DATA-EFFICIENT QUICKEST CHANGE DETECTION

BY

TAPOSH BANERJEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Venugopal V. Veeravalli, Chair
Professor Pierre Moulin
Assistant Professor Alejandro Dominguez-Garcia
Assistant Professor Georgios Fellouris

ABSTRACT

In the classical problem of quickest change detection, a decision maker observes a sequence of random variables. At some point of time, the distribution of the random variables changes abruptly. The objective is to detect this change in distribution with minimum possible delay, subject to a constraint on the false alarm rate. In many applications of quickest change detection, changes are rare and there is a cost associated with taking observations or acquiring data. For such applications, the classical quickest change detection model is no longer applicable. In this dissertation we extend the classical formulations by adding an additional penalty on the cost of observations used before the change point. The objective is to find a causal on-off observation control policy and a stopping time, to minimize the detection delay, subject to constraints on the false alarm rate and the cost of observations used before the change point. We show that two-threshold generalizations of the classical single-threshold tests are asymptotically optimal for the proposed formulations. The nature of optimality is strong in the sense that the false alarm rates of the two-threshold tests are at least as good as the false alarm rates of their classical counterparts. Also, the delays of the two-threshold tests are within a constant of the delays of their classical counterparts. These results indicate that an arbitrary but fixed fraction of observations can be skipped before change without any loss in asymptotic performance. A detailed performance analysis of these algorithms is provided, and guidelines are given for the design of the proposed tests, on the basis of the performance analysis. An important result obtained through this analysis is that the two constraints, on the false alarm rate and the cost of observations used before the change, can be met independent of each other. Numerical studies of these two-threshold algorithms also reveal that they have good trade-off curves, and perform significantly better than the approach of fractional sampling, where classical single threshold tests are used and the constraint on the cost of observations

is met by skipping observations randomly. We first study the problem in Bayesian and minimax settings and then extend the results to more general quickest change detection models, namely, model with unknown post-change distribution, a sensor network model, and a multi-channel model.

To Mummy, Papa, Smruti, Didi, Sonit and Tirumal for their love, support, and sacrifices. To my grandparents Dida, Nani, Dadu and Nana, for dedicating their lives toward a better future for their families.

ACKNOWLEDGMENTS

I would first like to thank Prof. Veeravalli for believing in me and giving me the opportunity to come to Illinois and work with him. I also thank him for teaching me every possible aspect of graduate level research including developing a positive attitude toward difficult problems, being enthusiastic about obtained results, effective way of communicating the research output to the research community, and the most important, how to do research. I am grateful to him for giving me the freedom to explore and work on topics of my interest. I also thank him for being so approachable and caring of all his students.

I am grateful to my dissertation committee members, Prof. Pierre Moulin, Prof. Alejandro Dominguez-Garcia and Prof. Georgios Fellouris for taking the time from their super busy schedules, and agreeing to be on my dissertation committee and providing valuable feedback. I am also grateful to Prof. Alexander Tartakovsky and Prof. Douglas Jones for agreeing to be on my prelim committee. I would like to thank Prof. Alexander Tartakovsky for his help and guidance throughout my Ph.D. studies.

I take this opportunity to thank every member of the Coordinated Science Laboratory (CSL) for making CSL such an exceptional place to work; so professional and yet so homely. I would like to express my gratitude to the entire CSL support staff, especially Ms. Barbara Horner and Ms. Peggy Wells, for their help and kindness. I admire their time management skills, professional work ethics and enthusiasm toward their work. I would also like to thank the support staff in the ECE department, especially Ms. Laurie Fisher and Ms. Jan Progen, for their help and support. I thank Ms. Laurie Fisher for patiently answering each and every email that I have sent her in the last five years, and Ms. Jan Progen for carefully reading my dissertation and providing invaluable feedback.

I would also like to thank all the people responsible for the funds that I have

received in the last five years: my advisor for writing such great proposals, the National Science Foundation and other funding agencies for accepting the proposals, and finally, the U.S. tax payers who are indirectly contributing to the funds.

I have been very fortunate to be a part of the ECE department of the Indian Institute of Science (IISc), Bangalore. I thank all the faculty members for creating and maintaining such a great learning environment. I also thank the Tata group for creating such a great institution and the Government of India for funding it. I also thank the Government of India and the tax payers of India for the funds/subsidies because of which I could enjoy affordable yet quality primary and higher education all these years.

I would like to express my gratitude to Prof. Vinod Sharma and Prof. Anurag Kumar from IISc for giving me the opportunity to work with them. I would especially like to thank Prof. Vinod Sharma for being such a great mentor all these years. I would also like to thank Dr. Arzad Kherani for guiding me and mentoring me during my initial years in academic life.

I thank all my teachers for teaching me the fundamentals and all the great authors who wrote insightful books.

No acknowledgment for an achievement in professional life is complete without acknowledging the contributions of family and friends. And it must start with my parents. I cannot thank my parents enough for being so loving, caring, for being so patient with me all these years, and for everything else they have done for me. They have worked hard all their lives so that I and my elder sister can lead a happy life. I especially thank my mother for selflessly dedicating her life for our upbringing.

I could not have finished the dissertation without the love and support of my wife Smruti. Her love, care, friendship, wisdom, sweetness, innocence, enthusiasm, and the ability to enjoy every small moment in life has kept me afloat all these years. I apologize to her for not being able to spend quality time with her, and thank her for her big-heartedness.

I thank my elder sister Mrini for her love and affection and for all the wonderful memories from childhood. I thank Sonit for being such a great and sweet nephew, and Tirumal for being such a loving and caring brother-in-law. I thank all my relatives for giving me a life full of love and happy memories: my grandparents, Dida, Nani, Dadu and Nana; my uncles and aunts, Masi and Uncle, Mama and Maima, Choto Kaka and Kakima, Bodo

Kaka and Kakima, Pisi and Pisomoshai; my cousins, Dada and Mouli, Mitu Didi and Alok, Babli and Tublu, Shubho and Saurabh, Mithun and Rini, and Sriti; and last but not the least my nieces Prisha, Piu and Tia. I thank my in-laws Bou, Nana, Smita Didi, Himanshu Bhaiya and Kutu for their love and support all these years. I also thank Tirumal's parents, his sisters, and the rest of the family for being so loving and caring. I thank Sachin, Amit and Rakesh for their friendship, love and support all these years, and giving me the true taste of friendship. I also thank Rahul, Kaushik, and Pradeep for my memorable days in IISc. Finally, I thank all the people whose names I could not mention and who have helped me at some point of time.

I have tremendous love and respect for all the people who have helped me in my quest for a better life. I sincerely apologize to anyone who feels that my words could not capture their efforts and influence on my academic life.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 DATA-EFFICIENT BAYESIAN QUICKEST CHANGE DETECTION 9
2.1 Problem Formulation 10
2.2 Classical Bayes QCD 13
2.3 The DE-Shiryaev Algorithm 16
2.4 Derivation of the DE-Shiryaev Algorithm 18
2.5 Asymptotic Optimality of the DE-Shiryaev Algorithm 19
2.6 Numerical Results 33
2.7 Existing Literature 37

CHAPTER 3 DATA-EFFICIENT MINIMAX QUICKEST CHANGE DETECTION 38
3.1 Problem Formulation 40
3.2 The DE-CuSum Algorithm 43
3.3 Analysis and Design of the DE-CuSum Algorithm 46
3.4 Numerical Results 61
3.5 Proofs of Various Results 64

CHAPTER 4 DATA-EFFICIENT MINIMAX QUICKEST CHANGE DETECTION WITH COMPOSITE POST-CHANGE HYPOTHESIS 68
4.1 Problem Formulation 69
4.2 Classical QCD with Unknown Post-Change Distribution . . . 71
4.3 QCD with Observation Control ($\beta < 1$), θ Known 75
4.4 The GDECuSum Algorithm 77
4.5 Asymptotic Optimality of the GDECuSum Algorithm 80
4.6 Numerical Results 90
4.7 Discussion on the Least Favorable Distribution 91

CHAPTER 5 DATA-EFFICIENT QUICKEST CHANGE DETECTION IN SENSOR NETWORKS 94
5.1 Problem Formulation 96

5.2 Quickest Change Detection in Sensor Networks: Existing Literature . . . 99
5.3 The DE-All Algorithm . . . 102
5.4 Asymptotic Optimality of the DE-All Algorithm . . . 103
5.5 Data-Efficient Algorithms for Sensor Networks . . . 104
5.6 Numerical Results . . . 106
5.7 Proofs of Various Results . . . 107

CHAPTER 6 DATA-EFFICIENT QUICKEST CHANGE DETECTION IN MULTI-CHANNEL SYSTEMS . . . 119
6.1 Problem Formulation . . . 120
6.2 Data-Efficient Algorithms for Multi-Channel Systems . . . 123
6.3 Asymptotic Optimality of the DE-Censor-Max Algorithm . . . 126
6.4 Numerical Results . . . 130

CHAPTER 7 CONCLUSIONS AND FUTURE WORK . . . 133

REFERENCES . . . 137

CHAPTER 1

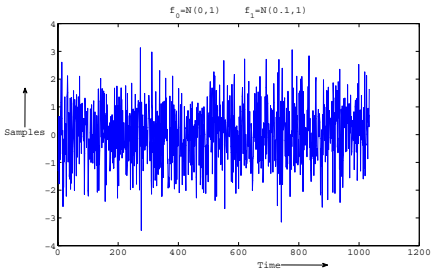
INTRODUCTION

In many engineering applications it is of interest to detect when a system being monitored abruptly moves from a normal state to an abnormal one. Applications include detection of the appearance of a sudden fault/stress in a system being monitored, e.g., bridges, historical monuments, power grids, bird/animal habitats, etc. Often in these applications the decision making has to be done in real time, by taking measurements sequentially. In statistics this detection problem is formulated within the framework of quickest change detection (QCD) [1], [2], [3], [4].

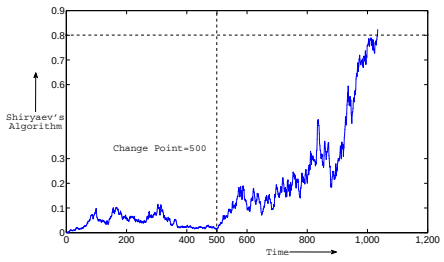
In the QCD problem, the objective is to detect an abrupt change in the distribution of a sequence of random variables. In the simplest of settings, the QCD problem is described as follows. A decision maker observes a sequence of random variables $\{X_n\}$, independent and identically distributed (i.i.d.) with a density function f_0. At some point of time γ, called the change point, the distribution of $\{X_n\}$ changes from f_0 to f_1, i.e., at and beyond γ, the random variables $\{X_n\}$ are i.i.d. with density f_1. The objective is to find a stopping time τ on the sequence $\{X_n\}$, i.e., a positive integer valued random variables such that $\mathbb{I}_{\{\tau=n\}}$ is a function of $\{X_1, \cdots, X_n\}$, $\forall n$. The stopping time τ is interpreted as the time at which a change is declared by the decision maker and the observation process $\{X_n\}$ is stopped and an alarm is raised. This stopping time has to be selected so as to detect the change in distribution of $\{X_n\}$ as quickly as possible, i.e., to minimize some version of the delay $\max\{0, \tau - \gamma\}$. The event $\{\tau < \gamma\}$ is called a false alarm and is not desirable. Hence, the delay has to be minimized subject to a constraint on the false alarm rate. The densities f_0 and f_1 may or may not be known. In the Bayesian setting of the problem the change point is modeled as a random variable Γ, and a prior distribution on Γ is assumed to be known. Otherwise the problem is studied in minimax settings, where the change point γ is modeled as an unknown constant. The quickest change

detection problem is also studied in various other settings: non-i.i.d. settings where the observations need not be i.i.d. conditioned on the change point, sensor network settings where the decision making is distributed, and the multi-channel settings where the change only affects a subset of a vector of observations, etc.; see [1], [2], and [3] for a review.

To motivate the need for quickest change detection algorithms, in Fig. 1.1a we plot a sample path of a stochastic sequence whose samples are distributed as $\mathcal{N}(0,1)$ before the change, and distributed as $\mathcal{N}(0.1,1)$ after the change. For illustration, we choose time slot 500 as the change point. As is evident from the figure, the change cannot be detected through visual inspection. In Fig. 1.1b, we plot the evolution of the *Shiryaev statistic* (to be discussed in Chapter 2), computed using the samples of Fig. 1.1a. As seen in Fig. 1.1b, the value of the Shiryaev statistic stays close to zero before the change point, and grows up to one after the change point. The change is detected at around the 1000th time slot by using a threshold of 0.8.



(a) Stochastic sequence with observations from $f_0 \sim \mathcal{N}(0,1)$ before the change (time slot 500), and with samples from $f_1 \sim \mathcal{N}(0.1,1)$ after the change.



(b) Evolution of the classical Shiryaev algorithm when applied to the samples given on the left. We see that the change is detected approximately at time slot 1000.

Figure 1.1: Detecting a change in the mean of a Gaussian random sequence.

For all the popular QCD formulations in the literature, the optimal stopping rule is similar to the Shiryaev algorithm in Fig. 1.1b. That is, the optimal stopping rule is a single threshold test, where a sequence of statistics is computed using the likelihood ratio of the observations, and a change is declared the first time the sequence of statistics crosses a threshold. The threshold is chosen to meet the constraint on the false alarm rate.

In this dissertation, we study the classical QCD problem with an additional constraint on the cost of observations used before the change point. The motivation for this study comes from the consideration of the following

engineering applications.

In many monitoring applications, for example infrastructure monitoring, environment monitoring, or habitat monitoring, especially of endangered species, surveillance is only possible through the use of inexpensive battery operated sensor nodes. This could be due to the high cost of employing a wired sensor network or a human observer, or the infeasibility of having a human intervention. For example in habitat monitoring of certain sea-birds as reported in [5], the very reason the birds chose the habitat was because of the absence of humans and predators around it. In these applications the sensors are typically deployed for long durations, possibility over months, and due to the constraint on energy, the most effective way to save energy at the sensors is to switch the sensor between on and off states. An energy-efficient quickest change detection algorithm can be employed here that can operate over months and trigger other more sophisticated and costly sensors, which are possibly power hungry, or more generally, trigger a larger part of the sensor network [6]. This change could be a fault in the structures in infrastructure monitoring [6], the arrival of the species to the habitat [5], etc.

In industrial quality control, statistical control charts are designed that can detect a sustained deviation of the industrial process from normal behavior [7]. Often there is a cost associated with acquiring the statistics for the control charts and it is of interest to consider designing *economic-statistical* control chart schemes [8, 7, 9]. The process control problem is fundamentally a quickest change detection problem, and it is therefore appropriate that economic-statistical schemes for process control are developed in this framework.

In most of the above mentioned or similar applications, changes are rare and quick detection is often required. So, ideally we would like to take as few observations as possible before change to reduce the observation cost, and skip as few as possible after change to minimize delay, while maintaining an acceptable probability of false alarm. In the literature on classical quickest change detection, only the trade-off between delay and false alarm is studied. Thus, while the cost of observations used after the change point is penalized, the cost of observations used before the change point is ignored. The goal in this dissertation is to develop a deeper understanding of the trade-off between delay, false alarm rate, and the cost of observation or information used before the change point, and to identify algorithms that have some

optimality property and are easy to design.

The dissertation is divided into five main chapters, Chapters 2-6. In each of the chapters, we consider a popular quickest change detection formulation from the literature. We then extend the classical formulation by including an additional constraint on the cost of observations used before the change point. We also introduce suitable metric that captures this cost in each setting. We seek control policies where causal on-off observation control is used in each time step to meet the constraint on the observation cost before change. The policy also includes a stopping time at which the change is declared. Since the policies are designed for efficient use of observations or data, we call this new problem data-efficient quickest change detection.

As discussed above, a typical algorithm from the classical quickest change detection literature is a single threshold test. In such a test all the observations are used for decision making and a sequence of statistics is computed using the likelihood ratio of the observations. A change is declared when the sequence of computed statistics is above a threshold. Popular examples include the Shiryaev algorithm [10] and the CuSum algorithm [11].

In this dissertation we propose two-threshold extensions of these classical single threshold tests; the DE-Shiryaev algorithm in Chapter 2 and the DE-CuSum algorithm in Chapter 3. In these two-threshold extensions that we propose there are two thresholds A and B, with $B < A$. Again, a sequence of statistics is computed using the likelihood ratio of the observations. As in the classical setting, a change is declared if the computed statistic is above the threshold A. If however the statistic is below A, then the observation in the next time step is taken only if the computed statistic is above the lower threshold B. When the statistic is below B, we also provide the recipe to update the statistic. In the Bayesian setting, it is updated using the prior on the change point. In non-Bayesian settings, the statistic is updated using a design parameter. The number of consecutive samples skipped then becomes a function of the undershoot of the statistics (which is also the likelihood ratio of the observations taken until that time).

More generally, we propose new data-efficient tests (see Chapters 2-6), where the feature of observation control is added to the classical tests. We provide performance analyses of these new tests, and using the performance analysis the parameters in the data-efficient algorithms can be designed. We then use the performance analysis to prove some optimality properties of

these new tests with respect to the new data-efficient formulations. The nature of the optimality proved is strong in the sense that the data-efficient tests have false alarm rates that are as good as the false alarm rates of their classical counterparts. Moreover, the average delay of these data-efficient tests is within a constant of the delay of their classical counterparts.

We also provide numerical and simulation results for these tests.

We now discuss the outline of the dissertation. In the classical literature the Bayesian formulation in the i.i.d. setting has provided valuable insights into the structure of the problem. These insights have played a key role in the extension of the results to other settings. Motivated by this in Chapter 2 we first study data-efficient quickest change detection in the Bayesian setting. In the rest of the chapters we use the insights obtained from Chapter 2 to extend the result from the Bayesian setting to other settings. We provide a brief overview of the context of each chapter.

1. Chapter 2, Bayesian setting: In the Bayesian setting the distribution of the change point is assumed to be known. In the classical Bayesian formulation studied in [10], the objective is to minimize the average detection delay subject to a constraint on the probability of false alarm. The optimal solution is a single threshold test popularly known as the Shiryaev test. In this chapter we extend the results from [10] by putting an additional constraint on the average number of observations used before the change point. We show that a two-threshold generalization of the Shiryaev test is asymptotically optimal for the proposed formulation as the probability of false alarm goes to zero. We call this two-threshold test the DE-Shiryaev algorithm.

2. Chapter 3, Minimax settings: If the distribution of the change point is not known then the classical quickest change detection problem is studied in the minimax settings of [12] and [13]. It is well know that the CuSum algorithm is asymptotically optimal for these formulations [11], [12], [14]. In this chapter we extend these minimax formulations by putting an additional constraint on the fraction of observations used before the change point. We propose a two-threshold generalization of the CuSum algorithm, which we term the DE-CuSum algorithm. We show that the DE-CuSum algorithm is asymptotically optimal for the extended formulations. The problem formulation, the structure

of the DE-CuSum algorithm, and the nature of its optimality, are all motivated by the Bayesian analysis provided in Chapter 2. The DE-CuSum algorithm plays a crucial rule in the rest of the dissertation.

3. Chapter 4, Minimax setting with composite post-change hypothesis: In this chapter we extend the results of Chapter 3 by allowing for the possibility that the post-change distribution is not known. In the classical formulations in the literature a generalized likelihood ratio test (GLRT) is asymptotically optimal under some conditions. We show that if the post-change family of distributions has a distribution that is least favorable in some sense, then it is possible to have efficient observation control in the GLRT test. Specifically, we propose an algorithm called the GDECuSum algorithm. In this algorithm observation control is implemented using the DE-CuSum algorithm applied to the least favorable distribution, and the GLRT statistic is computed using the available samples. We show that this test is asymptotically optimal.

4. Chapter 5, Sensor network setting: In this chapter we study data-efficient quickest change detection in a sensor network. Here, multiple sensors spread out in a geographical area are coordinated through a fusion center to detect the change. We propose data-efficient formulations for sensor networks and show that an algorithm called the DE-All algorithm, in which the DE-CuSum algorithm is used at each sensor, is asymptotically optimal.

5. Chapter 6, Multi-channel setting: In this chapter we consider the problem where there are multiple independent streams of observations and the change affects only a subset of streams. We propose two algorithms the DE-Censor-Max and the DE-Censor-Sum algorithms, and comment on their optimality properties. In both the algorithms, the DE-CuSum algorithm is applied to each independent stream.

6. Chapter 7, Conclusions and future work: In this chapter we conclude the dissertation and comment on future work.

Thus, a common theme in all the chapters is that the data-efficient tests are obtained by adding observation control to their classical counterparts. We will show that these data-efficient tests can be designed to meet the given

constraints on the false alarm rate and the observation cost, independent of each other. A rather surprising result in the dissertation is the nature of optimality of these algorithms. The performance analysis of various data-efficient algorithms will reveal that the performance of these data-efficient tests is as good as their classical counterpart. Thus, the optimality indicates that an arbitrary large but fixed fraction of observations can be skipped before change, without affecting the asymptotic performance.

In Table 1.1 we provide a glossary of symbols used in this dissertation.

Table 1.1: Glossary

Symbol	**Definition/Interpretation**
$o(1)$	$x = o(1)$ as $c \to c_0$, if $\forall \epsilon > 0$, $\exists \delta > 0$ s.t., $\lvert x \rvert \leq \epsilon$ if $\lvert c - c_0 \rvert < \delta$
$O(1)$	$x = O(1)$ as $c \to c_0$, if $\exists \epsilon > 0, \delta > 0$ s.t., $\lvert x \rvert \leq \epsilon$ if $\lvert c - c_0 \rvert < \delta$
$g(c) \sim h(c)$ as $c \to c_0$	$\lim_{c \to c_0} \frac{g(c)}{h(c)} = 1$ or $g(c) = h(c)(1 + o(1))$ as $c \to c_0$
$\mathbb{P}_n$ ($\mathbb{E}_n$)	Probability measure (expectation) when the change occurs at time n
$\mathbb{P}_\infty$ ($\mathbb{E}_\infty$)	Probability measure (expectation) when the change does not occur
ess sup X	$\inf\{K \in \mathbb{R} : \mathbb{P}(X > K) = 0\}$
$L(X)$	$\frac{f_1(X)}{f_0(X)}$
$\ell(X)$	$\log \frac{f_1(X)}{f_0(X)}$
$D(f_1 \parallel f_0)$	K-L Divergence between f_1 and f_0, defined as $\mathbb{E}_1 \left(\log \frac{f_1(X)}{f_0(X)} \right)$
$D(f_0 \parallel f_1)$	K-L Divergence between f_0 and f_1, defined as $\mathbb{E}_\infty \left(\log \frac{f_1(X)}{f_0(X)} \right)$
$(x)^+$	$\max\{x, 0\}$
$(x)^{h+}$	$\max\{x, -h\}$
$T(x)$	$\lceil (x)^{h+}/\mu \rceil$
S_n	$S_n = 1$ if X_n is used for decision making
Ψ	Policy for data-efficient quickest change detection $\{\tau, M_1, \cdots, M_\tau\}$
$\mathsf{ADD}(\Psi)$	$\sum_{n=0}^{\infty} \mathbb{P}(\Gamma = n)\ \mathbb{E}_n\left[(\tau - n)^+\right]$
$\mathsf{PFA}(\Psi)$	$\sum_{n=0}^{\infty} \mathbb{P}(\Gamma = n)\ \mathbb{P}_n(\tau < n)$
$\mathsf{FAR}(\Psi)$	$\frac{1}{\mathbb{E}_\infty[\tau]}$
$\mathsf{WADD}(\Psi)$	$\sup_{n \geq 1} \operatorname{ess\,sup} \mathbb{E}_n\left[(\tau - n)^+ \lvert I_{n-1}\right]$
$\mathsf{CADD}(\Psi)$	$\sup_{n \geq 1} \mathbb{E}_n[\tau - n \lvert \tau \geq n]$
$\mathsf{CPDC}(\Psi)$	$\limsup_n \frac{1}{n} \mathbb{E}_n \left[\sum_{k=1}^{n-1} S_k \middle\lvert \tau \geq n \right]$
$\mathsf{PDC}(\Psi)$	$\limsup_n \frac{1}{n} \mathbb{E}_n \left[\sum_{k=1}^{n-1} S_k \right]$
$\mathbb{I}_A$	Indicator function for event A
$\mathbb{E}[X; A]$	$\mathbb{E}[X \mathbb{I}_A]$

CHAPTER 2

DATA-EFFICIENT BAYESIAN QUICKEST CHANGE DETECTION

In this chapter we consider the quickest change detection problem in the Bayesian setting, where the change point is modeled as a random variable with a known distribution. In the classical Bayesian formulation studied by Shiryaev in [10], the objective is to detect a sudden change in the distribution of a sequence of random variables. This change has to be detected with minimum possible delay subject to a constraint on the probability of false alarm. In this chapter we extend Shiryaev's formulation by explicitly accounting for the cost of the observations used in the detection process. We capture the observation penalty (cost) through the average number of observations used before the change point, and allow for a dynamic control policy that determines whether or not a given observation is taken. The objective is to choose the observation control policy along with the stopping time, so that the average detection delay is minimized subject to constraints on the probability of false alarm and the observation cost.

For the classical formulation of Shiryaev, the optimal stopping rule is to compute the *a posteriori* probability that the change has already happened given the past observations, and stop the first time this probability is above a threshold. We will show in this chapter that a two-threshold generalization of the Shiryaev's test is optimal for a data-efficient formulation that we propose in this chapter. In fact, we will show that the performance of the two tests are asymptotically the same, as the probability of false alarm goes to zero. This result implies that one can skip an arbitrary fraction of samples before change, without sacrificing asymptotic performance.

2.1 Problem Formulation

A sequence of random variables $\{X_n\}$ is being observed. Initially, the random variables are i.i.d. with p.d.f. f_0. At some unknown point of time—denoted by γ and called the change point in the following—the density of the random variables changes to f_1. We denote by $\mathbb{P}_\gamma$ the underlying probability measure which governs such a sequence. We use $\mathbb{E}_\gamma$ to denote the expectation with respect to this probability measure. We use $\mathbb{P}_\infty$ ($\mathbb{E}_\infty$) to denote the probability measure (expectation) when the change never occurs (i.e., the random variable X_n has p.d.f. f_0, $\forall n$). Both f_0 and f_1 are known but the change point γ is unknown. We wish to detect this change in distribution as quickly as possible subject to a constraint on the false alarm rate.

In this chapter we consider the Bayesian version of the problem where the change point γ is modeled as a random variable Γ with a known prior distribution. It is further assumed that the $\Gamma \sim \text{Geom}(\rho)$, i.e., for $0 < \rho < 1$

$$\pi_n = \mathbb{P}\{\Gamma = n\} = \rho(1-\rho)^{n-1}\, \mathbb{I}_{\{n \geq 1\}}, \quad \pi_0 = 0, \tag{2.1}$$

where $\mathbb{I}_F$ represents the indicator of the event F. We use $\mathbb{P}_\pi$ to denote for every event A in the underlying σ-algebra

$$\mathbb{P}_\pi(A) = \sum_{\gamma=1}^{\infty} \pi_\gamma \mathbb{P}_\gamma(A).$$

We also use $\mathbb{E}_\pi$ to denote the expectation with respect to $\mathbb{P}_\pi$. For brevity of notation, we will suppress the subscript π when using $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$.

In the classical QCD formulation studied by Shiryaev in [10], the objective is to find a stopping time τ on the sequence $\{X_n\}$ (a positive integer valued random variable such that $\mathbb{I}_{\{\tau=n\}}$ is a measurable function of $X_1, \cdots, X_n$), so as to minimize a metric on the average delay, subject to a constraint on a metric on the false alarm rate. The delay metric used in [10] is the following average detection delay (ADD) metric:

$$\mathsf{ADD}(\tau) = \mathbb{E}[(\tau - \Gamma)^+]. \tag{2.2}$$

To capture the false alarm rate the metric used is the probability of false

alarm (PFA):

$$\mathsf{PFA}(\tau) = \mathbb{P}(\tau < \Gamma). \tag{2.3}$$

Thus, in the Shiryaev's formulation, the objective is to find a τ so as to minimize ADD subject to a constraint on the PFA.

In many applications the change occurs rarely, corresponding to a large γ. As a result, we also wish to control the number of observations used for decision making before γ. We are interested in control policies involving causal three-fold decision making at each time step. Specifically, based on the information available at time n, a decision has to be made whether to declare a change or to continue taking observations. If the decision is to continue, then a decision has to be made whether to use or skip the next observation for decision making.

Mathematically, let S_n denote the indicator random variable which is 1 if X_n is used for decision making. That is

$$S_n = \begin{cases} 1 & \text{if } X_n \text{ used for decision making} \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

The information available at time n is denoted by

$$\mathcal{I}_n = \{X_1^{(S_1)}, \cdots, X_n^{(S_n)}\},$$

where $X_k^{(S_k)} = X_k$ if $S_k = 1$, else X_k is absent from $\mathcal{I}_n$, and

$$S_n = \phi_n(\mathcal{I}_{n-1}).$$

Here, ϕ_n denotes the control map. Let τ be a stopping time for the sequence $\{\mathcal{I}_n\}$. A control policy is the collection

$$\Psi = \{\tau, \phi_1, \cdots, \phi_\tau\}.$$

We seek policies of type Ψ to minimize ADD subject to a constraint on the PFA and a constraint on the average number of observations taken before change. To capture the latter, we now propose a new metric, the average

number of observations (ANO):

$$\mathsf{ANO}(\Psi) = \mathbb{E}\left[\sum_{n=1}^{\tau\wedge(\Gamma-1)} S_n\right]. \tag{2.5}$$

Thus, the metric ANO captures the average number of observations used before the change point Γ. We note that for any Ψ, $\mathsf{ANO}(\Psi) \leq \mathbb{E}[\Gamma - 1]$.

The data-efficient extension of the classical Bayesian problem of Shiryaev that we study in this chapter is:

Problem 2.1.1.

$$\begin{aligned} \underset{\Psi}{\textit{minimize}} \quad & \mathsf{ADD}(\Psi), \\ \textit{subject to} \quad & \mathsf{PFA}(\Psi) \leq \alpha, \\ \textit{and} \quad & \mathsf{ANO}(\Psi) \leq \beta. \end{aligned} \tag{2.6}$$

Here, α and β, with $0 \leq \alpha \leq 1$ and $0 \leq \beta$, are given constraints.

When $\beta \geq \mathbb{E}[\Gamma]-1$, Problem 2.1.1 reduces to the classical Bayesian quickest change detection problem from [10].

In Section 2.2, we review the solution to the classical formulation of Shiryaev, which in the following we call the Shiryaev's test/algorithm. We also review the performance analysis of the Shiryaev's test as studied in [15]. We will see that the Shiryaev's algorithm cannot be a solution to Problem 2.1.1, especially for small β. Hence, in this chapter we propose a two-threshold generalization of the Shiryaev algorithm and show that it is asymptotically optimal for the proposed formulation, for each β, as $\alpha \to 0$. We call that algorithm the data-efficient Shiryaev (DE-Shiryaev) algorithm. In Section 2.3 we propose the algorithm and discuss its evolution and properties. In Section 2.4 we provide a dynamic programming based justification for the algorithm. In Section 2.5 we provide a detailed performance analysis of the DE-Shiryaev algorithm and prove its asymptotic optimality by relating its performance to that of the Shiryaev algorithm. In Section 2.6 we compare the performance of the DE-Shiryaev algorithm with the approach of fractional sampling, in which the Shiryaev algorithm is used, and the constraint on the cost of observations used before the change is met by skipping samples randomly, independent of the observation process. We will show that the

DE-Shiryaev algorithm provides significant gain in performance as compared to the approach of fractional sampling.

We will assume throughout that the moments of $\log[f_1(X)/f_0(X)]$ under both $\mathbb{P}_1$ and $\mathbb{P}_\infty$, up to the second order, are finite and positive.

2.2 Classical Bayes QCD

The optimal stopping rule for Shiryaev's problem is

$$\tau_{\mathrm{s}} = \inf\{n \geq 1 : \mathbb{P}(\Gamma \leq n | X_1, \cdots, X_n) > A\}, \tag{2.7}$$

where the threshold $A < 1$, and is chosen to meet the constraint on α with equality. Let $p_n^{\mathrm{s}} = \mathbb{P}(\Gamma \leq n | X_1, \cdots, X_n)$. Then the probability p_n^{s} can be updated using the following recursions:

$$\begin{aligned} p_0^{\mathrm{s}} &= 0 \\ p_n^{\mathrm{s}} &= \frac{\tilde{p}_{n-1}^{\mathrm{s}} L(X_n)}{\tilde{p}_{n-1}^{\mathrm{s}} L(X_n) + (1 - \tilde{p}_{n-1}^{\mathrm{s}})}, \end{aligned} \tag{2.8}$$

with $\tilde{p}_n^{\mathrm{s}} = p_n^{\mathrm{s}} + (1 - p_n^{\mathrm{s}})\rho$ and $L(X_n) = f_1(X_n)/f_0(X_n)$.

Thus, in the Shiryaev algorithm, all the samples are taken, i.e., $S_n = 1$, $\forall n$, and a change is declared the first time the *a posteriori* probability that the change has already happened, given all the observations until that time, crosses the threshold A.

We now review the performance analysis of the Shiryaev algorithm from [15]. We first prove a simple but important result that will be used in the analysis of both the Shiryaev as well as the DE-Shiryaev algorithms.

Lemma 2.2.1. *For any policy Ψ such that $\tau < \infty$ a.s., we have*

$$\mathbb{P}(\tau < \Gamma) = \mathbb{E}[1 - p_\tau]. \tag{2.9}$$

Proof. By the law of iterated expectation and because τ is a stopping time

w.r.t. $\{\mathcal{I}_n\}$

$$\begin{aligned}\mathbb{P}(\tau < \Gamma) = \mathbb{E}[\mathbb{I}_{\{\tau<\Gamma\}}] &= \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{I}_{\{\tau<\Gamma\}}\ \mathbb{I}_{\{\tau=n\}}] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{E}[\mathbb{I}_{\{n<\Gamma\}}\ \mathbb{I}_{\{\tau=n\}}|\mathcal{I}_n]] \\ &= \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{I}_{\{\tau=n\}}\mathbb{E}[\mathbb{I}_{\{n<\Gamma\}}|\mathcal{I}_n]] \\ &= \sum_{n=1}^{\infty} \mathbb{E}[\mathbb{I}_{\{\tau=n\}}(1-p_n)] = \mathbb{E}[1-p_\tau].\end{aligned}$$

□

One way to use (2.9) is to show that $\mathbb{E}[\tau] < \infty$. This in turn can be shown by using the following inequality:

$$\mathbb{E}[\tau] \leq \mathbb{E}[\tau - \Gamma|\tau \geq \Gamma] + \frac{1}{\rho}. \tag{2.10}$$

This is true because

$$\begin{aligned}\mathbb{E}[\tau] &= \mathbb{E}[\tau - \Gamma + \Gamma] = \mathbb{E}[\tau - \Gamma] + \mathbb{E}[\Gamma] \\ &= \mathbb{E}[\tau - \Gamma] + \frac{1}{\rho} \\ &\leq \mathbb{E}[\tau - \Gamma|\tau \geq \Gamma]\mathbb{P}(\tau \geq \Gamma) + \frac{1}{\rho} \\ &\leq \mathbb{E}[\tau - \Gamma|\tau \geq \Gamma] + \frac{1}{\rho},\end{aligned} \tag{2.11}$$

where the last two equalities are true because $\mathbb{E}[\tau - \Gamma|\tau < \Gamma] < 0$, and $\mathbb{P}(\tau \geq \Gamma) \leq 1$. Thus, if $\mathbb{E}[\tau - \Gamma|\tau \geq \Gamma] < \infty$, then $\mathbb{E}[\tau] < \infty$, and Lemma 2.2.1 is applicable. This observation will be used in the following.

We next show that $\mathbb{E}_1[\tau_{\mathrm{s}}] < \infty$ and provide an asymptotic expression. This will be used to show that $\mathbb{E}[\tau_{\mathrm{s}} - \Gamma|\tau_{\mathrm{s}} \geq \Gamma] < \infty$.

Let

$$Z_n^{\mathrm{s}} = \log \frac{p_n^{\mathrm{s}}}{(1-p_n^{\mathrm{s}})\ \rho}.$$

It is easy to see that there is a one-to-one mapping between Z_n^{s} and p_n^{s}. Hence, if $a = \log \frac{A}{(1-A)\rho}$, then

$$\tau_{\mathrm{s}} = \inf\{n \geq 1 : Z_n^{\mathrm{s}} > a\}.$$

Lemma 2.2.2 ([15]). *For any a as defined above,*

$$\mathbb{E}_1[\tau_{\mathrm{S}}] < \infty. \tag{2.12}$$

Further,

$$\mathbb{E}_1[\tau_{\mathrm{S}}] \sim \frac{\log(a)}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)) \text{ as } a \to \infty. \tag{2.13}$$

Proof. Note that

$$p_n^{\mathrm{s}} = \frac{\sum_{k=1}^{n}(1-\rho)^{k-1}\rho\prod_{i=k}^{n}L(X_i)}{\sum_{k=1}^{n}(1-\rho)^{k-1}\rho\prod_{i=k}^{n}L(X_i) + (1-\rho)^n}.$$

Hence, the statistic Z_n^{s} can be written as

$$Z_n^{\mathrm{s}} = Y_n + n|\log(1-\rho)| + \log\left(1 + \sum_{k=1}^{n-1}(1-\rho)^k e^{-Y_k}\right), \tag{2.14}$$

where $Y_n = \sum_{k=1}^{n}\log L(X_k)$. It is shown in [15] that the rightmost term in (2.14) is *slowly changing* (see [16]). Thus the statistic Z_n^{s} can be written as the sum of a random walk $Y_n + n|\log(1-\rho)|$ and a slowly changing term. Also, the slowly changing term is clearly positive. Hence, by Theorem 4.4 of [16], we have

$$\mathbb{E}_1[\tau_{\mathrm{S}}] \sim \frac{|\log(a)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)) \text{ as } a \to \infty.$$

This implicitly shows that $\mathbb{E}_1[\tau_{\mathrm{S}}] < \infty$ for a fixed $a = \log\frac{A}{(1-A)\rho}$. □

We thus have the following result for the Shiryaev algorithm.

Lemma 2.2.3 ([15]). *Setting $A_\alpha = 1 - \alpha$ ensures that*

$$\mathbb{P}(\tau_{\mathrm{S}} < \Gamma) = \mathbb{E}[1 - p^s_{\tau_{\mathrm{S}}}] \le 1 - A_\alpha \le \alpha, \tag{2.15}$$

and

$$\mathbb{E}_1[\tau_{\mathrm{S}}] \sim \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)) \text{ as } \alpha \to 0. \tag{2.16}$$

Proof. By sample-pathwise arguments, it is easy to see that for the Shiryaev

algorithm,

$$\mathbb{E}[\tau_{\rm S} - \Gamma | \tau_{\rm S} \geq \Gamma] \leq \mathbb{E}_1[\tau_{\rm S}].$$

Now $\mathbb{E}_1[\tau_{\rm S}] < \infty$ from Lemma 2.2.2. It thus follows from (2.11) that $\mathbb{E}[\tau_{\rm S}] < \infty$. This in turn implies that $\tau_{\rm S} < \infty$ a.s. Hence, by Lemma 2.2.1, since $p^{\rm s}_{\tau_{\rm S}} > A$ by definition, we have

$$\mathbb{P}(\tau_{\rm S} < \Gamma) = \mathbb{E}[1 - p^{\rm s}_{\tau_{\rm S}}] \leq 1 - A_\alpha \leq \alpha.$$

The delay result follows by substituting the value of $a = \log \frac{A_\alpha}{(1-A_\alpha)\rho}$ in (2.13). □

We note that

$$\mathsf{ADD}(\tau_{\rm S}) = \mathbb{E}[(\tau_{\rm S} - \Gamma)^+] = \mathbb{E}[\tau_{\rm S} - \Gamma | \tau_{\rm S} \geq \Gamma]\mathbb{P}(\tau_{\rm S} \geq \Gamma) \leq \mathbb{E}_1[\tau_{\rm S}].$$

Thus, the expression on the right-hand side of (2.16) gives an upper bound on the asymptotic performance of the Shiryaev algorithm. It is in fact shown in [15] that $\frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}$ is also a lower bound on the $\mathsf{ADD}(\tau)$ of any stopping rule τ, satisfying $\mathsf{PFA}(\tau) \leq \alpha$, as $\alpha \to 0$. Thus, $\mathsf{ADD}(\tau_{\rm S}) \sim \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}$, as $\alpha \to 0$. We state this as a theorem.

Theorem 2.2.1 ([15])**.** *As $\alpha \to 0$,*

$$\inf_{\tau:\mathsf{PFA}(\tau)\leq\alpha} \mathsf{ADD}(\tau) \sim \mathsf{ADD}(\tau_{\rm S}) \sim \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)). \quad (2.17)$$

But note that when α is small, $\tau_{\rm S} \geq \Gamma$ with high probability, and $\mathsf{ANO}(\tau_{\rm S}) \approx \mathbb{E}[\Gamma] - 1$. Hence, the Shiryaev algorithm cannot be a solution to the Problem 2.1.1 in (2.6), if β is small. We will show in Section 2.5 that $\frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}$ is also the first-order asymptotic ADD of the DE-Shiryaev algorithm, for any fixed β, as $\alpha \to 0$. Thus, in this sense, the DE-Shiryaev algorithm is asymptotically optimal, for each fixed β, as $\alpha \to 0$.

