and additive changes, hence the test for the change time is reduced to a single authorization even in this more general situation.

In the case of detecting changes in model eigenstructure such as changes in AR or ARMA models or equivalently in the state transition matrix of a state-space representation of a model, the problem of the joint estimation of the change time ν and the changing parameters is more complicated. At this point we need to distinguish between two types of situations: in the *first* case, if the signal or system is known to have the same behavior as an AR or ARMA process, then the model is descriptive enough for its parameters behavior to be detected (Basseville, 1988). The *second* case reflects a situation where the signal or system is not known and the main issue is to detect changes in the eigenstructure, then the AR or ARMA models are nothing but a tool for detection of such changes.

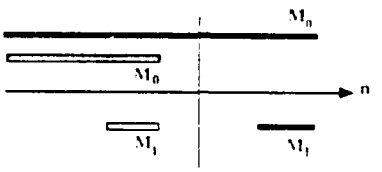
2. Simplification of GLR Tests (Two Models Approach)

In the case of segmentation of signals resulting from AR models such as speech segmentation (Andre-Obrecht, 1988) the detection of abrupt changes in the AR parameters is performed via the comparison between a long-term model M_0 identified in a growing window and a short-term model M_1 identified in a sliding window of fixed length. (See Figure 3.1) This method is shown to be a simplification of the GLR test since for implementing the GLR, the maximization over θ (the AR vector parameter) is no longer explicit because the change is not additive on the observation. Moreover in the case of ARMA models, the cumsum is no longer linear in the parameter, therefore the test becomes quite expensive since for *each* possible change time r we need to use the data $\{r, r+1, ..., n\}$ for identifying the AR model M_1 after the change and compute the log-likelihood ratio cumsum S_r^n, then maximize over r. In the case of AR models, this method is not only expensive but leads also to boundary problems (Deshayes and Picard, 1986).

The two model approach simplifies the GLR test by using a fixed length sliding window as opposed to varying length windows needed to implement the GLR test. Different statistical distance measures between the long-term and the short-level models were proposed by Appel and Brandt (1983) and Segen and Sanderson (1980), Basseville and Benveniste (1983, 1986), Ishii and Iwata (1979) and Andre-Obrecht (1988). Most of these measures are based upon innovation testing which in turn is based upon the conditional distribution of the observations.



a . Scheme for the GLR test.



b . Using two models, one "global" and one local.

Figure 3.1. Schemes for
a. GLR test
b. Two Model Method

3. The Statistical Local Approach

Another approach for overcoming the drawbacks of the GLR tests is known as the "**local approach**" and has been introduced in change detection problems by Nikiforov (1983, 1986) for on-line detection for AR models.

The original idea of Nikiforov consists of looking for *small* changes in AR or ARMA models and using the Taylor expansion of the log-likelihood function. His method results in a statistic function

$$g(x_n) = \frac{d}{d\theta} \log P_\theta(x_n | x_{n-1} \ldots) \Big|_{\theta = \theta_0}$$

In other words, instead of monitoring the observation process $\{x_n\}$ or the innovation process, the local approach monitors $g(x)$. The key result (Deshays and Picard, 1986) is that there exists a central limit theorem for $g(x)$, for which any change in θ is reflected in a change in the mean of $g(x)$ for which Page's test or the GLR tests can be used. Nikiforov derived two algorithms based upon cumsum tests for two different priors about the change directions. Different applications for these algorithms are described in Nikiforov (Nikiforov, 1986). Another use of these methods is in the area of recursive parameter identification. Benveniste (Benveniste, 1987) has shown that for any general recursive parameter identification algorithm

$$\theta_n = \theta_{n-1} + \gamma_n H_n(\theta_{n-1}, x_n),$$

where γ_n denotes the varying gain and H_n denotes the statistic, applying the local approach to the statistics $H_n(\theta_0, x_n)$ where θ_0 is a reference model, enables one to transform the problem of changes in the parameter vector θ into the problem of detecting a change in the mean value of an asymptotically Gaussian distributed process which is a cumsum of the function $H(\cdot)$.

Finally, Basseville (Basseville, 1987) and Benveniste (Benveniste, 1987) introduced another use of the local approach technique. In the case of detecting changes in the AR part of a multivariable ARMA process having unknown and time varying MA coefficients. Because the Fisher information matrix for an ARMA process is not block diagonal with respect to the AR and MA parameters (because of the coupling between the unknown monitored parameters and the unknown changing MA parameters), neither the likelihood function nor its Taylor's expansion (local approach) can be used. By using instrumental statistics on the observations (Benveniste and

Basseville and Moustakides, 1987), the changes in the AR portion are reflected in changes in the mean of the instrumental statistics. By looking for "small" changes in the AR coefficients, the local approach statistic, i.e., Taylor's expansion of the instrumental statistics results in an χ^2 test

$$U_n^T \Sigma_n^{-1} U_n \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

where U_n is the instrumental statistic vector (which is asymptotically Gaussian) and Σ_n is its covariance matrix.

B. ORGANIZATION OF THIS CHAPTER

In the introduction section, different competitive methods of Page's test were briefly described. Some of them enables one to detect (or estimate) the change time together with estimation of the changed parameters. Now, we will only be concerned with the quickest disorder (change) detection problem. In the case of implementing the log-likelihood nonlinearity, the Page test is the optimal (quickest detector) for the disorder problem but assumes that the observations are i.i.d. distributed with one distribution before the disorder and another distribution after the disorder. However, in the case that the i.i.d. assumption does not hold, other detection schemes "tuned" to the specific problem may perform better than the suboptimal Page test. Despite these concerns about Page's test, the test will be shown to detect the change instants occurring at random times very efficiently. This chapter focuses on general implementation of Page's test for both parametric and non-parametric detection, and evaluation of the test's performance for the implemented nonlinearities, by using the results in Chapter II.

Section C introduces the case of detecting jumps in the mean of Gaussian distributed observations. Both upward and downward directions are considered as well as the case when the change magnitude is unknown. Page's test implemented for this problem, (derived from the on-line point of view) and the GLR test (derived from the off-line point of view) are shown to be the same.

In Section D, performance evaluation for Page's test is evaluated for different nonlinearities in Gaussian and Gauss-Gauss mixture noise environments. In particular we are interested in the cases where the lower asymptotic performance bound $\underline{\eta}$ is tight (i.e., $\eta = \underline{\eta}$). In the **parametric framework**, the problem of detecting changes in the mean and variance of Gaussian observations is shown to result in $\eta = \underline{\eta}$ for which the performance measure is easily computed. As a second example we consider a suboptimal implementation of Page's test where the distribution after the disorder is not known and by the use of composite hypothesis technique, a new test is derived. This local optimum detector is based on Wolcin's test (Wolcin, 1983) and a modification (Broder, 1990), and is modified to detect energy changes occurring within frequency "windows." New performance results are obtained and shown to be consistent with the simulation results. Finally, in the **nonparametric framework**, the sign test is analyzed by using results from random walk theory, and shown to have the property $\eta = \underline{\eta}$.

Section E summarizes the main results of this chapter.

C. DETECTING JUMPS IN THE MEAN

We begin in this section with the simplest application of the Page test, namely the problem of a change in the mean of independent identically distributed Gaussian random variables. This problem is an important one since, as will be shown in the sequel, many complicated problems involving abrupt changes in the eigenstructure (parameter changes) can be converted to the problem of change in the mean. Two cases are considered: the first, when the means before and after the change are known, and secondly when the means and therefore the change magnitude is unknown.

1. Known Means before and after the Change

Let $\{e_n\}$ be a Gaussian white noise sequence with variance σ^2, and let $\{x_n\}$ be the observation sequence such that

$$x_n = \mu_n + e_n \qquad n = 1, 2, \ldots, N$$

where:

$$\mu_n = \begin{cases} \mu_0 & \text{if } n \leq \nu\text{-}1 \\ \mu_1 & \text{if } n \geq \nu. \end{cases}$$

Consider now the likelihood ratio test between the "no change" hypothesis:

$$H_0\colon\ \nu > N$$

versus the "change" hypothesis:

$$H_1\colon\ \nu \leq N.$$

Thus, the log-likelihood ratio between these two hypotheses has the following form (Basseville, 1988):

$$L(\nu)=\frac{\prod_{k=1}^{\nu-1}P_0(x_k)\cdot\prod_{k=\nu}^{N}P_1(x_k)}{\prod_{k=1}^{N}P_0(x_k)}=\prod_{k=\nu}^{N}\frac{P_1(x_k)}{P_0(x_k)}, \qquad (3-1)$$

therefore, its logarithm is given by:

$$\ln L(\nu)=\frac{\mu_1-\mu_0}{\sigma^2}\sum_{k=\nu}^{N}\left(x_k-\frac{\mu_0+\mu_1}{2}\right)=\frac{1}{\sigma^2}S_\nu^N(\mu_0,\Delta) \qquad (3-2)$$

where $\Delta=\mu_1-\mu_0$ is the magnitude of the jump, and

$$S_i^j(\mu,\Delta)=\Delta\cdot\sum_{k=i}^{j}\left(x_k-\mu-\frac{\Delta}{2}\right).$$

Replacing the unknown jump time ν by its maximum likelihood estimate under H_1 yields:

$$\hat{\nu}=\arg\min_{1\le\nu\le N}\left\{\prod_{k=0}^{\nu-1}P_0(x_k)\cdot\prod_{k=\nu}^{N}P_1(x_k)\right\}=\arg\min_{1\le\nu\le N}S_\nu^N(\mu_0,\Delta).$$

Thus, we get the following change detector:

$$g_N\triangleq L(\hat{\nu})=\max_{\nu}S_\nu^N(\mu_0,\Delta)\underset{H_0}{\overset{H_1}{\gtrless}}a \qquad (3-3)$$

where a is a threshold properly chosen as addressed in Section C.

This detector can be described also as follows: detection occurs the first time at which

$$g_N=S_1^N(\mu_0,\Delta)-\min_{1\le k\le N}S_1^k(\mu_0,\Delta)>a \qquad (3-4)$$

which is nothing but the Page-Hinkley stopping rule or cumsum algorithm, and may be computed in the following recursive manner:

$$g_n = \left(g_{n-1} + x_n - \mu_0 - \frac{\Delta}{2} \right)^+ . \qquad (3-5)$$

Thus, both **Page's stopping rule** (derived from the on-line viewpoint) and the **generalized likelihood ratio (GLR) test** (derived from the off-line viewpoint) are identical. The behavior of the Page-Hinkley stopping rule is depicted in Figure 3.2.

2. Unknown Magnitude of Change

In this case we may assume that μ_0 is known, but μ_1 is not. A minimum jump magnitude $\Delta_{\min}$ to be detected is fixed *a priori*. Two tests are running in parallel corresponding to two possible directions (increasing or decreasing mean).

For detecting a decrease in the mean we determine the stopping time N by observing when the maxima process drops down by a, the detection threshold (see Figure 3.2).

$$N = \inf\left\{ n: \quad \max_{1 \le k \le n} S_k - S_n \ge a \right\}$$

where

$$S_n = \sum_{k=1}^{n} \left(x_k - \mu_0 + \frac{\Delta_{\min}}{2} \right) \qquad (3-6)$$

$$S_0 = 0.$$

Similarly for detecting an increase in the mean we define

$$N = \inf\left\{ n: \quad \min_{1 \le k \le n} S_k - S_n \ge a \right\}$$

where

$$S_n = \sum_{k=1}^{n} \left(x_k - \mu_0 - \frac{\Delta_{\min}}{2} \right) \qquad (3-7)$$

$$S_0 = 0$$

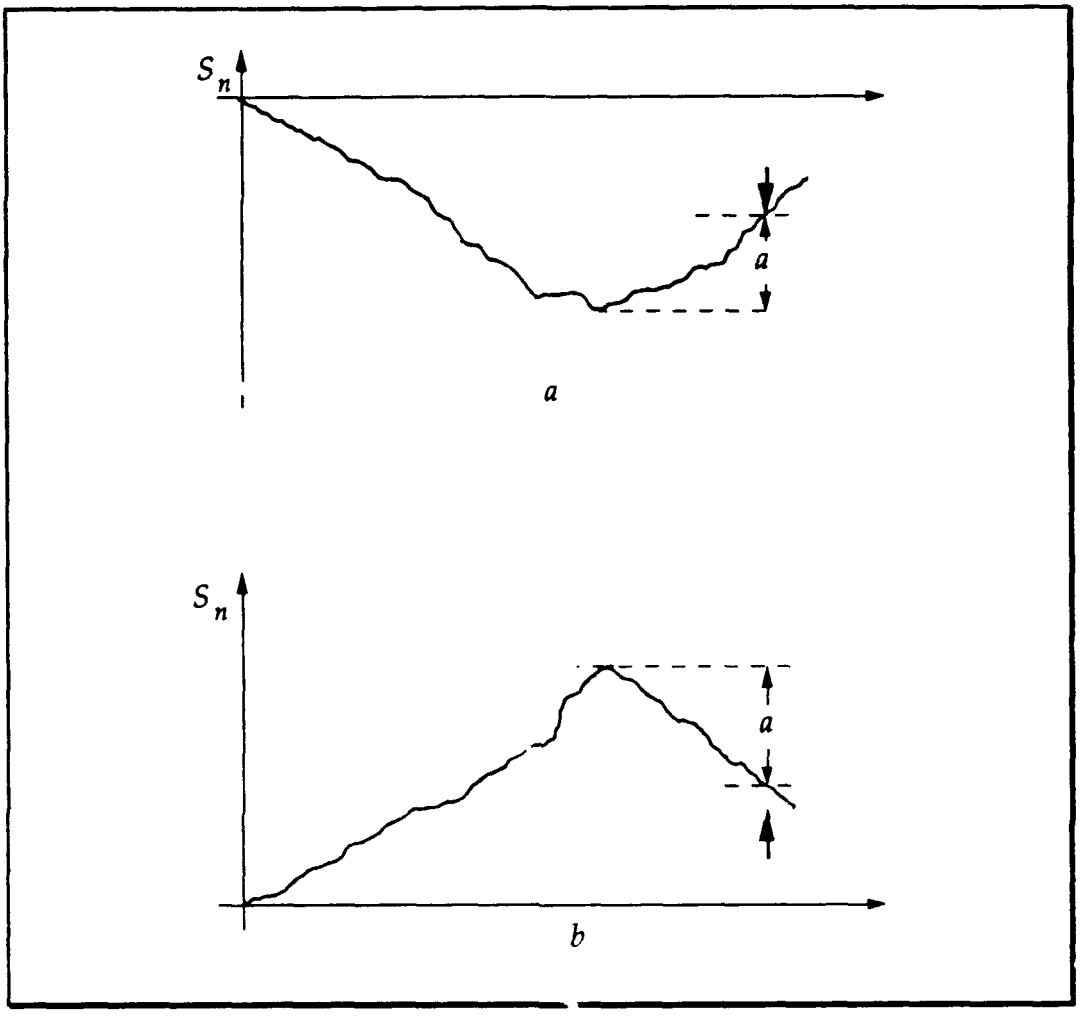


Figure 3.2. Page-Hinkley Stopping Rule as the Process of Global Minima (for a) and Global Maxima (for b)
a. Detecting Upward Change
b. Detecting Downward Change

The change time ν is estimated to be the last maximum time before detection. Similarly, the change time ν is estimated to be the last minimum before the stopping time. Notice that this test corresponds to a linear transformation $g(x) = x$ as described in Chapter II with bias terms: $\mu_0 \pm k$. Figure 3.3 illustrates the recursive version of Page's test when $\Delta_{\min}$, the

unknown change magnitude is set to be 0.2, the change time is at 100 and the SNR of the input signal is –3dB.

D. PERFORMANCE EVALUATION OF CUMSUM PROCEDURES

As shown in Chapter II, we characterize the performance by the mean time between false alarms T, and the mean delay for detection D. The asymptotic ratio between log T and D was shown to be defined by (2–53) and (2–55):

$$\eta = \lim_{\substack{T\to\infty \\ a\to\infty}} \frac{\log T(a)}{D(a)} \geq h(\theta_0)\cdot E\{g(x)|\theta_1\} = \underline{\eta} \qquad (3-8)$$

where a is the threshold of the test. This relationship is influenced by two factors: The first is the transformation or nonlinearity $g(x)$. The second is the statistical properties of the observation *before* the change (the root $h(\theta_0)$ is a function of the SNR) and *after* the change $\left(E\{g(x)\,|\,\theta_1\}\right)$.

In the sequel, several nonlinearities are presented and analyzed in different noise environments. Special attention is given to these situations which result in equality in (3–8), namely:

$$\lim_{\substack{T\to\infty \\ a\to\infty}} \frac{\log T(a)}{D(a)} = h(\theta_0)\cdot E\{g(x)|\theta_1\}$$

resulting in an easy way to calculate the asymptotic performance of the detector.

Notice that for the cases where (3-8) is an equality, the relationship $(\log T)/D$ enables a comparison of Lorden bounds (2–46), (2–47) and Wald's bounds (2–48), (2–50) for the pair (T,D) with the correct performance measure

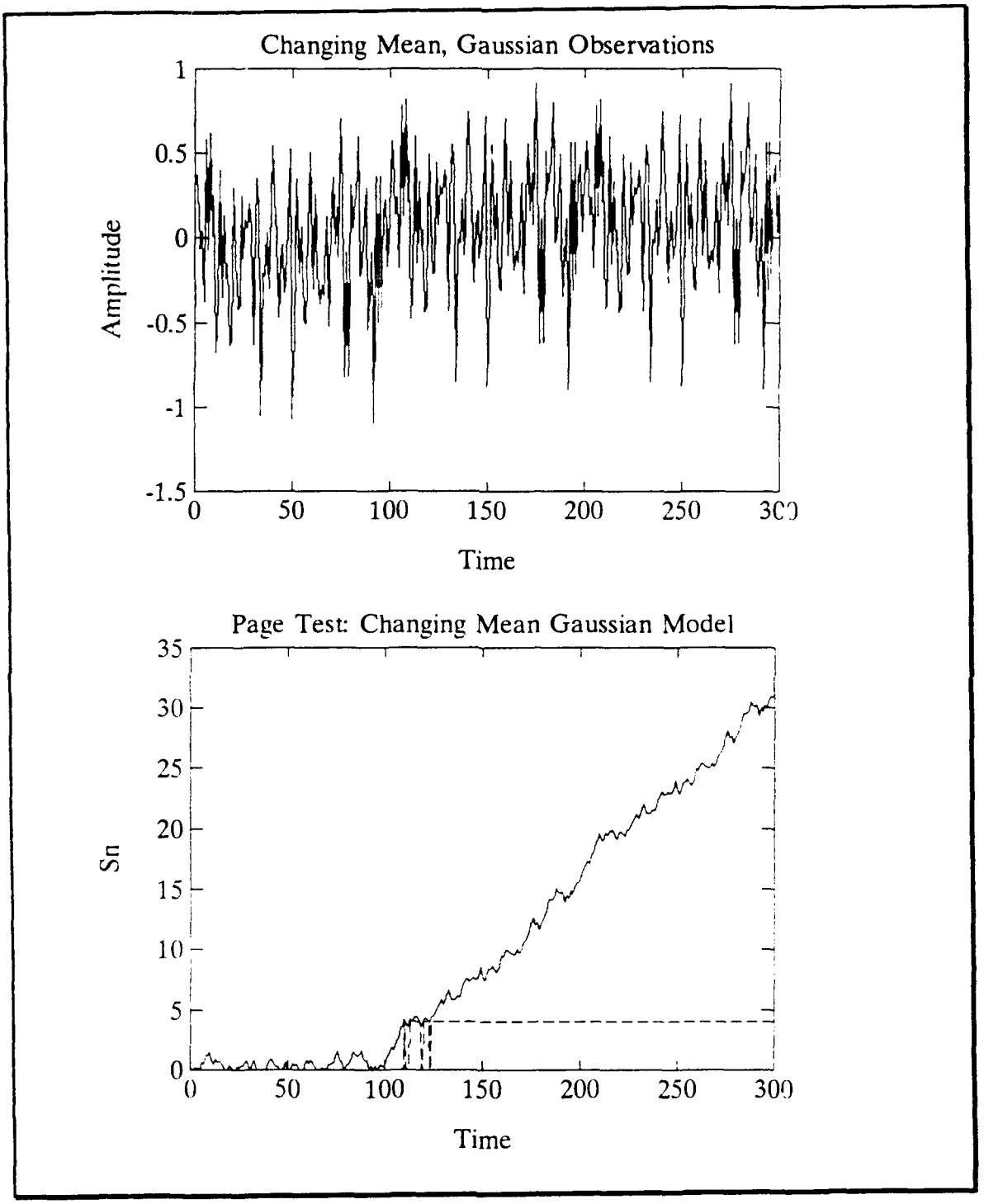


Figure 3.3. Detecting a Change in the Mean of Gaussian Observations using Page's Test

(3-10). Both bounds are based on the root of the moment generating function $h(\theta_0)$ and the statistics after the disorder $E\{g(x) \mid \theta_1\}$. The following subsections present some examples for which performance curves specified by calculating pairs of (T,D) for many values of a, the stopping boundary, and k, the bias term. Thus, the use of the approximating equations for (T,D) enables us to find the pair (a,k) for a given performance requirement (T,D).

1. Parametric Detection

For parametric detection schemes it is assumed that the general form of the statistics before and after the change is known. If the parameters after the change are not known, composite testing techniques could be used as shown in the sequel.

To illustrate the performance curves, we consider the situations where the noise distributions before and after the change are both Gaussian

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-x^2 / 2\sigma^2\right)$$

and also the case where both densities are Gauss-Gauss mixtures (Kassam, 1987):

$$P(x) = (1-\varepsilon)\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-x^2/2\sigma_0^2} + \varepsilon \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-x^2/2\sigma_1^2} \qquad (3-9)$$

with variance $\sigma^2 = (1-\varepsilon)\sigma_0^2 + \varepsilon\sigma_1^2$.

The Gauss-Gauss mixture density is the first two terms in Middleton's Class A model where the noise density function is modeled by an infinite weighted sum of Gaussian densities with decreasing weights and increasing variances, and has been used to model interfering waveforms

(pulses) and narrowband noise. The parameter ε indicates the amount of contamination and is typically in the range (0,0.25). For small enough values of ε, the behavior of $P(x)$ near the origin is dominated by that of σ_1^2. For large values of $|x|$, σ_0^2. dominates the behavior of $P(x)$ since its tails decay at a slower rate than do those of σ_0^2. Thus, the relative strength of the contamination is given by the power ratio $\gamma^2 \triangleq \sigma_1^2 / \sigma_0^2$. Adjusting the parameters (ε,γ) we can determine the performance of the cumsum procedures for a wide range of distributions including those with heavy tails.

A second disorder situation results in the assumption that before the disorder $P_0(x)$ is Gaussian while after the disorder $P_1(x)$ is a Gauss-Gauss mixture. We consider the linear detector $g(x) = x$, and the nonlinear log-likelihood detector and the local optimal energy detector $g(x) = x-1$.

a. Detecting Disorder in Gaussian Measurements

If $g(x)$ is the log-likelihood nonlinearity, then it has been shown in Chapter II that in the limiting situation the bound is tight, i.e., $= \eta = \underline{\eta}$ where

$$\eta = \lim_{T\to\infty} \frac{\log T}{D} \approx I(\theta_1,\theta_0)$$

where $I(\theta_1,\theta_0)$ is the Kullback-Liebler number defined in (2–23). In the case of a change in the mean, i.e., $X_i \sim N(\mu_0,\sigma^2)$ for $i < \nu$, $X_i \sim N(\mu_1,\sigma^2)$ for $i \geq \nu$, the Kullback-Liebler information number is given by (Therrien, 1989)

$$\begin{aligned} \eta &= \lim_{T\to\infty} (\log T / D) \\ &= I(\theta_0,\theta_1) \\ &= (\Delta_\mu)^2 / 2\sigma^2 \end{aligned} \qquad (3-10)$$

with

$$\Delta\mu = \mu_1 - \mu_0.$$

Thus the result can be directly related to the signal to noise ratio $\Delta\mu/\sigma$. Notice that this result is consistent with the result obtained in (3-2) for detecting jumps in the mean of i.i.d. Gaussian observations using the nonlinearity

$$g(x) = \frac{\Delta\mu}{\sigma^2}(x - \mu_0 - \Delta\mu/2)$$

which results from the log-likelihood ratio test. For this nonlinearity:

$$\begin{aligned}\eta &= E\{g(x)|\theta_1\} \\ &= \frac{\Delta\mu}{\sigma^2}(\mu_1 - \mu_0 - \Delta\mu/2) \\ &= (\Delta\mu)^2/2\sigma^2.\end{aligned}$$

In the case of detecting a change in the variance of zero mean Gaussian i.i.d. observations, the log-likelihood ratio results in a square law type detector and is given by

$$g(x) = cx^2 + \ln\gamma$$

where

$$c = \frac{1}{2}\frac{\sigma_1^2 - \sigma_0^2}{\sigma_1^2\sigma_0^2}, \qquad \gamma = \frac{\sigma_0}{\sigma_1}.$$

Notice that for detecting an upward change ($\gamma < 1$), c is positive and $\ln\gamma$ is negative, while for detecting a downward change ($\gamma > 1$), c is negative and $\ln\gamma$ is positive. This explains the behavior of the Page test as illustrated in Figure 3.5. Thus, the performance measure is given by

$$\eta = E\{g(x)|\theta_1\}$$
$$= c\sigma_1^2 + \ln\gamma$$
$$= \frac{1}{2}\left[\gamma^{-2} - 1\right] + \ln\gamma \qquad (3-11)$$

which is as expected the Kullback-Liebler information number for this case.

The closed form in which the asymptotic performance measure is given allows one to compute easily the performance curves.

Figure 3.4 illustrates the performance curves when detecting a disorder in the mean of Gaussian measurement (as illustrated by Figure 3.3) using the optimal nonlinearity $g(x) = \frac{\Delta\mu}{\sigma^2}\left(x - \mu_0 - \Delta\mu/2\right)$ for different signal to noise ratios (Equation 3-10). The predicted results obtained for the delay as a function of a given SNR agrees with the simulation results shown in Figure 3.3 within a tolerance of up to 10 samples.

Figure 3.5 illustrates a changing variance Gaussian signal with $\gamma = 1.2$ (downward change), and change time at 150. Also, the optimal Page test using the square law nonlinearity $g(x) = cx^2 + \ln\gamma$ applied to this signal is shown. Notice that in this case of $\gamma > 1$, $F\{g(x) \mid \theta_0\} < 0$ while $E\{g(x) \mid \theta_1\} > 0$ as needed.

Figure 3.6 illustrates the performance curves for the square law detector nonlinearity (3-11). Notice that when $\sigma_1 \rightarrow \sigma_0$ which means that the changes become undetectable, the bound given by Equation (3-11) turn to be noninformative since $\eta \rightarrow 0$. Thus, the delays obtained for values of γ approaching 1 are higher than those obtained for values of γ which are distant from 1. Notice also the bell curve shape of the ARL function for this detection scheme (which is consistent with the example shown in Figure 2.9).

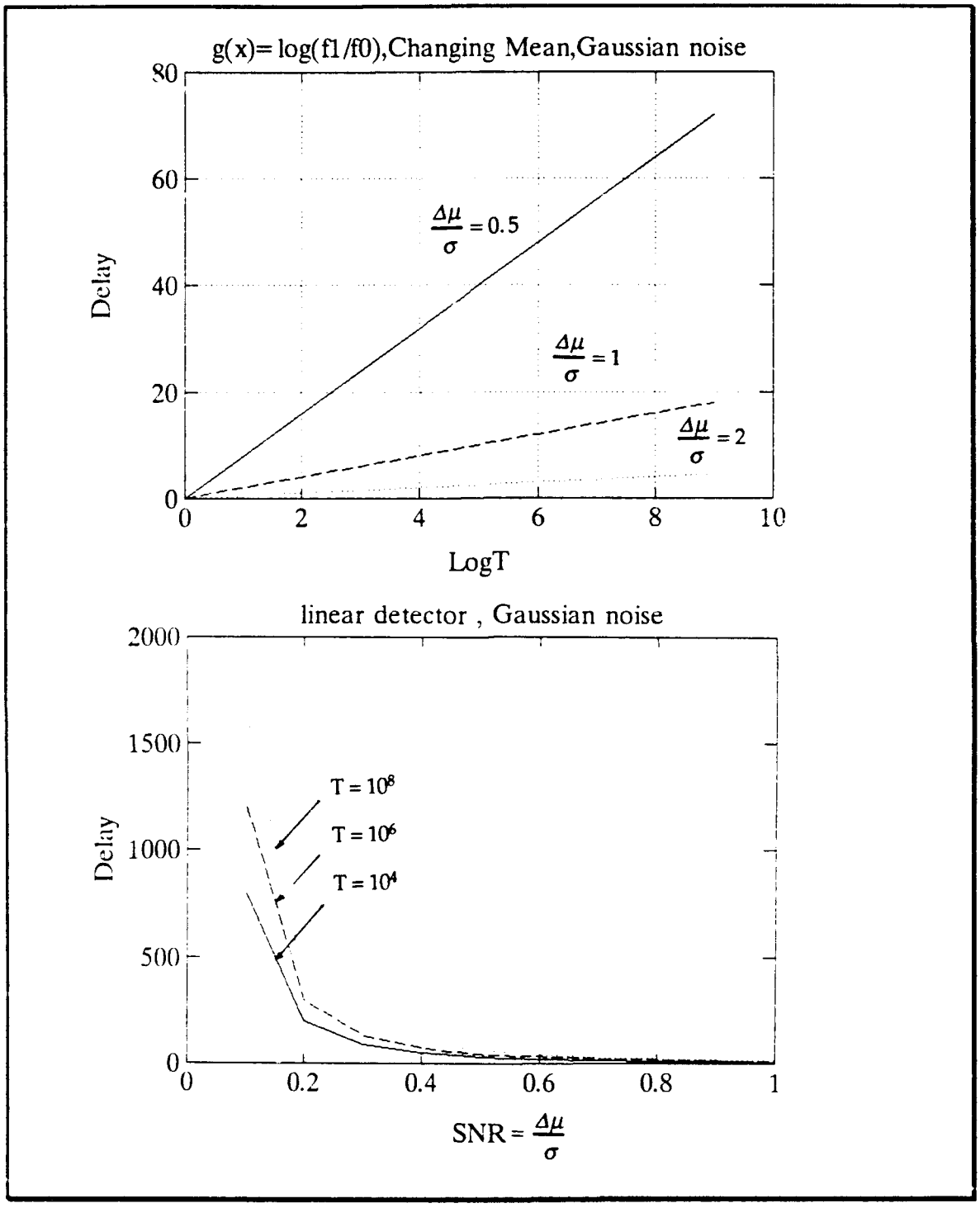


Figure 3.4. Performance Curves for Page's Test Implemented with the Linear Detector $g(x) = \frac{\Delta\mu}{\sigma^2}(x - \mu_0 - \Delta\mu/2)$

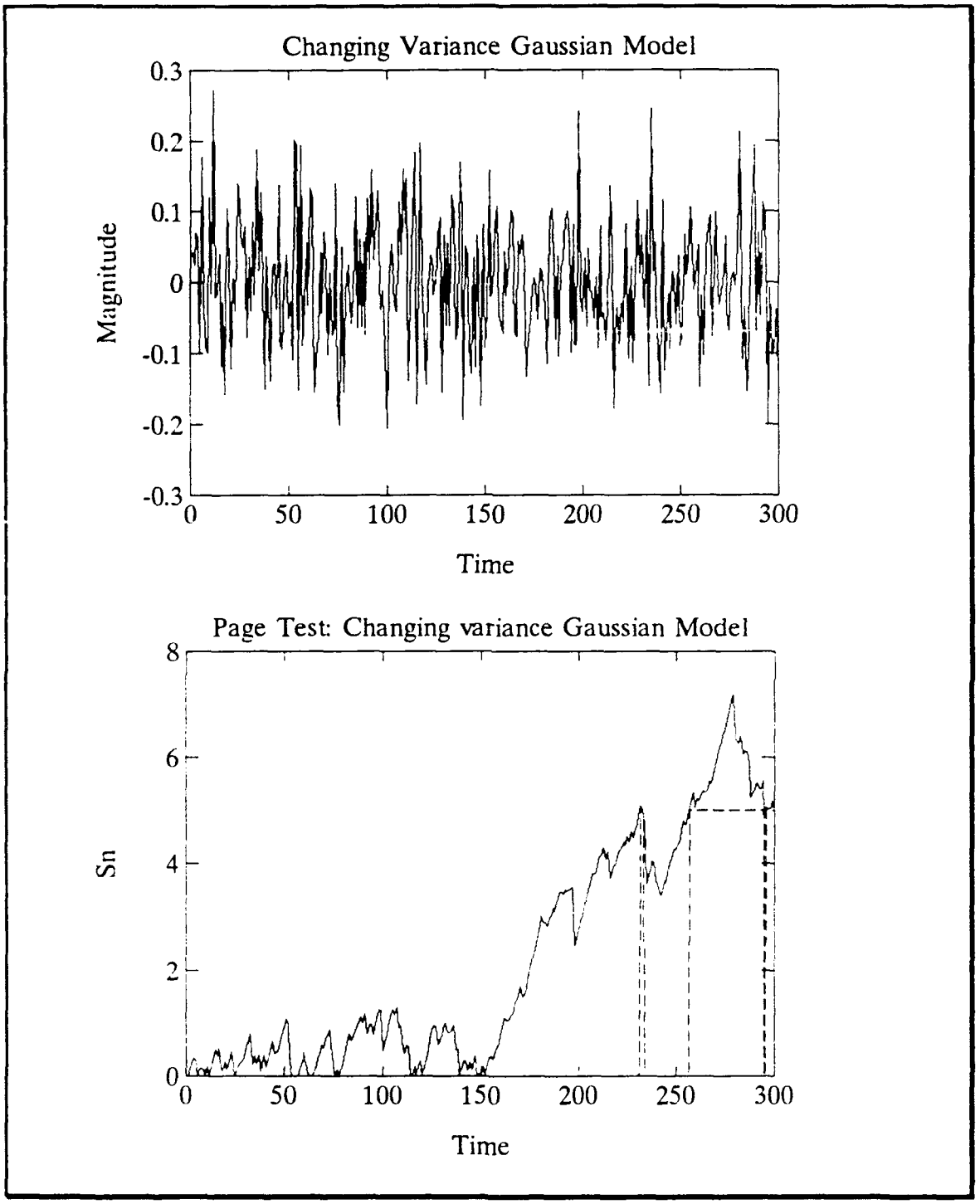


Figure 3.5. Changing Variance Gaussian Signal and the Corresponding Page Test Implemented with the Square Law Nonlinearity

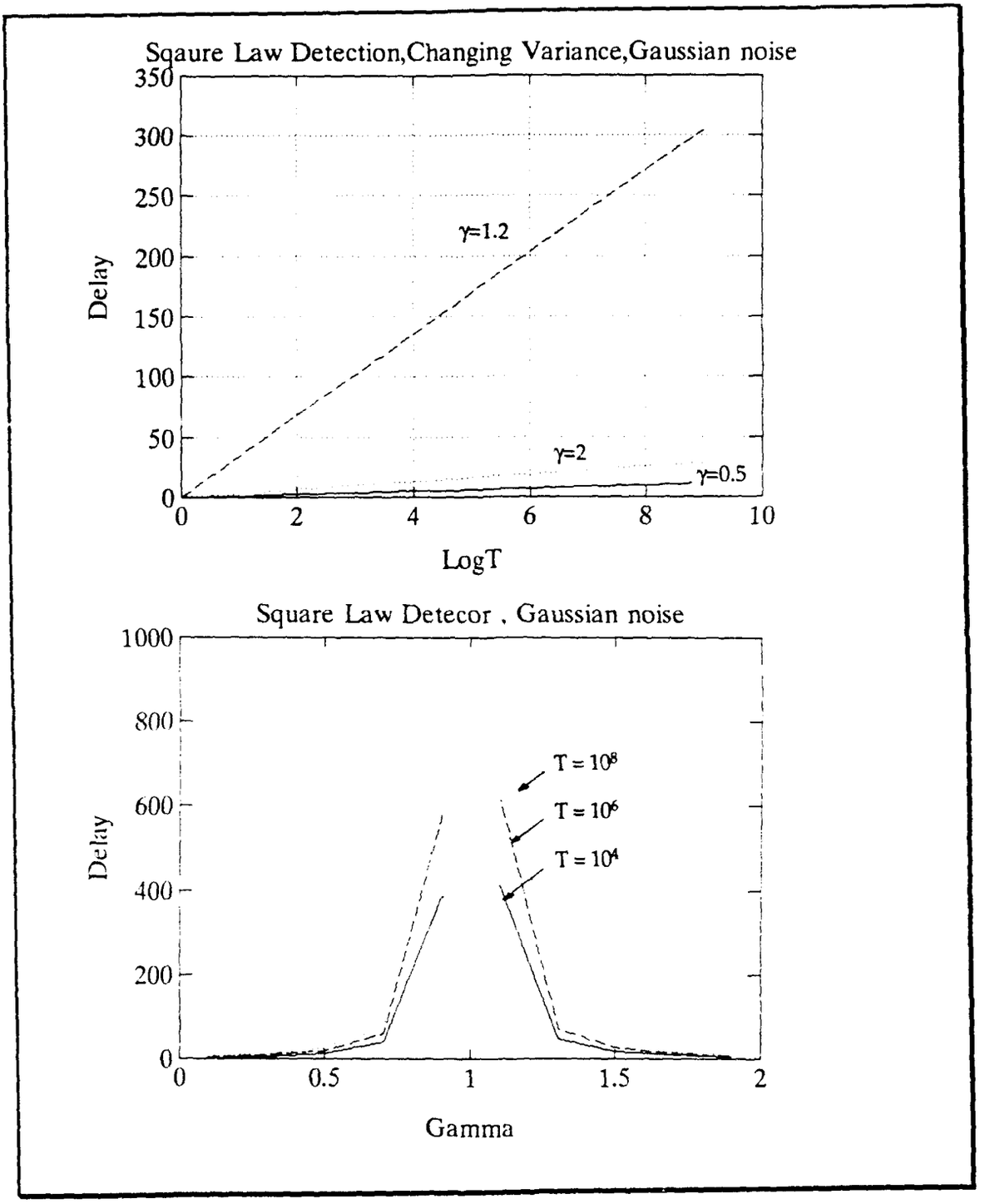


Figure 3.6. Performance Curve Obtained for Page's Test Implemented with the Square Law Detector $g(x) = cx^2+\ln\gamma$

b. *Locally Optimum Energy Detector for Spectral Signatures*

Consider the case where we observe the energy spectral density of a signal. Under the "no change" hypothesis we assume without loss of generality that the background noise process $\{y\}$ is normalized (i.e. $\sigma = 1$) White Gaussian Noise (WGN) and is grouped in disjoint blocks of M points for processing via the Discrete Fourier Transform (DFT). Hereby we assume that the sample blocks are mutually independent. The squared magnitudes of the M complex outputs of the DFT are computed and these random outputs denoted by $\{X_{i,m}\}$, $i = 1, 2, \ldots$, $m = 1, 2, \ldots, M$ where i is the block number and m is the frequency bin number, form the Periodogram and are available as the observations for the detection procedures. Namely,

$$\{X_{i,m}\} = \left|\mathrm{DFT}\{y_{i,m}\}\right|^2$$

where

$$y_{i,m} = y(iM + m) \quad i = 1,2,\ldots \quad m = 1,2,\ldots,M.$$

Hereby, we are interested in detecting a change within a specific frequency bin, while the method described here can be also used to detect a change from block to block as was done by Broder (Broder, 1990), thus, our method modifies Wolcin's method (Wolcin, 1983) by looking for a change in an orthogonal direction (frequency) to the direction (block) used by Wolcin's. Moreover, we use a narrow "window" of frequencies to detect changes within several frequency bins in order to detect a certain spectral signature.

