

**EFFICIENT ACTIVE QUICKEST DETECTION FOR STREAMING DATA  
UNDER SAMPLING CONTROL**

A Dissertation  
Presented to  
The Academic Faculty

By

Qunzhi Xu

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2024

© Qunzhi Xu 2024

**EFFICIENT ACTIVE QUICKEST DETECTION FOR STREAMING DATA  
UNDER SAMPLING CONTROL**

Thesis committee:

Dr. Yajun Mei (Advisor)  
H. Milton Stewart School of Industrial and  
Systems Engineering  
*Georgia Institute of Technology*

Dr. Roshan Joseph  
H. Milton Stewart School of Industrial and  
Systems Engineering  
*Georgia Institute of Technology*

Dr. Jianjun Shi (Co-advisor)  
H. Milton Stewart School of Industrial and  
Systems Engineering  
*Georgia Institute of Technology*

Dr. George V. Moustakides  
Department of Electrical and Computer  
Engineering  
*University of Patras*

Dr. Jie Chen  
Department of Population Health Sciences  
*Augusta University*

Dr. Tuo Zhao  
H. Milton Stewart School of Industrial and  
Systems Engineering  
*Georgia Institute of Technology*

Date approved: June 17, 2024

For my mom and dad, Qin and Yongjun

## ACKNOWLEDGMENTS

First and foremost, I would like to express my deep gratitude to my advisor Dr. Yajun Mei for his invaluable guidance and unconditional support during my Ph.D. journey. It is hard to imagine this dissertation without his guidance and support. In five years, he taught me to be an independent researcher, an encouraging mentor, and an engaging teacher. The lessons learned will travel with me wherever life takes me.

I would like to thank my co-advisor Dr. Jianjun Shi for his valuable insights and wise suggestions on my research and my future goals. One important lesson I learnt from him is how to identify and choose good research topics. Discussions with Dr. Shi always broaden my perspectives and provide me with a clearer vision of my future paths.

I would like to thank Dr. Roshan Joseph and Dr. Tuo Zhao from Georgia Tech for teaching me courses during my PhD journey and serving in my dissertation committee. Specifically, I have taken ISyE 6420 (Bayesian Statistics) by Dr. Joseph and ISyE 8803 (Special Topics on OR) by Dr. Zhao, and I learned a lot from their invaluable instructions in the courses. In addition to their teaching, I am particularly grateful to Dr. Zhao for his continuous support in my machine-learning research over these five years. Collaborating with Dr. Zhao on numerous projects has always been a pleasure for me, and his willingness to share his ideas with great patience and his profound insights into new research problems have been a constant source of inspiration.

I would like to thank Dr. Jie Chen from Augusta University and Dr. George V. Moustakides from University of Patras for serving on my committee. Dr. Chen has provided me with numerous opportunities to present my research at workshops and conferences, where I not only won prizes and awards but also gained valuable experience. Dr. Moustakides has been a pivotal figure in my research on change-point detection problems. Working with an expert in this field from the beginning of my research has been incredibly fortunate. His guidance on writing theoretical papers and implementing algorithms efficiently has been

invaluable. Our discussions on the mathematical details of change-point problems have greatly enhanced my understanding and skills.

Special thanks go to Dr. Zemin Zheng and Dr. Qunqiang Feng for their guidance when I was a student at the University of Science and Technology of China. Without them and their support, I would have never considered a life in academia. My academic journey started with them.

Last but not least, I would like to express my most profound appreciation to my mom, Qin, my father, Yongjun, and the rest of my family for their unconditional love, encouragement, and patience. I could not have traveled this path without them.

## TABLE OF CONTENTS

|   |     |
|---|-----|
| <b>Acknowledgments</b> . . . . .  | iv  |
| <b>List of Tables</b> . . . . .   | x   |
| <b>List of Figures</b> . . . . .  | xi  |
| <b>Summary</b> . . . . .  | xii |
| <b>Chapter 1: Introduction</b> . . . . .  | 1   |
| 1.1 Background and Motivation . . . . .   | 1   |
| 1.2 Overview of the Dissertation . . . . .  | 3   |
| 1.2.1 Asymptotic Optimality Theory for Active Quickest Detection with<br>One Affected Stream . . . . .                      | 3   |
| 1.2.2 Quickest Detection in High-Dimensional Linear Models via Im-<br>plicit Regularization . . . . .                       | 3   |
| 1.2.3 Efficient Active Quickest Detection for Multi-Stream Data . . . . .   | 4   |
| <b>Chapter 2: Asymptotic Optimality Theory for Active Quickest Detection with<br/>        One Affected Stream</b> . . . . . | 5   |
| 2.1 Introduction . . . . .  | 5   |
| 2.2 Problem Formulation . . . . .   | 8   |
| 2.2.1 Mathematical Formulation . . . . .  | 8   |

|       |  |           |
|-------|--|-----------|
| 2.2.2 | Review of Change-Point Methods . . . . .   | 10        |
| 2.2.3 | Review of Sampling Policies . . . . .  | 11        |
| 2.3   | Main Results . . . . .   | 13        |
| 2.3.1 | Candidate Sampling/Stopping Strategy . . . . .   | 13        |
| 2.3.2 | Finite-Sample Properties . . . . .   | 16        |
| 2.3.3 | Second-Order Asymptotic Optimality . . . . .   | 17        |
| 2.4   | Remarks . . . . .  | 18        |
| 2.5   | Numerical Results . . . . .  | 21        |
| 2.6   | Conclusion . . . . .   | 23        |
| 2.7   | Proof of the Main Theorem . . . . .  | 23        |
| <br>  |  |           |
|       | <b>Chapter 3: Quickest Detection in High-Dimensional Linear Models via Implicit<br/>Regularization . . . . .</b> | <b>33</b> |
| 3.1   | Introduction . . . . .   | 33        |
| 3.2   | Problem Formulation and Background . . . . .   | 35        |
| 3.2.1 | Mathematical Formulation . . . . .   | 35        |
| 3.2.2 | Review of Signal Recovery Methods in Linear Models . . . . .   | 36        |
| 3.3   | Implicit Regularization-Based Quickest Detection . . . . .   | 38        |
| 3.3.1 | Estimators $\hat{\beta}_t$ . . . . .   | 38        |
| 3.3.2 | Monitoring Statistics $W_t$ . . . . .  | 39        |
| 3.3.3 | Candidate change point $M(t)$ . . . . .  | 40        |
| 3.4   | Theoretical Properties . . . . .   | 40        |
| 3.5   | Numerical Results . . . . .  | 43        |
| 3.6   | Conclusions . . . . .  | 45        |

|   |   |           |
|---|---|-----------|
| 3.7   | Technical Details . . . . .   | 45        |
| 3.7.1   | Theoretical Properties of Implicit Regularization in Linear Model . .   | 45        |
| 3.7.2   | RIP conditions, Random matrix, and Concentration Inequalities . .   | 48        |
| 3.7.3   | Proof of Detection Delay Relationship When Change Occurs at $\nu = 0$   | 49        |
| 3.7.4   | Proof of Detection Delay Relationship When Change Occurs at $\nu = k$   | 60        |
| <b>Chapter 4: Efficient Active Quickest Detection for Multi-Stream Data . . . . .</b> |   | <b>68</b> |
| 4.1   | Asymptotic Optimality Theory for Active Quickest Detection with Un-<br>known Post-Change Parameters . . . . . | 69        |
| 4.1.1   | Problem Formulation and Methods Review . . . . .  | 69        |
| 4.1.2   | Our Proposed Algorithm . . . . .  | 72        |
| 4.1.3   | Numerical Results . . . . .   | 77        |
| 4.1.4   | Proof of the Main Theorem . . . . .   | 78        |
| 4.2   | Active Quickest Detection When Monitoring Multi-streams with Two Af-<br>fected Streams . . . . .              | 91        |
| 4.2.1   | Problem Formulation and Method Review . . . . .   | 91        |
| 4.2.2   | Our Proposed Algorithm . . . . .  | 94        |
| 4.2.3   | Asymptotic Optimality . . . . .   | 97        |
| 4.2.4   | Simulation . . . . .  | 99        |
| 4.2.5   | Technical Details . . . . .   | 101       |
| 4.3   | Impact of Compensation Coefficients on Active Quickest Detection . . . . .                                    | 104       |
| 4.3.1   | Background . . . . .  | 104       |
| 4.3.2   | Our Main Results . . . . .  | 107       |
| 4.3.3   | Numerical Studies . . . . .   | 111       |

|   |                                     |            |
|---|-------------------------------------|------------|
| 4.3.4   | Conclusions . . . . .               | 114        |
| 4.3.5   | Proof of the Main Theorem . . . . . | 115        |
| <b>Chapter 5: Conclusions and Future Research . . . . .</b> |                                     | <b>134</b> |
| 5.1   | Conclusions . . . . .               | 134        |
| 5.2   | Future Research . . . . .           | 135        |
| <b>References . . . . .</b>                                 |                                     | <b>137</b> |
| <b>Vita . . . . .</b>                                       |                                     | <b>144</b> |

## LIST OF TABLES

|     |  |     |
|-----|--|-----|
| 3.1 | Detection Delay of $T, T_{\text{CUSUM}}, \tilde{T}$ for $p = 200$ . . . . .                        | 44  |
| 3.2 | Detection Delay of $T, T_{\text{CUSUM}}, \tilde{T}$ for $p = 400$ . . . . .                        | 44  |
| 4.1 | Comparison of Detection Delay in Normal Distribution . . . . .                                     | 78  |
| 4.2 | Comparison of Detection Delay in Exponential Distribution . . . . .                                | 78  |
| 4.3 | Comparison of Detection Delay of $T = T_0$ or $T_1, T_{\text{SUM}}$ and $T_{\text{MAX}}$ . . . . . | 100 |
| 4.4 | Detection delay for $\Delta = 0, 0.03, 0.07$ when the change occurs to Stream 1 . . . . .          | 113 |

## LIST OF FIGURES

|     |  |     |
|-----|--|-----|
| 2.1 | Average detection delay as a function of average false alarm period for proposed $T_{MSP}$ (blue), full-sampling $T_{Full}$ (red) and optimum CUSUM $T_p$ (black) for $p = 2, 3, 5$ for detecting a change in the mean of a Gaussian from 0 to $\mu > 0$ . Top figures correspond to $\mu = 0.5$ and bottom to $\mu = 1$ . . . | 22  |
| 4.1 | The ARL to false alarm for different $\Delta$ . . . . .  | 112 |
| 4.2 | Detection delay relationship for $\Delta = 0$ (blue) and $\Delta = 0.03$ (red) . . . . .   | 113 |

## SUMMARY

Quickest detection has a wide range of real-world applications in industrial quality control, biosurveillance, network security, etc. Under a general setting, there are  $p$  local streams in a system, and at some unknown time  $\nu$ , an occurring event impacts  $s$  of the available streams by changing the distribution of their samples. In many applications, one often faces the sampling control constraint in the sense of allowing only to sample from  $q$  of the  $p$  local streams at each time instant. We call this “Active Quickest Detection”. The objective of active quickest detection is to decide how to adaptively sample partial data from these  $p$  local streams and how to use the observed partial data to raise a global alarm as quickly as possible once the change occurs subject to both the false alarm and sampling control constraints.