2.3 The DE-Shiryaev Algorithm

We now describe the DE-Shiryaev algorithm. Define,

$$p_n = \mathbb{P}\left(\Gamma \leq n \mid \mathcal{I}_n\right).$$

Algorithm 2.3.1 (DE-Shiryaev: Ψ_{D})**.** *Start with* $p_0 = 0$ *and use the following control, with* $0 \leq B < A$*, for* $n \geq 1$:

$$S_n = \phi_n(p_{n-1}) = \begin{cases} 0 & \text{if } p_{n-1} < B \\ 1 & \text{if } p_{n-1} \geq B \end{cases}, \tag{2.18}$$

$$\tau_{\text{D}} = \inf\{n \geq 1 : p_n > A\}.$$

The probability p_n *is updated using the following recursions*:

$$p_n = \begin{cases} \tilde{p}_{n-1} & \text{if } S_n = 0 \\ \frac{\tilde{p}_{n-1} L(X_n)}{\tilde{p}_{n-1} L(X_n) + (1 - \tilde{p}_{n-1})} & \text{if } S_n = 1 \end{cases}$$

with $\tilde{p}_n = p_n + (1 - p_n)\rho$ *and* $L(X_n) = f_1(X_n)/f_0(X_n)$.

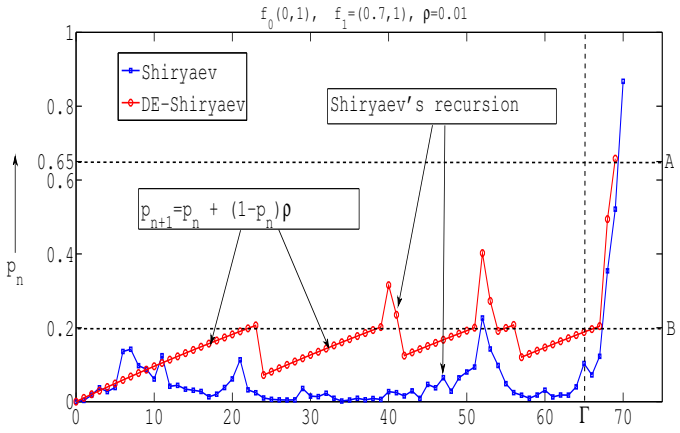


Figure 2.1: Typical evolution of the Shiryaev and the DE-Shiryaev algorithms, applied to the same set of observations, with thresholds $A = 0.65$ and $B = 0.2$.

With $B = 0$ the DE-Shiryaev algorithm reduces to the Shiryaev algorithm. When $B > 0$, the statistic evolves in a manner similar to the Shiryaev statistic as long as the statistic is above B. Thus, a change is declared when $p_n > A$. However, when p_n goes below B, p_n is updated using the prior on the change point ρ (p_n increases monotonically in this regime), and observations are skipped as long as p_n is below B. Thus, in the DE-Shiryaev algorithm observations are taken only if the *a posteriori* probability that the change has already occurred given the available information $\mathcal{I}_n$, exceeds a threshold B. A change is declared only if the probability is above another threshold

$A > B$. Since $p_0 = 0 < B$, few initial observations are skipped even before the observation process begins. Note that except for these initial skipped observations, the number of consecutive observations skipped at any time is a function of the undershoot of the statistic p_n (which is a function of the likelihood ratio of the observations), when it goes below B, and the geometric parameter ρ. In Fig. 2.1 we have plotted typical evolution of the DE-Shiryaev algorithm and the Shiryaev algorithm, applied to the same set of samples.

The DE-Shiryaev algorithm has the following interesting interpretation. Let

$$t(B) = \inf\{n \geq 1 : p_n > B\}. \tag{2.19}$$

At and beyond $t(B)$, whenever p_n crosses B from below, it does so with an overshoot that is bounded by ρ. This is because

$$p_{n+1} - p_n = (1 - p_n)\rho \leq \rho.$$

For small values of ρ, this overshoot is essentially zero, and the evolution of p_n is roughly statistically independent of its past evolution. Thus, beyond $t(B)$, the evolution of p_n can be seen as a sequence of *two-sided* statistically independent tests, each two-sided test being a test for sequential hypothesis testing between "H_0 = pre-change", and "H_1 = post-change". If the decision in the two-sided test is H_0, then observations are skipped depending on the likelihood ratio of the observations (the undershoot), and the two-sided test is repeated on the observations taken beyond the skipped observations. The change is declared the first time the decision in a two-sided test is H_1.

2.4 Derivation of the DE-Shiryaev Algorithm

The motivation for this algorithm comes from the fact that p_n is a sufficient statistics for a Lagrangian relaxation of Problem (2.1.1). This relaxed problem can be studied using dynamic programming, and numerical studies of the resulting Bellman equation show that the DE-Shiryaev algorithm is optimal for a wide choice of system parameters; see [17] for details. For an analytical justification see Section 2.5.

2.5 Asymptotic Optimality of the DE-Shiryaev Algorithm

In this section we prove the asymptotic optimality of the DE-Shiryaev algorithm. We note that from (2.11) we have

$$\mathbb{E}[\tau_{\text{D}}] \leq \mathbb{E}[\tau_{\text{D}} - \Gamma | \tau_{\text{D}} \geq \Gamma] + \frac{1}{\rho}. \tag{2.20}$$

Thus, if we can show that the right-hand side of the previous equation is finite then we can invoke Lemma 2.2.1 for the PFA result. We now obtain an upper bound on $\mathbb{E}[\tau_{\text{D}} - \Gamma | \tau_{\text{D}} \geq \Gamma]$. This is, in fact, a major result of this chapter.

Theorem 2.5.1. *For any fixed thresholds A and B,*

$$\mathbb{E}[\tau_{\text{D}} - \Gamma | \tau_{\text{D}} \geq \Gamma] \ \leq \ \mathbb{E}_1[\tau_{\text{S}}] + K_{DS}, \tag{2.21}$$

where K_{DS} is a constant that is a function of ρ, f_0, f_1 and B, but is not a function of the threshold A.

Thus, the conditional delay of the DE-Shiryaev algorithm is within a constant of the delay of the Shiryaev algorithm when the change occurs at time 1. We postpone the proof of Theorem 2.5.1 until the end of this section and study its implications.

Now, from Theorem 2.5.1 we have

$$\begin{aligned} \mathbb{E}[\tau_{\text{D}}] &\leq \mathbb{E}[\tau_{\text{D}} - \Gamma | \tau_{\text{D}} \geq \Gamma] + \frac{1}{\rho} \\ &\leq \mathbb{E}_1[\tau_{\text{S}}] + K_{\text{DS}} + \frac{1}{\rho} \\ &< \infty. \end{aligned} \tag{2.22}$$

Thus, $\tau_{\text{D}} < \infty$ a.s., and the Lemma 2.2.1 is applicable. Combining this with (2.13) we have the following corollary.

Corollary 2.5.1.1. *For any fixed B, if $A = 1 - \alpha$, then*

$$\text{PFA}(\tau_{\text{D}}) = \mathbb{E}[1 - p_{\tau_{\text{D}}}] \leq 1 - A \leq \alpha. \tag{2.23}$$

Further as $\alpha \to 0$*,*

$$\mathsf{ADD}(\tau_{\mathrm{D}}) \leq \mathbb{E}[\tau_{\mathrm{D}} - \Gamma | \tau_{\mathrm{D}} \geq \Gamma] \leq \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)). \quad (2.24)$$

We remark that even though the result in (2.24) shows that the performance of the DE-Shiryaev algorithm is asymptotically the same as that of the Shiryaev algorithm, it does not necessarily prove the asymptotic optimality of the algorithm w.r.t. Problem 2.1.1. This is because unless we choose the threshold B carefully, the ANO constraint can be violated, as $A \to 1$ or $\alpha \to 0$. The fact that such a B exists is proved in Lemma 2.5.1. For that recall from (2.19)

$$t(B) = \inf\{n \geq 1 : p_n > B\}.$$

Lemma 2.5.1. *There exists a* B_β *such that* $\forall A > B_\beta$*,*

$$\mathsf{ANO}(\tau_{\mathrm{D}}) \leq \beta.$$

Proof. We note that

$$\begin{aligned}
\mathsf{ANO}(\tau_{\mathrm{D}}) = \mathbb{E}\left[\sum_{n=1}^{\tau \wedge (\Gamma-1)} S_n\right] &= \mathbb{P}(\Gamma > t(B))\, \mathbb{E}\left[\sum_{n=1}^{\tau \wedge (\Gamma-1)} S_n \,\Big|\, \Gamma > t(B)\right] \\
&\leq \mathbb{P}(\Gamma > t(B))\, \mathbb{E}\left[\sum_{n=1}^{\Gamma} S_n \,\Big|\, \Gamma > t(B)\right] \\
&\leq \mathbb{P}(\Gamma > t(B))\, \mathbb{E}\left[\Gamma \,\Big|\, \Gamma > t(B)\right] \\
&= \mathbb{E}\left[\Gamma\, \mathbb{I}_{\{\Gamma > t(B)\}}\right].
\end{aligned} \quad (2.25)$$

Thus, by selecting $B = B_\beta$ such that $\mathbb{E}\left[\Gamma \mathbb{I}_{\{\Gamma > t(B)\}}\right] < \beta$, we complete the proof. □

We thus have the following optimality result due to Theorem 2.2.1.

Theorem 2.5.2. *If* $B = B_\beta$ *and* $A = A_\alpha = 1 - \alpha$*, then*

$$\begin{aligned}
&\mathsf{ANO}(\tau_{\mathrm{D}}) \leq \beta, \text{ for } \alpha \text{ small enough,} \\
&\mathsf{PFA}(\tau_{\mathrm{D}}) \leq \alpha, \text{ and} \\
&\mathsf{ADD}(\tau_{\mathrm{D}}) \sim \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)), \text{ as } \alpha \to 0\ .
\end{aligned} \quad (2.26)$$

Thus, the DE-Shiryaev algorithm is first-order asymptotically optimal for Problem 2.1.1, for each fixed β, as $\alpha \to 0$. We note that Theorem 2.5.2 implies that for α small enough, the two constraints on the PFA and the ANO can be met independent of each other.

We now provide the proof of Theorem 2.5.1.

Proof of Theorem 2.5.1. Recall that we wish to prove (see (2.21))

$$\mathbb{E}[\tau_{\mathrm{D}} - \Gamma | \tau_{\mathrm{D}} \geq \Gamma] \;\leq\; \mathbb{E}_1[\tau_{\mathrm{S}}] + K_{\mathrm{DS}}. \tag{2.27}$$

We first note that

$$\begin{aligned}
\mathbb{E}[\tau_{\mathrm{D}} - \Gamma | \tau_{\mathrm{D}} \geq \Gamma] &= \sum_{n=1}^{\infty} \mathbb{E}[(\tau_{\mathrm{D}} - \Gamma)\, \mathbb{I}_{\{\Gamma = n\}} | \tau_{\mathrm{D}} \geq \Gamma] \\
&= \sum_{n=1}^{\infty} \mathbb{E}[\tau_{\mathrm{D}} - n | \mathbb{I}_{\{\Gamma = n\}} \mathbb{I}_{\{\tau_{\mathrm{D}} \geq \Gamma\}}] \mathbb{P}(\Gamma = n | \tau_{\mathrm{D}} \geq \Gamma) \qquad (2.28) \\
&= \sum_{n=1}^{\infty} \mathbb{E}_n[\tau_{\mathrm{D}} - n | \tau_{\mathrm{D}} \geq n] \mathbb{P}(\Gamma = n | \tau_{\mathrm{D}} \geq \Gamma).
\end{aligned}$$

Thus, it is enough to get the bound specified in (2.27) on $\mathbb{E}_n[\tau_{\mathrm{D}} - n | \tau_{\mathrm{D}} \geq n]$. Toward this end we obtain a bound on $\mathbb{E}_n[(\tau_{\mathrm{D}} - n)^+ | \mathcal{I}_{n-1}]$. That is, we show that

$$\mathbb{E}_n[(\tau_{\mathrm{D}} - n)^+ | \mathcal{I}_{n-1}] \leq \mathbb{E}_1[\tau_{\mathrm{S}}] + K_{\mathrm{DS}}, \tag{2.29}$$

where K_{DS} is a constant not a function of the threshold A, change time n, and the conditioning $\mathcal{I}_{n-1}$.

Let

$$\tau_{\mathrm{D}}(p) = \inf\{n \geq 1 : p_n > A; p_0 = p\}$$

be the time for the DE-Shiryaev statistic to cross the threshold A starting at $p_0 = p$. Let $\mathcal{I}_{n-1} = i_{n-1}$ be such that $p_{n-1} = p \in [B, A)$. Then,

$$\mathbb{E}_n[\tau_{\mathrm{D}} - n | \mathcal{I}_{n-1} = i_{n-1}] = \mathbb{E}_1[\tau_{\mathrm{D}}(p)] - 1.$$

Note that the sojourn of the statistic p_n to A may include alternate sojourns of the statistic below and above the threshold B; see Fig. 2.1. And the time to hit A can be written as the sum of such times. Motivated by this we define

a set of new variables. Let

$$\tau_1(p) = \inf\{n \geq 1 : p_n > A \text{ or } p_n < B; \text{ with } p_0 = p\}.$$

This is the first time for the DE-Shiryaev statistic, starting at $p_0 = p$, to either hit A or go below B. On paths over which $p_{\tau_1} < B$, we know that a number of consecutive samples are skipped depending on the undershoot of the observations. Let $t_1(p)$ be the number of consecutive samples skipped after $\tau_1(p)$ on such paths. On such paths again, let

$$\tau_2(p) = \inf\{n > \tau_1(p) + t_1(p) : p_n > A \text{ or } p_n < B\}.$$

Thus, on paths such that $p_{\tau_1} < B$, after the times $\tau_1(p)$ and the number of skipped samples $t_1(p)$, the statistic p_n reaches B from below. The time $\tau_2(p)$ is the first time for p_n to either cross A or go below B, after time $\tau_1(p)+t_1(p)$. We define, $t_2(p)$, $\tau_3(p)$, etc. similarly. Next let

$$N(p) = \inf\{k \geq 1 : p_{\tau_k} > A\}.$$

For simplicity we introduce the notion of "cycles", "success" and "failure". With reference to the definitions of $\tau_k(p)$'s above, we say that a success has occurred if the statistic p_n, starting with $p_0 = p$, crosses A before going below B. In that case we also say that the number of cycles to A is 1. If on the other hand, the statistic p_n goes below B before it crosses A, we say a failure has occurred. Once p_n cross B from below, we say that a new cycle has started. The number of cycles is 2, if now the statistic p_n crosses A without ever going below B. Thus, $N(p)$ is the number of cycles to success. With this terminology in mind we write

$$\begin{aligned}\mathbb{P}_1(N(p) \geq k) = \mathbb{P}_1(\text{fail in 1st cycle})\, \mathbb{P}_1(\text{fail in 2nd cycle}|\ \text{fail in 1st cycle }) \\ \cdots \mathbb{P}_1(\text{fail in } k-1^{st} \text{ cycle}|\text{fail in all previous}).\end{aligned}$$

Now,

$$\begin{aligned}&\mathbb{P}_1(\text{fail in } i^{th} \text{ cycle}|\text{fail in all previous}) \\ &\qquad = 1 - \mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}).\end{aligned}$$

To obtain a bound on these probabilities, we now define

$$R_n = \frac{p_n}{(1-p_n)\rho}.$$

We note that this is the Shiryaev-Roberts statistic [1]. As long as $p_n \geq B$, the statistic evolves according to the Shiryaev recursion. As a result $Z_n = \log R_n$ has the expression

$$Z_n = \log R_n = Y_n + n|\log(1-\rho)| + \log\left(1 + R_0 + \sum_{k=1}^{n-1}(1-\rho)^k e^{-Y_i}\right), \quad (2.30)$$

where $Y_n = \sum_{i=1}^{n} \log L(X_i)$. Thus, in each cycle, when the DE-Shiryaev statistic is above B, it evolves according to the expression in (2.30) starting with some non-negative R_0. Since $R_0 \geq 0$, the only way a failure can happen in a cycle is if the random walk Y_n takes negative values. Thus,

$$\mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}) \geq \mathbb{P}_1(Y_n \geq 0,\ \forall n).$$

From Corollary 2.4, on p. 22 of [16], it is well known that $\mathbb{P}_1(Y_n \geq 0,\ \forall n) > 0$. Also, note that this probability is not a function of the initial point p. Thus, if

$$q = \mathbb{P}_1(Y_n \geq 0,\ \forall n),$$

then

$$\mathbb{P}_1(N(p) \geq k) \leq (1-q)^{k-1},$$

and

$$\mathbb{E}_1[N(p)] = \sum_{k=1}^{\infty} \mathbb{P}_1(N(p) \geq k) \leq \sum_{k=1}^{\infty}(1-q)^{k-1} = \frac{1}{q} < \infty. \quad (2.31)$$

The previous equation implies that

$$N(p) < \infty \text{ a.s. under } \mathbb{P}_1. \quad (2.32)$$

Let $\lambda_1(p) = \tau_1(p)$, $\lambda_2(p) = \tau_2(p) - \tau_1(p) - t_1(p)$, etc, be the lengths of the

sojourns of the statistic p_n above B. Then clearly we have

$$\tau_{\rm D}(p) = \sum_{n=1}^{N(p)} \lambda_k(p) + \sum_{n=1}^{N(p)-1} t_k(p),$$

and hence,

$$\begin{aligned} \mathbb{E}_1[\tau_{\rm D}(p)] &= \mathbb{E}_1\left[\sum_{n=1}^{N(p)} \lambda_k(p)\right] + \mathbb{E}_1\left[\sum_{n=1}^{N(p)-1} t_k(p)\right] \\ &\leq \mathbb{E}_1\left[\sum_{n=1}^{N(p)} \lambda_k(p)\right] + \frac{t(B)}{q}. \end{aligned} \tag{2.33}$$

Here, $t(B)$ is as defined in (2.19), and the inequality is true because

$$t_k(p) \leq t(B), \text{ for any } k, p,$$

and because of (2.31).

We now make an important observation. We note that the expression $\mathbb{E}_1\left[\sum_{n=1}^{N(p)} \lambda_k(p)\right]$ is the time for the Shiryaev statistic $p_n^{\rm s}$ to cross A from below, starting with $p_0^{\rm s} = p$, with the difference that each time the statistic goes below B, to say $x < B$, the statistic is reset to a value in $[B, A)$. This value corresponds to the overshoot of the DE-Shiryaev statistic p_n when it crosses B from below using the recursion $p_{n+1} = p_n + (1-p_n)\rho$, starting with x. By sample-pathwise arguments, it is easy to see that the latter delay is upper bounded on an average by the following stopping time,

$$\hat{\tau}(p) = \inf\{n \geq 1 : p_n > A; p_0 = p; \text{ each time } p_n < B, \text{ it is reset to } B\}.$$

Thus,

$$\mathbb{E}_1\left[\sum_{n=1}^{N(p)} \lambda_k(p)\right] \leq \mathbb{E}_1[\hat{\tau}(p)] \leq \mathbb{E}_1[\hat{\tau}(B)],$$

where the last inequality follows again by sample-pathwise arguments. Again by sample-pathwise arguments it is easy to see that

$$\mathbb{E}_1[\hat{\tau}(B)] \leq \mathbb{E}_1[\tau_{\rm S}(B)] \leq \mathbb{E}_1[\tau_{\rm S}]. \tag{2.34}$$

In (2.34) $\tau_{\rm S}(B)$ is the time for the Shiryaev statistic to cross A starting at B (with no resetting of statistics). Thus, we have

$$\mathbb{E}_1\left[\sum_{n=1}^{N(p)}\lambda_k(p)\right] \leq \mathbb{E}_1[\tau_{\rm S}]. \tag{2.35}$$

We note that the right-hand side of (2.35) is not a function of the initial value p. But, it does depend on the initial assumption that $p \in [B, A)$.

Thus, from (2.33) we have for $p \in [B, A)$,

$$\mathbb{E}_1[\tau_{\rm D}(p)] \leq \mathbb{E}_1[\tau_{\rm S}] + \frac{t(B)}{q}. \tag{2.36}$$

Going back to (2.29) we note that if $\mathcal{I}_{n-1} = i_{n-1}$ is such that $p_{n-1} = p \in [B, A)$, then

$$\mathbb{E}_n[\tau_{\rm D} - n|\mathcal{I}_{n-1} = i_{n-1}] \leq \mathbb{E}_1[\tau_{\rm S}] + \frac{t(B)}{q}. \tag{2.37}$$

If $\mathcal{I}_{n-1} = i_{n-1}$ is such that $p_{n-1} = p < B$, then the time to cross A for the DE-Shiryaev statistic will be equal to the time taken for the statistic to cross B from below, plus a time which corresponds to the left-hand side of (2.37) (with initial point p now corresponding to the overshoot of the DE-Shiryaev statistic when it crosses B from below). This latter time is again on an average bounded by $\mathbb{E}_1[\tau_{\rm S}] + \frac{t(B)}{q}$. Thus, we can write,

$$\mathbb{E}_n[(\tau_{\rm D} - n)^+|\mathcal{I}_{n-1}] \leq \mathbb{E}_1[\tau_{\rm S}] + \frac{t(B)}{q} + t(B). \tag{2.38}$$

Note that the right-hand side of (2.38) is no more a function of the conditioning $\mathcal{I}_{n-1}$. The proof is complete if we define

$$K_{\rm DS} = \frac{t(B)}{q} + t(B),$$

and average the left-hand side of (2.38) over the set $\tau_{\rm D} \geq n$, and then invoke (2.28). □

2.5.1 The Nonarithmetic Case

In this section we obtain a stronger result on the PFA of the DE-Shiryaev algorithm for the case when the $\log L(X)$ is nonarithmetic. A random variable is called nonarithmetic if it is not a lattice random variable. A lattice random variable X is such that there exists $d > 0$,

$$\mathbb{P}(X \in \{kd; k \text{ integer } \}) = 1.$$

We note that in Theorem 2.5.2 it is only guaranteed that setting $A = 1 - \alpha$ implies $\mathsf{PFA}(\tau_{\mathrm{D}}) \leq \alpha$. Thus, we can use the same threshold for both the Shiryaev algorithm and the DE-Shiryaev algorithm to meet the PFA constraint. However, the actual PFA of the two algorithms can in general be different, with one being much smaller than α than the other. In this subsection we show however that if $\log L(X)$ is nonarithmetic, then the ratio of the PFAs of the two algorithms goes to 1. Thus, in the nonarithmetic case, not only the delays but also the false alarms of the two algorithms are also asymptotically the same. This result is stronger than mere first-order asymptotic optimality of the DE-Shiryaev algorithm.

We use the tools from nonlinear renewal theory [16]. The key is the distribution of the overshoot when the statistic $Z_n = \log \frac{p_n}{(1-p_n)}$, for a fixed $b = \log \frac{B}{(1-B)}$, crosses the threshold $a = \log \frac{A}{(1-A)}$, for a large a. To establish the role of the overshoot distribution, we obtain an expression for PFA as a function of the overshoot when Z_k crosses a from below.

Lemma 2.5.2. *For any policy Ψ such that*

$$\tau = \inf\{n \geq 1 : p_n > A\},$$

and $\tau < \infty$ a.s., we have

$$\text{PFA} \;\; = \;\; \mathbb{E}[1 - p_\tau] = e^{-a}\mathbb{E}[e^{-(Z_\tau - a)} | \tau \geq \Gamma](1 + o(1)) \;\; \textit{as } a \;\; \to \infty.$$

Proof. Since, $p_\tau > A$ implies $Z_\tau > a$, we have,

$$\frac{1}{1 + e^{-Z_\tau}} \geq \frac{1}{1 + e^{-a}}.$$

The required result is obtained by obtaining upper and lower bounds on PFA

as follows.

$$\mathsf{PFA} = \mathbb{E}[1 - p_\tau] = \mathbb{E}\left[\frac{1}{1+e^{Z_\tau}}\right] \leq \mathbb{E}\left[e^{-Z_\tau}\right].$$

Also,

$$\begin{aligned}\mathsf{PFA} = \mathbb{E}[1 - p_\tau] = \mathbb{E}\left[\frac{1}{1+e^{Z_\tau}}\right] &= \mathbb{E}\left[\frac{1}{e^{Z_\tau}}\frac{1}{1+e^{-Z_\tau}}\right] \\ &\geq \mathbb{E}\left[\frac{1}{e^{Z_\tau}}\frac{1}{1+e^{-a}}\right] \\ &= \mathbb{E}\left[e^{-Z_\tau}\right](1+o(1)) \text{ as } a \to \infty.\end{aligned}$$

Thus,

$$\mathsf{PFA} = \mathbb{E}[e^{-Z_\tau}](1+o(1)) = e^{-a}\mathbb{E}[e^{-(Z_\tau - a)}](1+o(1)) \text{ as } a \to \infty.$$

Now note that,

$$\begin{aligned}\mathbb{E}[e^{-(Z_\tau - a)}] &= \mathbb{E}[e^{-(Z_\tau - a)}|\tau \geq \Gamma](1 - \mathbb{P}(\tau < \Gamma)) + \mathbb{E}[e^{-(Z_\tau - a)}|\tau < \Gamma]\mathbb{P}(\tau < \Gamma) \\ &= \mathbb{E}[e^{-(Z_\tau - a)}|\tau \geq \Gamma] \\ &\quad + \left(\mathbb{E}[e^{-(Z_\tau - a)}|\tau < \Gamma] - \mathbb{E}[e^{-(Z_\tau - a)}|\tau \geq \Gamma]\right)\mathbb{P}(\tau < \Gamma).\end{aligned}$$

Since, $\mathbb{P}(\tau < \Gamma) = \mathbb{E}[1 - p_\tau] \leq 1 - A \leq e^{-a}$, and e^{-x} is bounded by one, we can write

$$\mathbb{E}[e^{-(Z_\tau - a)}] = \mathbb{E}[e^{-(Z_\tau - a)}|\tau \geq \Gamma] + o(1) \text{ as } a \to \infty.$$

Hence,

$$\mathsf{PFA} = e^{-a}\mathbb{E}[e^{-(Z_\tau - a)}|\tau \geq \Gamma](1+o(1)) \text{ as } a \to \infty.$$

This proves the lemma. □

From Lemma 2.5.2, it is evident that PFA for the DE-Shiryaev algorithm depends on the overshoot when Z_k crosses a as $a \to \infty$. Before we study this overshoot distribution we recall that the DE-Shiryaev statistic p_n has the following recursions:

$$p_n = \tilde{p}_{n-1} \text{ if } S_n = 0, \tag{2.39}$$

and

$$p_n = \frac{\tilde{p}_{n-1} L(X_n)}{\tilde{p}_{n-1} L(X_n) + (1 - \tilde{p}_{n-1})} \text{ if } S_n = 1, \tag{2.40}$$

with $\tilde{p}_n = p_n + (1 - p_n)\rho$ and $L(X_n) = f_1(X_n)/f_0(X_n)$. We recall that $Z_n = \log \frac{p_n}{(1-p_n)}$ and write such a recursion for Z_n by combining the recursions in (2.40) and (2.39):

$$\begin{aligned} Z_{n+1} &= Z_n + \mathbb{I}_{\{S_{n+1}=1\}} \log L(X_{n+1}) + |\log(1-\rho)| + \log\left(1 + e^{-Z_n}\rho\right) \\ &= Z_n + \mathbb{I}_{\{Z_n \geq b\}} \log L(X_{n+1}) + |\log(1-\rho)| + \log\left(1 + e^{-Z_n}\rho\right). \end{aligned} \tag{2.41}$$

In analyzing the trajectory of Z_n, it is useful to allow for an arbitrary random starting point Z_0. By defining $Y_k = \log L(X_k) + |\log(1-\rho)|$ and expanding the recursion (2.41), we can write an expression for Z_n:

$$\begin{aligned} Z_n &= \sum_{k=1}^{n} Y_k + Z_0 + \sum_{k=0}^{n-1} \log\left(1 + e^{-Z_k}\rho\right) - \sum_{k=1}^{n} \mathbb{I}_{\{Z_{k-1}<b\}} \log L(X_k) \\ &= \sum_{k=1}^{n} Y_k + \eta_n. \end{aligned} \tag{2.42}$$

Here η_n is used to represent all the terms other than the first term $\sum_{k=1}^{n} Y_k$ in (2.42):

$$\eta_n = Z_0 + \sum_{k=0}^{n-1} \log\left(1 + e^{-Z_k}\rho\right) - \sum_{k=1}^{n} \mathbb{I}_{\{Z_{k-1}<b\}} \log L(X_k). \tag{2.43}$$

Recall that we are interested in the overshoot distribution when Z_n crosses a large boundary. According to nonlinear renewal theory, this overshoot distribution is the same as the overshoot distribution of the random walk $\sum_{k=1}^{n} Y_k$, when the latter crosses a large threshold, provided that the sequence $\{\eta_n\}$ is *slowly changing*. As defined in [18], η_n is a slowly changing sequence if

$$n^{-1} \max\{|\eta_1|, \ldots, |\eta_n|\} \xrightarrow[i.p.]{n\to\infty} 0, \tag{2.44}$$

and for every $\epsilon > 0$, there exists n^* and $\delta > 0$ such that for all $n \geq n^*$

$$\mathbb{P}\{\max_{1 \leq k \leq n\delta} |\eta_{n+k} - \eta_n| > \epsilon\} < \epsilon. \tag{2.45}$$

We now show that $\{\eta_n\}$ is indeed a slowly changing sequence and hence, the

asymptotic overshoot distribution of Z_n is the same as that of the random walk $\sum_{k=1}^{n} Y_k$.

Theorem 2.5.3. *Let $R(x)$ be the asymptotic distribution of the overshoot when the random walk $\sum_{k=1}^{n} Y_k$ crosses a large positive boundary under $\mathbb{P}_1$. Then for fixed ρ, b, we have the following:*

1. *$\{\eta_n\}$ is a slowly changing sequence under $\mathbb{P}_1$.*
2. *Conditioned on $\Gamma = \gamma$ and $\tau_{\text{D}} \geq \gamma$, $R(x)$ is the distribution of $Z_{\tau_{\text{D}}} - a$ as $a \to \infty$, i.e.,*

$$\lim_{a\to\infty} \mathbb{P}_\gamma \left[Z_{\tau_{\text{D}}} - a \leq x | \tau_{\text{D}} \geq \gamma \right] = R(x). \tag{2.46}$$

3. *Also,*

$$\lim_{a\to\infty} \mathbb{E}[e^{-(Z_{\tau_{\text{D}}} - a)} | \tau_{\text{D}} \geq \Gamma] = \int_0^\infty e^{-x} dR(x).$$

Proof. We first show that η_n with $b = -\infty$, and Z_0 a random variable, is a slowly changing sequence.

When $b = -\infty$, the statistic Z_n evolves as in the classical Shiryaev algorithm, and it is easy to see that in this case:

$$\eta_n \quad = \quad \left[Z_0 + \sum_{k=0}^{n-1} \log\left(1 + e^{-Z_k}\rho\right) \right].$$

If we start with the Shiryaev-Robert statistic $R_n = \frac{p_n}{1-p_n}$ and write an expression for it by expanding its recursion, as we did for Z_n above, and then we take log of that expression (note that Z_n is $\log R_n$), then we will get another expression for η_n:

$$\eta_n \quad = \quad \log \left[e^{Z_0} + \sum_{k=0}^{n-1} \rho(1-\rho)^k \prod_{i=1}^{k} \frac{f_0(X_i)}{f_1(X_i)} \right].$$

Since the two η_ns should be the same, we have

$$\begin{aligned} \eta_n \quad &= \quad \left[Z_0 + \sum_{k=0}^{n-1} \log\left(1 + e^{-Z_k}\rho\right) \right] \\ &= \quad \log \left[e^{Z_0} + \sum_{k=0}^{n-1} \rho(1-\rho)^k \prod_{i=1}^{k} \frac{f_0(X_i)}{f_1(X_i)} \right]. \end{aligned}$$

Since the terms in the summation are non-negative, we have

$$\eta_n \xrightarrow[n\to\infty]{} \log\left[e^{Z_0} + \sum_{k=0}^{\infty} \rho(1-\rho)^k \prod_{i=1}^{k} \frac{f_0(X_i)}{f_1(X_i)}\right] = \left[Z_0 + \sum_{k=0}^{\infty} \log\left(1 + e^{-Z_k}\rho\right)\right].$$

Define

$$\eta(Z_0) \stackrel{\triangle}{=} \log\left[e^{Z_0} + \sum_{k=0}^{\infty} \rho(1-\rho)^k \prod_{i=1}^{k} \frac{f_0(X_i)}{f_1(X_i)}\right].$$

Note that $\eta(Z_0)$ as a function of Z_0 is well defined and finite under $\mathbb{P}_1$. This is because by Jensen's inequality, for $Z_0 = z_0$,

$$\begin{aligned}
\mathbb{E}[\eta(z_0)] &\leq \log\left[e^{z_0} + \sum_{k=0}^{\infty} \rho(1-\rho)^k \mathbb{E}_1\left(\prod_{i=1}^{k} \frac{f_0(X_i)}{f_1(X_i)}\right)\right] \\
&= \log\left[e^{z_0} + \sum_{k=0}^{\infty} \rho(1-\rho)^k\right] = \log\left(e^{z_0} + 1\right).
\end{aligned}$$

Thus

$$\eta_n \xrightarrow[b=-\infty]{\mathbb{P}_1-a.s.} \eta(Z_0) = Z_0 + \sum_{k=0}^{\infty} \log\left(1 + e^{-Z_k}\rho\right). \tag{2.47}$$

This implies $\sum_{k=0}^{\infty} \log\left(1 + e^{-Z_k}\rho\right)$ converges a.s. for i.i.d. $\{X_k\}$ and $b = -\infty$. This series will also converge with probability 1 if we condition on a set with positive probability.

Let change happen at $\Gamma = \gamma$. We set $Z_0 = Z_\Gamma = Z_\gamma$ and assume that $\{X_n\}$, $n \geq 1$ have density f_1, which would happen after Γ. We first show that starting with the above Z_0, the sequence η_n generated in (2.43) is slowly changing.

To verify the first condition (2.44), from (2.43) note that,

$$\begin{aligned}
&n^{-1} \max\{|\eta_1|, \ldots, |\eta_n|\} \\
&\quad \leq n^{-1}\left[|Z_0| + \sum_{k=0}^{n-1} \log\left(1 + e^{-Z_k}\rho\right) + \sum_{k=1}^{n} \left(|\log L(X_k)|\right) \mathbb{I}_{\{Z_{k-1}<b\}}\right].
\end{aligned}$$

Since, $Z_k \to \infty$ a.s., $\log\left(1 + e^{-Z_k}\rho\right) \to 0$, also, $\mathbb{I}_{\{Z_k<b\}} \to 0$ a.s. Thus both the sequences $\left\{\log\left(1 + e^{-Z_k}\rho\right)\right\}$ and $\left\{\left(|\log L(X_k)|\right) \mathbb{I}_{\{Z_{k-1}<b\}}\right\}$ are Cesaro summable and have Cesaro sum of zero. Thus the term inside the square bracket above, when divided by n, goes to zero a.s. and hence also in probability. Thus the first condition is verified.

To verify the second condition (2.45), we first obtain a bound on $|\eta_{n+k}-\eta_n|$.

$$|\eta_{n+k}-\eta_n| \leq \sum_{i=n}^{n+k-1} \log\left(1+e^{-Z_i}\rho\right) + \sum_{i=n+1}^{n+k} \left(|\log L(X_i)|\right) \mathbb{I}_{\{Z_{i-1}<b\}}.$$

Thus,

$$\max_{1\leq k\leq n\delta} |\eta_{n+k}-\eta_n| \leq \sum_{i=n}^{n+n\delta-1} \log\left(1+e^{-Z_i}\rho\right) + \sum_{i=n+1}^{n+n\delta} \left(|\log L(X_i)|\right) \mathbb{I}_{\{Z_{i-1}<b\}} \stackrel{\triangle}{=} d_n^1+d_n^2.$$

Here, for convenience of computation, we use d_n^1 and d_n^2 to represent the first and second partial sums respectively. Now,

$$\mathbb{P}\{\max_{1\leq k\leq n\delta} |\eta_{n+k}-\eta_n| > \epsilon\} \leq \mathbb{P}(d_n^1+d_n^2>\epsilon),$$

and we bound the probability $\mathbb{P}(d_n^1+d_n^2>\epsilon)$ as follows.

On the event that $E \stackrel{\triangle}{=} \{Z_k \geq b, \forall k \geq 0\}$, d_n^2 is identically zero, thus for n large enough,

$$\mathbb{P}(d_n^1+d_n^2>\epsilon|E) = \mathbb{P}(d_n^1>\epsilon|E) < \epsilon.$$

This is because d_n^1 behaves like a partial sum of a series of type in (2.47). Since the series in (2.47) converges if random variables are generated i.i.d. f_1, it will also converge if conditioned on the event E. Thus, the partial sum d_n^1 converges to 0 almost surely, and hence converges to 0 in probability, i.e., $\mathbb{P}(d_n^1>\epsilon|E) \to 0$. Select, $n=n_1^*$ such that $\forall n > n_1^*$, $\mathbb{P}(d_n^1>\epsilon|E)<\epsilon$.

Define

$$L_Z = \sup\{k \geq 1 : Z_{k-1} < b, Z_k \geq b\},$$

with $L_Z=\infty$ if no such k exists. On the event E', which is the compliment of E, L_Z is a.s. finite. Then, by noting that $d_n^2=0$ for $L_Z<n$, we get for n

large enough,

$$\begin{aligned}
\mathbb{P}(d_n^1 + d_n^2 > \epsilon | E') &\stackrel{\triangle}{=} \mathbb{P}_{E'}(d_n^1 + d_n^2 > \epsilon) \\
&\leq \mathbb{P}_{E'}(d_n^1 + d_n^2 > \epsilon; L_Z \geq n) + \mathbb{P}_{E'}(d_n^1 + d_n^2 > \epsilon; L_Z < n) \\
&\leq \mathbb{P}_{E'}(L_Z \geq n) + \mathbb{P}_{E'}(d_n^1 + d_n^2 > \epsilon; L_Z < n) \\
&= \mathbb{P}_{E'}(L_Z \geq n) + \mathbb{P}_{E'}(d_n^1 > \epsilon; L_Z < n) \\
&\leq \mathbb{P}_{E'}(L_Z \geq n) + \mathbb{P}_{E'}(d_n^1 > \epsilon | L_Z < n) \\
&< \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

Since, L_Z is almost surely finite, $\mathbb{P}_{E'}(L_Z \geq n) \to 0$ as $n \to \infty$. Thus we can select $n = n_2^*$ such that $\forall n > n_2^*$, $\mathbb{P}_{E'}(L_Z \geq n) < \epsilon/2$. For the second term, note that conditioned on $L_Z < n$, d_n^1 behaves like a partial sum of a series of type in (2.47), with Z_0 replaced by Z_{L_Z}. Since the series in (2.47) converges if random variables are generated i.i.d. f_1 beyond L_Z, it will also converge if conditioned on the event $\{L_Z < n\}$. Thus, the partial sum d_n^1 converges to 0 almost surely, and hence converges to 0 in probability, i.e., $\mathbb{P}_{E'}(d_n^1 > \epsilon | L_Z < n) \to 0$. Select, $n = n_3^*$ such that $\forall n > n_3^*$, $\mathbb{P}(d_n^1 > \epsilon | L_Z < n) < \epsilon/2$. Then $n^* = \max\{n_1^*, n_2^*, n_3^*\}$, is the desired n^* and pick any $\delta > 0$. Then for $n > n^*$,

$$\begin{aligned}
\mathbb{P}(d_n^1 + d_n^2 > \epsilon) &= \mathbb{P}(d_n^1 + d_n^2 > \epsilon | E)\mathbb{P}(E) + \mathbb{P}(d_n^1 + d_n^2 > \epsilon | E')\mathbb{P}(E') \\
&< \epsilon \mathbb{P}(E) + \epsilon \mathbb{P}(E') < \epsilon.
\end{aligned}$$

Thus, the sequence $\{\eta_n\}$ is slowly changing. Since the sequence η_n is slowly changing, according to [18], the asymptotic distribution of the overshoot when Z_n crosses a large boundary under f_1 is $R(x)$. Thus we have the following result,

$$\lim_{a \to \infty} \mathbb{P}_\gamma \left[Z_{\tau_{\mathrm{D}}} - a \leq x | \tau_{\mathrm{D}} \geq \gamma \right] = R(x).$$

Since $e^{-(Z_\tau - a)}$ is a bounded continuous function of the overshoot, we have by convergence of measure arguments that

$$\lim_{a \to \infty} \mathbb{E}_\gamma [e^{-(Z_{\tau_{\mathrm{D}}} - a)} | \tau_{\mathrm{D}} \geq \gamma] = \int_0^\infty e^{-x} dR(x).$$

To prove that the result is true even when averaged over the distribution

of Γ, we note that

$$\mathbb{E}[e^{-(Z_{\tau_{\mathrm{D}}}-a)}|\tau_{\mathrm{D}} \geq \Gamma] = \sum_{\gamma=1}^{\infty} \mathbb{E}_{\gamma}[e^{-(Z_{\tau_{\mathrm{D}}}-a)}|\tau_{\mathrm{D}} \geq \gamma]\ \mathbb{P}(\Gamma = \gamma|\tau_{\mathrm{D}} \geq \Gamma).$$

Since, $\mathbb{E}_{\gamma}[e^{-(Z_{\tau_{\mathrm{D}}}-a)}|\tau_{\mathrm{D}} \geq \gamma] \to \int_0^{\infty} e^{-x} dR(x)$ and $\mathbb{P}(\Gamma = \gamma|\tau_{\mathrm{D}} \geq \Gamma) \to \mathbb{P}(\Gamma = \gamma)$, the result follows by dominated convergence theorem. □

Thus, we have the following result.