Under the white Gaussian noise assumption, the variables $\{X_{i,m}\}$ except the first one $m = 1$ the first frequency bin, are independent and

identically distributed with exponential distribution and unity mean, having the χ_1^2 distribution (Kay, 1988)

$$P_0(X_{i,m}) = \exp\{-X_{i,m}\}.$$

Under the change hypothesis, the distribution of $\{X_{i,m}\}$ containing the signal in addition to the WGN, will also be presumed to be exponential but now with mean $\mu_{i,m} > 1$. This is due to the fact that under the change hypothesis X_{im} has a noncentral χ_1^2 distribution with a noncentral parameter (Whalen, 1971) $\lambda > 0$, thus the mean μ of the non central χ_1^2 distribution is given by

$$\mu = \lambda + 1 > 1.$$

Thus, if we assume that after the disorder $\mu_{i,m}$ does not depend on i, we have

$$P_1(X_{i,m}) = \mu_m^{-1} \exp\{-X_{im} / \mu_m\}.$$

This is the case when the signal itself is also a Gaussian signal which is independent of the background WGN. Hence, the original hypothesis testing of

$$H_0: \{y_i\} \sim \text{WGN}$$

versus

$$H_1: \{y_i\} \sim \text{Gaussian signal} + \text{WGN}$$

in the signal domain, is equivalent to the hypothesis testing in the spectral domain.

$$
\begin{aligned}
H_0: \quad & X_{i,m} \sim \exp\{-X_{i,m}\} && i = 1,2,\ldots, \quad m = 1,\ldots,M \\
& \text{versus} && \qquad (3\text{-}12) \\
H_1: \quad & X_{i,m} \sim \mu_m^{-1} \exp\{-X_{i,m} / \mu_m\} && i = 1,2,\ldots, \quad m = 1,\ldots,M \\
& \mu_m > 1.
\end{aligned}
$$

Because the parameters $\{\mu_m\}$ are not known *a priori*, Page's test with the optimal nonlinearity the log-likelihood ratio cannot be implemented. Thus, we will use composite hypothesis techniques such as the Locally Most Powerful (LMP) test statistic. In Chapter II we introduced the local optimum nonlinearity.

$$
\begin{aligned}
g_{\ell o}(x) &= \frac{d}{d\theta} P(x;\theta) / P(x;\theta) \\
&= \frac{d}{d\theta} \ln \frac{P(x;\theta+\Delta\theta)}{P(x;\theta)} \Bigg|_{\Delta\theta = 0}.
\end{aligned}
$$

where $P(x;\theta)$ denotes the observations density conditioned on the parameter θ. This test measures small deviations from the "null" hypothesis, hence, as was shown in Chapter II, it maximizes the efficacy (incremental signal to noise ratio) of the test.

Using this function for testing between the hypotheses $\mu = 1$ and $\mu > 1$ for the case of univariate exponential distributions yields the following nonlinearity

$$
\begin{aligned}
g_{\ell o}(x) &= \lim_{\mu \downarrow 1} \frac{d}{d\mu} \log \frac{P(x|\mu > 1)}{P(x|\mu = 1)} \\
&= \lim_{\mu \downarrow 1} \frac{d}{d\mu} \log \frac{\mu^{-1} \exp\{-x/\mu\}}{\exp\{-x\}} \\
&= \lim_{\mu \downarrow 1} \frac{d}{d\mu} \left(-\log \mu - x/\mu + x\right) \\
&= \lim_{\mu \downarrow 1} \left(x/\mu^2 - \mu^{-1}\right) \\
&= x - 1.
\end{aligned}
$$

Thus, $g_{\ell o}(x)$ does not depend on μ after the change, this results in a locally most powerful test for all $\mu > 1$.

Implementing Page's test for bin number m yields

$$\tilde{S}_{i,m} = \max\left\{0, \tilde{S}_{i-1,m} + g\left(X_{i,m}\right)\right\}$$

$$\tilde{S}_{0,m} = 0$$

where
$$g\left(X_{i,m}\right) = X_{i,m} - 1 - k \qquad (3-13)$$

where k is a positive parameter or reference value needed to bias the test for the null hypothesis, such that $E\{g(x_{i,m}) \mid \mu_m = 1\} < 0$, since Page's test performs better when the mean of the nonlinearity before the change takes place is negative as opposed to zero.

At this point it is important to notice a robustness property of this detector. Since the method is based upon detecting changes in the energy (periodogram), and since it is assumed that the disorder is independent of the background noise, the presence of the signal with a certain frequency component will increase the total energy in the corresponding frequency bin

which is to be detected. Hence, the underlying signal model should not assume a specific model for the signal.

The performance of Page's test (3–13) is determined by the parameters a and k. Hence, with two degrees of freedom the test results in many pairs (a,k) that yield the same performance. The problem is to find a specific pair which results in a high detection probability. In order to determine the performance in this situation, notice that since the parameter μ is not known a priori, the performance measure η cannot be determined since $E\{g(x) \mid \theta_1\}$ is not explicitly known. Thus, we shall use Lorden's bounds (2–46), (2–47) and Wald's bounds (2–48), (2–50) to obtain informative bounds for the false alarm rate. In order to obtain these bounds it is necessary to find the root h of the moment generating function identity (2–11) before the disorder (Broder, 1990).

$$
\begin{aligned}
1 &= E\left\{\exp\left\{h \cdot g\left(X_{i,m}\right)\right\} \middle| \mu_m = 1\right\} \quad m = 1,\ldots,M. \\
&= E\left\{\exp\left\{h \cdot \left[X_{i,m} - 1 - k\right]\right\} \middle| \mu_m = 1\right\} \\
&= \exp\{-h(1+k)\} \cdot E\left\{\exp\left\{hX_{i,m}\right\} \middle| \mu_m = 1\right\} \\
&= \exp\{-h(1+k)\} / (1-h). \qquad (3-14)
\end{aligned}
$$

The root is shown to be a function of the bias term k. Figure 3.7 illustrates this relationship. Notice that the root h does not depend on the DFT length. This fact will be shown to be the key to the surprising observation shown in the sequel that the SNR per bin does not depend on the DFT length.

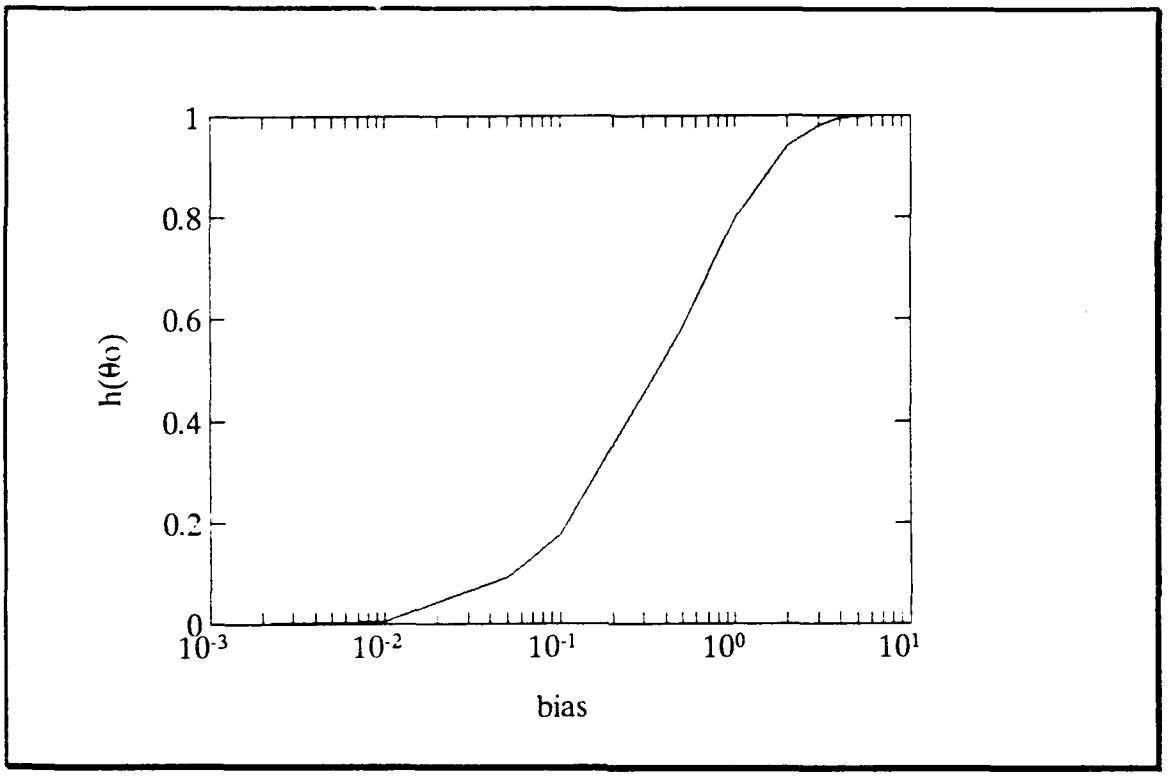


Figure 3.7. The Root of the Moment Generating Function Identity (2–11) for $g(x) = x-1-k$ as a Function of the Bias Term k

Notice that the root is upper bounded by $h < 1$. Recall that by (2-46) $T \geq \exp\{h(\theta_0) \cdot a\}$. Thus, large values of h are desired. Recall also that η is lower bounded by $\underline{\eta}$. Consequently, for a given false alarm rate, a larger $\underline{\eta}$ corresponds to a smaller delay. Thus, from (2–55) it is clear that larger values of h are desired, which means that biasing the test with larger values is favorable.

In order to improve the poor statistical properties of the periodogram (standard deviation of the order of the mean), a window of length $W = 3$ that groups the expected frequency bin and the two neighboring frequency bins was taken. Thus the statistic function $g(x)$ was modified to

$$g(X_{i,m}) = \frac{1}{3}\sum_{m=m_\ell}^{m_\ell+2}(X_{i,m} - 1 - k) \qquad (3-15)$$

where $m_\ell, m_{\ell+1}, m_{\ell+2}$, are the frequency bins used by the window. A typical time/frequency sample grid is shown in Figure 3.8.

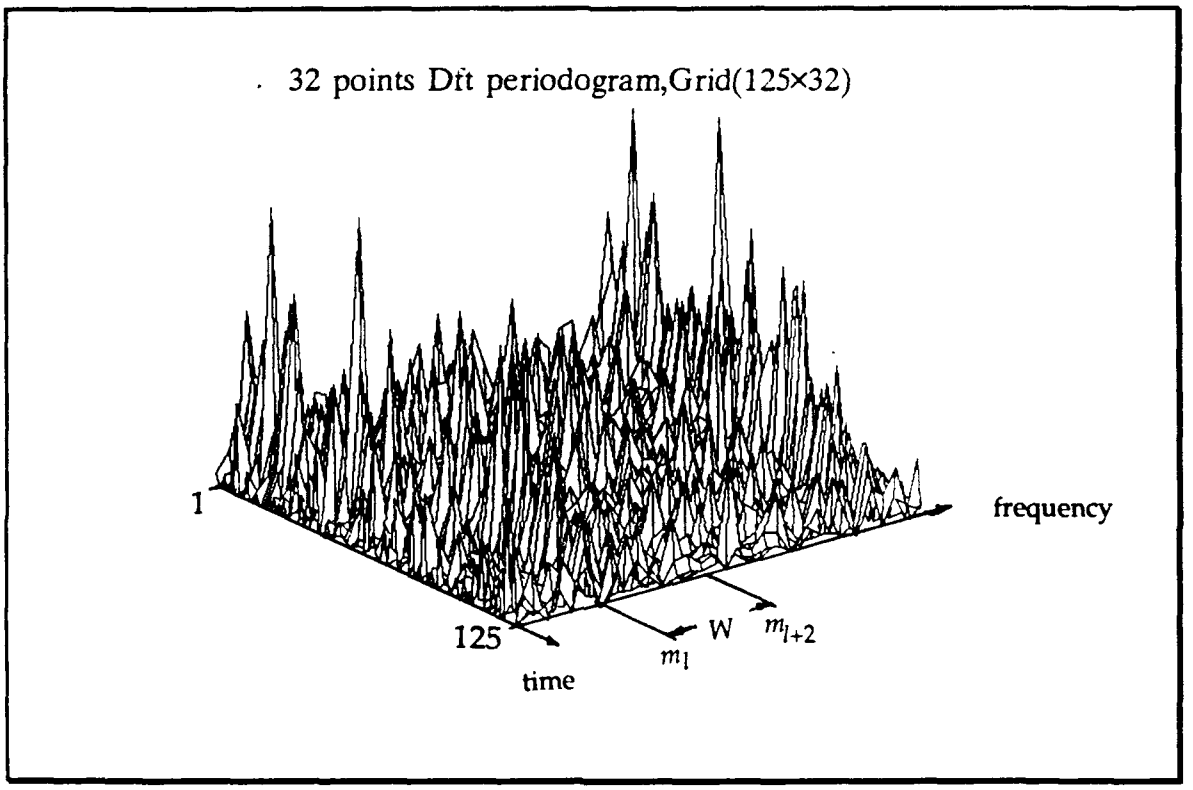


Figure 3.8. Time/Frequency Sample Grid

Notice that in this case the root location depends on the window length since for that case the moment generating function has the form

$$1 = E\left\{\exp\left\{h\frac{1}{3}\sum_{m=m_\ell}^{m_{\ell+2}}\left(X_{i,m}-1-k\right)\right\}\middle|\mu_{i,m}=1\right\}$$

$$= E\left\{\prod_{m=m_\ell}^{m_{\ell+2}}\exp\left\{\frac{h}{3}\left(X_{i,m}-1-k\right)\right\}\middle|\mu_{i,m}=1\right\}$$

$$= \left(E\left\{\exp\left\{\frac{h}{3}\left(X_{i,m}-1-k\right)\middle|\mu_{i,m}=1\right\}\right\}\right)^3$$

$$= \frac{3}{3-h}\exp\left\{-\frac{h}{3}(1+k)\right\}. \qquad (3-16)$$

Figure 3.9 illustrates the root location for the given window $W = 3$.

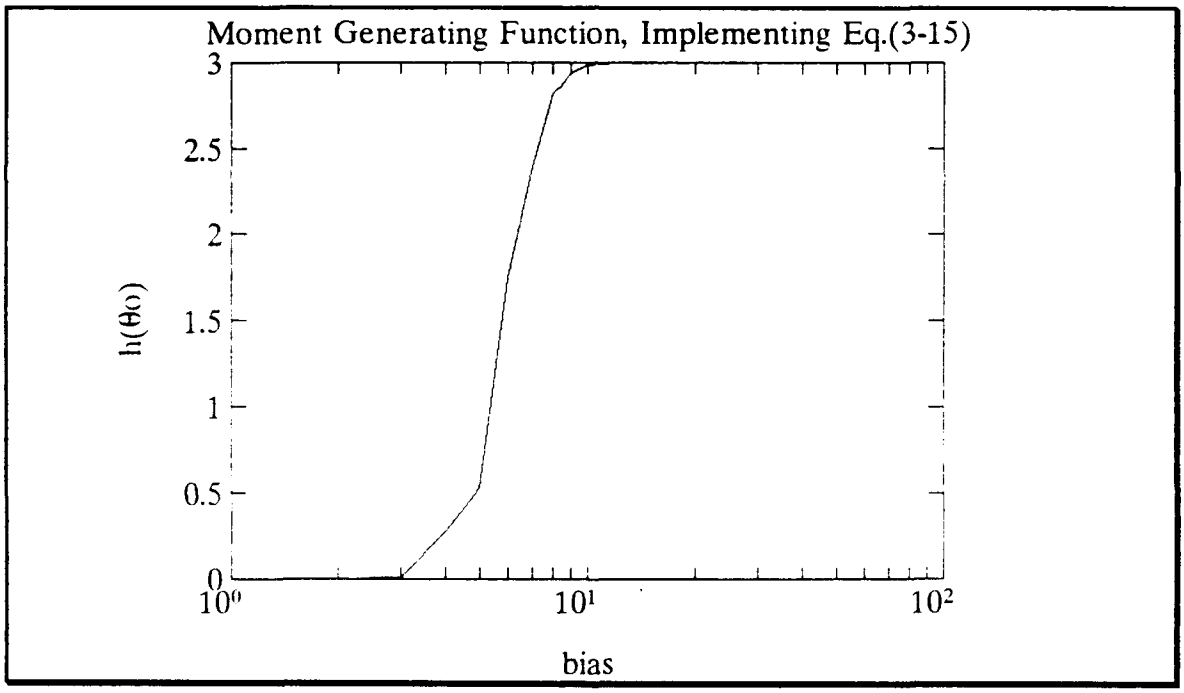


Figure 3.9. Root of Moment Generating Function for

$$g(x) = \frac{1}{3}\sum_{m=m_\ell}^{m_{\ell+2}}\left(X_{i,m}-1-k\right)$$

Notice that the root is upper bounded by 3. This implies that the mean time between false alarms T will be larger in this case than the previous one, since with the same bias level a higher root value is obtained. However, if the averaging of the window frequencies were done by the function

$$g(X_{i,m}) = \sum_{m=m_\ell}^{m_{\ell+2}} (X_{i,m} - 1 - k) \qquad (3-17)$$

the root location would be the same as in (3-14), i.e., upper bounded by 1. This may imply that the averaging method (3-15) performs better than the others.

The problem is that we would like to determine the performance with a given pair of (a,k), but since the function $g(x)$ was based upon a suboptimal hypothesis test, only bounds (Lorden and Wald) can be derived. To resolve this problem the following method is presented.

Consider that we are given the desired mean time between false alarm T and some minimum value for μ, say $\mu_{\min}(>1)$ of μ_m for which we can test. In this situation, Page's test using the optimal nonlinearity, the Log-Likelihood Ratio (LLR) can be implemented. This results in

$$g(x) = (1-\lambda)x + \log\lambda \qquad (3-18)$$

where $$\lambda = \mu_{\min}^{-1}$$

and Page's test (3–13) is implemented with the function (3–15) and a new threshold a'. In order to find the relationship between the pairs (a,k) and (a',k) of the tests (3–13) and (3–15) we will use the following analysis:

$$N_{\ell o} = \inf\{i: \quad \tilde{S}_i \geq a\}, \qquad \tilde{S}_i \text{ implemented with } (3-13)$$

$$N_{LLR} = \inf\{i: \quad \tilde{S}_i \geq a'\}, \qquad \tilde{S}_i \text{ implemented with } (3-15)$$

Hence, we obtain

$$N_{LLR} = \inf\{i: \quad \tilde{S}_{i-1} + X_{i,m} + \log[\lambda_m / (1-\lambda_m)] \geq a'/(1-\lambda_m)\}$$

$$N_{\ell o} = \inf\{i: \quad \tilde{S}_{i-1} + X_{i,m} - 1 - k \geq a\}. \qquad (3-19)$$

To achieve the same performance requires that the following relationships will sustain

$$a = a'/1-\lambda_m$$

$$k = \log[\lambda_m / (1-\lambda_m)] - 1. \qquad (3-20)$$

Notice now that for the log-likelihood ratio function $h(\theta_0) \equiv 1$. Thus, for the given average time between false alarms, T, equation (2–46) becomes

$$T \geq \exp a'.$$

Hence, the following procedure can be implemented:

- given T, the threshold a' which guarantees that requirement is given by

$$a' = \ln T, \qquad (3–21)$$

- use (3–20) to find both the threshold a and the bias k needed for implementing the local optimum test given T and μ_m. Hence, this test is now "tuned" for the desired performance.

To summarize, this procedure allows the use of optimal nonlinearity in order to find the specific pair (a,k) needed to achieve the performance requirement for the energy detector.

A second and even more practical way to determine the test parameters (a,k) is by using the SNR per bin which is required to meet the performance requirements. Hereby, the notation S relates to the signal and N to the noise (in the spectrum domain). Decomposition of the data yields (provided that energy exists only in one of the frequency bins)

$$S+N: \quad E\{g(x)|\theta_1\} = \frac{1}{3}\sum_{m=m_\ell}^{m_\ell+2}\left(E\{X_{i,m}\}-1-k\right)$$
$$= \frac{\mu_m - 1}{3} - k$$
$$= \underline{\eta} / h(k).$$

Thus,

$$\text{SNR} = \mu_m - 1$$
$$= 3\left[\frac{\underline{\eta}}{h(k)} + k\right]. \qquad (3-22)$$

Notice that k and $h(k)$ were determined to achieve a given lower bound for T, thus, $\underline{\eta}$ given by (3–8) determines the asymptotic ratio for the desired pair (T,D). Hence, using equation (3–22) enables us to find the corresponding SNR per bin which is required to achieve the desired performance. Figure 3.10 shows the SNR required per bin as a function of the bias term k for different values of the asymptotic measure $\underline{\eta}$. Notice that each given k corresponds to a certain T, thus, the corresponding delay value, D, is found from the graph by using the assigned $\underline{\eta}$ needed for certain SNR.

Analyzing (3–22) reveals an important result. Larger values for $\underline{\eta}$ correspond to a smaller delay, D, in detecting a disorder. Thus, larger values

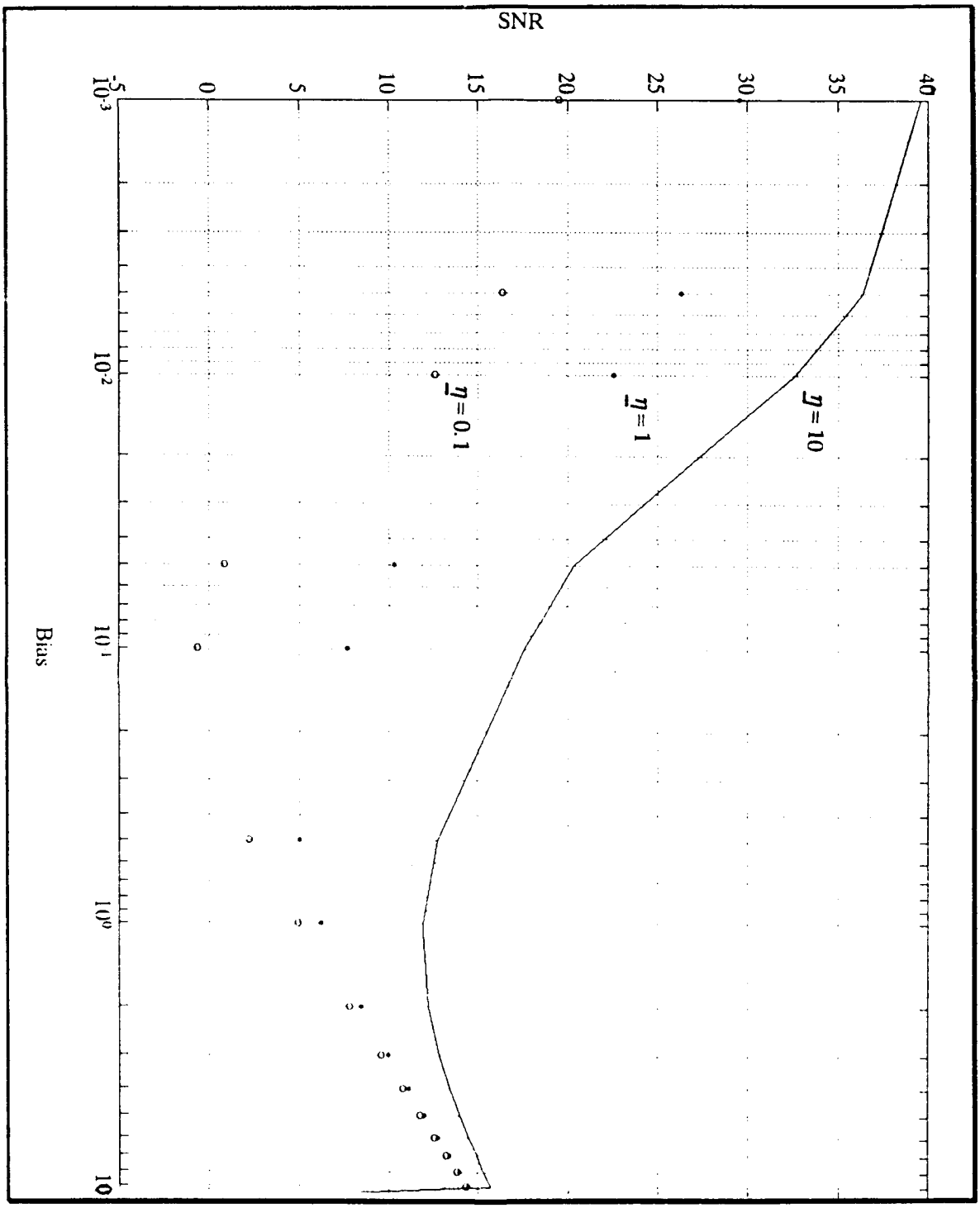


Figure 3.10. SNR per Bin for Energy Detector as a Function of the Bias k Implementing Nonlinearity (3-15)

of k are needed. If the function (3-13) had to be used, the SNR per frequency bin would remain the same. Thus, using (3-15) does not improve the minimal SNR required per bin to achieve some level of detection probability, but improves the overall performance by having a lower false alarm rate. However if we implement nonlinearity (3-17), decomposition of the signal and noise yields (provided that energy exists only in one of the frequency bins)

$$\begin{aligned} E\{g(X_{i,m})|\theta_1\} &= \sum_{m=m_\ell}^{m_\ell+2} \left(E\{X_{i,m}\} - 1 - k\right) \\ &= \mu - 1 - 3k \\ &= \frac{\eta}{h(k)} \end{aligned}$$

hence, the minimal SNR per frequency bin is given by

$$\begin{aligned} SNR &= \mu - 1 \\ &= \frac{\eta}{h(k)} + 3k. \end{aligned} \qquad (3-23)$$

Figure 3.11 illustrates the SNR function as a function of the bias k for the nonlinearity (3.17). Hence, there is an SNR improvement of the order of 1–3dB. This is a surprising result because one would expect that since the root for (3–17) is upper bounded by one as opposed to the root of (3- 15) which is upper bounded by 3, the overall performance of (3–15) will be better. Thus, a tradeoff between the delay and the minimal SNR required for detection is determined by the bias k. Moreover, analyzing (3–23) reveals an important result. Larger values for $\underline{\eta}$ correspond to lower delay and better detector

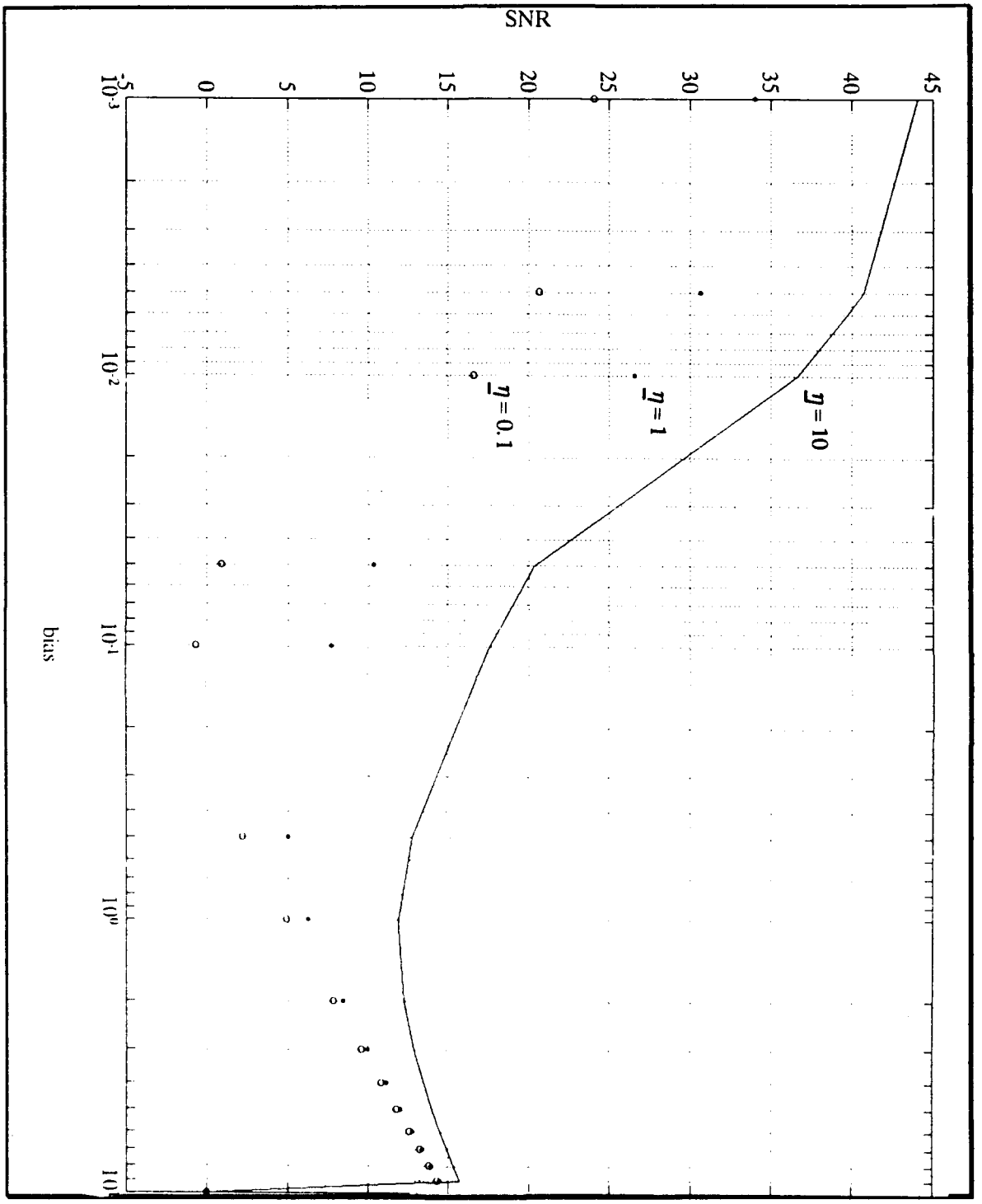


Figure 3.11. SNR per Bin as a Function of the Bias k Implementing Nonlinearity (3–17)

performance, thus, larger values of k are needed. But this is opposed to having lower values of k which are needed to obtain the required SNR per bin. Hence, the chosen bias term k should reflect a tradeoff between these two conflicting requirements.

Simulation results were done by using the function (3–17) data records of length 4000 samples where the change point was at sample 2000 (i.e., middle of the record). The Nyquist frequency used was 500Hz, and at the change point the transition was from 62Hz to 156Hz. We used two algorithms, one of which uses a 32-point DFT producing a time/frequency grid of (125×32) points and the other uses a 128-point DFT producing a time/frequency grid of (30×128) points, where the corresponding processing gains are 12dB and 18dB respectively. Hence, using Figure 3.10 allows one to predict the detection performance. An incoming signal with input SNR of –3dB cannot be detected by using a 32-point DFT since the output SNR is 9dB, which is below the minimum SNR per bin required for detection. By using a 128-point DFT, the output SNR is 15dB, which is about 3dB above the minimal SNR required for detection. The same analysis done by using signals with input SNR of –6dB reveals that the 32-point DFT cannot detect the changes, while a 128-point DFT copes with the detection successfully. Figure 3.12 illustrates the time/frequency grid for the case of using a 32-point DFT with input SNR of –6dB, while Figure 3.13 illustrates Page's test implemented on bins 19, 20, 21 (bin 20 being the 156Hz bin) by using a 128-point DFT to detect energy at 156Hz with input SNR of –3dB and –6dB respectively. Similarly, Figure 3.14 illustrates Page's test implemented on bins 4, 5, 6 (bin 5 being the 156Hz bin) by using a 32-point DFT with input SNR of –3dB and –6dB respectively.