In this dissertation, we present our results in five different directions on active quickest detection. The first one is the simplest scenario when there is only one affected local stream and we are only able to sample from one stream at a time. This is investigated in Chapter 2, where we not only present an efficient algorithm based on myopic or greedy sampling, but also establish a second-order asymptotically optimum property. To the best of our knowledge, while such scenario was investigated as early as 1963 by Shiryaev, this is the first asymptotic optimum result in the literature.

The second direction is to apply random sampling matrices in compressed sensing and implicit regularization in neural networks for active quickest detection. The underlying mathematical model is to detect a sparse change of underlying coefficients in high-dimensional linear regression models - one of the most fundamental predictive models. By over-parametrizing linear regression models with  $p$  coefficients as two-layers neural networks with  $2p$  weights, we develop gradient-descent-based quickest detection algorithms for linear regression models that can recover the sparse post-change with high accuracy, and provide theoretical guarantees on the average run length to false alarm and detection delay.

Our approach can be easily extended to the context of monitoring other advanced predictive machine learning or artificial intelligent algorithms for future research directions.

Three more subtle directions are further presented in Chapter 4 for active quickest detection. In Chapter 4.1, we consider the scenario when the post-change distributions involve unknown parameters. An efficient greedy-cyclic-sampling-based quickest detection algorithm is proposed, and we show that our proposed algorithm is first-order asymptotically optimal in the sense of minimizing the detection delay under both false alarm and sampling control constraints. In Chapter 4.2, we extend our results of  $r = s = 1$  in Chapter 1 to a more complicated scenario when  $r = s = 2$ . We combine three ideas to develop efficient active quickest detection algorithms: (1) win-stay, lose-switch sampling strategy; (2) local CUSUM statistics for local monitoring; and (3) the SUM-Shrinkage technique to fuse local statistics into a global decision. Theoretical guarantees on the ARL to false alarm and detection delay are provided. In Chapter 4.3, we revisited the TRAS algorithm proposed in [1] which incorporates a novel idea of compensation coefficients for unobserved data streams. However, it is unclear how to choose the compensation coefficients suitably from theoretical point of view so as to balance the trade-off between the detection delay and false alarm. In this context, we investigate the impact of compensation coefficients on TRAS algorithm. Our main contributions are two-folded. On the one hand, under the general setting, we prove that if compensation coefficient is larger than  $O(q/(p - q))$ , then the TRAS algorithm is not efficient in the sense of having a larger detection delay than the optimal lower bound. On the other hand, under the special case of  $q = s = 1$ , if compensation coefficient is smaller than  $O(1/(p - 1))$ , then the TRAS algorithm is efficient.

# CHAPTER 1

## INTRODUCTION

### 1.1 Background and Motivation

*Active learning* has received extensive attentions in the modern big data age, partly due to the resource constraints either at the data acquisition level or at the communication level. Often it is required that the decision maker actively selects partial samples from the underlying raw data so as to make right decision with as few attempts as possible. Some examples of active learning include searching and tracking targets [2, 3, 4], active sequential multiple testing [5, 6, 7, 8], etc.

One important active learning problem when monitoring streaming data is active quickest detection problem, which has a wide range of real-world applications in industrial quality control, biosurveillance, network security, etc. Under a general setting, we assume that a system is operated under resource constraints that prevent the simultaneous use of all resources all of the time, and there are  $p$  streams in the system, e.g., different features from principle component analysis, wavelets, or deep neural networks [9, 10, 11], different sensors in sensor management [12], different work stations in an assembly line in manufacturing [13], or different locations in biosurveillance, search and tracking target [14]. The decision maker is responsible to sample observations from these processes in a sequential manner and there are constraints on the number of processes that can be observed. Assume that at some unknown time  $\nu$ , an unusual event (e.g., target, manufacturing defect, or disease outbreak) occurs and yields a subset of anomalous processes in the sense of changing the distributions of the data from these anomalous processes. One is then interested in raising an alarm as quickly as possible once the change occurs (and ideally accurately identifying the subset of anomalous processes) subject to the sampling or resource constraints.

When all streams are passively observed at each time, this is the standard passive quickest detection problem for streaming data. It was first investigated theoretically in [15] on how to combine local monitoring procedures together suitably to control the global false alarm rate when the pre-change and possible post-change distributions of data from all processes are completely specified. Later [16] and [17] extended to more complicated scenarios when the post-change distributions are Gaussian distributions but with unknown postchange means. For more research on passive quickest detection problem, see [18, 19, 20].

When there are sampling or resource constraints, the decision maker is responsible to actively choose which streams to observe, enhancing her or his information about the occurring event or anomalous processes rapidly while accounting for the penalty of wrong declaration. We refer to this as the active quickest detection problem, as compared to the standard passive quickest detection problem. This problem was first investigated by [21] in the context of detecting signals when a radar sensor can only monitor one out of  $p$  directions one at a time. In recent years the topic of active quickest detection receives more attentions, not only from the theoretical viewpoint, see [22, 23, 24], but also from the applied statistics viewpoint, see [25, 26, 27].

This dissertation focuses on making comprehensive progress on methodology, theory, and application of active quickest detection problem io multi-stream data under the sampling or resource constraints. Our specific research aims are to design new algorithms with theoretical guarantees and develop an asymptotic optimality theory to characterize sharp information bound.

## 1.2 Overview of the Dissertation

### 1.2.1 Asymptotic Optimality Theory for Active Quickest Detection with One Affected Stream

The goal of this chapter is to develop an asymptotic optimality theory for active quickest detection problem under a simple scenario when there is only one affected stream. Our main result consists in proving that a simple myopic sampling scheme is second-order asymptotically optimum when the number  $p$  of streams is fixed. The main idea of the proposed detection strategy consists in exploring each local process periodically and decide whether or not a change took place. If we decide positively then we stop and raise a global alarm while if the decision is negative we switch to the next process. To the best of our knowledge, a second-order asymptotically optimum result is proved for the first time for the problem of multi-stream data monitoring under sampling control when time is discrete and the processes are inhomogeneous.

### 1.2.2 Quickest Detection in High-Dimensional Linear Models via Implicit Regularization

The goal of this chapter is to tackle the problem of detecting changes in high-dimensional linear regression models - one of the most fundamental predictive models. Our approach is inspired by recent research on (deep) neural networks, where optimization algorithms such as (stochastic) gradient descent hold *implicit regularization* properties, see [28], [29], etc. By treating linear regression models as two-layers neural networks, we develop efficient implicit regularization-based quickest detection for linear regression models. To the best of our knowledge, this is the first to apply implicit regularization in the text of quickest detection, and our ideas can be easily extended to the context of monitoring other advanced predictive machine learning or artificial intelligent algorithms.

### 1.2.3 Efficient Active Quickest Detection for Multi-Stream Data

The goal of this chapter is to investigate three research topics under the setting of active quickest detection problem in multi-stream data. In the first topic, we consider the scenario when the post-change distributions involve unknown parameters. An efficient greedy-cyclic-sampling-based quickest detection algorithm is proposed, and we show that our proposed algorithm is first-order asymptotically optimal in the sense of minimizing the detection delay under both false alarm and sampling control constraints.

In the second topic, we extend our results in Chapter 2 to a more complicated scenario when  $r = s = 2$ . We combine three ideas to develop efficient active quickest detection algorithms: (1) win-stay, lose-switch sampling strategy; (2) local CUSUM statistics for local monitoring; and (3) the SUM-Shrinkage technique to fuse local statistics into a global decision. Theoretical guarantees on the ARL to false alarm and detection delay are provided.

In the third topic, we revisited the TRAS algorithm proposed in [1] which incorporates a novel idea of compensation coefficients for unobserved data streams. However, it is unclear how to choose the compensation coefficients suitably from theoretical point of view so as to balance the trade-off between the detection delay and false alarm. In this context, we investigate the impact of compensation coefficients on TRAS algorithm. Our main contributions are two-folded. On the one hand, under the general setting, we prove that if compensation coefficient is larger than  $O(q/(p - q))$ , then the TRAS algorithm is not efficient in the sense of having a larger detection delay than the optimal lower bound. On the other hand, under the special case of  $q = s = 1$ , if compensation coefficient is smaller than  $O(1/(p - 1))$ , then the TRAS algorithm is efficient.

**CHAPTER 2**  
**ASYMPTOTIC OPTIMALITY THEORY FOR ACTIVE QUICKEST**  
**DETECTION WITH ONE AFFECTED STREAM**

This chapter is a collaborative work with Dr. Yajun Mei and Dr. George V. Moustakides, and part of the work was published in the journal IEEE Transactions on Information Theory 2021.

**2.1 Introduction**

Quickest detection for multi-stream data under sampling control has many important real-world applications such as quality control, surveillance or security, etc. Under a general setting, there are  $p$  processes or data streams available in a system, and at some unknown point in time, an occurring event impacts one of the available processes by changing the distribution of its samples. Unlike the conventional problem where one samples simultaneously all streams, under a sampling constraint scenario we are allowed to sample only one of the  $p$  local streams at each time. This constraint may be imposed due to sampling costs or limitations to the on-line processing power. We understand that in addition to the usual problem of developing a *stopping* strategy for signaling the detection of the change we must also provide a *sampling strategy* for the sampling of the available streams.

Under the classical passive learning set-up when we can sample all streams simultaneously without any constraint, much research has been done for the dimensional  $p = 1$ , see books [30, 31, 32, 33, 34, 35] for a review. When  $p$  is large, it was first investigated theoretically in [15] on how to combine local monitoring procedures together suitably to control the global false alarm rate when the pre-change and possible post-change distributions of data from all streams are completely specified. Later [16] and [17] extended to more complicated scenarios when the post-change distributions are Gaussian distributions

but with unknown postchange means.

When there is sampling control as to which process must be sampled at each time instant, literature is rather limited. The only existing result is [1] which proposes an algorithm that performs reasonably well in simulations but has no theoretical justification to support it. Related work is also [36] which proposes a data-efficient sampling technique but applies to a single data stream.

We should mention that sampling control has also been extensively used in two other well-known problems: sequential hypothesis testing and the multi-armed bandit problems. Pioneering work in sequential hypothesis testing for  $p = 2$  processes, when the data are Bernoulli distributed, can be found in [37]. Other works on sequential hypothesis testing with sampling control include [38, 39, 40, 41, 42, 43] and more recently [8, 7, 44, 45]. There has also been progress on the “multi-stream” case of sequential hypothesis testing with sampling control, where there are multiple data streams, some of which are anomalous, and the goal is to accurately pick out the anomalous ones among them, while observing only a subset of the streams at each time-step, see [46, 47, 48, 49, 50]. Sampling control is also investigated in the multi-armed bandit problem (MAB), which was originally introduced by [51] that addresses the dilemma between “exploration” and “exploitation”. Later, [52], and [53] derived ground-breaking fundamental results on the classical MAB. In recent years, computer scientists, particularly those in reinforcement learning or machine learning, follow this approach to rejuvenate MAB, and compare the regret of their proposed sampling strategies or algorithms with the information bounds of Lai and Robbins with some (possibly non-trivial) twists, see [54, 55, 56, 57, 58].