Theorem 2.5.4. *For a fixed b and ρ,*

$$\mathrm{PFA}(\tau_{\mathrm{D}}) = \left(e^{-a} \int_0^{\infty} e^{-x} dR(x)\right)(1 + o(1)) \ \textit{as}\ a \to \infty. \tag{2.48}$$

We note that the right-hand side in (2.48) is also the PFA of the Shiryaev algorithm [15]. This proves our claim that the ratio of the PFA of the Shiryaev algorithm and the DE-Shiryaev algorithm goes to 1.

We thus have the following theorem:

Theorem 2.5.5. *If $B = B_{\beta}$ and $A = A_{\alpha} = 1 - \alpha$, then*

$$\begin{aligned}
&\mathsf{ANO}(\tau_{\mathrm{D}}) \leq \beta, \ \textit{for}\ \alpha\ \textit{small enough},\\
&\mathsf{PFA}(\tau_{\mathrm{D}}) \leq \alpha, \ \textit{and}\\
&\mathsf{ADD}(\tau_{\mathrm{D}}) \sim \mathsf{ADD}(\tau_{\mathrm{S}}) \sim \frac{|\log(\alpha)|}{D(f_1 \,||\, f_0) + |\log(1-\rho)|}(1 + o(1)), \ \textit{as}\ \alpha \to 0\ .
\end{aligned} \tag{2.49}$$

If further $\log L(X)$ *is non-arithmetic, then for a fixed β,*

$$\mathsf{PFA}(\tau_{\mathrm{D}}) \sim \mathsf{PFA}(\tau_{\mathrm{S}}) \sim \alpha \int_0^{\infty} e^{-x} dR(x) \ \textit{as}\ \alpha \to 0, \tag{2.50}$$

2.6 Numerical Results

In Section 2.5 we obtained asymptotic expressions for the ADD and the PFA of the DE-Shiryaev algorithm as a function of the system parameters: the threshold a, the densities f_0 and f_1, and the prior ρ. In this section we provide simulation results to show the accuracy of the asymptotic expressions for moderate values of the false alarm constraint α. We also compare the

performance of the DE-Shiryaev algorithm with other tests. The observations are assumed to be Gaussian with $f_0 \sim \mathcal{N}(0,1)$, and $f_1 \sim \mathcal{N}(\theta,1)$, $\theta > 0$, for the simulations and analysis. In the simulations, the PFA values are computed using the expression $\mathbb{E}[1-p_\tau]$. This guarantees a faster convergence for small values of PFA.

In Fig. 2.2 we first compare the performance of the Shiryaev algorithm, the DE-Shiryaev algorithm and the fractional sampling scheme, for $\beta = 50$. In the fractional sampling scheme, the Shiryaev algorithm is used and observations are skipped by tossing a biased coin (with probability of success 50/99), without looking at the state of the system. When an observation is skipped in the fractional sampling scheme, the Shiryaev statistic is updated using the prior on change point. In the figure we see a substantial gap in performance between the DE-Shiryaev algorithm and the fractional sampling scheme.

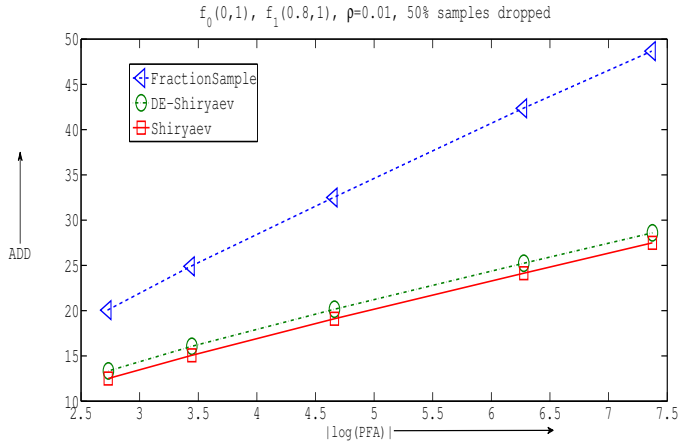


Figure 2.2: Comparative performance of schemes for $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.8,1)$, and $\rho = 0.01$.

By Theorem 2.5.4, we have the following approximation for PFA:

$$\text{PFA} \approx e^{-a} \int_0^\infty e^{-x} dR(x).$$

We note that $\int_0^\infty e^{-x} dR(x)$ can be computed numerically, at least for Gaussian observations [18]. In this section we provide numerical results to show the accuracy of the above expression for PFA.

In Table 2.1 we compare the analytical approximation with the PFA obtained using simulations of the DE-Shiryaev algorithm for various choices of

ρ, thresholds a, and $b = \log \frac{B}{(1-B)}$, and post change mean θ. From the table we see that the analytical approximation is quite good.

Table 2.1: PFA: for $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(\theta,1)$

θ	ρ	a	b	PFA Simulations	PFA Analysis
0.4	0.01	3.0	0	3.78×10^{-2}	3.94×10^{-2}
0.4	0.01	6.0	2.0	1.955×10^{-3}	1.96×10^{-3}
0.75	0.01	9.0	-2.0	7.968×10^{-5}	7.964×10^{-5}
2.0	0.01	5.0	-4.0	2.15×10^{-3}	2.155×10^{-3}
0.75	0.005	7.6	3.0	3.231×10^{-4}	3.235×10^{-4}
0.75	0.1	4.0	-3.0	1.143×10^{-2}	1.157×10^{-2}

In Table 2.2, we show that PFA is not a function of b for large values of a. We fix $a = 4.6$, and increase b from -2.2 to 0.85. We notice that PFA is unchanged in simulations when b is changed this way. This is also captured by the analysis and it is quite accurate.

Table 2.2: PFA for $\rho = 0.01$, $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.75,1)$

a	b	Simulations	Analysis
4.6	-2.2	6.44×10^{-3}	6.48×10^{-3}
4.6	-1.5	6.44×10^{-3}	6.48×10^{-3}
4.6	-0.85	6.44×10^{-3}	6.48×10^{-3}
4.6	0	6.44×10^{-3}	6.48×10^{-3}
4.6	0.85	6.44×10^{-3}	6.48×10^{-3}

By Theorem 2.5.4 again a first-order approximation for ADD of the DE-Shiryaev algorithm is:

$$\mathsf{ADD} \approx \left\lceil \frac{a}{D(f_1, f_0) + |\log(1-\rho)|} \right\rceil. \tag{2.51}$$

This is also the first-order approximation for the ADD of the Shiryaev algorithm, and gives a good estimate of the delay when PFA is small.

For the Shiryaev algorithm, (2.51) provides a very good estimate of the delay even for moderate values of PFA; see [15]. In case of the DE-Shiryaev algorithm, the accuracy of (2.51) depends on the choice of b and hence on the constraint β, as having $b > -\infty$ increases the delay. Before we demonstrate this through numerical and simulation results we introduce the following

concept:

$$\mathsf{ANO}\% = \text{ ANO } \textit{expressed as a percentage of } \mathbb{E}[\Gamma]. \tag{2.52}$$

For example, if $\rho = 0.05$, and for some choice of system parameters $\mathsf{ANO} = 15$, then $\mathsf{ANO}\% = 15 * 0.05 = 75\%$. Thus, the concept of $\mathsf{ANO}\%$ captures the reduction in the average number of observations used before change by employing the DE-Shiryaev algorithm.

In Table 2.3 we provide various numerical examples where (2.51) is a good approximation for ADD. Since (2.51) is a good approximation for the Shiryaev delay as well, it follows that, for these parameter values, the delay of the DE-Shiryaev algorithm is approximately equal to the delay of the Shiryaev algorithm. It might be intuitive that if we are aiming for large $\mathsf{ANO}\%$ values of say 90%, then the delay of the two algorithms will be close to each other. But from the values in Table 2.3 we infer that it is possible to achieve considerably smaller values of $\mathsf{ANO}\%$ without significantly affecting the delay.

Table 2.3: $f_0 \sim \mathcal{N}(0,1),\ f_1 \sim \mathcal{N}(\theta,1)$

				ADD		PFA		ANO%
θ	ρ	a	b	Simulations $\mathbb{E}[\tau - \Gamma \vert \tau \geq \Gamma]$	Analysis (2.51)	Simulations	Analysis	
0.4	0.01	8.5	-2.2	104.9	111.7	1.608×10^{-4}	1.608×10^{-4}	66%
0.75	0.01	6.467	-2.2	32.3	29.5	1.002×10^{-3}	1.004×10^{-3}	35%
2.0	0.01	7.5	-4.0	6.1	6.23	1.77×10^{-4}	1.768×10^{-4}	43%
0.75	0.005	8.7	-3.0	42.6	40.4	1.076×10^{-4}	1.076×10^{-4}	77%
0.75	0.1	8.5	0.0	23.9	22.18	1.286×10^{-4}	1.285×10^{-4}	26%

However, if the $\mathsf{ANO}\%$ value is small, then this means that the value of b is large, and further that the delay is large. In this case, it might happen that (2.51) is a good approximation only for values of PFA which are very small. This is demonstrated in Table 2.4. It is clear from the table that, for the parameter values considered, estimating the delay with less than 10% error is only possible at PFA values of the order of $\mathsf{PFA} \approx 10^{-22}$.

This motivates the need for a more accurate estimate of the delay. Please see [17] for details.

Table 2.4: $\rho = 0.05,\ f_0 \sim \mathcal{N}(0,1),\ f_1 \sim \mathcal{N}(0.75,1)$

a	b	Simulations ADD	Analysis (2.51)	ANO%	PFA
5.0	1.0	30	13	7.5%	4.3×10^{-3}
9.0	1.0	42	25	7.5%	7.9×10^{-5}
13.0	1.0	54	37	7.5%	1.4×10^{-6}
18.0	1.0	69	52	7.5%	9.7×10^{-9}
50.0	1.0	165	149	7.5%	1.23×10^{-22}

2.7 Existing Literature

The problem of quickest change detection with cost considerations is also studied in the quality control literature in the context of sampling rate control and sampling size control [19], [9], [7]. However, a systematic study of the quickest change detection problem with observation control in the framework of classical quickest change detection is not available in the literature

The problem of controlling the long-term average operational cost of a machine in the context of destructive testing is studied by Girshik and Rubin in [19]. In fact, both the Shiryaev and the DE-Shiryaev algorithms were first proposed in [19], and their optimality proved for different machine repair problems. However, in the absence of a suitable common problem formulation and absence of performance analysis of these algorithms, the relationship between the two algorithms is not obvious from the results in [19].

The problem of quickest change detection with observation control in the context of sensor networks is studied in [20]. However, the structure of the optimal observation control is studied only numerically.

CHAPTER 3

DATA-EFFICIENT MINIMAX QUICKEST CHANGE DETECTION

In Chapter 2 we studied data-efficient quickest change detection in the Bayesian setting. There we modeled the change point as a random variable with a known distribution, and extended the classical quickest change detection problem studied by Shiryaev by introducing an additional constraint on the cost of observations used before the change point. In this chapter we relax the assumption that the distribution of change point is known and study data-efficient quickest change detection in non-Bayesian settings.

The classical quickest change detection problem in non-Bayesian settings (without any data-efficiency or observation control) is studied in [12], [13], [21], [22], [14], and [23]. In these works, in the absence of the knowledge of the distribution of the change point, the change point is modeled as an unknown constant. In this non-Bayesian setting, the quickest change detection problem is studied in two different minimax settings introduced in [12] and [13]. The objective in these minimax settings is to minimize some version of the *worst case average delay*, subject to a constraint on the *mean time to false alarm*. The results from these papers show that, variants of the Shiryaev-Roberts algorithm [24], the latter being derived from the Shiryaev algorithm by setting the geometric parameter to zero, and the CuSum algorithm [11], are asymptotically optimal for both the minimax formulations, as the mean time to false alarm goes to infinity.

Recall that, for the i.i.d. model, and for geometrically distributed change point, we showed in Chapter 2 that a two-threshold generalization of the classical single threshold Shiryaev test is asymptotically optimal, for the data-efficient formulation we proposed there. We called the two-threshold generalization, the DE-Shiryaev algorithm. In the DE-Shiryaev algorithm, the *a posteriori* probability that the change has already happened conditioned on available information, is computed at each time step, and the change is declared the first time this probability crosses a threshold A. When the a

posteriori probability is below this threshold A, observations are taken only when this probability is above another threshold $B < A$. When an observation is skipped, the *a posteriori* probability is updated using the prior on the change point random variable. We also showed that, for reasonable values of the false alarm constraint and the observation cost constraint, these two thresholds can be selected independent of each other: the upper threshold A can be selected directly from the false alarm constraint and the lower threshold B can be selected directly from the observation cost constraint. Finally, we showed that the DE-Shiryaev algorithm achieves a significant gain in performance over the approach of *fractional sampling*, where the Shiryaev algorithm is used and an observation is skipped based on the outcome of a coin toss.

In this chapter we study the data-efficient quickest change detection problem in a non-Bayesian setting, by introducing an additional constraint on the cost of observations used in the detection process, in the minimax settings of [12] and [13]. We first use the insights from the Bayesian analysis from Chapter 2 to propose a metric for data efficiency in the absence of knowledge of the distribution on the change point. This metric is the fraction of time samples are taken before change. We then propose extensions of the minimax formulations in [12] and [13] by introducing an additional constraint on data efficiency in these formulations. The objective in these formulations is to find a stopping time and an on-off observation control policy to minimize a version of the worst case average delay, subject to constraints on the mean time to false alarm and the fraction of time observations are taken before change. Then, motivated by the structure of the DE-Shiryaev algorithm, we propose an extension of the CuSum algorithm from [11]. We call this extension the DE-CuSum algorithm. We show that the DE-CuSum algorithm inherits the good properties of the DE-Shiryaev algorithm. That is, the DE-CuSum algorithm is asymptotically optimal, is easy to design, and provides substantial performance improvements over the approach of fractional sampling, where the CuSum algorithm is used and observations are skipped based on the outcome of a sequence of coin tosses, independent of the observations process.

3.1 Problem Formulation

In the absence of a prior knowledge on the distribution of the change point, as is standard in classical quickest change detection literature, we model the change point as an unknown constant γ. As a result, the quantities $\mathsf{ADD}, \mathsf{PFA}, \mathsf{ANO}$ defined in Chapter 2 are not well defined. Thus, we need new metrics to capture the false alarm rate, delay and data-efficiency. We will reuse the notation from Chapter 2. Specifically,

$$S_n = \begin{cases} 1 & \text{if } X_n \text{ used for decision making} \\ 0 & \text{otherwise.} \end{cases}$$

The information available at time n is denoted by

$$\mathcal{I}_n = \{X_1^{(S_1)}, \cdots, X_n^{(S_n)}\},$$

where $X_k^{(S_k)} = X_k$ if $S_k = 1$, else X_k is absent from $\mathcal{I}_n$, and

$$S_n = \phi_n(\mathcal{I}_{n-1}).$$

Here, ϕ_n denotes the control map. Let τ be a stopping time for the sequence $\{\mathcal{I}_n\}$. A control policy is the collection

$$\Psi = \{\tau, \phi_1, \cdots, \phi_\tau\}.$$

For false alarm, we consider the metric used in [12] and [13], the mean time to false alarm or its reciprocal, the false alarm rate:

$$\mathsf{FAR}(\Psi) = \frac{1}{\mathbb{E}_\infty[\tau]}. \tag{3.1}$$

For delay we consider two possibilities: the minimax setting of Pollak [13], where the delay metric is the following supremum over time of the conditional delay[1]

$$\mathsf{CADD}(\Psi) = \sup_\gamma \quad \mathbb{E}_\gamma[\tau - \gamma | \tau \geq \gamma], \tag{3.2}$$

or the minimax setting of Lorden [12], where the delay metric is the supre-

[1]We are only interested in those policies for which the CADD is well defined.

mum over time of the essential supremum of the conditional delay[2]

$$\mathsf{WADD}(\Psi) = \sup_{\gamma} \operatorname{ess\,sup}\ \mathbb{E}_\gamma \left[(\tau - \gamma)^+ | \mathcal{I}_{\gamma-1}\right]. \tag{3.3}$$

Since $\{\tau \geq \gamma\}$ belongs to the sigma algebra generated by $\mathcal{I}_{\gamma-1}$, we have

$$\mathsf{CADD}(\Psi) \leq \mathsf{WADD}(\Psi).$$

We next propose a metric for data-efficiency in a non-Bayesian setting. In Chapter 2, we saw that in the DE-Shiryaev algorithm, observation cost constraint is met using an initial wait, and by controlling the fraction of time observations are taken, after the initial wait. In the absence of prior statistical knowledge on the change point such an initial wait cannot be justified. This motivates us to seek control policies that can meet a constraint on the fraction of time observations are taken before change. We propose the following duty cycle based observation cost metric, Conditional Pre-change Duty Cycle (CPDC):

$$\begin{aligned} \mathsf{CPDC}(\Psi) &= \limsup_{\gamma\to\infty} \frac{1}{\gamma}\mathbb{E}_\gamma \left[\sum_{k=1}^{\gamma-1} S_k \Big| \tau \geq \gamma\right] \\ &= \limsup_{\gamma\to\infty} \frac{1}{\gamma}\mathbb{E}_\infty \left[\sum_{k=1}^{\gamma-1} S_k \Big| \tau \geq \gamma\right]. \end{aligned} \tag{3.4}$$

Clearly, $\mathsf{CPDC} \leq 1$.

We now discuss why we use lim sup rather than sup in defining CPDC. In all reasonable policies Ψ, S_1 will typically be set to 1. As mentioned earlier, this is because an initial wait cannot be justified without a prior statistical knowledge of the change point. As a result, in (3.4), we cannot replace the lim sup by sup, because the latter would give us a CPDC value of 1. Even otherwise, without any prior knowledge on the change point, it is reasonable to assume that the value of γ is large corresponding to a rare change, and hence the CPDC metric defined in (3.4) is a reasonable metric for our problem.

If in a policy all the observations are used for decision making, then the

[2]The delay metric considered in [12] and [21] is actually $\sup_\gamma \operatorname{ess\,sup}\ \mathbb{E}_\gamma\left[(\tau-\gamma+1)^+|\mathcal{I}_{\gamma-1}\right]$. However, these two metrics are equivalent as the WADD goes to infinity.

CPDC for that policy is 1. If every alternate observation is used, then the CPDC $= 0.5$.

Our first minimax formulation is the following data-efficient extension of Pollak [13]:

Problem 3.1.1.

$$\begin{aligned} \underset{\Psi}{\textit{minimize}} \quad & \mathsf{CADD}(\Psi), \\ \textit{subject to} \quad & \mathsf{FAR}(\Psi) \leq \alpha, \\ \textit{and} \quad & \mathsf{CPDC}(\Psi) \leq \beta, \end{aligned} \tag{3.5}$$

where $0 \leq \alpha, \beta \leq 1$ *are given constraints.*

We are also interested in the data-efficient extension of the minimax formulation of Lorden [12]:

Problem 3.1.2.

$$\begin{aligned} \underset{\Psi}{\textit{minimize}} \quad & \mathsf{WADD}(\Psi), \\ \textit{subject to} \quad & \mathsf{FAR}(\Psi) \leq \alpha, \\ \textit{and} \quad & \mathsf{CPDC}(\Psi) \leq \beta, \end{aligned} \tag{3.6}$$

where $0 \leq \alpha, \beta \leq 1$ *are given constraints.*

With $\beta = 1$, Problem 3.1.1 reduces to the minimax formulation of Pollak in [13], and Problem 3.1.2 reduces to the minimax formulation of Lorden in [12].

In [11], the following algorithm called the CuSum algorithm is proposed:

Algorithm 3.1.1 (CuSum: Ψ_{C})**.** *Start with* $C_0 = 0$*, and update the statistic* C_n *as*

$$C_{n+1} = \left(C_n + \ell(X_{n+1})\right)^+,$$

where $(x)^+ = \max\{0, x\}$ *and* $\ell(X) = \log \frac{f_1(X)}{f_0(X)}$*. Stop at*

$$\tau_{\mathrm{C}} = \inf\{n \geq 1 : C_n > A\}.$$

Thus, in the CuSum algorithm, the log likelihood ratio of the observations is accumulated over time. If the accumulated log likelihood ratio becomes

negative, it is reset to zero. The CuSum algorithm can also be seen as a sequence of two-sided SPRTs; see [18]. It is shown by Lai in [14] that the CuSum algorithm is asymptotically optimal for both Problem 3.1.1 and Problem 3.1.2, with $\beta = 1$, as $\alpha \to 0$ (see Section 3.3.2 for precise statements about the CuSum algorithm).

However, note that the CPDC for the DE-CuSum algorithm is equal to 1. Hence, it cannot be a solution to Problem 3.1.2 and Problem 3.1.1, if $\beta < 1$. In the following we propose the DE-CuSum algorithm, an extension of the CuSum algorithm for the data-efficient setting, and show that it is asymptotically optimal, for each fixed β, as $\alpha \to 0$; see Section 3.3.5. The extension is motivated by the Bayesian analysis from Chapter 2.

3.2 The DE-CuSum Algorithm

We now present the DE-CuSum algorithm.

Algorithm 3.2.1 ($\mathrm{DE-CuSum}$: $\Psi_{\mathrm{w}}(A, \mu, h)$)**.** *Start with $W_0 = 0$ and fix $\mu > 0$, $A > 0$ and $h \geq 0$. For $n \geq 0$ use the following control:*

$$S_{n+1} = \begin{cases} 0 & \text{if } W_n < 0 \\ 1 & \text{if } W_n \geq 0 \end{cases},$$
$$\tau_{\mathrm{w}} = \inf\{n \geq 1 : W_n > A\}.$$

The statistic W_n is updated using the following recursions:

$$W_{n+1} = \begin{cases} \min\{W_n + \mu, 0\} & \text{if } S_{n+1} = 0 \\ (W_n + \ell(X_{n+1}))^{h+} & \text{if } S_{n+1} = 1, \end{cases}$$

where $(x)^{h+} = \max\{x, -h\}$ and $\ell(X) = \log \frac{f_1(X)}{f_0(X)}$.

When $h = \infty$, the DE-CuSum algorithm works as follows. The statistic W_n starts at 0, and evolves according to the CuSum algorithm until it goes below 0. When W_n goes below 0, it does so with an undershoot. Beyond this, W_n is incremented deterministically (by using the recursion $W_{n+1} = W_n + \mu$), and observations are skipped until W_n crosses 0 from below. As a consequence, the number of observations that are skipped is determined by the undershoot

(log likelihood ratio of the observations) as well as the parameter μ. When W_n crosses 0 from below, it is reset to 0 (this is the mathematical version of the statement that beyond the skipped samples, the DE-CuSum statistic is computed using fresh samples). Once $W_n = 0$, the process renews itself and continues to evolve this way until $W_n > A$, at which time a change is declared.

If $h < \infty$, W_n is truncated to $-h$ when W_n goes below 0 from above. In other words, the undershoot is reset to $-h$ if its magnitude is larger than h. A finite value of h guarantees that the number of consecutive samples skipped is bounded by $\frac{h}{\mu} + 1$. The parameter h can be selected based on practical considerations. This feature will also be crucial to the delay analysis of the DE-CuSum algorithm in Section 3.3.4.

If $h = 0$, the DE-CuSum statistic W_n never becomes negative and hence reduces to the CuSum statistic and evolves as: $W_0 = 0$, and for $n \geq 0$,

$$W_{n+1} = \max\{0, W_n + \ell(X_{n+1})\}.$$

Thus, μ is a substitute for the Bayesian prior ρ that is used in the DE-Shiryaev algorithm described in Chapter 2. But unlike ρ which represents a prior statistical knowledge of the change point, μ is a design parameter. An appropriate value of μ is selected to meet the constraint on CPDC; see Section 3.3.1 for details.

The evolution of the DE-CuSum algorithm is plotted in Fig. 3.1. In analogy with the evolution of the DE-Shiryaev algorithm, the DE-CuSum algorithm can also be seen as a sequence of independent two-sided tests. In each two-sided test a Sequential Probability Ratio Test (SPRT) [25], with log boundaries A and 0, is used to distinguish between the two hypotheses "H_0 = pre-change" and "H_1 = post-change". If the decision in the SPRT is in favor of H_0, then samples are skipped based on the likelihood ratio of all the observations taken in the SPRT. A change is declared the first time the decision in the sequence of SPRTs is in favor of H_1. If $h = 0$, no samples are skipped and the DE-CuSum reduces to the CuSum algorithm, i.e., to a sequence of SPRTs (also see [18]).

Unless it is required to have a bound on the maximum number of samples skipped, the DE-CuSum algorithm can be controlled by just two-parameters: A and μ. We will show in Section 3.3.1 that these two parameters can be

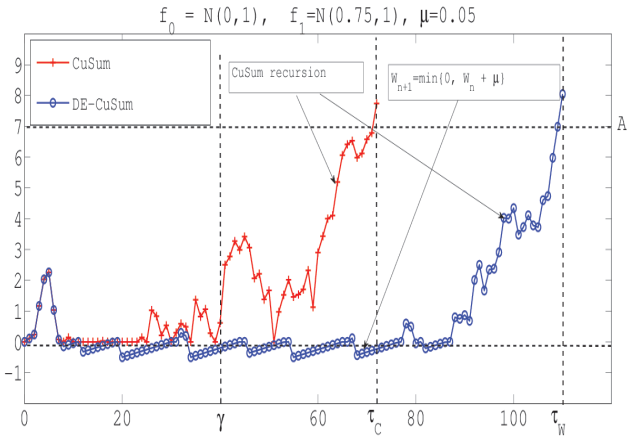


Figure 3.1: Typical evolution of the CuSum and the DE-CuSum algorithms applied to the same set of samples. Parameters used: $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.75,1)$, $\gamma = 40$, $A = 7$, and $\mu = 0.05$. With $h = 0.5$, the undershoots are truncated at -0.5.

selected independent of each other directly from the constraints. That is, the threshold A can be selected so that $\mathsf{FAR} \leq \alpha$ independent of the value of μ. Also, it is possible to select a value of μ such that $\mathsf{CPDC} \leq \beta$ independent of the choice of A.

Remark 3.2.1. *With the way the DE-CuSum algorithm is defined, we will see in the following that it may not be possible to meet* CPDC *constraints that are close to 1, with equality. We ignore this issue in what follows, as in many practical settings the preferred value of* CPDC *would be closer to 0 than 1. But, we remark that the DE-CuSum algorithm can be easily modified to achieve* CPDC *values that are close to 1 by resetting W_n to zero if the undershoot is smaller than a pre-designed threshold. See [26].*

Remark 3.2.2. *One can also modify the Shiryaev-Roberts algorithm [24] and obtain a two-threshold version of it, with an upper threshold used for stopping and a lower threshold used for on-off observation control. Also note that the SPRTs of the two-sides tests in the DE-CuSum algorithm have a lower threshold of 0. One can also propose variants of the DE-CuSum algorithm with a negative lower threshold for the SPRTs.*

Remark 3.2.3. *For the CuSum algorithm, the supremum in* (3.2) *and* (3.3) *is achieved when the change is applied at time $\gamma = 1$ (see also* (3.19)*). This is useful from the point of view of simulating the test. However, in the data-efficient setting, since the information vector also contains information about*

missed samples, the worst case change point in (3.2) *would depend on the observation control and may not be* $n = 1$*. But note that in the DE-CuSum algorithm, the test statistic evolves as a Markov process. As a result, the worst case usually occurs in the initial slots, before the process hits stationarity. This is useful from the point of view of simulating the algorithm. In the analysis of the DE-CuSum algorithm provided in Section 3.3, we will see that the* WADD *of the DE-CuSum algorithm is equal to its delay when change occurs at* $\gamma = 1$*, plus a constant. Similarly, even if computing the* CPDC *may be a bit difficult using simulations, we will provide a simple numerically-computable upper bound on the* CPDC *of the DE-CuSum algorithm that can be used to ensure that the* CPDC *constraint is satisfied. We will also provide a simple approximation to the* CPDC*. This approximation can be used to approximately satisfy the constraint on the* CPDC*.*

3.3 Analysis and Design of the DE-CuSum Algorithm

The identification or intepretation of the DE-CuSum algorithm as a sequence of two-sided tests will now be used in this section to perform its asymptotic analysis. In the rest of the chapter, we use $L(X)$ and $\ell(X)$ to denote $\frac{f_1(X)}{f_0(X)}$ and $\log \frac{f_1(X)}{f_0(X)}$, respectively.

Recall that the DE-CuSum algorithm can be seen as a sequence of two-sided tests, each two-sided test contains an SPRT and a possible sojourn below zero. The length of the latter is dependent on the likelihood ratio of the observations. To capture these quantities mathematically we now define some new variables.

Define the stopping time for an SPRT

$$\lambda_{\text{A}} \triangleq \inf\{n \geq 1 : \sum_{k=1}^{n} \ell(X_k) \notin [0, A]\}. \tag{3.7}$$

To capture the sojourn time below 0, define for $x < 0$

$$T(x) = \lceil |(x)^{h+}|/\mu \rceil. \tag{3.8}$$

Note that $T(0) = 0$. We also define the stopping time for the two-sided test

$$\Lambda_{\text{A}} = \lambda_{\text{A}} + T(W_{\lambda_{\text{A}}})\, \mathbb{I}_{\{W_{\lambda_{\text{A}}} < 0\}}. \tag{3.9}$$

Let λ_∞ and Λ_∞ be the variables λ_{A} and Λ_{A}, respectively, when the threshold $A = \infty$.

The variables λ_{A}, Λ_{A} and $T(x)$ should be interpreted as follows. The DE-CuSum algorithm can be seen as a sequence of two-sided tests, with the stopping time of each two-sided test distributed accordingly to the law of Λ_{A}. Each of the above two-sided tests consists of an SPRT with stopping time distributed accordingly to the law of λ_{A}, and a sojourn of length $T(W_{\lambda_{\text{A}}})$ corresponding to the time for which the statistic W_n is below 0, provided that at the stopping time for the SPRT, the accumulated log likelihood is negative, i.e., the event $\{W_{\lambda_{\text{A}}} < 0\}$ happens. See Fig. 3.2. In the figure, $\Lambda_1, \Lambda_2, \cdots$ are random variables distributed accordingly to the law of Λ_{A}, and $\lambda_1, \lambda_2, \cdots$ are random variables distributed accordingly to the law of λ_{A}.

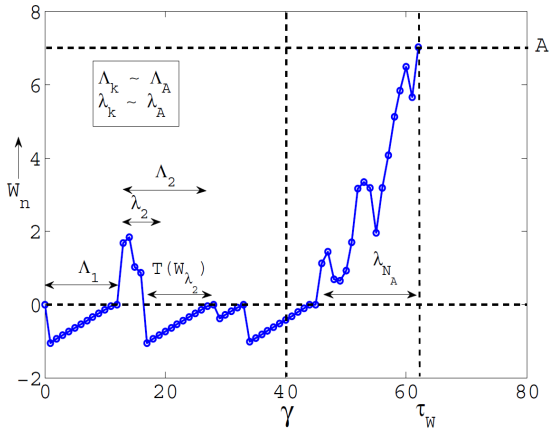


Figure 3.2: Evolution of W_n for $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.75,1)$, and $\Gamma = 40$, with $A = 7$, $h = \infty$, and $\mu = 0.1$. The two-sided tests with distribution of Λ_{A} are shown in the figure. Also shown are the two components of Λ_{A}: λ_{A} and $T(x)$.

The CuSum algorithm can also be seen as a sequence of SPRTs, with the stopping time of each SPRT distributed according to the law of λ_{A} (see [18]).

We now provide some results on the mean of λ_{A} and $T(x)$ that will be used in the analysis of the DE-CuSum algorithm in Sections 3.3.1, 3.3.3 and 3.3.4.

If $0 < D(f_0 \parallel f_1) < \infty$, then from Corollary 2.4 in [16],

$$\mathbb{E}_\infty[\lambda_\infty] < \infty, \tag{3.10}$$

and by Wald's lemma

$$\mathbb{E}_\infty[|W_{\lambda_\infty}|] = D(f_0 \parallel f_1)\, \mathbb{E}_\infty[\lambda_\infty] < \infty. \tag{3.11}$$

Also for $h \geq 0$

$$\mathbb{E}_\infty[|W^{h+}_{\lambda_\infty}|] \leq \mathbb{E}_\infty[|W_{\lambda_\infty}|] < \infty, \tag{3.12}$$

where the finiteness follows from (3.11).

In the next lemma we show that the quantity $\mathbb{E}_\infty[\lambda_{\text{A}}|W_{\lambda_{\text{A}}} < 0]$ is finite for every A and provide a finite upper bound to it that is not a function of the threshold A. This result will be used in the CPDC analysis in Section 3.3.1.

Lemma 3.3.1. *If $0 < D(f_0 \,||\, f_1) < \infty$, then for any A, $\mathbb{E}_\infty[\lambda_{\text{A}}|W_{\lambda_{\text{A}}} < 0]$ is well defined and finite:*

$$\mathbb{E}_\infty[\lambda_{\text{A}}|W_{\lambda_{\text{A}}} < 0] \leq \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{P}_\infty(\ell(X_1) < 0)} < \infty.$$

Proof. The proof of the first inequality is provided in Section 3.5. The second inequality is true by (3.10) and because $\mathbb{P}_\infty(\ell(X_1) < 0) > 0$. □

In the next lemma we provide upper and lower bounds on $\mathbb{E}_\infty[T(W_{\lambda_{\text{A}}})|W_{\lambda_{\text{A}}} < 0]$ that are not a function of the threshold A. The upper bound will be useful in the FAR analysis in Section 3.3.3, and the lower bound will be useful in the CPDC analysis in Section 3.3.1. Define

$$T_L^{(\infty)} = \frac{\mathbb{E}_\infty[|\ell(X_1)^{h+}| \,;\, \{\ell(X_1) < 0\}]}{\mu}, \tag{3.13}$$

and

$$T_U^{(\infty)} = \frac{\mathbb{E}_\infty[|W^{h+}_{\lambda_\infty}|]}{\mu\, \mathbb{P}_\infty(\ell(X_1) < 0)} + 1. \tag{3.14}$$

Lemma 3.3.2. *If $0 < D(f_0 \parallel f_1) < \infty$ and $\mu > 0$, then*

$$T_L^{(\infty)} \leq\ \mathbb{E}_\infty[T(W_{\lambda_{\text{A}}})\Big|W_{\lambda_{\text{A}}} < 0]\ \leq\ T_U^{(\infty)}. \tag{3.15}$$

Moreover, $T_U^{(\infty)} < \infty$*, and if* $h > 0$*, then* $T_L^{(\infty)} > 0$*.*

Proof. The proof is provided in the Section 3.5. □

3.3.1 Meeting the CPDC Constraint

In this section we show that the CPDC metric is well defined for the DE-CuSum algorithm. In general $\mathsf{CPDC}(\Psi_{\mathrm{W}})$ will depend on both A and μ (apart from the obvious dependence on f_0 and f_1). But, we show that it is possible to choose a value of μ that ensures that the CPDC constraint of β can be met independent of the choice of A. The latter is crucial to the asymptotic optimality proof of the DE-CuSum algorithm provided later in Section 3.3.5.

Theorem 3.3.1. *For fixed values of* A*,* h*, and* $\mu > 0$*, if* $0 < D(f_0 \,||\, f_1) < \infty$*, then*

$$\mathsf{CPDC}(\Psi_{\mathrm{W}}(A,\mu,h)) = \frac{\mathbb{E}_\infty[\lambda_{\mathrm{A}} | W_{\lambda_{\mathrm{A}}} < 0]}{\mathbb{E}_\infty[\lambda_{\mathrm{A}} | W_{\lambda_{\mathrm{A}}} < 0] + \mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}}) \,\Big|\, W_{\lambda_{\mathrm{A}}} < 0]}. \tag{3.16}$$

Proof. Consider an alternating renewal process $\{V_n, U_n\}$, i.e, a renewal process with renewal times $\{V_1, V_1+U_1, V_1+U_1+V_2, \cdots\}$, with $\{V_n\}$ i.i.d. with distribution of λ_{A} conditioned on $\{W_{\lambda_{\mathrm{A}}} < 0\}$, and $\{U_n\}$ i.i.d. with distribution of $T(W_{\lambda_{\mathrm{A}}})$ conditioned on $\{W_{\lambda_{\mathrm{A}}} < 0\}$. Thus,

$$\mathbb{E}_\infty[V_1] = \mathbb{E}_\infty[\lambda_{\mathrm{A}} | W_{\lambda_{\mathrm{A}}} < 0],$$

and

$$\mathbb{E}_\infty[U_1] = \mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}}) \,\Big|\, W_{\lambda_{\mathrm{A}}} < 0].$$

Both the means are finite by Lemma 3.3.1 and Lemma 3.3.2.

At time n assign a reward of $R_n = 1$ if the renewal cycle in progress has the law of V_1, set $R_n = 0$ otherwise. Then by renewal reward theorem,

$$\frac{1}{n}\mathbb{E}_\infty\left[\sum_{k=1}^{n-1} R_k\right] \to \frac{\mathbb{E}_\infty[V_1]}{\mathbb{E}_\infty[V_1] + \mathbb{E}_\infty[U_1]}.$$

On $\{\tau_{\mathrm{W}} \geq n\}$, the total number of observations taken until time $n-1$ has the same distribution as the total reward for the alternating renewal process defined above. Hence, the expected value of the average reward for both the sequences must have the same limit:

$$\begin{aligned}\mathsf{CPDC}(\Psi_{\mathrm{W}}(A,\mu,h)) &= \lim_{n\to\infty}\frac{1}{n}\mathbb{E}_n\left[\sum_{k=1}^{n-1} S_k \Big| \tau_{\mathrm{W}} \geq n\right] \\ &= \frac{\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}}<0]}{\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}}<0]+\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})|W_{\lambda_{\mathrm{A}}}<0]}.\end{aligned} \tag{3.17}$$

□

If $h=0$, then $\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})|W_{\lambda_{\mathrm{A}}}<0]=0$ and we get the CPDC of the CuSum algorithm that is equal to 1.

As can be seen from (3.16), CPDC is a function of A as well as that of h and μ. We now show that for any A and $h>0$, the DE-CuSum algorithm can be designed to meet any CPDC constraint of β. Moreover, for a given $h>0$, a value of μ can always be selected such that the CPDC constraint of β is met independent of the choice of A. The latter is convenient not only from a practical point of view, but will also help in the asymptotic optimality proof of the DE-CuSum algorithm in Section 3.3.5.

Theorem 3.3.2. *For the DE-CuSum algorithm, for any choice of A and $h>0$, if $0<D(f_0\ ||\ f_1)<\infty$, then we can always choose a value of μ to meet any given* CPDC *constraint of β. Moreover, for any fixed value of $h>0$, there exists a value of μ say $\mu^*(h)$ such that for every A,*

$$\mathsf{CPDC}(\Psi_{\mathrm{W}}(A,\mu^*,h)) \leq \beta.$$

In fact any μ that satisfies

$$\mu \leq \frac{\mathbb{E}_\infty[|\ell(X_1)^{h+}| \,\Big|\, \ell(X_1)<0]\,\mathbb{P}_\infty(\ell(X_1)<0)^2}{\mathbb{E}_\infty[\lambda_\infty]}\frac{\beta}{1-\beta},$$

can be used as μ^*.

Proof. Note that $\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}}<0]$ is not affected by the choice of h and μ.

Moreover, from Lemma 3.3.2

$$\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}}) \;\Big|\; W_{\lambda_{\mathrm{A}}} < 0] \geq T_L^{(\infty)}.$$

Thus from (3.13), for a given A and h,

$$\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})|W_{\lambda_{\mathrm{A}}} < 0] \to \infty \text{ as } \mu \to 0.$$

Therefore, we can always select a μ small enough so that the CPDC is smaller than the given constraint of β.

Next, our aim is to find a μ^* such that for every A

$$\frac{\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}} < 0]}{\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}} < 0] + \mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})) \;\Big|\; W_{\lambda_{\mathrm{A}}} < 0]} \leq \beta.$$

Since CPDC increases as $\mathbb{E}_\infty[\lambda_{\mathrm{A}}|W_{\lambda_{\mathrm{A}}} < 0]$ increases and $\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})\ W_{\lambda_{\mathrm{A}}} < 0]$ decreases, we have from Lemma 3.3.1 and Lemma 3.3.2,

$$\mathsf{CPDC}(\Psi_{\mathrm{W}}) \leq \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + T_L^{(\infty)}\ \mathbb{P}_\infty(\ell(X_1) < 0)}. \tag{3.18}$$

Then, the theorem is proved if we select μ such that the right-hand side of (3.18) is less than β or using (3.13), a μ that satisfies

$$\mu \leq \frac{\mathbb{E}_\infty[|\ell(X_1)^{h+}| \;\Big|\; \ell(X_1) < 0]\ \mathbb{P}_\infty(\ell(X) < 0)^2}{\mathbb{E}_\infty[\lambda_\infty]} \frac{\beta}{1-\beta}.$$

□

While the existence of μ^* proved by Theorem 3.3.2 is critical for asymptotic optimality of the DE-CuSum algorithm, the estimate it provides when substituted for μ in (3.16) may be a bit conservative. In Section 3.3.6 we provide a good approximation to CPDC that can be used to choose the value of μ in practice. In Section 3.4 we provide numerical results showing the accuracy of the approximation.