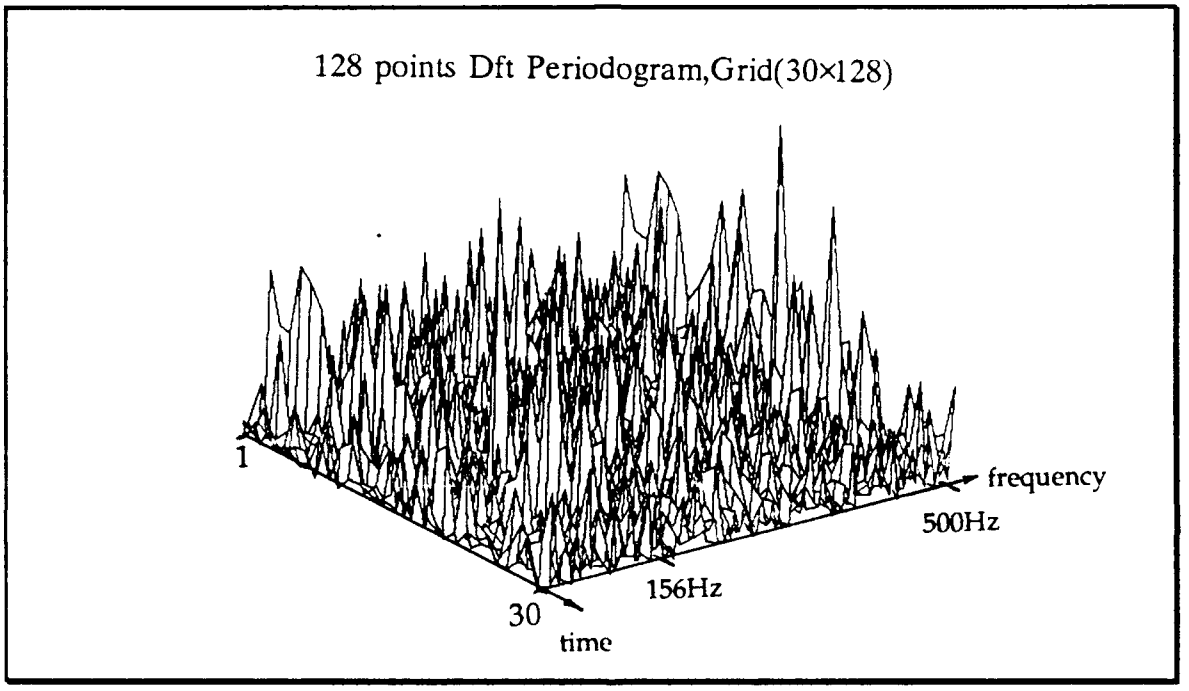


Figure 3.12. Typical Time/Frequency Grid of (30×128) Points. 128 Point DFT, Input SNR = –6dB

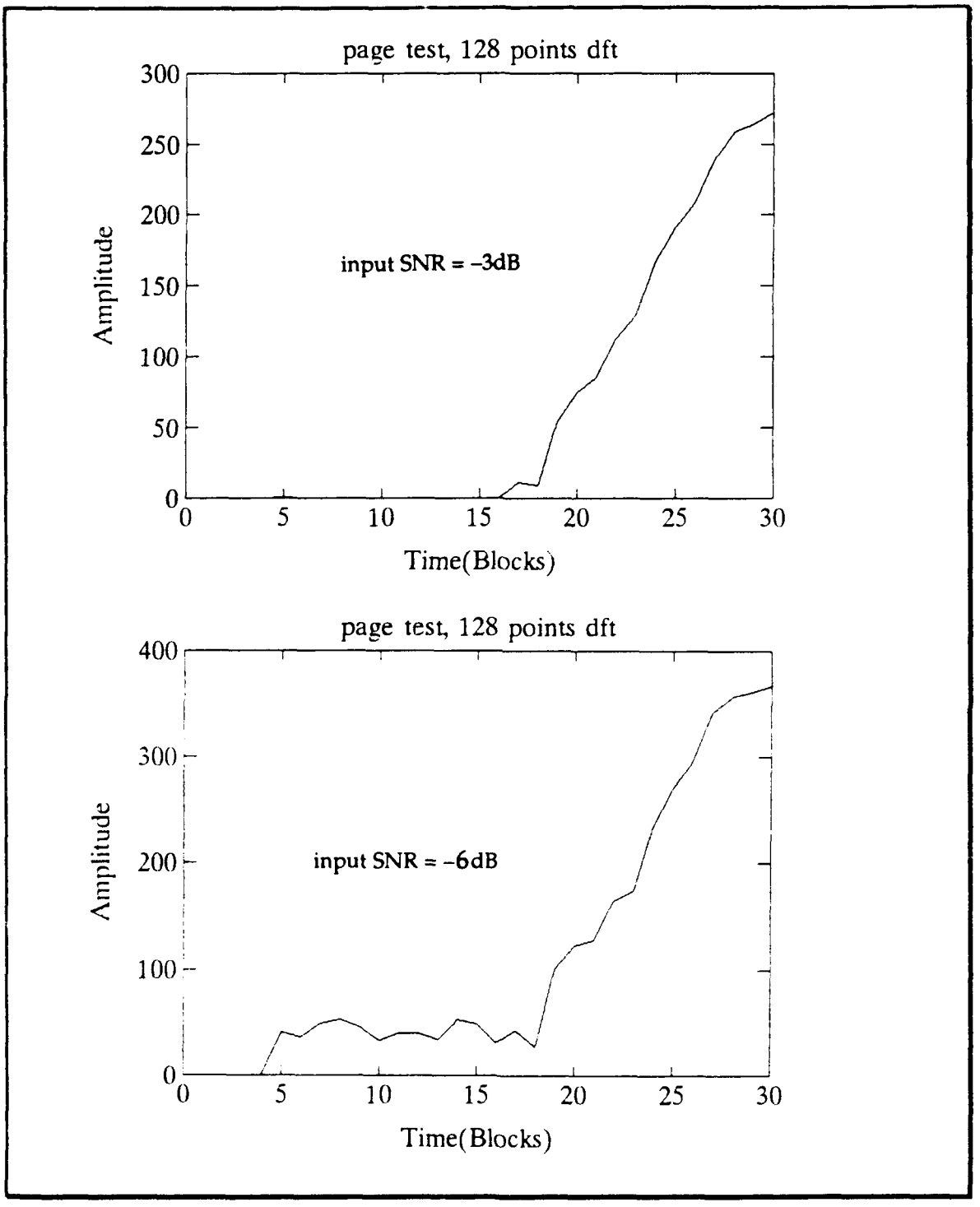


Figure 3.13. Page's Test Implemented on Bins 19, 20, 21 of a 128-Point DFT

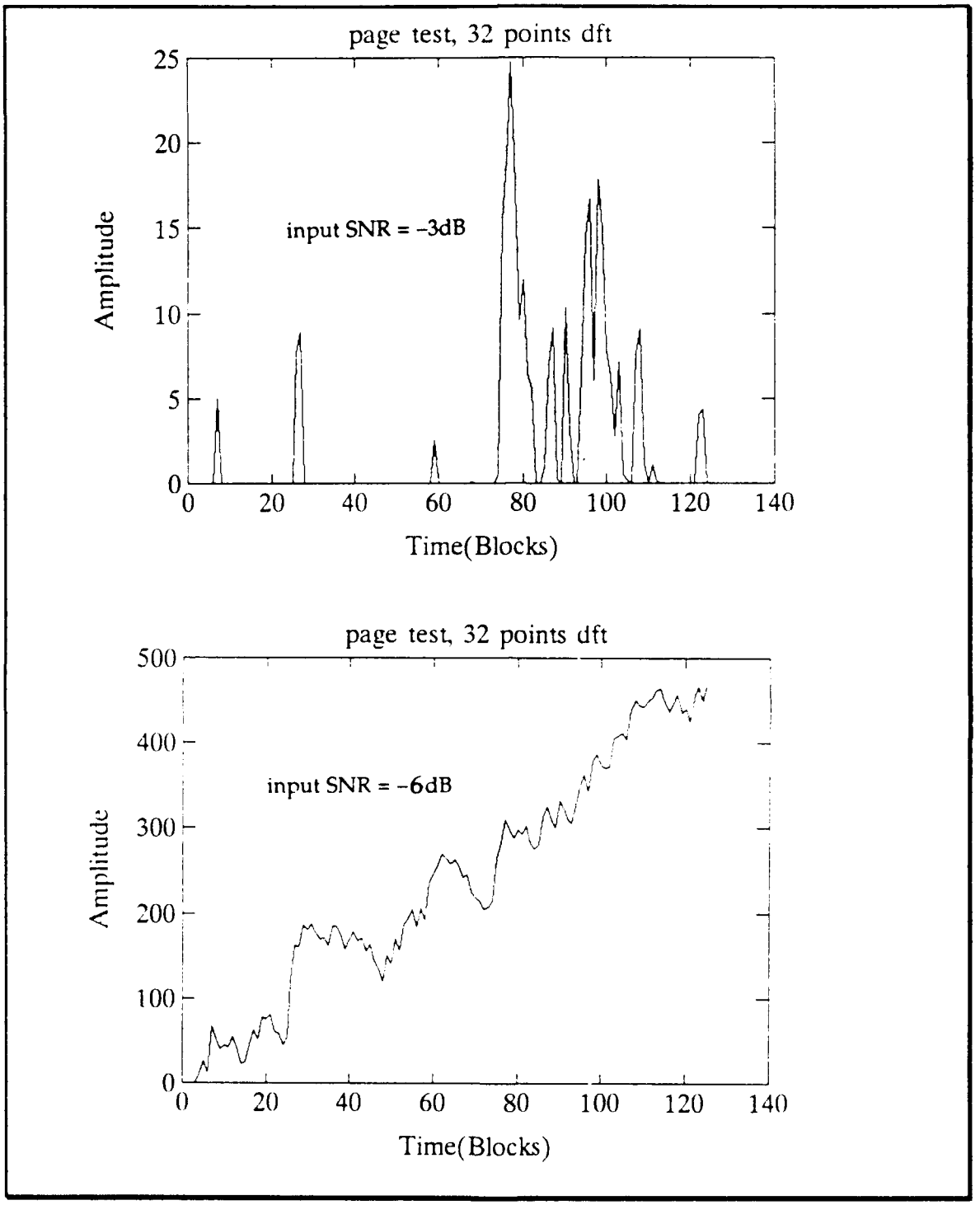


Figure 3.14. Page's Test Implemented on Bins 4, 5, 6 of a 32-Point DFT

In order to compare the detection performance of the Page test with a conventional detection scheme we refer to Whalen (Whalen, 1971) in which the performance (Receiver Operating Characteristic—ROC) for detecting M independent sinewave samples in white Gaussian noise by using a linear detector (which is the locally optimum detector for Gaussian signals, see Kassum, 1988), is analyzed. Even though the detection is not based on energy, the comparison presented in the sequel indicates better performance of our method. Figure 3.13 illustrates the Page detector implemented on a 128-point DFT. For an incoming signal with SNR of –6dB the delay for detection is 4 blocks and the minimum SNR required for detection (using the proper bias value to minimize the SNR) is about 12dB for η=10 and about 6dB-8dB for η=1. The corresponding bounds for the false alarm rate are 10^{-40} and 10^{-4} respectively. Figure 3.15 illustrates the ROC for a linear detector for detecting four independent samples (equivalent to delay in detection of four blocks) of a sinewave in white Gaussian noise (Whalen, 1971, p. 250) where the parameter is the SNR required for detection. For this classical detection scheme the ROC is in terms of P_{FA} versus P_D. Thus, to compare the performance of these two methods we refer only to values of $P_D \rightarrow 1$ to reflect that the detection is almost surely certain. Figure 3.15 illustrates that for values of 6dB–8dB the performance of the linear detector is very poor since the P_{FA} is in the order of 10^{-1}-10^{-3} respectively, while for the Page test it is at least 10^{-4}. Furthermore, for the linear detector as the P_{FA} is lowered, at a given (fixed) SNR the P_D decreases, while for the Page detector, equivalently lower P_{FA} (corresponding to higher mean time between false alarms) requires a higher threshold and reflects in a higher delay but still, the detection is guaranteed. In the operating ranges of above 9dB (which is the typical operating range for this type of detection) the Page test is shown to have better performance than the conventional linear detection.

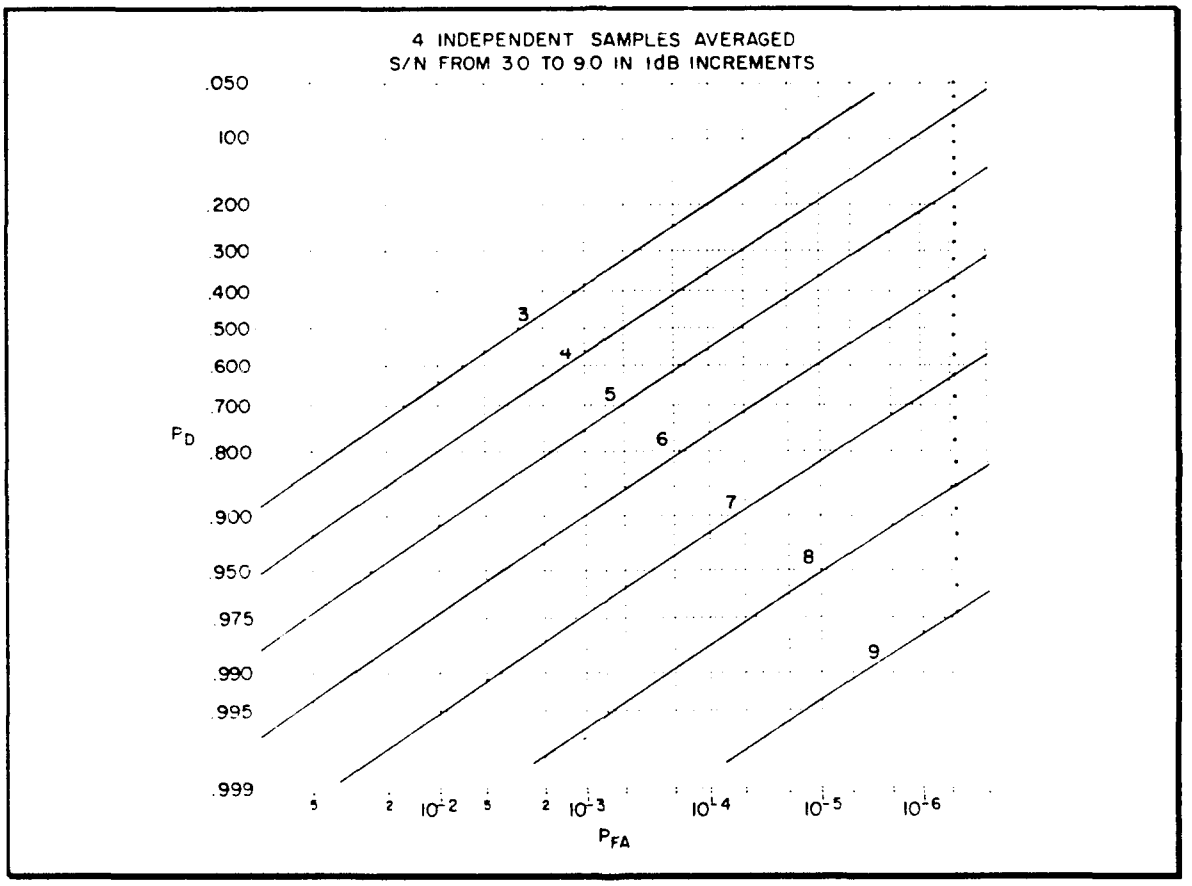


Figure 3.15. ROC for Detecting Sinewaves in White Gaussian Noise (four samples averaged). From Whalen, 1971.

2. Non-Parametric Detection

Hereby we will consider only the sign detector defined as

$$g(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ -1 & \text{for } x < 0 \end{cases}.$$

This nonlinearity is sometimes also referred to as random walk nonlinearity since the output $g(x)$ is a random walk. Thus, results from random walk theory can be used. Define:

$$p(\theta) = \Pr\{g(x) = 1 | \theta\}$$

$$q(\theta) = \Pr\{g(x) = -1 | \theta\}.$$

If $p(\theta) \neq q(\theta)$ there is a positive probability that the process will drift to $+\infty$ if $p(\theta) > q(\theta)$ (and to $-\infty$ if $p(\theta) < q(\theta)$). Thus assuming that $E\{g(x) \mid \theta_0\} < 0$ yields $p(\theta_0) < q(\theta_0)$ while assuming $E\{g(x) \mid \theta_1\} > 0$ results in $p(\theta_1) > q(\theta_1)$.

The moment generating identity is given by

$$E\left\{\exp\{h(\theta_0) \cdot g(x)\} | \theta_0\right\} = p(\theta_0)\exp\{h(\theta_0) \cdot (+1)\} + q(\theta_0)\exp\{h(\theta_0) \cdot (-1)\}$$

$$= 1.$$

Consider $h(\theta_0) = \ln\frac{q(\theta_0)}{p(\theta_0)}$, thus,

$$E\left\{\exp\left\{\ln\frac{q(\theta_0)}{p(\theta_0)} \cdot g(x)\right\} | \theta_0\right\} = p(\theta_0) \cdot \frac{q(\theta_0)}{p(\theta_0)} + q(\theta_0) \cdot \frac{p(\theta_0)}{q(\theta_0)}$$

$$= p(\theta_0) + q(\theta_0) = 1.$$

Hence,

$$h(\theta_0) = \ln \frac{q(\theta_0)}{p(\theta_0)}$$

and

$$E\{g(x)|\theta_1\} = p(\theta_1) - q(\theta_1) > 0.$$

The result is that the lower bound on the performance measures is given by

$$\underline{\eta} = [p(\theta_1) - q(\theta_1)] \log \frac{q(\theta_0)}{p(\theta_0)}.$$

In order to evaluate the performance measure η, we will use results from random walk theory (see Karlin and Taylor, 1984, p. 109) for the approximation of the ARL function of Page's test as done by Broder (Broder, 1990).

$$\mathrm{ARL}(\theta) = \frac{1}{q(\theta) - p(\theta)} \left[\frac{1 - \left(\frac{q(\theta)}{p(\theta)} \right)^a}{\frac{p(\theta)}{q(\theta)} - 1} - a \right] \quad \text{if } p(\theta) \neq q(\theta). \qquad (3-24)$$

Since under θ_0, $q(\theta_0) > p(\theta_0)$, the average time between false alarms for large a can be approximated as

$$T \approx \frac{1}{q(\theta_0)-p(\theta_0)} \cdot \frac{-\left[\frac{q(\theta_0)}{p(\theta_0)}\right]^a}{\frac{p(\theta_0)}{q(\theta_0)}-1}$$

$$= q(\theta_0)\frac{\left[\frac{q(\theta_0)}{p(\theta_0)}\right]^a}{\left[p(\theta_0)-q(\theta_0)\right]^2}$$

$$\approx \left[\frac{q(\theta_0)}{p(\theta_0)}\right]^a.$$

Under θ_1, $p(\theta_1) > q(\theta_1)$, hence, the average delay for large a is given by

$$D \approx \frac{a}{p(\theta_1)-q(\theta_1)}$$

hence, the performance bound η is given by (Broder, 1990)

$$\eta = \lim_{a\to\infty} \frac{\log T}{D}$$

$$\approx \frac{\log\left[\frac{q(\theta_0)}{p(\theta_0)}\right]^a}{a/\left[p(\theta_1)-q(\theta_1)\right]}$$

$$= \left[p(\theta_1)-q(\theta_1)\right]\log\frac{q(\theta_0)}{p(\theta_0)}$$

$$= \underline{\eta}.$$

Using this result allows the comparison of Lorden and Wald bounds with the approximated results from random walk theory.

For the simulation results we considered the symmetric additive signal in noise situation, i.e.

$$P(x|\theta_i) = \begin{cases} P(x-\theta) & i=0 \\ P(x+\theta) & i=1 \end{cases}$$

and the noise environments considered were Gaussian and a Gauss-Gauss mixture. In order to calculate the $p(\theta)$ and $q(\theta)$ parameters as a function of the signal and noise parameters consider the following

$$\Pr\{g(x) = \pm 1\} = \Pr\left\{x \gtrless 0\right\}.$$

Thus, by knowing the mean of the incoming signal we can use the complementary error function to derive both the Gauss and Gauss-Gauss mixture cases, as shown in Figure 3.16. For the Gauss-Gauss mixture

$$p(\theta_1) = (1-\varepsilon)p_1(\theta_1) + \varepsilon p_1(\theta_1)$$

$$q(\theta_1) = 1 - p(\theta_1)$$

for which $p(\theta_1) < q(\theta)$. It follows from the symmetric signal assumption ($\mu_0 = -\mu_1$) that

$$p(\theta_0) = q(\theta_1)$$

$$q(\theta_0) = p(\theta_1)$$

which results in the desired situation of $q(\theta_0) > p(\theta_0)$.

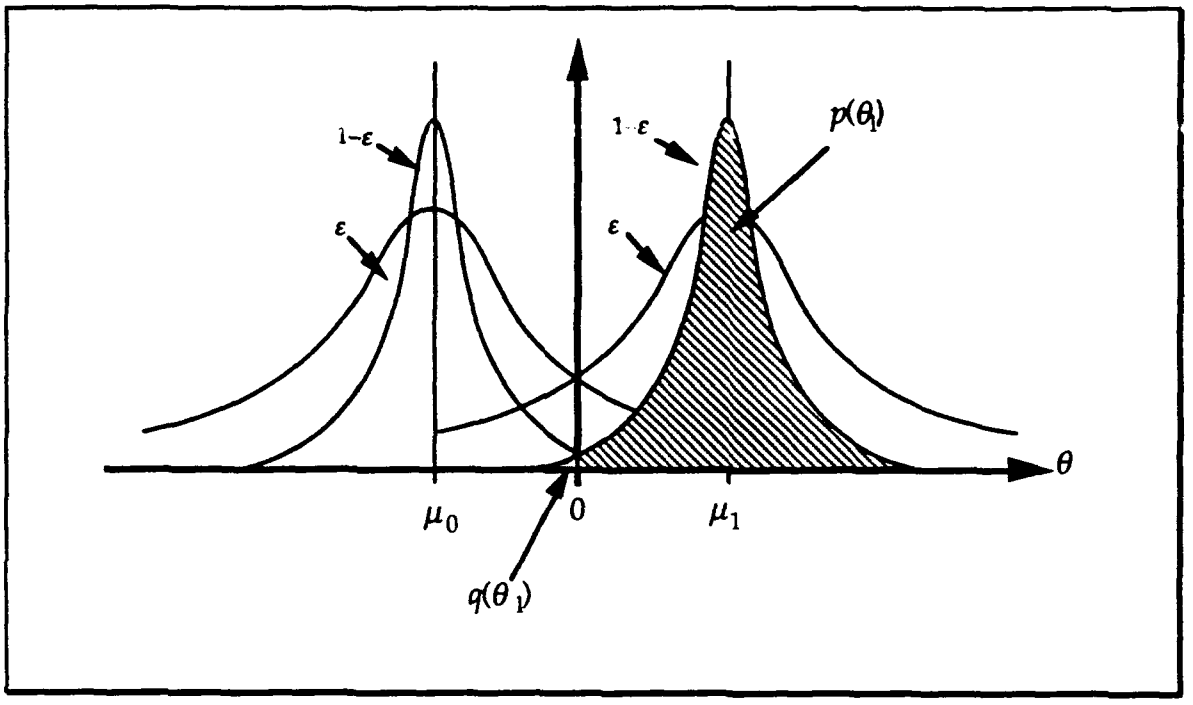


Figure 3.16. $p(\theta)$ as a Function of the Signal and Noise Parameters (Symmetric Case, Gauss-Gauss Mixture)

As shown in the previous example, the root of the moment generating function is needed for Lorden's and Wald's approximation. Figure 3.17 illustrates the root position for different pair values of (p,q). We see that as $p(\theta_0) < 0.5$ becomes larger, the root is smaller which indicates that for a given false alarm rate the delay for detection will be larger due to the fact that $p(\theta_0)$ approaches $q(\theta_0)$, resulting in a difficult decision situation. In the neighborhood where $q(\theta_0)$ is slightly larger than $p(\theta_0)$ $E\{g(x) \mid \theta_0\} \approx 0$. In this situation, biasing the test is needed since the root approaches zero and the bound η is not informative anymore.

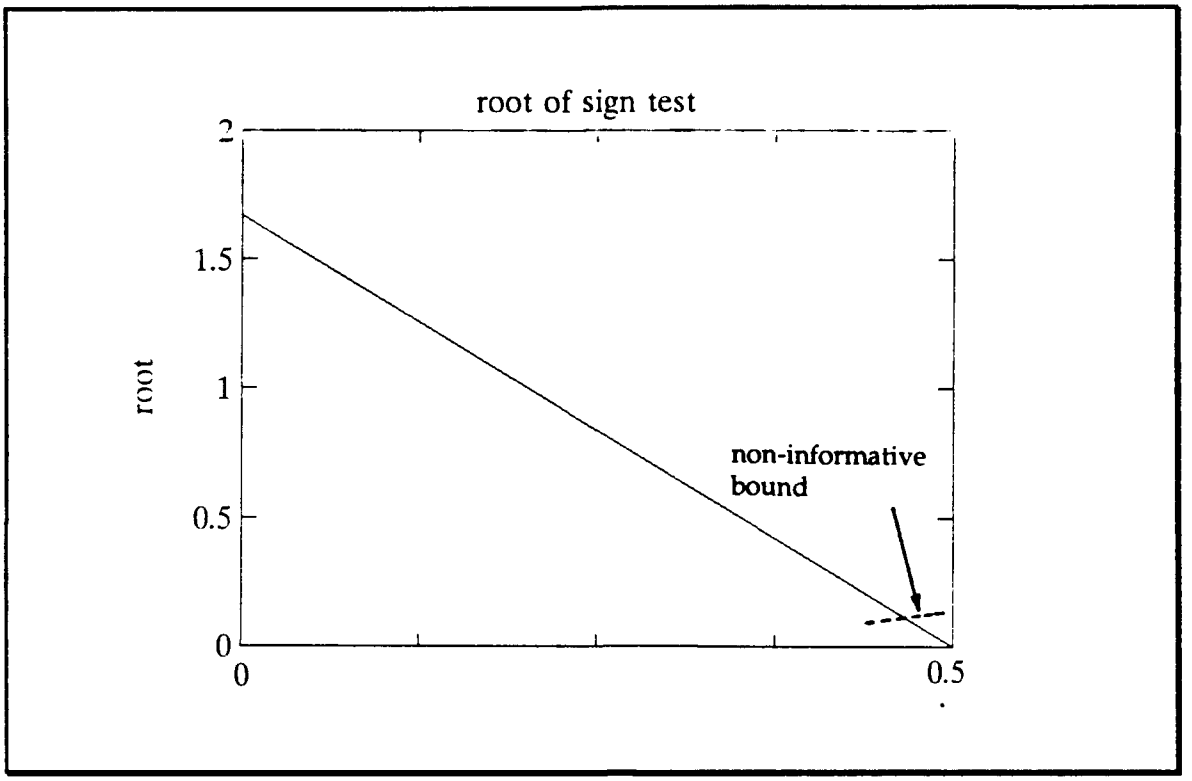


Figure 3.17. The Root of the Moment Generating Function for the Sign Test

Figure 3.18 illustrates Lorden's, Wald's, and the random walk approximation (3-20) as functions of the threshold a for a certain case where before the disorder the difference between $p(\theta_0)$ and $q(\theta_0)$ is large enough. The results indicate good detection bounds. Figure 3.19 illustrates the same approximations but now when $q(\theta_0)$ approaches $p(\theta_0)$, the degradation in performance is shown to be in the order of several magnitudes. The values for $q(\theta_0)$ and $p(\theta_0)$ were chosen to simulate two cases of Gauss-Gauss mixtures, resulting in $p(\theta_0) = 0.15$, $q(\theta_0) = 0.85$ for the first case, and $p(\theta_0) = 0.4$, $q(\theta_0) = 0.6$ for the second case.

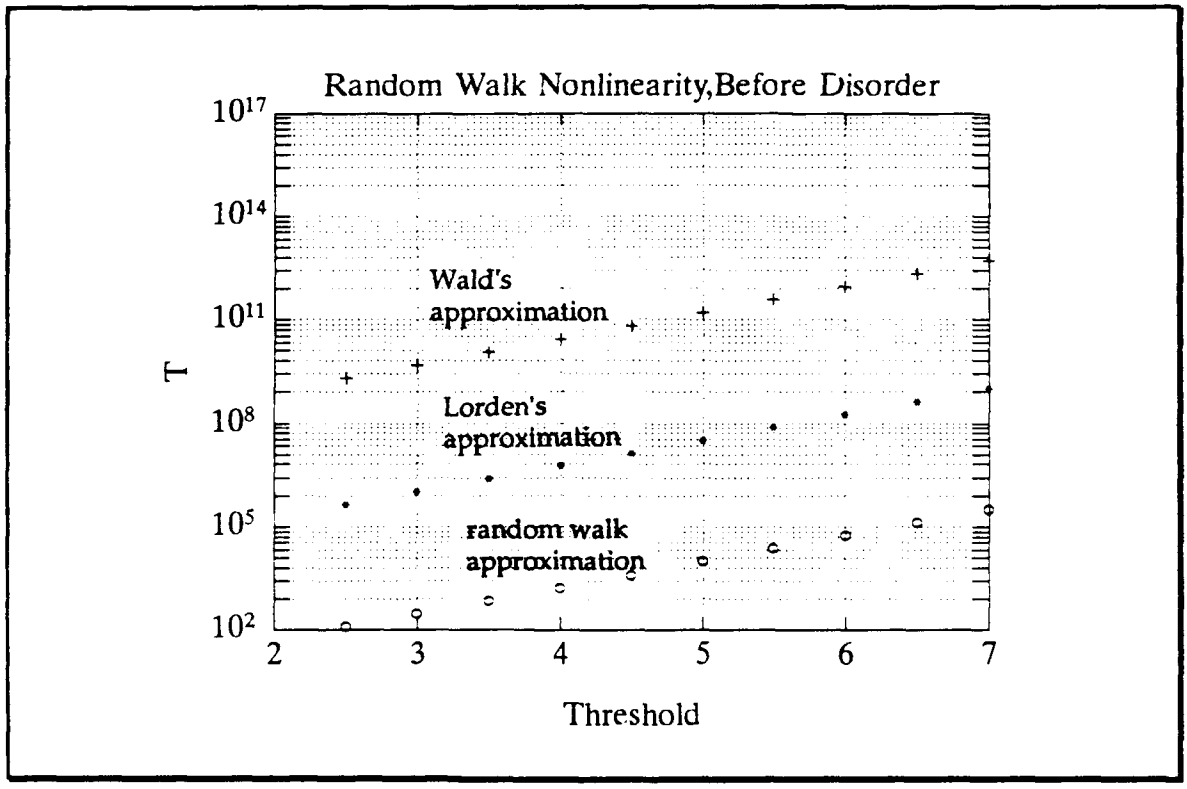


Figure 3.18. Sign Test. Mean Time between False Alarms for $p(\theta_0) = 0.15$, $q(\theta_0) = 0.85$

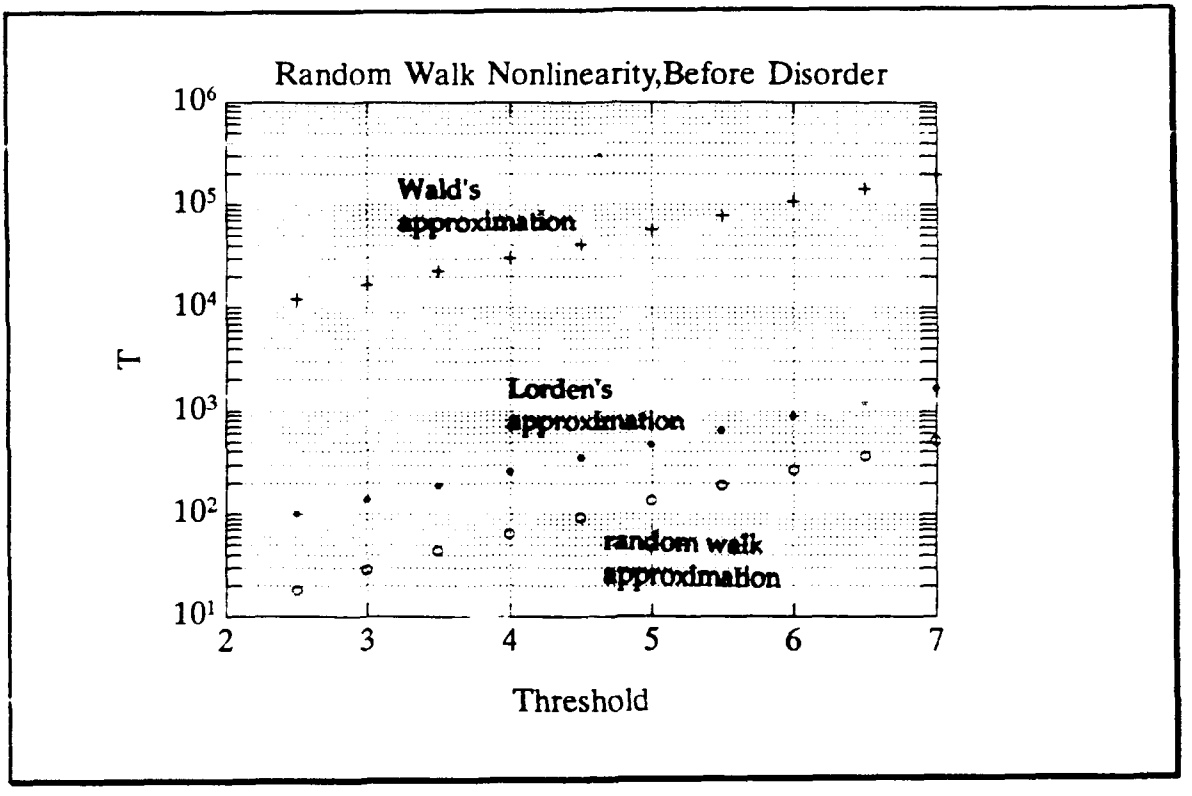


Figure 3.19. Sign Test. Mean time between False Alarms for $p(\theta_0) = 0.4$, $q(\theta_0) = 0.6$

To analyze the delay for detection we use a similar technique, but since we now explore the situation after the disorder, we consider the two corresponding cases where $p(\theta_1)$ is larger than $q(\theta_1)$ and where $p(\theta_1)$ approaches $q(\theta_1)$. The results are similar to those obtained in the case of the false alarm rate and are shown in Figures 3.20 and 3.21 in the form of performance curves for Page's Test implemented with the sign detector. As in the previous case, $p(\theta_1)$ and $q(\theta_1)$ correspond to the same Gauss-Gauss mixture parameters.

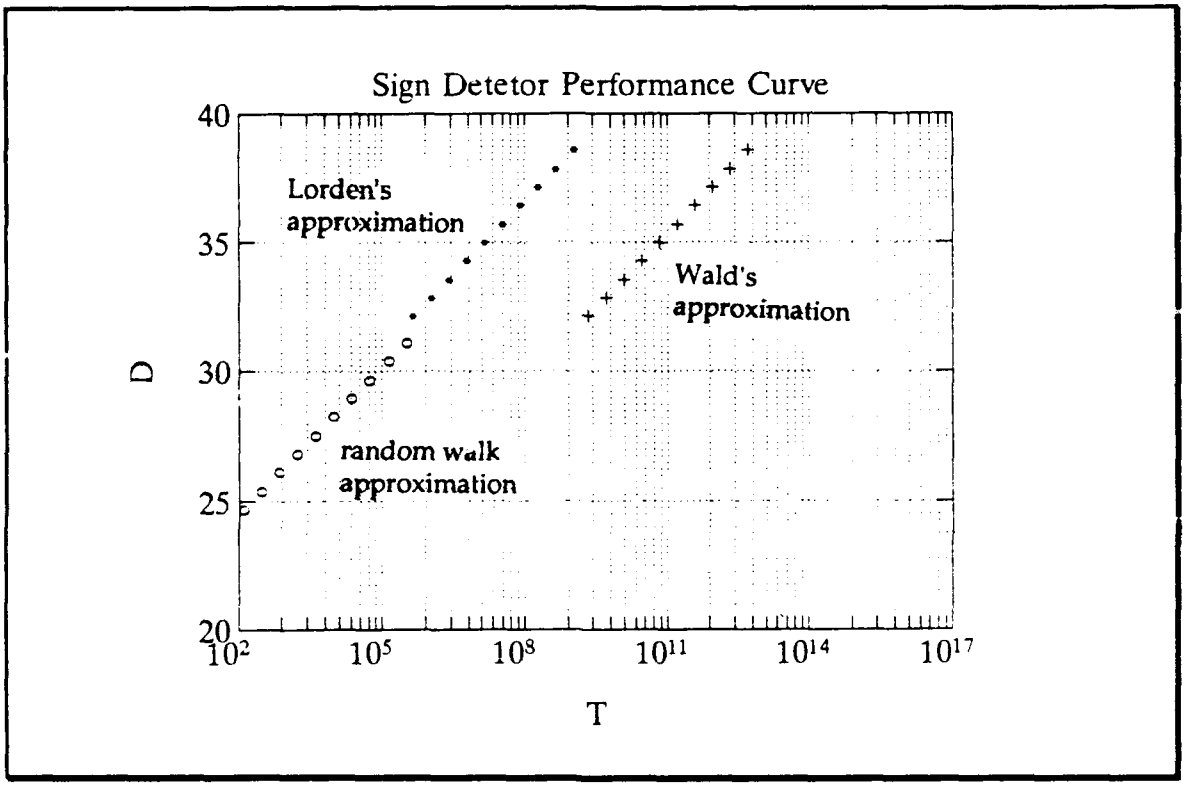


Figure 3.20. Performance Curves for the Sign Detector
$p(\theta_0) = 0.15 \quad q(\theta_0) = 0.85$

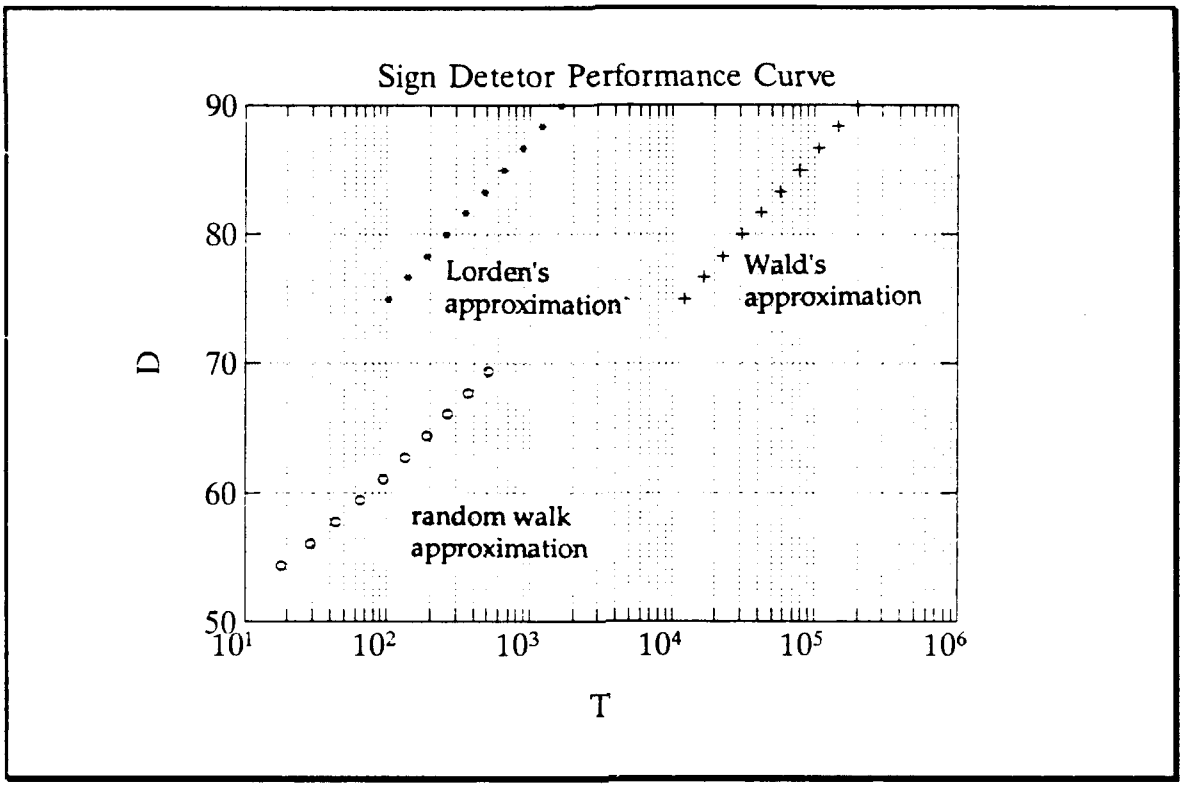


Figure 3.21. Performance Curves for the Sign Detector
$p(\theta_0) = 0.4 \quad q(\theta_0) = 0.6$

E. SUMMARY

In this chapter we have described the problem of the change detection and of the joint estimation of the change time and the model parameters. Within this framework, only the problem of the quickest detection has been investigated by using Page's test. In the parametric framework, the linear detector and the square law detector were shown to be optimal in the sense of quickest detection of changes in the mean and variance of Gaussian observations. In both cases performance measures were derived and shown to be consistent with the actual results of simulations. A new algorithm for detecting changes in the spectral energy was implemented based on locally

optimal testing and shown to be consistent with the analytical performance results obtained for this test. The bias of the test was shown to reflect a tradeoff between the detector performance and the minimal SNR required for detection. Finally, the issue of non-parametric detection was investigated by implementing Page's test with the sign nonlinearity and testing the performance under Gauss and Gauss-Gauss mixture noise distributions.

IV. BROWNIAN MOTION APPROXIMATION TO CUMSUM PROCEDURES

A. INTRODUCTION

In sequential analysis additional simplification results from approximating sums of independent random variables $\sum_{i=1}^{n} x_i$ in discrete time by a Brownian motion process $\{B(t),\ t \geq 0\}$ in continuous time. Moreover, for cases where the observations do not form a Gaussian process, the discrete time process can be approximated by a Brownian motion process which is Gaussian. For further discussion on this subject see Reynolds (Reynolds, 1975).