Sequential hypothesis testing for the *homogeneous* case and for general  $p$  was considered in [59] but with the special requirement of identifying only a single process under the alternative hypothesis. The same switching and stopping strategy we propose in our current work, is optimum for this problem as well and, remarkably, the optimality turns out to be exact. Later in [50] we find an asymptotic optimality theory for sequential hypothesis

testing problems when one is allowed to sample one out of  $p$  streams at each time. Finally, important results for the multi-armed bandit problem are offered in [52] where the first asymptotic optimality theory is developed.

The problem we are attempting to solve in this work has two major differences compared to sequential hypothesis testing and the multi-armed bandit problem: a) data collected from unaffected process provide no useful information for the affected stream, and b) data collected from the process to be affected before the change also provide no information about detection.

In this work we focus on an asymptotic optimality theory for sequential change-point detection for multi-stream data under sampling control. Our main result consists in proving that a simple myopic sampling scheme is second-order asymptotically optimum when the number  $M$  of processes is fixed. The main idea of the proposed detection strategy consists in exploring each local process periodically and decide whether or not a change took place. If we decide positively then we stop and raise a global alarm while if the decision is negative we switch to the next process. To the best of our knowledge, a second-order asymptotically optimum result is proved for the first time for the problem of multi-stream data monitoring under sampling control when time is discrete and the processes are inhomogeneous.

As one of our reviewers correctly pointed out, our proposed myopic sampling strategy is similar to the “cyclic-return system of observations over  $p$  directions” scheme proposed in [21] (see also [60]). We should emphasize that the mathematical formulation, models and technical details in [21] exhibit significant differences compared to the ones adopted here. Indeed, in [21] it is assumed that the change occurs with the same probability in any of the  $M$  local streams or directions (namely, there exists a prior). In our study we adopt a worst-case scenario which is consistent with Lorden’s single stream min-max approach. In [21] the analysis is focused on continuous-time and continuous path (Wiener) processes with all local streams or directions being homogeneous. In our work we adopt discrete-time and inhomogeneous processes across different local streams. These essential dissimilarities

require the employment of alternative analytical tools in order to be able to tackle problems that are not present in Wiener processes as, for example, overshoots. We must mention that the absence of overshoots in the continuous-time and continuous-path case allows for the derivation of fairly accurate expressions for the average detection delay and average time between false alarms, while in discrete time the analysis can only provide bounds of the appropriate order of magnitude. What is also interesting is that the proposed scheme in [21] turns out to be a repeated CUSUM test even though CUSUM, at the time, was not yet known for its optimality properties. Finally, we should add that we believe that under a more complicated setup where one is allowed to observe more than one local streams simultaneously, it is unclear how the methodology in [21] can be extended. On the other hand, our approach, as we discuss in Section IV, does not seem to have this problem.

The remainder of our paper is organized as follows. In section 2.2, we mathematically formulate the problem of interest and review existing methods. In section 2.3, we present our candidate sampling/ stopping strategy and prove the second-order asymptotic optimality property of our scheme when  $M$  is fixed. In section 2.4 we provide certain remarks concerning possible extensions to more complicated scenarios and discuss the corresponding challenges. Numerical simulations are offered in section 2.5 to illustrate the agreement between theory and practice. Finally, in section 2.6 we draw our conclusions and discuss future research topics. Technical proofs appear in section 2.7.

## **2.2 Problem Formulation**

### 2.2.1 Mathematical Formulation

Suppose there are  $p$  statistically independent local streams in a system, and denote with  $X_t^i$  the observation from the  $i$ -th stream at time  $t$ , where  $i = 1, \dots, p$  and  $t = 1, 2, \dots$ . Initially, the system is in the in-control state and  $\{X_t^i\}$  from the  $i$ th stream produces i.i.d. samples following the density  $f_i(X)$ . At some unknown time  $\tau$ , an event occurs which leads the system out-of-control with one of its  $p$  processes, say, the  $i$ -th, changing to i.i.d. samples

following a new density  $g_i(X)$ . Specifically, if the  $i$ th data stream is affected, then

$$X_t^i \sim \begin{cases} f_i(X), & \text{if } t \leq \tau \\ g_i(X), & \text{if } t > \tau, \end{cases} \quad (2.1)$$

while  $X_t^j \sim f_j(X)$  for  $j \neq i$  and all  $t > 0$ .

Under the classical setup when the full data information is available at each time  $t$  we observe the complete set of  $p$  samples  $\{X_t^1, \dots, X_t^p\}$ . However, when we adopt a sampling control policy then we are allowed to access only *one* of these  $p$  data points. This clearly requires the definitions of a sequence of *sampling indices*  $\{R_t\}$  with  $R_t \in \{1, \dots, p\}$ .  $R_t$  is random and points to the local stream that must be sampled during the *next* time instant  $t + 1$ . The sampling constraint can be expressed as

$$\mathbb{1}_{\{R_t=1\}} + \dots + \mathbb{1}_{\{R_t=p\}} = 1, \quad (2.2)$$

for all times  $t = 1, 2, \dots$ , where  $\mathbb{1}_A$  denotes the indicator function of the event  $A$ .

As we discuss next, due to the existence of various possibilities, we need to introduce several sequences of sigma-algebras (filtrations). With  $\{\mathcal{F}_t^i\}$  we denote the filtration generated by the  $i$ th stream, namely,  $\mathcal{F}_t^i = \sigma\{X_1^i, \dots, X_t^i\}$ . Then we define the filtration  $\{\mathcal{F}_t\}$  containing the complete information where  $\mathcal{F}_t = \mathcal{F}_t^1 \cup \dots \cup \mathcal{F}_t^p$ . Finally, by sampling one out of  $p$  local streams at each time instant we generate the filtration  $\{\mathcal{F}_t^R\}$  with  $\mathcal{F}_t^R = \sigma\{X_1^{R_0}, \dots, X_t^{R_{t-1}}\}$ . A sequential change-point detection procedure under sampling control contains two components, the sampling policy  $\{R_t\}$  and the stopping time  $T$ . For the sampling policy each  $R_t$  is  $\mathcal{F}_t^R$ -measurable (we use the already available samples up to time  $t$  to decide which stream to sample at the next time instant  $t + 1$ ). The stopping time  $T$  is  $\{\mathcal{F}_t^R\}$ -adapted (uses all samples up to and including time  $t$  in order to decide whether to stop and raise an alarm at  $t$  or continue sampling according to  $R_t$ ).

Denote by  $P_t^i(\cdot), E_t^i[\cdot]$  the probability measure and the corresponding expectation in-

duced by the change occurring at Process  $i$  at time  $\tau = t$ , and denote by  $P_\infty(\cdot), E_\infty[\cdot]$  the probability measure and the corresponding expectation under the nominal probability measure (namely when the change occurs at  $\infty$ ). Following the min-max approach suggested by [61], our aim is to minimize

$$D_i(T) = \sup_{t \geq 0} \text{ess sup } E_t^i[T - t | \mathcal{F}_t^i, T > t], \quad (2.3)$$

for all affected data stream  $i$ , subject to  $E_\infty[T] \geq \gamma > 1$ , where  $T$  and  $\{R_t\}$  are defined over the appropriate filtration. As we can see the false alarm constraint requires the average false alarms period to be no less than some prescribed value  $\gamma > 1$ .

For the constrained problem in (Equation 2.3) it is very unrealistic to expect that we can find a *single* sampling/stopping policy capable of *exactly* optimizing it, simultaneously for all  $i = 1, \dots, p$ . The goal of this work is to show that in fact such a possibility exists (by proposing a particular solution) but within the class of policies that are *second-order asymptotically optimum*.

### 2.2.2 Review of Change-Point Methods

Focusing on solving (Equation 2.3) with  $D_i(T)$  defined in (Equation 2.3) we first consider the case where there is a genie that provides the index  $i$  of the stream where the change occurs. If we know  $i$  then there is no reason to sample any other streams, consequently  $R_t = i$  at all times and we can limit  $T$  to be  $\{\mathcal{F}_t^i\}$ -adapted. In this case it is well known that the optimum stopping time is the CUSUM defined as

$$T_i(A_i) = \inf\{t > 0 : W_t^i \geq A_i\}, \quad (2.4)$$

where  $W_t^i$  is the CUSUM statistic [62] that satisfies the recursion

$$W_t^i = \max\{W_{t-1}^i, 0\} + \log \frac{g_i(X_t^i)}{f_i(X_t^i)}, \quad (2.5)$$

for  $t > 0$  and is initialized with  $W_0^i = 0$ . Threshold  $A_i$  is selected so that the false alarm constraint is met with equality. Proof for first-order asymptotic optimality was offered in [61] while exact optimality can be found in [63]. It is clear that there is no detection strategy that can outperform the CUSUM test that knows where (but not when) the change occurs.

A more practically interesting scenario consists in having access to the complete data set but not knowing where and when the change occurs. Again, there is no need to specify a sampling strategy  $\{R_t\}$  since we sample all local streams. Clearly  $T$  is now  $\{\mathcal{F}_t\}$ -adapted, namely the stopping time uses all the information up to time  $t$  to decide whether to stop at  $t$  or not. In this case we run local CUSUMs in parallel, one for each local stream and as it is proven in [35] we raise an alarm whenever one of the  $p$  stopping times stops. More specifically if  $T_i$  is defined as in (Equation 2.4) then the combination

$$T_{\text{Full}} = \min\{T_1(A), \dots, T_p(A)\}, \quad (2.6)$$

with all CUSUM having the same threshold  $A$ , is asymptotically optimum in the sense that it solves the problem defined in (Equation 2.3) asymptotically as  $\gamma \rightarrow \infty$ . In fact we can rewrite  $T_{\text{Full}}(A)$  as

$$T_{\text{Full}}(A) = \inf\{t > 0 : \max_{1 \leq i \leq p} W_t^i \geq A\}, \quad (2.7)$$

where, essentially, in the test we apply the generalized likelihood ratio with respect to index  $i$ . In Theorems 9.2.1 and 9.2.2 of [35], it is shown that  $T_{\text{Full}}(A)$  is second-order asymptotically optimal when the number  $p$  of local streams is fixed. This means that we minimize each detection delay  $D_i(T)$  up to an  $O(1)$  quantity which is independent from the false alarm constraint parameter  $\gamma$  in (Equation 2.3).

### 2.2.3 Review of Sampling Policies

Under sampling control there is the need to define a sampling policy  $\{R_t\}$  since, as we explained, at each time instant we are allowed to sample only one out of the  $p$  local streams.