Remark 3.3.1. *By Theorem 3.3.2, for any value of h, we can select a value of μ small enough, so that any* CPDC *constraint close to zero can be met with equality. However, meeting the* CPDC *constraint with equality may not*

be possible if β is close to 1. This is because if $h \neq 0$ then

$$\mathsf{CPDC}(\Psi_{\mathrm{W}}) \leq \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + \mathbb{P}_\infty(\ell(X) < 0)} < 1.$$

However, as mentioned earlier, for most practical applications β will be closer to zero than 1, and hence this issue will not be encountered. If β closer to 1 is indeed desired then the DE-CuSum algorithm can be easily modified to address this issue (by skipping samples only when the undershoot is larger than a pre-designed threshold).

3.3.2 Analysis of the CuSum Algorithm

In the sections to follow, we will express the performance of the DE-CuSum algorithm in terms of the performance of the CuSum algorithm. Therefore, in this section we summarize the performance of the CuSum algorithm.

It is easy to show (see [1]) that

$$\mathsf{CADD}(\Psi_{\mathrm{C}}) = \mathsf{WADD}(\Psi_{\mathrm{C}}) = \mathbb{E}_1[\tau_{\mathrm{C}} - 1]. \tag{3.19}$$

From [12], if $0 < D(f_1 \mid\mid f_0) < \infty$, then $\mathbb{E}_1[\tau_{\mathrm{C}}] < \infty$. Moreover, if $\{\lambda_1, \lambda_2, \cdots\}$ are i.i.d. random variables each with distribution of λ_{A}, then by Wald's lemma [18]

$$\mathbb{E}_1[\tau_{\mathrm{C}}] = \mathbb{E}_1\left[\sum_{k=1}^{N} \lambda_k\right] = \mathbb{E}_1[N]\ \mathbb{E}_1[\lambda_{\mathrm{A}}], \tag{3.20}$$

where N is the number of two-sided tests (SPRTs)–each with distribution of λ_{A}–executed before the change is declared.

It is also shown in [12] that $0 < D(f_1 \mid\mid f_0) < \infty$ is also sufficient to guarantee $\mathbb{E}_\infty[\tau_{\mathrm{C}}] < \infty$ and $\mathsf{FAR}(\Psi_{\mathrm{C}}) > 0$. Moreover,

$$\mathbb{E}_\infty[\tau_{\mathrm{C}}] = \mathbb{E}_\infty\left[\sum_{k=1}^{N} \lambda_k\right] = \mathbb{E}_\infty[N]\ \mathbb{E}_\infty[\lambda_{\mathrm{A}}]. \tag{3.21}$$

The proof of the following theorem can be found in [12] and [14].

Theorem 3.3.3. *If $0 < D(f_1 \mid\mid f_0) < \infty$, then with $A = \log \frac{1}{\alpha}$,*

$$\mathsf{FAR}(\Psi_{\mathrm{C}}) \leq \alpha,$$

and as $\alpha \to 0$,

$$\mathsf{CADD}(\Psi_{\mathrm{C}}) = \mathsf{WADD}(\Psi_{\mathrm{C}}) = \mathbb{E}_1[\tau_{\mathrm{C}} - 1] \sim \frac{|\log \alpha|}{D(f_1 \mid\mid f_0)}.$$

Thus, the CuSum algorithm is asymptotically optimal for both Problem 3.1.2 and Problem 3.1.1, with $\beta = 1$, *as* $\alpha \to 0$. *This is because for any stopping time* τ *with* $\mathsf{FAR}(\tau) \leq \alpha$,

$$\mathsf{WADD}(\tau) \geq \mathsf{CADD}(\tau) \geq \frac{|\log \alpha|}{D(f_1 \mid\mid f_0)}\Big(1 + o(1)\Big), \tag{3.22}$$

as $\alpha \to 0$.

3.3.3 FAR for the DE-CuSum Algorithm

In this section we characterize the false alarm rate of the DE-CuSum algorithm. In the next lemma we show that for a fixed A, μ and h, if the DE-CuSum algorithm and the CuSum algorithm are applied to the same sequence of random variables, then sample-pathwise, the DE-CuSum statistic W_n is always below the CuSum statistic C_n. Thus, the DE-CuSum algorithm crosses the threshold A only after the CuSum algorithm has crossed it.

Lemma 3.3.3. *Under any* $\mathbb{P}_n$, $n \geq 1$ *and under* $\mathbb{P}_\infty$,

$$C_n \geq W_n.$$

Thus

$$\tau_{\mathrm{C}} \leq \tau_{\mathrm{W}}.$$

Proof. Note that initially $C_n = W_n$ until both the statistics become negative. When W_n goes below zero, it is incremented by μ until it reaches zero from below, at which time it is reset to zero. Since all the samples are taken in the CuSum algorithm, the time at which W_n reaches 0 from below, $C_n \geq 0$. That point onward, since the same sequence of observations are used to compute C_n and W_n, $C_n \geq W_n$, until W_n goes below 0. One can now repeat the arguments provided until now to claim that C_n will continue to stay above W_n throughout. Thus, if a sequence of samples causes the statistic of the DE-CuSum algorithm to go above A, then since all the samples are utilized

in the CuSum algorithm, the same sequence must also cause the CuSum statistic to go above A. □

It follows as a corollary of Lemma 3.3.3 that

$$\mathbb{E}_\infty[\tau_{\mathrm{C}}] \leq \mathbb{E}_\infty[\tau_{\mathrm{W}}].$$

The following theorem shows that these quantities are finite and also provides an estimate for $\mathsf{FAR}(\Psi_{\mathrm{W}})$.

Theorem 3.3.4. *For any fixed h (including $h = \infty$) and $\mu > 0$, if*

$$0 < D(f_0 \mid\mid f_1) < \infty \quad \text{and} \quad 0 < D(f_1 \mid\mid f_0) < \infty,$$

then with $A = \log \frac{1}{\alpha}$,

$$\mathsf{FAR}(\Psi_{\mathrm{W}}) \leq \mathsf{FAR}(\Psi_{\mathrm{C}}) \leq \alpha.$$

Moreover, for any A

$$\mathbb{E}_\infty[\tau_{\mathrm{W}}] = \mathbb{E}_\infty[\tau_{\mathrm{C}}] + \frac{\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})\, \mathbb{I}_{\{W_{\lambda_{\mathrm{A}}} < 0\}}]}{\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)} \tag{3.23}$$

and as $A \to \infty$,

$$\frac{\mathsf{FAR}(\Psi_{\mathrm{W}})}{\mathsf{FAR}(\Psi_{\mathrm{C}})} \to \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + \mathbb{E}_\infty[T(W_{\lambda_\infty})]}, \tag{3.24}$$

where λ_∞ is the variable λ_{A} with $A = \infty$. The limit in (3.24) *is strictly less than 1 if $h > 0$.*

Proof. For a fixed A, let N_{A} be the number of two-sided tests of distribution Λ_{A} executed before the change is declared in the DE-CuSum algorithm. Then, if $\{\Lambda_1, \Lambda_2, \cdots\}$ is a sequence of random variables each with distribution of Λ_{A}, then

$$\mathbb{E}_\infty[\tau_{\mathrm{W}}] = \mathbb{E}_\infty\left[\sum_{k=1}^{N_{\mathrm{A}}} \Lambda_k\right].$$

Because of the renewal nature of the DE-CuSum algorithm,

$$\mathbb{E}_\infty[N_{\mathrm{A}}] = \mathbb{E}_\infty[N],$$

where N is the number of SPRTs used in the CuSum algorithm. Thus from (3.21),

$$\mathbb{E}_\infty[N_{\text{A}}] = \mathbb{E}_\infty[N] \leq \mathbb{E}_\infty[\tau_{\text{C}}] < \infty.$$

Further from (3.9),

$$\mathbb{E}_\infty[\Lambda_{\text{A}}] = \mathbb{E}_\infty[\lambda_{\text{A}}] + \mathbb{E}_\infty[T(W_{\lambda_{\text{A}}})\, \mathbb{I}_{\{W_{\lambda_{\text{A}}}<0\}}]. \tag{3.25}$$

From (3.21) again

$$\mathbb{E}_\infty[\lambda_{\text{A}}] \leq \mathbb{E}_\infty[\tau_{\text{C}}] < \infty.$$

Moreover from Lemma 3.3.2

$$\begin{aligned}\mathbb{E}_\infty[T(W_{\lambda_{\text{A}}})\, \mathbb{I}_{\{W_{\lambda_{\text{A}}}<0\}}] &\leq \mathbb{E}_\infty[T(W_{\lambda_{\text{A}}}) \mid W_{\lambda_{\text{A}}} < 0] \\ &\leq T_U^{(\infty)} < \infty.\end{aligned}$$

Thus, $\mathbb{E}_\infty[\Lambda_{\text{A}}] < \infty$ and

$$\mathbb{E}_\infty[\tau_{\text{W}}] = \mathbb{E}_\infty\left[\sum_{k=1}^{N_{\text{A}}} \Lambda_k\right] = \mathbb{E}_\infty[N_{\text{A}}]\, \mathbb{E}_\infty[\Lambda_{\text{A}}] < \infty.$$

Hence,

$$\begin{aligned}\mathbb{E}_\infty[\tau_{\text{W}}] &= \mathbb{E}_\infty[N_{\text{A}}]\, \mathbb{E}_\infty[\Lambda_{\text{A}}] \\ &= \mathbb{E}_\infty[N]\, \mathbb{E}_\infty[\Lambda_{\text{A}}] \\ &\geq \mathbb{E}_\infty[N]\, \mathbb{E}_\infty[\lambda_{\text{A}}] \\ &= \mathbb{E}_\infty[\tau_{\text{C}}].\end{aligned}$$

We note that the statement also follows as a corollary of Lemma 3.3.3 and Theorem 3.3.3. And with $A = \log \frac{1}{\alpha}$,

$$\mathsf{FAR}(\Psi_{\text{W}}) \leq \mathsf{FAR}(\Psi_{\text{C}}) \leq \alpha.$$

Since, N_{A} is Geom($\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)$), (3.23) follows from (3.25) and (3.21):

$$\begin{aligned}\mathbb{E}_\infty[\tau_{\mathrm{W}}] &= \frac{\mathbb{E}_\infty[\Lambda_{\mathrm{A}}]}{\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)} \\ &= \frac{\mathbb{E}_\infty[\lambda_{\mathrm{A}}]}{\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)} + \frac{\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})\, \mathbb{I}_{\{W_{\lambda_{\mathrm{A}}} < 0\}}]}{\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)} \\ &= \mathbb{E}_\infty[\tau_{\mathrm{C}}] + \frac{\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}})\, \mathbb{I}_{\{W_{\lambda_{\mathrm{A}}} < 0\}}]}{\mathbb{P}_\infty(W_{\lambda_{\mathrm{A}}} > 0)}.\end{aligned} \tag{3.26}$$

Further, since $\mathbb{E}_\infty[N_{\mathrm{A}}] = \mathbb{E}_\infty[N]$, we have

$$\frac{\mathbb{E}_\infty[\tau_{\mathrm{C}}]}{\mathbb{E}_\infty[\tau_{\mathrm{W}}]} = \frac{\mathbb{E}_\infty[N]\, \mathbb{E}_\infty[\lambda_{\mathrm{A}}]}{\mathbb{E}_\infty[N_{\mathrm{A}}]\, \mathbb{E}_\infty[\Lambda_{\mathrm{A}}]} = \frac{\mathbb{E}_\infty[\lambda_{\mathrm{A}}]}{\mathbb{E}_\infty[\Lambda_{\mathrm{A}}]}.$$

Since, $\lambda_{\mathrm{A}} \uparrow \lambda_\infty$ and $\Lambda_{\mathrm{A}} \uparrow \Lambda_\infty$, we have by monotone convergence theorem, as $A \to \infty$,

$$\frac{\mathbb{E}_\infty[\tau_{\mathrm{C}}]}{\mathbb{E}_\infty[\tau_{\mathrm{W}}]} \to \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\Lambda_\infty]} = \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + \mathbb{E}_\infty[T(W_{\lambda_\infty})]}.$$

The limit is clearly less than 1 if $h > 0$. □

Remark 3.3.2. *Thus, unlike the Bayesian setting where the* PFA *of the DE-Shiryaev algorithm converges to the* PFA *of the Shiryaev algorithm (for the nonarithmetic case), here, the* FAR *of the DE-CuSum algorithm is strictly less than the* FAR *of the CuSum algorithm. Moreover, for large A, the right side of* (3.24) *is approximately the* CPDC *achieved. Thus,* (3.24) *shows that, asymptotically as $A \to \infty$, the ratio of the* FAR*s is approximately equal to the* CPDC*. This also shows that one can set the threshold in the DE-CuSum algorithm to a value smaller than $A = \frac{1}{\alpha}$ to meet the* FAR *constraint with equality. This latter fact will be used in obtaining the numerical results in Section 3.4.*

3.3.4 CADD and WADD of the DE-CuSum Algorithm

We now provide the delay analysis of the DE-CuSum algorithm. The main content of Theorem 3.3.5 below is, that for each value of A, the WADD of the DE-CuSum algorithm is within a constant of the corresponding performance of the CuSum algorithm. This constant is independent of the choice of A, and as a result the WADD of the two algorithms are asymptotically the same.

Because of the virtue of the relation $\mathsf{CADD} \leq \mathsf{WADD}$ for any policy, and due to the fact that $\mathsf{CADD}(\tau_{\mathrm{C}}) = \mathsf{WADD}(\tau_{\mathrm{C}})$, even the CADD of the DE-CuSum algorithm is within a constant of the CADD of the CuSum algorithm. A critical assumption made in these results is that $h < \infty$, i.e., the number of consecutive samples skipped is bounded by $\lceil h/\mu \rceil < \infty$.

The results provided below depend on the following fundamental lemma. In the lemma it is shown that when the change happens at $\gamma = 1$, then the average delay of the DE-CuSum algorithm starting with $W_0 = x > 0$, is upper bounded by the average delay of the algorithm when $W_0 = 0$, plus a constant which is $\lceil h/\mu \rceil < \infty$.

Let

$$\tau_{\mathrm{W}}(x) = \inf\{n \geq 1 : W_n > A; W_0 = x\}. \tag{3.27}$$

Here, W_n is the DE-CuSum statistic and evolves according the description of the algorithm in Section 3.2. Thus, $\tau_{\mathrm{W}}(x)$ is the first time for the DE-CuSum algorithm to cross A, when starting at $W_0 = x$. Clearly, $\tau_{\mathrm{W}}(x) = \tau_{\mathrm{W}}$ if $x = 0$.

Lemma 3.3.4. *Let $0 < D(f_1 \mid\mid f_0) < \infty$ and $0 \leq x < A$, and moreover $h < \infty$, then*

$$\mathbb{E}_1[\tau_{\mathrm{W}}(x)] \leq \mathbb{E}_1[\tau_{\mathrm{W}}] + \lceil h/\mu \rceil.$$

Proof. The proof is provided in Section 3.5. □

We now express the WADD of the DE-CuSum algorithm in terms of the WADD of the CuSum algorithm.

Theorem 3.3.5. *Let*

$$0 < D(f_1 \mid\mid f_0) < \infty.$$

Then, for fixed values of $\mu > 0$ and $h < \infty$, and for each A,

$$\mathsf{WADD}(\Psi_{\mathrm{W}}) \leq \mathsf{WADD}(\Psi_{\mathrm{C}}) + K_{DC},$$

and

$$\mathsf{CADD}(\Psi_{\mathrm{W}}) \leq \mathsf{CADD}(\Psi_{\mathrm{C}}) + K_{DC},$$

where K_{DC} is a constant not a function of A. Thus, as $A \to \infty$,

$$\mathsf{WADD}(\Psi_{\mathrm{W}}) \leq \mathsf{WADD}(\Psi_{\mathrm{C}}) + O(1),$$

and

$$\mathsf{CADD}(\Psi_{\mathrm{W}}) \leq \mathsf{CADD}(\Psi_{\mathrm{C}}) + O(1).$$

Proof. From Lemma 3.3.4, it follows that for $n > 1$

$$\operatorname{ess\,sup}\ \mathbb{E}_\gamma \left[(\tau_{\mathrm{W}} - \gamma)^+ | \mathcal{I}_{\gamma-1}\right] \leq \lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\mathrm{W}}]. \tag{3.28}$$

This is because if $\mathcal{I}_{\gamma-1}$ is such that $W_{\gamma-1} = x \in [0, A)$, then the inequality in (3.28) is true for each such x due to Lemma 3.3.4. On the other hand if $\mathcal{I}_{\gamma-1}$ is such that $W_{\gamma-1} = x < 0$, then the time to hit A is the time spent by the DE-CuSum statistic to reach 0 from below, plus $\mathbb{E}_1[\tau_{\mathrm{W}}]$. The latter is true because a renewal occurs once the DE-CuSum statistic reaches 0 from below. Under the assumption that $h < \infty$, the time to hit 0 starting from $x < 0$ is clearly bounded by $\lceil h/\mu \rceil$.

Since the right-hand side in (3.28) is not a function of γ and it is greater than $\mathbb{E}_1[\tau_{\mathrm{W}} - 1]$ (corresponding to $\gamma = 1$), we have

$$\mathsf{WADD}(\Psi_{\mathrm{W}}) \leq \lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\mathrm{W}}].$$

Thus, from the proof of Theorem 3.3.4 (see (3.26)) with $\mathbb{E}_\infty$ replaced by $\mathbb{E}_1$ we have

$$\mathbb{E}_1[\tau_{\mathrm{W}}] = \mathbb{E}_1[\tau_{\mathrm{C}}] + \frac{\mathbb{E}_1[T(W_{\lambda_{\mathrm{A}}})\, \mathbb{I}_{\{W_{\lambda_{\mathrm{A}}} < 0\}}]}{\mathbb{P}_1(W_{\lambda_{\mathrm{A}}} > 0)}.$$

But,

$$\mathbb{E}_1[T(W_{\lambda_{\mathrm{A}}})\, \mathbb{I}_{\{W_{\lambda_{\mathrm{A}}} < 0\}}] \leq \lceil h/\mu \rceil,$$

and from [16] we have

$$\mathbb{P}_1(W_{\lambda_{\mathrm{A}}} > 0) \geq \mathbb{P}_1(W_{\lambda_\infty} > 0) > 0.$$

Thus, from (3.19) we have

$$\begin{aligned}\mathbb{E}_1[\tau_{\mathrm{W}}] &\leq \mathbb{E}_1[\tau_{\mathrm{C}}] + \frac{\lceil h/\mu \rceil}{\mathbb{P}_1(W_{\lambda_\infty} > 0)} \\ &= \mathsf{WADD}(\Psi_{\mathrm{C}}) + \frac{\lceil h/\mu \rceil}{\mathbb{P}_1(W_{\lambda_\infty} > 0)} + 1,\end{aligned}$$

and we have

$$\mathsf{WADD}(\Psi_{\mathrm{W}}) \leq \mathsf{WADD}(\Psi_{\mathrm{C}}) + \frac{\lceil h/\mu \rceil}{\mathbb{P}_1(W_{\lambda_\infty} > 0)} + \lceil h/\mu \rceil + 1.$$

This proves the theorem for the WADD.

For the CADD note that

$$\mathsf{CADD}(\Psi_{\mathrm{W}}) \leq \mathsf{WADD}(\Psi_{\mathrm{W}}) \leq \mathsf{WADD}(\Psi_{\mathrm{C}}) + K_{\mathrm{DC}} = \mathsf{CADD}(\Psi_{\mathrm{C}}) + K_{\mathrm{DC}}.$$

The proof of the theorem is complete. □

The following corollary follows easily from Theorem 3.3.3 and Theorem 3.3.5.

Corollary 3.3.5.1. *If $0 < D(f_1 \mid\mid f_0) < \infty$, then for fixed values of μ and $h < \infty$, as $A \to \infty$,*

$$\mathsf{CADD}(\Psi_{\mathrm{W}}) \sim \frac{A}{D(f_1 \mid\mid f_0)},$$

and

$$\mathsf{WADD}(\Psi_{\mathrm{W}}) \sim \frac{A}{D(f_1 \mid\mid f_0)}.$$

3.3.5 Asymptotic Optimality of the DE-CuSum Algorithm

We now use the results from the previous sections to show that the DE-CuSum algorithm is asymptotically optimal.

In the following theorem it is shown that for a given CPDC constraint of β, the DE-CuSum algorithm is asymptotically optimal for both Problem 3.1.2 and Problem 3.1.1, as $\alpha \to 0$, for the following reasons:

- The CPDC of the DE-CuSum algorithm can be designed to meet the constraint independent of the choice of A.
- The CADD and WADD of the DE-CuSum algorithm approaches the corresponding performances of the CuSum algorithm.
- The FAR of the DE-CuSum algorithm is always better than that of the CuSum algorithm.
- The CuSum algorithm is asymptotically optimal for both Problem 3.1.2 and Problem 3.1.1, with $\beta = 1$, as $\alpha \to 0$.

Theorem 3.3.6. *Let $0 < D(f_1 \,||\, f_0) < \infty$ and $0 < D(f_0 \,||\, f_1) < \infty$. For a given α, set $A = \log \frac{1}{\alpha}$, then for any choice of h and μ,*

$$\mathsf{FAR}(\Psi_{\mathrm{W}}) \leq \mathsf{FAR}(\Psi_{\mathrm{C}}) \leq \alpha.$$

For a given β, and for any given h, it is possible to select $\mu = \mu^*(h)$ such that $\forall A$, and hence even with $A = \log \frac{1}{\alpha}$ for any value of $\alpha > 0$,

$$\mathsf{CPDC}(\Psi_{\mathrm{W}}) \leq \beta.$$

Moreover, for each fixed β, for any $h < \infty$ and with $\mu^*(h)$ selected to meet this CPDC *constraint of β, as $\alpha \to 0$ (or $A \to \infty$ because $A = \log \frac{1}{\alpha}$),*

$$\mathsf{CADD}(\Psi_{\mathrm{W}}(\log \frac{1}{\alpha}, h, \mu^*(h))) \sim \mathsf{CADD}(\Psi_{\mathrm{C}}) \sim \frac{|\log \alpha|}{D(f_1 \,||\, f_0)},$$

and

$$\mathsf{WADD}(\Psi_{\mathrm{W}}(\log \frac{1}{\alpha}, h, \mu^*(h))) \sim \mathsf{WADD}(\Psi_{\mathrm{C}}) \sim \frac{|\log \alpha|}{D(f_1 \,||\, f_0)}.$$

Proof. The result on FAR follows from Theorem 3.3.4. The fact that one can select a $\mu = \mu^*(h)$ to meet the CPDC constraint independent of the choice of A follows from Theorem 3.3.2. Finally, the delay asymptotics follow from Theorem 3.3.5 and Corollary 3.3.5.1. □

Since, by Theorem 3.3.3, $\frac{|\log \alpha|}{D(f_1 \,||\, f_0)}$ is the best possible asymptotics performance for any given FAR constraint of α, Theorem 3.3.6 establishes the asymptotic optimality of the DE-CuSum algorithm for both Problem 3.1.1 and Problem 3.1.2, for each fixed β, as $\alpha \to 0$. We also note that the ability to set the CPDC independent of the threshold A is critical to the optimality result. In general, as we vary the FAR for a policy, its CPDC can change and can even violate the constraint of β.

We also note that the DE-CuSum algorithm is second-order asymptotically optimal (within a constant of the optimal, see [23]) for the Lorden's criterion. This is because the WADD of the DE-CuSum algorithm is within a constant of the WADD of the CuSum algorithm, and the latter is exactly optimal for the Lorden's criterion.

3.3.6 Design of the DE-CuSum Algorithm

We now discuss how to set the parameters μ, h and A so as to meet a given FAR constraint of α and a CPDC constraint of β.

Theorem 3.3.4 provides the guideline for choosing A: for any h, μ,

$$\text{if} \quad A = \log \frac{1}{\alpha} \qquad \text{then} \quad \mathsf{FAR}(\Psi_{\mathrm{W}}) \leq \alpha.$$

As discussed earlier, the CPDC constraint can be satisfied independent of the false alarm constraint. However, the estimate provided in Theorem 3.3.2 of the CPDC can be conservative. For practical purposes, we suggest using the following approximation for CPDC (obtained in the limit as $A \to \infty$):

$$\mathsf{CPDC} \approx \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + \mathbb{E}_\infty[T(W_{\lambda_\infty})]}. \tag{3.29}$$

For large values of A, (3.29) will indeed provide a good estimate of the CPDC. We note that $\mathbb{E}_\infty[\lambda_\infty]$ can be computed numerically; see Corollary 2.4 in [16]. The quantity $\mathbb{E}_\infty[T(W_{\lambda_\infty})]$ can be computed using Monte Carlo simulations.

If $h = \infty$, or for h large, then using (3.11) we can further simplify (3.29) to

$$\mathsf{CPDC} \approx \frac{\mathbb{E}_\infty[\lambda_\infty]}{\mathbb{E}_\infty[\lambda_\infty] + \frac{\mathbb{E}_\infty[|W_{\lambda_\infty}|]}{\mu}} = \frac{\mu}{\mu + D(f_0 \,||\, f_1)}. \tag{3.30}$$

Thus, to ensure $\mathsf{CPDC} \leq \beta$, the approximation in (3.30) suggests selecting μ such that

$$\mu \leq \frac{\beta}{1-\beta} D(f_0 \,||\, f_1).$$

In Section 3.4 we provide numerical results in which we show that the approximation (3.30) indeed provides a good estimate of the CPDC when h is large.

3.4 Numerical Results

The asymptotic optimality of the DE-CuSum algorithm for all β does not guarantee good performance for moderate values of FAR. In Fig. 3.3, we plot the trade-off curves for the CuSum algorithm and the DE-CuSum algorithm, obtained using simulations. We plot the performance of the DE-CuSum

algorithm for two different CPDC constraints: $\beta = 0.5$ and $\beta = 0.25$. For simplicity we restrict ourself to the CADD performance for $h = \infty$ in this section. Similar performance comparisons can be obtained for both CADD and WADD with $h < \infty$.

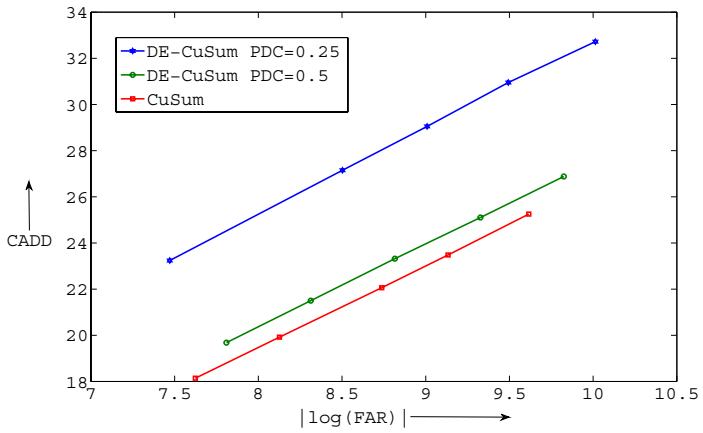


Figure 3.3: Trade-off curves for the DE-CuSum algorithm for CPDC $= 0.25, 0.5$, with $f_0 \sim \mathcal{N}(0,1)$ and $f_1 \sim \mathcal{N}(0.75,1)$.

Each of the curves for the DE-CuSum algorithm in Fig. 3.3 is obtained in the following way. Five different threshold values for A were arbitrarily selected. For each threshold value, a large value of γ was chosen, and the DE-CuSum algorithm was simulated and the fraction of time the observations are taken before change was computed. Specifically, γ was increased in the multiples of 100 and an estimate of the CPDC was obtained by Monte Carlo simulations. The value of μ was so chosen that the CPDC value obtained in simulations was slightly below the constraint $\beta = 0.5$ or 0.25. For this value of μ and for the chosen threshold, the FAR was computed by selecting the change time to be $\gamma = \infty$ (generating random numbers from $f_0 \sim \mathcal{N}(0,1)$). The CADD was then computed for the above choice of μ and A by varying the value of γ from $1, 2, \ldots$ and recording the maximum of the conditional delay. The maximum was achieved in the first five slots.

As can be seen from the figure, a CPDC of 0.5 (using only 50% of the samples in the long run) can be achieved using the DE-CuSum algorithm with a small penalty on the delay. If we wish to achieve a CPDC of 0.25, then we have to incur a significant penalty (of approximately six slots in Fig. 3.3). But, note that the difference of delay with the CuSum algorithm remains fixed as FAR $\to 0$. This is due to the result reported in Theorem 3.3.5 and

this is precisely the reason the DE-CuSum algorithm is asymptotic optimal. The trade-off between CADD and FAR is a function of the K-L divergence between p.d.f. f_1 and p.d.f. f_0: the larger the K-L divergence the more is the fraction of samples that can be dropped for a given loss in delay performance.

In Fig. 3.4 we compare the performance of the DE-CuSum algorithm with the *fraction sampling* scheme, in which, to achieve a CPDC of β, the CuSum algorithm is employed, and a sample is chosen with probability β for decision making. Note that this scheme skips samples without exploiting any knowledge about the state of the system. As seen in Fig. 3.4, the DE-CuSum algorithm performs considerably better than the fractional sampling scheme. Thus, the trade-off curves show that the DE-CuSum algorithm has good performance even for moderate FAR, when the CPDC constraint is moderate.

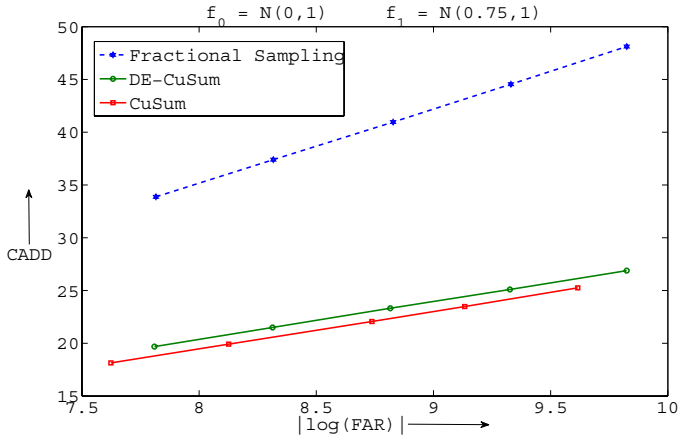


Figure 3.4: Comparative performance of the DE-CuSum algorithm with the CuSum algorithm and the fractional-sampling scheme: CPDC $= 0.5$, with $f_0 \sim \mathcal{N}(0,1)$ and $f_1 \sim \mathcal{N}(0.75,1)$.

We now provide numerical results that show that (3.30) provides a good estimate for the CPDC. We use the following parameters: $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.75,1)$ and set $h = \infty$. In Table 3.1a, we fix the value of μ and vary A and compare the CPDC obtained using simulations and the one obtained using (3.30), that is using the approximation CPDC $\approx \frac{\mu}{\mu + D(f_0||f_1)}$. We see that the approximation becomes more accurate as A increases. We also note that the CPDC obtained using simulations does not converge to $\frac{\mu}{\mu+D(f_0||f_1)}$, even as A becomes large, because of the effect of the presence of a ceiling function in the CPDC expression; see (3.8) and (3.16).

In Table 3.1b, we next fix a large value of A, specifically $A = 6$, for which the CPDC approximation is most accurate in Table 3.1a, and check the accuracy of the approximation $\frac{\mu}{\mu+D(f_0||f_1)}$ by varying μ. We see in the table that the approximation is more accurate for small values of μ. This is due to the fact that the effect of the ceiling function in the CPDC (3.8), (3.16) is negligible when μ is small.

Table 3.1: Comparison of CPDC obtained using simulations with the approximation (3.30) for $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(0.75,1)$ and $h = \infty$.

(a) Fixed μ

		CPDC	
A	μ	Simulations	Approx (3.30) $\frac{\mu}{\mu+D(f_0\|\|f_1)}$
1	0.1	0.16	0.26
2	0.1	0.20	0.26
3	0.1	0.22	0.26
4	0.1	0.238	0.26
6	0.1	0.248	0.26

(b) Fixed A

		CPDC	
A	μ	Simulations	Approx (3.30) $\frac{\mu}{\mu+D(f_0\|\|f_1)}$
6	0.01	0.033	0.034
6	0.05	0.145	0.151
6	0.2	0.37	0.41
6	0.3	0.46	0.51
6	0.4	0.51	0.58
6	0.6	0.58	0.68

3.5 Proofs of Various Results

Proof of Lemma 3.3.1. If $0 < D(f_0 \,||\, f_1) < \infty$, then $\mathbb{E}_\infty[\lambda_\infty] < \infty$. Thus, $\mathbb{P}_\infty(\lambda_\infty < \infty) = 1$. Also,

$$\mathbb{P}_\infty(W_{\lambda_A} < 0) > \mathbb{P}_\infty(\ell(X_1) < 0) > 0.$$

Thus, $\mathbb{E}_\infty[\lambda_{\text{A}} \Big| W_{\lambda_{\text{A}}} < 0]$ is well defined and

$$\begin{aligned}\mathbb{E}_\infty[\lambda_\infty] &\geq \mathbb{E}_\infty[\lambda_{\text{A}}] \\ &\geq \mathbb{E}_\infty[\lambda_{\text{A}} \Big| W_{\lambda_{\text{A}}} < 0]\, \mathbb{P}_\infty(W_{\lambda_{\text{A}}} < 0). \\ &\geq \mathbb{E}_\infty[\lambda_{\text{A}} \Big| W_{\lambda_{\text{A}}} < 0]\, \mathbb{P}_\infty(\ell(X_1) < 0).\end{aligned}$$

This proves the lemma. □

Proof of Lemma 3.3.2. Again note that

$$\mathbb{P}_\infty(W_{\lambda_{\text{A}}} < 0) > \mathbb{P}_\infty(\ell(X_1) < 0) > 0.$$

Thus, $\mathbb{E}_\infty[T(W_{\lambda_{\text{A}}}) \Big| W_{\lambda_{\text{A}}} < 0]$ is well defined.

Since $T(x) = \lceil |x^{h+}|/\mu \rceil$, we have

$$\frac{|x^{h+}|}{\mu} \leq T(x) \leq \frac{|x^{h+}|}{\mu} + 1.$$

We will use this simple inequality to obtain the upper and lower bounds.

We first obtain the upper bound. Clearly,

$$\mathbb{E}_\infty[T(W_{\lambda_{\text{A}}}) \Big| W_{\lambda_{\text{A}}} < 0] \leq \frac{\mathbb{E}_\infty[|W^{h+}_{\lambda_{\text{A}}}| \;\Big|\; W_{\lambda_{\text{A}}} < 0]}{\mu} + 1.$$

An upper bound for the right-hand side of the above equation is easily obtained. First note that from (3.12)

$$\mathbb{E}_\infty[|W^{h+}_{\lambda_\infty}|] \;\leq\; \mathbb{E}_\infty[|W_{\lambda_\infty}|] \;<\; \infty,$$

and

$$\begin{aligned}\mathbb{E}_\infty[|W^{h+}_{\lambda_\infty}|] &\geq \mathbb{E}_\infty[|W^{h+}_{\lambda_{\text{A}}}| \;\Big|\; W_{\lambda_{\text{A}}} < 0]\, \mathbb{P}_\infty(W_{\lambda_{\text{A}}} < 0). \\ &\geq \mathbb{E}_\infty[|W^{h+}_{\lambda_{\text{A}}}| \;\Big|\; W_{\lambda_{\text{A}}} < 0]\, \mathbb{P}_\infty(\ell(X_1) < 0).\end{aligned}$$

This completes the proof for the upper bound.

For the lower bound we have

$$\begin{aligned}
&\mathbb{E}_\infty[T(W_{\lambda_{\mathrm{A}}}) \,\Big|\, W_{\lambda_{\mathrm{A}}} < 0] \\
&\geq \frac{\mathbb{E}_\infty[|W^{h+}_{\lambda_{\mathrm{A}}}| \,\Big|\, W_{\lambda_{\mathrm{A}}} < 0]}{\mu} \\
&\geq \frac{\mathbb{E}_\infty[|W^{h+}_{\lambda_{\mathrm{A}}}| \,;\{\ell(X_1) < 0\} \,\Big|\, W_{\lambda_{\mathrm{A}}} < 0]}{\mu} \\
&= \frac{\mathbb{E}_\infty[|\ell(X_1)^{h+}| \,\Big|\, \ell(X_1) < 0]}{\mu}\, \mathbb{P}_\infty(\ell(X_1) < 0).
\end{aligned}$$

In the above equation we have used the fact that the unconditional probability $\mathbb{P}_\infty(\ell(X_1) < 0)$ is smaller than the conditional one $\mathbb{P}_\infty(\ell(X_1) < 0 | W_{\lambda_{\mathrm{A}}} < 0)$. □

Proof of Lemma 3.3.4. Let

$$\tau_{\mathrm{C}}(x) = \inf\{n \geq 1 : C_n > A; C_0 = x\}.$$

Here, C_n is the CuSum statistic and evolves according to the description of the algorithm in Algorithm 3.1.1. Thus, $\tau_{\mathrm{C}}(x)$ is the first time for the CuSum algorithm to cross A, when starting at $C_0 = x$. Clearly, $\tau_{\mathrm{C}}(x) = \tau_{\mathrm{C}}$ if $x = 0$. It is easy to see by sample-pathwise arguments that

$$\mathbb{E}_1[\tau_{\mathrm{C}}(x)] \leq \mathbb{E}_1[\tau_{\mathrm{C}}].$$

The proof depends on the above inequality.

Let $\mathcal{A}_x$ be the event that the CuSum statistic, starting with $C_0 = x$, touches zero before crossing the upper threshold A. Let $q_x = \mathbb{P}_1(\mathcal{A}_x)$. Then,

$$\mathbb{E}_1[\tau_{\mathrm{C}}(x)] = \mathbb{E}_1[\tau_{\mathrm{C}}(x); \mathcal{A}_x] + \mathbb{E}_1[\tau_{\mathrm{C}}(x); \mathcal{A}'_x] \leq \mathbb{E}_1[\tau_{\mathrm{C}}].$$

Note that

$$\mathbb{E}_1[\tau_{\mathrm{C}}(x); \mathcal{A}'_x] = \mathbb{E}_1[\tau_{\mathrm{W}}(x); \mathcal{A}'_x].$$

We call this common constant $\mathtt{t}_1$. Also note that on $\mathcal{A}_x$, the average time taken to reach 0 is the same for both the CuSum and the DE-CuSum algo-

rithm. We call this common average conditional delay by t_2. Thus,

$$\mathbb{E}_1[\tau_{\mathrm{C}}(x)] = (\mathsf{t}_1)(1-q_x) + q_x(\mathsf{t}_2 + \mathbb{E}_1[\tau_{\mathrm{C}}]) \leq \mathbb{E}_1[\tau_{\mathrm{C}}].$$

The equality in the above equation is true because, once the DE-CuSum statistic reaches zero, it is reset to zero and the average delay that point onwards is $\mathbb{E}_1[\tau_{\mathrm{C}}]$.

Then for any $\mathsf{t}_3 \geq \mathbb{E}_1[\tau_{\mathrm{C}}]$ we have

$$(\mathsf{t}_1)(1-q_x) + q_x(\mathsf{t}_2 + \mathsf{t}_3) \leq \mathsf{t}_3.$$

This is because for $\mathsf{t}_3 \geq \mathbb{E}_1[\tau_{\mathrm{C}}]$

$$\begin{aligned}
(\mathsf{t}_1)(1-q_x) \quad &+ \quad q_x(\mathsf{t}_2 + \mathsf{t}_3) \\
&= \quad (\mathsf{t}_1)(1-q_x) + q_x(\mathsf{t}_2 + \mathbb{E}_1[\tau_{\mathrm{C}}] + \mathsf{t}_3 - \mathbb{E}_1[\tau_{\mathrm{C}}]) \\
&\leq \quad \mathbb{E}_1[\tau_{\mathrm{C}}] + q_x(\mathsf{t}_3 - \mathbb{E}_1[\tau_{\mathrm{C}}]) \\
&\leq \quad \mathsf{t}_3.
\end{aligned}$$

It is easy to see that

$$\mathbb{E}_1[\tau_{\mathrm{W}}(x)] \leq (\mathsf{t}_1)(1-q_x) + q_x(\mathsf{t}_2 + \lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\mathrm{W}}]).$$

This is because on $\mathcal{A}_x$, the average delay of the DE-CuSum algorithm is the average time to reach 0, which is t_2, plus the average time spent below 0 due to the undershoot, which is bounded from above by $\lceil h/\mu \rceil$, plus the average delay after the sojourn below 0, which is $\mathbb{E}_1[\tau_{\mathrm{W}}]$. The latter is due to the renewal nature of the DE-CuSum algorithm. Since $\lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\mathrm{W}}] \geq \mathbb{E}_1[\tau_{\mathrm{C}}]$, the lemma is proved if we set $\mathsf{t}_3 = \lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\mathrm{W}}]$. □

CHAPTER 4

DATA-EFFICIENT MINIMAX QUICKEST CHANGE DETECTION WITH COMPOSITE POST-CHANGE HYPOTHESIS

In this chapter we extend the results of Chapter 3 to the case when the post-change distribution is unknown.

The classical problem of detecting a change when the post-change distribution is unknown (and with no observation control) has been well studied in the literature. In the parametric setting, where the post-change distribution is assumed to belong to a parametric family, there are three main approaches: generalized likelihood ratio (GLR) based, mixture based and adaptive estimates based approaches. In the nonparametric setting, one approach has been to take a robust approach to the QCD problem. See [1], [2] and [3] for a review.

In this chapter we combine the ideas from Chapter 3 and from the QCD literature for the case where the post-change distribution is unknown to study DE-QCD problems when the post-change distribution is unknown. We assume that the post-change family of distributions has a *least favorable* member (see Assumption 4.4.1 for a precise definition). Based on this assumption we propose an algorithm called the generalized data-efficient cumulative sum (GDECuSum) algorithm. In this algorithm on-off observation control is performed using the DE-CuSum algorithm designed for the least favorable distribution, and the change is detected using a GLRT based CuSum algorithm [11], [12], [27]. We show that if the post-change family of distributions is finite or if both the pre- and post-change distributions belong to a one-parameter exponential family, then the GDECuSum algorithm is asymptotically optimal for a modified version of the problem formulation from Chapter 3, uniformly over all possible post-change distributions.