To understand the motivation of the use of the Brownian motion process as a continuous approximation to the random walk (which describes the cumsum procedures), let $x_1, x_2, \ldots$ be independent and normally distributed with mean μ and unit variance. If $\{B(t),\ t \geq 0\}$ is a Brownian motion with drift μ, then $S_n = \sum_{i=1}^{n} x_i$ and $B(n)$, $n = 0, 1, \ldots$ have the same joint distribution. The analogy is clear: Brownian motion is an interpolation of the discrete time random walk S_n which preserves the Gaussian distributions to the extent that a random walk process is approximately normally distributed for large n. Thus, the Brownian motion process may be used as an asymptotic approximation to a large class of random walks and hence of log-likelihood ratios. A good reference for a detailed discussion of this point is Siegmund (Siegmund, 1985). This chapter concentrates primarily on Brownian motion approximations to cumsum procedures (specifically the Page test). A

continuous Brownian motion process $\{B(t),\ t \geq 0\}$ is used as an approximation to cumulative sums $\Sigma(g(x) \pm k)$ which form Page's test. The original problem of detecting a disorder as described in Chapter II, is now modified in the sense that it can be viewed as a shift in the drift of a Brownian motion approximating a cumsum procedure.

1. Problem Statement

Let ν be the time of shift and let $\mu > 0$ be the amount of shift in the drift of a standard Brownian motion $\{B(t),\ t \geq 0\}$, $B(0) = 0$. Consider the observation process

$$W(t) = \mu(t-\nu)^{+} + B(t), \qquad \mu > 0.$$

Thus the observation process is a Brownian process with drift 0 up to the point of shift ν, and μ after that.

The Page test applied to Brownian motion is defined as follows: stop at the smallest t for which the one-sided test with boundaries 0 and a stops. The test is repeated if the lower boundary 0 is reached before a. Define the stopping role to be as

$$N = inf\{t:\ S(t) \geq a\}$$

where

$$S(t) = \left(W(t) + kt\right) - \min_{0 \leq s \leq t}\left(W(s) + ks\right)$$

for detecting a one-sided positive shift in a drift and

$$S(t) = \max_{0 \leq s \leq t}\left(W(s) + ks\right) - \left(W(t) + kt\right)$$

for detecting a one-sided negative shift in a drift. The variable k is the reference value or the bias of the test. (Recall the fact from Chapter II that it is advantageous to bias the test). Hence, this procedure has two degrees of freedom, k and a to achieve a given desired performance.

2. Organization of this Chapter

The primary goal of this chapter is to analyze the performance of the Page test using the Brownian motion approximation, namely, the evaluation of the Average Run Length (ARL) function under the disorder (Delay) and under no disorder hypothesis (mean time between false alarms). These approximations will be compared with the results obtained in Chapter III, and a new error (bias) term which enables the "training" of the Brownian motion parameters (drift and variance) and Page's test parameters (k and a) will be presented.

In Section B, general theory about diffusion processes and the related stopping time problems is presented. The first threshold crossing time and hitting probabilities are shown to be reduced to solving 2nd order differential equations. The relation to the Page test is introduced and a new bias term which enables the comparison of the accuracy of the calculation is presented.

Section C deals with the approximation to the ARL functions of the cumsum procedure and an explicit form for the bias is calculated.

Simulation results are presented in Section D and compared to simulation results presented in Chapter III. Also, a new error (bias) term which enables the "training" of the Brownian motion parameters and Page's test parameters is introduced.

A short summary is presented in Section E.

B. GENERAL THEORY ABOUT DIFFUSION PROCESSES AND RELATED STOPPING TIME PROBLEMS

In this section general properties of diffusion processes will be presented. It will be shown that many functionals, including the first threshold crossing time and associated probabilities, boundary behavior properties and stationary distributions of cumsum procedures, can be approximated by using one-dimensional diffusions.

1. General Description and Definitions

Definition (Karlin and Taylor, 1981). A continuous time parameter stochastic process which possesses the (strong) Markov property and for which the sample paths $X(t)$ are (almost always) continuous functions of t is called a *diffusion* process.

Consider a diffusion process $\{X(t), t \geq 0\}$ whose state-space is an interval I with endpoints $l < r$. Such a process is said to be *regular* if starting from any point in the interior of I, any other point in the interior of I may be reached with non-zero probability. Henceforth, without further mention, we shall consider only regular diffusion processes.

Dynkin Condition (Karlin and Taylor, 1981): A sufficient condition for a standard process $X(t)$ to be a diffusion process is the Dynkin condition:

$$\lim_{h \downarrow 0} \frac{1}{h} \Pr\{|X(t+h) - X(t)| > \varepsilon | X(t) = x\} = 0 \qquad (4-1)$$

for all x in I. □

This relation asserts that large displacements of order exceeding a fixed ε, are very unlikely over sufficiently small time intervals. This is in fact

a formalization of the property that the sample paths of the process are continuous.

All diffusion processes are characterized by the mean and the variance of the infinitesimal increments. Let $\Delta X(t)$ be the increment in the process accrued over a time interval of length h, (i.e., $\Delta X(t) = X(t+h)-X(t)$), then

$$\lim_{h\downarrow 0}\frac{1}{h}E\{\Delta X(t)|X(t)=x\}=\mu(x,t)$$

and (4-2)

$$\lim_{h\downarrow 0}\frac{1}{h}E\{\Delta X(t)^2|X(t)=x\}=\sigma^2(x,t).$$

The functions $\mu(x,t)$ and $\sigma^2(x,t)$ are called the **drift** and **diffusion** parameters, respectively. In the time homogeneous case, the functions $\mu(x,t) = \mu(x)$ and $\sigma^2(x,t) = \sigma^2(x)$ are both independent of t.

A **Brownian motion process** (sometimes called the **Wiener** process) is a regular process on the state-space I with parameters $\mu(x) = 0$ and $\sigma^2(x) = \sigma^2$ for all x. Adding a trend μt to the Brownian motion $B(t)$ produces a Brownian motion with drift $B(t) + \mu t$. In this case, the drift parameter is μ, while the diffusion parameter remains σ^2.

The Brownian motion process $\{B(t), t \geq 0\}$ has the following properties:

- $B(0) = 0$.
- $\{B(t), t \geq 0\}$ has stationary and independent increments.
- for every $t \geq 0$, $B(t)$ is normally distributed with mean 0 and variance c^2t.

When $c = 1$, the process is called the **Standard** Brownian motion. Notice that any Brownian motion can be converted to the standard process by scaling via $B(t)/c$.

The behavior of the diffusion process $X_t = \{X(t), t \geq 0\}$ can be modeled by nonlinear stochastic differential equations of the form

$$dX_t = \mu(X_t,t)dt + \sigma(X_t, t)dB_t, \qquad t \geq 0$$

with initial condition X_0, where μ and σ are the drift and diffusion parameters as defined by (4–2) and where $\{B_t, t \geq 0\}$ is a standard Brownian motion. Thus, dB_t has the interpretation as a "white" noise driver. This notation is shorthand for the integral equation

$$X_t = X_0 + \int_0^t \mu(X_s,s)ds + \int_0^t \sigma(X_s,s)dB_s.$$

This integral representation of a diffusion process demonstrates the *Markov* property of the diffusion. That is, given X_s, for each $s > 0$ $\{X_t, t > s\}$ and $\{X_t, 0 \leq t < s\}$ are independent. This property is easy to see since for any $t \geq s \geq 0$, we can write

$$X_t = X_s + \int_s^t \mu(X_u,u)du + \int_s^t \sigma(X_u,u)dB_u.$$

This equation indicates that $\{X_t, t > s\}$ can be constructed completely from X_s and $\{B_u, t \geq u \geq s\}$. Thus, with X_s fixed, $\{X_t, t \geq s\}$ is generated independently of $\{X_t, t < s\}$ since $\{B_t - B_s, t > s\}$ is independent of all the past.

The following theorem determines the parameters of $Y(t) = g[X(t)]$, *where $X(t)$ is a regular diffusion process.*

Theorem (Karlin and Taylor, 1981): Let $\{X(t), t \geq 0\}$ be a regular diffusion process with parameters $\mu(x)$ and $\sigma^2(x)$ whose state-space is defined on $I = (l,r)$.

Let g be a strictly monotone function on I with continuous second derivative $g''(x)$ for $l < x < r$. Then $Y(t) = g[X(t)]$ defines a regular diffusion process on I with the parameters

$$\mu_Y(y) = \frac{1}{2}\sigma^2(x)g''(x) + \mu(x)\cdot g'(x)$$
$$\sigma_Y^2(y) = \sigma^2(x)\cdot[g'(x)]^2. \qquad (4-3)$$

2. Stopping Time Functionals of Diffusion Processes

In this section we analyze stopping time problems using properties of diffusion processes. It is assumed that $\{X(t), t \geq 0\}$ is a regular, time homogeneous diffusion process. Let a and b be fixed, subject to $l < b < a < r$, and let $T(z) = T_z$ be the hitting time of z defined by

$$T_z = \begin{cases} \infty & \text{if } X(t) \neq z \qquad \forall t > 0 \\ \inf\{t \geq 0; X(t) = z\} & \text{otherwise.} \end{cases}$$

We use the notation

$$T^* = T_{a,b} = T(a,b) = \min\{T(a), T(b)\}$$

to denote the first time $X(t) = a$ or $X(t) = b$. For processes starting at $X(0) = x$ in (a,b), this is the same as the *exit time* of the interval (a,b):

$$T(a,b) = \inf\{t \geq 0; X(t) \notin (a,b)\}, \quad X(0) = x \in (a,b).$$

a. Stopping Time Related Problems

This section concentrates on three problems related to the first hitting time of a diffusion which are relevant in the case of the cumsum procedure.

Problem 1. Find

$$u(x) = \Pr\{T(a) < T(b) | X(0) = x\} \quad b < x < a \tag{4-4}$$

that is, the probability that the process reaches a before b.

Problem 2. Find

$$v(x) = E\{T^* | X(0) = x\} \quad b < x < a \tag{4-5}$$

which is the mean time to reach either a *or* b.

Problem 3. For a bounded and continuous function g, find

$$w(x) = E\left\{\int_0^{T^*} g(X(s))ds | X(0) = x\right\} \quad b < x < a. \tag{4-6}$$

Since the sample paths of the diffusion processes are continuous (4–1), the integral $A = \int_0^{T^*} g(X(s))ds$ is defined. If $g(x)$ represents a cost rate incurred whenever the process is in state x, then A would be the total cost up to the time when either a or b was first reached. If $g(x) = 1$ for all x, then $A = T^*$, the time to reach a or b, so that problem 2 can be considered as a special case of problem 3.

b. Solutions of the Stopping Time Problems

A convenient reference for the solution of these three problems is Karlin and Taylor (1981, Ch. 15), where it is shown that $u(x)$, $v(x)$, and $w(x)$ possess two bounded derivatives for $b < x < a$, and that these functions satisfy the following differential equations:

Solution Equation for Problem 1

$$0 = \mu(x)\frac{du}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2u}{dx^2} \qquad \text{for} \quad b < x < a, \quad u(b) = 0, u(a) = 1. \tag{4-7}$$

Solution Equation for Problem 2

$$-1 = \mu(x)\frac{dv}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2v}{dx^2} \qquad \text{for} \quad b < x < a, \quad v(b) = v(a) = 0. \quad (4-8)$$

Solution Equation for Problem 3

$$-g(x) = \mu(x)\frac{dw}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2w}{dx^2} \qquad \text{for} \quad b < x < a, \quad w(b) = w(a) = 0. \quad (4-9)$$

In order to solve these three problems we need to use several new functions. Let

$$s(x) = \exp\left\{-\int^x \frac{2\mu(\xi)}{\sigma^2(\xi)}d\xi\right\} \quad \text{for} \quad l < x < r \qquad (4-10)$$

be the **scale density** of the process. The use of an indefinite integral will become clear later. Next, the **scale function** of the process is defined by

$$S(x) = \int^x s(\eta)d\eta \qquad (4-11)$$

and finally, the **speed density** is given by

$$m(x) = 1/\left[\sigma^2(x)\cdot s(x)\right] \quad \text{for} \quad l < x < r.$$

Using these definitions, the solution for Problem 1, namely the probability of hitting a before b is given by:

$$u(x) = \frac{S(x) - S(b)}{S(a) - S(b)} \quad b < x < a. \qquad (4-12)$$

The solution for Problem 3 is given as

$$w(x) = 2\Big\{u(x)\cdot\int_x^a [S(a)-S(\xi)]\,m(\xi)g(\xi)d\xi + [1-u(x)]\int_b^x [S(\xi)-S(b)]\,m(\xi)g(\xi)d\xi.\Big\} \qquad (4-13)$$

The solution for Problem 2 is obtained by letting $g(\xi) \equiv 1$.

Notice that the solution for $w(x)$ can also be written as:

$$w(x) = \int_b^a G(x,\xi)g(\xi)d\xi, \qquad (4-14)$$

where:

$$G(x,\xi) = \begin{cases} 2\dfrac{[S(x)-S(b)][S(a)-S(\xi)]}{S(a)-S(b)}\cdot\dfrac{1}{\sigma^2(\xi)s(\xi)} & b \le x \le \xi \le a \\ 2\dfrac{[S(a)-S(x)][S(\xi)-S(b)]}{S(a)-S(b)}\cdot\dfrac{1}{\sigma^2(\xi)s(\xi)} & b \le \xi \le x \le a. \end{cases} \qquad (4-15)$$

The function $G(x,\xi)$ is called the **Green function** of the process on the interval $[b,a]$.

Determining the mean time prior to T^* that the process spends in the interval $[\xi, \xi+\Delta)$ is equivalent to evaluating

$$w(x) = E\Big\{\int_0^{T^*} g(X(s))ds \,\Big|\, X(0) = x\Big\}$$

for

$$g(x) = \begin{cases} 1 & \xi \le x \le \xi + \Delta \\ 0 & \text{otherwise} \end{cases}$$

and following the format of (4–14), this is

$$w(x) = v(x) = E\{\Delta T \,|\, X(0) = x\} = \int_\xi^{\xi+\Delta} G(x,\eta)d\eta \qquad (4-16)$$

we see from (4–16) that $G(x,\xi)d\xi$ measures the mean time ΔT prior to T^* that the process spends in the infinitesimal interval $[\xi,\xi + d\xi]$ given by $X(0) = x$.

c. *Some Examples of Functional Calculations*

Given the solutions (4–12), (4–13) and (4–16), some cases of interest will be examined.

(1) Standard Brownian Motion. Let $\{X(t), t \geq 0\}$ be a standard Brownian motion with parameters $\mu(x) \equiv 0$, $\sigma^2(x) \equiv 1$. Then,

$$s(x) = \exp\left\{-2\int^x \frac{\mu(\xi)}{\sigma^2(\xi)}d\xi\right\} = 1.$$

The scale measure is given by

$$S(x) = x.$$

Thus, $u(x)$, the probability of hitting a prior to b, with initial state x, is

$$u(x) = \frac{x-b}{a-b} \quad b \leq x \leq a. \tag{4–17}$$

The speed density in this case is

$$m(\xi) = \frac{1}{s(\xi)} = 1$$

and the Green function (4–15) for the interval $[b,a]$ is

$$G(x,\xi) = \begin{cases} \dfrac{2(x-b)(a-\xi)}{(a-b)}, & b \leq x \leq \xi \leq a \\ \dfrac{2(\xi-b)(a-x)}{(a-b)}, & b \leq \xi \leq x \leq a. \end{cases}$$

Direct calculation from (4–14) gives

$$v(x) = E\{T_{a,b} | X(0) = x\} = \int_b^a G(x,\xi)d\xi$$

$$= (x-b)(a-x) \quad b \le x \le a. \tag{4-18}$$

Remark. A process $\{X(t)$ whose scale function is linear $S(x) = x$, is said to be of *natural* or *canonical scale* since the hitting probability (4–17) is proportional to actual distances.

Notice that the scale function can be used to rescale the state-space (l,r) in terms of probabilities of achieving various levels, and this use motivates the name. If a point x_0 is fixed as the origin, we can easily determine a new scale function by performing a translation, causing $S(x_0) = 0$ and form a process $Y(t) = S(X(t))$ on the interval $(S(l), S(r))$. Since S is strictly monotone and twice differentiable, the use of Theorem (4–3) establishes the infinitesimal parameters of the process $\{Y(t)\}$:

$$\mu_Y(y) = \frac{1}{2}\sigma^2(x)S''(x) + \mu(x)S'(x)$$

and

$$\sigma_Y^2(y) = \sigma^2(x)\cdot[S'(x)]^2 = \sigma^2(x)\, s^2(x) \quad \text{where} \quad y = S(x).$$

The scale measure for $\{Y(t)\}$ process is $S_Y(y) = y$, thus, the use of the scale function enables one to transform a process to a natural scale.

(2) Brownian Motion with Drift. If $\{X(t), t \ge 0\}$ is Brownian motion with nonzero drift $\mu(x) \equiv \mu$ and variance σ^2, then:

$$s(x) = \exp(-2\mu x/\sigma^2) \tag{4–19}$$

$$S(x) = A\, \exp(-2\mu x/\sigma^2) + B \quad (A \text{ and } B \text{ constants}),$$

and

$$u(x) = \frac{e^{-2\mu x/\sigma^2} - e^{-2\mu b/\sigma^2}}{e^{-2\mu a/\sigma^2} - e^{-2\mu b/\sigma^2}} \quad b \le x \le a. \qquad (4-20)$$

3. Instantaneous Return Processes and the Relation to Page's Cumsum Procedure

This section introduces a certain boundary behavior of the diffusion process that defines an Instantaneous Return process. This process is shown to describe any cumsum procedure and forms the basis for the approximated ARL function. It enables also the derivation of a new bias term which is used to evaluate the accuracy of the approximation.

a. Instantaneous Return Processes (Karlin and Taylor, 1981)

Consider a diffusion $\{X(t),\ t \ge 0\}$ on the state-space $I = (l,r)$ and let $l < b < a < r$. A return process $Z(t)$ relative to $[b, a]$ shown in Figure 4.1 and is defined as follows: Starting at a point x_0 in (b, a), the process is returned instantaneously to x_0 whenever b or a is reached. After such a return, the subsequent process behaves just like $X(t)$. This process is repeated at each attainment of level b or a.

The resulting process $Z(t)$ consists of recurrent cycles of random time duration $T_1, T_2, T_3, \ldots$, where T_i are independently and identically distributed, with the same distribution as $T_{a,b} = \min\{T_a, T_b\}$, the first exit time from the interval (b, a), starting from x_0 (stationary process). It follows from (4–16) that

$$E\{T_i | X(0) = x_0\} = \int_b^a G(x_0, \xi) d\xi \qquad (4-21)$$

where $G(x_0,\xi)$ is the Green function of the process $X(t)$ relative to the state-space (b,a).

Let $P(t,y)$ be the density function of $Z(t)$. Thus,

$$P(t,y)dy = \Pr\{y \le Z(t) \le y + dy | Z(0) = x_0\}.$$

Define the limiting density of $Z(t)$ as

$$\alpha(y|x) = \lim_{t\to\infty} P(t,y). \qquad (4-22)$$

To do so, consider an interval $[y_1,y_2]$ such that $b < y_1 < y_2 < a$ and define the indicator process $\{I(t),\ t \ge 0\}$ by

$$I(t) = \begin{cases} 1 & \text{if } y_1 \le Z(t) < y_2 \\ 0 & \text{otherwise.} \end{cases}$$

from Figure 4.1, we see that

$$\Pr\{I(t) = 1\} = E\{I(t)\} = \int_{y_1}^{y_2} P(t,y)dy. \qquad (4-23)$$

Recalling the renewal theorem (Ross, 1989) (Feller, 1971), we can deduce that

$$\lim_{t\to\infty} \Pr\{I(t) = 1\} = \frac{E\{\text{time spent in } (y_1,y_2) \text{ in a cycle} | Z(0) = x_0\}}{E\{\text{time duration of a cycle} | Z(0) = x_0\}}$$

$$= \frac{\int_{y_1}^{y_2} G(x_0,\xi)d\xi}{\int_b^a G(x_0,\xi)d\xi}. \qquad (4-24)$$

Using (4–23) we get

$$\lim_{t\to\infty} \Pr\{I(t)=1\} = \lim_{t\to\infty} \int_{y_1}^{y_2} p(t,y)dy$$
$$= \frac{\int_{y_1}^{y_2} G(x_0,\xi)d\xi}{\int_b^a G(x,\xi)d\xi}.$$

Since this holds for every $y_1, y_2 \in (a, b)$, it follows that

$$\alpha(y|x_0) = \lim_{t\to\infty} P(t,y)$$
$$= \frac{G(x_0,y)}{\int_b^a G(x,\xi)d\xi} \quad b < y < a. \qquad (4-25)$$

The stationary density of the instantaneous return process $\alpha(y \mid x_0)$, can be interpreted as the proportion of the mean time spent at state y in one cycle T_i.

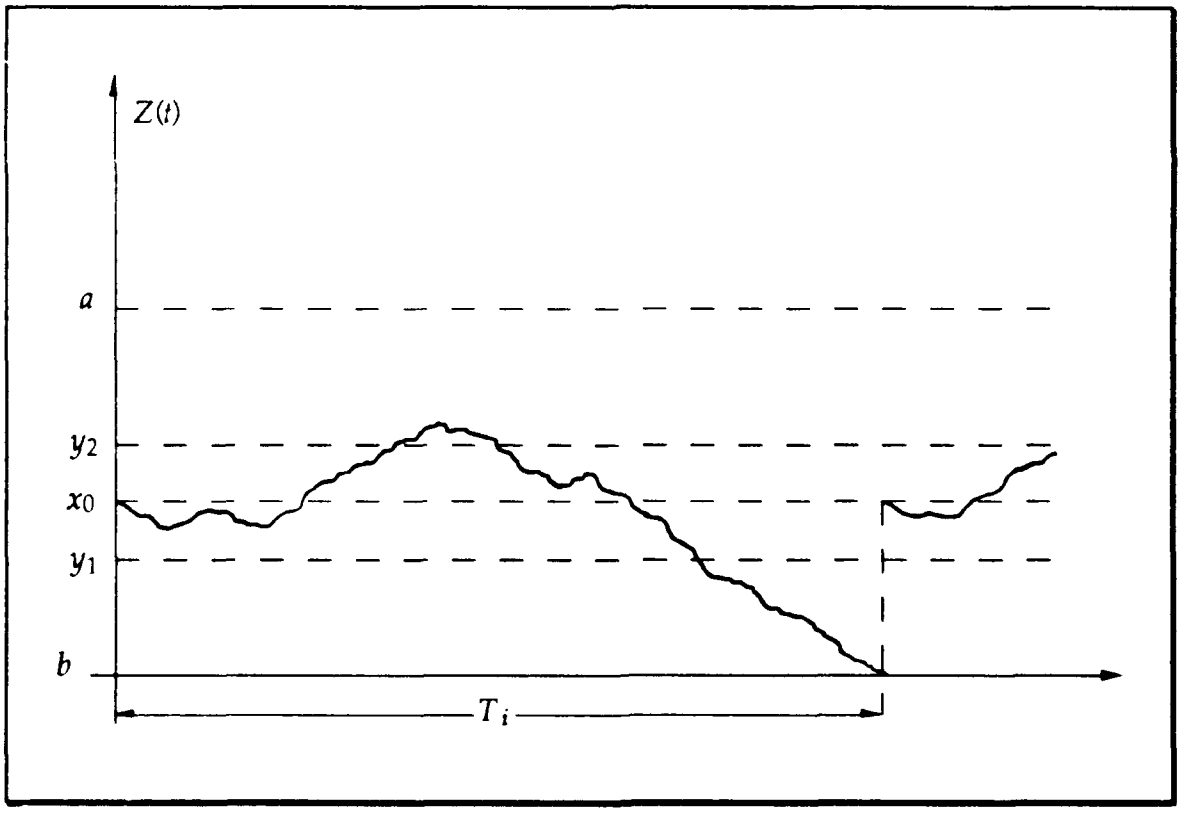


Figure 4.1. Instantaneous Return Process

b. Relation to the Cumsum Procedure

Let N denote the stopping rule based on the cumsum test with reference value k stopping boundary a and restarting boundary b,

$$N = \inf\{t: X(t) \geq a\} \qquad (4-26)$$

where

$$X(t) = (W(t) + kt) - \min_{0 \leq s \leq t} (W(s) + ks)$$

is based on the observation process

$$W(t) = \mu(t - v)^{+} + B(t), \quad \mu > 0 \qquad (4-27)$$

where $(x)^{+} = \max(0,x)$, and μ defines the amount of shift in the drift of a standard Brownian motion $B(t)$ with $B(0) = 0$, at the point of shift v. The reference value k is chosen to minimize the Delay for detection.

Before the shift occurs, the reference value guarantees that the test will hit the lower boundary and cause a restart. Each restart will force the process to return to the initial state x_0 and start once again, thus, the restart process can be considered as causing an instantaneous return process.

Notice that before the shift occurs, the process $W(t)$ is a Brownian motion with drift k, while after the shift (change) in drift occurs, $W(t)$ is a Brownian motion with drift $\mu+k$. Let L be the number of restarts before the shift, and let $\{N_i\}$ be the corresponding run length intervals of the test until the shift is detected ($i = 1, ..., L+1$). Hereby, we follow the analysis as given by Srivastava and Wu (Srivastava and Wu, 1990).

$$\sum_{i=1}^{L} N_i < v < \sum_{i=1}^{L+1} N_i.$$

The average delay time is given by

$$D(t) = E_t\left\{\sum_{i=1}^{L+1} N_i - t\right\}$$

where $E_t\{\cdot\}$ denotes the expectation when the shift occurs at a fixed time t. The asymptotic average delay time or the stationary average run length is defined as

$$\overline{\text{ARL}} = \lim_{t\to\infty} D(t). \qquad (4-28)$$

We denote $\text{ARL}_\mu(x_0)$ as the Average Run Length of the diffusion $X(t)$ when the shift in the mean is μ at the initial state $X_0 = x_0$. Similarly, $\text{ARL}_0(0) \equiv \text{ARL}_0$ denotes the ARL under no change, namely, the ARL when there is no shift in the drift and the initial state is zero. Hence, ARL_0 is the mean time between false alarms (with initial state zero).

Under our assumptions, the instantaneous return process caused by the restart process will be at some stationary state, say y, when the shift occurs. Denote the stationary density of this state y as $\alpha(y \mid x_0)$. Figure 4.2 is the appropriate picture to guide the analysis. Suppose that we use this state y as a new initial state for the detecting process with shifted mean to find $\text{ARL}_\mu(y)$. Thus, the stationary average delay time (4–28) is given by

$$\overline{\text{ARL}}_\mu(x_0) = \int \text{ARL}_\mu(y)\alpha\left(y \mid x_0\right)dy. \qquad (4-29)$$

Notice that $\overline{\text{ARL}}_\mu(x_0)$ can be interpreted in two ways. First as a weighted average of ARLs under disorder over the set of all possible initial states y taking into account the effect of the distribution of run length before the disorder. Second, time-wise, $\overline{\text{ARL}}_\mu(x_0)$ takes the weighted average of all

possible places of shift (since for each realization of $X(t)$, each different y is related to a different shift time), conditioned that the shift occurred. Since $\mathrm{ARL}_\mu(y)$ is a decreasing function of y, we obtain that

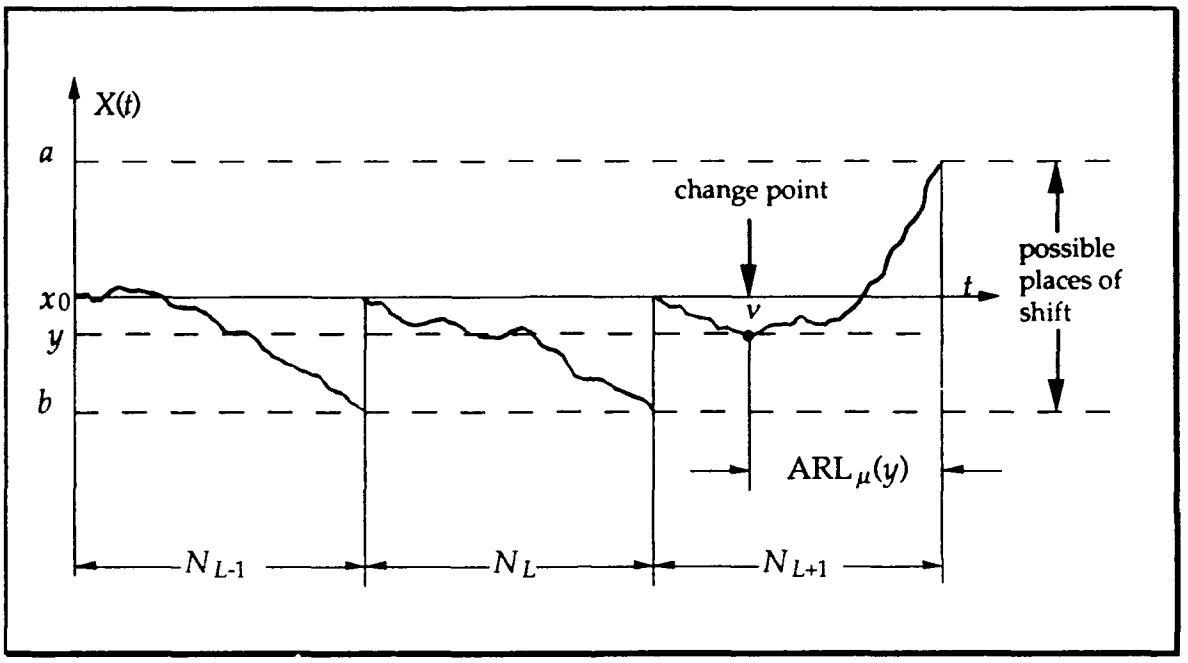


Figure 4.2. The Cumsum Process $X(t)$ as an Instantaneous Return Process

$$\overline{\mathrm{ARL}}_\mu(x_0) < \mathrm{ARL}_\mu(y).$$

Since the location of the change point v is not known inside the last run length interval N_{L+1}, the approximated ARL should take into account all the possible places of shift within the last run length interval, thus, the approximated stationary ARL under change (Delay) is obtained by (4-28) and (4-29) while the bias of the approximation can be obtained by

$$\mathrm{bias}(x_0,\mu) = \mathrm{ARL}_\mu(x_0) - \overline{\mathrm{ARL}}_\mu(x_0). \qquad (4-30)$$

Hence, (4-30) "measures" the effect of the point of shift in the limiting situation for the cumsum procedure. In the following section, we will use

the theory of this section to derive the diffusion approximation to ARL_0, $ARL_\mu(y)$ and $\overline{ARL}$ for the cumsum procedure. These approximations will be used to compare and measure the accuracy of the theoretical results derived in Chapters II and III.

C. BROWNIAN APPROXIMATIONS TO THE ARL FUNCTIONS OF THE CUMSUM PROCEDURES

The approximation to the run length functions for the one-sided Page test for an increase in the drift, will be obtained with the aid of the following two lemmas. Before presenting the lemmas, one key principle of the diffusion process which is relevant in our case needs to be addressed. This will be done in the following section.

1. The Reflection Principle (Karlin and Taylor, 1968)

A Brownian motion with a reflecting boundary at zero behaves as a standard Brownian motion in the interior of its domain $(0,\infty)$. However, when it reaches its zero boundary, then the sample path returns to the interior in a manner of that of a light wave reflection from a mirror. In general, consider $\{Z(t), t \geq 0\}$ with $Z(0) = 0$ and $Z(t) > a$ $(a > 0)$. Since $Z(t)$ is continuous and $Z(0) = 0$, there exists a random time τ at which $Z(t)$ firsts attains the value a. For $t > \tau$, we reflect $Z(t)$ about the line $z = a$ to obtain

$$X(t) = \begin{cases} Z(t) & \text{for } t < \tau \\ a - [Z(t) - a] & \text{for } t > \tau \end{cases} \qquad (4-31)$$

(see Figure (4.3). Note that $X(t) < a$ since $Z(T) > a$. Because the probability law of the path for $t < \tau$, given $X(\tau) = a$, is symmetrical with respect to the values

$x > a$ and $x < a$ and independent of the history prior to time τ, the reflection argument displays for every sample path with $Z(T) > a$, two sample paths $X(t)$ and $Z(t)$ with the same probability of occurrence.

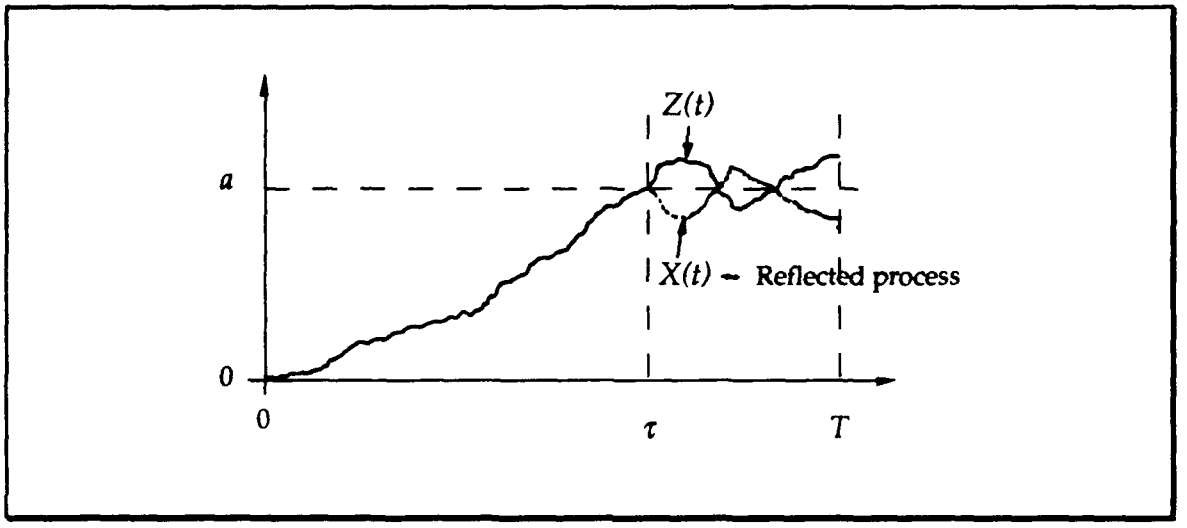


Figure 4.3. The Reflection Principle about Line *a*

The following lemma establishes the fact that the Page cumsum procedure ($X(t)$ given by (4.26)) with boundaries $(0,a)$ results in a Brownian motion with an absorbing barrier at a and a reflecting barrier at 0. The second lemma uses the fact that the reflecting barrier is at 0 to obtain the result that before the disorder, the process $X(t)$ with a reflecting barrier at 0, can be viewed as the absolute value process (set $a = 0$ in (4–31)). Thus, the reflecting boundary phenomenon is equivalent to setting $X(t) = |Z(t)|$.

Lemma 1 (Bagshaw and Johnson, 1975)

Before the shift occurs, the process $X(t)$ given by (4–26), has the same probability law as a Brownian motion $W(t)$ given by (4–27) with drift k and a reflecting barrier at 0. □

Lemma 1 applies to any diffusion type process. It states that *before* the shift occurs, $X(t) = (W(t) + kt) - \min_{0 \le s \le t}(W(s) + ks)$ and $|W(t)|$ have the same distribution. Moreover, the distribution of the first passage time of $X(t)$ to a can be determined by finding the distribution of the first passage time of a process with a reflecting barrier at zero to an absorbing barrier at a. Thus, it is clear that *after* the shift occurs, $X(t)$ and $|W(t)|$ do not have the same distribution (since a is an absorbing barrier).

Using the results of lemma 1, two alternative methods can be used to get the desired approximation for the ARL function. The following two subsections describe these methods.

2. Direct Calculation of the ARL Function via the Functional (4–8)

Let $\{X(t)\}$ be a Brownian motion on $I = [0,\infty)$ with drift μ and variance parameter σ^2, where 0 is a reflecting boundary. Let T_a be the hitting time to level $a > 0$, and set $v(x) = E\{T_a \mid X(0) = x\}$ for $0 \le x \le a$. Then, $v(x)$ is obtained by solving the differential equation (4–8) and is given by (Bagshaw and Johnson, 1975) and (Karlin and Taylor, 1981):

$$v(x) = \begin{cases} \dfrac{1}{\mu}\left[(a-x) - \dfrac{1}{2\gamma}\left(e^{-2\gamma x} - e^{-2\gamma a}\right)\right] & \mu \neq 0 \\ \dfrac{a^2 - x^2}{\sigma^2} & \mu = 0 \end{cases} \qquad (4-32)$$

where

$$\gamma = \mu / \sigma^2.$$

Recall that before the shift occurs, $X(t)$ is a Brownian motion with drift k, and since $X_0 = 0$ (Page's cumsum test), the ARL before the shift is obtained from (4–32) as follows by setting $\mu = k$

$$ARL_0 = ARL_0(0) = \begin{cases} \frac{1}{k}\left[a - \frac{1}{2\gamma}\left(1 - e^{-2\gamma a}\right)\right] & k \neq 0 \\ \frac{a^2}{\sigma^2} & k = 0 \end{cases} \qquad (4-33)$$

where

$$\gamma = k / \sigma^2.$$

After the shift, $X(t)$ is a Brownian motion with drift $\mu+k$, and since the initial state is given by $X_0 = y$ (see Figure 4.2), the ARL after the shift is given by

$$ARL_\mu(y) = \frac{1}{(\mu + k)}\left[a - y - \frac{1}{2\gamma^*}\left(e^{-2\gamma^* y} - e^{-2\gamma^* a}\right)\right] \qquad 0 \le y \le a \qquad (4-34)$$

where

$$\gamma^* = (\mu + k) / \sigma^2.$$

3. Calculation of the ARL and $\overline{ARL}$ Functions using the Green Function

Lemma 1 established the result that before the shift occurs, $X(t)$ has the same probability law as a Brownian motion with drift k and a reflecting boundary 0. The following lemma uses this result to transform the reflected Brownian motion into another diffusion process, for which we can use theory established in the last section, namely, the use of the Green function to derive the ARL function.

Lemma 2 (Karlin and Taylor, 1968)

Let $X(t)$ be a Brownian motion with reference value k (bias) as defined in (4-26). Then, before the shift occurs, $X(t)$ has the same probability law as the process $|\tilde{W}(t)|$, where $\tilde{W}(t)$ is a Brownian motion with parameters

$$\mu_{\tilde{W}}(z) = (\text{sign } z) \cdot k$$

$$\sigma^2_{\tilde{W}}(z) = \sigma^2_x(|z|) = \text{constant.}$$

for all z in the state-space I. □

Thus, the reflecting barrier phenomenon is equivalent to setting $X(t) = |\tilde{W}(t)|$, where $\tilde{W}(t)$ is a Brownian motion on $(-a, a)$ having parameters given by Lemma 2. Hence, the stopping rule (4–26) can be modified as

$$N = \inf\{t: |\tilde{W}(t)| \geq a\}$$

which is the first exit time from the interval $(-a, a)$. Thus, the reflected Brownian motion which describes Page's cumsum procedure is transformed to a nonreflected Brownian motion to which we can apply the results obtained for regular diffusions.