Consequently let us review sampling possibilities that are widely adopted in the literature.

The most straightforward policy consists in sampling each local stream periodically meaning that  $R_t = t \bmod p + 1$  where we visit each local stream deterministically every  $p$  samples. It is not difficult to show that this sampling strategy will lead to a detection delay which is  $p$  times larger than the optimum which, as mentioned, is enjoyed by the CUSUM stopping time  $T_i$  that knows where the change occurs. Clearly, this observation makes unrealistic any expectation for establishing second-order (in fact even first-order) asymptotic optimality with this form of sampling.

An alternative widely used policy is the myopic (or greedy) sampling policy (MSP) which samples the local stream exhibiting the maximal immediate reward. A general implementation of the myopic sampling policy is to define local statistics  $\widetilde{W}_t^i, i = 1, \dots, p$  that summarize the immediate sampling rewards for each process at time  $t$  and then sample the process with the largest local statistic  $\widetilde{W}_t^i$  (with random sampling or pre-assigned order in case of ties). This type of sampling is frequent in the multi-armed bandit problem and there is extensive literature as to which is the most suitable selection for the local statistic  $\widetilde{W}_t^i$ .

Our intention is to use the myopic sampling policy for the problem of interest. It is clear that a natural candidate for the local statistic  $\widetilde{W}_t^i$  is the CUSUM statistic defined in (Equation 2.5). However, in order for our sampling/stopping policy to be complete, we need to explicitly specify three points that are unclear: (i) How should we update the local statistic of a process not being sampled. (ii) How should we break ties when the largest local statistic occurs at multiple processes. This is particularly important when all local CUSUM statistics become simultaneously 0 (which is a very frequent event under the nominal regime). (iii) When should we raise a global alarm. While it is possible to give intuitively meaningful answers to these three points, the challenge is to accompany them with theoretical justification capable of establishing the desired form of asymptotic optimality.

## 2.3 Main Results

In this section we introduce our candidate sampling/stopping strategy and establish its asymptotic optimality characteristics. Again we break our presentation into parts: in the first part we define our candidate stopping time  $T_{\text{MSP}}$  implemented with the help of the myopic sampling policy. We also provide the answers to the three questions in the previous section. In the second part, we study the non-asymptotic properties of the false-alarm and detection delay of  $T_{\text{MSP}}$ . We conclude the presentation of our main results by establishing the second-order asymptotic optimality of our candidate test in the third part.

### 2.3.1 Candidate Sampling/Stopping Strategy

At a high level, our algorithm is based on the myopic sampling policy and mimics the full-sampling method  $T_{\text{Full}}(A)$  in (Equation 2.7) under the sampling control constraint. Here we exploit the prior knowledge that there is only one local stream which changes, and thus propose to sample each local stream until we are confident to decide whether a change has occurred or not. If we detect a change, then we stop and raise a global alarm. If we decide there is no change, then we deterministically switch to the next local stream to sample and repeat the previous step. Switching follows a periodic pattern starting from the first local stream and going to the  $p$ -th local stream. When we reach the  $p$ -th local stream and decide to switch again we simply restart from the first local stream. We repeat these steps until we raise an alarm.

Let us now define our scheme rigorously through the recursive definition of the statistics  $\widetilde{W}_t^i$ ,  $i = 1, \dots, p$  and the sampling sequence  $\{R_t\}$ . At time  $t - 1$  assume we already have available  $\widetilde{W}_{t-1}^i, i = 1, \dots, p$  and  $R_{t-1}$  with the latter pointing to the local stream to be

sampled at time  $t$ . For the local statistics we have the recursions

$$\begin{aligned}\widetilde{W}_t^i &= \max\{\widetilde{W}_{t-1}^i, 0\} + \mathbb{1}_{\{i=R_{t-1}\}} \log \frac{g_i(X_t^i)}{f_i(X_t^i)} \\ &= \begin{cases} \max\{\widetilde{W}_{t-1}^i, 0\}, & \text{if } i \neq R_{t-1} \\ \max\{\widetilde{W}_{t-1}^i, 0\} + \log \frac{g_i(X_t^i)}{f_i(X_t^i)}, & \text{if } i = R_{t-1}, \end{cases}\end{aligned}\quad (2.8)$$

and for the sampling sequence

$$R_t = \begin{cases} R_{t-1} & \text{if } \widetilde{W}_t^{R_{t-1}} > 0 \\ R_{t-1} \bmod p + 1 & \text{if } \widetilde{W}_t^{R_{t-1}} \leq 0. \end{cases}\quad (2.9)$$

For all  $i = 1, \dots, p$  we initialize with  $\widetilde{W}_0^i = 0$  and  $R_0 = 1$ . Similarly to the full-sampling method  $\mathsf{T}_{\text{Full}}(A)$  in (Equation 2.7), we propose the stopping time  $\mathsf{T}_{\text{MSP}}$  defined as

$$\mathsf{T}_{\text{MSP}}(A) = \inf \left\{ t > 0 : \max_{1 \leq i \leq p} \widetilde{W}_t^i \geq A \right\}, \quad (2.10)$$

where threshold  $A$  is selected to meet the false alarm constraint.

We observe that we continue sampling the  $i$ -th local stream as long as  $\widetilde{W}_t^i > 0$  and switch to the next stream when  $\widetilde{W}_t^i \leq 0$ . It is easy to see that we can equivalently define  $\mathsf{T}_{\text{MSP}}$  with a single test statistic that satisfies the update

$$\widetilde{W}_t = \max\{\widetilde{W}_{t-1}, 0\} + \log \frac{g_{R_{t-1}}(X_t^{R_{t-1}})}{f_{R_{t-1}}(X_t^{R_{t-1}})}, \quad (2.11)$$

with  $\widetilde{W}_0 = 0$  and the sampling policy

$$R_t = \begin{cases} R_{t-1} & \text{if } \widetilde{W}_t > 0 \\ R_{t-1} \bmod p + 1 & \text{if } \widetilde{W}_t \leq 0, \end{cases}\quad (2.12)$$

with  $R_0 = 1$ . The stopping time  $T_{\text{MSP}}$  can then be equivalently written as

$$T_{\text{MSP}}(A) = \inf \left\{ t > 0 : \widetilde{W}_t \geq A \right\}. \quad (2.13)$$

The recursion in (Equation 2.11) and the definition of our stopping time in (Equation 2.13) are clearly very CUSUM-like, the only difference being that instead of always sampling the same process which is the practice in the regular CUSUM, every time the test statistic  $\widetilde{W}_t$  falls below 0, we switch to testing the next process by restarting and forgetting the whole past. Let us summarize the proposed scheme.

**Step 1:** Sample the first local stream until  $\widetilde{W}_t \notin (0, A)$ . If  $\widetilde{W}_t \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t \leq 0$ , we switch to sampling the second local stream.

**Step 2:** Sample the second local stream until  $\widetilde{W}_t \notin (0, A)$ . If  $\widetilde{W}_t \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t \leq 0$ , we switch to sampling the third local stream.

⋮

**Step  $p$ :** Sample the  $p$ -th local stream until  $\widetilde{W}_t \notin (0, A)$ . If  $\widetilde{W}_t \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t \leq 0$ , we switch to sampling the first local stream

**Step  $p + 1$ :** Go back to Step 1.

The reason we expect that  $T_{\text{MSP}}$  defined in (Equation 2.10) will enjoy second-order optimality properties is because when we start sampling a local stream we practically apply a sequential probability ratio test (SPRT) with the lower threshold set to 0. We recall that the classical CUSUM is also a *repeated* SPRT test with lower threshold equal to 0 only, as mentioned, it is always applied onto the *same* local stream. Here, what we propose is that every time we restart the SPRT we switch to the next stream. Streams that do not change or the stream which will change but is still under the pre-change state drive the SPRT to 0 very quickly with short random periods. When we hit post-change data then with high

probability the corresponding SPRT will remain at this local stream and drive its statistic towards the high threshold to raise an alarm.

### 2.3.2 Finite-Sample Properties

To establish the desired asymptotic optimality characteristic for  $T_{\text{MSP}}$  we first need to introduce certain finite-sample properties. We start by making some standard assumptions encountered in the classical sequential detection literature. For  $i = 1, \dots, p$  we have

$$\begin{aligned} \text{(A1): } I_{\infty}^i &= \int \log \frac{f_i(X)}{g_i(X)} f_i(X) dX > 0, \\ I_0^i &= \int \log \frac{g_i(X)}{f_i(X)} g_i(X) dX > 0, \\ \text{(A2): } J_{\infty}^i &= \int \left( \log \frac{f_i(X)}{g_i(X)} \right)^2 f_i(X) dX < \infty, \\ J_0^i &= \int \left( \log \frac{g_i(X)}{f_i(X)} \right)^2 g_i(X) dX < \infty, \end{aligned}$$

In other words we define the information numbers and make the Assumption (A1) that they are bounded away from 0 meaning that pre- and post-change densities must be essentially different. Assumption (A2) is technical and states that the second moments of the log-likelihood ratios are bounded away from  $\infty$ . Clearly (A2) implies that the information numbers are also bounded away from  $\infty$ .

To establish the second-order asymptotic optimality, we need to compare the performance of our proposed stopping time  $T_{\text{MSP}}$  against the *optimum performance* delivered by the CUSUM stopping time  $T_i(A_i)$ . Since there are no exact formulas for both schemes we present useful estimates that will allow us to achieve our goal. We start with the CUSUM test for which the next lemma provides the required estimates most of which are already established in the literature.

**LEMMA 1** *Under Assumptions (A1), (A2), the CUSUM tests  $T_i(A)$ ,  $i = 1, \dots, p$ , satisfy*

the following bounds for the average period of false alarms

$$e^A \leq \mathbb{E}_\infty[\mathbb{T}_i(A)] \leq \mathcal{C}_i e^A, \quad (2.14)$$

while for the worst-case average detection delays

$$\frac{A}{l_0^i} - \mathcal{L}_i \leq D_i(\mathbb{T}_i(A)) \leq \frac{A}{l_0^i} + \mathcal{U}_i. \quad (2.15)$$

The quantities  $\mathcal{C}_i, \mathcal{U}_i, \mathcal{L}_i$  are positive constants that depend only on  $f_i, g_i$  and not on  $A$ .

The next theorem provides corresponding estimates for the proposed stopping time  $\mathbb{T}_{\text{MSP}}$ .

**THEOREM 1** *Under Assumptions (A1), (A2) the proposed stopping time  $\mathbb{T}_{\text{MSP}}(A)$  defined in (Equation 2.10) satisfies the following lower bound for the average false alarm period*

$$e^A \leq \mathbb{E}_\infty[\mathbb{T}_{\text{MSP}}], \quad (2.16)$$

while for the worst-case average detection delays for  $i = 1, \dots, p$ , satisfy the upper bounds

$$D_i(\mathbb{T}_{\text{MSP}}) \leq \frac{A}{l_0^i} + \mathcal{U}_i + \mathcal{D}(p-1), \quad (2.17)$$

where  $\mathcal{U}_i$  are the constants from (Equation 2.15) in Lemma Lemma 1 and  $\mathcal{D}$  is a constant that depends on all  $\{f_i, g_i\}$ ,  $i = 1, \dots, p$  but not on  $A$ .