The assumption that the post-change distribution belongs to a finite set of distributions is satisfied in many practical applications. For example, it is satisfied in the problem of detecting a power line outage in a power grid [4], or in a multi-channel scenario where the observations are vector valued and

a change affects the distribution of only a subset of the components (each component for example may correspond to the output of a distinct sensor on a sensor board) [28], [29]. Also, see [27] for a possible scenario.

4.1 Problem Formulation

A sequence of random variables $\{X_n\}$ is being observed. Initially, the random variables are i.i.d. with p.d.f. f_0. At the change point γ the density of the random variables changes to f_θ, $\theta \in \Theta$. That is, we assume that the post-change distribution belongs to a parametric family of distributions parameterized by θ. Both θ and γ are unknown. We assume that $f_0 \neq f_\theta$ for all $\theta \in \Theta$. We denote by $\mathbb{P}_\gamma^\theta$ the underlying probability measure which governs such a sequence. We use $\mathbb{E}_\gamma^\theta$ to denote the expectation with respect to this probability measure. We use $\mathbb{P}_\infty$ ($\mathbb{E}_\infty$) to denote the probability measure (expectation) when the change never occurs (i.e., the random variable X_n has p.d.f. f_0, $\forall n$). We wish to detect this change in distribution as quickly as possible subject to a constraint on the false alarm rate.

We will reuse the notation from the previous chapters. Specifically,

$$S_n = \begin{cases} 1 & \text{if } X_n \text{ used for decision making} \\ 0 & \text{otherwise.} \end{cases}$$

The information available at time n is denote by

$$\mathcal{I}_n = \{X_1^{(S_1)}, \cdots, X_n^{(S_n)}\},$$

where $X_k^{(S_k)} = X_k$ if $S_k = 1$, else X_k is absent from $\mathcal{I}_n$, and

$$S_n = \phi_n(\mathcal{I}_{n-1}).$$

Here, ϕ_n denotes the control map. Let τ be a stopping time for the sequence $\{\mathcal{I}_n\}$. A control policy is the collection

$$\Psi = \{\tau, \phi_1, \cdots, \phi_\tau\}.$$

As in Chapter 3, for delay we choose the following conditional average

detection delay metric (CADD) of Pollak [13]:

$$\mathsf{CADD}^{\theta}(\Psi) := \sup_{\gamma \geq 1} \mathbb{E}_{\gamma}^{\theta}[\tau - \gamma | \tau \geq \gamma]. \tag{4.1}$$

Note that the CADD is now a function of the post-change parameter θ.

For false alarm we choose the metric of false alarm rate (FAR) [12], [13]:

$$\mathsf{FAR}(\Psi) := \frac{1}{\mathbb{E}_{\infty}[\tau]}. \tag{4.2}$$

To capture the cost of observations used before γ, we use the following variation of the duty cycle metric CPDC proposed in Chapter 3, the Pre-change Duty Cycle (PDC) metric:[1]

$$\begin{aligned} \mathsf{PDC}(\Psi) &:= \limsup_{\gamma \to \infty} \mathbb{E}_{\gamma}^{\theta} \left[\frac{1}{\gamma} \sum_{n=1}^{\gamma-1} S_n \right] \\ &= \limsup_{\gamma \to \infty} \mathbb{E}_{\infty} \left[\frac{1}{\gamma} \sum_{n=1}^{\gamma-1} S_n \right]. \end{aligned} \tag{4.3}$$

Note that both the FAR and the PDC are *not* a function of the post-change parameter θ.

The first problem that we are interested in is the following:

Problem 4.1.1.

$$\begin{aligned} \min_{\Psi} \quad & \mathsf{CADD}^{\theta}(\Psi) \\ \textit{subj. to} \quad & \mathsf{FAR}(\Psi) \leq \alpha, \\ \textit{and} \quad & \mathsf{PDC}(\Psi) \leq \beta, \end{aligned}$$

where, $0 \leq \alpha, \beta \leq 1$ *are given constraints.*

Again, as in Chapter 3, we are also interested in the problem where the CADD in Problem 4.1.1 is replaced by the following worst case average detection delay (WADD) metric of Lorden [12],

$$\mathsf{WADD}^{\theta}(\Psi) := \sup_{\gamma \geq 1} \operatorname{ess\,sup} \mathbb{E}_{\gamma}^{\theta}[(\tau - \gamma)^{+} | \mathcal{I}_{\gamma-1}], \tag{4.4}$$

where, $x^{+} := \max\{0, x\}$:

[1] The definition of CPDC used in Chapter 3 had an extra conditioning on $\{\tau \geq \gamma\}$.

Problem 4.1.2.

$$\begin{aligned} \min_{\Psi} \quad & \mathsf{WADD}^{\theta}(\Psi) \\ \textit{subj. to} \quad & \mathsf{FAR}(\Psi) \leq \alpha, \\ \textit{and} \quad & \mathsf{PDC}(\Psi) \leq \beta, \end{aligned}$$

where, $0 \leq \alpha, \beta \leq 1$ *are given constraints.*

We recall that for any policy Ψ,

$$\mathsf{CADD}^{\theta}(\Psi) \leq \mathsf{WADD}^{\theta}(\Psi). \tag{4.5}$$

Our objective is to find an algorithm that is a solution to both Problem 4.1.1 and Problem 4.1.2 uniformly for each $\theta \in \Theta$. However, it is not clear if such a solution exists, even with $\beta = 1$. As a result we seek a solution that is asymptotically optimal, for a given β, for each θ, as $\alpha \to 0$.

4.2 Classical QCD with Unknown Post-Change Distribution

In this section we review the results from [27], [12] that are relevant to this chapter.

We first review the lower bound on the performance of any test for an FAR of α. This result was also discussed in (3.22). Let

$$\Delta_{\alpha} := \{\Psi : \mathsf{FAR}(\Psi) \leq \alpha\}.$$

When the post-change density is f_{θ}, a universal lower bound on the CADD over the class Δ_{α} is given by [14]

$$\inf_{\Psi \in \Delta_{\alpha}} \mathsf{CADD}^{\theta}(\Psi) \geq \frac{|\log \alpha|}{D(f_{\theta} \mid\mid f_0)}(1 + o(1)) \text{ as } \alpha \to 0. \tag{4.6}$$

By (4.5), this is a lower bound on WADD^{θ} as well.

4.2.1 QCD with No Observation Control ($\beta = 1$), θ Known

We first consider the case when the post-change distribution is known to be f_θ, i.e., when the post-change parameter θ is known, and when there is no observation control, i.e., when $\beta = 1$, in Problem 4.1.1 and Problem 4.1.2. Then the lower bound (4.6) is achieved by the cumulative sum (CuSum) algorithm [11], [12]; also see Algorithm (3.1.1). We note that the CuSum algorithm can also be defined as follows:

$$\begin{aligned} C_n(\theta) &= \max_{1\le k\le n+1} \sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \quad \text{for } n \ge 1, \\ \tau_{\mathrm{C}}(\theta) &= \inf\{n \ge 1 : C_n(\theta) \ge A\}. \end{aligned} \tag{4.7}$$

The statistic $C_n(\theta)$ can be computed recursively (see Algorithm 3.1.1):

$$\begin{aligned} C_0(\theta) &= 0, \\ C_n(\theta) &= \left(C_{n-1}(\theta) + \log \frac{f_\theta(X_n)}{f_0(X_n)} \right)^+ \text{ for } n \ge 1. \end{aligned} \tag{4.8}$$

We also note that the CuSum algorithm can also be written as

$$\begin{aligned} \hat{C}_n(\theta) &= \max_{1\le k\le n} \sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \quad \text{for } n \ge 1, \\ \tau_{\mathrm{C}}(\theta) &= \inf\{n \ge 1 : \hat{C}_n(\theta) \ge A\}. \end{aligned} \tag{4.9}$$

The difference between the statistics $\hat{C}_n$ and C_n is that the former can take negative values.

As we saw in Chapter 3 in Theorem 3.3.3, the CuSum algorithm is asymptotically optimal for both Problem 4.1.1 and Problem 4.1.2 (with θ known and $\beta = 1$) due to (4.5) and because of the following result: setting $A = \log 1/\alpha$ in (4.7) ensures that [12]

$$\begin{aligned} \mathsf{FAR}(\tau_{\mathrm{C}}(\theta)) &\le \alpha, \\ \mathsf{WADD}^{\theta(\theta)}(\tau_{\mathrm{C}}) &\le \frac{|\log \alpha|}{D(f_\theta \,||\, f_0)}(1 + o(1)) \text{ as } \alpha \to 0. \end{aligned} \tag{4.10}$$

We note that the PDC of the CuSum algorithm is equal to 1.

4.2.2 QCD with No Observation Control ($\beta = 1$), θ Unknown

We next consider the case when the post-change distribution is unknown, i.e., when the post-change parameter θ is unknown, and again there is no observation control, i.e., $\beta = 1$ in Problem 4.1.1 and Problem 4.1.2. A natural extension of the CuSum algorithm for this case is the generalized likelihood ratio based CuSum algorithm. We refer to the algorithm as the GCuSum algorithm and it is defined as follows:

$$\begin{aligned} G_n &= \max_{1\le k\le n} \sup_{\theta\in\Theta'(\alpha)} \sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \quad \text{for } n \ge 1, \\ \tau_{\text{GC}} &= \inf\{n \ge 1 : G_n(\theta) \ge A\}, \end{aligned} \tag{4.11}$$

where, $\Theta'(\alpha) \subset \Theta$ can be a function of α, and is either equal to Θ, or is allowed to be arbitrarily close and grow to Θ as $\alpha \to 0$. The GCuSum algorithm has the following interpretation. To detect a change when the post-change parameter is unknown, a family of CuSum algorithms are executed in parallel, one for each post-change parameter. A change is declared the first time a change is detected in any one of the CuSum algorithms. It can be shown that

$$\mathsf{WADD}^\theta(\tau_{\text{GC}}) = \mathsf{CADD}^\theta(\tau_{\text{GC}}) = \mathbb{E}_1^\theta[\tau_{\text{GC}} - 1]. \tag{4.12}$$

We also note that the PDC of the GCuSum algorithm is equal to 1.

The asymptotic optimality of the GCuSum algorithm is known for example in the following two cases: when the post-change family is finite [27], and when the pre- and post-change distributions belong to a one-parameter exponential family [12].

When the post-change set Θ is finite, i.e,

$$\Theta = \{\theta_1, \cdots, \theta_M\},$$

the GCuSum algorithm with $\Theta'(\alpha) = \Theta$ reduces to the following algorithm

$$\tau_{\text{GC}} = \inf\left\{n \ge 1 : \max_{1\le k\le M} C_n(\theta_k) \ge A\right\}, \tag{4.13}$$

where $C_n(\theta_k)$ is the CuSum statistic (4.8) evaluated for $\theta = \theta_k$. Equation

(4.13) can also be written as

$$\tau_{\text{GC}} = \min_{1 \le k \le M} \tau_{\text{C}}(\theta_k). \tag{4.14}$$

In the following we refer to the GCuSum algorithm with Θ finite as the MCuSum algorithm. Thus, the GCuSum algorithm (4.11) has a recursive implementation in this case.[2] The asymptotic optimality of the GCuSum algorithm, with Θ finite, is proved in [27]. Specifically, setting $A = \log M/\alpha$ in (4.13) ensures that

$$\begin{aligned} &\mathsf{FAR}(\tau_{\text{GC}}) \le \alpha, \\ &\mathsf{WADD}^{\theta_k}(\tau_{\text{GC}}) \le \frac{|\log \alpha|}{D(f_{\theta_k} \,||\, f_0)}(1 + o(1)) \text{ as } \alpha \to 0, \text{ for } 1 \le k \le M. \end{aligned} \tag{4.15}$$

Thus, due to (4.15) and (4.6), the GCuSum algorithm is asymptotically optimal for both Problem 4.1.1 and Problem 4.1.2, with $\beta = 1$, as $\alpha \to 0$, uniformly over θ_k, $1 \le k \le M$.

Now consider the case when the pre- and post-change distributions belong to an exponential family such that

$$f_\theta(x) = \exp(\theta x - b(\theta)) f_0(x), \theta \in \Theta, \tag{4.16}$$

where, Θ is an interval on the real line not containing 0, i.e., $\Theta = [\theta_\ell, \theta_u] \backslash \{0\}$, and $b(0) = 0$. As claimed in [12], this model can be used to represent a much broader class of one-parameter exponential family. For this case, the asymptotic optimality of the GCuSum algorithm is studied in [12]. Specifically, with $\epsilon > 0$, $\Theta'(\alpha) = \{\theta \in [\theta_\ell, \theta_u] : |\theta| > \epsilon\}$ and setting $A = A_\alpha \simeq \log 1/\alpha$ ensures

$$\begin{aligned} &\mathsf{FAR}(\tau_{\text{GC}}) \le \alpha(1 + o(1)), \text{ as } \alpha \to 0 \\ &\mathsf{WADD}^{\theta}(\tau_{\text{GC}}) \le \frac{|\log \alpha|}{D(f_\theta \,||\, f_0)}(1 + o(1)) \text{ as } \alpha \to 0, \text{ for all } \theta \in \Theta'(\alpha). \end{aligned} \tag{4.17}$$

Here, ϵ is allowed to decrease to zero as $\alpha \to 0$. As a result, each $\theta \in \Theta$ is covered eventually. Thus, to detect a change with θ very close to 0, we must operate at low false alarm rates.

[2] We note however that while C_n is a non-negative statistic, the statistic G_n can take negative values.

We remark on the differences between the results in (4.15) and (4.17). While (4.15) is valid only with Θ finite, the pre- and post-change distributions are allowed to be arbitrary, and the FAR result is non-asymptotic. On the other hand, in (4.17), the distributions are restricted to an exponential family, and the FAR result is asymptotic, but the parameter set Θ is allowed to be uncountably infinite.

4.3 QCD with Observation Control ($\beta < 1$), θ Known

For the case when θ is known and $\beta < 1$, in Chapter 3, we proposed the DE-CuSum algorithm which is a two-threshold modification of the CuSum algorithm (4.8) and showed that it is asymptotically optimal for a variation of both Problem 4.1.1 and Problem 4.1.2 (with CPDC as the duty cycle metric), for each β, as $\alpha \to 0$. Since the duty cycle metric PDC is different here, in this section we prove the asymptotic optimality of the DE-CuSum algorithm with this new definition of the duty cycle metric.

We first write the DE-CuSum algorithm in a different form (see Algorithm 3.2.1). This will be useful in the Section 4.4.

$$
\begin{aligned}
&\text{If } W_{n-1}(\theta) \geq 0, \\
&\qquad S_n = 1 \\
&\qquad W_n(\theta) = \max\left\{-h, \max_{1\leq k\leq n} \sum_{i=k}^{n} \log \frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})}\right\}. \\
&\text{If } W_{n-1}(\theta) < 0, \\
&\qquad S_n = 0, \\
&\qquad W_n(\theta) = \min\{0, W_{n-1}(\theta) + \mu\}. \\
&\text{Stop at} \\
&\qquad \tau_{\mathrm{w}}(\theta) = \inf\{n \geq 1 : W_n(\theta) \geq A\}.
\end{aligned}
\tag{4.18}
$$

where $\frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})} = 1$ if $S_i = 0$. Note that this way of writing the DE-CuSum algorithm is similar to the description of the CuSum algorithm in (4.9), whereas the description of the DE-CuSum algorithm provided in Algorithm 3.2.1 is closer to the recursive description of the CuSum algorithm in (4.8).

For the theorem below, we need the following definition. We define the ladder variable [16]

$$\tau_-(\theta) = \inf\left\{ n \geq 1 : \sum_{k=1}^{n} \log \frac{f_\theta(X_k)}{f_0(X_k)} < 0 \right\}.$$

Then note that $W_{\tau_-}(\theta)$ is the ladder height. Recall that $(x)^{h+} = \max\{x, -h\}$.

Theorem 4.3.1. *When the post-change density f_θ is fixed and known, and $\mu > 0$, $h < \infty$, and $A = |\log \alpha|$, we have*

$$\begin{aligned} \mathsf{FAR}(\tau_{\mathrm{W}}(\theta)) &\leq \mathsf{FAR}(\tau_{\mathrm{C}}(\theta)) \leq \alpha, \\ \mathsf{PDC}(\tau_{\mathrm{W}}(\theta)) &= \frac{\mathbb{E}_\infty[\tau_-(\theta)]}{\mathbb{E}_\infty[\tau_-(\theta)] + \mathbb{E}_\infty[\lceil |W_{\tau_-}(\theta)^{h+}|/\mu \rceil]}, \\ \mathsf{WADD}^\theta(\tau_{\mathrm{W}}(\theta)) &\sim \mathsf{WADD}^\theta(\tau_{\mathrm{C}}(\theta)) \sim \frac{|\log \alpha|}{D(f_\theta \mid\mid f_0)}(1 + o(1)) \text{ as } \alpha \to 0. \end{aligned} \tag{4.19}$$

If $h = \infty$, then

$$\mathsf{PDC}(\tau_{\mathrm{W}}(\theta)) \leq \frac{\mu}{\mu + D(f_0 \mid\mid f_\theta)}. \tag{4.20}$$

Proof. The proofs for the FAR and WADD analysis are identical to that provided in [30]. For the PDC we have the following proof. If S_n is treated as a reward for an on-off renewal process with the on time distributed according to the law of τ_-, and the off time distributed according to the law of $\lceil |W_{\tau_-}|/\mu \rceil$ (with truncation taken into account if $h < \infty$). Then, by the renewal reward theorem we have

$$\mathsf{PDC}(\tau_{\mathrm{W}}) = \frac{\mathbb{E}_\infty[\tau_-]}{\mathbb{E}_\infty[\tau_-] + \mathbb{E}_\infty[\lceil |W_{\tau_-}^{h+}/\mu| \rceil]}.$$

This proves (4.19).

If $h = \infty$, then (4.20) follows from the above equation because $x \leq \lceil x \rceil$, and from the Wald's lemma: $\mathbb{E}_\infty[|W_{\tau_-}|] = \mathbb{E}_\infty[\tau_-]\ D(f_0 \mid\mid f_\theta)$ [16]. □

We note that the expression for the PDC is not a function of the threshold A. Also, for any given $h > 0$, the smaller the value of the parameter μ, the smaller the PDC.

With $A = |\log \alpha|$ and μ and h set to achieve the PDC constraint of β (independent of the choice of A), the WADD of the DECuSum algorithm achieves the lower bound (4.6). Hence, we have from (4.5) that the algorithm

is asymptotically optimal for both Problem 4.1.1 and Problem 4.1.2, for the given β, as $\alpha \to 0$. Thus, the pre-change observation control can be executed, i.e., any arbitrary but fixed fraction of samples can be dropped before change, without any loss in the asymptotic performance.

4.4 The GDECuSum Algorithm

In this section we propose the main algorithm of this chapter, the GDE-CuSum algorithm. This algorithm can be used for the case when the post-change distribution is not known, and there is a need to perform on-off observation control, which is the object of study in this chapter. Mathematically, $\beta < 1$ in Problem 4.1.1 and Problem 4.1.2, and θ is unknown.

We now make the important assumption that there exists $\theta^* \in \Theta$ such that f_{θ^*} is the least favorable distribution among the family $\{f_\theta\}$, in a sense defined by the following assumption:

Assumption 4.4.1. *For each $\theta \in \Theta$,*

$$\mathbb{E}_1^\theta \left[\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)} \right] = D(f_\theta \mid\mid f_0) - D(f_\theta \mid\mid f_{\theta^*}) > 0.$$

The assumption is satisfied for example when the law of $\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)}$ under $\{f_\theta\}$ is stochastically bounded by its law under f_{θ^*} (see Definition 1 in [31]), i.e.,

$$\mathbb{P}_1^\theta \left(\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)} > x \right) \geq \mathbb{P}_1^{\theta^*} \left(\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)} > x \right), \quad \forall \theta \in \Theta.$$

The latter condition is satisfied for example in the following cases:

1. Θ is finite, $\Theta = \{\theta_1, \cdots, \theta_M\}$, $f_0 = \mathcal{N}(0,1)$, $f_{\theta_k} = \mathcal{N}(\theta_k, 1)$, with $0 < \theta_1 < \theta_2 < \cdots < \theta_M$, and $\theta^* = \theta_1$.

2. $\{f_\theta\}$ and f_0 belong to an exponential family such that $f_0 = \mathcal{N}(0,1)$, $f_\theta = \mathcal{N}(\theta, 1)$, with $\theta \in [0.2, 1]$, and $\theta^* = 0.2$.

We now propose the GDECuSum algorithm. In the GDECuSum algorithm also, just like the GCuSum algorithm (4.11), a family of algorithms are executed in parallel, one for each post-change parameter, with the difference that the CuSum algorithm corresponding to the parameter $\theta = \theta^*$ is replaced

by the DECuSum algorithm. Also, the CuSum algorithms corresponding to $\theta \neq \theta^*$ are updated only when samples are taken. Essentially, the post-change density that is least favorable is used for observation control—which by Assumption 4.4.1 is f_{θ^*}—while all the f_θs are used for change detection.

The GDECuSum algorithm is described as follows.

Algorithm 4.4.1. *Fix $\mu > 0$ and $h \geq 0$,*

Compute for each $n \geq 1$,

$$\bar{G}_n = \max_{1 \leq k \leq n} \sup_{\theta \in \Theta} \sum_{i=k}^{n} \log \frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})}.$$

If $W_{n-1}(\theta^*) \geq 0$,

$$S_n = 1$$

$$W_n(\theta^*) = \max \left\{ -h, \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log \frac{f_{\theta^*}(X_i^{(S_i)})}{f_0(X_i^{(S_i)})} \right\}. \tag{4.21}$$

If $W_{n-1}(\theta^*) < 0$,

$$S_n = 0$$

$$W_n(\theta^*) = \min\{0, W_{n-1}(\theta^*) + \mu\}.$$

Stop at

$$\tau_{\text{GD}} = \inf\{n \geq 1 : \bar{G}_n \geq A\}.$$

The evolution of the GDECuSum algorithm can be described as follows. In this algorithm two statistics $\bar{G}_n$ and $W_n(\theta^*)$ are computed in parallel. While the statistic $\bar{G}_n$ is used to detect the change, the statistic $W_n(\theta^*)$ is used for observation control. Specifically, the statistic $W_n(\theta^*)$ is updated using the DECuSum algorithm (4.18). The statistic $\bar{G}_n$ is updated using the GCuSum algorithm (4.11) with the difference that when $W_n(\theta^*) < 0$, the statistic $\bar{G}_n$ is not updated and set to its value on the previous time instant. This is because when $W_n(\theta^*) < 0$, it is incremented by μ at each time instant, and observations are skipped until $W_n(\theta^*)$ reaches 0 from below. In the absence of any new observation, the GCuSum statistic $\bar{G}_n$ cannot be updated. In this algorithm, by design, while $W_n(\theta^*) < 0$, the GCuSum statistic $\bar{G}_n$ is set to its value in the last time instant. This is ensured by the definition $\frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})} = 1$ if $S_i = 0$.

Assumption 4.4.1 is critical to the working of this algorithm. By this

assumption the mean of the log likelihood ratio between f_{θ^*} and f_0 is positive for every possible post-change distribution. This is because for $\theta \in \Theta$,

$$\mathbb{E}_1^{\theta}\left[\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)}\right] = D(f_\theta \mid\mid f_0) - D(f_\theta \mid\mid f_{\theta^*}).$$

This ensures that after the change occurs, and after a finite number of samples (irrespective of the threshold A), the DECuSum statistic $W_n(\theta^*)$ always remains positive and no more observations are skipped. This allows the statistic $\bar{G}_n$ to grow with the right "slope". If the Assumption 4.4.1 is violated, and the post-change parameter is $\theta \neq \theta^*$, then the statistic $W_n(\theta^*)$ will be below zero for a longer duration of time, and this time grows to infinity as the threshold $A \to \infty$. Thus, essentially, the growth of the GCuSum statistic will be intercepted by multiple sojourns of the statistic $W_n(\theta^*)$ below zero. As a result, the change will still be detected, but with a delay larger than the lower bound (4.6).

For $\Theta = \{\theta_1, \cdots, \theta_M\}$ with $\theta^* = \theta_1$, the GDECuSum algorithm has a recursive implementation [3]

$$\begin{aligned}
&\text{If } W_{n-1}(\theta_1) \geq 0,\\
&\qquad S_n = 1,\\
&\qquad W_n(\theta_1) = \left(W_{n-1}(\theta_1) + \log \frac{f_{\theta_1}(X_n)}{f_0(X_n)}\right)^{h+}.\\
&\text{If } W_{n-1}(\theta_1) < 0,\\
&\qquad S_n = 0,\\
&\qquad W_n(\theta_1) = \min\{0, W_{n-1}(\theta_1) + \mu\}.\\
&\text{For } k \geq 2,\\
&\qquad \bar{C}_0(\theta_k) = 0,\\
&\qquad \bar{C}_n(\theta_k) = \left(\bar{C}_{n-1}(\theta_k) + \log \frac{f_{\theta_k}(X_n^{(S_n)})}{f_0(X_n^{(S_n)})}\right)^{+}.\\
&\text{Stop at,}\\
&\tau_{\text{GD}} = \inf\{n \geq 1 : \max\{W_n(\theta_1), \max_{2 \leq k \leq M} \bar{C}_n(\theta_k)\} \geq A\}.
\end{aligned} \tag{4.22}$$

[3] Again note that the statistics $\{\bar{C}_k\}_{k=2}^M$ here are non-negative while $\bar{G}_n$ is allowed to take negative values.

Thus, for Θ finite, the GDECuSum algorithm is equivalent to executing M recursive algorithms in parallel. One is the DE-CuSum algorithm using the least favorable distribution, and the rest $M-1$ algorithms are the CuSum algorithms. Note that when the DE-CuSum statistic $W_n(\theta_1) < 0$, the CuSum statistics $\{\bar{C}_n(\theta_k)\}_{k=2}^{M}$ are set to their values in the last time instant. For the case of finite Θ, we refer to the GDECuSum algorithm by the MDECuSum algorithm. In Fig. 4.1 we plot the evolution of the GDECuSum algorithm (or the MDECuSum algorithm) for $f_0 = \mathcal{N}(0,1)$, $f_{\theta_1} = \mathcal{N}(0.4,1)$, $f_{\theta_2} = \mathcal{N}(0.6,1)$, $f_{\theta_3} = \mathcal{N}(0.8,1)$, $f_{\theta_4} = \mathcal{N}(1,1)$, $\mu = 0.18$, and $h = 10$. The post-change parameter is $\theta = \theta_2 = 0.6$.

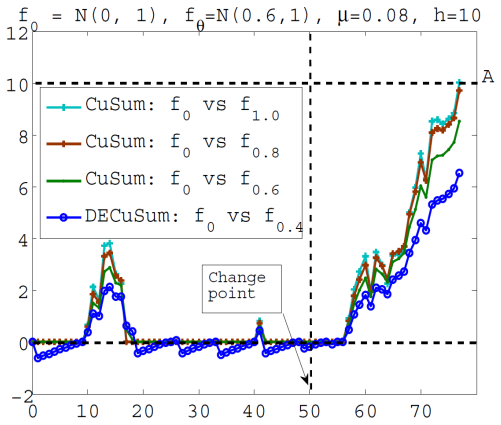


Figure 4.1: Evolution of the GDECuSum algorithm for $f_0 = \mathcal{N}(0,1)$, $f_{\theta_1} = \mathcal{N}(0.4,1)$, $f_{\theta_2} = \mathcal{N}(0.6,1)$, $f_{\theta_3} = \mathcal{N}(0.8,1)$, $f_{\theta_4} = \mathcal{N}(1,1)$, $\mu = 0.18$, and $h = 10$. The post-change parameter is $\theta = \theta_2 = 0.6$.

4.5 Asymptotic Optimality of the GDECuSum Algorithm

The evolution of the GDECuSum algorithm is statistically identical to that of the GCuSum algorithm, except of the possible sojourns of the statistic $W_n(\theta^*)$ below 0. Also, the sojourn time of $W_n(\theta^*)$ below zero is completely specified by the DECuSum algorithm. These two facts will now be used to express the performance of the GDECuSum algorithm in terms of the performance of the GCuSum algorithm and the DECuSum algorithm.

Let

$$\tau_{\mathrm{W}}(\theta^*) = \inf\{n \geq 1 : W_n(\theta^*) \geq A\},$$

the first time the statistic $W_n(\theta^*)$ crosses the threshold A.

Theorem 4.5.1. *Under the Assumption 4.4.1, for any fixed $\mu > 0$ and $h \geq 0$ and A we have*

$$\begin{aligned} \mathsf{FAR}(\tau_{\mathrm{GD}}) &\leq \mathsf{FAR}(\tau_{\mathrm{GC}}), \\ \mathsf{PDC}(\tau_{\mathrm{GD}}) &= \mathsf{PDC}(\tau_{\mathrm{W}}(\theta^*)), \end{aligned} \tag{4.23}$$

and for any $\mu > 0$ and $h < \infty$, and any $A \geq 0$,

$$\mathsf{WADD}^{\theta}(\tau_{\mathrm{GD}}) \leq \mathsf{WADD}^{\theta}(\tau_{\mathrm{GC}}) + K_{\mathrm{GD}}, \tag{4.24}$$

where K_{GD} is a constant that is a function of μ and h, but is not a function of A. As a result, for any $\mu > 0$ and $h < \infty$, we have

$$\begin{aligned} \mathsf{WADD}^{\theta}(\tau_{\mathrm{GD}}) \sim \mathsf{WADD}^{\theta}(\tau_{\mathrm{GC}}) \sim \frac{A}{D(f_\theta \,||\, f_0)}(1 + o(1)) \\ \text{as } A \to \infty, \text{ for each } \theta \in \Theta. \end{aligned} \tag{4.25}$$

We will provide the proof of the theorem at the end of this section. But, before that we will discuss its implications. From the theorem we see that, the GDECuSum algorithm can be designed to satisfy any arbitrary PDC constraint of β, independent of the choice of A. Also, the FAR of the GDECuSum algorithm is at least as good as that of the GCuSum algorithm. Finally, the WADD of the GDECuSum algorithm is within a constant of the WADD of the GCuSum algorithm. From (4.5) and (4.12) we have

$$\mathsf{CADD}^{\theta}(\tau_{\mathrm{GD}}) \leq \mathsf{WADD}^{\theta}(\tau_{\mathrm{GD}}) \leq \mathsf{WADD}^{\theta}(\tau_{\mathrm{GC}}) + K_{\mathrm{GD}} = \mathsf{CADD}^{\theta}(\tau_{\mathrm{GC}}) + K_{\mathrm{GD}}.$$

Thus, the GDECuSum algorithm will be asymptotically optimal if the GCuSum algorithm is asymptotically optimal. This is formally stated in the next corollary.

Corollary 4.5.1.1. *If the GCuSum algorithm is uniformly asymptotic optimal for a parametric family, then under the conditions of the theorem and if $h < \infty$, the GDECuSum algorithm is also uniformly asymptotically optimal,*

for each β, as $\alpha \to 0$.

Since the GCuSum algorithm is asymptotically optimal (with $\beta = 1$) for the two special classes of $\{f_\theta\}$: finite and exponential, the GDECuSum algorithm is also asymptotically optimal (for each fixed β) in these two cases. These are stated as corollaries below.

For a finite family we have the following result.

Corollary 4.5.1.2. *If Θ if finite, $\Theta = \{\theta_1, \cdots, \theta_M\}$, and Assumption 4.4.1 is satisfied for some $\theta^* \in \Theta$. Then, for any fixed $\mu > 0$ and $h \geq 0$ and $A = \log M/\alpha$ we have*

$$\begin{aligned} \mathsf{FAR}(\tau_{\mathrm{GD}}) &\leq \mathsf{FAR}(\tau_{\mathrm{GC}}) \leq \alpha, \\ \mathsf{PDC}(\tau_{\mathrm{GD}}) &= \mathsf{PDC}(\tau_{\mathrm{W}}(\theta^*)). \end{aligned} \tag{4.26}$$

Also, if $\mu > 0$ and $h < \infty$, then

$$\begin{aligned} \mathsf{WADD}^\theta(\tau_{\mathrm{GD}}) \sim \mathsf{WADD}^\theta(\tau_{\mathrm{GC}}) \sim \frac{|\log \alpha|}{D(f_{\theta_k} \mid\mid f_0)}(1 + o(1)) \\ \text{as } \alpha \to 0, \text{ for each } \theta_k, k = 1, \cdots, M. \end{aligned} \tag{4.27}$$

Proof. The result follows from (4.15) and Theorem 4.5.1. □

For an exponential family, we have the following result.

Corollary 4.5.1.3. *If $\{f_\theta\}$, f_0 belong to a one-parameter exponential family, i.e., if the following is satisfied,*

$$f_\theta(x) = \exp(\theta x - b(\theta)) f_0(x), \text{ for } \theta \in \Theta,$$

where, $\Theta = [\theta_\ell, \theta_u]$, with $0 < \theta_\ell < \theta_u$, and $b(0) = 0$. Also, Assumption 4.4.1 is satisfied for some $\theta^* \in \Theta$. Then, for any fixed $\mu > 0$, $h \geq 0$ and $A = A_\alpha \simeq \log 1/\alpha$ we have

$$\begin{aligned} \mathsf{FAR}(\tau_{\mathrm{GD}}) \leq \mathsf{FAR}(\tau_{\mathrm{GC}}) \leq \alpha(1 + o(1)), \text{ as } \alpha \to 0 \\ \mathsf{PDC}(\tau_{\mathrm{GD}}) = \mathsf{PDC}(\tau_{\mathrm{W}}(\theta^*)). \end{aligned} \tag{4.28}$$

And if $h < \infty$, then

$$\mathsf{WADD}^{\theta}(\tau_{\mathrm{GD}}) \sim \mathsf{WADD}^{\theta}(\tau_{\mathrm{GC}}) \sim \frac{|\log \alpha|}{D(f_\theta \mid\mid f_0)}(1 + o(1)) \quad \text{as } \alpha \to 0, \text{ for each } \theta \in [\theta_\ell, \theta_u]. \tag{4.29}$$

Proof. The result follows from (4.17) and Theorem 4.5.1. □

Since, the GDECuSum algorithm achieves the lower bound (4.6), the algorithm is asymptotically optimal for the two cases specified in the corollaries above, for a given β, uniformly over $\theta \in \Theta$, as $\alpha \to 0$.

Proof of Theorem 4.5.1. We recall that the GCuSum algorithm is the GLRT based test discussed in (4.11), and the GDECuSum algorithm is its data-efficient modification discussed in (4.21), where the observation control is executed based on the least favorable distribution f_{θ^*}.

We wish to prove (4.23) and (4.24), i.e., for any $\mu > 0$, $h \geq 0$, and A,

$$\mathsf{FAR}(\tau_{\mathrm{GD}}) \leq \mathsf{FAR}(\tau_{\mathrm{GC}}),$$
$$\mathsf{PDC}(\tau_{\mathrm{GD}}) = \mathsf{PDC}(\tau_{\mathrm{W}}(\theta^*)),$$

and for any $\mu > 0$ and $h < \infty$, and any A,

$$\mathsf{WADD}^{\theta}(\tau_{\mathrm{GD}}) \leq \mathsf{WADD}^{\theta}(\tau_{\mathrm{GC}}) + K_{\mathrm{GD}},$$

where K_{GD} is a constant that is a function of μ and h, but is not a function of A.

The PDC result follows from the PDC result proved in Theorem 4.3.1 because the observation control is governed by the statistic $W_n(\theta^*)$. We now prove the FAR and the WADD results. Both the results are based on the idea that the evolution of the GDECuSum algorithm is statistically identical to that of the GCuSum algorithm τ_{GC}, except of the possible sojourns of the statistic $W_n(\theta^*)$ below 0. Under $\mathbb{P}_\infty$, the sojourns of the statistic $W_n(\theta^*)$ below 0 only lead to larger mean time to false alarm for the GDECuSum algorithm. On the other hand, under each $\mathbb{P}_1^\theta$, the average number of times the statistic $W_n(\theta^*)$ goes below 0 is bounded by a constant, not a function of A. This is due to the fact that f_{θ^*} is the least favorable distribution, and as a result the drift of $W_n(\theta^*)$ is positive. Since $h < \infty$, the mean time

spent by the statistic $W_n(\theta^*)$ each time it goes below 0, it bounded by $\lceil h/\mu \rceil$. Thus, the total average mean time spent by the statistic $W_n(\theta^*)$ below 0 is bounded above by a constant. This in turn guarantees that the delay of the GDECuSum algorithm is within a constant of the GCuSum algorithm. The rest of the proof below formalizes these arguments.

We start by writing the stopping time τ_{GC} as a sum of a random number of stopping times. Such a representation is critical to this proof. Toward this end we define a set of new stopping variables. Similar variables were defined in the proof of Theorem 2.5.1 as well. Let $w \in [0, A)$, and define

$$\tau_1(w) = \inf \left\{ n \geq 1 : G_n > A \text{ or } \ w + \sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} < 0 \right\}.$$

This is the first time for either the GCuSum statistic G_n to hit A or the random walk $\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ to go below $-w$.

On paths over which $G_{\tau_1(w)} < A$, let

$$\tau_2(w) = \inf \left\{ n > \tau_1(w) : G_n > A \text{ or } \sum_{k=\tau_1(w)+1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} < 0 \right\}.$$

Thus, on paths such that $G_{\tau_1(w)} < A$, after the time $\tau_1(w)$, the time $\tau_2(w)$ is the first time for G_n to either cross A or the random walk $\sum_{k=\tau_1(w)+1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ to go below 0. We define, $\tau_3(w)$, etc. similarly. Next let,

$$N(w) = \inf\{k \geq 1 : G_{\tau_k} > A\}.$$

For simplicity we introduce the notion of "cycles", "success" and "failure". With reference to the definitions of $\tau_k(w)$'s above, we say that a success has occurred if the statistic G_n crosses A before the random walk $\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ goes below $-w$. In that case we also say that the number of cycles to A is 1. If on the other hand, the random walk $\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ goes below $-w$ before G_n crosses A, we say a failure has occurred. The number of cycles is 2, if now the statistic G_n crosses A before the random walk $\sum_{k=\tau_1(w)+1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ goes below 0. Thus, $N(w)$ is the number of cycles to success.

We note that for any given θ,

$$N(w) \leq \tau_{\text{GC}} \leq \tau_{\text{C}}(\theta).$$

This is because each cycle has length at least 1, and $\tau_{\text{C}}(\theta)$ is nothing but the τ_{GC} without the sup over Θ. Since, $\tau_{\text{C}}(\theta)$ is finite a.s. under both $\mathbb{P}_\infty$ and $\mathbb{P}_1^\theta$, for each $\theta \in \Theta$ (see Lorden [12]), even $N(w) < \infty$ a.s. under both $\mathbb{P}_\infty$ and $\mathbb{P}_1^\theta$, for any $\theta \in \Theta$.

Define $\lambda_1(w) = \tau_1(w)$, $\lambda_2(w) = \tau_2(w) - \tau_1(w)$, etc. Then we in fact have

$$\tau_{\text{GC}} = \sum_{k=1}^{N(w)} \lambda_k(w). \tag{4.30}$$

An important point to observe here is that while the terms on the right-hand side depend on w, their sum does not and equals τ_{GC}.

We now bound the mean of $N(w)$ under $\mathbb{P}_1^\theta$ by a number that is not a function of w and threshold A. With the identity

$$\mathbb{E}_1^\theta[N(w)] = \sum_{k=1}^{\infty} \mathbb{P}_1^\theta(N(w) \geq k)$$

in mind, and using the terminology of cycles, success and failure just defined, we write

$$\begin{aligned}\mathbb{P}_1^\theta(N(w) \geq k) = \mathbb{P}_1^\theta(&\text{fail in 1st cycle}) \\ &\cdots \mathbb{P}_1^\theta(\text{fail in } k-1^{st} \text{ cycle}|\text{fail in all previous}).\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{P}_1^\theta(&\text{fail in } i^{th} \text{ cycle}|\text{fail in all previous}) \\ &= 1 - \mathbb{P}_1^\theta(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}).\end{aligned}$$

We claim that

$$\mathbb{P}_1^\theta(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}) \geq \mathbb{P}_1^\theta\left(\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \geq 0,\ \forall n\right). \tag{4.31}$$

From [16] it is well known that $\mathbb{P}_1^\theta(\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \geq 0,\ \forall n) > 0$. This is because under θ, by the Assumption 4.4.1, the drift of the random walk

$\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}$ is positive. Thus, if

$$q_\theta = \mathbb{P}_1^\theta \left(\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \geq 0, \ \forall n \right),$$

then,

$$\mathbb{P}_1^\theta(N(w) \geq k) \leq (1 - q_\theta)^{k-1}.$$

Note that the right-hand side is not a function of the initial point w, nor is a function of the threshold A. Hence,

$$\mathbb{E}_1^\theta[N(w)] = \sum_{k=1}^{\infty} \mathbb{P}_1^\theta(N(w) \geq k) \leq \sum_{k=1}^{\infty} (1 - q_\theta)^{k-1} = \frac{1}{q_\theta} < \infty. \tag{4.32}$$

To prove the above claim (4.31) we note that

$$\begin{aligned}
&\mathbb{P}_1^\theta(\text{success in } 1^{st} \text{ cycle}) = \mathbb{P}_1^\theta(G_{\tau_1(w)} > A) \\
&\quad = \mathbb{P}_1^\theta(\text{Statistic } G_n \text{ reaches } A \text{ before } \textstyle\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } -w) \\
&\quad = \mathbb{P}_1^\theta \Big(\max_{1 \leq k \leq n} \ \sup_{\theta \in \Theta} \ \textstyle\sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \text{ reaches } A \\
&\qquad\qquad \text{before } \textstyle\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } -w \Big) \\
&\quad \geq \mathbb{P}_1^\theta \Big(\max_{1 \leq k \leq n} \ \textstyle\sum_{i=k}^{n} \log \frac{f_{\theta^*}(X_i)}{f_0(X_i)} \text{ reaches } A \\
&\qquad\qquad \text{before } \textstyle\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } -w \Big) \\
&\quad \geq \mathbb{P}_1^\theta \Big(\textstyle\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ reaches } A \\
&\qquad\qquad \text{before } \textstyle\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } -w \Big) \\
&\quad \geq \mathbb{P}_1^\theta \left(\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \geq 0, \ \forall n \right) = q_\theta.
\end{aligned} \tag{4.33}$$

Here, the first inequality follows because $\theta^* \in \Theta$ over which the supremum is being taken. The last inequality follows because $\sum_{k=1}^{n} \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \to \infty$ a.s. under $\mathbb{P}_1^\theta$ since θ^* is least favorable.