Recall the definition of the Green function given by (4–15). Then, for the process $X(t)$ in terms of the process $\tilde{W}(t)$, the scale density function (4–10) is given by

$$s(z) = e^{-\int^z \frac{2\mu(\xi)}{\sigma^2(\xi)} d\xi} \qquad -a \leq z \leq a.$$

$$= e^{-2|z|k/\sigma^2}$$

Since the initial value in our case (Page's Procedure) is $x = 0$, we get the Green function for the cumsum procedure (with $b = -a$) for the no change hypothesis:

$$G(0,z) = \begin{cases} 2\dfrac{\int_{-a}^{0} e^{-2|u|k/\sigma^2}du \cdot \int_{z}^{a} e^{-2|u|k/\sigma^2}du}{\left[\int_{-a}^{a} e^{-2|u|k/\sigma^2}\right]\sigma^2 \cdot e^{-2|z|k/\sigma^2}} & -a \le 0 \le z \le a \\ 2\dfrac{\int_{0}^{a} e^{-2|u|k/\sigma^2}du \cdot \int_{-a}^{z} e^{-2|u|k/\sigma^2}du}{\left[\int_{-a}^{a} e^{-2|u|k/\sigma^2}du\right]\sigma^2 \cdot e^{-2|z|k/\sigma^2}} & -a \le z \le 0 \le a \end{cases}$$

Now define $\gamma = k/\sigma^2$

$$= 2\frac{e^{2|z|\gamma}\int_{-a}^{\min(0,z)} e^{-2|u|\gamma}du \cdot \int_{\max(z,0)}^{a} e^{-2|u|\gamma}du}{\sigma^2\int_{-a}^{a} e^{-2|u|\gamma}du}, \qquad -a \le z \le a$$

$$= \frac{e^{2|z|\gamma}}{\sigma^2} \cdot \int_{|z|}^{a} e^{-2u\gamma}du.$$

$$= \frac{1 - e^{2(|z|-a)\gamma}}{2k} \qquad -a \le z \le a. \quad (4-35)$$

This result agrees with the result shown by Srivastava and Wu (Srivastava and Wu, 1990) except that by (4–35) it is assumed that the process has a general diffusion parameter σ^2. Hence, from (4–16), ARL_0 is given by

$$ARL_0 = \int_{-a}^{a} G(0,z)dz$$

and a direct calculation yields the same result as given by (4–33). Using these results, we get from (4–25) the stationary density of the process $X(t)$ defined by (4–26) is given by (y is a stationary state of the process $X(t)$)

$$\alpha(y|x=0)=\frac{G(0,y)}{\int_0^a G(0,y)dy} \qquad 0\le y\le a$$

$$=\frac{1-e^{-2\gamma(a-y)}}{a-\dfrac{1-e^{-2a\gamma}}{2\gamma}}.$$

Thus, using (4–34), $\overline{\text{ARL}}$ is obtained as

$$\overline{\text{ARL}}_\mu(0)=\int_0^a \text{ARL}_\mu(y)\cdot\alpha(y|x=0)dy$$

$$=\frac{1}{2\left[a-\dfrac{1-e^{-2a\gamma}}{2\gamma}\right]\cdot(\mu+k)\gamma^*}\int_0^a\left[1-e^{-2\gamma(a-y)}\right]\left[(a-y)2\gamma^*-e^{-2\gamma^* y}+e^{-2\gamma^* a}\right]dy$$

$$=\frac{\gamma^*\left[a^2-\dfrac{1}{2\gamma^2}\left(1-(1+2\gamma a)e^{-2\gamma a}\right)\right]-e^{-2\gamma^* a}\left[\dfrac{1-e^{-2\gamma a}}{2\gamma}+\dfrac{e^{2\gamma^* a}-1}{2\gamma^*}-\dfrac{e^{2(\gamma^*-\gamma)a}+2(\gamma^*-\gamma)a-1}{2(\gamma^*-\gamma)}\right]}{(\gamma^*/\gamma)(\mu+k)\left[e^{-2\gamma a}+2\gamma a-1\right]}$$

(4–36)

where:

$$\gamma=k/\sigma^2 \qquad \text{(before shift)}$$

$$\gamma^*=(\mu+k)/\sigma^2 \qquad \text{(after shift).}$$

Having calculated $ARL_\mu(y)$ and $\overline{ARL}_\mu(0)$ yields an analytical approximation to the bias as defined in (4–30).

D. RESULTS

Using equations (4-33) and (4-36) for calculating ARL_0 and $ARL_\mu(0)$ respectively, the error (bias) term has been calculated via equation (4-30):

$$\text{bias } (0,\mu) = ARL_\mu(0) - \overline{ARL}_\mu(0).$$

For the symmetric case $k = -\mu/2$, $\gamma = -\mu/2\sigma^2$ (before change) and $\gamma^* = \mu/2\sigma^2$ (after change). The reason for this assumption is that it has been shown (Bagshaw and Johnson, 1975) that this is the optimal reference value if the objective is to minimize the ARL_μ function.

Figure 4.4 illustrates the bias term as a function of the drift. For lower values of the reference value k the bias term is in the order of about 10 samples.

Figure 4.5 illustrates the effect of the initial value on the delay of the cumsum procedure for an initial value of $y = 5$.

Figure 4.6 illustrates the ARL function for both the delay and the mean time between false alarms, as obtained by using the Brownian motion approximations.

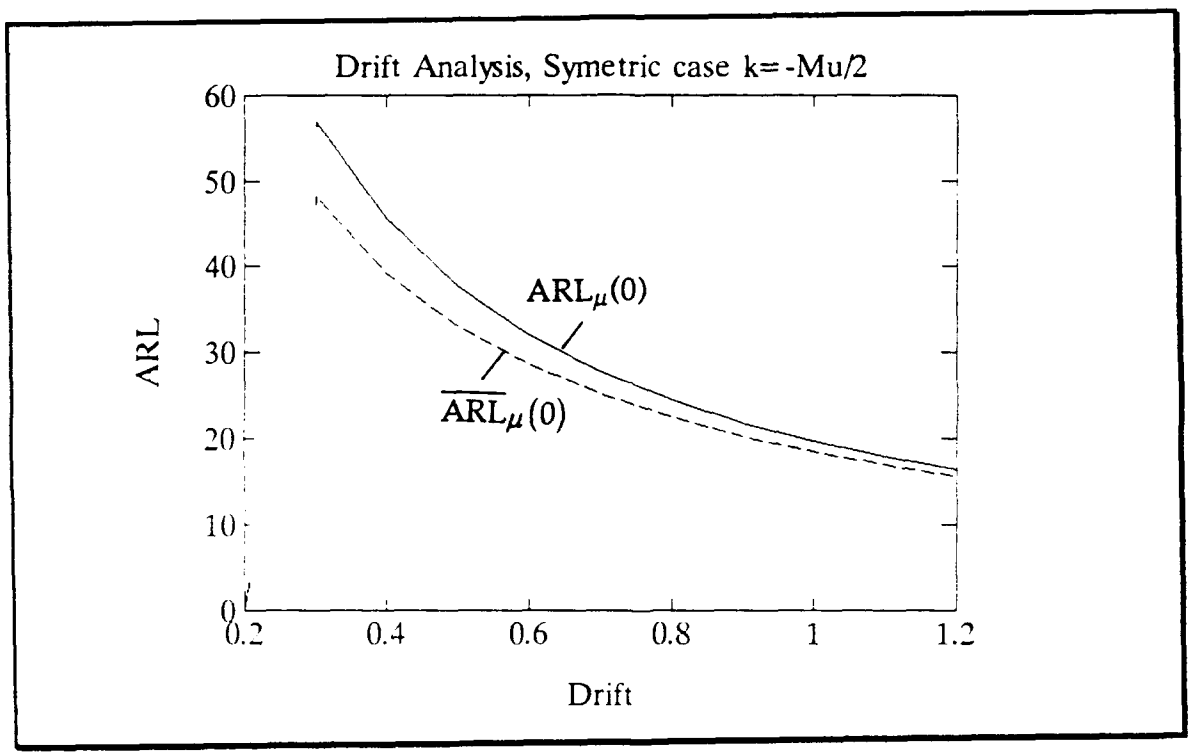


Figure 4.4. The Bias Term as a Function of the Drift. Symmetric Case $k = -\mu/2$

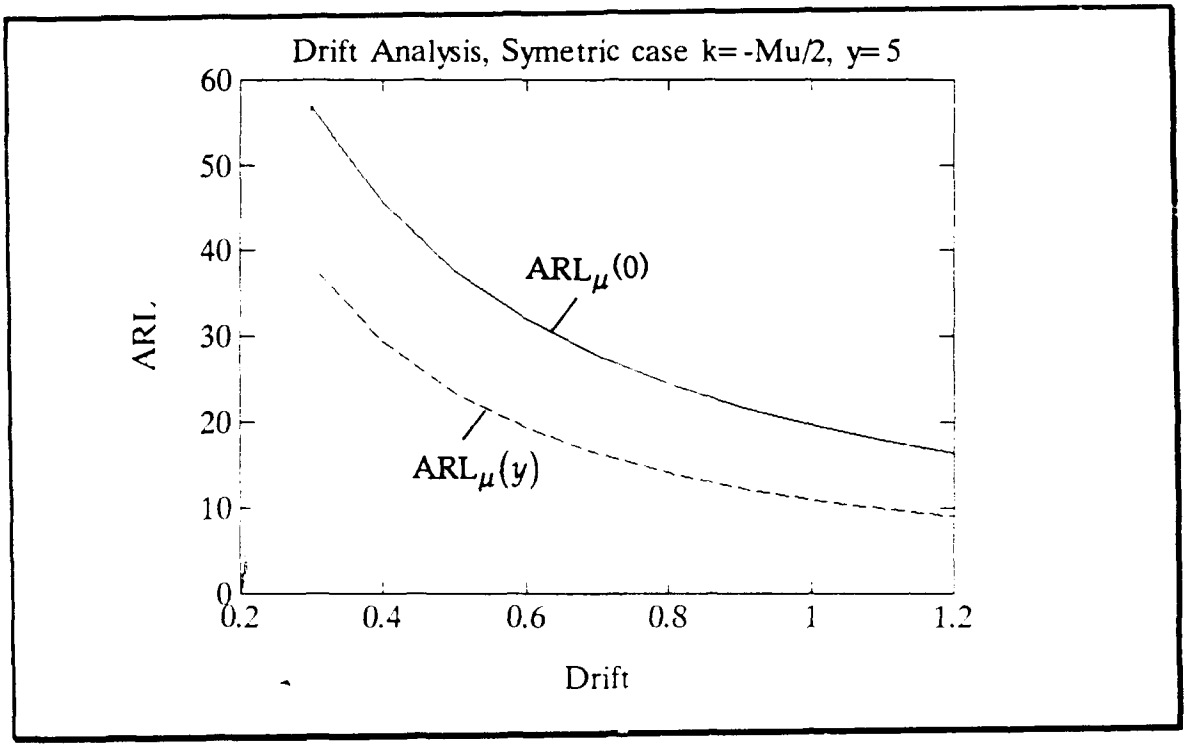


Figure 4.5. The Effect of the Initial Point y on the Delay Symmetric Case $k = -\mu/2$

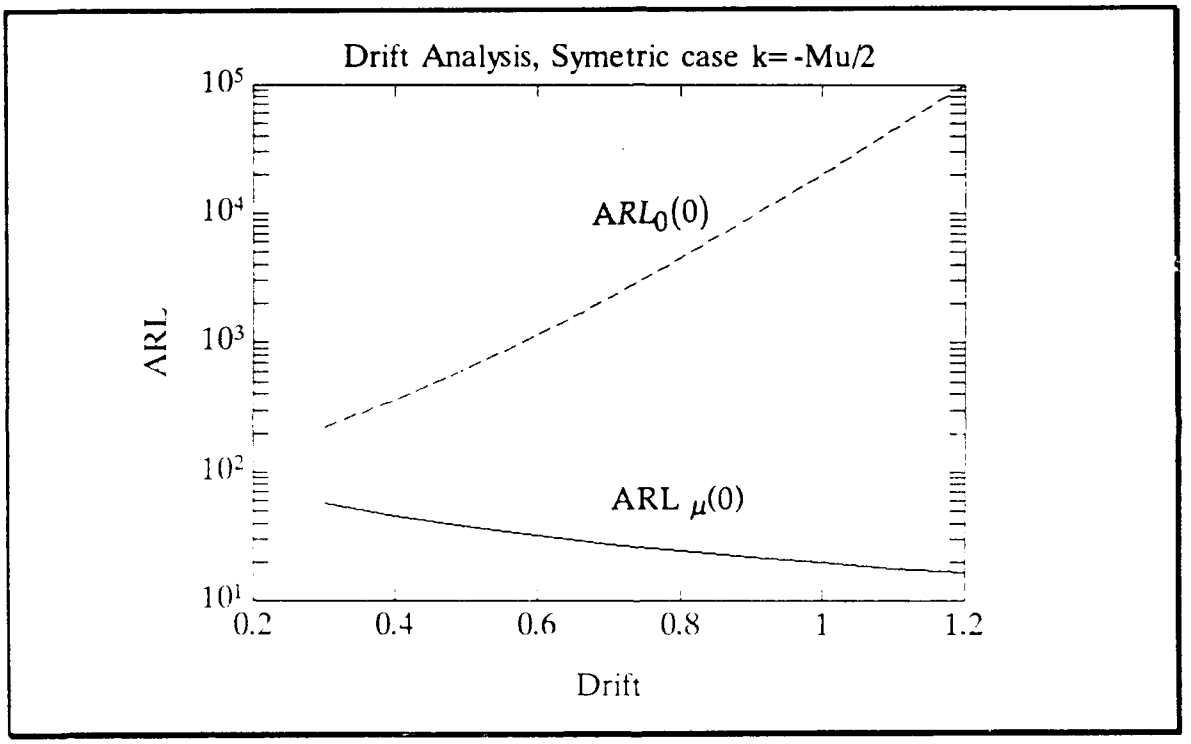


Figure 4.6. The $ARL_0(0)$ and $ARL_\mu(0)$ Functions as a Function of the Drift

E. SUMMARY

In this chapter an additional viewpoint to the analysis of cumsum procedures was introduced by using the Brownian motion approximations for stopping times. The problem of determining the probability, the average, and some general cost function of the stopping time was shown to be reduced to a closed form.

Next, the behavior of the diffusion process was investigated for two cases. In the first case, the cumsum test was shown to be modeled as a diffusion instantaneous return process which enabled the derivation of the stationary density of the diffusion, thus representing the density of the cumsum process. In the second case, the behavior of the diffusion process near a reflecting boundary was investigated and shown to be the key to determining the approximation of the ARL function for cumsum procedures. Finally, a new error (bias) term was developed allowing one to predict the average error in the delay for detection. Also, a new procedure of "tuning" the diffusion parameters to a given problem was introduced. The drift parameter was shown (as expected) to be the most influential parameter for the approximation.

V. QUICKEST DISORDER DETECTION METHODS: THE BAYESIAN FRAMEWORK

A. INTRODUCTION

Consider once again the disorder formulation of (1-2), where the observations $x_1, x_2, \ldots$ are i.i.d. random variables, such that up to a certain time $\nu > 1$, $x_1, \ldots, x_{\nu-1}$ are identically distributed with distribution $P_0(x)$, while x_ν, $X_{\nu+1}, \ldots$, are identically distributed with another distribution $P_1(x)$, where $P_0(x)$ and $P_1(x)$ do not depend on ν. In the non-Bayesian formulation the random time ν is considered as a *parameter*, and this formulation leads to classical problems of hypothesis testing. By the Bayesian approach, the parameter ν is considered as a *random variable* with a certain distribution. As in the non-Bayesian approach, we shall be concerned mainly with the problem of how to use the observations to determine as quickly as possible the time ν, or the "disorder" situation, for a given false alarm ratio. Shiryayev (Shiryayev, 1978) and Roberts (Roberts, 1966) independently proposed an approach similar to cumsum procedures. We shall refer only to Shiryayev and use his notation.

Shiryayev solved the problem of quickest disorder detection subject to a constraint on the probability of false alarms $\Pr\{N < \nu\} \leq \alpha$ for all ν (where N is the stopping time) in the Bayesian framework.

The following section gives a short presentation of his work and some important results which will be used next to establish new results.

This chapter is organized as follows: In Section B we introduce Shiryayev's results which are relevant to our case, and form the basic

underlying observation process which is used to solve the Bayes version of the cumsum procedures. Section C presents a new approach to evaluate the performance of the Bayes version for cumsum procedures. The analysis is based on the Shiryayev optimal Bayes solution (Shiryayev, 1978) and the modification of the double procedure algorithm of Assaf and Ritov (Assaf and Ritov, 1988) and uses Brownian motion approximations to solve the Bayes problem.

Finally, Section D contains a short summary of the results.

B. BAYESIAN APPROACH TO CUMSUM PROCEDURES APPROXIMATED BY BROWNIAN MOTION

1. Problem Formulation

As mentioned in the introduction, we will follow the work done by Shiryayev (Shiryayev, 1978), thereby, a new derivation of the performance of the cumsum procedure will be introduced in the Bayesian framework, using some of Shiryayev's results. The problem will be presented in terms of a Brownian motion process which approximates the cumsum behavior (see Chapter IV).

Consider a Brownian motion process $\{W(t), t \geq 0\}$, which during the time interval $[0,\nu]$ has zero drift, and during (ν,∞) has drift $\mu > 0$, where $\nu \leq \infty$ and μ are unknown parameters. The process $W(t)$ satisfies the stochastic differential equation

$$dW_t = \mu(t-\nu)^+ dt + \sigma dB_t, \quad w_0 = 0$$

where $(a)^+ = \max(0,a)$ and $B(t)$ is a standard Brownian motion with $B(0) = 0$. In other words, the structure of the observed process is

$$W_t = \begin{cases} \sigma B_t & t < \nu \\ \mu(t-\nu) + \sigma B_t & t \geq \nu \end{cases} \tag{5-1}$$

where ν is considered as the (unknown) "disorder" time in which a disorder takes place in the observed process, and the local drift shifts from zero to μ.

In what follows we assume that ν is a non negative random variable with *a priori* distribution

$$\Pr\{\nu = 0\} = p_0, \quad \Pr\{\nu \geq t | \nu > 0\} = e^{-\lambda t} \tag{5-2}$$

where p_0 and λ are known constants. Let N be the stopping variable which defines a certain class of detection rules ϕ. The class ϕ of those solution rules for which $N \in \phi$ is finite with probability one, is denoted by Δ.

For every $\phi \in \Delta$, let

$$R(\phi, p_0) = \Pr\{N < \nu\} + c \cdot E\left\{(N-\nu)^+\right\}$$

be the risk consisting of the probability of a false alarm $\Pr\{N \leq \nu\}$ and the average delay of detecting the disorder correctly, $E\{N-\nu \mid N \geq \nu\}$. The cost of one observation is assumed to be $c \geq 0$. Thus, the cost of the false alarm compared to the cost of the delay in detection is determined by the value of c. Define

$$\rho(N^*) = \inf_{\phi \in \Delta} R(\phi, p_0). \tag{5-3}$$

where N^* is the optimum stopping rule which minimizes the cost function. Hence, the problem can be slightly changed, i.e., to find among all the rules

$\phi \in \Delta$ with a given probability of false alarm $\alpha = \Pr\{N < \nu\}$, a rule $N^* \in \phi$ which guarantees the minimum of the mean time of delay, if the detection was correctly done, i.e., such that

$$D(\alpha) = \inf_{\phi \in \Delta} E\{N^* - \nu | N^* \geq \nu\} \qquad (5-4)$$

where N^* is called the **Bayes time**. Thus, the Bayes problem of quickest detection can be formulated in the following way:

For a given false alarm probability $\alpha = \Pr\{N \leq \nu\}$, find the observation method with the minimum average delay, (which minimizes the risk (5-3)). The following theorem establishes the optimum observation method in the class of decision functions $\phi \in \Delta$.

2. The Optimum Bayes Solution

Theorem (Shiryayev, 1978): For a given false alarm probability $Pr\{N < \nu\} \leq \alpha \leq 1$, the optimum observation method for the problem of minimizing the average delay as defined by (5-4) consists of observing the process

$$Z_t = \Pr\{\nu \leq t | W_s, \quad s \leq t\} \qquad (5-5)$$

with the initial condition $Z_0 = p_0 \leq 1$ and deciding that a disorder is present when a threshold $a \leq 1$ is first attained. Hence the stopping rule is given by

$$N = \inf\{t: \quad t \geq 0, \quad Z_t \geq a\}$$

where $a = 1-\alpha$. The process Z_t satisfies the following differential equation:

$$dZ_t = \lambda(1 - Z_t)dt + \frac{\mu}{\sigma} Z_t(1 - Z_t)d\bar{B}_t \qquad Z_0 = z. \qquad (5-\epsilon)$$

□

For the proof of optimality of N, see Shiryayev (Shiryayev, 1978, Ch. 4). The theorem gives two important results: *First*, the structure of optimal Bayes time N^* solutions consists of observing the current posteriori probability that the change has already occurred. The process W_t is observed until the process Z_t reaches (at time N^*) for the first time a certain level a. *Second*, Z_t is a diffusion process with time homogeneous coefficients given by

$$\begin{aligned} \mu(z) &= \lambda(1-z) \\ \sigma^2(z) &= \left[(\mu/\sigma)\cdot z(1-z)\right]^2 \end{aligned} \tag{5-7}$$

where μ and σ are the time homogeneous coefficients of the observation process (5-1). Notice that when $\lambda \to 0$, i.e., when the mean time at which the disorder occurs $E\{\nu\} = \lambda^{-1}$ tends to infinity, hence $\mu(z) = 0$ and the diffusion Z_t has a zero drift. Notice also that in this case it is natural to assume that $\alpha \to 1$. This situation indicates that the disorder appears on the background of an established stationary regime. Shiryayev solved the problem of quickest detection under this assumption. For a given mean time between false alarms T, under the optimal method of observation, the mean delay time $D(T)$ is given by

$$\begin{aligned} D(T) &= E\{N-\nu \mid N \geq \nu\} \\ &\approx \frac{1}{\gamma}\{\log(\gamma T) - 1 - C\} \end{aligned} \tag{5-8}$$

where

$$\gamma = \mu^2/2\sigma^2$$
$$C = 0.577\ldots = \text{Euler constant}$$
$$T = (1-\alpha)/\lambda$$

Notice that the assumption that the disorder is preceded by a long process of observation in which a stationary regime is established implies that $\lambda \to 0$, $\alpha \to 1$, but such that $T = (1-\alpha)/\lambda$ is fixed.

The results established in this section, in particular the diffusion type behavior of the optimal observation method, namely, the posterior probability of change, motivates a new formulation of the quickest detection of cumsum procedures and will be presented in the next section. The analysis will make use of the theory of diffusion processes established in Chapter IV.

C. THE BAYES SOLUTION TO CUMSUM PROCEDURES

The framework set by Shiryayev enables a convenient formulation of quickest detection problem for cumsum in the Bayesian framework. Hereby, there analysis of Shiryayev (Shiryayev, 1978) and Assaf and Ritov (Assaf and Ritov, 1988) is modified to obtain a new performance analysis of the optimal Bayesian stopping time solution for cumsum procedures.

1. Problem Formulation

The behavior of cumsum procedures as processes which exhibit renewal properties (Chapter II) and which can be described by instantaneous return processes (Chapter IV) establishes the observation that for a general cumsum procedure, the process of local minima (or local maxima) results in **regimes** (i.e. periods between successive local minima points) in which the diffusion approximation has a certain drift. The disorder occurs in one of the regimes, where the diffusion approximation will exhibit a change in the drift. The problem of quickest detection is concerned with the minimization of the average number of bad regimes which are mistakenly accepted during one

cycle, i.e., between two successive alarms. Notice that if the disorder occurred in the last regime in the cycle, then the average delay is given by $\overline{\overline{\mathrm{ARL}}}_{\mu}(x_0)$ as defined in (4-29). Let L be the number of regimes in one cycle. Thus, the first $L-1$ regimes are accepted, each time the present regime is accepted the test continues to the next regime, while the last one is rejected and produces the alarm. Let X_i be the set of observations within the regimes, i.e., X_1 denotes the observation set within regime 1, etc. Assume that the (true) change occurred in regime ν. Thus, $X_0, X_1, \ldots, X_{\nu-1}$ are independently distributed according to some P_0 while $x_\nu, x_{\nu+1}, \ldots$, are independently distributed to some P_1.

Assumption 1: Both P_0 and P_1 are the normal distributions with known means μ_0 and μ_1 and common variance σ^2.

Assumption 2: It is assumed that the change occurs only between regimes and not within a regime. This assumption can be justified by the fact that by using the ladder variable approach it was shown in Chapter II that the process of local minima reflects the set of time instants which are more likely to be the change points. Moreover, it was shown (2-33) that the actual number of regimes within a cycle is geometrically distributed. Following Assumption 2 we establish the last assumption.

Assumption 3: The change regime ν has a prior which is geometrically distributed with a known parameter $0 < p < 1$, i.e., $\Pr\{\nu = n\} = p \cdot q^{n-1}$ for $n \geq 1$.

Let L be a stopping time for declaring a change. Let $\alpha = \Pr\{L \leq \nu\}$ be the probability of false alarm and let $D = E\{L-\nu\}^+$ be the expected number of regimes which are mistakenly accepted in one cycle (i.e., which are mistakenly identified as regimes containing "no change" information).

Consider the following version of the optimal problem: find a stopping time L which minimizes D subject to the constraints α and $ARL_0(\cdot)$, where α is a given probability of false alarm and $ARL_0(\cdot)$ is the mean time between false alarms defined in (4-29)

2. Optimal Bayes Solution

The solution of the optimal Bayes problem is given by (5-5) and is denoted as the Z process. For any regime, Z_ℓ is the "current" posterior probability that the change has already occurred given the first ℓ regimes.

$$Z_\ell = \Pr\{\nu \le \ell | X_0, X_1, \ldots, X_\ell\} \quad \ell = 0,1,2,.... \tag{5-9}$$

Due to Shiryayev results, (5-9) is defined as the observed process which behaves like a diffusion process with time homogeneous coefficients given by

$$\begin{aligned} \mu(z) &= 0 \\ \sigma^2(z) &= [(\Delta\mu / \sigma) z(1-z)]^2 \end{aligned} \tag{5-10}$$

where $\Delta\mu = \mu_1 - \mu_0$.

Notice that the underlying model assumes that within a regime the cumsum behaves like a Brownian motion with drift parameter μ_0 or μ_1 and variance parameter σ^2. The corresponding observed process (5-9) has zero drift parameter due to the fact that it is assumed that the change does not occur within the regime. Notice that this assumption results in a natural scaled diffusion whose scale function is linear, $S(z) = z$.

The observed Z_ℓ process is defined on the state-space $I = (0,1)$. Let $0 < b < a < 1$. The optimal stopping rule is defined as follows:

- accept the present regime and move to the next one whenever $Z_\ell \le b$.

- continue sampling within the present regime as long as $b < Z_\ell < a$.
- reject the present regime and declare a disorder as soon as $Z_\ell \geq a$.

Thus, the stopping regime is given by

$$L = \inf\{\ell: Z_\ell \geq a\}.$$

The initial value of Z at the first regime is $z_0 = p$ while a decision to accept a certain regime and to move to a next one results in an initial condition

$$\bar{z}_0 = b + p(1-b). \qquad (5-11)$$

This result is due to the fact that for any regime in the cycle except the last one the test is terminated at the lower boundary, $Z_\ell = b \quad 0 \leq \ell \leq L-1$. We obtain (5–11) by using the law of total probability. All the regimes following the first one have the same probabilistic behavior. Thus, their initial z values are given by (5–11). See Figure 5.1 for a pictorial illustration.

3. Cumsum Performance Analysis

The goal of this section is to find the relationship between the test parameters a, b, and p and the delay D and the probability of false alarm α. To start the analysis we need to find $E\{L\}$, the average number of regimes within a cycle. Note that L is modified geometrically distributed since L is a mixture of two random variables, the first of which is identically zero and the second of which is geometric (Assaf and Ritov, 1988). Hence, the probability of success p should be calculated when the initial value is $\bar{z}_0$ and the probability of failure q should be calculated with initial value z_0.

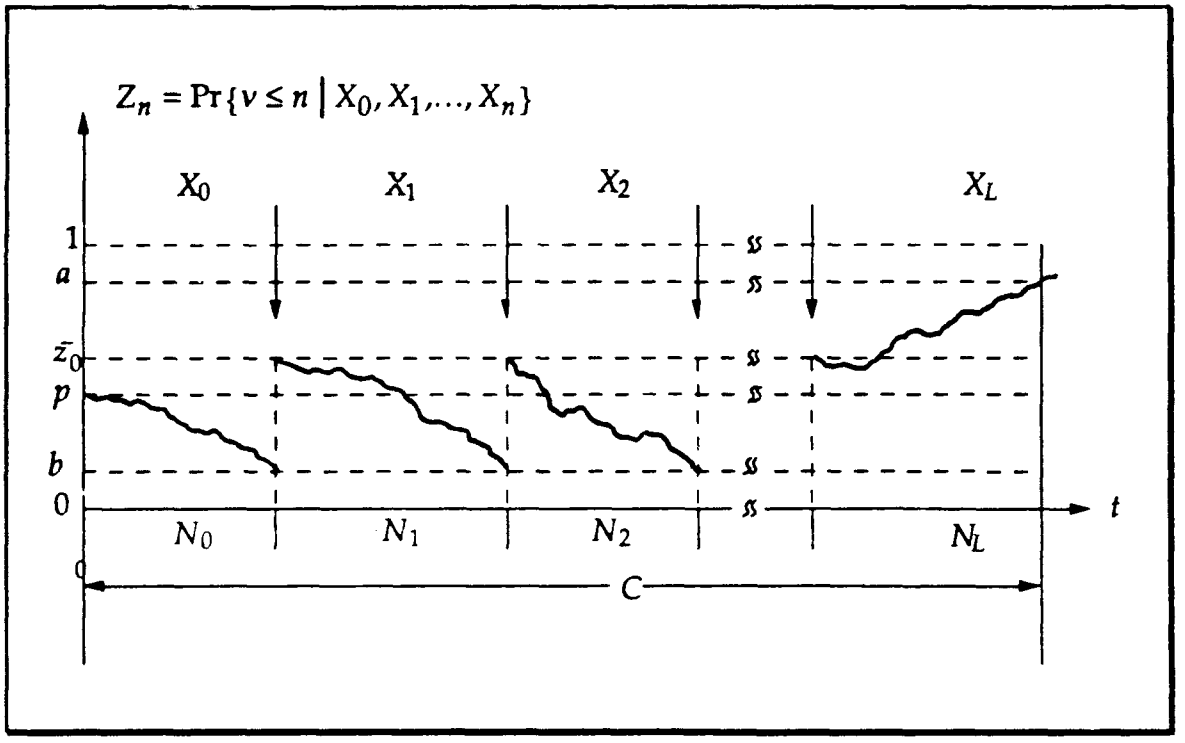


Figure 5.1. The Observed Diffusion Process $z(n)$

$$E\{L\} = \frac{q}{p}$$

$$= \frac{\Pr\{z(n) = b\}}{\Pr\{z(n) = a\}}$$

$$= \frac{\Pr\{\text{hitting } b \text{ before } a \mid \text{initial regime value } = z_0\}}{\Pr\{\text{hitting } a \text{ before } b \mid \text{initial regime value } = \bar{z}_0\}}.$$

Recall the results obtained for a general diffusion process Z_t for solving for the probability of hitting the boundary a before b as given by equation (4-4). The solution is given by (4–12).

$$u(z_0) = \Pr\{T(a) \le T(b) \mid Z(0) = z_0\} \quad b < z_0 < a,$$

hence, $E[L]$ can be obtained by

$$E\{L\} = \frac{1-u(z_0)}{u(\bar{z}_0)}. \qquad (5-12)$$

Since the observed diffusion Z_ℓ has zero drift coefficient within any regime, the Z process has a natural scale function $S(z) = z$. Thus, using (4–17) we obtain

$$u(z_0) = \frac{z_0-b}{a-b} \quad b \le z_0 \le a$$

$$u(\bar{z}_0)) = \frac{\bar{z}_0-b}{a-b} \quad b \le \bar{z}_0 \le a$$

which results in

$$E\{L\} = (a-z_0)/(\bar{z}_0-b).$$

Using $\bar{z}_0 = b + p(1-b)$, the expected number of regimes per cycle is given by

$$E\{L\} = (a-z_0)/p(1-b). \qquad (5-13)$$

Having derived an explicit form for the of average number of regimes per cycle $E\{t\}$, enables one to show the relationship between α (probability of false alarms) and D (delay) with the test parameters a, b, and p.

To compute α, notice that when the observed process $Z_\ell = \Pr\{\nu \le \ell \mid X_0, ..., X_l\}$ crosses the upper boundary a and causes an alarm, then $Z_\ell = a$ or $\Pr\{\nu \le \ell \mid X_0, ..., X_l\} = a$, thus it follows that $1-\alpha = a$, or

$$\alpha = 1-a. \qquad (5-14)$$

To compute $D = E\{L-\nu\}^+$, notice that Z_ℓ is the expected value, using posterior information of the indicator function $I_{(\nu \le \ell)}$ (Shiryayev, 1978), i.e.

$$I_{(\nu\le\ell)} = \begin{cases} 1 & \nu \le \ell \\ 0 & \nu > \ell \end{cases}$$

hence

$$\Pr\left\{I_{(\nu\le\ell)} = 1 | X_0,\ldots,X_\ell\right\} = \Pr\left\{\nu \le \ell | X_0,\ldots X_\ell\right\}$$

$$= Z_\ell.$$

Thus,

$$D = E\{L-\nu\}^+$$

$$= \sum_{\ell=1}^{L} E\left\{I_{(\nu\le\ell)} | X_0,\ldots,X_\ell\right\}$$

$$= E\left\{\sum_{\ell=1}^{L} Z_\ell\right\}. \qquad (5-15)$$

We obtain (5-15) which is consistent with Shiryayev's result, but here the derivation is done in a much simpler way. Since for $0 \le \ell \le L-1$ the Z process terminates by crossing the lower boundary, thus $Z_\ell = b$ for $\ell = 0, \ldots, L-1$. Hence,

$$D = E\left\{\sum_{\ell=1}^{L} Z_\ell\right\}$$

$$= E\left\{E\left\{\sum_{\ell=1}^{L} Z_\ell | L\right\}\right\}$$

$$= E\{L \cdot E\{Z_\ell\}\}$$

$$= E\{L\} \cdot b$$

$$= (a - z_0)b \,/\, p(1-b). \qquad (5-16)$$

4. Asymptotic Analysis

In this section we are concerned with the asymptotic analysis of the delay D as $p \to 0$. Since $1/p$ determines the average rate of changes, this asymptotic analysis will indicate the performance of the optimal algorithm when the rate of changes is small. Hereby we shall consider the constrained version of minimizing D subject to given values of $\Pr\{\nu \le L\} = \alpha$ and $\mathrm{ARL}_0(\bar{z}_0) = T$, the regime time.