### 2.3.3 Second-Order Asymptotic Optimality

Using Lemma 1 and Theorem 1 we are now able to establish the second-order asymptotic optimality property for  $\mathbb{T}_{\text{MSP}}$ . The optimum detection delay grows to infinity as the false alarm parameter  $\gamma \rightarrow \infty$ . Our intention is to show that  $\mathbb{T}_{\text{MSP}}$  has a detection delay which, for each  $i$ , grows to infinity at the same rate as the optimum CUSUM test that knows where

the change occurs. More specifically we will show that the two performances can differ at most by a bounded constant which does not depend on  $\gamma$ .

If we know that the change is going to occur at the  $i$ -th local stream then the best detection delay performance, as we mentioned, is delivered by the CUSUM stopping time  $T_i(A_i)$  with  $A_i$  selected so that the false alarm constraint is satisfied with equality. In fact there is absolutely no other stopping time that can enjoy better performance since  $T_i(A_i)$  uses information which is absolutely relevant to the corresponding detection task while any other stopping time defined on a different filtration will use information that is not related to the change at the  $i$ -th local stream. We have the following corollary that establishes the second-order optimality of  $T_{\text{MSP}}$  *simultaneously* for all  $i = 1, \dots, p$ .

**COROLLARY 1** *Let  $A = \log \gamma$ , then our proposed stopping time  $T_{\text{MSP}}(A)$  defined in (Equation 2.10) satisfies both the false alarm and the sampling control constraint. If for each  $i = 1, \dots, p$ , we have the optimum CUSUM tests  $T_i(A_i)$  with  $A_i$  selected to satisfy the false alarm constraint with equality, then*

$$0 \leq D_i(T_{\text{MSP}}(A)) - D_i(T_i(A_i)) \leq Cp, \quad (2.18)$$

for proper constant  $C$ .

## 2.4 Remarks

Let us now discuss possible extensions and corresponding challenges we may encounter.

1. Our theoretical results still hold if one uses the slightly different form of the CUSUM statistic  $\hat{W}_t = \max\{\tilde{W}_t, 0\}$ . We have the following update of the test statistic for this version

$$\hat{W}_t = \max \left\{ \hat{W}_{t-1} + \log \frac{g_{R_{t-1}}(X_t^{R_{t-1}})}{f_{R_{t-1}}(X_t^{R_{t-1}})}, 0 \right\} \quad (2.19)$$

while the sampling policy becomes

$$R_t = \begin{cases} R_{t-1} & \text{if } \hat{W}_t > 0 \\ R_{t-1} \bmod M + 1 & \text{if } \hat{W}_t = 0, \end{cases} \quad (2.20)$$

As we can see by comparing (Equation 2.11) to (Equation 2.19) in this version the attractive linearity property no longer holds for  $\hat{W}_t$ .

2. The second-order asymptotic optimality property is assured under the assumption that the number  $M$  of processes is fixed. Since our finite-sampling estimates hold for any given  $M$  we can easily deduce that our proposed scheme can still enjoy *first-order* asymptotically optimality when  $M \rightarrow \infty$  provided that the constants entering in the estimates of the CUSUMs and the proposed  $T_{\text{MSP}}$  are uniformly bounded in  $M$  (for example when all processes follow the same pre- and post-change density) while the number of processes grows to  $\infty$  as  $M = o(\log \gamma)$ .
3. A possible extension is to allow sampling of more than one processes at each time instant. This can be particularly advantageous when the number  $M$  of processes is large or when more than one processes may be affected (change) simultaneously. The sampling control in (Equation 2.2) can therefore be relaxed to

$$\mathbb{1}_{\{R_t=1\}} + \cdots + \mathbb{1}_{\{R_t=M\}} = Q, \quad (2.21)$$

for all times  $t = 1, 2, \dots$  and some integer  $Q \in \{1, \dots, M\}$ . In particular, when  $Q = M$ , this corresponds to the full-sampling scenario. We can now extend the myopic sampling policy to this more general case by sampling the  $Q$  processes corresponding to the  $Q$  largest  $\widetilde{W}_t^i$  values. Unfortunately, analyzing this more general sampling scheme is not as simple as the case  $Q = 1$  we already examined. For the analysis we note that we have now  $Q$  SPRTs running in parallel which, unfortunately, are not

synchronized. It is in fact this last observation that makes the analysis challenging.

4. The previous general case enjoys a considerable simplification when we can have a change in only one process. A possible solution strategy consists in adopting a block sampling policy that satisfies the sampling control requirement (Equation 2.2) as follows: We divide the  $M$  processes into  $Q$  blocks, where each block contains roughly  $\lceil \frac{M}{Q} \rceil$  local processes. In each block, we sample cyclically following the procedure proposed for the case  $Q = 1$  and we update the statistic of each process as in (Equation 2.11). We raise a global alarm when any statistic crosses the upper threshold.

Regarding the analysis, denote by  $\tilde{T}_{\text{MSP}}$  the corresponding block-sampling-based scheme subject to the sampling control in (Equation 2.21). Then it can be shown that

$$\begin{aligned} \frac{e^A}{Q} &\leq \mathbb{E}_\infty[\tilde{T}_{\text{MSP}}] \\ D_i(\tilde{T}_{\text{MSP}}) &\leq \frac{A}{l_0^i} + \mathcal{U}_i + \mathcal{D} \frac{M}{Q}. \end{aligned} \tag{2.22}$$

Quantity  $\mathcal{D}$  is the constant already introduced in (Equation 2.17) of Theorem Theorem 1 and it is related to all  $\{f_i, g_i\}$ ,  $i = 1, \dots, M$ . To satisfy the false-alarm constraint in (Equation 2.3) we select  $A = \log \gamma + \log Q$  and this yields

$$D_i(\tilde{T}_{\text{MSP}}) \leq \frac{\log \gamma}{l_0^i} + \frac{\log Q}{l_0^i} + \mathcal{U}_i + \mathcal{D} \frac{M}{Q}, \tag{2.23}$$

which can ensure second-order asymptotic optimality when  $M$  is fixed. In the case where we allow  $M, Q \rightarrow \infty$  it is possible to enjoy first-order asymptotic optimality if the following two rates are satisfied:  $\log Q = o(\log \gamma)$  and  $\frac{M}{Q} = o(\log \gamma)$ . The latter is a clear improvement over the rate  $M = o(\log \gamma)$  required when  $Q = 1$ .

5. Under a high-dimensional setting where  $M \gg \log \gamma$  if multiple processes undergo a change, myopic sampling policy might turn out to be overly greedy and it might

be necessary to encourage the exploration of processes that have not been sampled. A possible means to achieve this (see [1]) is to introduce a compensation coefficient  $\Delta \geq 0$  to the local processes that are not being sampled. In other words modify the update in (Equation 2.8) as follows

$$\widetilde{W}_t^i = \begin{cases} \max\{\widetilde{W}_{t-1}^i, 0\} + \Delta, & \text{if } i \neq R_{t-1} \\ \max\{\widetilde{W}_{t-1}^i, 0\} + \log \frac{g_i(X_t^i)}{f_i(X_t^i)}, & \text{if } i = R_{t-1}, \end{cases} \quad (2.24)$$

which increases the chance of the myopic sampling policy to select unobserved processes. It is shown in [1] numerically that a suitable choice of  $\Delta > 0$  can significantly improve performance, but it is still an open problem the rigorous theoretical analysis of this sampling scheme.

## 2.5 Numerical Results

In this section, we conduct Monte Carlo simulation studies to corroborate our theoretical results. Assume  $f_i = f \sim N(0, 1)$ , and  $g_i = g \sim N(\mu, 1)$ . We consider  $\mu = 0.5, 1$  and  $p = 2, 3, 5$  and simulate all six combinations of  $\mu$  and  $p$  values. In each of these cases, we report the detection delay of our proposed stopping time  $T_{\text{MSP}}$  under the sampling control  $Q = 1$  (blue) and we compare it against the optimum CUSUM procedure  $T_i$  which knows where the change occurs (black) and  $T_{\text{Full}}$  which has access to the full data set but does not know where the change occurs (red).

All competing schemes have either a CUSUM or a CUSUM-like update consequently we can safely claim that the worst case scenario for the detection delay is when the change occurs at  $\tau = 0$ . Furthermore, regarding our scheme, we consider the least favorable scenario of the change occurring in the  $p$ th stream which is the last to be sampled by our method. This clearly adds an extra initial delay until our test rejects the first  $p - 1$  local streams and starts sampling the correct stream. We compare the detection delay of the three tests as a function of the average false alarm period. This is achieved by performing

100,000 Monte Carlo independent runs for different values of the threshold  $A$ .

From Fig. Figure 2.1, it is clear that the gap between the three curves remains bounded for all values of  $\gamma$  (actually the curves become parallel suggesting that the gap tends to a constant). This is consistent with our theoretical result that the proposed scheme is second-order asymptotically optimum when  $p$  is fixed. We also observe that the gap between the blue and the black curve increases with  $p$  which is again consistent with the upper bound recorded in (Equation 2.18) in Corollary 1. A similar phenomenon occurs for  $T_{\text{Full}}$  but, as we can see, it is far less pronounced which is of course expected since for this test we sample all processes simultaneously and therefore there is no initial delay to start sampling the affected process.

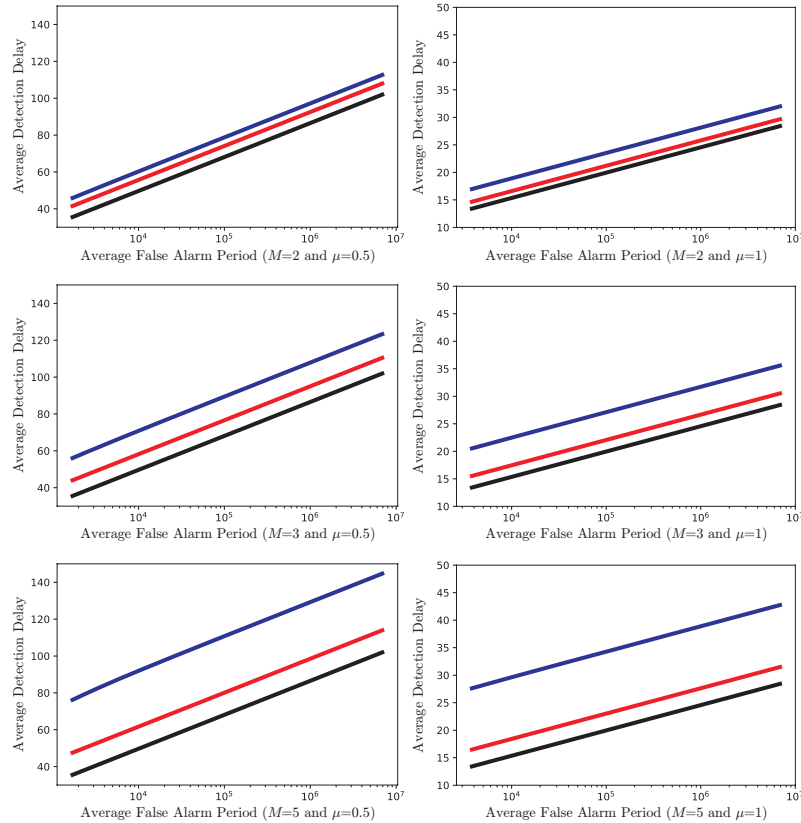


Figure 2.1: Average detection delay as a function of average false alarm period for proposed  $T_{\text{MSP}}$  (blue), full-sampling  $T_{\text{Full}}$  (red) and optimum CUSUM  $T_p$  (black) for  $p = 2, 3, 5$  for detecting a change in the mean of a Gaussian from 0 to  $\mu > 0$ . Top figures correspond to  $\mu = 0.5$  and bottom to  $\mu = 1$ .