For the second cycle note that

$$\begin{aligned}
&\mathbb{P}_1^\theta(\text{success in } 2^{nd} \text{ cycle}|\text{failure in first}) = \mathbb{P}_1^\theta\left(G_{\tau_2(w)} > A | G_{\tau_1(w)} < A)\right) \\
&\quad = \mathbb{P}_1^\theta\left(\text{Statistic } G_n,\, n > \tau_1(w), \text{ reaches } A \right. \\
&\qquad\qquad \left. \text{before } \textstyle\sum_{k=\tau_1(w)+1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } 0 \;\middle|\; G_{\tau_1(w)} < A\right) \\
&\quad = \mathbb{P}_1^\theta\left(\max_{1\le k\le n} \sup_{\theta\in\Theta} \textstyle\sum_{i=k}^n \log \frac{f_\theta(X_i)}{f_0(X_i)},\, n > \tau_1(w), \text{ reaches } A \right. \\
&\qquad\qquad \left. \text{before } \textstyle\sum_{k=\tau_1(w)+1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } 0 \;\middle|\; G_{\tau_1(w)} < A\right) \\
&\quad \ge \mathbb{P}_1^\theta\left(\max_{\tau_1(w)<k\le n} \textstyle\sum_{i=k}^n \log \frac{f_{\theta^*}(X_i)}{f_0(X_i)}, \text{ for } n > \tau_1(w), \text{ reaches } A \right. \\
&\qquad\qquad \left. \text{before } \textstyle\sum_{k=\tau_1(w)+1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } 0 | G_{\tau_1(w)} < A\right) \\
&\quad \ge \mathbb{P}_1^\theta\left(\textstyle\sum_{k=\tau_1(w)+1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)}, \text{ for } n > \tau_1(w), \text{ reaches } A \right. \\
&\qquad\qquad \left. \text{before } \textstyle\sum_{k=\tau_1(w)+1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } 0 | G_{\tau_1(w)} < A\right) \\
&\quad = \mathbb{P}_1^\theta\left(\textstyle\sum_{k=1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ reaches } A \right. \\
&\qquad\qquad \left. \text{before } \textstyle\sum_{k=1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \text{ goes below } 0\right) \\
&\quad \ge \mathbb{P}_1^\theta\left(\sum_{k=1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \ge 0,\ \forall n\right) = q_\theta.
\end{aligned}$$

Almost identical arguments for the other cycles proves the claim that

$$\mathbb{P}_1^\theta(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}) \ge \mathbb{P}_1^\theta\left(\sum_{k=1}^n \log \frac{f_{\theta^*}(X_k)}{f_0(X_k)} \ge 0,\ \forall n\right),$$

and hence it follows that

$$\mathbb{E}_1^\theta[N(w)] \le \frac{1}{q_\theta} < \infty.$$

Let

$$\tau_{\text{GD}}(w) = \inf\{n \ge 1 : \bar{G}_n > A, \text{ with } W_0(\theta^*) = w\}.$$

Clearly, $\tau_{\text{GD}} = \tau_{\text{GD}}(0)$.

Just like we did for τ_{GC}, we now write the time $\tau_{\text{GD}}(w)$ as a sum of stopping times. We will then draw parallels between representation of this type for τ_{GC} and $\tau_{\text{GD}}(w)$ to prove the theorem.

Note that the sojourn of the statistic $\bar{G}_n$ to A may include alternate so-

journs of the statistic $W_n(\theta^*)$ above and below 0. Motivated by this we define a set of new variables. Let $w \in [0, A)$, and define

$$\bar{\tau}_1(w) = \inf\left\{n \geq 1 : \bar{G}_n > A \text{ or } \ W_n(\theta^*) < 0; \text{ starting with } W_0(\theta^*) = w\right\}.$$

This is the first time for either the GDECuSum statistic $\bar{G}_n$ to hit A or the DE-CuSum statistic $W_n(\theta^*)$ to go below 0, starting with $W_0(\theta^*) = w$. On paths over which $\bar{G}_{\bar{\tau}_1(w)} < A$, let $t_1(w)$ be the number of consecutive samples skipped after $\bar{\tau}_1(w)$ using the DE-CuSum statistic. On such paths again, let

$$\bar{\tau}_2(w) = \inf\left\{n > \bar{\tau}_1(w) + t_1(w) : \bar{G}_n > A \text{ or } W_n(\theta^*) < 0\right\}.$$

Thus, on paths such that $\bar{G}_{\bar{\tau}_1(w)} < A$, after the time $\bar{\tau}_1(w) + t_1(w)$, the time $\bar{\tau}_2(w)$ is the first time for $\bar{G}_n$ to either cross A or the DE-CuSum statistic $W_n(\theta^*)$ to go below 0. We define, $t_2(w)$, $\bar{\tau}_3(w)$, etc. similarly. Next let

$$\bar{N}(w) = \inf\{n \geq 1 : \bar{G}_{\bar{\tau}_n} > A\}.$$

We also define $\bar{\lambda}_1(w) = \bar{\tau}_1(w)$, $\bar{\lambda}_2(w) = \bar{\tau}_2(w) - \bar{\tau}_1(w) - t_1(w)$, etc. We now make an important observation. We observe that

$$\begin{aligned} \bar{N}(w) &\stackrel{d}{=} N(w) \\ \bar{\lambda}_k(w) &\stackrel{d}{=} \lambda_k(w), \ \forall k. \end{aligned} \tag{4.34}$$

Then we have

$$\bar{N}(w) < \infty \text{ a.s. under both } \mathbb{P}_\infty \text{ and } \mathbb{P}_1^\theta \text{ for each } \theta \in \Theta,$$

and

$$\tau_{\text{GD}}(w) = \sum_{k=1}^{\bar{N}(w)} \bar{\lambda}_k(w) + \sum_{k=1}^{\bar{N}(w)-1} t_k(w).$$

We are now ready to prove the FAR result. Using (4.34) and (4.30), and

the observation following (4.30), we have

$$\begin{aligned}
\mathbb{E}_\infty[\tau_{\text{GD}}] = \mathbb{E}_\infty[\tau_{\text{GD}}(0)] &= \mathbb{E}_\infty\left[\sum_{n=1}^{\bar{N}(0)} \bar{\lambda}_k(0)\right] + \mathbb{E}_\infty\left[\sum_{n=1}^{\bar{N}(0)-1} t_k(0)\right] \\
&= \mathbb{E}_\infty\left[\sum_{n=1}^{N(0)} \lambda_k(0)\right] + \mathbb{E}_\infty\left[\sum_{n=1}^{N(0)-1} t_k(0)\right] \\
&= \mathbb{E}_\infty\left[\tau_{\text{GC}}\right] + \mathbb{E}_\infty\left[\sum_{n=1}^{N(0)-1} t_k(0)\right] \\
&\geq \mathbb{E}_\infty\left[\tau_{\text{GC}}\right].
\end{aligned} \tag{4.35}$$

For the WADD we have for each $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}_1^\theta[\tau_{\text{GD}}(w)] &= \mathbb{E}_1^\theta\left[\sum_{n=1}^{\bar{N}(w)} \bar{\lambda}_k(w)\right] + \mathbb{E}_1^\theta\left[\sum_{n=1}^{\bar{N}(w)-1} t_k(w)\right] \\
&= \mathbb{E}_1^\theta\left[\sum_{n=1}^{N(w)} \lambda_k(w)\right] + \mathbb{E}_1^\theta\left[\sum_{n=1}^{\bar{N}(w)-1} t_k(w)\right] \\
&= \mathbb{E}_1^\theta\left[\tau_{\text{GC}}\right] + \mathbb{E}_1^\theta\left[\sum_{n=1}^{\bar{N}(w)-1} t_k(w)\right] \\
&\leq \mathbb{E}_1^\theta\left[\tau_{\text{GC}}\right] + \mathbb{E}_1^\theta\left[\bar{N}(w) - 1\right]\lceil h/\mu \rceil \\
&= \mathbb{E}_1^\theta\left[\tau_{\text{GC}}\right] + \mathbb{E}_1^\theta\left[N(w) - 1\right]\lceil h/\mu \rceil \\
&\leq \mathbb{E}_1^\theta[\tau_{\text{GC}}] + \frac{1}{q_\theta}\lceil h/\mu \rceil.
\end{aligned} \tag{4.36}$$

In (4.36) we have used the fact that

$$t_k(w) \leq \lceil h/\mu \rceil,\ \forall w \in [0, A), \forall k,$$

and the upper bound obtained on $\mathbb{E}_1^\theta[N(w)]$. Also note that the right-hand side is not a function of w, but does depend on the assumption that $w \in [0, A)$. We now obtain an upper bound on $\mathbb{E}_\gamma^\theta[(\tau_{\text{GD}} - \gamma)^+ | \mathcal{I}_{\gamma-1}]$.

If $\mathcal{I}_{\gamma-1} = i_{\gamma-1}$ is such that $W_{\gamma-1} = w \in [0, A)$, then

$$\mathbb{E}_\gamma^\theta[(\tau_{\text{GD}} - \gamma)^+ | \mathcal{I}_{\gamma-1} = i_{\gamma-1}] \leq \mathbb{E}_1[\tau_{\text{GD}}(w)].$$

This is because for $n \geq \gamma$

$$\max_{1\leq k\leq n} \sup_{\theta\in\Theta} \sum_{i=k}^{n} \log \frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})} \geq \max_{\gamma\leq k\leq n} \sup_{\theta\in\Theta} \sum_{i=k}^{n} \log \frac{f_\theta(X_i^{(S_i)})}{f_0(X_i^{(S_i)})}. \tag{4.37}$$

Thus, if $\mathcal{I}_{\gamma-1} = i_{\gamma-1}$ is such that $W_{\gamma-1} = w \in [0, A)$, then using (4.36) we have

$$\mathbb{E}_\gamma^\theta[(\tau_{\text{GD}} - \gamma)^+ | \mathcal{I}_{\gamma-1} = i_{\gamma-1}] \leq \mathbb{E}_1[\tau_{\text{GD}}(w)] \leq \mathbb{E}_1^\theta[\tau_{\text{GC}}] + \frac{1}{q_\theta}\lceil h/\mu \rceil. \tag{4.38}$$

On the other hand, if $\mathcal{I}_{\gamma-1} = i_{\gamma-1}$ is such that $W_{\gamma-1} = w < 0$, then the time to cross A for the GDECuSum statistic will be equal to the time taken for the statistic to cross 0 from below, plus a time bounded by $\mathbb{E}_1[\tau_{\text{GD}}(0)]$, where again we have used (4.37). Thus, we can write,

$$\begin{aligned} \mathbb{E}_\gamma^\theta[(\tau_{\text{GD}} - \gamma)^+ | \mathcal{I}_{\gamma-1} = i_{\gamma-1}] &\leq \lceil h/\mu \rceil + \mathbb{E}_1[\tau_{\text{GD}}(0)] \\ &\leq \mathbb{E}_1^\theta[\tau_{\text{GC}}] + (\frac{1}{q_\theta} + 1)\lceil h/\mu \rceil. \end{aligned} \tag{4.39}$$

Thus, we can write

$$\begin{aligned} \mathbb{E}_\gamma^\theta[(\tau_{\text{GD}} - \gamma)^+ | \mathcal{I}_{\gamma-1}] &\leq \mathbb{E}_1^\theta[\tau_{\text{GC}}] + (\frac{1}{q_\theta} + 1)\lceil h/\mu \rceil \\ &= \mathsf{WADD}^\theta(\tau_{\text{GC}}) + (\frac{1}{q_\theta} + 1)\lceil h/\mu \rceil + 1. \end{aligned} \tag{4.40}$$

Note that the right-hand side is no more a function of the conditioning $\mathcal{I}_{\gamma-1}$. The proof is complete if we define

$$K_{\text{GD}} = (\frac{1}{q_\theta} + 1)\lceil h/\mu \rceil + 1,$$

and take the essential supremum on the left-hand side. □

4.6 Numerical Results

In Fig. 4.2 we plot the CADD–FAR trade-off curves obtained using simulations for the GDECuSum algorithm (4.21), the GCuSum algorithm (4.11), and the fractional sampling scheme. In the latter, the GCuSum algorithm is

used and observations are skipped randomly, independent of the observation process. The simulation set used is: $M = 4$, $f_0 = \mathcal{N}(0,1)$, $f_{\theta_1} = \mathcal{N}(0.4,1)$, $f_{\theta_2} = \mathcal{N}(0.6,1)$, $f_{\theta_3} = \mathcal{N}(0.8,1)$, $f_{\theta_4} = \mathcal{N}(1,1)$, $\mu = 0.08$ and $h = \infty$. The post-change parameter is $\theta = \theta_2 = 0.6$, and the value of μ is chosen using (4.20) and (4.23) to achieve a $\mathsf{PDC} = 0.5$ (skipping/saving 50% of the samples). To achieve a PDC of 0.5 through the fractional sampling scheme, every alternate sample is skipped in the GCuSum algorithm. In the figure we see that skipping samples randomly results in a twofold increase in delay as compared to that of the GCuSum algorithm. However, if we use the GDECuSum algorithm and use the state of the system to skip observations, then there is a small and constant penalty on the delay, as compared to the performance of the GCuSum algorithm. Thus, the GDECuSum algorithm provides a significant gain in performance as compared to the fractional sampling scheme.

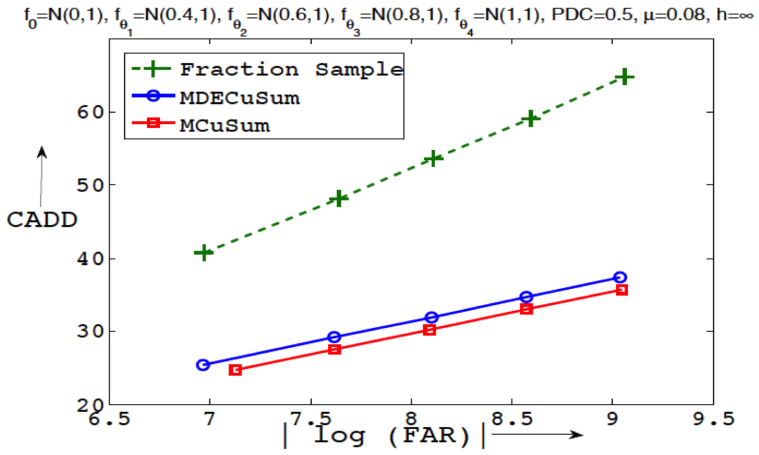


Figure 4.2: Comparative performance of the GDECuSum algorithm, the GCuSum algorithm, and the fractional sampling scheme. The post-change parameter is $\theta = \theta_2 = 0.6$.

4.7 Discussion on the Least Favorable Distribution

In this chapter we modified the GLRT based GCuSum algorithm by introduced observation control based on the DE-CuSum algorithm. We showed that the new data-efficient algorithm, the GDECuSum algorithm, is asymptotically optimal for the proposed data-efficient quickest change detection formulations, Problem 4.1.1 and Problem 4.1.2. See Theorem 4.5.1.

In the proof of Theorem 4.5.1 we used the main assumption of this chapter, that there is a distribution f_{θ^*} that is least favorable in the sense of Assumption 4.4.1. That is, for each $\theta \in \Theta$,

$$\mathbb{E}_1^\theta \left[\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)} \right] > 0.$$

The positive mean of the log likelihood ratio $\log \frac{f_{\theta^*}(X_1)}{f_0(X_1)}$ under each θ ensures that after a finite number of time slots, no observations are skipped using the DE-CuSum algorithm, and the change is detected efficiently.

However, for a given parametric family, there may not be a distribution that satisfies Assumption 4.4.1. In such a case, the results of this chapter are also valid for any distribution g that satisfies this assumption, [4] i.e.,

$$\mathbb{E}_1^\theta \left[\log \frac{g(X_1)}{f_0(X_1)} \right] > 0.$$

Thus, as long as such a distribution exists, we can design the DE-CuSum algorithm using the distribution g and the positive drift in the last equation will ensure that the GDECuSum with this new modification is still asymptotically optimal. We however note that in the proof of Theorem 4.5.1 we used the fact that $\theta^* \in \Theta$. Since g may not be in the parametric family, the proof needs to be modified. This can be accomplished by replacing the arguments in (4.33) with

$$\begin{aligned}
&\mathbb{P}_1^\theta(\text{success in } 1^{st} \text{ cycle}) = \mathbb{P}_1^\theta(G_{\tau_1(w)} > A) \\
&\quad = \mathbb{P}_1^\theta(\text{Statistic } G_n \text{ reaches } A \text{ before } \textstyle\sum_{k=1}^n \log \frac{g(X_k)}{f_0(X_k)} \text{ goes below } -w) \\
&\quad = \mathbb{P}_1^\theta \Big(\max\nolimits_{1 \le k \le n} \ \sup\nolimits_{\theta \in \Theta} \ \textstyle\sum_{i=k}^n \log \frac{f_\theta(X_i)}{f_0(X_i)} \text{ reaches } A \\
&\qquad\qquad \text{before } \textstyle\sum_{k=1}^n \log \frac{g(X_k)}{f_0(X_k)} \text{ goes below } -w \Big) \\
&\quad \ge \mathbb{P}_1^\theta \left(\sum_{k=1}^n \log \frac{g(X_k)}{f_0(X_k)} \ge 0, \ \forall n \right).
\end{aligned} \tag{4.41}$$

The last quantity is positive because $\mathbb{E}_1^\theta \left[\log \frac{g(X_1)}{f_0(X_1)} \right] > 0$ [16].

We note that in (4.41), we can also replace the GLRT statistic G_n by

[4] We thank Dr. Sirin Nitinawarat for bringing this point to our attention.

a mixture based statistic. And if the mixture distribution is chosen in an optimal way, then data-efficient extension of mixture based tests can also be shown to have asymptotic optimality properties.

Finally, recall that unless the post-change distribution belongs to a finite family, the GDECuSum algorithm does not have a recursive implementation. In the classical QCD literature, this problem is addressed by proposing window based tests; see Lai [14]. One can also study data-efficient extensions of such window based GLRT and mixture based tests.

CHAPTER 5

DATA-EFFICIENT QUICKEST CHANGE DETECTION IN SENSOR NETWORKS

In Chapters 2-4 we introduced the concept of data-efficiency via observation control in the classical problem of quickest change detection. We studied the problem in Bayesian (Chapter 2) and minimax settings (Chapters 3 and 4). In the latter case, we even allowed the post-change distribution to be unknown (Chapter 4). However, in all the results in the previous chapters, we assumed that we have a single observation sequence and a single decision maker. In many engineering applications of change detection, e.g., surveillance/monitoring of infrastructure (bridges, historic buildings, etc) or animal/bird habitat using sensor networks, the decision making is often implemented in a distributed fashion. Motivated by such applications, in this chapter we study data-efficient quickest change detection in sensor networks.

In a typical application of change detection involving a sensor network, multiple geographically distributed sensors are deployed to observe a phenomenon. At the sensors observations/measurements are taken periodically. At some point of time some property of the observations at the sensors changes. The objective is to detect this change as quickly as possible.

In this chapter, we study the above detection problem in the framework of decentralized quickest change detection introduced in [32] and further studied in [33] and [34]. In the model studied in [32], [33] and [34], the observations at the sensors are modeled as random variables. At each time step a processed version of the observations is transmitted from the sensors to a common decision node, called the fusion center. At some point of time, called the change point, the distribution of the random variables observed at *all* the sensors changes. The objective is to find a stopping time on the information received at the fusion center, so as to detect the change in distribution as quickly as possible (with minimum possible delay), subject to a constraint on the false alarm rate. The observations are independent across the sensors, and independent before and after the change point, conditioned on the change

point. The pre- and post-change distributions are assumed to be known.

In many applications of quickest change detection, including those mentioned above, changes are rare and acquiring data or taking observations is costly, e.g., the cost of batteries in sensor networks or the cost of communication between the sensors and the fusion center. In [32], [33] and [34], the cost of communication is controlled by quantizing or censoring observations/statistic at the sensors. However, the cost of taking observations at the sensors is not taken into account. Motivated by this, we study quickest change detection in sensor networks with an additional constraint on the cost of observations used at each sensor.

One way to detect a change in the sensor network model discussed above is to use the *Centralized CuSum* algorithm. In this algorithm, all the observations are taken at each sensor, raw observations are transmitted from each sensor to the fusion center. At the fusion center a CuSum is applied to all the received observations. The Centralized CuSum algorithm is clearly globally asymptotically optimal since the problem is simply of detecting a change in a vector sequence of observations (and hence reduces to the classical QCD problem). The problem is more interesting when sending raw observations from the sensors to the fusion center is not permitted, and at each sensor quantization of observations is enforced. A major result in this case is due to Mei in [33]. In [33], the following ALL scheme is proposed. In this scheme, a CuSum is applied locally at each sensor. A "1" is transmitted from a sensor to a fusion center each time the local CuSum statistic is above a threshold. A change is declared at the fusion center when a "1" is received from *all* the sensors at the same time. It is shown in [33] that the delay of the ALL scheme is asymptotically of the same order good as the delay of the Centralized CuSum scheme (for the same false alarm rate constraint, the ratio of their delay goes to 1), as the false alarm rate goes to zero.

In this chapter we introduce observation control in the ALL scheme of Mei by replacing the CuSum algorithm at each sensor by the DE-CuSum algorithm. We call the scheme the DE-All algorithm. We propose extensions of data-efficient formulations, Problem 4.1.2 and Problem 4.1.1, to sensor networks, and show that the DE-All scheme is globally asymptotically optimal for these formulations. By global asymptotic optimality we mean that the ratio of the delay the DE-All scheme and the Centralized CuSum scheme goes to 1 as the false alarm rate goes to zero. Thus, one can skip an arbitrary

but fixed fraction of samples before change, and transmit very rarely to the fusion center (just send an occasional "1"), thus conserving significantly the cost of battery, and yet perform as well (asymptotically up to first order) as the Centralized CuSum algorithm.

We also propose two additional algorithms for sensor networks, one distributed and one centralized, and compare the performances of all the three algorithms.

5.1 Problem Formulation

The sensor network is assumed to consist of L sensors and a central decision maker called the fusion center. The sensors are indexed by the index $\ell \in \{1, \cdots, L\}$. In the following we say sensor ℓ to refer to the sensor indexed by ℓ. At sensor ℓ the sequence $\{X_{n,\ell}\}_{n\geq 1}$ is observed, where n is the time index. At some unknown time γ, the distribution of $\{X_{n,\ell}\}$ changes from $f_{0,\ell}$ to say $f_{1,\ell}$, $\forall \ell$. The random variables $\{X_{n,\ell}\}$ are independent across indices n and ℓ conditioned on γ. The distributions $f_{0,\ell}$ and $f_{1,\ell}$ are assumed to be known.

We now discuss the type of policies considered in this chapter. The policies are similar to the one considered for data-efficient settings in the previous chapters, but extended to the sensor network setting. In the quickest change detection models studied in [32], [33] and [34], observations are taken at each sensor at all times. Here we consider policies in which on-off observation control is employed at each sensor. At sensor ℓ, at each time $n, n \geq 0$, a decision is made as to whether to *take* or *skip* the observation at time $n+1$ at that sensor. Let $S_{n,\ell}$ be the indicator random variable such that

$$S_{n,\ell} = \begin{cases} 1 & \text{if } X_{n,\ell} \text{ is used for decision making at sensor } \ell \\ 0 & \text{otherwise.} \end{cases}$$

Let $\phi_{n,\ell}$ be the observation control law at sensor ℓ, i.e.,

$$S_{n+1,\ell} = \phi_{n,\ell}(\mathcal{I}_{n,\ell}),$$

where $\mathcal{I}_{n,\ell} = \left[S_{1,\ell}, \ldots, S_{n,\ell}, X_{1,\ell}^{(S_{1,\ell})}, \ldots, X_{n,\ell}^{(S_{n,\ell})}\right]$. Here, $X_{n,\ell}^{(S_{n,\ell})} = X_{1,\ell}$ if $S_{1,\ell} = 1$, otherwise $X_{1,\ell}$ is absent from the information vector $\mathcal{I}_{n,\ell}$. Thus,

the decision to take or skip a sample at sensor ℓ is based on its past information. Let

$$Y_{n,\ell} = g_{n,\ell}(\mathcal{I}_{n,\ell})$$

be the information transmitted from sensor ℓ to the fusion center. If no information is transmitted to the fusion center, then $Y_{n,\ell} = \text{NULL}$, which is treated as zero at the fusion center. Here, $g_{n,\ell}$ is the transmission control law at sensor ℓ. Let

$$\boldsymbol{Y}_n = \{Y_{n,1}, \cdots, Y_{n,L}\}$$

be the information received at the fusion center at time n, and let τ be a stopping time on the sequence $\{\boldsymbol{Y}_n\}$.

Let

$$\boldsymbol{\phi_n} = \{\phi_{n,1}, \cdots, \phi_{n,L}\},$$

denote the observation control law at time n, and let

$$\boldsymbol{g}_n = \{g_{n,1}, \cdots, g_{n,L}\},$$

denote the transmission control law at time n. For data-efficient quickest change detection in sensor networks we consider the policy of type Π defined as

$$\Pi = \{\tau, \{\boldsymbol{\phi}_0, \cdots, \boldsymbol{\phi}_{\tau-1}\}, \{\boldsymbol{g}_1, \cdots, \boldsymbol{g}_\tau\}\}.$$

To capture the cost of observations used at each sensor before change, we use the Pre-Change Duty Cycle (PDC) metric introduced in Chapter 4. The PDC_ℓ, the PDC for sensor ℓ is defined as

$$\mathsf{PDC}_\ell(\Pi) = \limsup_{\gamma\to\infty} \frac{1}{\gamma}\mathbb{E}_\infty\left[\sum_{k=1}^{\gamma-1} S_{k,\ell}\right]. \tag{5.1}$$

Thus, PDC_ℓ is the fraction of time observations are taken before change at sensor ℓ. If all the observations are used at sensor ℓ, then $\mathsf{PDC}_\ell = 1$. If every second sample is skipped at sensor ℓ, then $\mathsf{PDC}_\ell = 0.5$.

We now propose data-efficient extensions of Problem 4.1.2 and Problem 4.1.1 for sensor networks. Let

$$\boldsymbol{\mathcal{I}}_n = \{\mathcal{I}_{n,1}, \cdots, \mathcal{I}_{n,L}\},$$

be the information available at time n across the sensor network. For extension of Problem 4.1.2, we consider the delay and false alarm metrics used in [12]: the Worst case Average Detection Delay (WADD)

$$\mathsf{WADD}(\Pi) = \sup_{\gamma} \operatorname{ess\,sup}\ \mathbb{E}_\gamma \left[(\tau-\gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1}\right], \tag{5.2}$$

and the False Alarm Rate (FAR)

$$\mathsf{FAR}(\Pi) = 1/\mathbb{E}_\infty [\tau]. \tag{5.3}$$

The extension of Problem 4.1.2 to sensor networks is

Problem 5.1.1.

$$\begin{array}{ll} \underset{\Pi}{\textit{minimize}} & \mathsf{WADD}(\Pi), \\ \textit{subject to} & \mathsf{FAR}(\Pi) \leq \alpha, \\ & \mathsf{PDC}_\ell(\Pi) \leq \beta_\ell, \quad \textit{for} \ \ \ell = 1, \cdots, L. \end{array} \tag{5.4}$$

Here, $0 \leq \alpha, \beta_\ell \leq 1$*, for* $\ell = 1, \cdots, L$*, are given constraints.*

We also consider the extension of Problem 4.1.1, where instead of WADD, the CADD metric

$$\mathsf{CADD}(\Pi) = \sup_{\gamma} \ \ \mathbb{E}_\gamma [\tau - \gamma | \tau \geq \gamma] \tag{5.5}$$

is used:

Problem 5.1.2.

$$\begin{array}{ll} \underset{\Pi}{\textit{minimize}} & \mathsf{CADD}(\Pi), \\ \textit{subject to} & \mathsf{FAR}(\Pi) \leq \alpha, \\ & \mathsf{PDC}_\ell(\Pi) \leq \beta_\ell, \quad \textit{for} \ \ \ell = 1, \cdots, L. \end{array} \tag{5.6}$$

Here, $0 \leq \alpha, \beta_\ell \leq 1$*, for* $\ell = 1, \cdots, L$*, are given constraints.*

The lower bound developed in [14] can be specialized to sensor networks. Let

$$\Delta_\alpha = \{\Pi : \mathsf{FAR}(\Pi) \leq \alpha\}.$$

Theorem 5.1.1 ([14]). *As* $\alpha \to 0$,

$$\inf_{\Pi \in \Delta_\alpha} \mathsf{CADD}(\Pi) \geq \frac{|\log \alpha|}{\sum_{\ell=1}^{L} D(f_{1,\ell} \,||\, f_{0,\ell})}(1 + o(1)). \tag{5.7}$$

Since $\mathsf{WADD}(\Pi) \geq \mathsf{CADD}(\Pi)$, we also have as $\alpha \to 0$,

$$\inf_{\Pi \in \Delta_\alpha} \mathsf{WADD}(\Pi) \geq \frac{|\log \alpha|}{\sum_{\ell=1}^{L} D(f_{1,\ell} \,||\, f_{0,\ell})}(1 + o(1)). \tag{5.8}$$

We note that the lower bound on the WADD was first obtained in [12].

We will be particularly interested in policies such that the information transmitted from the sensors to the fusion center at any time is a binary digit. That is we are primarily interested in policies in the class

$$\Delta_{(\alpha,\beta)}^{\{0,1\}} = \{\Pi : \mathsf{FAR}(\Pi) \leq \alpha;\ \mathsf{PDC}_\ell \leq \beta_\ell \ \text{ and } \ Y_{n,\ell} \in \{0,1\},\ \forall n, \ell\}. \tag{5.9}$$

The interest in the policies in the set $\Delta_{(\alpha,\beta)}^{\{0,1\}}$, where only a binary number is sent to the fusion center, stems from the fact that in these policies the information transmitted to the fusion center is the minimal. Thus, it represents in some sense the maximum possible compression of the transmitted information. The main objective of this chapter is to show that an algorithm from this class can be globally asymptotically optimal.

Specifically, we will propose an algorithm, called the DE-All algorithm, from the class $\Delta_{(\alpha,\beta)}^{\{0,1\}}$, and show that it is asymptotically optimal for both Problem 5.1.1 and Problem 5.1.2, i.e., the performance of the DE-All algorithm achieves the lower bound of [14] given in Theorem 5.1.1, for each fixed set of $\{\beta_\ell\}$, as $\alpha \to 0$.

5.2 Quickest Change Detection in Sensor Networks: Existing Literature

In this section we provide a brief overview of the existing literature relevant to this chapter.

We first describe the Centralized CuSum algorithm in a mathematically precise way.

Algorithm 5.2.1 (Centralized CuSum)**.** *Fix a threshold $A \geq 0$.*

1. *Use all the observations at the sensors, i.e.,*

$$S_{n,\ell} = 1, \ \forall n, \ell.$$

2. *Raw observations are transmitted from the sensors to the fusion center at each time step, i.e.,*

$$Y_{n,\ell} = X_{n,\ell} \ \forall n, \ell.$$

3. *The CuSum algorithm (see Algorithm* (3.1.1)*) is applied to the vector of observations received at the fusion center. That is, at the fusion center, the sequence $\{V_n\}$ is computed according to the following recursion: $V_0 = 0$, and for $n \geq 0$,*

$$V_{n+1} = \max \left\{ 0, \ V_n + \sum_{\ell=1}^{L} \log \frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})} \right\}. \tag{5.10}$$

A change is declared the first time V_n is above a threshold $A > 0$:

$$\tau_{\text{CC}} = \inf \{ n \geq 1 : V_n > A \}.$$

It is well known [12] that the performance of the Centralized CuSum algorithm is asymptotically equal to the lower bound provided in Theorem 5.1.1; see Theorem 3.3.3.

We now describe the ALL algorithm from [33]. Let

$$d_\ell = \frac{D(f_{1,\ell} \mid\mid f_{0,\ell})}{\sum_{k=1}^{L} D(f_{1,k} \mid\mid f_{0,k})}.$$

Algorithm 5.2.2 (ALL)**.** *Start with $C_{0,\ell} = 0$, $\forall \ell$, and fix $A \geq 0$.*

1. *At each sensor ℓ the CuSum statistic is computed over time:*

$$C_{n+1,\ell} = \max \left\{ 0, \ C_{n,\ell} + \log \frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})} \right\}.$$

Thus $S_{n,\ell} = 1 \ \forall n, \ell$.

2. *A "1" is transmitted from sensor ℓ to the fusion center if the CuSum statistic is above a threshold $d_\ell A$, i.e,*

$$Y_{n,\ell} = \mathbb{I}_{\{C_{n,\ell} > d_\ell A\}}.$$

3. *A change is declared when a "1" is transmitted from all the sensors at the same time, i.e.,*

$$\tau_{\text{All}} = \inf\{n \geq 1 : Y_{n,\ell} = 1 \ \forall \ell\}.$$

The ALL algorithm has a surprising optimality property proved in [33].

Theorem 5.2.1 ([33])**.** *If the absolute moments up to the third order of* $\log L(X)$ *are finite and positive, then with* $A = |\log \alpha|$*, we have* $\alpha \to 0$*.*

$$\begin{aligned} \mathsf{FAR}(\tau_{\text{All}}) &\leq \alpha(1 + o(1)), \\ \mathsf{WADD}(\tau_{\text{All}}) &\leq \frac{|\log \alpha|}{\sum_{\ell=1}^{L} D(f_{1,\ell} \,||\, f_{0,\ell})}(1 + o(1)). \end{aligned} \tag{5.11}$$

Thus, the ALL scheme achieves the asymptotic lower bound in Theorem 5.1.1, which is also the performance of the Centralized CuSum algorithm. In this sense, the ALL scheme is globally asymptotically optimal as the false alarm rate goes to zero. It is important to note that such an optimality is obtained by sending such a minimal amount of information (binary digits) from the sensors to the fusion center.

However, we note that $\mathsf{PDC}_\ell = 1$, $\forall \ell$, for both the Centralized CuSum algorithm and the ALL algorithm. Hence, neither the Centralized CuSum algorithm nor the ALL algorithm are asymptotically optimal for Problem 5.1.1 and Problem 5.1.2, when $\beta_\ell < 1$, for any ℓ.

Consider a policy in which, at each sensor every n^{th} sample is used, and raw observations are transmitted from each sensor to the fusion center, each time an observation is taken. At the fusion center, the CuSum algorithm, as defined above, is applied to the received samples. In this policy, the PDC_ℓ achieved is equal to $1/n$, $\forall \ell$. Using this scheme, any given constraints on the PDC_ℓ can be achieved by using every n^{th} sample, and by choosing a suitably large n. However, the detection delay for this scheme would be approximately n times that of the delay for the Centralized CuSum algorithm, for the same

false alarm rate.

Motivated by the results on the DE-CuSum algorithm from the previous chapters it is interesting to ask if the global asymptotic optimality of the ALL scheme can be retained if we replace the CuSum algorithm at each sensor by the DE-CuSum algorithm. We will show below that such an optimality result is indeed true. Specifically, we will propose an algorithm called the DE-All based on the DE-CuSum algorithm (see Algorithm 3.2.1) and show that the proposed algorithm can be designed to satisfy any FAR constraint of α, and PDC_ℓ constraints of $\{\beta_\ell\}$. The DE-All algorithm belong to the class $\Delta^{\{0,1\}}_{(\alpha,\beta)}$. Also, the WADD, and hence the CADD, of these algorithms equals the lower bound in Theorem 5.1.1, as $\alpha \to 0$.

5.3 The DE-All Algorithm

In the DE-All algorithm, the DE-CuSum algorithm (see Algorithm 3.2.1) is used at each sensor, and a "1" is transmitted each time the DE-CuSum statistic at any sensor is above a threshold. A change is declared the first time a "1" is received at the fusion center from *all* the sensors at the same time.

We use $W_{n,\ell}$ to denote the DE-CuSum statistic at sensor ℓ. Recall that

$$d_\ell = \frac{D(f_{1,\ell} \,||\, f_{0,\ell})}{\sum_{k=1}^{L} D(f_{1,k} \,||\, f_{0,k})}.$$

Algorithm 5.3.1 $(\mathrm{DE}-\mathrm{All})$**.** *Start with $W_{0,\ell} = 0 \ \forall \ell$. Fix $\mu_\ell > 0$, $h_\ell \geq 0$, and $A \geq 0$. For $n \geq 0$ use the following control:*

1. *Use the DE-CuSum algorithm at each sensor ℓ, i.e., update the statistics $\{W_{n,\ell}\}_{\ell=1}^{L}$ for $n \geq 1$ using*

$$\begin{aligned} S_{n+1,\ell} &= 1 \text{ only if } W_{n,\ell} \geq 0 \\ W_{n+1,\ell} &= \min\{W_{n,\ell} + \mu_\ell, 0\} \quad \text{if } S_{n+1,\ell} = 0 \\ &= \left(W_{n,\ell} + \log \frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})}\right)^{h+} \quad \text{if } S_{n+1} = 1 \end{aligned}$$

where $(x)^{h+} = \max\{x, -h\}$.

2. Transmit

$$Y_{n,\ell} = \mathbb{I}_{\{W_{n,\ell} > d_\ell A\}}.$$

3. At the fusion center stop at

$$\tau_{\text{DE-All}} = \inf\{n \geq 1 : Y_{n,\ell} = 1 \text{ for all } \ell \in \{1, \cdots, L\}\}.$$

If $h_\ell = 0$ $\forall \ell$ then the DE-CuSum algorithm used at each sensor reduces to the CuSum algorithm. Hence, the DE-All algorithm reduces to the ALL algorithm; see Algorithm 5.2.2.

5.4 Asymptotic Optimality of the DE-All Algorithm

In this section we prove the asymptotic optimality of the DE-All algorithm proposed in Section 5.3.

We define the ladder variable [16] corresponding to sensor ℓ:

$$\tau_{\ell-} = \inf\left\{ n \geq 1 : \sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} < 0 \right\}.$$

Then note that $W_{\tau_{\ell-}}$ is the ladder height.

Theorem 5.4.1. *Let moments of up to third order for the K-L divergences at each sensor be finite and positive. Let $\mu_\ell > 0$, $h_\ell < \infty$, $\forall \ell$, and $A = |\log \alpha|$. Then we have*

$$\begin{aligned}
\mathsf{FAR}(\Pi_{\text{DE-All}}) &\leq \mathsf{FAR}(\Pi_{\text{All}}) = \alpha(1 + o(1)), \text{ as } \alpha \to 0, \\
\mathsf{PDC}_\ell(\Pi_{\text{DE-All}}) &= \frac{\mathbb{E}_\infty[\tau_{\ell-}]}{\mathbb{E}_\infty[\tau_{\ell-}] + \mathbb{E}_\infty[\lceil |W^{h_\ell+}_{\tau_{\ell-}}| / \mu_\ell \rceil]}, \\
\mathsf{WADD}(\Pi_{\text{DE-All}}) &= \frac{|\log \alpha|}{\sum_{\ell=1}^{L} D(f_{1,\ell} \,||\, f_{0,\ell})}(1 + o(1)) \text{ as } \alpha \to 0.
\end{aligned} \tag{5.12}$$

If $h_\ell = \infty$, $\forall \ell$, then

$$\mathsf{PDC}_\ell(\Pi_{\text{DE-All}}) \leq \frac{\mu_\ell}{\mu_\ell + D(f_{0,\ell} \,||\, f_{1,\ell})}. \tag{5.13}$$

Proof. The FAR result follows from Lemma 3.3.3 and the FAR result of Π_{All} from Theorem 5.2.1. The results on PDC follows from Theorem 4.3.1. The

proof of the WADD is based on the techniques used in [33] and the properties of the DE-CuSum algorithm. The details are provided in Section 5.7. □

Since CADD $\leq$ WADD, we also have under the same assumptions as in Theorem 5.4.1

$$\mathsf{CADD}(\Pi_{\mathrm{DE-All}}) \leq \frac{|\log \alpha|}{\sum_{\ell=1}^{L} D(f_{1,\ell} \,||\, f_{0,\ell})}(1 + o(1)) \text{ as } \alpha \to 0. \tag{5.14}$$

The statements in Theorem 5.4.1 prove that the DE − All algorithm is asymptotically optimal for both Problem 5.1.1 and Problem 5.1.2, for each given $\{\beta_\ell\}$, as $\alpha \to 0$. This is because the WADD of $\Pi_{\mathrm{DE-All}}$ is asymptotically equal to the lower bound provided in From Theorem 5.1.1, as $\alpha \to 0$, and the PDC_ℓ is not a function of threshold A. Hence, the PDC_ℓ constraints can be satisfied independent of the FAR constraint α.