The analysis starts with computation of the average cycle time $E\{C\}$ which is needed to analyze the asymptotic average delay. The diffusion type behavior of the observed process Z_ℓ enables the use of techniques introduced in Chapter IV to obtain the result for $E\{C\}$. Finally, we obtain an asymptotic approximation for the average delay.

a. Calculation of the Mean Cycle Time E{C}

Figure 5.1 is the appropriate picture to guide the following analysis. To compute the expected cycle time, $E\{C\}$, notice that the first regime

run length in a cycle starts with initial condition z_0, while all the following regimes start with initial condition $\bar{z}_0$ given by (5–11). Let N_1 be the sampling time of the first regime in a cycle, then $E\{N_1\}$ is the expected time it takes the diffusion to reach b or a. Similarly, let $\overline{N}$ be the time needed for the diffusion starting at $\bar{z}_0$ to reach b or a. The total run length of a cycle is given by $C = \sum_{i=1}^{L} N_i$ where L is the stopping regime and with $N_1, N_2, \ldots$, independent and identically distributed like $\overline{N}$. Applying Wald's equation and using (5-13) we obtain

$$E\{C\} = E\{N_0\} + (E\{L\} - 1) \cdot E\{\overline{N}\}$$

$$= E\{N_0\} + [(a - z_0)/p(1-b) - 1]E\{\overline{N}\}. \qquad (5-17)$$

To make the computation simpler we consider the long run situation, using the simplification $z_0 = \bar{z}_0$. In this case (5–17) becomes

$$E\{C\} = E\{L\} \cdot E\{\overline{N}\}$$

$$= [(a - z_0)/p(1-b)] \cdot E\{\overline{N}\}$$

$$= [(a - z_0)/p(1-b)] \cdot \mathrm{ARL}_0(\bar{z}_0). \qquad (5-18)$$

The last result is due to the fact that $E\{\overline{N}\}$ is the average regime time which is by definition equal to $\mathrm{ARL}_0(\bar{z}_0)$ since within the regime the drift coefficient is zero.

b. Calculation of $ARL_0(z)$

The observed diffusion Z_t is in natural scale since the scale measure is linear (see 4–17), i.e.,

$$S(z) = z.$$

Using the expression of the variance coefficient (5–7) we obtain from (4–15) the Green function for the observed diffusion Z_t

$$G(z,\xi)=\begin{cases}2\dfrac{(z-b)(a-\xi)}{(\Delta\mu/\sigma)^2(a-b)\xi^2(1-\xi)^2} & b\le z\le\xi\le a\\ 2\dfrac{(\xi-b)(a-z)}{(\Delta\mu/\sigma)^2(a-b)\xi^2(1-\xi)^2} & b\le\xi\le z\le a.\end{cases}$$

The Average Run Length $ARL_0(z)$ is given (Assaf and Ritov, 1988),

$$ARL_0(z)=E\{\overline{N}\}$$

$$=\int_b^a G(z,\xi)d\xi$$

$$=\frac{1}{(\Delta\mu/2\sigma)^2(a-b)}\left[(z-b)(2a-1)\log\left\{\frac{a(1-z)}{(1-a)z}\right\}+(a-z)(1-2b)\log\left\{\frac{z(1-b)}{b(1-z)}\right\}\right]$$

(5–19)

For the limiting situation

$$\lim_{p\to 0}\bar{z}_0=\lim_{p\to 0}\{b+p(1-b)\}\to b$$

it follows that in this situation $ARL_0(b)\to 0$ (as anticipated). Thus, it follows that for the constraint $ARL_0(z)=T$ to be satisfied, we need $b\to 0$ resulting in 0/0 situation.

$$\lim_{p\to 0} \mathrm{ARL}_0(\bar{z}_0) = \lim_{\substack{p\to 0 \\ b\to 0}} \frac{1}{(\Delta\mu/2\sigma)^2(a-b)}\left[(a-b-p(1-b))(1-2b)\log\left\{\frac{(b+p(1-b))(1-b)}{b(1-b-p(1-b))}\right\}\right]$$

$$= \lim_{\substack{p\to 0 \\ b\to 0}} \frac{1}{(\Delta\mu/2\sigma)^2 a}\left[a\,\log(1+p/b)\right]$$

$$\approx \frac{1}{(\Delta\mu/2\sigma)^2}\log(1+p/b). \qquad (5-20)$$

Notice that $\mathrm{ARL}_0(\bar{z}_0)$ approaches in the limit to a finite value.

c. *Asymptotic Delay*

For the limiting situation we also obtain the following approximations:

$$\lim_{\substack{\bar{z}\to 0 \\ b\to 0}} E\{L\} = \lim_{\substack{\bar{z}\to 0 \\ b\to 0}} (1-\bar{z}_0)/p(1-b)$$

$$= a/p$$

$$= (1-\alpha)/p \qquad (5-21)$$

and

$$\lim_{\substack{\bar{z}\to 0 \\ b\to 0}} E\{C\} = (1-\alpha)\cdot \mathrm{ARL}_0(0)/p \qquad (5-22)$$

and for the constraint $\mathrm{ARL}_0(0) = T$ we need

$$p/b = e^{T(\Delta\mu)^2/2\sigma^2} - 1$$

$$\approx e^{T(\Delta\mu)^2/2\sigma^2}. \qquad (5-23)$$

Substituting (5–23) in equation (5–16) for D, we obtain

$$\lim_{z_0 \to 0} D = ab / p$$

$$= (1-\alpha)b / p$$

$$= (1-\alpha) / \left(e^{T(\Delta\mu)^2 / 2\sigma^2} \right). \qquad (5-24)$$

Hence, the asymptotic average delay is given in terms of the constraints α and T and the signal parameters $\Delta\mu = \mu_1 - \mu_0$ and σ^2.

Since the ratio $\Delta\mu / \sigma$ can describe a measure for signal to noise ratio, the average delay (5-24) can also be described as

$$D \cong (1-\alpha) / \left(e^{(T/2)(\mathrm{SNR})^2} \right). \qquad (5-25)$$

Notice that in the limiting situation the average delay does not depend on p Once again, as for Shiryayev's result (5-8), as $p \to 0$, $\alpha \to 1$, and the delay D approaches the limit to a finite value.

D. SUMMARY

The fact the cumsum procedure can be viewed as a process of local minima (or respectively, maxima) enabled the use of the Brownian motion approixmation to the optimal observation process Z_ℓ. With the aid of these tools, $\mathrm{ARL}_0(\cdot)$ given by equation (5-20) and the asymptotic delay (5-25) were derived and shown to reach finite values.

VI. DETECTION-ESTIMATION ALGORITHM FOR NOISY DATA WITH ABRUPT CHANGES (DISCONTINUITIES) MODELED BY THE PIECEWISE STATE-SPACE MODEL

A. INTRODUCTION

Until now, all of the chapters dealt with problems of disorder as defined for Types 1, 2, and 3 (see Chapter 1). In this chapter we present a Type 4 problem, namely, an initial condition disruption problem. The use of state-space models as descriptive models for the initial condition disruption allows the *joint* estimation of the change time ν and the state-space parameter representing the observed signal. This methods seems to be efficient compared to GLR methods for certain classes of problems since the Kalman filter gains and covariance matrix can be computed off-line if the state-space matrices do not change in time. However, this is not the case for AR or ARMA modeling in the state-space format.

The problem of detection-estimation or detection-smoothing of signals with time-varying statistical characteristics is of great interest in many areas of signal processing. In many cases, prior knowledge of the signal characteristics can be used to model (using model-based techniques) the non-stationary behavior. In this section, the statistical changes are modeled by piecewise deterministic state-space equations with random initial conditions, and measurements corrupted by additive Gaussian white noise (Cristi, 1988). A particularly interesting class is the case of signals representable by Auto-Regressive models with piecewise constant coefficients (Andre-Obrecht, 1988).

Also, the class of PSK (Phase Shift Keying) signals enters this category, where the phase of the sinusoidal carrier is shifted according to the information (Point, 1987). For the PSK the phase shift of the sinusoidal carrier can be modeled by change of initial condition of a state-space model that describes the sinusoid. The goal is a non-coherent detection scheme that will recover the piecewise constant phase.

For such classes of signal models, we can approach the combined detection-estimation problem as a combination of: a) detection of the transition points, in order to **segment** the data field into compact regions having similar characteristics (for example constant phase in the PSK signal), and b) **filtering** within the regions to reconstruct the original signal. In the estimation framework the *joint* estimation of the change time and the model parameters can be achieved.

Previous works (Cristi, 1990) used techniques based upon the combination of Markov Random Fields (MRF) models, with Recursive Least Squares (RLS) algorithms in order to estimate the model parameters within the regions for 1D or 2D fields. Another approach (Point, 1987) used Kalman filtering techniques in order to estimate the change instants in PSK signals. Hereby we present a new technique based on Kalman filtering techniques, which calculates the joint distribution of the measurements and the change process (defined as the transition process) over a finite length window. The approach presented in this section is based upon a Maximum a-Posteriori (**MAP**) framework (Cristi and Aviv, 1991). The signal of interest is described by a piece wise state-space modeling, with initial conditions set at the beginning of each interval. By applying the Kalman

filtering technique, the two tasks of segmentation and filtering over the segmented regions can be achieved. The method leads to hypothesis testing. Moreover, if the state-space matrices do not change in time, the algorithm can be implemented at a low computation cost.

This chapter is organized as follows: Section B presents the model and the assumptions used to describe the prior needed for the algorithm. Section C presents the algorithm derivation, and finally, Section D presents simulation results. A short summary is given in Section E.

B. MODEL DESCRIPTION

1. Problem Statement

Consider the state-space model

$$\mathbf{x}(n+1) = \begin{cases} \mathbf{A}\mathbf{x}(n) + \mathbf{B}\mathbf{v}(n) & \text{no change} \\ \mathbf{x}_{IC}(n) & \text{change in initial conditions.} \end{cases}$$

$$y(n) = \mathbf{c}\,\mathbf{x}(n) + w(n) \qquad (6-1)$$

where $\mathbf{x}_{I.C}(n)$ is the initial condition vector at instant n, and $\mathbf{A},\mathbf{B}$ are known matrices, $\mathbf{c}$ is a known vector, $\mathbf{v}(n)$ and $w(n)$ are i.i.d. white Gaussian drivers with zero mean and known covariance $\mathbf{Q}$ and σ^2.

$$\mathbf{v} \sim N(0,\mathbf{Q})$$

$$w \sim N(0,\sigma^2).$$

Notice the doubly stochastic nature of the process $\{\mathbf{x}\}$. In this respect the process can be described as a combination of two models: one, modeling the regions corresponding to the initial condition, and one for state-space model itself.

Let $\gamma = \{\gamma(n), n > 0\}$, $\gamma(n) \in \{0,1\}$ be defined as the process of *transitions*, i.e.,

$$\gamma(n) = \begin{cases} 0 & \text{if } \mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{v}(n) \\ 1 & \text{if } \mathbf{x}(n+1) = \mathbf{x}_{IC} \end{cases} .$$

2. Model Assumptions

Assumptions on $\mathbf{x}_{IC}(n)$ and $\gamma(n)$ are needed.

(1) Assumptions on $\mathbf{x}_{IC}(n)$ are independence

$$P(\mathbf{x}_{IC}(n) \mid \mathbf{x}_{IC}(n-1) \ldots \mathbf{x}_{IC}(0)) = P(\mathbf{x}_{IC}(n))$$

and that $P(\mathbf{x}_{IC}(n))$ is Gaussian with known mean and variance.

$$P(\mathbf{x}_{IC}(n)) \approx N(\bar{\mathbf{x}}_{-1}, \mathbf{P}_{-1}),$$

where $\bar{\mathbf{x}}_{-1}$ is the a priori mean and $\mathbf{P}_{-1}$ is the covariance matrix of the vector $\mathbf{x}_{IC}$.

(2) Assumptions on the transition process $\gamma(n)$ are

- There exists an integer d such that at most only *one* transition occurs in the process γ during *any* interval $[t-d,t]$.
- The process γ is assumed to be d–Markov, in the sense that

$$P(\gamma(t) \mid \gamma(t-1) \ldots \gamma(0)) = P(\gamma(t) \mid \gamma(t-1) \ldots \gamma(t-d))$$

for all $t > d$, which implies that the statistics of γ are known from the last d samples.

3. Probabilistic Model for the Transition Process $\gamma(n)$

In order to assign a probability measure to γ, define the following "truncated" sequence:

$$\gamma_{t,d+1} = [\gamma(t-d) \ldots \gamma(t)]$$

where the pair $(t,d+1)$ defines the truncation boundaries, t defines the time index of the starting element, and $d+1$ defines the sequence length. In a similar way we can define any other truncated sequence.

The vector of possible realizations of γ is defined by

$$\boldsymbol{\varepsilon}_j^T = [\varepsilon_j(0) \ldots \varepsilon_j(d)]^T \in \{0,1\}^{d+1} \quad j = -1, \ldots, d$$

where

$$\varepsilon_j(i) = \begin{cases} 1 & \text{if } i = j \quad i = 0,\ldots,d \\ 0 & \text{if } i \neq j \quad j = 1,\ldots,d \end{cases}$$

and

$$\boldsymbol{\varepsilon}_{-1} = 0.$$

The possible $d+2$ realizations of the vector $\boldsymbol{\varepsilon}_j$ are of the form that at most only a single "1" can be present at any i location $(0 \leq i \leq d)$ corresponding to a change at location i. Thus, $\boldsymbol{\varepsilon}_s$ results in a change inside the window $\boldsymbol{\gamma}_{t,d+1}$ at location $(t\text{-}s)$, while $\boldsymbol{\varepsilon}_{-1}$ is by definition the no change vector.

From the assumption of γ being a Markov process with realizations $\boldsymbol{\varepsilon}_j$, we can determine its probabilistic model as

$$P(\chi(t) = 0 \,|\, \boldsymbol{\gamma}_{t-1,d+1}) = \begin{cases} 1 & \text{if } \boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_j \quad j \neq -1 \\ P_0 & \text{if } \boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_{-1} \end{cases} \tag{6-2}$$

The reason behind this equation is the fact that $\chi(t) = 0$ with probability 1 if a transition exists in the interval $[t-d-1, t-1]$ which defines the previous "sliding window." If there was no change in the previous window $(\boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_{-1})$, then $\chi(t) = 0$ with probability P_0, thus $\chi(t) = 1$ with probability $P_1 = 1-P_0$. Figure 6–1 shows realizations of the process γ.

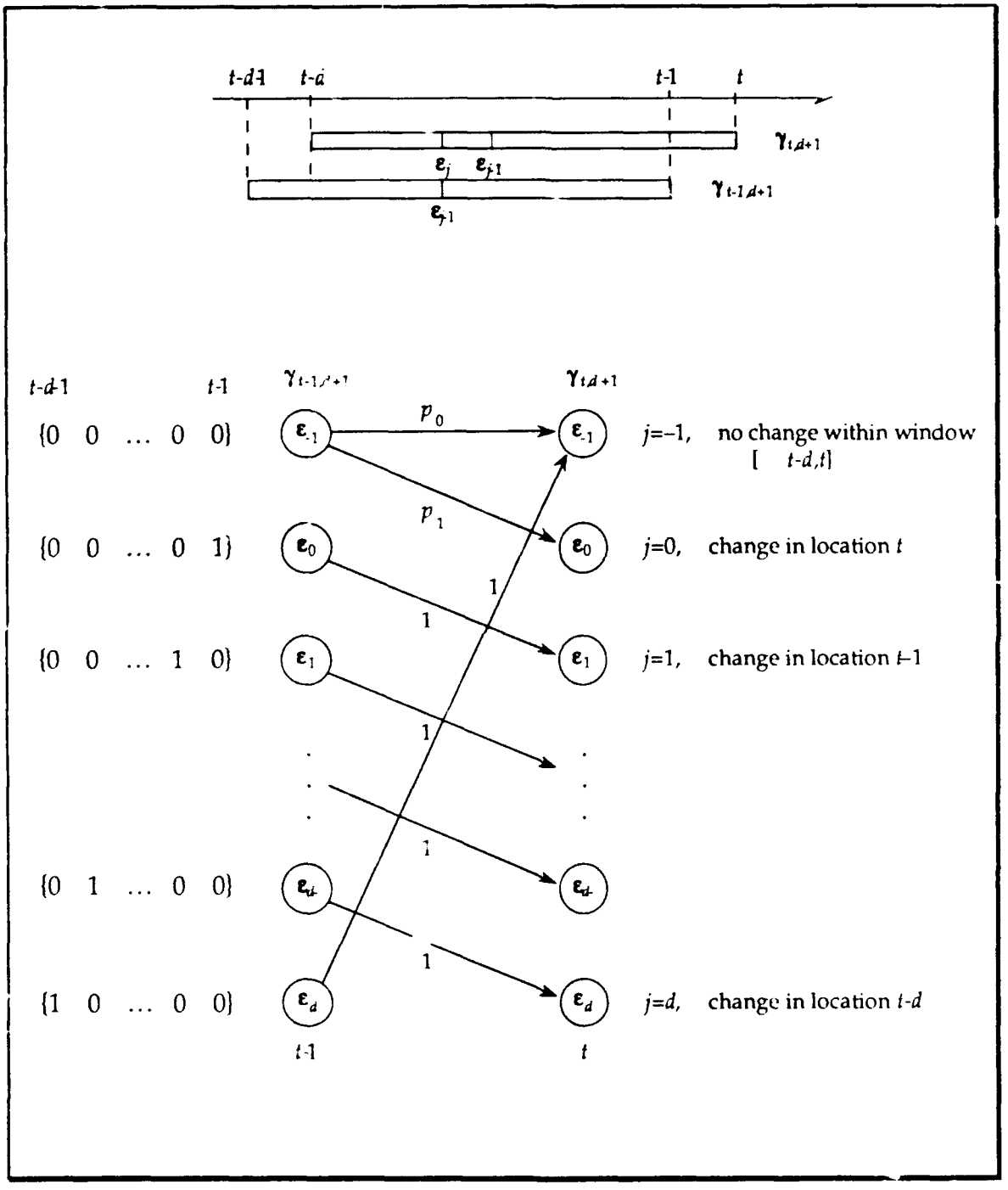


Figure 6.1. Realization Map of the Process $\{\gamma\}$

C. THE DETECTION-ESTIMATION RECURSIVE ALGORITHM

In this section the detection-estimation algorithm will be presented based on a Maximum a Posteriori (MAP) probability approach in order to extract the transition process $\{\gamma\}$ from the observations $\{y\}$ and estimate the process $\{x\}$. Notice that the transition process $\{\gamma\}$ defines the regions of the same probabilistic nature (constant phase in the case of PSK signals), resulting in the segmentation task. Within this framework, the transition points t are indicated by the process $\{\gamma\}$ and they correspond to $\gamma(t) = 1$. The algorithm presented here is based on a "sliding window" of length $d+1$ over which the likelihood of the transitions is recursively computed using the following lemma 1.

1. Basic Lemma

Lemma 1. Define $\mathbf{Y}_t = [y(0) ..., y(t)]$

$\gamma_t = [\gamma(0) ..., \gamma(t)]$

$\mathbf{Y}_{t,d} = [y(t-d-1) ..., y(t)]$

and similarly:

$\mathbf{Y}_{t,d+1} = [y(t-d) ..., y(t)]$

$\gamma_{t,d+1} = [\gamma(t-d) ..., \gamma(t)]$,

$\mathbf{Y}_{t-1,d+1} = [y(t-d-1) ..., y(t-1)]$

$\mathbf{Y}_{t-1,d} = [y(t-d) ..., y(t-1)]$

then

$$P\left(\mathbf{Y}_{t,d+1},\boldsymbol{\gamma}_{t,d+1}\middle|\mathbf{Y}_{t-d-1},\hat{\boldsymbol{\gamma}}_{t-d-1}\right)=$$

$$=\frac{P\left(y(t)\middle|\mathbf{Y}_{t-1},\boldsymbol{\gamma}_{t,d+1}\right)\cdot P\left(\gamma(t)\middle|\boldsymbol{\gamma}_{t-1,d+1},\mathbf{Y}_{t-d-1}\right)}{P\left(y(t-d-1)\middle|\mathbf{Y}_{t-d-2},\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\cdot P\left(\hat{\gamma}(t-d-1)\middle|\hat{\boldsymbol{\gamma}}_{t-d-2},\mathbf{Y}_{t-d-1}\right)}\times$$

$$P\left(\mathbf{Y}_{t-1,d+1},\boldsymbol{\gamma}_{t-1,d+1}\middle|\mathbf{Y}_{t-d-2},\hat{\boldsymbol{\gamma}}_{t-d-2}\right) \qquad (6-3)$$

Proof: By Bayesian factorization:

$$P\left(\mathbf{Y}_{t,d+1},\boldsymbol{\gamma}_{t,d+1}\middle|\mathbf{Y}_{t-d-1},\hat{\boldsymbol{\gamma}}_{t-d-1}\right)=$$

$$=P\left(\mathbf{Y}_{t,d+1}\middle|\ \boldsymbol{\gamma}_{t,d+1};\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\cdot P\left(\boldsymbol{\gamma}_{t,d+1}\middle|\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right) \qquad (6-4)$$

The left-most probability term can be further factored:

$$P\left(\mathbf{Y}_{t,d+1}\middle|\ \boldsymbol{\gamma}_{t,d+1};\mathbf{Y}_{t-d-1},\hat{\boldsymbol{\gamma}}_{t-d-1}\right)=$$

$$=\frac{P\left(y(t)\middle|\mathbf{Y}_{t-1};\boldsymbol{\gamma}_{t,d+1},\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\cdot P\left(\mathbf{Y}_{t-1};\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}{P\left(\boldsymbol{\gamma}_{t,d+1};\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}=$$

$$=P\left(y(t)\middle|\mathbf{Y}_{t-1},\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\times$$

$$\times\frac{P\left(\mathbf{Y}_{t-1,d+1}\middle|\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\cdot P\left(\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}{P\left(y(t-d-1)\middle|\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)\cdot P\left(\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t,d+1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}$$

because of the Markov property of the $\gamma(t)$ process, it is clear that in the conditioned probability terms, once $\boldsymbol{\gamma}_{t,d+1}$ is known, then, $\hat{\boldsymbol{\gamma}}_{t-d-1}$ is redundant. Furthermore, since $\boldsymbol{\gamma}_{t-1,d+1}$ is independent of $\gamma(t)$, the conditioned probability term in the numerator becomes:

$$P\left(\mathbf{Y}_{t-1,d+1}\middle|\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t-1,d}\right)\cdot$$

Notice also that in the denominator $y(t-d-1)$ is independent of $\gamma_{t,d+1}$, hence, the left-most expression in (6–4) becomes:

$$\frac{P\left(y(t)\middle|\mathbf{Y}_{t-1},\boldsymbol{\gamma}_{t,d+1}\right)}{P\left(y(t-d-1)\middle|\mathbf{Y}_{t-d-2};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}\cdot P\left(\mathbf{Y}_{t-1,d+1}\middle|\mathbf{Y}_{t-d-2};\boldsymbol{\gamma}_{t-1,d}\right). \qquad (6-5)$$

The right-most probability term in (6–4) can be also further factored:

$$P\left(\boldsymbol{\gamma}_{t,d+1}\middle|\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)=$$

$$=\frac{P\left(\gamma(t)\middle|\boldsymbol{\gamma}_{t-1,d};\hat{\boldsymbol{\gamma}}_{t-d-1};\mathbf{Y}_{t-d-1}\right)\cdot P\left(\boldsymbol{\gamma}_{t-1,d};\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}{P\left(\mathbf{Y}_{t-d-1};\hat{\boldsymbol{\gamma}}_{t-d-1}\right)}$$

$$=P\left(\gamma(t)\middle|\boldsymbol{\gamma}_{t-1,d};\hat{\boldsymbol{\gamma}}_{t-d-1};\mathbf{Y}_{t-d-1}\right)\cdot P\left(\boldsymbol{\gamma}_{t-1,d};\hat{\gamma}(t-d-1)\middle|\hat{\boldsymbol{\gamma}}_{t-d-2};\mathbf{Y}_{t-d-1}\right)\times$$

$$\times\frac{P\left(\hat{\boldsymbol{\gamma}}_{t-d-2};\mathbf{Y}_{t-d-1}\right)}{P\left(\hat{\boldsymbol{\gamma}}_{t-d-1};\mathbf{Y}_{t-d-1}\right)} \qquad (6-6)$$

$$=\frac{P\left(\gamma(t)\middle|\boldsymbol{\gamma}_{t-1,d};\hat{\boldsymbol{\gamma}}_{t-d-1},\mathbf{Y}_{t-d-1}\right)}{P\left(\hat{\gamma}(t-d-1)\middle|\hat{\boldsymbol{\gamma}}_{t-d-2};\mathbf{Y}_{t-d-1}\right)}\cdot P\left(\boldsymbol{\gamma}_{t-1,d};\hat{\gamma}(t-d-1)\middle|\hat{\boldsymbol{\gamma}}_{t-d-2};\mathbf{Y}_{t-d-1}\right)$$

Therefore, inserting (6–5) and (6–6) into (6–4) yields the desired recursion (6-3). □

By the recursion (6–3) we can update the statistics of the processes $\{y\}$ and $\{\gamma\}$ over the interval $[t-d,t]$ conditioned on past values of these processes.

2. Likelihood Function of the Transition Process

Using the probability model of the last section, the transition process $\{\gamma\}$ can be estimated at time $(t-d)$, i.e., the edge of the window, on the basis of the observations up to time t, (i.e., using the data within the window $[t-d,t]$), by using the Kalman filter technique.

The rationale behind this can be explained as follows: Suppose there was no (true) change in the signal's model. If the initial condition of the filter

is changed at any time instant within the interval $[t-d,t]$, then there will be a probabilistic mismatch between the true signal and the estimated one (See Figure 6–2a). Suppose now that the initial condition is changed at the same time instant the true change occurred, then, by forcing the change to be evaluated at the edge of the sliding window, namely at time $t-d$, will create a probabilistic match between the true signal and the predicted signal that relies on the maximum number of available observations (d) (See Figure 6-2b.).

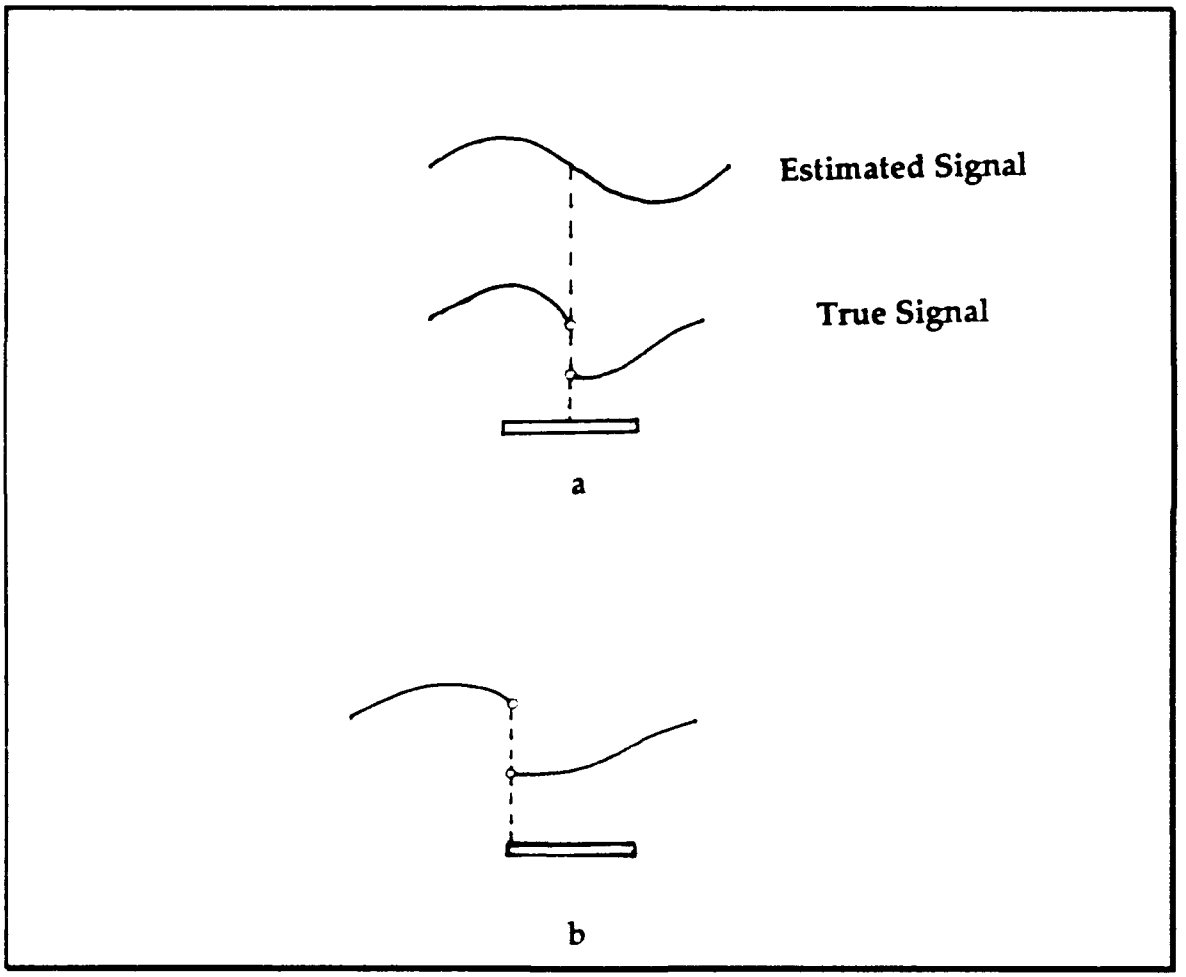


Figure 6.2. Relationship between Estimated Signal and True Signal
a. Probabilistic Mismatch
b. Probabilistic Match

The proposed algorithm can be viewed in light of this interpretation as follows: at *each* time instant t we calculate $d+2$ likelihood terms $l_t(\varepsilon_j)$, of all the possible realizations of $\{\gamma\}$, each one of these realizations is associated with the assumption that the change occurred at the $d+1$ possible locations within the window $[t-d,t]$, and one corresponds to the "no change" hypothesis (see Figure 6.3).

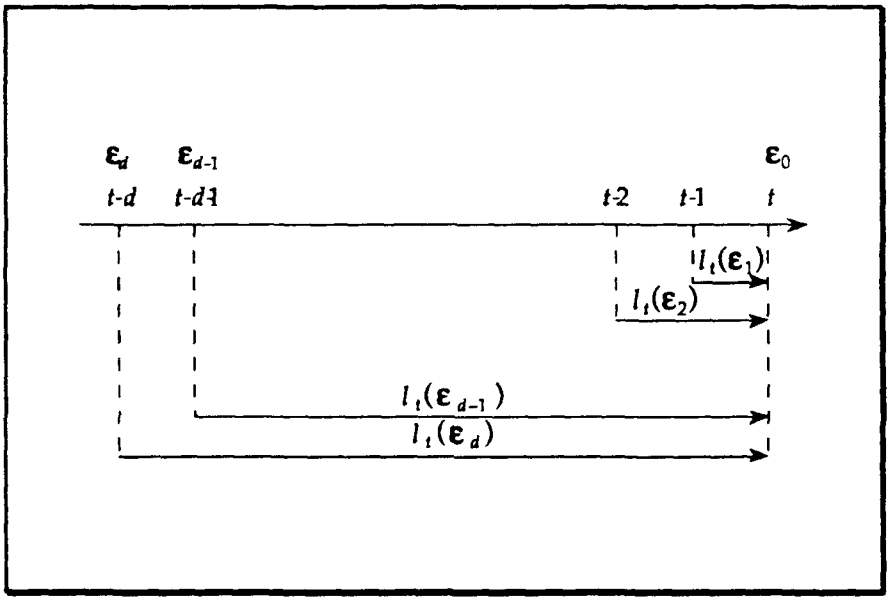


Figure 6.3. Calculating of $d+2$ Likelihood Terms

The algorithm "looks" at all the possible realizations of $\{\gamma\}$ within the window $[t-d,t]$ and decides about a change in a way which will be described in the next section.

3. Recursive Detection

The recursive detection algorithm is based upon the likelihood terms as described in the last section.

Define the MAP estimate of γ as:

$$\hat{\gamma}(t-d) = \arg\max_{\gamma(t-d)} \left\{ P\left(\gamma(t-d) \middle| \mathbf{Y}_t, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) \right\}. \qquad (6-7)$$

By standard Bayesian factorization, it is easy to see that

$$P\left(\mathbf{Y}_{t,d+1}; \gamma(t-d) \middle| \hat{\boldsymbol{\gamma}}_{t-d-1}; \mathbf{Y}_{t-d-1}\right)$$

$$= P\left(\gamma(t-d) \middle| \mathbf{Y}_t, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) \cdot P\left(\mathbf{Y}_{t,d} \middle| \mathbf{Y}_{t-d-1}, \hat{\boldsymbol{\gamma}}_{t-d-1}\right)$$

Since the rightmost term does not depend on $\gamma(t-d)$, it is easy to see that maximization of (6–7) is equivalent to the maximization of

$$\hat{\gamma}(t-d) = \arg\max_{\gamma(t-d)} \left\{ P\left(\mathbf{Y}_{t,d+1}, \gamma(t-d) \middle| \mathbf{Y}_{t-d-1}, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) \right\} \qquad (6-8)$$

The likelihood term in (6–8) can be recursively determined from the probability relation given by Lemma (6–3).

In order to achieve the maximization efficiently, recall that $\boldsymbol{\gamma}_{t,d}$ assumes only the realizations $\boldsymbol{\varepsilon}_j (j = -1, \ldots, d)$ since, at most one transition occurs within any interval $[t-d,t]$, thus, the probability of "no change at $(t-d)$" $(\gamma(t-d) = 0)$ is the union of all the possible mutually exclusive events of a change occurring at each of the other time instants within the window $(j = 1, \ldots, d-1)$ including the event of no change $(j = -1)$.

The hypothesis of no change is given by

$$\mathbf{H}_0: \quad P\left(\mathbf{Y}_{t,d+1}, \gamma(t-d) = 0 \middle| \mathbf{Y}_{t-d-1}, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) = \sum_{j=-1}^{d-1} l_t\left(\boldsymbol{\varepsilon}_j\right) \qquad (6-9)$$

and the hypothesis of change is given by

$$\mathbf{H}_1: \quad P\left(\mathbf{Y}_{t,d+1}, \gamma(t-d) = 1 \middle| \mathbf{Y}_{t-d-1}, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) = l_t\left(\varepsilon_d\right) \qquad (6-10)$$

where

$$l_t(\varepsilon_j) = P\left(\mathbf{Y}_{t,d+1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_j \middle| \mathbf{Y}_{t-d-1}, \hat{\boldsymbol{\gamma}}_{t-d-1}\right) \qquad j = -1,\ldots,d. \qquad (6-11)$$

Hence, the maximization (6–7) becomes hypothesis testing problem

$$\sum_{j=-1}^{d-1} l_t(\boldsymbol{\varepsilon}_j) \underset{H_1}{\overset{H_0}{\gtrless}} l_t(\boldsymbol{\varepsilon}_d) \qquad (6-12)$$

where each likelihood term is given by (6–11) and calculated via the recursion (6–3). Equation (6–12) evaluates the likelihood $l_t(\boldsymbol{\varepsilon}_d)$ of a change at $t-d$ against all the other possible changes within the window. Hence, the vector $\boldsymbol{\varepsilon}_j$ can also be viewed as an **indicator vector** of the change assumption (or re-initialization location of the Kalman filter).

4. Implementation

By careful examination of the recursion (6–3), it is clear that the denominator is constant with respect to the transition sequence $\boldsymbol{\gamma}_{t,d}$. Furthermore the right–most term in the numerator is given by (see Figure 6-1)

$$P\left(\gamma(t)\middle|\boldsymbol{\gamma}_{t-1,d+1}; \mathbf{Y}_{t-d-1}\right) = \begin{cases} 1 & \text{if } \boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_j \qquad j \neq -1 \\ P_0 & \text{if } \boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_{-1} \qquad \gamma(t) = 0 \\ P_1 & \text{if } \boldsymbol{\gamma}_{t-1,d+1} = \boldsymbol{\varepsilon}_{-1} \qquad \gamma(t) = 1 \end{cases}$$

Now, notice that the left-most term in the numerator of (6–3) is computed directly from the Kalman filter equations, since $P\left(y(t)\middle|\mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_j\right)$ is equal to $P\left(y(t)\middle|\mathbf{Y}_{t-1}\right)$ for $j \geq 1$, given that the filter was reinitialized at time $t-j$. Thus, the "past" sequence $\mathbf{Y}_{t-1}$ is the "truncated past": $\{y(t-1) \ldots y(t-j)\}$ and contains all the past observations since the filter's initialization. Hence, calculating (6-11) via (6–3) becomes:

$$l_t(\boldsymbol{\varepsilon}_j) = C\, P\big(y(t)\big|\mathbf{Y}_{t-1,j}\big) P\big(\gamma(t)\big|\boldsymbol{\gamma}_{t-1,d+1};\mathbf{Y}_{t-1,d+1}\big)\, l_{t-1}(\boldsymbol{\varepsilon}_{j-1}) \qquad (6-13)$$

$$j = 1,\ldots,d.$$

where C is a constant independent of $\boldsymbol{\gamma}_{t,d}$.

Using the realization map of $\{\gamma\}$ (Figure 6.1), the update phase of the algorithm can be calculated as follows:

Update:

1. If $\hat{\gamma}(t-d-1) = 0$, then

$j = -1$: $l_t(\boldsymbol{\varepsilon}_{-1}) = C \cdot P(y(t) \mid \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_{-1}) \cdot p_0 \cdot l_{t-1}(\boldsymbol{\varepsilon}_{-1})$

$j = 0$: $l_t(\boldsymbol{\varepsilon}_0) = C \cdot P(y(t) \mid \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_0) \cdot p_1 \cdot l_{t-1}(\boldsymbol{\varepsilon}_{-1})$ (6–14)

for $1 \le j \le d$

$l_t(\boldsymbol{\varepsilon}_j) = C \cdot P(y(t) \mid \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_j) \cdot l_{t-1}(\boldsymbol{\varepsilon}_{j-1})$

2. If $\hat{\gamma}(t-d-1) = 1$, then

$l_t(\boldsymbol{\varepsilon}_{-1}) = C \cdot P(y(t) \mid \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_{-1}) \cdot l_{t-1}(\boldsymbol{\varepsilon}_d)$ (6–15)

$l_t(\boldsymbol{\varepsilon}_j) = 0 \qquad \forall j \neq -1.$

Change detection:

at each time instant t check for:

$$\sum_{j=-1}^{d-1} l_t(\boldsymbol{\varepsilon}_j) \;\begin{matrix} \hat{\gamma}(t-d) = 0 \text{ no change} \\ > \\ < \\ \hat{\gamma}(t-d) = 1 \text{ change} \end{matrix}\; l_t(\boldsymbol{\varepsilon}_d) \qquad (6-16)$$

The procedure introduced above can be represented on the graph shown in Figure 6.4.

At each time instant t, the nodes marked as $j = -1, 0, \ldots, d$ refer to the realizations of $\boldsymbol{\gamma}_{t,d+1}$ and the corresponding likelihood terms $l_t(\boldsymbol{\varepsilon}_j)$. These realizations of the likelihood terms are updated at each node according to equations (6–14) and (6–15).