## 2.6 Conclusion

Sequential change-point detection with sampling control is an important and challenging topic with many applications. In this work we developed a detection strategy based on the simple idea of myopic sampling. Interestingly with this simple sampling scheme we are able to establish second-order asymptotic optimality when the number  $M$  of processes is fixed and first-order when  $M$  increases to infinity but at a proper rate as compared to the false alarm parameter. Future work may include extension to the case where we have multiple processes that simultaneously undergo a change and we are allowed to sample more than one processes simultaneously at each time instant.

## 2.7 Proof of the Main Theorem

Even though the estimates of Lemma 1 are already established elsewhere, for completeness we highlight the corresponding proofs referencing the original work where these proofs appear. The reason in doing so is to present explicit formulas for the various constants that appear in these estimates.

Important elements for demonstrating the necessary bounds constitute the *ladder variables*, see [30], Chapter VIII.4, defined as

$$\begin{aligned}\tau_-^i &= \inf \left\{ t > 0 : S_t^i = \sum_{\ell=1}^t \log \frac{g_i(X_\ell^i)}{f_i(X_\ell^i)} \leq 0 \right\}, \\ \tau_+^i &= \inf \left\{ t > 0 : S_t^i = \sum_{\ell=1}^t \log \frac{g_i(X_\ell^i)}{f_i(X_\ell^i)} > 0 \right\}.\end{aligned}\tag{2.25}$$

We also need to define a third stopping time for  $A \geq 0$

$$\tau_A^i = \inf \left\{ t > 0 : S_t^i = \sum_{\ell=1}^t \log \frac{g_i(X_\ell^i)}{f_i(X_\ell^i)} \geq A \right\}.\tag{2.26}$$

From classical renewal theory and under Assumptions (A1), (A2), we have that  $E_0^i[\tau_+^i], E_\infty^i[\tau_-^i]$

are both finite constants that depend only on  $\{f_i, g_i\}$ , see [41], Theorem C, p. 748 or [64]. This is true because the statistic  $S_t^i$  is a random walk with positive drift  $l_0^i > 0$  under the post-change regime and negative drift  $-l_\infty^i < 0$  under the nominal regime. We also have the following interesting equalities from [30], Corollary 8.39

$$\mathbb{E}_0^i[\tau_+]^i = \frac{1}{\mathbb{P}_0^i(\tau_-^i = \infty)}, \quad \mathbb{E}_\infty^i[\tau_-^i] = \frac{1}{\mathbb{P}_\infty^i(\tau_+^i = \infty)}, \quad (2.27)$$

while for  $\tau_A^i$  a useful estimate [64], Theorem 1 regarding the overshoot of the threshold  $A$

$$\mathbb{E}_0^i[S_{\tau_A^i}^i - A] \leq \frac{J_0^i}{l_0^i}, \quad (2.28)$$

with the upper bound being constant and valid for all  $A \geq 0$ .

We would like to relate the three stopping times  $\tau_-^i, \tau_+^i, \tau_A^i$  to the SPRT  $\mathcal{T}^i$  defined in (Equation 2.37). The first observation is that the SPRT can be written as  $\mathcal{T}^i = \min\{\tau_-^i, \tau_A^i\}$ . Regarding the sequential hypothesis testing procedure implemented by the SPRT we have the two error probabilities  $\alpha_i = \mathbb{P}_\infty^i(S_{\mathcal{T}^i}^i \geq A)$  which is the Type-I error and  $\beta_i = \mathbb{P}_0^i(S_{\mathcal{T}^i} \leq 0)$  which is the Type-II. For  $\tau_-^i$  we can immediately conclude that for  $i = 1, \dots, p$  we have  $\mathcal{T}^i = \min\{\tau_-^i, \tau_A^i\} \leq \tau_-^i$ . We also note that the event  $\{S_{\mathcal{T}^i}^i \leq 0\}$  implies  $\{\tau_-^i < \infty\}$ , namely that  $\tau_-^i$  will also stop. However, it is clear that we can have the latter event occurring without the former. This suggests for the Type-II error that  $\beta_i = \mathbb{P}_0^i(S_{\mathcal{T}^i}^i \leq 0) \leq \mathbb{P}_0^i(\tau_-^i < \infty)$  and in combination with (Equation 2.27) we have

$$\frac{1}{1 - \beta_i} \leq \frac{1}{1 - \mathbb{P}_0^i(\tau_-^i < \infty)} = \frac{1}{\mathbb{P}_0^i(\tau_-^i = \infty)} = \mathbb{E}_0^i[\tau_+]^i. \quad (2.29)$$

Since CUSUM is a repeated SPRT applied to the same process we have [30]

$$\begin{aligned} \mathbb{E}_\infty[\mathbb{T}_i(A)] &= \frac{\mathbb{E}_\infty[\mathcal{T}^i]}{\alpha_i} \\ D_i(\mathbb{T}_i(A)) &= \mathbb{E}_0^i[\mathbb{T}_i(A)] = \frac{\mathbb{E}_0^i[\mathcal{T}^i]}{1 - \beta_i}. \end{aligned} \quad (2.30)$$

In (Equation 2.30) we also indicated the well known fact that the worst-case average detection delay for the classical CUSUM occurs for  $\tau = 0$  when its test statistic is initialized with  $W_0^i = 0$ . Focusing on the average time to false alarm  $E_\infty[\mathbb{T}_i(A)]$ , to establish the lower bound in (Equation 2.14) we observe that  $\mathcal{T}^i \geq 1$  since we take at least one sample, consequently  $E_\infty[\mathcal{T}^i] \geq 1$ . Furthermore, from the estimates provided by Wald [65] we know that  $\alpha_i \leq e^{-A}$ . Substituting in (Equation 2.30) proves the lower bound. The upper bound requires more work. From  $\mathcal{T}^i \leq \tau_-^i$  we have  $E_\infty[\mathcal{T}^i] \leq E_\infty[\tau_-^i]$  which, as we argued, is bounded. We now need a lower bound for  $\alpha_i$ . We can write

$$\begin{aligned}
\alpha_i &= P_\infty(S_{\mathcal{T}^i}^i \geq A) \\
&= E_0^i[e^{-S_{\mathcal{T}^i}^i} \mathbb{1}_{\{S_{\mathcal{T}^i}^i \geq A\}}] \\
&= e^{-A} E_0^i[e^{-(S_{\mathcal{T}^i}^i - A)} | S_{\mathcal{T}^i}^i \geq A] (1 - \beta_i) \\
&\geq e^{-A} e^{-E_0^i[S_{\mathcal{T}^i}^i - A | S_{\mathcal{T}^i}^i \geq A]} (1 - \beta_i) \\
&= e^{-A} \exp \left\{ -\frac{E_0^i[(S_{\mathcal{T}^i}^i - A) \mathbb{1}_{\{S_{\mathcal{T}^i}^i \geq A\}}]}{1 - \beta_i} \right\} (1 - \beta_i), \tag{2.31}
\end{aligned}$$

where we used Jensen's inequality. As mentioned the SPRT satisfies  $\mathcal{T}^i = \min\{\tau_-^i, \tau_A^i\}$ . From this equality we conclude that  $(S_{\mathcal{T}^i}^i - A) \mathbb{1}_{\{S_{\mathcal{T}^i}^i \geq A\}} = (S_{\tau_A^i}^i - A) \mathbb{1}_{\{\tau_A^i < \tau_-^i\}} \leq (S_{\tau_A^i}^i - A)$  and therefore, using (Equation 2.28), we obtain  $E_0^i[(S_{\mathcal{T}^i}^i - A) \mathbb{1}_{\{S_{\mathcal{T}^i}^i \geq A\}}] \leq E_0^i[S_{\tau_A^i}^i - A] \leq \frac{J_0^i}{I_0^i}$ . Additionally, as we argued above, we have  $1 - \beta_i \geq P_0^i(\tau_-^i = \infty) = \frac{1}{E_0^i[\tau_+^i]}$ . If we substitute in (Equation 2.31) we strengthen the inequality and we obtain a lower bound for  $\alpha_i$ . Substituting this lower bound in (Equation 2.30) provides the desired upper bound

$$E_\infty[\mathbb{T}_i(A)] \leq \left( E_\infty[\tau_-^i] E_0^i[\tau_+^i] e^{-E_0^i[\tau_+^i] \frac{J_0^i}{I_0^i}} \right) e^A, \tag{2.32}$$

and therefore we can define  $\mathcal{C}_i = E_\infty[\tau_-^i] E_0^i[\tau_+^i] e^{-E_0^i[\tau_+^i] \frac{J_0^i}{I_0^i}}$ .