5.5 Data-Efficient Algorithms for Sensor Networks

In this section we propose two more algorithms that can be used to detect the change in a data-efficient way in a sensor network. The first one, the DE-Dist algorithm is a distributed algorithm, in the sense that the observation control is executed locally at each sensor. The second one, called the Serialized-DE-CuSum algorithm is a centralized control based algorithm in which the observations from all the sensors are combined to execute the observation control.

5.5.1 The DE-Dist Algorithm

In the DE-All algorithm information is transmitted very rarely to the fusion center and that too in the form of "1"s and "0"s. This essentially means that the decision on the change is effectively taken at the sensors, since the change is declared at the fusion center the first time the change is "sensed" by all the sensors simultaneously. Although the DE-All algorithm is asymptotically optimal, the optimality is of the first order. A careful look at the proof of Theorem 5.4.1 and the proof of Theorem 5.2.1 in [33] also reveals that the $\mathsf{WADD}(\Pi_{\mathrm{DE-All}})$ has a $\sqrt{|\log \alpha|}$ term in addition to the $|\log \alpha|$ term. This

results in poor performance for moderate values of FAR. In applications where the cost of communication is not severe, transmission can be allowed to happen more frequently. Also, since in modern communication networks information is sent in packets, instead of sending just binary digits, there is a possibility of sending more information per packet from the sensors to the fusion center. In this case, a different fusion technique can also be employed to improve on the performance. Motivated by these observations we propose the DE-Dist algorithm.

In the DE-Dist algorithm, the DE-CuSum algorithm is used at each sensor for observation control. If an observation is taken at a sensor, then the observation is transmitted to the fusion center. At the fusion center, the CuSum algorithm is applied to the information received from the sensors to detect the change. For a similar algorithm technique see [35] and [36].

Mathematically, the DE-Dist algorithm can be written as follows.

Algorithm 5.5.1 (DE-Dist)**.** *Start with* $W_{0,\ell} = 0\ \forall \ell$. *Fix* $\mu_\ell > 0$, $h_\ell \geq 0$ *and* $A \geq 0$. *For* $n \geq 0$*:*

1. *Use the DE-CuSum algorithm at each sensor* ℓ*, i.e., update the statistics* $\{W_{n,\ell}\}_{\ell=1}^{L}$ *for* $n \geq 1$ *using*

$$\begin{aligned} S_{n+1,\ell} &= 1 \text{ only if } W_{n,\ell} \geq 0 \\ W_{n+1,\ell} &= \min\{W_{n,\ell} + \mu_\ell, 0\} \quad \text{if } S_{n+1,\ell} = 0 \\ &= \left(W_{n,\ell} + \log \frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})}\right)^{h+} \quad \text{if } S_{n+1} = 1 \end{aligned}$$

   where $(x)^{h+} = \max\{x, -h\}$.

2. *Transmit* $Y_{n,\ell} = \log \frac{f_{1,\ell}(X_{n,\ell})}{f_{0,\ell}(X_{n,\ell})}$ *if* $S_{n,\ell} = 1$.

3. *At the fusion center compute the statistics* $\{F_n\}$ *using the CuSum recursion:* $F_0 = 0$ *and for* $n \geq 0$*,*

$$F_{n+1} = \max\{0, F_n + \sum_{\ell=1}^{L} Y_{n,\ell}\}. \tag{5.15}$$

   Stop and declare change at

$$\tau_{\text{DE-Dist}} = \inf\{n \geq 1 : F_n > A\}. \tag{5.16}$$

We will see in Section 5.6 that the DE-Dist algorithm significantly outperforms the DE-All algorithm. Since, the DE-All is asymptotically optimal, we conjecture that the DE-Dist algorithm is asymptotically optimal as well.

5.5.2 The Serialized-DE-CuSum Algorithm

We now propose an algorithm called the Serialized-DE-CuSum algorithm. It is a centralized control based algorithm in which the fusion center executes both the observation control and also makes the decision on stopping.

In the Serialized-DE-CuSum algorithm, the observation control is implemented by serializing the observations from the sensors. That is the sequence $\{X_{1,1}, X_{1,2}, \cdots, X_{1,L}, X_{2,1}, X_{2,2}, \cdots, X_{2,L}, \ldots\}$ is considered, and the DE-CuSum algorithm (with a fixed μ and h) is applied to this serialized sequence. If $\hat{W}_{n,\ell}$ is the DE-CuSum statistic computed using the above serialized observations sequence, then a change is declared at

$$\tau_{\text{SD}} = \inf\{n \geq 1 : \hat{W}_{n,L} > D\}.$$

If $f_{0,\ell} = f_0$ and $f_{1,\ell} = f_1$, i.e., the pre- and post-change distributions are the same for all the sensors, then it can be shown that setting $A = \log \frac{L}{\alpha}$ ensures that $\mathsf{FAR}(\tau_S) \leq \alpha$. Also the delay is equal to the lower bound of Theorem 5.1.1. Moreover, it can also be shown that a PDC constraint if suitably defined can be satisfied independent of the FAR constraint. However, note that since the observation control is executed by the fusion center, this algorithm is not a policy of type II considered in Problem 5.1.1 and Problem 5.1.2. Thus, we cannot claim asymptotic optimality with respect to these formulations. This algorithm will be essentially used as a benchmark for performance obtained in a data-efficient setting.

5.6 Numerical Results

In Fig. 5.1 we compare the CADD performance as a function of the FAR, of the DE-All algorithm with that of the DE-Dist algorithm and the Serialized-DE-CuSum algorithm. The parameters used in the simulations are: $L = 10$, $f_{0,\ell} = \mathcal{N}(0, 1)$, $\forall \ell$, and $f_{1,\ell} = \mathcal{N}(0.2, 1)$, $\forall \ell$. We see that the DE-Dist algo-

rithm performs better than the DE-All algorithm, by virtue of transmitting more information and using better fusion technique. Also, the Serialized-DE-CuSum algorithm performs better than both the DE-All algorithm and the DE-Dist algorithm. This is expected as in the Serialized-DE-CuSum the observation control is executed by the fusion center.

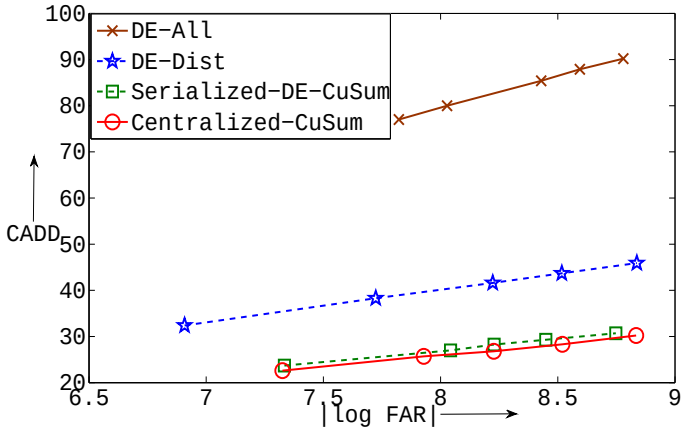


Figure 5.1: Trade-off curves for the algorithms studied: $h_\ell = 10$, and μ_ℓ is selected to satisfy the PDC constraint of 0.5.

In Fig. 5.2 we compare the CADD performance as a function of the FAR, of the ALL scheme, the DE-All algorithm, and the fractional sampling scheme. In the fractional sampling scheme, the ALL scheme is used and to meet the constraint on PDC_ℓ, samples are skipped randomly locally at each sensor.

The parameters used in the simulations are: $L = 10$, $f_0 = f_{0,\ell} = \mathcal{N}(0, 1)$, $\forall \ell$, and $f_1 = f_{1,\ell} = \mathcal{N}(0.4, 1)$, $\forall \ell$. The values of $\mu = \mu_\ell = 0.2$, and $h = h_\ell = 20$ are used to satisfy a PDC_ℓ constraint of 0.65 for each ℓ.

As shown in the figure the DE-All algorithm provides significant gain as compared to the fractional sampling scheme. In general, the gap in performance between the DE-All scheme and the fractional sampling scheme increases as a function of the Kullback-Leibler divergence between the pre- and post-change distributions.

5.7 Proofs of Various Results

We first define some quantities and set the notation to be used in the proof of Theorem 5.4.1. Let $\tau_{\mathrm{w},\ell}(x, y)$ be the time taken for the DE-CuSum statistic

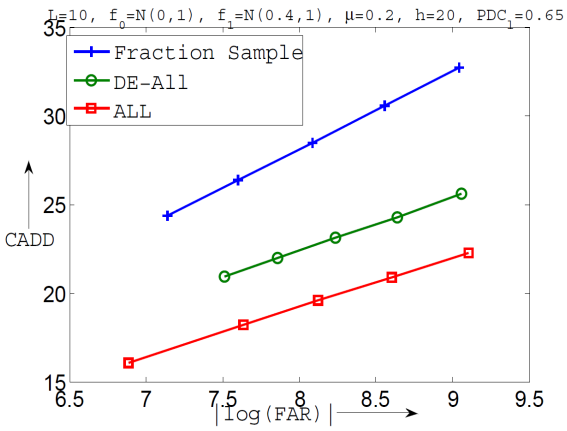


Figure 5.2: Trade-off curves for the algorithms studied: $L = 10$, $f_0 = f_{0,\ell} = \mathcal{N}(0,1)$, $\forall \ell$, and $f_1 = f_{1,\ell} = \mathcal{N}(0.4,1)$, $\forall \ell$. The values of $\mu = \mu_\ell = 0.2$, and $h = h_\ell = 20$ are used to satisfy a PDC_ℓ constraint of 0.65 for each ℓ.

at sensor ℓ to reach y, starting at $W_{0,\ell} = x$. Formally, for $x < y$ let

$$\tau_{\mathrm{W},\ell}(x,y) = \inf\{n \geq 1 : W_{n,\ell} > y; W_{0,\ell} = x\}.$$

If $x \geq y$, then $\tau_{\mathrm{W},\ell}(x,y) = 0$. Similarly define

$$\tau_{\mathrm{C},\ell}(x,y) = \inf\{n \geq 1 : C_{n,\ell} > y; C_{0,\ell} = x\},$$

where $C_{n,\ell}$ is the CuSum statistic at sensor ℓ when a CuSum algorithm is employed at sensor ℓ. Also define the corresponding time for a random walk to move from x to y:

$$\tau_{\mathrm{R},\ell}(x,y) = \inf\{n \geq 1 : x + \sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} > y\}.$$

Let $\nu_{\mathrm{W},\ell}(y)$ be the last time below y for the DE-CuSum statistic, i.e., for $y \geq 0$,

$$\nu_{\mathrm{W},\ell}(y) = \sup\{n \geq 1 : W_{n,\ell} \leq y; W_{0,\ell} = y\}. \tag{5.17}$$

Similarly define the last exit times for the CuSum algorithm

$$\nu_{\mathrm{C},\ell}(y) = \sup\{n \geq 1 : C_{n,\ell} \leq y; C_{0,\ell} = y\},$$

and for the random walk

$$\nu_{\mathrm{R},\ell} = \sup\{n \geq 1 : \sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \leq 0\}.$$

For simplicity we refer to the stopping for the DE-All algorithm simply by τ_a.

Proofs of Various Results. Our proof follows the outline of the proof of Theorem 3 in [33], but the details here are slightly more involved.

We obtain an upper bound on $\mathbb{E}_\gamma\left[(\tau_a - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1}\right]$ that is not a function of γ and the conditioning $\boldsymbol{\mathcal{I}}_{\gamma-1}$, and that scales as the lower bound in Theorem 5.1.1. The theorem is then established if we then take the essential supremum and then the supremum over γ.

Let $\boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1}$ be such that $W_{\gamma-1,\ell} = w_\ell$, $w_\ell \in [-h_\ell, \infty)$, $\forall \ell$. We first note that

$$\mathbb{E}_\gamma\left[(\tau_a - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1}\right] \leq \mathbb{E}_1\left[\max_{1\leq\ell\leq L}\{\tau_{\mathrm{W},\ell}(w_\ell, d_\ell A) + \nu_{\mathrm{W},\ell}(W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)})\}\right]. \tag{5.18}$$

By definition $W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)} \geq d_\ell A$. See Fig. 5.3 for a typical evolution of the DE-CuSum statistic at a sensor ℓ, showing the first passage time $\tau_{\mathrm{W},\ell}(w_\ell, d_\ell D)$, and the last exit time $\nu_{\mathrm{W},\ell}(y_\ell)$, where $y_\ell := W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)}$ for simplicity of representation.

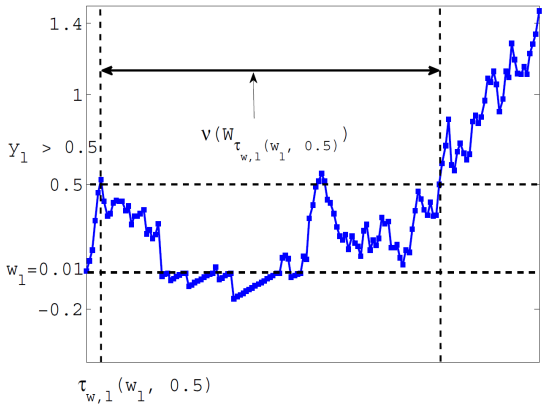


Figure 5.3: Typical evolution of the DE-CuSum algorithm showing the first passage time $\tau_{\mathrm{W},\ell}$, and the last exit time $\nu_{\mathrm{W},\ell}$ with $w_\ell = 0.01$, $d_\ell A = 0.5$.

It is easy to see that

$$\begin{aligned}&\mathbb{E}_\gamma\left[(\tau_a-\gamma)^+|\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}\right]\\&\quad\leq\ \mathbb{E}_1\left[\max_{1\leq\ell\leq L}\{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)\}\right]+\mathbb{E}_1\left[\max_{1\leq\ell\leq L}\{\nu_{\mathrm{W},\ell}(W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)})\}\right].\end{aligned}\tag{5.19}$$

We now show that the second term on the right-hand side of (5.19) is bounded by a constant, and the first term on the right-hand side of (5.19) is $\frac{A}{\sum_{\ell=1}^{L}D(f_{1,\ell}\,||\,f_{0,\ell})}+O(\sqrt{A})$.

For the second term, from Lemma 5.7.1 below, we have

$$\mathbb{E}_1\left[\max_{1\leq\ell\leq L}\{\nu_{\mathrm{W},\ell}(W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)})\}\right]\leq\sum_{\ell=1}^{L}\mathbb{E}_1\left[\nu_{\mathrm{W},\ell}(W_{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)})\right]\leq LK_3,\tag{5.20}$$

where K_3 is a constant, not a function of the conditioning w_ℓ, d_ℓ, $\forall\ell$, and the threshold A. Thus (5.19) can be written as

$$\mathbb{E}_\gamma\left[(\tau_a-\gamma)^+|\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}\right]\leq\ \mathbb{E}_1\left[\max_{1\leq\ell\leq L}\{\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)\}\right]+LK_3.\tag{5.21}$$

For the first term on the right, we write the random variable $\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)$ in terms of $\tau_{\mathrm{C},\ell}(w_\ell,d_\ell A)$. We first assume that $0\leq w_\ell\leq d_\ell A$. Note that $\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)$ is the time for the DE-CuSum statistic $W_{n,\ell}$ to reach $d_\ell A$ starting with $W_{0,\ell}=w_\ell$. And this time to hit $d_\ell A$ may have multiple sojourns of the statistic $W_{n,\ell}$ below 0. Thus, the time $\tau_{\mathrm{W},\ell}(w_\ell,d_\ell A)$ can be written as the sum of random times. Motivated by this we define a set of new variables. In the following, we often suppress the dependence on the index ℓ for simplicity.

Let

$$\tau_1(w_\ell)=\inf\{n\geq 1: W_{n,\ell}\notin[0,d_\ell A]\text{ with }W_{0,\ell}=w_\ell\}.$$

This is the first time for the DE-CuSum statistic, starting at $W_{0,\ell}=w_\ell$, to either hit $d_\ell A$ or go below 0. On paths over which $W_{\tau_1,\ell}<0$, we know that a number of consecutive samples are skipped depending on the undershoot of the observations. Let $t_1(w_\ell)$ be the number of consecutive samples skipped after $\tau_1(w_\ell)$ on such paths. On such paths again, let

$$\tau_2(w_\ell)=\inf\{n>\tau_1(w_\ell)+t_1(w_\ell): W_{n,\ell}\notin[0,d_\ell A]\}.$$

Thus, on paths such that $W_{\tau_1,\ell} < 0$, after the times $\tau_1(w_\ell)$ and the number of skipped samples $t_1(w_\ell)$, the statistic $W_{n,\ell}$ reaches 0 from below. The time $\tau_2(w_\ell)$ is the first time for $W_{n,\ell}$ to either cross A or go below 0, after time $\tau_1(w_\ell) + t_1(w_\ell)$. We define, $t_2(w_\ell)$, $\tau_3(w_\ell)$, etc. similarly. Next let

$$N_\ell(w_\ell) = \inf\{k \geq 1 : W_{\tau_k,\ell} > d_\ell A\}.$$

For simplicity we introduce the notion of "cycles", "success" and "failure". With reference to the definitions of $\tau_k(w_\ell)$'s above, we say that a success has occurred if the statistic $W_{n,\ell}$, starting with $W_{0,\ell} = w_\ell$, crosses $d_\ell A$ before it goes below 0. In that case we also say that the number of cycles to $d_\ell A$ is 1. If on the other hand, the statistic $W_{n,\ell}$ goes below 0 before it crosses $d_\ell A$, we say a failure has occurred. On paths such that $W_{\tau_1,\ell} < 0$, and after the times $\tau_1(w_\ell)$ and the number of skipped samples $t_1(w_\ell)$, the statistic $W_{n,\ell}$ reaches 0 from below. We say that the number of cycles is 2, if now the statistic $W_{n,\ell}$ crosses $d_\ell A$ before it goes below 0. Thus, $N_\ell(w_\ell)$ is the number of cycles to success at sensor ℓ.

Let

$$q_\ell = \mathbb{P}_1 \left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0, \ \forall n \right).$$

From [16] it is well known that $q_\ell > 0$. We claim that

$$\mathbb{E}_1[N_\ell(w_\ell)] \leq \frac{1}{q_\ell}. \tag{5.22}$$

Thus, $N_\ell(w_\ell) < \infty$, a.s. $\mathbb{P}_1$.

If (5.22) is indeed true then we can define $\lambda_1(w_\ell) = \tau_1(w_\ell)$, $\lambda_2(w_\ell) = \tau_2(w_\ell) - \tau_1(w_\ell) - t_1(w_\ell)$, etc., to be the lengths of the sojourns of the statistic $W_{n,\ell}$ above 0. Then clearly we have

$$\tau_{\mathrm{w},\ell}(w_\ell, d_\ell A) = \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) + \sum_{k=1}^{N_\ell(w_\ell)-1} t_k(w_\ell).$$

If $w_\ell < 0$, then note that there will be an additional initial sojourn of the statistic $W_{n,\ell}$ below 0, equal to $\lceil |w_\ell^{h_\ell +}| / \mu_\ell \rceil$. This is followed by a delay term

which corresponds to $w_\ell = 0$. Thus, in this case we can write

$$\tau_{\mathrm{W},\ell}(w_\ell, d_\ell A) = \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) + \sum_{k=1}^{N_\ell(w_\ell)} t_k(w_\ell).$$

Such a statement is also valid even if $w_\ell > A$ because the right-hand side of the above equation is positive.

Substituting this in (5.21) we have

$$\begin{aligned}
&\mathbb{E}_\gamma \left[(\tau_a - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1} \right] \\
&\quad \leq \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \tau_{\mathrm{W},\ell}(w_\ell, d_\ell A) \right\} \right] + LK_3 \\
&\quad \leq \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) + \sum_{k=1}^{N_\ell(w_\ell)} t_k(w_\ell) \right\} \right] + LK_3 \\
&\quad \leq \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) \right\} \right] + \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \sum_{k=1}^{N_\ell(w_\ell)} t_k(w_\ell) \right\} \right] + LK_3 \\
&\quad \leq \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) \right\} \right] + \sum_{\ell=1}^{L} \frac{\lceil h_\ell / \mu_\ell \rceil}{q_\ell} + LK_3.
\end{aligned} \tag{5.23}$$

The last inequality is true because

$$t_k(w_\ell) \leq \lceil h_\ell / \mu_\ell \rceil, \ \forall w_\ell, k, \ell$$

and because of (5.22).

We now make an important observation. We observe that because of the i.i.d. nature of the observations

$$\tau_{\mathrm{C},\ell}(w_\ell, d_\ell A) \stackrel{d}{=} \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell),$$

where we have used the symbol $\stackrel{d}{=}$ to denote equality in distribution. Thus,

$$\mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \left\{ \sum_{k=1}^{N_\ell(w_\ell)} \lambda_k(w_\ell) \right\} \right] = \mathbb{E}_1 \left[\max_{1 \leq \ell \leq L} \tau_{\mathrm{C},\ell}(w_\ell, d_\ell A) \right].$$

But, by sample-pathwise arguments it follows that

$$\mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{C},\ell}(w_\ell, d_\ell A)\right] \le \mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{C},\ell}(0, d_\ell A)\right].$$

This gives us

$$\begin{aligned}
&\mathbb{E}_\gamma\left[(\tau_a-\gamma)^+|\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}\right]\\
&\quad\le\ \mathbb{E}_1\left[\max_{1\le\ell\le L}\left\{\sum_{k=1}^{N_\ell(w_\ell)}\lambda_k(w_\ell)\right\}\right]+\sum_{\ell=1}^{L}\frac{\lceil h_\ell/\mu_\ell\rceil}{q_\ell}+LK_3\\
&\quad=\ \mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{C},\ell}(w_\ell, d_\ell A)\right]+\sum_{\ell=1}^{L}\frac{\lceil h_\ell/\mu_\ell\rceil}{q_\ell}+LK_3 \qquad (5.24)\\
&\quad\le\ \mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{C},\ell}(0, d_\ell A)\right]+\sum_{\ell=1}^{L}\frac{\lceil h_\ell/\mu_\ell\rceil}{q_\ell}+LK_3\\
&\quad\le\ \mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{R},\ell}(0, d_\ell A)\right]+\sum_{\ell=1}^{L}\frac{\lceil h_\ell/\mu_\ell\rceil}{q_\ell}+LK_3.
\end{aligned}$$

We note that the right-hand side of (5.24) is not a function of γ and the conditioning $\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}$ anymore. The theorem thus follows by taking ess sup on the left-hand side followed by a sup over time index γ, and recalling the result from the proof of Theorem 3 of [33] that $\mathbb{E}_1\left[\max_{1\le\ell\le L}\tau_{\mathrm{R},\ell}(0,d_\ell A)\right]$ is of the order of $\frac{A}{\sum_{\ell=1}^{L}D(f_{1,\ell}\,||\,f_{0,\ell})}+O(\sqrt{A})$.

The proof of the theorem will be complete if we prove the claim (5.22).

With the identity

$$\mathbb{E}_1[N_\ell(w_\ell)]=\sum_{k=1}^{\infty}\mathbb{P}_1(N_\ell(w_\ell)\ge k)$$

in mind, and using the terminology of cycles, success and failure defined earlier, we write

$$\begin{aligned}
\mathbb{P}_1(N_\ell(w_\ell)\ge k) = \mathbb{P}_1(&\text{fail in 1st cycle})\ \mathbb{P}_1(\text{fail in } 2^{nd} \text{ cycle}|\text{fail in first cycle})\\
&\cdots\mathbb{P}_1(\text{fail in } k-1^{st} \text{ cycle}|\text{fail in all previous}).
\end{aligned}$$

Now,

$$\mathbb{P}_1(\text{fail in } i^{th} \text{ cycle}|\text{fail in all previous}) \\ = 1 - \mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}).$$

We note that

$$\begin{aligned} \mathbb{P}_1(\text{success in } 1^{st} \text{ cycle}) &= \mathbb{P}_1(W_{\tau_1,\ell} > A) \\ &= \mathbb{P}_1 \left(\text{Statistic } W_{n,\ell} \text{ starting with } W_{0,\ell} = w_\ell \right. \\ &\qquad \left. \text{reaches } d_\ell A \text{ before it goes below 0}\right) \qquad (5.25) \\ &\geq \mathbb{P}_1 \left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0, \ \forall n \right) = q_\ell. \end{aligned}$$

Here, the last inequality follows because $\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \to \infty$ a.s. under $\mathbb{P}_1$, and hence the statistic $W_{n,\ell}$ reaches $d_\ell A$ before actually never coming below w_ℓ, and hence reaches $d_\ell A$ before going below 0. Note that the lower bound is not a function of the starting point w_ℓ.

Similarly, for the second cycle

$$\begin{aligned} &\mathbb{P}_1(\text{success in } 2^{nd} \text{ cycle}|\text{failure in first}) = \mathbb{P}_1 \left(W_{\tau_2,\ell} > A | W_{\tau_1,\ell} < 0)\right) \\ &= \mathbb{P}_1 \left(\text{Statistic } W_{n,\ell}, \text{ for } n > \tau_1(w_\ell) + t_1(w_\ell) \right. \\ &\qquad \left. \text{reaches } d_\ell A \text{ before it goes below 0}\right) \\ &= \mathbb{P}_1 \left(\text{Statistic } W_{n,\ell} \text{ starting with } W_{0,\ell} = 0 \right. \\ &\qquad \left. \text{reaches } d_\ell A \text{ before it goes below 0}\right) \\ &\geq \mathbb{P}_1 \left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0, \ \forall n \right) = q_\ell. \end{aligned}$$

Almost identical arguments for the other cycles proves that

$$\mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}) \geq q_\ell, \ \forall i.$$

As a result we get

$$\mathbb{P}_1(N_\ell(w_\ell) \geq k) \leq (1 - q_\ell)^{k-1}.$$

Note that the right-hand side is not a function of the initial point w_ℓ, nor is

a function of the threshold A. Hence,

$$\mathbb{E}_1[N_\ell(w_\ell)] = \sum_{k=1}^{\infty} \mathbb{P}_1(N_\ell(w_\ell) \geq k) \leq \sum_{k=1}^{\infty} (1-q_\ell)^{k-1} = \frac{1}{q_\ell} < \infty. \tag{5.26}$$

This proves the claim in (5.22) and proves the theorem. □

Lemma 5.7.1. *Let $\nu_{\mathrm{W},\ell}(w_\ell)$ as defined in (5.17) be the last exit time of the DE-CuSum statistic at sensor ℓ of the interval $(-\infty, w_\ell]$. Then if the variance of the log likelihood ratio at sensor ℓ is finite, then for all $w_\ell \geq 0$, and every ℓ,*

$$\begin{aligned} \mathbb{E}_1[\nu_{\mathrm{W},\ell}(w_\ell)] &\leq \mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] + K_1 \\ &\leq \mathbb{E}_1[\nu_{\mathrm{R},\ell}(w_\ell)] + K_1 = K_3 < \infty, \end{aligned} \tag{5.27}$$

where K_1 and K_3 are finite positive constants.

Proof. Throughout the proof, we often suppress the dependence on the sensor index ℓ. The evolution of the DE-CuSum statistic from $n = 1$ until $\nu_{\mathrm{W},\ell}(w_\ell)$ can be described as follows. The DE-CuSum starts at w_ℓ, and initially evolves like the CuSum algorithm, until either it goes below 0, or grows to ∞ without ever coming back to 0. Let $\mathcal{A}_1$ be the set of paths such that the DE-CuSum statistic grows to infinity without ever touching 0. In Fig. 5.3 consider the evolution of the DE-CuSum statistic by considering the time $\tau_{\mathrm{W},\ell}(w_\ell, d_\ell A)$ as the origin or time $n = 0$. Then the sample shown in Fig. 5.3 is a path from the set $\mathcal{A}_1^c$, which is the complement of the set $\mathcal{A}_1$. We define

$$\begin{aligned} \nu_1(w_\ell) &= \sup\{n \geq 1 : W_{n,\ell} \leq w_\ell; W_{0,\ell} = w_\ell\} \quad \text{on } \mathcal{A}_1 \\ &= \inf\{n \geq 1 : W_{n,\ell} < 0;\} \quad \text{on } \mathcal{A}_1^c. \end{aligned} \tag{5.28}$$

Thus, on the set $\mathcal{A}_1$, $\nu_1(w_\ell)$ is the last exit time for the level w_ℓ, and on the set $\mathcal{A}_1^c$, $\nu_1(w_\ell)$ is the first time to hit 0. We note that $\nu_1(w_\ell)$ is not a stopping time.

On the set $\mathcal{A}_1^c$, the DE-CuSum statistic goes below 0. Let $t_1(w_\ell)$ be the time taken for the DE-CuSum statistic to grow up to 0, once it goes below 0 at $\nu_1(w_\ell)$. Beyond $\nu_1(w_\ell) + t_1(w_\ell)$, the evolution of the DE-CuSum statistic is similar. Either it grows up to ∞ (say on set of paths $\mathcal{A}_2$), or it goes below

0 (say on set of paths $\mathcal{A}_2^c$). Thus, we define the variable

$$\begin{aligned}\nu_2(w_\ell) &= \sup\{n > \nu_1(w_\ell) + t_1(w_\ell) : W_{n,\ell} \leq w_\ell\} \quad \text{on } \mathcal{A}_2 \\ &= \inf\{n > \nu_1(w_\ell) + t_1(w_\ell) : W_{n,\ell} < 0;\} \quad \text{on } \mathcal{A}_2^c\end{aligned} \tag{5.29}$$

The variables $t_3(w_\ell)$ and $\nu_3(w_\ell)$, etc., can be similarly defined. We note that the variables here are similar to that used in the proof of Theorem 5.4.1, but the variables $\nu_k(w_\ell)$s here are not stopping times.

Also, let

$$N_\ell^\nu(w_\ell) = \inf\{k \geq 1 : W_{\nu_k(w_\ell),\ell} \geq 0\}.$$

As done in the proof of Theorem 5.4.1, we define the notion of "cycles", "success" and "failure". With reference to the definitions of $\nu_k(w_\ell)$'s above, we say that a success has occurred if the statistic $W_{n,\ell}$, starting with $W_{0,\ell} = w_\ell$, grows to infinity before it goes below 0. In that case we also say that the number of cycles to the last exit time is 1. If on the other hand, the statistic $W_{n,\ell}$ goes below 0, we say a failure has occurred. On paths such that $W_{\tau_1,\ell} < 0$, and after the times $\tau_1(w_\ell)$ and the number of skipped samples $t_1(w_\ell)$, the statistic $W_{n,\ell}$ reaches 0 from below. We say that the number of cycles is 2, if now the statistic $W_{n,\ell}$ grows to infinity before it goes below 0. Thus, $N_\ell(w_\ell)$ is the number of cycles to success at sensor ℓ. See Fig. 5.3, where in the figure $N_\ell^\nu = 7$.

Let

$$q_\ell = \mathbb{P}_1\left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0,\ \forall n\right).$$

We now show that

$$\mathbb{E}_1[N_\ell^\nu(w)] \leq \frac{1}{q_\ell} < \infty.$$

The last strict inequality is true because $q_\ell > 0$; see [16].

With the identity

$$\mathbb{E}_1[N_\ell^\nu(w_\ell)] = \sum_{k=1}^{\infty} \mathbb{P}_1(N_\ell^\nu(w_\ell) \geq k)$$

in mind, and using the terminology of cycles, success and failure defined above (and which are different from those used in the proof of Theorem 5.4.1), we

write

$$\mathbb{P}_1(N_\ell^\nu(w_\ell) \geq k) = \mathbb{P}_1(\text{fail in } 1^{st} \text{ cycle})\, \mathbb{P}_1(\text{fail in } 2^{nd} \text{ cycle}|\text{fail in } 1^{st} \text{ cycle}) \\ \cdots \mathbb{P}_1(\text{fail in } k-1^{st} \text{ cycle}|\text{fail in all previous}).$$

Now,

$$\mathbb{P}_1(\text{fail in } i^{th} \text{ cycle}|\text{fail in all previous}) \\ = 1 - \mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}).$$

We note that

$$\begin{aligned} &\mathbb{P}_1(\text{success in } 1^{st} \text{ cycle}) = \mathbb{P}_1(W_{\nu_1,\ell} > 0) \\ &= \mathbb{P}_1\left(\text{Statistic } W_{n,\ell} \text{ starting with } W_{0,\ell} = w_\ell \right. \\ &\qquad \left. \text{grows to } \infty \text{ before it goes below } 0\right) \\ &\geq \mathbb{P}_1\left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0,\ \forall n\right) = q_\ell. \end{aligned} \tag{5.30}$$

Here, the last inequality follows because $\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \to \infty$ a.s. under $\mathbb{P}_1$, and hence the statistic $W_{n,\ell}$ grows to infinity before never coming below $w_\ell \geq 0$. Note that the lower bound is not a function of the starting point w_ℓ.

Similarly, for the second cycle

$$\begin{aligned} &\mathbb{P}_1(\text{success in } 2^{nd} \text{ cycle}|\text{failure in first}) = \mathbb{P}_1\left(W_{\nu_2,\ell} > 0 | W_{\nu_1,\ell} < 0\right)) \\ &= \mathbb{P}_1\left(\text{Statistic } W_{n,\ell}, \text{ for } n > \nu_1(w_\ell) + t_1(w_\ell), \right. \\ &\qquad \left. \text{grows to } \infty \text{ before it goes below } 0\right) \\ &= \mathbb{P}_1\left(\text{Statistic } W_{n,\ell} \text{ starting with } W_{0,\ell} = 0 \right. \\ &\qquad \left. \text{grows to } \infty \text{ before it goes below } 0\right) \\ &= \mathbb{P}_1\left(\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} \geq 0,\ \forall n\right) = q_\ell. \end{aligned}$$

Almost identical arguments for the other cycles proves that

$$\mathbb{P}_1(\text{success in } i^{th} \text{ cycle}|\text{fail in all previous}) \geq q_\ell,\ \forall i.$$

As a result we get

$$\mathbb{P}_1(N_\ell^\nu(w_\ell) \geq k) \leq (1 - q_\ell)^{k-1}.$$

Note that the right-hand side is not a function of the initial point w_ℓ, nor is a function of the threshold A. Hence,

$$\mathbb{E}_1[N_\ell^\nu(w_\ell)] = \sum_{k=1}^{\infty} \mathbb{P}_1(N_\ell(w_\ell) \geq k) \leq \sum_{k=1}^{\infty} (1-q_\ell)^{k-1} = \frac{1}{q_\ell} < \infty. \tag{5.31}$$

Thus, $N_\ell^\nu(w_\ell) < \infty$, a.s. under $\mathbb{P}_1$ and we can define the following random variables: $\lambda_1^\nu(w_\ell) = \nu_1(w_\ell)$, $\lambda_2^\nu(w_\ell) = \nu_2(w_\ell) - \nu_1(w_\ell) - t_1(w_\ell)$, etc, to be the lengths of the sojourns of the statistic $W_{n,\ell}$ above 0. Then, $\nu_{\mathrm{W},\ell}(w_\ell)$ can be written as

$$\mathbb{E}_1[\nu_{\mathrm{W},\ell}(w_\ell)] = \mathbb{E}_1\left[\sum_{k=1}^{N_\ell^\nu(w)} \lambda_1^\nu(w_\ell)\right] + \mathbb{E}_1\left[\sum_{k=1}^{N_\ell^\nu(w)-1} t_k(w)\right]. \tag{5.32}$$

We observe that because of the i.i.d. nature of the observations

$$\mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] = \mathbb{E}_1\left[\sum_{k=1}^{N_\ell^\nu(w)} \lambda_1^\nu(w_\ell)\right].$$

As a result,

$$\mathbb{E}_1[\nu_{\mathrm{W},\ell}(w)] = \mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] + \mathbb{E}_1\left[\sum_{k=1}^{N_\ell^\nu(w)-1} t_k(w)\right]. \tag{5.33}$$

Now, $t_k(w) \leq \lceil h_\ell/\mu_\ell \rceil$, for any k, w_ℓ, and every ℓ. Thus, we have

$$\begin{aligned} \mathbb{E}_1[\nu_{\mathrm{W},\ell}(w)] &\leq \mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] + \mathbb{E}_1[N_\ell^\nu(w)]\lceil h_\ell/\mu_\ell \rceil \\ &\leq \mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] + \frac{\lceil h_\ell/\mu_\ell \rceil}{q_\ell}. \end{aligned} \tag{5.34}$$

The first inequality of the lemma follows from by setting $K_1 = \frac{\lceil h_\ell/\mu_\ell \rceil}{q_\ell}$. The rest of the lemma follows by noting that by definition of the CuSum algorithm

$$\mathbb{E}_1[\nu_{\mathrm{C},\ell}(w_\ell)] \leq \mathbb{E}_1[\nu_{\mathrm{R},\ell}(w_\ell)],$$

and the latter is finite, and not a function of w_ℓ, provided the variance of the log likelihood ratio is finite; see [33]. □

CHAPTER 6

DATA-EFFICIENT QUICKEST CHANGE DETECTION IN MULTI-CHANNEL SYSTEMS

In this chapter we study data-efficient quickest change detection in a multi-channel system. In a multi-channel system, a decision maker observes multiple independent streams of observations simultaneously. At the change point, the distribution of observations in an *unknown* subset of streams changes. The objective is to detect this change as quickly as possible without the knowledge of the affected subset. This problem is encountered for example in sensor networks, where either a change affects an unknown subset of sensors onboard a sensor node, or affects an unknown subset of sensor nodes in the network. We study the problem in the more general setup of sensor networks. For the special case when the problem does not involve a sensor network, the communication between the sensors and the fusion center can be ignored.

We propose extensions of the minimax problem formulations from Chapter 5 by now explicitly considering the cost of communication between the sensors and the fusion center, before the change point. The QCD problem in a multi-channel setting is studied in [28], [29] and [37]. However, in these papers neither the cost of observations at the sensors, nor the cost of communication between the sensor nodes and the fusion center is taken into account. We note that the papers [32], [33], [34], and [37] did consider the communication constraint by restricting the amount of transmitted information to either one bit or few bits, but the constraint on communication was not part of the problem formulation itself. The quickest change detection problem where the cost of communication is considered explicitly is studied in [38] and [36]. However, in these papers, the cost of observations at the sensors is not taken into account. Also, the problems were not considered in a multi-channel setting, i.e., in these papers the change affects all the sensors at the time of change.

For the sensor networks multi-channel problem we propose two algorithms:

the DE-Censor-Max algorithm and the DE-Censor-Sum algorithm. In both the algorithms the DE-CuSum algorithm (Algorithm 3.2) is applied to each stream in parallel, or used at each sensor; thus ensuring data-efficiency at the sensors. Again, in both the algorithms, the local DE-CuSum statistics is transmitted from the sensors to the fusion center, if the local DE-CuSum statistic is above a certain threshold; this is censoring. In the DE-Censor-Max algorithm, a change is declared at the fusion center when the *maximum* of the transmitted DE-CuSum statistics across the streams is above a threshold. In the DE-Censor-Sum algorithm, a change is declared at the fusion center when the *sum* of the transmitted DE-CuSum statistics across the streams is above a threshold. We will provide detailed performance analysis of these algorithms. The analysis will reveal that the DE-Censor-Max algorithm is asymptotically optimal for the problems proposed (for each possible post-change scenario), when the change affects *exactly one* stream, as the false alarm rate goes to zero. Also, under an assumption on a result in [29], the DE-Censor-Sum algorithm is asymptotically optimal for the problems proposed (for each possible post-change scenario), as the false alarm rate goes to zero.

We note that the multi-channel problem is similar to the problem studied in Chapter 4 on the composite post-change hypothesis. The difference here is that there are multiple independent streams of observations and the observation control can be implemented at each stream independently.

6.1 Problem Formulation

For the sensor network multi-channel problem we consider the same type of policies that we considered for sensor networks in Chapter 5. We also consider the same metrics for delay, false alarm and the cost of observation at sensors. We will however introduce a new metric for the cost of communication between the sensors and the fusion center. We reproduce the definitions of policy and metrics of Chapter 5 here for completeness.

The sensor network is assumed to consist of L sensors and a central decision maker called the fusion center. The sensors are indexed by the index $\ell \in \{1, \cdots, L\}$. At sensor ℓ the sequence $\{X_{n,\ell}\}_{n\geq 1}$ is observed, where n is the time index. At γ, the distribution of $\{X_{n,\ell}\}$ in a subset

$\kappa = \{k_1, k_2, \cdots, k_m\} \subset \{1, 2, \cdots, L\}$ of the streams changes, from $f_{0,\ell}$ to say $f_{1,\ell}$. The random variables $\{X_{n,\ell}\}$ are independent across indices n and ℓ conditioned on γ and the affected subset κ. The distributions $f_{0,\ell}$ and $f_{1,\ell}$ are assumed to be known, but neither the affected subset κ nor its size m is known.