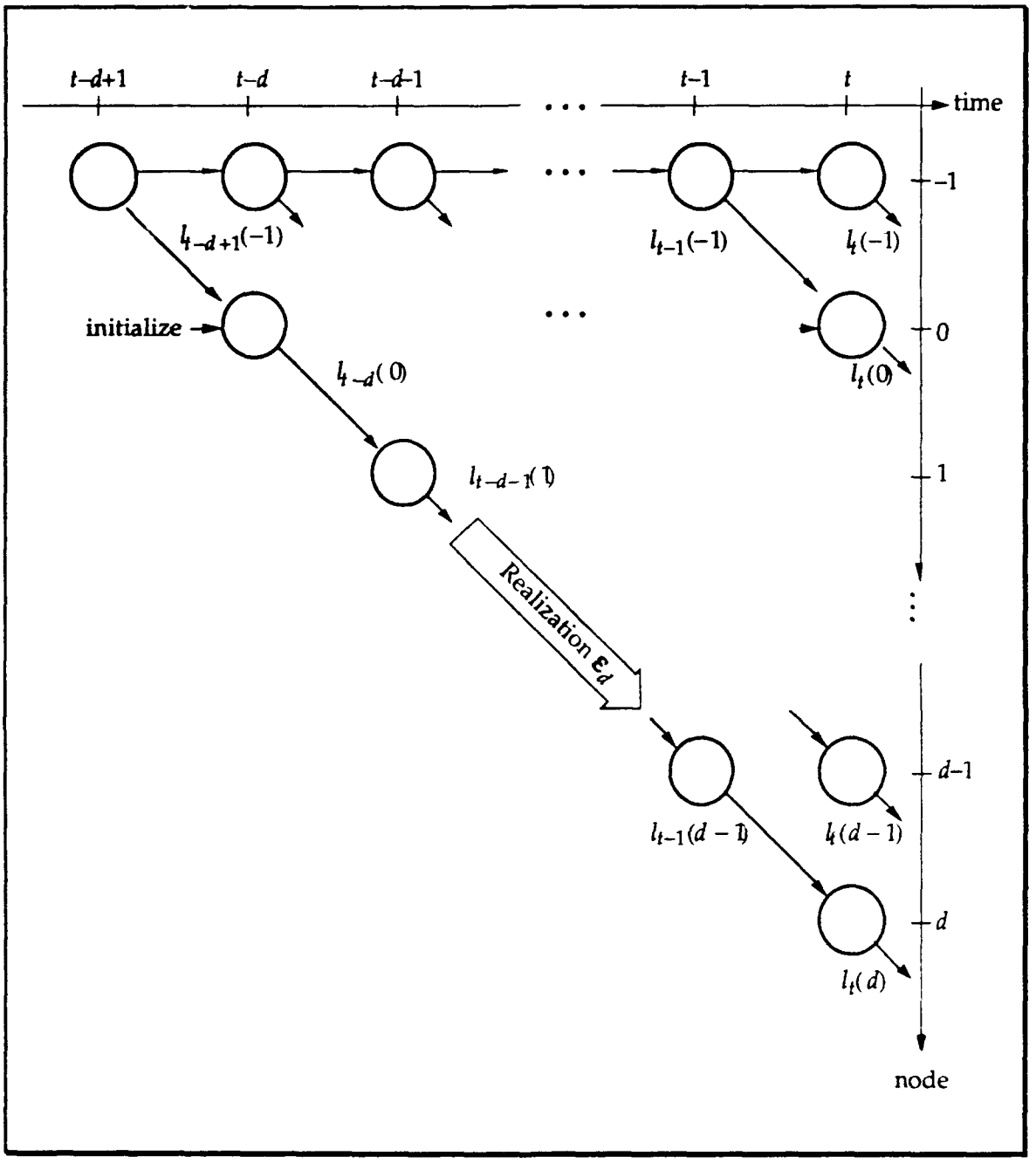


Figure 6.4. Flow Graph of the Detection Algorithm

Updating of these equations is done on the basis of the transition probability, i.e., the conditional probability terms $P(y(t) \mid \mathbf{Y}_{t-1,j})$ and is determined by the well known properties of the **Kalman filter** (Anderson and Moore, 1979)

Let $\mathbf{x}_j$ be the state vector of the state-space model at node j. At each node j of the graph, we update the estimate of $\mathbf{x}_j$, (i.e., $\hat{\mathbf{x}}_j$) and consequently the probability terms are as follows:

1. if $j = 1, ..., d$ then

Time update (use estimate of filter j–1):

$$\hat{\mathbf{x}}_j(t|t-1) = \mathbf{A}\hat{\mathbf{x}}_{j-1}(t-1|t-1)$$

$$\mathbf{V}_j(t|t-1) = \mathbf{A}\mathbf{V}_{j-1}(t-1|t-1)\mathbf{A}^T + \mathbf{B}\,\mathbf{Q}\,\mathbf{B}^T \qquad (6-17)$$

Observation Update (Filtering):

$$\hat{\mathbf{x}}_j(t|t) = \hat{\mathbf{x}}_j(t|t-1) + \mathbf{l}_j(t)\left[y(t) - \mathbf{c}^T\hat{\mathbf{x}}_j(t|t-1)\right]$$

$$\mathbf{l}_j(t) = \mathbf{V}_j(t|t-1)\cdot\mathbf{c}\cdot\left[\mathbf{c}^T\mathbf{V}_j(t|t-1)\mathbf{c} + \sigma^2\right]^{-1}$$

$$\mathbf{V}_j(t|t) = \left[\mathbf{I} - \mathbf{l}_j(t)\cdot\mathbf{c}^T\right]\mathbf{V}_j(t|t-1). \qquad (6-18)$$

2. if $j = 0$, then:

initialize filter.

$$\hat{\mathbf{x}}_0(t|t) = \bar{\mathbf{x}}_{-1}$$

$$\mathbf{V}_0(t|t) = \mathbf{P}_{-1}.$$

$\bar{\mathbf{x}}_{-1}$ and $\mathbf{P}_{-1}$ being the initial state and initial filter error covariance matrix respectively.

3. if $j = -1$, then:

$\hat{\mathbf{x}}_{-1}(t \mid t)$ is updated as in (6–17), (6–18) with the index change (see Figure 6.1)

$$\hat{\mathbf{x}}_{-1}(t|t) = \begin{cases} \hat{\mathbf{x}}_{-1}(t-1|t-1) & \text{if } \hat{\gamma}(t-d-1) = 0 \\ \hat{\mathbf{x}}_{d}(t-1|t-1) & \text{if } \hat{\gamma}(t-d-1) = 1. \end{cases}$$

The state-space representation and the Kalman filter yields an efficient algorithm for the desired transition distribution $P(y(t) \mid \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \boldsymbol{\varepsilon}_j)$,

$$f_t(j) \equiv P\left(y(t) \middle| \mathbf{Y}_{t-1}, \boldsymbol{\gamma}_{t,d+1} = \varepsilon_j\right) = \left(2\pi s_j(t)\right)^{-1/2} \exp\left[\frac{-1}{2s_j(t)}\left(y(t) - \mathbf{c}^T \hat{\mathbf{x}}_j(t|t-1)\right)^2\right]$$

$$s_j(t) = \mathbf{c}^T \mathbf{V}_j(t|t-1)\mathbf{c} + \sigma^2 \qquad (6-19)$$

namely, the desired distributions $f_t(j)$ are Gaussian with mean $\mathbf{c}^T \hat{\mathbf{x}}_j(t \mid t-1)$ and covariance $s_j(t)$. Hence, the likelihood terms (6–14) and (6–15) can be updated on-line by the following equation:

$$l_t\left(\boldsymbol{\varepsilon}_j\right) = C \cdot P_1 \prod_{i=0}^{j} P\left(y(t-i) \middle| \mathbf{Y}_{t-1-i}, \boldsymbol{\varepsilon}_{j-i}\right)$$

$$= C \cdot P_1 \prod_{i=0}^{j} f_{t-i}(j-i) \qquad 1 \le j \le d.$$

The presented algorithm lends itself to a parallel structure implementation. Furthermore, if the state-space matrices are not time dependent, the Kalman filter gains $\mathbf{l}_j$ and covariance matrices $\mathbf{V}j$ can be precomputed, since, in this case the filter's performance is a priori known and not data dependent. Hence, lookup tables can be prepared resulting in a simple computational cost algorithm.

D. RESULTS

The algorithm was implemented and tested on different data structures as piecewise constant models, PSK models and AR models in different signal to noise ratios. The results obtained show that the transitions are estimated by the algorithm.

Figure 6.5 illustrates the results obtained for detecting the transitions in piecewise constant signals represented by the state space model

$$x(n+1) = x(n)$$

$$y(n) = x(n) + w(n)$$

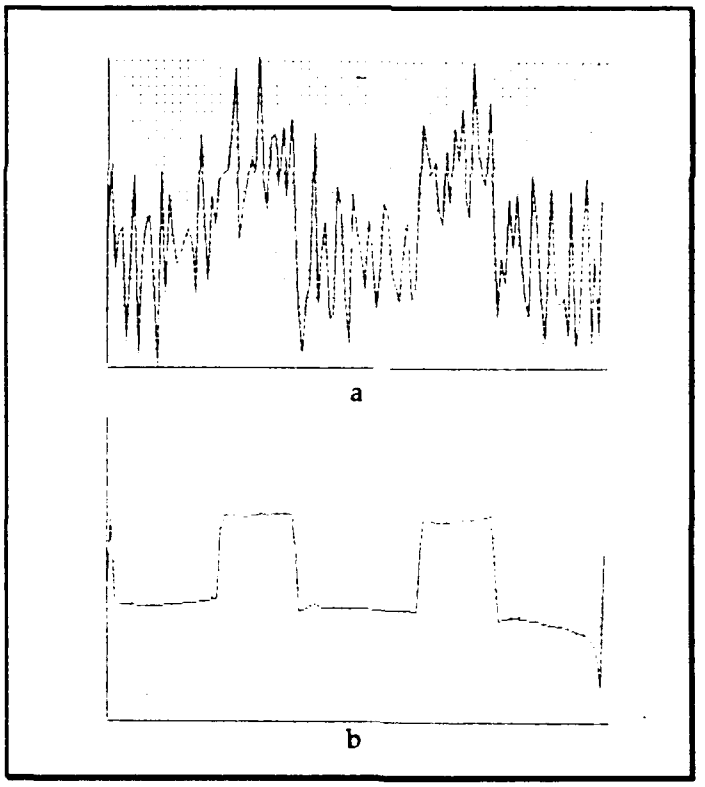


Figure 6.5. The Joint Detection-Estimation of a Piecewise Constant Signal.
a. Noisy Data
b. Filtered Data

Figure 6.6 illustrates the results obtained for an PSK signal with input SNR of of –3dB. Figure 6.7 illustrates the results obtained for an PSK signal with input SNR of –9dB.

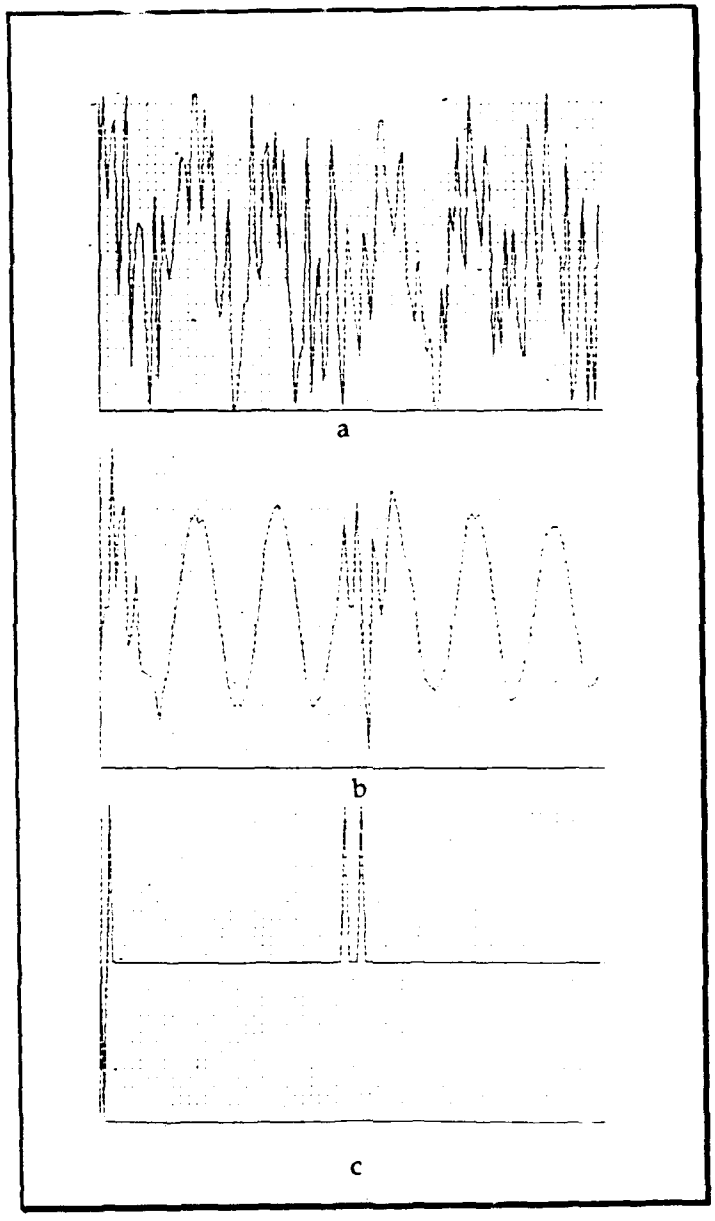


Figure 6.6. PSK Signal with Input SNR of –3dB,
a. Noisy Data
b. Filtered Data
c. Estimated Transitions

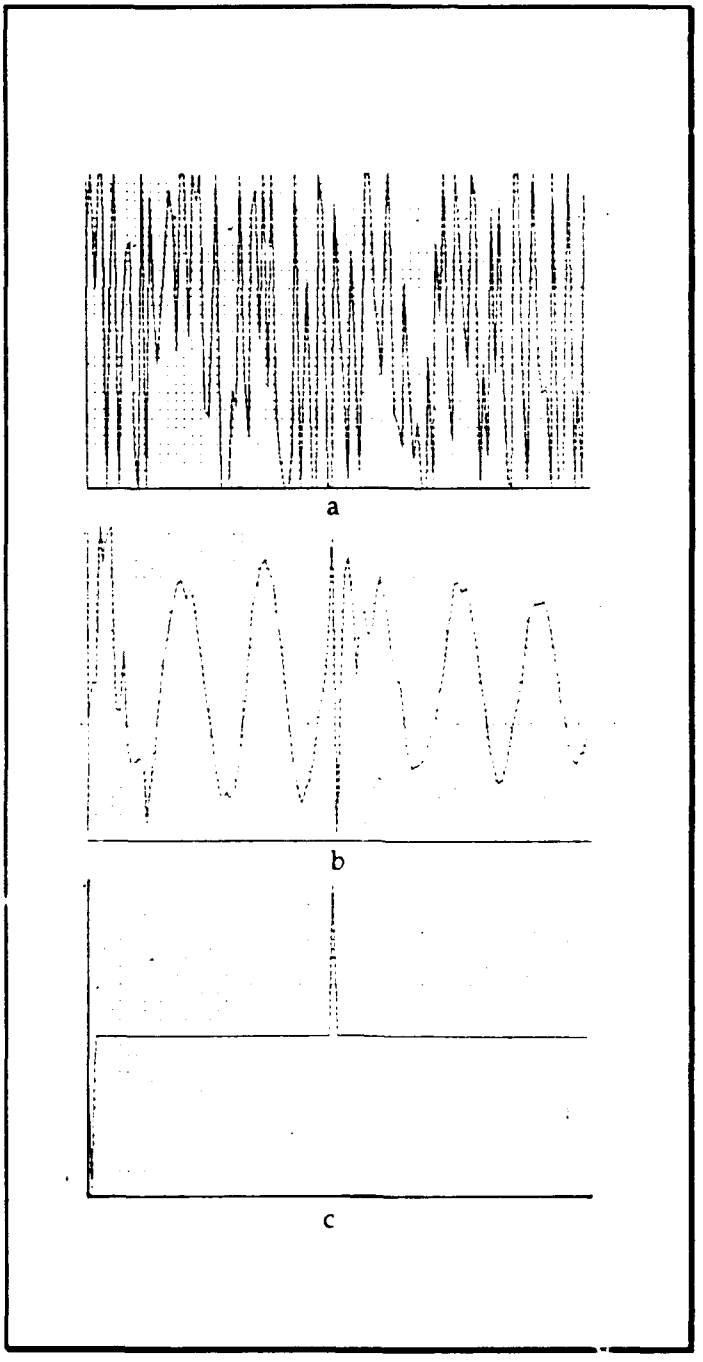


Figure 6.7. PSK Signal with Input SNR of –9dB,
a. Noisy Data
b. Filtered Data
c. Estimated Transitions

Figure 6.8 illustrates data obtained by using an AR model. Figure 6.9 illustrates the estimated transitions while Figures 6.10 and 6.11 illustrate the true and estimated AR parameters.

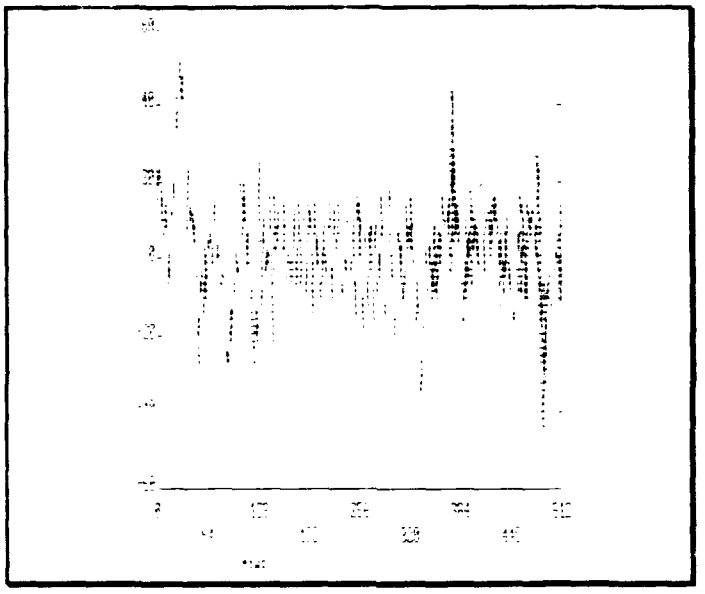


Figure 6.8. The Original AR Data

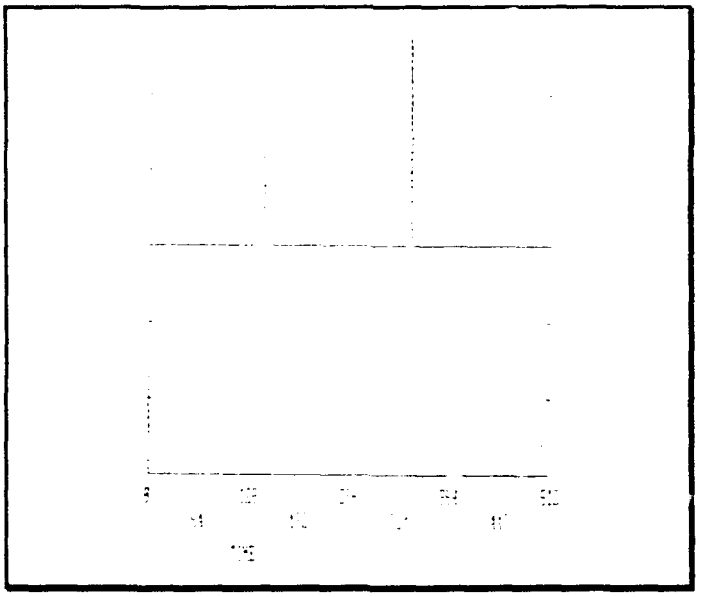


Figure 6.9. Estimated Transitions

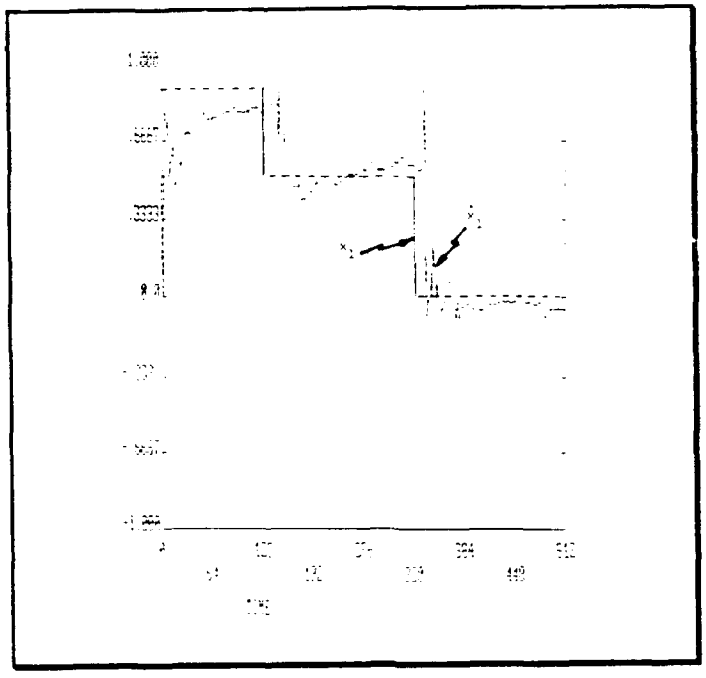


Figure 6.10. True X_1 and Estimated X_1 AR Parameter

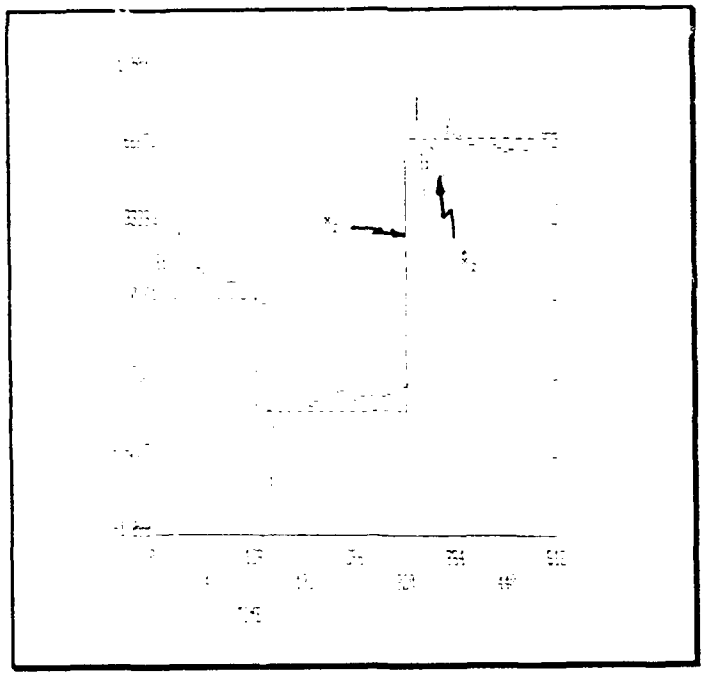


Figure 6.11. True X_2 and Estimated X_2 AR Parameter

E. SUMMARY

The problem of *joint* detection-estimation is addressed in this chapter. The state-space representation of the signal model allows joint detection estimation by using the Kalman filter properties. Furthermore, the detection is completely asynchronous. Since the algorithm is based on optimal estimation techniques, it is expected to be able to detect signals in the presence of very low SNR. However simulation results do not permit this type of conclusion. Detailed performance analysis of this algorithm is not available now and requires more research.

VII. CONCLUSIONS

This dissertation investigated different types of disorder problems by using sequential procedures for on-line implementation. The problem was considered within the framework of detecting changes in statistical models of an observed random process when the disorder can occur at unknown times. The focus of this work was on quickest detection methods for cumsum procedures implemented for different parametric and nonparametric nonlinearities. In this context, several issues remain unresolved, namely, for a multiple disorder problem or for transient detection a critical issue is the joint estimation of disorder time and the model parameters. There is much more to do in investigating this problem by implementing recursive identification procedures together with detection procedures.

In Chapter III, the concept of detecting energy changes in the Energy Spectral Density of a signal reflect different spectral signatures and is of interest in many applications. More work can still be done in the theoretical domain in order to examine the coupling effects between window sizes, averaging methods within a window with the root location and the minimal SNR needed for detection. Moreover, modern spectral energy estimators might be considered rather than the transitional periodogram.

The detection algorithm which was presented in Chapter VI has an advantage of being noncoherent with respect to coherent detectors for PSK type signals. Even though the algorithm is optimal in the sense that optimal techniques (Kalman filtering) were used, there is still room for investigating its performance as a function of window length. Also, the merits of this

approach should be compared to traditional detection methods of PSK signals to check the tradeoff between noncoherent versus coherent detection.

The disorder problem can be considered a local problem. Thus, conventional time frequency methods for detecting and estimating the change parameters have the problem of the tradeoff between the time-frequency resolution. It seems that the wavelet representation which has become popular recently has the potential to resolve this time-frequency resolution problem.

Finally, the research can be extended to the situation where the measurements are dependent (Sadowsky, 1989) for providing a parallel framework for evaluating Page test performance.

LIST OF REFERENCES

Anderson, B., and Moore, J.B., *Optimal Filtering, Prentice-Hall Information and System Sciences Series*, 1979.

Andre-Obrecht, R., "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," *IEEE Transactions on ASSP*, v. 36, no. 1, January 1988.

Appel, U., and Brandt, A., "Adaptive Sequential Segmentation of Piecewise Stationary Time series," *Information Sciences*, v. 29, pp. 27-56, 1983.

Assaf, D., and Ritov, Y., "A Double Sequential Procedure for Detecting a Change in Distribution," *Biometrica*, v. 75, no. 4, pp. 715-722, 1988.

Bagshaw, M., and Johnson, R.A., "The Effect of Serial Correlation on the Performance of Cumsum Tests II," Technometrics, v. 17, no. 1, February 1975.

Bagshaw, M., and Johnson, R. A., "Sequential Procedures for Detecting Parameter Changes in a Time Series Model," *Journal of the American Statistical Association*, September 1977.

Basseville, M., "Edge Detection using Sequential Methods for Change in Level, Part II: Sequential Detection of Change in Mean," *IEEE Transactions on ASSP*, v. 29, no. 1, February 1981.

Basseville, M., "On-line Detection of Jumps in Mean, (Chapter I)," "Detection of Abrupt Changes in Signals and Dynamical Systems," edited by M. Basseville and A. Benveniste, *Lecture Notes in Control and Information Sciences*, v. 77, Springer-Verlag, 1986.

Basseville, M., "The Two-models Approach for the On-line Detection of Changes in AR Processes," edited by M. Basseville and A. Benveniste, *Lecture Notes in Control and Information Sciences*, v. 77, Springer-Verlag, 1986.

Basseville, M., "Detecting Changes in Signals and Systems—A Survey," *Automatica*, v. 24, no. 3, 1988.

Basseville, M., and Benveniste, A., "Design and Comparative Study of some Sequential Jump Detection Algorithms for Digital Signals," *IEEE Transactions on ASSP*, v. 31, no. 3, June 1983.

Basseville, M., and Benveniste, A., "Sequential Segmentation of Nonstationary Digital Signals using Spectral Analysis," *Information Sciences*, v. 29, pp. 57-73, 1983.

Basseville, M., Benveniste, A., "Sequential Detection of Abrupt Changes in Spectral Characteristics of Digital Signals," *IEEE Transactions on Information Theory*, v. IT-29, no. 5, September 1983.

Basseville, M., Benveniste, A., Moustakides, G., "Detection and Diagnosis of Abrupt Changes in Model Characteristics of Non-stationary Digital Signals," *IEEE Transactions on Information Theory*, v. IT-32, no. 3, May 1986.

Benveniste, A., "Advanced Methods of Change Detection: An Overview," edited by M. Basseville and A. Benveniste, *Lecture Notes in Control and Information Sciences*, v. 77, Springer-Verlag, 1986.

Benveniste, A., Basseville, M., and Moustakides, G., "The Asymptotic Local Approach to Change Detection and Model Validation," *IEEE Transactions on Automatic Control*, v. AC-32, no. 7, June 1987.

Basseville, M., Benveniste, A., Moustakides, G., and Rougee, A. "Detection and Diagnosis of Changes in the Eigenstructure of Non-stationary Multivariable Systems," *Automatica*, v. 23, no. 4, July 1987.

Bojdecki, T., "On Optimal Stopping of a Sequence of Independent Random Variables—Probability Maximizing Approach," *Stochastic Processes and their Applications*, v. 6, pp. 153-156, 1978.

Bojdecki, T., and Hosza, J., "On a Generalized Disorder Problem," *Stochastic Processes and their Applications*, v. 18, pp. 349-359, 1984.

Broder, B., *Quickest Detection Procedures and Transient Signal Detection*, Ph.D Dissertation, Princeton, June 1990.

Brown, R. L, Durbin, J., and Evans, J. M., "Techniques for Testing the Constancy of Regression Relationships over Time," *J. Royal Statist. Soc., Ser. B.* v. 37, pp. 149-192, 1975.

Chow, E. Y., and Willsky, A. S. "Issues in the Development of a General Algorithm for Reliable Failure Detection," *Proc. 19th C.D.C.*, Albuquerque, NM, 1980.

Chow, E. Y., and Willsky, A. S., "Analytical Redundancy and the Design of Robust Failure Detection Systems," *IEEE Transactions on Automatic Control*, v. 29, pp. 603-615, 1984.

Chow, E. Y., and Willsky, A. S., "Analytical Redundancy and the Design of Robust Failure Detection Systems," *IEEE Transactions on Automatic Control*, 1986.

Cristi R., and Aviv, D., "Detection of Signals with Discontinuities," presented at the 1991 Midwest Conference of Signals and Systems.

Cristi, R., "Markov and Recursive Least Squares Methods for the Estimation of Data with Discontinuities," *IEEE ASSP Magazine*, v. 38, no. 11, November 1990.

Deekert, J. C., Desai, M. N., and Deyst, J. J. and Willsky, A. S., "F-8 DFBN Sensor Failure Identification using Analytic Redundancy," *IEEE Transactions on Automatic Control*, v. 22, pp. 795-803, 1977.

Deshayes, J., and Picard, D., "Off-line Statistical Analysis of Change-Point Models using Nonparametric and Likelihood methods," *Detection of Abrupt Changes in Signals and Dynamical Systems*, Springer-Verlag, 1986.

Dickinson, B. W., "Properties and Applications of Gaussian Autoregressive Processes in Detection Theory," *IEEE Transactions on Information Theory*, v. IT-27, no. 3, May 1981.

Feller, W., *An Introduction to Probability Theory and its Applications*, v. II, Wiley, 1971.

Gertler, J. J., "Survey of Model Based Failure Detection and Isolation in Complex Plants," *IEEE Control systems Magazine*, v. 8, no. 6, December 1988.

Gersch, W., and Kitagawa, G., "Smoothness Priors in Time Series," Naval Postgraduate School, Technical Report, April 1987.

Ghosh, B. K., *Sequential Tests of Statistical Hypotheses*, Addison-Wesley, 1970.

Gordon, K., and Smith, A. F. M., "Modeling and Monitoring Discontinuous Changes in Time Series," *Bayesian Analysis of Time Series and Dynamic Models*, edited by J. C. Spall, Marcel-Dekker, Statistics, v. 94, 1988.

Govindarajulu, Z., *Sequential Statistical Procedures,* Academic Press, v. 26 in *Probability and Mathematical Statistics,* 1975.

Hinkley, D. V., "Inference about the Change Point in a Sequence of Random Variables," *Biometrika,* v. 57, no. 1, pp. 1-17, 1970.

Hinkley, D. V., "Inference about the Change Point from Cumulative Sum Tests," *"Biometrica,* v. 58, no. 3, pp. 509-523, 1971.

Hinkley, D. V., "Time Ordered Classification," Biometrica, v. 59, no. 3, pp. 509-523, 1972.

Isermann, R., "Process Fault Detection Based on Modeling and Estimation Methods—A Survey," *Automatica,* v. 20, no. 4, July 1984.

Ishii, N., Iwata, A., and Suzumara, N., "Segmentation of Non-stationary Time Series," *International Journal of Systems Sciences,* v. 10, no. 8, 883-894, 1979.

Ishii, N., Iwata, A., and Suzumara, N., "Detection of Abrupt Change and Trend in the Time Series," *International Journal of Systems Sciences,* v. 11, no. 5, pp. 557-566, 1980.

Karlin, S., and Taylor, T. H., *A First Course in Stochastic Processes,* Academic Press, 1968.

Karlin, S., and Taylor, T. H., *A Second Course in Stochastic Processes,* Academic Press, 1981.

Karlin, S., and Taylor, T. H., *An Introduction to Stochastic Modeling,* Academic Press, 1984.

Kassam, S. A., *Signal Detection in Non-Gaussian Noise,* Springer-Verlag, *Texts in Electrical Engineering,* 1987.

Kay, S. M., *Modern Spectral Estimation, Theory and Applications,* Prentice-Hall, 1988.

Kazakos, P. P., and Kazakos, D., *Nonparametric Methods in Communication,* Marcel-Dekker, 1977.

Khan, R., "A Note on Page's Two-sided Cumulative Sum Procedure," *Biometrica,* v. 68, no. 3, pp. 717-719, 1981.

Lehoczky, J. P., "Formulas for Stopped Diffusion Processes with Stopping Times Based on the Maximum," *The Annals of Probability*, v. 5, no. 4, pp. 601-607, 1977.

Lehmann, E. L., *Testing Statistical Hypotheses*, Wiley, 1986.

Ljung, L., and Gunnarsson, S., "Survey Paper: Adaptation and Tracking in System Identification," *Automatica*, v. 26, no. 1, January 1990.

Lorden, G., "Procedures for Reacting to a Change in Distribution," *The Annals of mathematical Statistics*, v. 42, no. 6, pp. 1897-1908, 1971.

MacNeill, I. B., "Tests for Change of Parameter at Unknown Times and Distributions of some Related Functionals of Brownian Motion," *The Annals of Statistics*, v. 2, no. 5, pp. 950-962, 1974.

Moustakides, G. V., "Optimal Stopping Times for Detecting Changes in Distributions," *The Annals of Statistics*, v. 14, no. 4, pp. 1379-1387, 1986.

Moustakides, G. V., and Benveniste, A., "Detecting Changes in the AR Parameters of Nonstationary ARMA Process, *Stochastics*, v. 16, pp. 137-155, 1986.

Nikiforov, I.V., "Cumulative Sums for Detection of Changes in Random Process Characteristics," *Automation and Remote Control*, v. 40, no. 2, Part 1, February 1979.

Nikiforov, I.V., "Modification and Analysis of the Cumulative Sum Procedure," *Automation and Remote Control*, v. 41, no. 9, Part 1, September 1980.

Nikiforov, I.V., "Sequential Detection of Changes in Stochastic Systems," in *Detection of Abrupt Changes in Signals and Dynamical Systems*, M. Basseville and A. Benveniste eds., Springer-Verlag, New York 1986.

Nikiforov, I. V., and Tikhonov, I. N., "Application of Change Detection Theory to Seismic Signal Processing," in *Detection of Abrupt Changes in Signals and Dynamical Systems*, M. Basseville and A. Benveniste eds., Springer-Verlag, New York 1986.

Page, E. S., "Continuous Inspection Schemes," *Biometrica*, v. 41, pp. 100-115, 1954.

Pelkowitz, L., "The General Discrete Time Disorder Problem," *Stochastics*, v. 20, pp. 89-110.

Pena, d., and Guttman, I., "Bayesian Approach to Robustifying the Kalman Filter" in *Bayesian Analysis of Time Series and Dynamic Models*, edited by J. C. Spall, Marcel Dekker, in *Statistics*, v. 94, 1988.

Point, G., *Detection and Analysis of PSK Signals using Kalman Filtering Techniques*, Master's Thesis, Naval Postgraduate School, Monterey, California, 1987.

Pollak, M. and Siegmund, D., "A Diffusion Process and its Applications to Detecting a Change in the Drift of a Brownian Motion," *Biometrica*, v. 72, no. 2, pp. 267-280, 1985.

Pollak, M., "Optimal Detection of a Change in Distribution," *The Annals of Statistics*, v. 13, no. 1, pp. 206-227, 1985.

Poor, H. V., *An Introduction to Signal Detection and Estimation, Springer Texts in Electrical Engineering*, Springer-Verlag, 1988.

Reynolds, M. R., "Approximation to the Average Run Length in Cumulative Sum Control Charts," *Technometrics*, v. 17, no. 1, February 1975.

Roberts, S. W., "A Comparison of Some Control Chart Procedures," *Technometrics*, v. 8, no. 3, August, 1966.

Ross, S. M., *Introduction to Probability Models*, Academic Press, 1989.

Sadowsky, J.S., "A Dependent Data Extension of Wald's Identity and its Application to Sequential Test Performance Computation," *IEEE Transactions on Information Theory*, v. 35, no. 4, July, 1989.

Segen, J., and Sanderson, A. C., "Detecting Change on a Time Series," *IEEE Transactions on Information Theory*, v. IT-26, no. 2., March 1980.

Shiryayev, A. N., *On Optimum Methods in Quickest Detection Problems, Theory of Probability and its Applications*, v. VIII, no. 1, 1963.

Shiryayev, A. N., "Some Exact Formulas in a 'Disorder' Problem," *Theory of Probability and its Applications*, v. 10, no. 3, pp. 348-354, 1965.

Shiryayev, A. N., "Optimal Stopping Rules," *Applications of Mathematics*, Springer-Verlag, 1978.

Shumway, R. H., "Dynamic Linear Models with Switching," Technical Report 91, Division of Statistics, University of California, Davis, July 1990.

Siegmund, D., *Sequential analysis, Tests and Confidence Intervals, Springer Series in Statistics*, Springer-Verlag, 1985.

Srivastava, M.S. and Wu, Y., *Comparison of Cumsum, Ewma, and Shiryayev Roberts Procedures for Detecting a Shift in the Mean*, Technical Report 9017, Department of Statistics, University of Toronto, 1990.

Taylor, H. M., "A Stopped Brownian Motion Formula," *The Annals of Probability*, v. 3, no. 2, pp. 234-246, 1975.

Therrien, C. W., "A Sequential Approach to Target Discrimination," *IEEE Transactions on Aerospace and Electronic Systems*, v. AES-14, no. 3, May 1978.

Therrien, C. W., *Decision Estimation and Classification*, Wiley, 1989.

Therrien, C. W., *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, 1991.

Tugnait, J. K., "Adaptive Estimation and Identification for Discrete Systems with Markov Jump Parameters," *IEEE Transactions on Automatic Control*, v. AC-27, no. 5, October 1982.

Tugnait, J. K., "Detection and Estimation for Abruptly Changing Systems," *Automatica*, v. 18, no. 5, September 1982.

Wald, A., *Sequential Analysis*, Wiley, New York, 1947.

Whalen, A. D., *Detection of Signals in Noise*, Academic Press, 1971.

Willsky, A. S., "A Survey of Design Methods for Failure Detection in Dynamical Systems," *Automatica*, v. 12, pp. 601-611, 1976.

Willsky, A. S., "Detection of Abrupt Changes in Dynamic Systems," Chapter 3 in *Lectures Notes in Control and Information Sciences*, v. 77, Springer-Verlag, 1986.

Willsky, A. S., and Jones, H. L., "A Generalized Likelihood Ratio Approach to the Detection and Estimation of Jumps in Lienar Systems," *IEEE Transactions on Automatic Control*, February 1976.