Consider now the worst-case average detection delay  $D_i(\mathbb{T}_i(A))$ . To find the upper

bound in (Equation 2.15), we combine Corollary 1 from [64] with (Equation 2.29) and obtain

$$D_i(\mathbb{T}_i(A)) = \mathbb{E}_0^i[\mathbb{T}_i(A)] = \frac{\mathbb{E}_0^i[\mathcal{T}^i]}{1 - \beta_i} \leq \frac{A}{l_0^i} + \mathbb{E}_0^i[\tau_+^i] \frac{J_0^i}{l_0^i}, \quad (2.33)$$

suggesting that  $\mathcal{U}_i = \mathbb{E}_0^i[\tau_+^i] \frac{J_0^i}{l_0^i}$ , see also [61] and [66]. To find an equivalent lower bound we use Wald's lower bound for the expectation of the SPRT [65]

$$\mathbb{E}_0^i[\mathcal{T}^i] \geq \frac{1}{l_0^i} \left\{ (1 - \beta_i) \log \frac{1 - \beta_i}{\alpha_i} + \beta_i \log \frac{\beta_i}{1 - \alpha_i} \right\}, \quad (2.34)$$

from which we conclude

$$\begin{aligned} D_i(\mathbb{T}_i(A)) &= \frac{\mathbb{E}_0^i[\mathcal{T}^i]}{1 - \beta_i} \\ &\geq \frac{1}{l_0^i} \left\{ \log \frac{1 - \beta_i}{\alpha_i} + \frac{\beta_i}{1 - \beta_i} \log \frac{\beta_i}{1 - \alpha_i} \right\} \\ &= \frac{1}{l_0^i} \left\{ \log \frac{1}{\alpha_i} + \log(1 - \beta_i) + \frac{\beta_i}{1 - \beta_i} \log \beta_i \right\}, \end{aligned} \quad (2.35)$$

where the last inequality resulted by removing the nonnegative term  $\frac{\beta_i}{1 - \beta_i} \log \frac{1}{1 - \alpha_i}$ . Recall that  $\alpha_i \leq e^{-A}$  and  $\beta_i \leq 1 - \frac{1}{\mathbb{E}_0^i[\tau_+^i]}$ , furthermore by taking the derivative we can prove that the function  $\log(1 - \beta) + \frac{\beta}{1 - \beta} \log \beta$  is decreasing in  $\beta \in [0, 1]$ , consequently

$$D_i(\mathbb{T}_i(A)) \geq \frac{A}{l_0^i} - \mathcal{L}_i \quad (2.36)$$

with  $\mathcal{L}_i = \frac{1}{l_0^i} \left\{ \log(\mathbb{E}_0^i[\tau_+^i]) + (\mathbb{E}_0^i[\tau_+^i] - 1) \log\left(\frac{\mathbb{E}_0^i[\tau_+^i]}{1 - \mathbb{E}_0^i[\tau_+^i]}\right) \right\}$  which is clearly nonnegative since  $\tau_+^i \geq 1$ . This concludes the proof of the lemma.  $\blacksquare$

**PROOF OF THEOREM 1:** Let us start by considering the lower bound for the average false alarm period in (Equation 2.16). The key idea is to relate  $\mathbb{T}_{\text{MSP}}$  to the following *p prototype*

SPRTs: For  $i = 1, \dots, p$ , a prototype SPRT applied to the  $i$ -th local stream is defined as

$$\mathcal{T}^i = \inf \left\{ t > 0 : S_t^i = \sum_{\ell=1}^t \log \frac{g_i(X_\ell^i)}{f_i(X_\ell^i)} \notin (0, A) \right\}. \quad (2.37)$$

Consider now the sequence  $\{\mathcal{T}_\ell\}, \ell = 1, 2, \dots$ , of SPRTs applied to the data streams when we employ the periodic sampling policy. Then it is clear that each  $\mathcal{T}_\ell$  has the same distribution as a particular prototype SPRT  $\mathcal{T}^i$ . In fact  $\ell$  and  $i$  are related through the equation  $i = (\ell - 1) \bmod p + 1$  suggesting that the distributions of  $\{\mathcal{T}_1, \mathcal{T}_2, \dots\}$  change periodically with period  $p$ . In addition, if we define (a stopping time)  $k$  to be the first time the SPRT of the process being tested crosses the upper boundary  $A$ , then our proposed stopping time  $T_{\text{MSP}}$  can be written as the sum

$$T_{\text{MSP}} = \mathcal{T}_1 + \mathcal{T}_2 + \dots + \mathcal{T}_k = \sum_{\ell=1}^k \mathcal{T}_\ell. \quad (2.38)$$

Our intention is to analyze the proposed scheme  $T_{\text{MSP}}$  by using (Equation 2.38) while the main challenge comes from the fact that these SPRTs require different time steps. However, due to the periodic nature of their corresponding distributions it is possible to come up with interesting formulas. In particular recalling that with  $P_\infty(\cdot), E_\infty[\cdot]$  we denote the probability measure and the corresponding expectation under the pre-change regime then, using (Equation 2.38) we can write

$$\begin{aligned} E_\infty[T_{\text{MSP}}] &= E_\infty \left[ \sum_{\ell=1}^k \mathcal{T}_\ell \right] = \sum_{\ell=1}^{\infty} E_\infty[\mathcal{T}_\ell] P_\infty(k \geq \ell) \\ &= \Omega_\infty + \omega_\infty \Omega_\infty + \omega_\infty^2 \Omega_\infty + \dots = \frac{\Omega_\infty}{1 - \omega_\infty} \end{aligned} \quad (2.39)$$

where, as in Lemma 1,  $\alpha_i = P_\infty(S_{\mathcal{T}^i} \geq A)$  denotes the Type-I error probability of the

SPRT  $\mathcal{T}^i$  when applied onto the  $i$ -th stream as in (Equation 2.37) and where

$$\begin{aligned}\omega_\infty &= \mathbb{P}_\infty(k > p) = \prod_{m=1}^p (1 - \alpha_m), \\ \Omega_\infty &= \sum_{j=1}^p \mathbb{E}_\infty[\mathcal{T}^j] \prod_{m=1}^{j-1} (1 - \alpha_m),\end{aligned}\tag{2.40}$$

with  $\prod_a^b = 1$  when  $a > b$ . The formulas in (Equation 2.39) and (Equation 2.40) are a consequence of the fact that the delays  $\mathcal{T}_\ell$ ,  $\ell = 1, 2, \dots$  of the SPRTs comprising  $\mathsf{T}_{\text{MSP}}$  in (Equation 2.38), have the same distribution as the  $\mathcal{T}^i$ ,  $i = 1, \dots, p$  defined in (Equation 2.37) with the correspondence between the two sets of SPRTs being periodic with period  $p$  and of the form  $\ell \rightarrow i = (\ell - 1) \bmod p + 1$ . Finally, we must mention that in (Equation 2.39) the most crucial point is the fourth equality which is true because  $\mathcal{T}_\ell$  and  $\mathbb{1}_{\{k \geq \ell\}} = \mathbb{1}_{\{k > \ell - 1\}}$  are functions of *non-overlapping* data (the event  $\{k > \ell - 1\}$  depends on data used by  $\mathcal{T}_1, \dots, \mathcal{T}_{\ell-1}$ ) therefore, due to independence accros space and time, and the fact that each  $\mathcal{T}_\ell$  has no memory of past data we conclude that  $\mathcal{T}_\ell$  and  $\mathbb{1}_{\{k \geq \ell\}}$  are independent. As in Lemma 1 we note that  $\mathcal{T}_\ell \geq 1$  and  $\alpha_i \leq e^{-A}$ , therefore (Equation 2.40) implies

$$\begin{aligned}\omega_\infty &\geq (1 - e^{-A})^p, \\ \Omega_\infty &\geq \sum_{j=1}^p (1 - e^{-A})^{j-1} = \left(1 - (1 - e^{-A})^p\right) e^A.\end{aligned}\tag{2.41}$$

Using these lower bounds in (Equation 2.39) yields (Equation 2.16) and proves the desired false alarm estimate.

Consider now the second estimate depicted in (Equation 2.17) for the detection delay. Let us start with the scenario when change occurs at change time  $\tau = 0$ .

**When change occurs at time  $\tau = 0$  :**

Without loss of generality, we assume the change occurs to the  $p$ -th stream, so that our proposed algorithm will achieve the worst detection delay under the scenario when change occurs at time  $\tau = 0$ . We recall that  $\mathbb{P}_0^p(\cdot)$  denotes the probability measure induced by

the change occurring at the  $p$ -th local stream at time  $\tau = 0$  and  $\mathbb{E}_0^p[\cdot]$  the corresponding expectation. It is then clear that all local streams for  $i = 1, \dots, p-1$  are under the nominal regime while only the last process is under the alternative. Similarly to the previous case we can write

$$D_p(\mathbb{T}_{\text{MSP}}) = \mathbb{E}_0^p[\mathbb{T}_{\text{MSP}}] = \Omega_p + \omega_p \Omega_p + \omega_p^2 \Omega_p + \dots = \frac{\Omega_p}{1 - \omega_p} \quad (2.42)$$

where

$$\begin{aligned} \omega_p &= \mathbb{P}_0^p(k > p) = \beta_p \prod_{m=1}^{p-1} (1 - \alpha_m) \\ \Omega_p &= \mathbb{E}_0^p[\mathcal{T}^p] \prod_{m=1}^{p-1} (1 - \alpha_m) + \sum_{j=1}^{p-1} \mathbb{E}_\infty[\mathcal{T}^j] \prod_{m=1}^{j-1} (1 - \alpha_m), \end{aligned} \quad (2.43)$$

where, as in Lemma 1,  $\beta_i = \mathbb{P}_0^i(S_{\mathcal{T}^i} \leq 0)$  denotes the Type-II error probability. Since  $0 \leq \alpha_i, \beta_i \leq 1$ , using (Equation 2.42) and (Equation 2.43) we have the following upper bound

$$D_p(\mathbb{T}_{\text{MSP}}) \leq \frac{\mathbb{E}_0^p[\mathcal{T}^p]}{1 - \beta_p} + \frac{1}{1 - \beta_p} \sum_{j=1}^{p-1} \mathbb{E}_\infty[\mathcal{T}^j]. \quad (2.44)$$

In (Equation 2.44), in the first term of the right hand side, we recognize the detection delay of a CUSUM with threshold  $A$  applied solely to the  $p$ -th local stream. For this quantity we have the upper bound from (Equation 2.15) of Lemma Lemma 1 yielding

$$D_p(\mathbb{T}_{\text{MSP}}) \leq \frac{A}{l_0^p} + \mathcal{U}_p + \frac{1}{1 - \beta_p} \sum_{j=1}^{p-1} \mathbb{E}_\infty[\mathcal{T}^j]. \quad (2.45)$$

As pointed out in Lemma 1,  $\mathcal{U}_p$  is a constant that depends only on  $f_p$  and  $g_p$  and not on  $A$ .

Since the change may occur at any stream we can write similar estimates for any  $i$ -th

stream, specifically

$$D_i(\mathsf{T}_{\text{MSP}}) \leq \frac{A}{l_0^i} + \mathcal{U}_i + \frac{1}{1 - \beta_i} \sum_{j=1, j \neq i}^p \mathbb{E}_\infty[\mathcal{T}^j]. \quad (2.46)$$

To bound the  $p - 1$  terms in the sum in (Equation 2.46) we need to bound  $\beta_i$  and  $\mathbb{E}_\infty[\mathcal{T}^i]$  which can be accomplished by employing the ladder variables defined in (Equation 2.25). Combining (Equation 2.46) and (Equation 2.29) we can write

$$D_i(\mathsf{T}_{\text{MSP}}) \leq \frac{A}{l_0^i} + \mathcal{U}_i + \mathbb{E}_0^i[\tau_+^i] \sum_{j=1, j \neq i}^p \mathbb{E}_\infty[\tau_-^j]. \quad (2.47)$$

Since, as we argued in the proof of Lemma 1, both averages  $\mathbb{E}_0^i[\tau_+^i]$ ,  $\mathbb{E}_\infty[\tau_-^i]$  are finite under (A1), (A2), if we define

$$\mathcal{D} = \left( \max_{1 \leq i \leq p} \mathbb{E}_0^i[\tau_+^i] \right) \left( \max_{1 \leq i \leq p} \mathbb{E}_\infty[\tau_-^i] \right) \quad (2.48)$$

and use it to strengthen the inequality in (Equation 2.47) we obtain (Equation 2.17). This completes the proof.