Let $S_{n,\ell}$ be the indicator random variable such that

$$S_{n,\ell} = \begin{cases} 1 & \text{if } X_{n,\ell} \text{ is used for decision making at sensor } \ell \\ 0 & \text{otherwise.} \end{cases}$$

Let $\phi_{n,\ell}$ be the observation control law at sensor ℓ, i.e.,

$$S_{n+1,\ell} = \phi_{n,\ell}(\mathcal{I}_{n,\ell}),$$

where $\mathcal{I}_{n,\ell} = \left[S_{1,\ell}, \ldots, S_{n,\ell}, X_{1,\ell}^{(S_{1,\ell})}, \ldots, X_{n,\ell}^{(S_{n,\ell})}\right]$. Here, $X_{n,\ell}^{(S_{n,\ell})} = X_{1,\ell}$ if $S_{1,\ell} = 1$, otherwise $X_{1,\ell}$ is absent from the information vector $\mathcal{I}_{n,\ell}$. Thus, the decision to take or skip a sample at sensor ℓ is based on its past information. Let

$$\boldsymbol{\mathcal{I}}_n = \{\mathcal{I}_{n,1}, \cdots, \mathcal{I}_{n,L}\}$$

be the information available at time n across the sensor network.

Also let

$$Y_{n,\ell} = g_{n,\ell}(\mathcal{I}_{n,\ell})$$

be the information transmitted from sensor ℓ to the fusion center. If no information is transmitted to the fusion center, then $Y_{n,\ell} = \text{NULL}$, which is treated as zero at the fusion center. Here, $g_{n,\ell}$ is the transmission control law at sensor ℓ. Let

$$\boldsymbol{Y}_n = \{Y_{n,1}, \cdots, Y_{n,L}\}$$

be the information received at the fusion center at time n, and let τ be a stopping time on the sequence $\{\boldsymbol{Y}_n\}$.

Let

$$\boldsymbol{\phi_n} = \{\phi_{n,1}, \cdots, \phi_{n,L}\}$$

denote the observation control law at time n, and let

$$\boldsymbol{g}_n = \{g_{n,1}, \cdots, g_{n,L}\}$$

denote the transmission control law at time n. For data-efficient quickest change detection in sensor networks we consider the policy of type Π defined as

$$\Pi = \{\tau, \{\boldsymbol{\phi}_0, \cdots, \boldsymbol{\phi}_{\tau-1}\}, \{\boldsymbol{g}_1, \cdots, \boldsymbol{g}_\tau\}\}.$$

The PDC_ℓ, the PDC for sensor ℓ, is defined as

$$\mathsf{PDC}_\ell(\Pi) = \limsup_{\gamma\to\infty} \frac{1}{\gamma}\mathbb{E}_\infty\left[\sum_{k=1}^{\gamma-1} S_{k,\ell}\right]. \tag{6.1}$$

Thus, PDC_ℓ is the fraction of time observations are taken before change at sensor ℓ.

To capture the cost of communication between each sensor and the fusion center before change, we propose the Pre-Change Transmission Cost (PTC) metric. For the definition we define

$$T_{n,\ell} = \begin{cases} 1 & \text{if } Y_{n,\ell} \neq \text{NULL, i.e, some information} \\ & \text{is transmitted to the fusion center} \\ 0 & \text{otherwise.} \end{cases}$$

The Pre-change Transmission Cost at sensor ℓ (PTC_ℓ) is defined as

$$\mathsf{PTC}_\ell(\Pi) = \limsup_{\gamma\to\infty} \frac{1}{\gamma}\mathbb{E}_\infty\left[\sum_{k=1}^{\gamma-1} T_{k,\ell}\right]. \tag{6.2}$$

If in a policy every sample is taken and some information is transmitted at every time slot at all the sensors, then for that policy $\mathsf{PDC}_\ell = \mathsf{PTC}_\ell = 1$, $\forall \ell$. If transmissions happen from the sensors only in every alternate time slots, then $\mathsf{PTC}_\ell = 0.5$, $\forall \ell$.

The objective here is to solve the following extensions of Problem 5.1.1 and Problem 5.1.2:

Problem 6.1.1.

$$\begin{aligned} \underset{\Pi}{\textit{minimize}} \quad & \mathsf{WADD}(\Pi), \\ \textit{subject to} \quad & \mathsf{FAR}(\Pi) \leq \alpha, \\ & \mathsf{PDC}_\ell(\Pi) \leq \beta_\ell, \quad \textit{for} \;\; \ell = 1, \cdots, L, \\ \textit{and} \quad & \mathsf{PTC}_\ell(\Pi) \leq \sigma_\ell, \quad \textit{for} \;\; \ell = 1, \cdots, L, \end{aligned} \tag{6.3}$$

where $0 \leq \alpha, \beta_\ell, \sigma_\ell \leq 1$*, for* $\ell = 1, \cdots, L$*, are given constraints, and*

Problem 6.1.2.

$$\begin{array}{rl} \underset{\Pi}{\textit{minimize}} & \mathsf{CADD}(\Pi), \\ \textit{subject to} & \mathsf{FAR}(\Pi) \leq \alpha, \\ & \mathsf{PDC}_\ell(\Pi) \leq \beta_\ell, \quad \textit{for} \quad \ell = 1, \cdots, L, \\ \textit{and} & \mathsf{PTC}_\ell(\Pi) \leq \sigma_\ell, \quad \textit{for} \quad \ell = 1, \cdots, L, \end{array} \tag{6.4}$$

where $0 \leq \alpha, \beta_\ell, \sigma_\ell \leq 1$*, for* $\ell = 1, \cdots, L$*, are given constraints.*

The asymptotic lower bound developed in [14] can be specialized to the multi-channel setting. Let

$$\Delta_\alpha = \{\Pi : \mathsf{FAR}(\Pi) \leq \alpha\}.$$

Theorem 6.1.1 ([14])**.** *If the affected subset post-change is* $\kappa = \{k_1, k_2, \cdots, k_m\}$*, then as* $\alpha \to 0$*,*

$$\inf_{\Pi \in \Delta_\alpha} \mathsf{CADD}(\Pi) \geq \frac{|\log \alpha|}{\sum_{i=1}^m D(f_{1,k_i} \,||\, f_{0,k_i})}(1 + o(1)). \tag{6.5}$$

Since $\mathsf{WADD}(\Pi) \geq \mathsf{CADD}(\Pi)$, we also have as $\alpha \to 0$,

$$\inf_{\Pi \in \Delta_\alpha} \mathsf{WADD}(\Pi) \geq \frac{|\log \alpha|}{\sum_{i=1}^m D(f_{1,k_i} \,||\, f_{0,k_i})}(1 + o(1)). \tag{6.6}$$

6.2 Data-Efficient Algorithms for Multi-Channel Systems

In this section we propose two algorithms that can be used to detect the change in a data-efficient way in a multi-channel system. In both the algorithms the DE-CuSum algorithm (Algorithm 3.2) is used locally at each sensor. In the rest of the chapter we use $W_{n,\ell}$ to denote the DE-CuSum statistic at sensor ℓ at time n.

6.2.1 The DE-Censor-Max Algorithm

In the DE-Censor-Max algorithm, the DE-CuSum algorithm is used at each sensor ℓ. If the DE-CuSum statistic $W_{n,\ell}$ at a sensor is above a threshold D_ℓ, then the statistic is transmitted to the fusion center. A change is declared at the fusion center, if the maximum of the transmitted statistics from all the sensors is larger than another threshold A. Mathematically, the DE-Censor-Max algorithm is described as follows.

Algorithm 6.2.1 (DE-Censor-Max: Π_{DCM})**.** *Start with* $W_{0,\ell} = 0\ \forall \ell$. *Fix* $\mu_\ell > 0$, $h_\ell \geq 0$, $D_\ell \geq 0$ *and* $A \geq 0$. *For* $n \geq 0$ *use the following control:*

1. *Use the DE-CuSum algorithm at each sensor* ℓ, *i.e., update the statistics* $\{W_{n,\ell}\}_{\ell=1}^{L}$ *for* $n \geq 1$ *using*

$$\begin{aligned} S_{n+1,\ell} &= 1 \text{ only if } W_{n,\ell} \geq 0 \\ W_{n+1,\ell} &= \min\{W_{n,\ell} + \mu_\ell, 0\} \quad \text{if } S_{n+1,\ell} = 0 \\ &= \left(W_{n,\ell} + \log \frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})}\right)^{h+} \quad \text{if } S_{n+1} = 1, \end{aligned}$$

   where $(x)^{h+} = \max\{x, -h\}$.

2. *Transmit*

$$Y_{n,\ell} = W_{n,\ell} \mathbb{I}_{\{W_{n,\ell} > D_\ell\}},\ \forall \ell.$$

3. *At the fusion center stop at*

$$\tau_{\text{DCM}} = \inf\{n \geq 1 : \max_{\ell \in \{1,\cdots,L\}} Y_{n,\ell} > A\}.$$

With $D_\ell = 0$ and $h_\ell = 0$, $\forall \ell$, the DE-CuSum algorithm at each sensor reduces to the CuSum algorithm, and $Y_{n,\ell} = W_{n,\ell}\ \forall n, \ell$. In this case, the DE-Censor-Max algorithm reduces to the MAX algorithm proposed in [28].

We will show in the Section 6.3 that when exactly one of the L sensor is affected post-change, then this algorithm is uniformly asymptotically optimal for both Problem 6.1.1 and Problem 6.1.2 (achieves the lower bound provided in Theorem 6.1.1 for each κ), for each fixed $\{\beta_\ell\}$ and $\{\sigma_\ell\}$, as $\alpha \to 0$.

6.2.2 The DE-Censor-Sum Algorithm

Although the DE-Censor-Max algorithm is asymptotically optimal, we will show in Section 6.4 that it performs poorly when the size of the affected subset is large. To detect the change efficiently when the size of affected susbet is large, we propose the DE-Censor-Sum algorithm.

In the DE-Censor-Sum algorithm, the DE-CuSum algorithm is used at each sensor. If the DE-CuSum statistic at a sensor is above a threshold, then the statistic is transmitted to the fusion center. A change is declared at the fusion center, if the sum of the transmitted statistics from all the sensors is larger than another threshold. Mathematically, the DE-Censor-Sum algorithm is described as follows.

Algorithm 6.2.2 $(\mathrm{DE}-\mathrm{Censor}-\mathrm{Sum}\colon \Pi_{\mathrm{DCS}})$**.** *Start with* $W_{0,\ell}=0\ \forall \ell$*. Fix* $\mu_\ell>0$*,* $h_\ell\geq 0$*,* $D_\ell\geq 0$ *and* $A\geq 0$*. For* $n\geq 0$ *use the following control:*

1. *Use the DE-CuSum algorithm at each sensor* ℓ*, i.e., update the statistics* $\{W_{n,\ell}\}_{\ell=1}^{L}$ *for* $n\geq 1$ *using*

$$\begin{aligned} S_{n+1,\ell} &= 1 \ \textit{only if}\ W_{n,\ell}\geq 0 \\ W_{n+1,\ell} &= \min\{W_{n,\ell}+\mu_\ell, 0\} \quad \textit{if}\ S_{n+1,\ell}=0 \\ &= \left(W_{n,\ell}+\log\frac{f_{1,\ell}(X_{n+1,\ell})}{f_{0,\ell}(X_{n+1,\ell})}\right)^{h+} \quad \textit{if}\ S_{n+1}=1, \end{aligned}$$

   where $(x)^{h+}=\max\{x,-h\}$*.*

2. *Transmit*

$$Y_{n,\ell}=W_{n,\ell}\mathbb{I}_{\{W_{n,\ell}>D_\ell\}},\ \forall \ell.$$

3. *At the fusion center stop at*

$$\tau_{\mathrm{DCS}}=\inf\{n\geq 1:\sum_{\ell\in\{1,\cdots,L\}}Y_{n,\ell}>A\}.$$

With $D_\ell=0$ and $h_\ell=0$, $\forall \ell$, the DE-CuSum algorithm at each sensor reduces to the CuSum algorithm, and $Y_{n,\ell}=W_{n,\ell}\ \forall n,\ell$. In this case, the DE-Censor-Sum algorithm reduces to the N_{sum} algorithm proposed in [29]. If $h_\ell=0\ \forall \ell$ and $D>0$, the DE-Censor-Sum algorithm reduces to the N_{hard} algorithm proposed in [37]. The DE-Censor-Sum algorithm can easily be

modified to obtain data-efficient extensions of other algorithms proposed in [37].

We will provide a detailed performance analysis of the DE-Censor-Sum algorithm using which the threshold A and the parameter h_ℓ and μ_ℓ can be selected. We will use the performance analysis to show that, under an additional assumption on a result in [29], the DE-Censor-Sum algorithm is uniformly asymptotically optimal for both Problem 6.1.1 and Problem 6.1.2 (achieves the lower bound provided in Theorem 6.1.1 for each κ), for each fixed $\{\beta_\ell\}$ and $\{\sigma_\ell\}$, as $\alpha \to 0$.

6.3 Asymptotic Optimality of the DE-Censor-Max Algorithm

We now show that when exactly one of the sensor is affected post-change, then the DE-Censor-Max algorithm is asymptotically optimal for Problems 6.1.1 and Problem 6.1.2, for each fixed $\{\beta_\ell\}$, $\{\sigma_\ell\}$, as $\alpha \to 0$.

As in Section 5.4 we define the ladder variable [16] corresponding to sensor ℓ:

$$\tau_{\ell-} = \inf\left\{ n \geq 1 : \sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} < 0 \right\},$$

and note that $W_{\tau_{\ell-}}$ is the ladder height. Also, let

$$U_{D_\ell} = \left\{ \#\text{ times } : \sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})} > D_\ell \text{ before it is} < 0 \right\}.$$

Thus, U_{D_ℓ} is the number of times the random walk $\sum_{k=1}^{n} \log \frac{f_{1,\ell}(X_{k,\ell})}{f_{0,\ell}(X_{k,\ell})}$ is above D_ℓ before hitting 0. We note that $U_{D_\ell} = 0$ with a positive probability. We note that U_{D_ℓ} is also the number of times the DE-CuSum statistic $W_{n,\ell}$ is above D_ℓ before hitting 0.

Theorem 6.3.1. *Let*

$$0 < D(f_{1,\ell} \,||\, f_{0,\ell}) < \infty \quad \text{and} \quad 0 < D(f_{0,\ell} \,||\, f_{1,\ell}) < \infty \quad \forall \ell.$$

Let $\mu_\ell > 0$, $h_\ell < \infty$, $\forall \ell$, $D_\ell \geq 0$, and $A = \log \frac{L}{\alpha}$. If the change occurs in the

stream ℓ^*, then we have

$$\begin{aligned}
\mathsf{FAR}(\Pi_{\text{DCM}}) &\leq \alpha, \\
\mathsf{PDC}_\ell(\Pi_{\text{DCM}}) &= \frac{\mathbb{E}_\infty[\tau_{\ell-}]}{\mathbb{E}_\infty[\tau_{\ell-}] + \mathbb{E}_\infty[\lceil |W^{h_\ell+}_{\tau_{\ell-}}|/\mu_\ell \rceil]}, \ \forall \ell, \\
\mathsf{PTC}_\ell(\Pi_{\text{DCM}}) &= \frac{\mathbb{E}_\infty[U_{D_\ell}]}{\mathbb{E}_\infty[\tau_{\ell-}] + \mathbb{E}_\infty[\lceil |W^{h_\ell+}_{\tau_{\ell-}}|/\mu_\ell \rceil]}, \ \forall \ell, \\
\mathsf{WADD}(\Pi_{\text{DCM}}) &\leq \frac{|\log \alpha|}{D(f_{1,\ell^*} \mid\mid f_{0,\ell^*})}(1 + o(1)) \ \textit{as} \ \alpha \to 0.
\end{aligned} \tag{6.7}$$

If $h_\ell = \infty$, $\forall \ell$, then

$$\mathsf{PDC}_\ell(\Pi_{\text{DCM}}) \leq \frac{\mu_\ell}{\mu_\ell + D(f_{0,\ell} \mid\mid f_{1,\ell})}, \ \forall \ell. \tag{6.8}$$

Proof. The FAR result follows from Lemma 3.3.3 and Theorem 1 of [28]. The results on PDC_ℓ follows from Theorem 4.3.1. The results on PTC_ℓ follows also from the renewal reward theorem and the arguments are almost identical to those provide for PDC in Theorem 4.3.1. The delay proof is true because after change the max of statistics is greater than the statistics in which the change has taken place. Thus, the delay of the DE-Censor-Max algorithm is bounded from above by the delay of the DE-CuSum algorithm when applied to the affected sensor. Mathematically, the arguments is as follows.

We obtain an upper bound on $\mathbb{E}_\gamma\left[(\tau_{\text{DCM}} - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1}\right]$ that is not a function of γ and the conditioning $\boldsymbol{\mathcal{I}}_{\gamma-1}$, and that scales as the lower bound in Theorem 6.1.1. The theorem is then established if we then take the essential supremum and then the supremum over γ.

Let $\boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1}$ be such that $W_{\gamma-1,\ell} = x_\ell$, $x_\ell \in [-h_\ell, \infty)$. We first note that for $A > \max_\ell D_\ell$,

$$\mathbb{E}_\gamma\left[(\tau_{\text{DCM}} - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1}\right] \leq \mathbb{E}_1\left[\tau_{\text{W},\ell^*}(x_{\ell^*})\right], \tag{6.9}$$

where $\tau_{\text{W},\ell}(x_\ell)$ is the time for the DE-CuSum statistic $W_{n,\ell}$ to reach A starting with $W_{0,\ell} = x_\ell$; see (3.27). Then from Lemma 3.3.4 we have

$$\mathbb{E}_\gamma\left[(\tau_{\text{DCM}} - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1}\right] \leq \mathbb{E}_1\left[\tau_{\text{W},\ell^*}(x_{\ell^*})\right] \leq \mathbb{E}_1[\tau_{\text{W},\ell^*}] + \lceil h_{\ell^*}/\mu_{\ell^*} \rceil. \tag{6.10}$$

The result now follows from the proof of Theorem 3.3.5 on the DE-CuSum algorithm. □

Since $\mathsf{CADD} \leq \mathsf{WADD}$, we also have under the same assumptions as in Theorem 6.3.1

$$\mathsf{CADD}(\Pi_{\text{DCM}}) \leq \frac{|\log \alpha|}{D(f_{1,\ell^*} \,||\, f_{0,\ell^*})}(1 + o(1)) \text{ as } \alpha \to 0. \tag{6.11}$$

From Theorem 6.1.1, the WADD, and hence the CADD performance of the DE-Censor-Max algorithm is the best one can do when the change affects the stream ℓ^*, for given $\{\beta_\ell\}$ and $\{\sigma_\ell\}$, as $\alpha \to 0$. Also, the PDC_ℓ and the PTC_ℓ performances do not depend on the threshold A, thus the constraints $\{\beta_\ell\}$ and $\{\sigma_\ell\}$ can be satisfied independent of the FAR constraint α. Hence, the DE-Censor-Max algorithm is asymptotically optimal when the change affects exactly one stream, for both Problem 6.1.1 and Problem 6.1.2, for each given $\{\beta_\ell\}$ and $\{\sigma_\ell\}$, as $\alpha \to 0$.

6.3.1 Performance Analysis of the DE-Censor-Sum Algorithm

In this section we provide the performance analysis of the DE-Censor-Sum algorithm and then comment on its asymptotic optimality.

Theorem 6.3.2. *Let*

$$0 < D(f_{1,\ell} \,||\, f_{0,\ell}) < \infty \quad \textit{and} \quad 0 < D(f_{0,\ell} \,||\, f_{1,\ell}) < \infty \quad \forall \ell.$$

Let $\mu_\ell > 0$, $h_\ell < \infty$, $\forall \ell$, $D_\ell \geq 0$, and $A = L \log \frac{L}{\alpha}$. If the change affects the subset κ of streams, then we have

$$\begin{aligned}
\mathsf{FAR}(\Pi_{\text{DCS}}) &\leq \alpha, \\
\mathsf{PDC}_\ell(\Pi_{\text{DCS}}) &= \frac{\mathbb{E}_\infty[\tau_{\ell-}]}{\mathbb{E}_\infty[\tau_{\ell-}] + \mathbb{E}_\infty[\lceil |W^{h_\ell+}_{\tau_{\ell-}}|/\mu_\ell \rceil]}, \ \forall \ell, \\
\mathsf{PTC}_\ell(\Pi_{\text{DCS}}) &= \frac{\mathbb{E}_\infty[U_{D_\ell}]}{\mathbb{E}_\infty[\tau_{\ell-}] + \mathbb{E}_\infty[\lceil |W^{h_\ell+}_{\tau_{\ell-}}|/\mu_\ell \rceil]}, \ \forall \ell, \\
\mathsf{WADD}(\Pi_{\text{DCS}}) &\leq \frac{A}{\sum_{i=1}^m D(f_{1,k_i} \,||\, f_{0,k_i})}(1 + o(1)) \ \textit{as } A \to \infty.
\end{aligned} \tag{6.12}$$

If $h_\ell = \infty$, $\forall \ell$, then

$$\mathsf{PDC}_\ell(\Pi_{\text{DCS}}) \leq \frac{\mu_\ell}{\mu_\ell + D(f_{0,\ell} \,||\, f_{1,\ell})}, \ \forall \ell. \tag{6.13}$$

Proof. The proofs on PDC_ℓ and PTC_ℓ are identical to that provided in the Theorem 6.3.1.

For the FAR note that

$$\left\{ \sum_{\ell=1}^{L} W_{n,\ell} > A \right\} \subset \left\{ \max_{\ell \in \{1,\cdots,L\}} W_{n,\ell} > \frac{A}{L} \right\}.$$

For simplicity we write $\Pi_{\text{DCS}}(A)$ to represent DE-Censor-Sum algorithm when the threshold used at the fusion center is A. Similarly we use $\Pi_{\text{DCM}}(A/L)$ to represent DE-Censor-Max algorithm when the threshold used at the fusion center is A/L. Then the above subset relation implies

$$\mathsf{FAR}(\Pi_{\text{DCS}}(A)) \leq \mathsf{FAR}(\Pi_{\text{DCM}}(A/L)).$$

The FAR result follows because from Theorem 6.3.1 we have that

$$\mathsf{FAR}(\Pi_{\text{DCM}}(A/L)) \leq \alpha \quad \text{if} \quad A/L = \log L/\alpha.$$

For the WADD analysis, let $\tau_{\text{DCS}}(\kappa)$ denote the DE-Censor-Sum algorithm applied to only the streams in the affected subset κ. Further let $\boldsymbol{\mathcal{I}}_{\gamma-1}(\kappa)$ denote the information in the affected streams. Then

$$\begin{aligned} \mathbb{E}_\gamma \left[(\tau_{\text{DCS}} - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1} \right] &\leq \mathbb{E}_\gamma \left[(\tau_{\text{DCS}}(\kappa) - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1} = \boldsymbol{i}_{\gamma-1} \right] \\ &= \mathbb{E}_\gamma \left[(\tau_{\text{DCS}}(\kappa) - \gamma)^+ | \boldsymbol{\mathcal{I}}_{\gamma-1}(\kappa) = \boldsymbol{i}_{\gamma-1}(\kappa) \right]. \end{aligned} \tag{6.14}$$

Because of the above inequality, we can assume that the change affects all the subsets at the same time, i.e., $\kappa = \{1, \cdots, L\}$.

Now note that any A (see Algorithm 5.3.1)

$$\{W_{n,\ell} > d_\ell A, \ \forall \ell\} \subset \left\{ \sum_\ell W_{n,\ell} > \sum_\ell d_\ell A = A \right\}.$$

Hence, for A sufficiently large and from the proof of Theorem 5.4.1, we have

$$\begin{aligned}\mathbb{E}_\gamma\left[(\tau_{\text{DCS}}-\gamma)^+|\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}\right] &\leq \mathbb{E}_\gamma\left[(\tau_{\text{DE-All}}-\gamma)^+|\boldsymbol{\mathcal{I}}_{\gamma-1}=\boldsymbol{i}_{\gamma-1}\right] \\ &\leq \mathbb{E}_1\left[\max_{1\leq\ell\leq L}\tau_{\text{C},\ell}\right]+\text{ constant.}\end{aligned} \tag{6.15}$$

The proof of the theorem is now complete because we can now take ess sup and then sup over γ on the left-hand side. Then, from [33] it follows that $\mathbb{E}_1\left[\max_{1\leq\ell\leq L}\tau_{\text{C},\ell}\right]$ grows in the order $\frac{A}{\sum_{\ell=1}^{L}D(f_{1,\ell}\,||\,f_{0,\ell})}$. This when applied to the affected subset κ gives us the desired result on the WADD from (6.14). □

Note that the above theorem does not imply the asymptotic optimality of the DE-Censor-Sum algorithm, mainly due to the fact that the choice of the threshold is conservative. It only gives a delay bound of L times that of the lower bound in Theorem 6.1.1. However, if the threshold can be set to be of the order $\log 1/\alpha$ to satisfy the FAR constraint, then the above theorem establishes the uniform asymptotic optimality of the DE-Censor-Sum algorithm for each possible post-change distribution. It is claimed in [29] that such a result is indeed true. We thus have the following corollary.[1]

Corollary 6.3.2.1. *If Theorem 1 in [29] is indeed true, then under the conditions of Theorem 6.3.2 above, the DE-Censor-Sum algorithm is uniformly asymptotically optimal, for each possible κ, for each fixed $\{\beta_\ell\}$ and $\{\sigma_\ell\}$, as $\alpha\to 0$.*

6.4 Numerical Results

We first compare the performance of the DE-Censor-Sum algorithm, the DE-Censor-Max algorithm and the Centralized CuSum algorithm as a function of the number of affected stream. The Centralized CuSum scheme is designed by assuming that the affected subset post-change is known. We plot the CADD versus the number of affected stream comparison in Fig. 6.1 for the parameters: $\mathsf{FAR}=10^{-3}$, $L=100$, $f_{0,\ell}=f_0=\mathcal{N}(0,1)$, $\forall\ell$, $f_{1,\ell}=f_1=\mathcal{N}(0.5,1)$, $\forall\ell$, and for the $\{\mathsf{PDC}_\ell\}$ and $\{\mathsf{PTC}_\ell\}$ constraints of $\beta_\ell=\sigma_\ell=0.5$, $\forall\ell$. We also set the local thresholds $D_\ell=0$, $\forall\ell$. In the figure we see

[1] There is a gap in the proof of Theorem 1 in [29]. The result however is believed to be true.

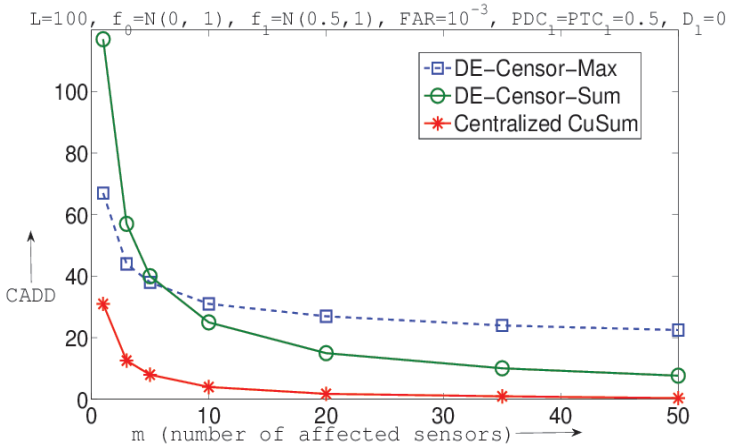


Figure 6.1: Comparison of DE-Censor-Sum algorithm, the DE-Censor-Max algorithm and the Centralized CuSum algorithm as a function of the number of affected stream.

that the DE-Censor-Max scheme outperforms the DE-Censor-Sum scheme when the number of affected streams is small. This is because the former is optimal when the number of affected stream is exactly one. However, when the number of affected streams is large, the DE-Censor-Sum algorithm outperforms the DE-Censor-Max algorithm. We note that this observation is consistent with the observations made in [29] regarding the comparison between the MAX and SUM algorithms.

In Fig. 6.2 we compare the CADD vs FAR performance of the DE-Censor-Sum algorithm with the fractional sampling scheme for $L = 10$, $f_{0,\ell} = \mathcal{N}(0,1)$, $\forall \ell$, $f_{1,\ell} = \mathcal{N}(0.2,1)$, $\forall \ell$, and for the $\{\mathsf{PDC}_\ell\}$ and $\{\mathsf{PTC}_\ell\}$ constraints of $\beta_\ell = \sigma_\ell = 0.5$ $\forall \ell$. We consider the post-change scenario when $m = 7$. We restrict our numerical study to the comparison of the CADD performance. Similar comparison can be obtained for the WADD as well.

In the fractional sampling scheme, the CuSum algorithm is used at each sensor, and samples are skipped based on the outcome of a sequence of fair coin tosses, independent of the observation process. If an observation is taken at a sensor, the CuSum statistic is transmitted to the fusion center. Thus, achieving the constraints on the $\{\mathsf{PDC}_\ell\}$ and $\{\mathsf{PTC}_\ell\}$. At the fusion center a change is declared the first time the sum of the CuSum statistics from all the sensors crosses a threshold. At the fusion center, in the absence of any transmission from a sensor, its CuSum statistics from the last time instant is used to compute the sum. For the DE-Censor-Sum algorithm, we set $D_\ell = 0$,

$\{h_\ell = h = 10\}$, $\forall \ell$, and use the approximation (6.8) to select μ_ℓ. This ensures that the $\{\mathsf{PDC}_\ell\}$ and $\{\mathsf{PTC}_\ell\}$ constraints are satisfied for the DE-Censor-Sum algorithm. In the figure we see that the DE-Censor-Sum algorithm provides a significant gain in performance as compared to the approach of fractional sampling.

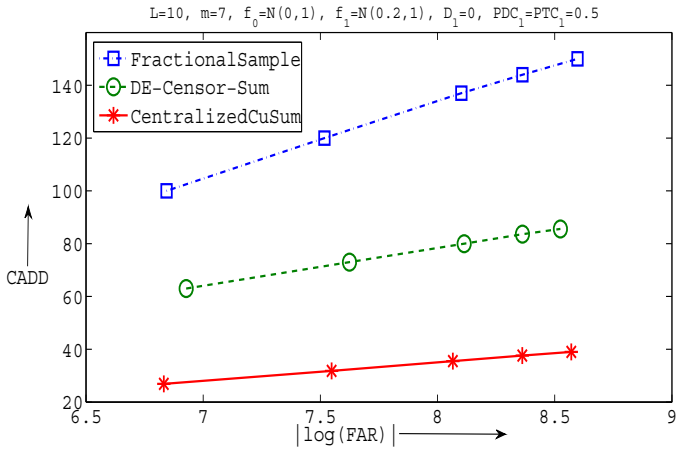


Figure 6.2: Comparison of the DE-Censor-Sum algorithm with the fractional sampling scheme.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this dissertation we studied data-efficient quickest change detection. The classical quickest change detection formulations are not suitable for applications where the change occurs rarely, and where taking observations before the change is costly. Thus, new formulations were needed where additional penalty is applied on the cost of observations used before the change point. We modified various classical quickest change detection formulations in the literature by adding such a penalty to the formulations.

We showed that on-off observation control can be introduced in the classical tests to make them data-efficient. In these new data-efficient tests, the likelihood ratio of the observations taken is used not only to detect change, but to also skip observations, if the observations provide a strong indication of no change. We showed that these new tests have the same asymptotic performance as the classical tests where all the observations are used for decision making. Thus, data-efficiency can be introduced without any loss in asymptotic performance, as the false alarm rate goes to zero. We now discuss possible future directions in which the theory discussed in this dissertation can be extended.

1. Non-i.i.d. settings: A major assumption throughout this dissertation has been the i.i.d. assumption, i.e., the observations are independent conditioned on the change point. In [15] and [14] the quickest change detection problem is studied in non-i.i.d. settings. In these works, conditions are identified under which generalizations of the Shiryaev and the CuSum algorithms are optimal in some non-i.i.d. settings. It would be interesting to check if some conditions can be identified under which the extension of the data-efficient algorithms studied here retain there optimality property in those non-i.i.d. settings.

2. Sensor networks: In Chapter 5, we studied the sensor network problem.

We proposed the DE-Dist algorithm that performs significantly better than the DE-All algorithm. There we conjectured that the DE-Dist algorithm is also asymptotically optimal given the fact that the DE-All is asymptotically optimal. An interesting problem for future work is to prove that this conjecture is indeed true. More generally, sensor network has a rich and complex literature where various different inference models are studied. It would be interesting to investigate the effect of observation control in such complex models.

3. Multi-channel setting: In Chapter 6 we studied the multi-channel problem. In this setting we only proved the asymptotic optimality of the DE-Censor-Max algorithm. It would be interesting to prove the optimality of the DE-Censor-Sum algorithm as well. In fact, it is still not known if there exists efficient and optimal algorithms in multi-channel setting, even in the classical setup.

4. Unknown post-change distribution: In Chapter 4 we studied the problem with unknown post-change distribution. We extended the GLRT based test to the data-efficient setting and proved its optimality under certain conditions. It would be interesting to prove optimality of the equivalent extension of mixture based tests. See the discussion at the end of Chapter 4. Also, we only concentrated on parametric setting and with unknown post-change distribution. It would be interesting to study data-efficiency when both the pre- and post-change distributions are not known, and/or the distributions do not belong to any parametric class. It is worth noting that one can design a DE-CuSum based test in one such scenario; see Fig. 7.1. Let $\{X_n\}$ be a sequence of random variable such that some function $g(.)$ of the observations changes in mean from μ_0 to say μ_1. If both the means are known then a CuSum algorithm to detect such a change would be

$$W_n = \left(W_{n-1} + g(X_n) - \frac{\mu_0 + \mu_1}{2} \right)^+ .$$

A DE-CuSum algorithm can be designed using this CuSum algorithm; see Fig. 7.1 with $g(x) = x$, $\mu_0 = 0$, and $\mu_1 = 0.7$.

5. Experiment design: The problem of on-off observation control belongs

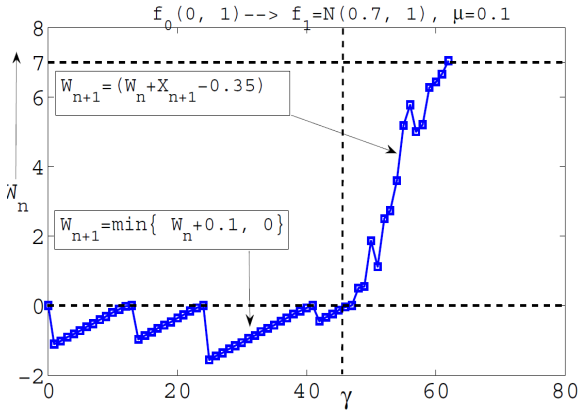


Figure 7.1: Nonparametric DECuSum.

to the more general problem of experiment design. It would be interesting to obtain lower bounds on the performance of any algorithm in this setting. Also, it would be interesting to identify algorithms that can achieve this lower bound. We note that the concept of DE-CuSum algorithm can be extended to design a test with experiment design; see Fig. 7.2. Consider a problem where there are four possible experiments to choose from. The experiments fetch observations with KL divergences decreasing in value, with the last experiment corresponds to no observation. A multi-threshold extension of the DE-CuSum algorithm can be used to choose the experiments. The evolution of such a test is shown in Fig. 7.2.

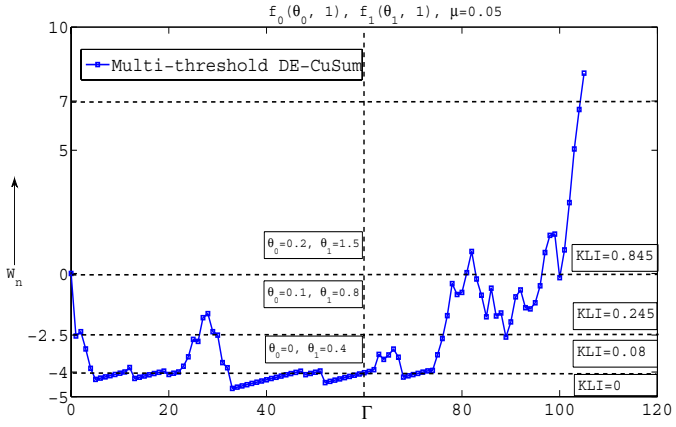


Figure 7.2: DECuSum with experiment design.

6. Continuous time: All the problems studied in the dissertation are discrete time problems, where the observations are collected in discrete time indices and the stopping variable is only allowed to stop at those times. There is a rich literature on quickest change detection in continuous time [2]. It would be interesting to investigate data-efficiency in a continuous time setting.

7. Fault isolation: The quickest change detection problem is also studied in combination with fault isolation in the literature; see [39], [4] and the references therein. In this problem the post-change hypothesis is composite, and the objective for the decision maker is not only to detect the change, but also to isolate the possible post-change distribution at the time of stopping. It would be interesting to investigate the effect of data-efficiency on the rate of false isolation.

REFERENCES

[1] V. V. Veeravalli and T. Banerjee, *Quickest Change Detection*. Elsevier: E-reference Signal Processing, 2013, http://arxiv.org/abs/1210.5552.

[2] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2009.

[3] A. G. Tartakovsky, I. V. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Change-Point Detection*, ser. Statistics. CRC Press, 2014.

[4] T. Banerjee, Y. C. Chen, A. D. Dominguez-Garcia, and V. V. Veeravalli, "Power system line outage detection and identification – A quickest change detection approach," in *IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.

[5] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, ser. WSNA '02. New York, NY, USA: ACM, Sep. 2002, pp. 88–97.

[6] J. A. Rice, K. Mechitov, S. Sim, T. Nagayama, S. Jang, R. Kim, B. F. Spencer, G. Agha, and Y. Fujino, "Flexible smart sensor framework for autonomous structural health monitoring," *Smart Structures and Systems*, vol. 6, no. 5-6, pp. 423–438, 2010.

[7] Z. G. Stoumbos, M. R. Reynolds, T. P. Ryan, and W. H. Woodall, "The state of statistical process control as we proceed into the 21st century," *J. Amer. Statist. Assoc.*, vol. 95, no. 451, pp. 992–998, Sep. 2000.

[8] V. Makis, "Multivariate Bayesian control chart," *Operations Research*, vol. 56, no. 2, pp. 487–496, Mar. 2008.

[9] G. Tagaras, "A survey of recent developments in the design of adaptive control charts," *Journal of Quality Technology*, vol. 30, no. 3, pp. 212–231, July 1998.

[10] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Prob and App.*, vol. 8, pp. 22–46, 1963.

[11] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, June 1954.

[12] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, Dec. 1971.

[13] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, no. 1, pp. 206–227, Mar. 1985.

[14] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.

[15] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of quickest change detection," *SIAM Theory of Prob. and App.*, vol. 49, no. 3, pp. 458–497, Sep. 2005.

[16] M. Woodroofe, *Nonlinear Renewal Theory in Sequential Analysis*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1982.

[17] T. Banerjee and V. V. Veeravalli, "Data-efficient quickest change detection with on-off observation control," *Sequential Analysis*, vol. 31, no. 1, pp. 40–77, Feb. 2012.

[18] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, ser. Springer series in statistics. Springer-Verlag, 1985.

[19] M. A. Girshick and H. Rubin, "A Bayes approach to a quality control model," *Ann. Math. Statist.*, vol. 23, no. 1, pp. 114–125, 1952.

[20] K. Premkumar and A. Kumar, "Optimal sleep-wake scheduling for quickest intrusion detection using wireless sensor networks," in *IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2008, pp. 1400–1408.

[21] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, Dec. 1986.

[22] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *Ann. Statist.*, vol. 18, no. 3, pp. 1464–1469, Nov. 1990.

[23] A. G. Tartakovsky, M. Pollak, and A. Polunchenko, "Third-order asymptotic optimality of the generalized Shiryaev-Roberts changepoint detection procedures," *ArXiv e-prints*, May 2010.

[24] S. W. Roberts, "A comparison of some control chart procedures," *Technometrics*, vol. 8, no. 3, pp. 411–430, Aug. 1966.

[25] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Statist.*, vol. 19, no. 3, pp. pp. 326–339, 1948.

[26] T. Banerjee and V. V. Veeravalli, "Data-efficient minimax quickest change detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2012.

[27] A. G. Tartakovsky and A. S. Polunchenko, "Quickest changepoint detection in distributed multisensor systems under unknown parameters," in *Proc. of the 11th IEEE International Conference on Information Fusion*, July 2008.

[28] A. G. Tartakovsky and V. V. Veeravalli, "An efficient sequential procedure for detecting changes in multichannel and distributed systems," in *IEEE International Conference on Information Fusion*, vol. 1, Annapolis, MD, July 2002, pp. 41–48.

[29] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, Apr. 2010.

[30] T. Banerjee and V. V. Veeravalli, "Data-efficient quickest change detection in minimax settings," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6917–6931, Oct. 2013.

[31] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1604 –1614, Mar. 2011.

[32] V. V. Veeravalli, "Decentralized quickest change detection," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1657–1665, May 2001.

[33] Y. Mei, "Information bounds and quickest change detection in decentralized decision systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2669–2681, July 2005.

[34] A. G. Tartakovsky and V. V. Veeravalli, "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Analysis*, vol. 27, no. 4, pp. 441–475, Oct. 2008.

[35] T. Banerjee, V. Kavitha, and V. Sharma, "Energy efficient change detection over a mac using physical layer fusion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 2501–2504.

[36] T. Banerjee, V. Sharma, V. Kavitha, and A. K. Jayaprakasam, "Generalized analysis of a distributed energy efficient algorithm for change detection," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 91–101, Jan. 2011.

[37] Y. Mei, "Quickest detection in censoring sensor networks," in *IEEE International Symposium on Information Theory (ISIT)*, Aug. 2011, pp. 2148–2152.

[38] L. Zacharias and R. Sundaresan, "Decentralized sequential change detection using physical layer fusion," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4999–5008, Dec. 2008.

[39] A. G. Tartakovsky, "Multidecision quickest change-point detection: Previous achievements and open problems," *Sequential Analysis*, vol. 27, no. 2, pp. 201–231, Apr. 2008.