Wolcin, J. J., "Maximum Likelihood Detection of Transient Signals using Sequential Short-time Power Spectra," *TM 831138*, Naval Underwater Systems Center, August 1983.

APPENDIX. BASIC CONCEPTS OF HYPOTHESIS TESTING AND DETECTION THEORY

(FROM KASSAM, 1988)

Let $X = (X_1, X_2, ..., X_n)$ be a random vector of observations with joint probability density function (pdf) $P_{\mathbf{x}}(\mathbf{x} \mid \theta)$, where θ is a parameter of the density function. Any specific realization $\mathbf{x} = (x_1, x_2, ..., x_n)$ of X will be a point in $\Re^n$, where $\Re$ is the set of all real numbers. In binary hypothesis-testing problems we have to decide between one of two hypotheses, which we will label as H_0 and H_1 about the pdf $P_{\mathbf{x}}(\mathbf{x} \mid \theta)$, given an observation vector in $\Re^n$. Let $\boldsymbol{\Theta}$ be the set of all possible values of θ; we usually identify H_0 with one subset $\boldsymbol{\Theta}_{H_0}$ of θ values and H_1 with a disjoint subset $\boldsymbol{\Theta}_{H_1}$, so that $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{H_0} \cup \boldsymbol{\Theta}_{H_1}$. This may be expressed formally as

$$H_0: \ X \text{ has pdf } P_{\mathbf{x}}(\mathbf{x}|\theta) \text{ with } \theta \in \boldsymbol{\Theta}_{H_0} \qquad (A-1)$$

$$H_1: \ X \text{ has pdf } P_{\mathbf{x}}(\mathbf{x}|\theta) \text{ with } \theta \in \boldsymbol{\Theta}_{H_1}. \qquad (A-2)$$

If $\boldsymbol{\Theta}_{H_0}$ and $\boldsymbol{\Theta}_{H_1}$ are made up of single elements, say $\boldsymbol{\Theta}_{H_0}$ and $\boldsymbol{\Theta}_{H_1}$, respectively, we say that the hypotheses are **simple**; otherwise the hypotheses are **composite.** If $\boldsymbol{\Theta}$ can be viewed as a subset of $\Re^p$ for a finite integer p, the pdf $P_{\mathbf{x}}(\mathbf{x} \mid \theta)$ is completely specified by the finite number p of real components of θ, and we say that our hypotheses are **parametric**.

A *test* for the hypothesis H_0 against H_1 may be specified as a partition of the same space $S = \Re^n$ of observations into disjoint subsets S_{H_0} and S_{H_1}, so that x falling in S_{H_0} leads to acceptance of H_0, with H_1 accepted otherwise. This may also be expressed by a *test function* which is defined to have value

$\delta(x) = 1$ for $x \in S_{H_1}$ and value $\delta(x) = 0$ for $x \in S_{H_0}$. The value of the test function is defined to be the probability with which the hypothesis H_1, the **alternative hypothesis**, is accepted. The hypothesis H_0 is called the **null hypothesis.**

More generally, the test function can be allowed to take on probability values in the closed interval [0,1]. A test based on a test function taking on values inside [0,1] is called a **randomized test**.

The **power function** $\mathcal{P}(\theta \mid \delta)$ of a test based on a test function δ is defined for $\theta \in \Theta_{H_0} \cup \Theta_{H_1}$ as

$$\mathcal{P}(\theta|\delta) = E\{\delta(x)|\theta\} \qquad (A\text{-}3)$$

$$= \int_{\Re^n} \delta(x) P_x(x|\theta) dx.$$

Thus it is the probability with which the test will accept the alternative hypothesis H_1 for any particular parameter value θ. When θ is in Θ_{H_0} the value of $\mathcal{P}(\theta \mid \delta)$ gives the probability of an error, that of accepting H_1 when H_0 is correct. This is called a *type I error* or the *probability of false alarm*, and depends on the particular value of θ in Θ_{H_0}. The **size** of a test is the quantity

$$\alpha = \sup_{\theta \in \Theta_{H_0}} \mathcal{P}(\theta|\delta) \qquad (A\text{-}4)$$

which may be considered as being the best upper bound on the type I error probability of the test.

Similarly, we define the **Operating Characteristic** (OC) of a test $Q(\theta \mid \delta)$, based on a test function δ, as the probability with which a test will accept the null hypothesis H_0 for any particular parameter value θ. When θ is in Θ_{H_0} the value of $Q(\theta \mid \delta)$ gives the **confidence** $(1-\alpha)$, that of accepting H_0 when H_0

is correct. When θ is in Θ_{H_1} the value of $Q(\theta \mid \delta)$ gives the *probability of miss* (β), that of accepting H_0 when H_1 is correct. This is called a *type II error* and depends on the particular value of θ in Θ_{H_1} Figure A.1 illustrates a typical OC function of a test.

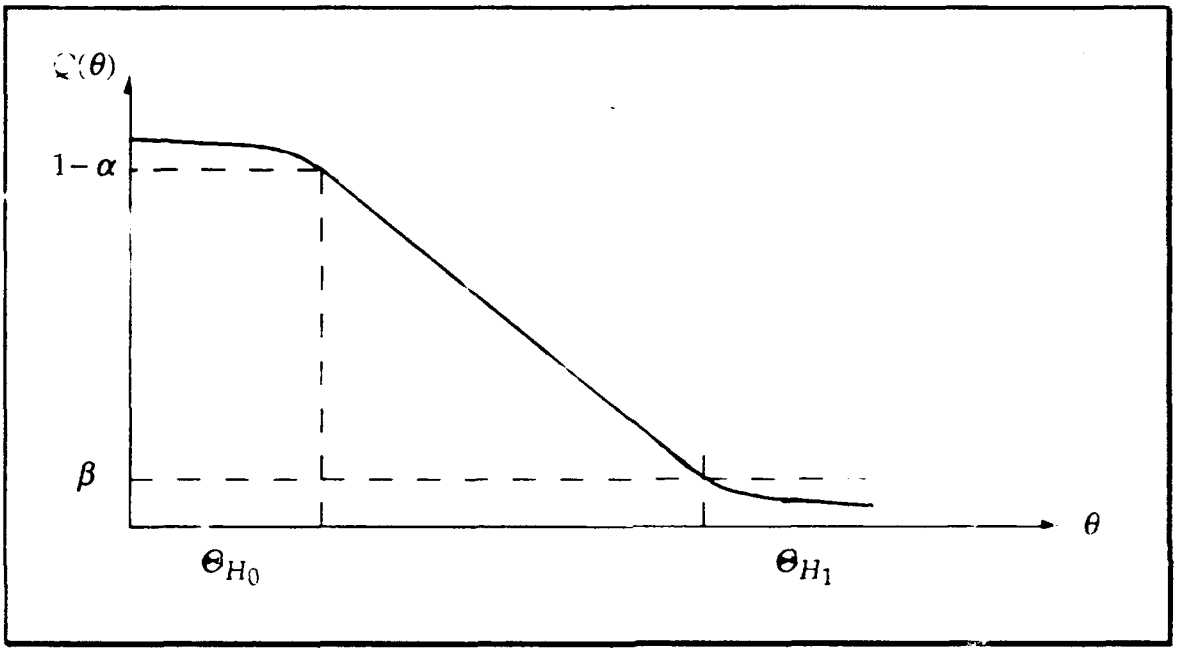


Figure A.1. A Typical Operating Characteristic Function of a Test

In signal detection the null hypothesis is often a noise-only hypothesis, and the alternative hypothesis expresses the presence of a signal in the observations. For a detector D implementing a test function $\delta(\mathbf{x})$ the power function evaluated for any θ *in* Θ gives a **probability of detection** of the signal. Thus, we will use the notation $\mathcal{P}(\theta \mid D)$ for the power function of a detector D, and in discussing the probability of detection at a particular value of the parameter θ in Θ_{H_1} (or for a simple alternative hypothesis H_1) we will use for it the notation P_D. The size of a detector is often called its **false-alarm**

probability. This usage is encountered specifically when the noise-only null hypothesis is simple, and the notation for this probability is P_{FA}.

A. MOST POWERFUL TESTS AND THE NEYMAN-PEARSON LEMMA

Given a problem of binary hypothesis testing such as defined by (A-1) and (A-2), the question arises as to how one may define and then construct an optimum test. Ideally, one would like to have a test for which the power function $\mathcal{P}(\theta \mid \delta)$ has values close to zero for θ in Θ_{H_0}, and has values close to unity for θ in Θ_{H_1}. These are, however, conflicting requirements. We can instead impose the condition that the size α of any acceptable test be no larger than some reasonable level α_0, and subject to this condition look for a test for which $\mathcal{P}(\theta \mid \delta)$, evaluated at a particular value θ_{H_1} of θ in Θ_{H_1}, has its largest possible value. Such a test is *most powerful* at level α_0 in testing H_0 against the simple alternative $\theta = \theta_{H_1}$ in Θ_{H_1}; its test function $\delta^*(\mathbf{x})$ satisfies

$$\sup_{\theta \in \Theta_{H_0}} \mathcal{P}(\theta \mid \delta^*) \leq \alpha_0 \qquad (A-5)$$

$$\mathcal{P}\left(\theta_{H_1} \mid \delta^*\right) \geq \mathcal{P}\left(\theta_{H_1} \mid \delta\right) \qquad (A-6)$$

for all other test function $\delta(\mathbf{x})$ of size less than or equal to α_0. In most cases of interest a most powerful level α_0 test satisfies (A-5) with equality, so that its size is $\alpha = \alpha_0$.

For a simple null hypothesis H_0 when $\theta = \theta_{H_0}$ is the only parameter value in Θ_{H_0}, the condition (A-5) becomes $\mathcal{P}(\theta_{H_1} \mid \delta^*) \leq \alpha_0$ or $P_{FA} \leq 0$, subject to which P_D at $\theta = \theta_{H_1}$ is maximized. For this problem of testing a simple H_0 against a simple H_1, a fundamental result of Neyman and Pearson (called the

Neyman-Pearson lemma) gives the structure of the most powerful test. We state the result here as a theorem:

Theorem 1: Let $\delta(\mathbf{x})$ be a test function of a the form

$$\delta(\mathbf{x}) = \begin{cases} 1 & , \quad P_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_1}\right) > tP_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_0}\right) \\ r(\mathbf{x}) & , \quad P_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_1}\right) = tP_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_0}\right) \\ 0 & , \quad P_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_1}\right) < tP_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_0}\right) \end{cases} \qquad (A-7)$$

for some constant $t \geq 0$ and some function $r(\mathbf{x})$ taking on values in [0,1]. Then the resulting test is most powerful at level equal to its size for H_0: $\theta = \theta_{H_0}$ versus H_1: $\theta = \theta_{H_1}$.

In addition to the above sufficient condition for a most powerful test it can be shown that conversely, if a test is known to be most powerful at level equal to its size, then its test function must be of the form (A-7) except perhaps on a set of $\mathbf{x}$ values of probability measure zero. Additionally, we may always require $r(\mathbf{x})$ in (A-7) to be a constant r in [0,1]. Finally, we note that we are always guaranteed the existence of such a test for H_0 versus H_1, of given size α [Lehmann, 1959, Ch. 3].

From the above result we see that generally the structure of a most powerful test may be described as one comparing a *likelihood ratio* to constant *threshold*,

$$\frac{P_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_1}\right)}{P_{\mathbf{x}}\left(\mathbf{x}|\theta_{H_0}\right)} > t \qquad (A-8)$$

in deciding if the alternative H_1 is to be accepted. If the likelihood ratio on the left-hand side of (A-8) equals the threshold value t, the alternative H_1 may be accepted with some probability r (the randomization probability). The

constants t and r may be evaluated to obtain a desired size α using knowledge of the distribution function of the likelihood ratio under H_0.

When the alternative hypothesis H_1 is composite we may look for a test which is **uniformly most powerful (UMP)** in testing H_0 against H_1, that is, one which is most powerful for H_0 against each $\theta = \theta_{H_1}$ in Θ_{H_1}. While UMP tests can be found in some cases, notably in many situations involving Gaussian noise in signal detection, such tests do not exist for many other problems of interest. One option in such situations is to place further restrictions on the class of acceptable or admissible tests in defining a most powerful test; for example, a requirement of **unbiasedness** or of **invariance** may be imposed [Lehmann, 1959, Ch. 4-6]. As an alternative, other performance criteria based on the power function may be employed. We will consider one such criterion, leading to **locally optimum** or **locally most powerful** tests for composite alternatives, in the next section. One approach to obtaining reasonable tests for composite hypotheses is to use maximum-likelihood estimates $\hat{\theta}_{H_0}$ and $\hat{\theta}_{H_1}$ of the parameter θ, obtained under the constraints of $\theta \in \Theta_{H_0}$ and $\theta \in \Theta_{H_1}$, respectively, in place of θ_{H_0} and θ_{H_1} in (A-8). The resulting test is called a **generalized likelihood ratio** (GLR) test or simply a **likelihood ratio** test.

B. LOCAL OPTIMALITY AND THE GENERALIZED NEYMAN-PEARSON LEMMA

Let us now consider the approach to construction of tests for composite alternative hypotheses. In this approach attention is concentrated on alternatives $\theta = \theta_{H_1}$, in Θ_{H_1}, which are close, in the sense of a metric or

distance, to the null-hypothesis parameter value $\theta = \theta_{H_0}$. Specifically, let θ be a real-valued parameter with value $\theta = \theta_0$ defining the simple null hypothesis and let $\theta > \theta_0$ define the composite alternative hypothesis. Consider the class of all tests based on test functions $\delta(x)$ of a particular desired size α for $\theta = \theta_0$ against $\theta > \theta_0$, and assume that the power function $\mathcal{P}(\theta \mid \delta)$ of these tests are continuous and also continuously differentiable at $\theta = \theta_0$. Then if we are interested primarily in performance for alternatives which are close to the null hypothesis, we can use as a measure of performance the slope of the power function at $\theta = \theta_0$, that is

$$\mathcal{P}'(\theta_0 \mid \delta) = \mathcal{P}'(\theta \mid \delta)\Big|_{\theta=\theta_0} \qquad (A-9)$$

$$= \frac{d}{d\theta}\mathcal{P}(\theta \mid \delta)\Big|_{\theta=\theta_0}$$

From among our class of tests of size α, the test based on $\delta^*(x)$ which uniquely maximizes $\mathcal{P}'(\theta_0 \mid \theta)$ has a power function satisfying

$$\mathcal{P}'(\theta \mid \delta^*) \geq \mathcal{P}(\theta \mid \delta), \quad \theta_0 < \theta < \theta_{\max} \qquad (A-10)$$

for some $\theta_{\max} > \theta_0$. Such a test is called a **locally most powerful** or **locally optimum** (LO) test for $\theta = \theta_0$ against $\theta > \theta_0$. It is clearly of interest in situations such as the weak-signal case in signal detection, when the alternative-hypothesis parameter values of primary concern are those which define pdf's $P_x(x \mid \theta)$ close to the null-hypothesis noise-only pdf $P_x(x \mid \theta_{H_0})$.

The following generalization of the Neyman-Pearson fundamental result of Theorem 1 can be used to obtain the structure of an LO test:

Theorem 2: Let $g(\mathbf{x})$ and $h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_m(\mathbf{x})$ be real-valued and integrable functions defined on $\mathfrak{R}^n$. Let an integrable function $\delta(\mathbf{x})$ on $\mathfrak{R}^n$ have the characteristics

$$\delta(\mathbf{x}) = \begin{cases} 1 & , \quad g(\mathbf{x}) > \sum_{i=1}^{m} t_i\, h_i(\mathbf{x}) \\ r(\mathbf{x}) & , \quad g(\mathbf{x}) = \sum_{i=1}^{m} t_i\, h_i(\mathbf{x}) \\ 0 & , \quad g(\mathbf{x}) < \sum_{i=1}^{m} t_i\, h_i(\mathbf{x}) \end{cases} \qquad (A-11)$$

for a set of constants $t_i \geq 0$, $i = 1, 2, \ldots, m$, and where $0 \leq r(\mathbf{x}) \leq 1$. Define, for $i = 1, 2, \ldots, m$, the quantities

$$\alpha_i = \int_{\mathfrak{R}^n} \delta(\mathbf{x}) h_i(\mathbf{x}) d\mathbf{x}. \qquad (A-12)$$

Then from within the class of all test functions satisfying the m constraints (A-12), the function $\delta(\mathbf{x})$ defined by (A-11) maximizes $\int_{\mathfrak{R}^n} \delta(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$.

A more complete version of the above theorem, and its proof, may be found in [Lehmann, 1959, Ch. 3]; Ferguson [1967, Ch. 5] also discusses the use of this result.

To use the above result in finding an LO test for $\theta = \theta_0$ against $\theta > \theta_0$ defining Θ_{H_0} and Θ_{H_1} in (A-1) and (A-2), respectively, let us write (A-9) explicitly as

$$\begin{aligned} \mathcal{P}'(\theta_0 | \delta) &= \frac{d}{d\theta} \int_{\mathfrak{R}^n} \delta(\mathbf{x}) P_{\mathbf{x}}(\mathbf{x}|\theta) d\mathbf{x} \Big|_{\theta=\theta_0} \\ &= \int_{\mathfrak{R}^n} \delta(\mathbf{x}) \frac{d}{d\theta} P_{\mathbf{x}}(\mathbf{x}|\theta) \Big|_{\theta=\theta_0} d\mathbf{x} \end{aligned} \qquad (A-13)$$

assuming that our pdf's are such as to allow the interchange of the order in which limits and integration operations are performed. Taking $m = 1$ and identifying $h_1(\mathbf{x})$ with $P_{\mathbf{x}}(\mathbf{x} \mid \theta_0)$ and $g(\mathbf{x})$ with $\frac{d}{d\theta} P_X(\mathbf{x}|\theta)\Big|_{\theta=\theta_0}$ in Theorem 2, we are led to the locally optimum test which accepts the alternative H_1: $\theta \geq \theta_0$ when

$$\frac{\frac{d}{d\theta} P_{\mathbf{x}}(\mathbf{x}|\theta)\Big|_{\theta=\theta_0}}{P_{\mathbf{x}}(\mathbf{x}|\theta_0)} > t \qquad (A-14)$$

where t is the test threshold value which results in a size-α test satisfying

$$E\{\delta(\mathbf{X})|H:\theta = \theta_0\} = \alpha. \qquad (A-15)$$

The test of (A-14) may also be expressed as one accepting the alternative when

$$\frac{d}{d\theta} \ln\{P_{\mathbf{x}}(\mathbf{x}|\theta)\}\Big|_{\theta=\theta_0} > t. \qquad (A-16)$$

Theorem 2 may also be used to obtain tests maximizing the *second* derivative $\mathcal{P}'(\theta_0 \mid \delta)$ at $\theta = \theta_0$. This would be appropriate to attempt if it so happens that $\mathcal{P}'(\theta_0 \mid \delta) = 0$ for all size-α tests for a given problem. The condition $\mathcal{P}'(\theta_0 \mid \delta) = 0$ will occur if $\frac{d}{d\theta} P_{\mathbf{x}}(\mathbf{x}|\theta)\Big|_{\theta=\theta_0}$ is zero, assuming the requisite regularity conditions mentioned above. In this case Theorem 2 can be applied to obtain the locally optimum test accepting the alternative hypothesis H_1: $\theta > \theta_0$ when

$$\frac{\frac{d^2}{d\theta^2} P_{\mathbf{x}}(\mathbf{x}|\theta)\Big|_{\theta=\theta_0}}{P_{\mathbf{x}}(\mathbf{x}|\theta)} > t. \qquad (A-17)$$

One type of problem for which Theorem 2 is useful in characterizing locally optimum tests is that of testing $\theta = \theta_0$ against the two-sided alternative hypothesis $\theta \neq \theta_0$. We have previously mentioned that one can impose the condition of **unbiasedness** on the allowable tests for a problem. Unbiasedness of a size-α test for the hypotheses H_0 and H_1 of (A-1) and (A-2) means that the test satisfies

$$\mathcal{P}(\theta|\delta) \leq \alpha, \quad \text{all } \theta \in \Theta_{H_0} \tag{A-18}$$

$$\mathcal{P}(\theta|\delta) \leq \alpha, \quad \text{all } \theta \in \Theta_{H_1} \tag{A-19}$$

so that the detection probability for any $\theta_{H_1} \in \Theta_{H_1}$, is never less than the size α. For the two-sided alternative hypothesis $\theta \neq \theta_0$, suppose the pdf's $P_{\mathbf{x}}(\mathbf{x} \mid \theta)$ are sufficiently regular so that the power functions of all tests are twice continuously differentiable at $\theta = \theta_0$. Then it follows that for any **unbiased** size-α test we will have $\mathcal{P}(\theta_0|\delta) = \alpha$ and $\mathcal{P}'(\theta_0 \mid \delta) = 0$. Thus, the test function of a locally optimum unbiased test can be characterized by using these two constraints and maximizing $\mathcal{P}''(\theta_0 \mid \delta)$ in Theorem 2. Another interpretation of the above approach for the two-sided alternative hypothesis is that the quantity $\omega = (\theta - \theta_0)^2$ may then be used as a measure of the distance of any alternative hypothesis from the null hypothesis $\theta = \theta_0$. We have

$$\frac{d}{d\omega}\mathcal{P}(\theta|\delta)\Big|_{\omega=0} = \frac{1}{2(\theta-\theta_0)}\frac{d}{d\theta}\mathcal{P}(\theta|\delta)\Big|_{\theta=\theta_0}$$

$$= \frac{1}{2}\mathcal{P}''(\theta_0|\delta) \tag{A-20}$$

if $\mathcal{P}'(\theta_0 \mid \delta)$ is zero, for sufficiently regular pdf's $P_{\mathbf{x}}(\mathbf{x} \mid \theta)$. Thus if $\mathcal{P}'(\theta_0 \mid \delta)$ is zero for a class of size-α tests, then maximization of $\mathcal{P}''(\theta_0 \mid \delta)$ leads to a test which is locally optimum within that class.

In this appendix we are concerned with problems where the noise density function P is completely specified, as a special case of the general *parametric* problem where P may have a finite number of unknown parameters (such as the noise variance). Our detection problem can be formulated as a statistical hypothesis-testing problem of choosing between a **null hypothesis** H_0 and an **alternative hypothesis** H_1 describing the joint density function P_X of the observation vector **X**, with

$$H_0: \quad P_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{n} P(x_i) \qquad (A-21)$$

$$H_1: \quad P_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{n} P(x_i - \theta s_i), \ \mathbf{s} \text{ specified, any } \theta > 0. \qquad (A-22)$$

Here s is the vector $(s_1, s_2, \ldots, s_n)$ of signal components. Note that we are considering parametric hypotheses which completely define P_X to within a finite number of unknown parameters (here with only $\theta > 0$ unknown under the alternative hypothesis). Let us now proceed to obtain the structures of tests for H_0 versus H_1.

C. LOCALLY OPTIMUM DETECTION AND ASYMPTOTIC OPTIMALITY

Since the alternative hypothesis H_1 is not a simple hypothesis, the signal amplitude value being unspecified, we cannot apply directly the fundamental lemma of Neyman and Pearson to obtain the structure of the optimum detector for the detection problem. For non-Gaussian noise densities it is also

generally impossible to obtain UMP tests for the composite alternative hypothesis H_1.

To illustrate the difficulty, consider the special case where P is specified to be the *double-exponential* noise density function defined by

$$P(x) = \frac{a}{2} e^{-a|x|}, a \geq 0. \qquad (A-23)$$

The likelihood ratio for testing H_0 versus H_1 for a particular value $\theta = \theta_0 > 0$, is

$$L(\mathbf{X}) = \prod_{i=1}^{n} \frac{P(x_i - \theta_0 s_i)}{P(x_i)}. \qquad (A-24)$$

This now becomes

$$L(\mathbf{X}) = e^{-a\sum_{i=1}^{n}(|x_i - \theta_0 s_i| - |x_i|)} \qquad (A-25)$$

giving

$$\ln L(\mathbf{X}) = a\sum_{i=1}^{n}(|x_i| - |x_i - \theta_0 s_i|). \qquad (A-26)$$

Thus for *given* $\theta = \theta_0$, the test based on

$$\tilde{\lambda}(\mathbf{X}) = \sum_{i=1}^{n}(|x_i| - |x_i - \theta_0 s_i|) \qquad (A-27)$$

is an optimum test, since the constant a is positive. The optimum detector therefore has a test function defined by

$$\delta(\mathbf{X}) = \begin{cases} 1 & , \quad \tilde{\lambda}(\mathbf{X}) > t \\ r & , \quad \tilde{\lambda}(\mathbf{X}) = t \\ 0 & , \quad \tilde{\lambda}(\mathbf{X}) < t \end{cases} \qquad (A-28)$$

where the threshold t and randomization probability r are chosen to obtain the desired value for the false-alarm probability P_{FA}, so that the equation

$$E\{\delta(\mathbf{X})|H_1\} = P_{FA} \qquad (A-29)$$

is satisfied. Notice that we do not need randomization at $\tilde{\lambda}(\mathbf{X}) = t$ if this event has zero probability under H_0.

We can express $\tilde{\lambda}(\mathbf{X})$ of (2-19) in the form

$$\tilde{\lambda}(\mathbf{X}) = \sum_{i=1}^{n} l(x_i; \theta_0 s_i) \qquad (A-30)$$

where the characteristic l is defined by

$$l(x; \theta_s) = |x| - |x - \theta_s|. \qquad (A-31)$$

This is shown in Figure A.2 as a function of x and depends strongly on θ, so that $\tilde{\lambda}(\mathbf{X})$ cannot be expressed in a simpler form decoupling θ_0 and the x_i. For an implementation of the test statistic $\tilde{\lambda}(\mathbf{X})$ the value θ_0 of θ must be known, and a UMP test does not exist for this problem for $n > 1$.

One approach we might take in the above case is to use a **generalized likelihood ratio** (GLR) test, here obtained by using as the test statistic $\tilde{\lambda}(\mathbf{X})$ of (A-27) with θ_0 replaced by its **maximum likelihood** (ML) estimate under the alternative hypothesis H_1. This maximum-likelihood estimate $\hat{\theta}_{ML}$ is given implicitly as the solution of the equation

$$\sum_{i=1}^{n} s_i \operatorname{sgn}\left(x_i - \hat{\theta}_{ML}\, s_i\right) = 0 \qquad (A-32)$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1 & , \quad x > 0 \\ 0 & , \quad x = 0 \\ -1 & , \quad x < 0 \end{cases} \qquad (A-33)$$

provided that the solution turns out to be non-negative; otherwise, $\theta_{ML} = 0$. Thus the implementation of the GLR test is not simple. In addition, the distribution of the GLR test statistic under the null hypothesis is not easily obtained.

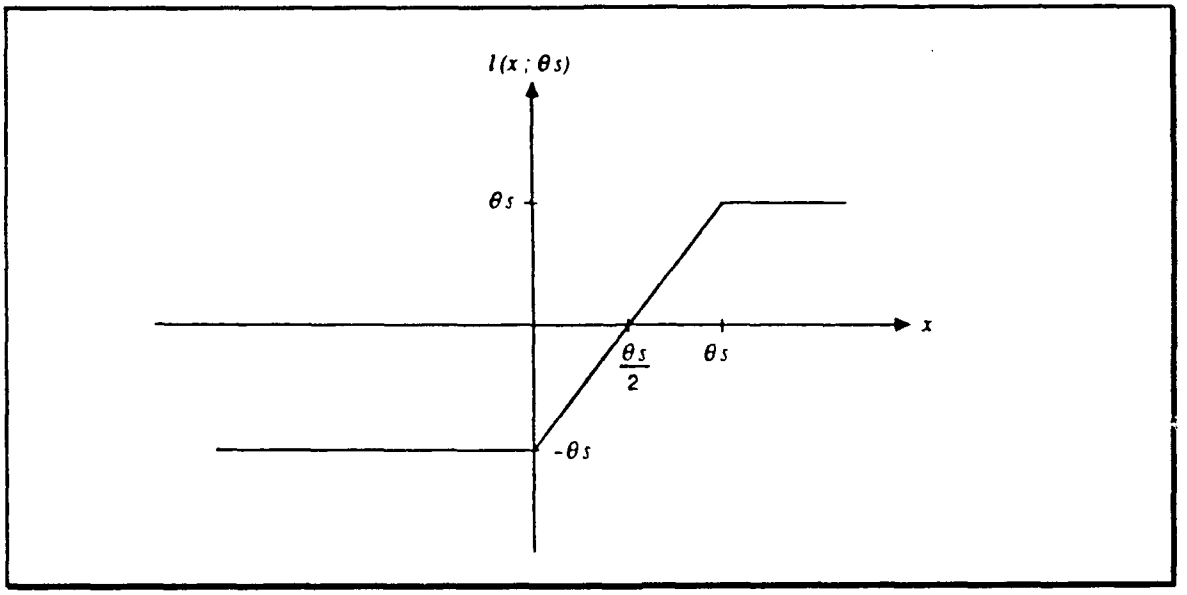


Figure A.2. The Characteristic $l(x;\theta_s)$ of Equation A-31

In the general case, for any noise density function P, the optimum detector for *given* $\theta = \theta_0 > 0$ under H_1 can be based on the test statistic

$$\lambda(\mathbf{X}) = \ln L(\mathbf{X})$$

$$= \sum_{i=1}^{n} \ln \frac{P(x_i - \theta_0 s_i)}{P(x_i)} \qquad (A-34)$$

which is of the form of $\tilde{\lambda}(\mathbf{X})$ of (A-30). But again, θ_0 must be specified and the detector will be optimum only for a signal with that amplitude. The GLR detector can be obtained if the ML estimate $\hat{\theta}_{ML}$ of θ can be found under the constraint that $\hat{\theta}_{ML}$ be non-negative. Once again, in general this will not lead to an easily implemented and easily analyzed system.

D. LOCALLY OPTIMUM DETECTORS

The above discussion shows that we have to search further in order to obtain reasonable schemes for detection of a known signal of unspecified amplitude in additive non-Gaussian noise. By a "reasonable" scheme we mean a detector that is practical to implement and relatively easy to analyze for performance, which should be acceptable for the anticipated range of input signal amplitudes. Fortunately, there is one performance criterion with respect to which it is possible to derive a simple and useful canonical structure for the optimum detector for our detection problem. This is the criterion of *local* detection power, and leads to detectors which are said to be **locally optimum**.

A **locally optimum** (LO) or locally most powerful detector is one which maximizes the *slope* of the detector power function at the **origin** ($\theta = 0$), from among the class of all detectors which have its false alarm probability. Let Δ_α be the class of detectors of size α for H_0 versus H_1. In our notation any detector D in Δ_α is based on a test function $\delta(\mathbf{X})$ for which

$$E\{\delta(\mathbf{X})|H_1\} = \alpha. \qquad (A-35)$$

Let $\mathcal{P}_d(\theta|D)$ be the power function of detector D, that is,

$$\mathcal{P}_d(\theta|D) = E\{\delta(\mathbf{X})|H_1\}. \qquad (A-36)$$

Formally, an LO detector D_{LO} of size α is a detector in Δ_α which satisfies

$$\max_{D\in\Delta_\alpha} \frac{d}{d\theta}\mathcal{P}_d(\theta|D)\Big|_{\theta=0} = \frac{d}{d\theta}\mathcal{P}_d(\theta|D_{LO})\Big|_{\theta=0}. \qquad (A-37)$$

It would be appropriate to use a locally optimum detector when one is interested primarily in detecting *weak* signals, for which θ under the alternative hypothesis H_1 remains close to zero. The idea is that an LO detector has a larger slope for its power function at $\theta = 0$ than any other detector D of the same size which is not an LO detector, and this will ensure that the power of the LO detector will be larger than that of the other detector at least for θ in some non-null interval $(0,\theta_{max})$, with θ_{max} depending on D. This is illustrated in Figure A.3. Note that if an LO detector is not unique, then one may be better than another for $\theta > 0$. There is good reason to be concerned primarily with weak-signal detection. It is the weak signal that one has the most difficulty in detecting, whereas most *ad hoc* detection schemes should perform adequately for strong signals; after all, the detection probability is upper bounded by unity.

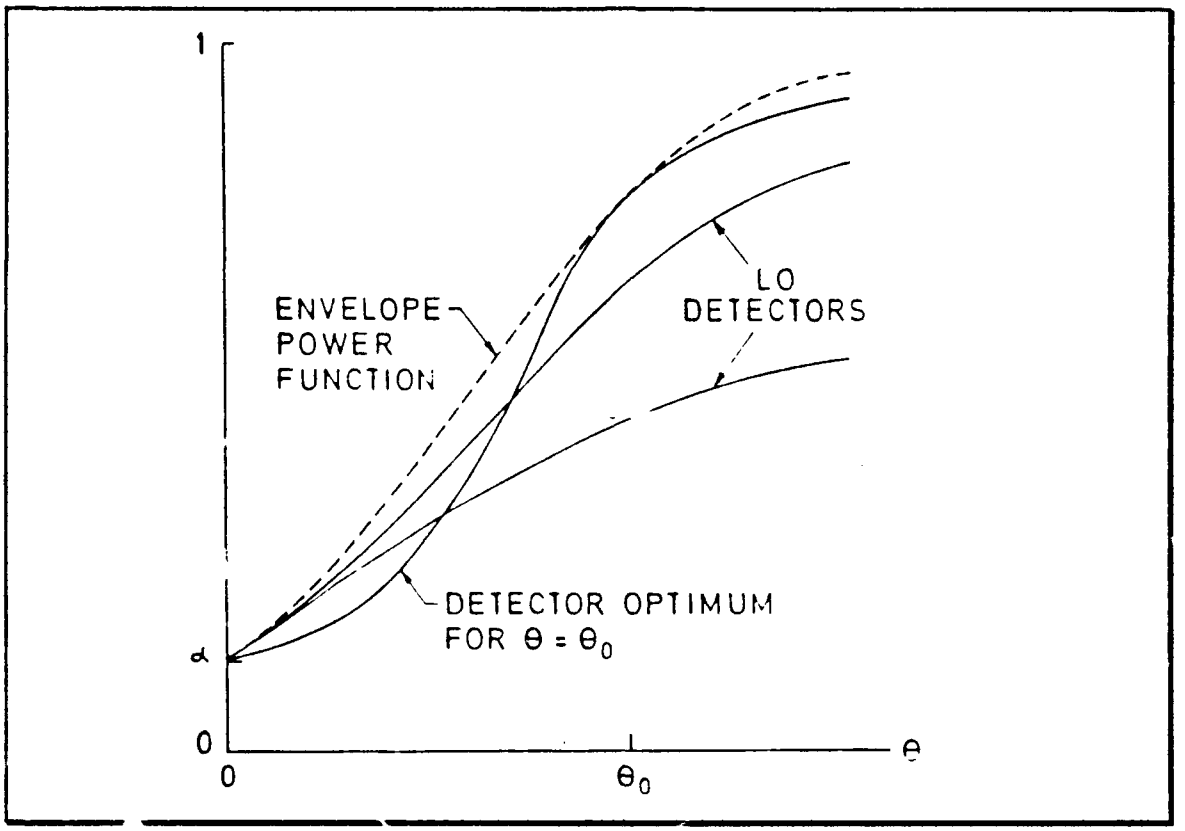


Figure A.3. Power Functions of Optimum and LO Detectors

To obtain explicitly the canonical form of the LO detector, we can apply the generalized Neyman-Pearson lemma of Section A.2. Now the power function of a detector D based on a test function $\delta(\mathbf{X})$ is

$$\mathcal{P}_d(\theta|D) = \int_{\Re^n} \delta(\mathbf{x}) \prod_{i=1}^{n} P(x_i - \theta s_i) d\mathbf{x} \qquad (A-38)$$

where the integration is over the n-dimensional Euclidean space $\Re^n$. The regularity Assumptions allow us to get

$$\left.\frac{d}{d\theta}\mathcal{P}_d(\theta|D)\right|_{\theta=0} = \int_{\Re^n} \delta(x)\left.\frac{d}{d\theta}\prod_{i=1}^{n} P(x_i - \theta s_i)\right|_{\theta=0} dx$$

$$= \int_{\Re^n} \delta(x)\left[\sum_{i=1}^{n} -s_i \frac{P'(x_i)}{P(x_i)}\right]\prod_{i=1}^{n} P(x_i)dx$$

$$= E\left\{\delta(\mathbf{X})\left[\sum_{i=1}^{n} -s_i \frac{P'(x_i)}{P'(x_i)}\right]\middle| H_0\right\} \qquad (A-39)$$

from this it follows, from the generalized Neyman-Pearson lemma, that a locally optimum detector $D_{\ell o}$ is based on the test statistic

$$\lambda_{\ell o}(\mathbf{X}) = -\sum_{i=1}^{n} s_i \frac{P'(x_i)}{P(x_i)}$$

$$= \sum_{i=1}^{n} s_i\, g_{Lo}(x_i) \qquad (A-40)$$

where $g_{\ell o}$ is the function defined by

$$g_{\ell o}(x) = -\frac{P'(x)}{P(x)}. \qquad (A-41)$$

Note that we may express $\lambda_{\ell o}(\mathbf{X})$ as

$$\lambda_{\ell o}(\mathbf{X}) = \sum_{i=1}^{n} \left.\frac{d}{d\theta} \ln P(x_i - \theta\, s_i)\right|_{\theta=0}$$

$$= \left.\frac{d}{d\theta}\sum_{i=1}^{n} \ln \frac{P(x_i - \theta\, s_i)}{P(x_i)}\right|_{\theta=0} \qquad (A-42)$$

from which the LO detector test statistic (multiplied by θ) is seen to be a first-order approximation of the optimum detector test statistic given by A-34.

For the double-exponential noise density of (A-23) we find that g_{ℓ_0} is given by

$$g_{\ell_0}(x) = a\,\mathrm{sgn}(x). \qquad (A-43)$$

Note that the *optimum* detector for $\theta = \theta_0$ in this case is based on the test statistic $\tilde{\lambda}(\mathbf{X})$ of (A-27). Similarly, for a zero mean Gaussian density with variance σ^2 we have

$$g_{\ell_0}(x) = \frac{x}{\sigma^2}. \qquad (A-44)$$