**When change occurs at general time  $\tau \geq 1$  :**

We now consider the more complicated scenario when change occurs at general time  $\tau \geq 1$ , and we need a series of lemmas.

**LEMMA 2** *For any non-negative random variable  $X$  and two disjoint non-null events  $H_1$  and  $H_2$ , we have*

$$\mathbb{E}(X|H_1 \cup H_2) \leq \mathbb{E}(X|H_1) + \mathbb{E}(X|H_2) \quad (2.49)$$

PROOF OF LEMMA 2:

$$\begin{aligned}
\mathbb{E}(X|H_1 \cup H_2) &= \frac{\mathbb{E}(X \mathbb{1}_{\{H_1 \cup H_2\}})}{\mathbb{P}(H_1 \cup H_2)} \\
&= \frac{\mathbb{E}(X \mathbb{1}_{\{H_1\}}) + \mathbb{E}(X \mathbb{1}_{\{H_2\}})}{\mathbb{P}(H_1) + \mathbb{P}(H_2)} \\
&\leq \frac{\mathbb{E}(X \mathbb{1}_{\{H_1\}})}{\mathbb{P}(H_1)} + \frac{\mathbb{E}(X \mathbb{1}_{\{H_2\}})}{\mathbb{P}(H_2)} \\
&\leq \mathbb{E}(X|H_1) + \mathbb{E}(X|H_2).
\end{aligned}$$

LEMMA 3 *Consider the following SPRT:*

$$T_W = \inf\{t > 0 : S'_t = W + \sum_{\ell=1}^t \log \frac{g(X_\ell)}{f(X_\ell)} \notin (0, A)\}, \quad (2.50)$$

for some  $W, A, g, f$ . Assume that  $X_\ell$  are i.i.d., and we have:

$$\begin{aligned}
\mathbb{E}_f[T_W] &\leq C_1 W + C_2, \\
\mathbb{E}_g[T_W] &\leq \frac{\mathbb{P}_g(S'_{T_W} \geq A)A}{I(g, f)} + C_3,
\end{aligned}$$

for some constants  $C_1, C_2, C_3 > 0$  that are not related to  $W, A$ .

PROOF OF LEMMA 3: Standard analysis of SPRT in sequential analysis.

Without loss of generality, we assume the change occurs to the  $j$ -th stream for some  $j \in \{1, \dots, p\}$ , and we have:

$$\begin{aligned}
\mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau] &= \mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau = j] \mathbb{P}_\tau^j(R_\tau = j) \\
&\quad + \mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau \neq j] \mathbb{P}_\tau^j(R_\tau \neq j).
\end{aligned}$$

We then define the following SPRT which corresponds to the detection process when the

change occurs at time  $\tau$  and the sensor points to the stream  $R_\tau$  at time  $\tau + 1$ :

$$T_\tau = \inf\{t > 0 : S_t^\tau = W_\tau + \sum_{\ell=1}^t \log \frac{g_{R_\tau}(X_\ell^{R_\tau})}{f_{R_\tau}(X_\ell^{R_\tau})} \notin (0, A)\}.$$

For the term  $\mathbb{E}_t^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau = j]$ , we have:

$$\mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau = j] \leq \mathbb{E}_\tau^j[T_\tau | \mathcal{F}_\tau^j, T > \tau, R_\tau = j] + \mathbb{P}_\tau^j(S_{T_\tau}^\tau \geq A) \mathbb{E}_0^j[T].$$

This is because once the sensor switches to other streams after change time  $\tau$ , the remaining detection process will be irrelevant to  $\mathcal{F}_\tau^j$  and returns to the simple scenario when change occurs at time  $\tau = 0$ . Based on Lemma 3 we obtain that

$$\mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau = j] \leq \frac{A}{\mathfrak{l}_0^p} + C_4(p - 1) + C_5,$$

for some constants  $C_4, C_5$ . For the term  $\mathbb{E}_t^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau \neq j]$ , similarly we have

$$\mathbb{E}_\tau^j[T - \tau | \mathcal{F}_\tau^j, T > \tau, R_\tau \neq j] \leq \frac{A}{\mathfrak{l}_0^p} + C_4(p - 1) + C_1 \mathbb{E}_\tau^j[W_\tau | \mathcal{F}_\tau^j, T > \tau, R_\tau \neq j] + C_5,$$

and

$$\begin{aligned} \mathbb{E}_\tau^j[W_\tau | \mathcal{F}_\tau^j, T > \tau, R_\tau \neq j] &= \mathbb{E}_\tau^j[W_\tau | T > \tau, R_\tau \neq j] \\ &\leq \sum_{i \neq j} \mathbb{E}_\tau^j[W_\tau | T > \tau, R_\tau = i] \\ &\leq C_6(p - 1), \end{aligned}$$

for some constant  $C_6$ , where the first step results from the fact that the value of  $W_\tau$  is not observed from  $\mathcal{F}_\tau^j$ , and the second step is proved in Lemma E.4 in [67].  $\blacksquare$

## CHAPTER 3

# QUICKEST DETECTION IN HIGH-DIMENSIONAL LINEAR MODELS VIA IMPLICIT REGULARIZATION

This paper is a collaborative work with Dr. Yi Yu and Dr. Yajun Mei.

### 3.1 Introduction

While predictive machine learning and artificial intelligent algorithms have been widely used in many real world applications, it is often unclear what the lifetime of such algorithms are or when one needs to update those algorithms. For instance, while the COVID-19 pandemic yielded a booming housing market, Zillow, a major real estate web company, actually lost billion of dollars in its house-flipping business, Zillow Offers, partly because its algorithm did not update quickly to reflect the market and thus yields to over-estimated values of the homes it bought. As a result, Zillow was forced to cut 25% of its workforce and closed down the business of Zillow Offers, see [68].

In this paper, we tackle the problem of detecting changes in high-dimensional linear regression models - one of the most fundamental predictive models. Our approach is inspired by recent research on (deep) neural networks, where optimization algorithms such as (stochastic) gradient descent hold *implicit regularization* properties, see [28], [29], etc. By treating linear regression models as two-layers neural networks, we develop efficient implicit regularization-based quickest detection for linear regression models. To the best of our knowledge, this is the first to apply implicit regularization in the text of quickest detection, and our ideas can be easily extended to the context of monitoring other advanced predictive machine learning or artificial intelligent algorithms.

To be specific, we assume that we observe a sequence  $\{(y_t, X_t)\}_{t \in \mathbb{Z}_+} \subset \mathbb{R}^m \times \mathbb{R}^p$  over

time  $t \in \mathbb{Z}_+$  which can be modeled as

$$y_t = X_t \beta_t + \epsilon_t, \quad (3.1)$$

where  $\beta_t \in \mathbb{R}^p$  is a  $p$ -dimensional vector. The system is initially in control and the parameter vector takes a constant value  $\beta_t = \beta_0$ . At some unknown time  $\nu \in \mathbb{Z}_+$ , an event occurs and changes the regression coefficient vector to  $\beta_1 \in \mathbb{R}^p$ ,  $\beta_1 \neq \beta_0$ . In the literature, such time  $\nu$  is often called the “change point”. The primary goal is to develop an efficient algorithm to raise an alarm as quickly as possible once it occurs at change point  $\nu$  based on the observed data sequence  $\{(y_t, X_t)\}_{t \in \mathbb{Z}_+}$ .

It is useful to point out that statistical research on online/sequential monitoring of change points in high-dimensional linear model has been studied in the statistical literature, but most through *explicit regularization* methods such as LASSO proposed in [69]. For instance, [70] developed a LASSO-based multivariate statistical process control (SPC) methodology. [71] developed residual-based detection statistics via LASSO estimator. Here we take a different approach by adopting the implicit regularization method, which might allow us to handle more complicated models or algorithms.

The rest of the paper is organized as follows: in section 3.2 we formulate the problem and review the implicit regularization methods in linear regression model. In section 3.3 we propose the general algorithm and the efficient implementation via implicit regularization. In section 3.4 we develop the theoretical properties of the proposed algorithm and in section 3.5 we conduct several numerical studies to validate our results. We conclude our paper in section 3.6 and present the technical details in section 3.7.

## 3.2 Problem Formulation and Background

### 3.2.1 Mathematical Formulation

Assume that we observe a sequence of data streams  $\{(y_t, X_t)\}_{t \in \mathbb{Z}_+} \subset \mathbb{R}^m \times \mathbb{R}^p$  from model (Equation 3.1), where the noise vector  $\epsilon_t \in \mathbb{R}^m \sim N(0, \sigma^2 I)$ , and the regression coefficient  $\beta_t$  change from a pre-specified vector  $\beta_0$  to an unknown vector  $\beta^* \neq \beta_0$  at some unknown change point  $\nu$ , i.e.

$$\beta_t = \begin{cases} \beta_0, & t \leq \nu, \\ \beta^*, & t > \nu. \end{cases} \quad (3.2)$$

The known value of  $\beta_0$  is reasonable in many applications such as the quality control in manufacturing engineering. Without loss of generality, we assume  $\beta_0 = 0$  since we can monitor  $y_t - X_t \beta_0$  instead of  $y_t$ .

We consider the random design case, where each component  $(X_t)_{i,j}$  of the matrix  $X_t$  is assumed to be independent and identically distributed random variables with the standard Gaussian distribution (i.e.,  $(X_t)_{i,j} \sim N(0, 1)$ ), for all  $i = 1, \dots, m, j = 1, \dots, p$  and  $t = 1, \dots, \infty$ . In addition, we assume that the change is entrywise sparse, and define the sparsity  $s^* = \|\beta^* - \beta_0\|_\infty$ , where  $\|\cdot\|_\infty$  denotes the  $L$ -infinity norm.

Our goal is to develop an efficient algorithm to detect the change based on the observed data  $\{(y_t, X_t)\}_{t \in \mathbb{Z}_+}$  as quickly as possible. An algorithm for quickest detection problem can be characterized by a stopping time  $T$  with respect to the observed data sequence, where  $T = n$  means that we raise a global alarm at time  $n$ . Denote by  $P_\nu$  and  $E_\nu$  the probability measure and expectation when change occurs at time  $t$ . Denote by  $P_\infty$  and  $E_\infty$  the probability measure and expectation when there are no changes, or equivalently, when the change occurs at time  $\infty$ . Motivated by [72], the detection delay of a stopping time  $T$  can be evaluated by the following worst case detection delay conditioned on  $T > \nu$ :

$$D(T) = \sup_{\nu \geq 1} E_\nu[T - \nu \mid T > \nu], \quad (3.3)$$

subject to the average run length (ARL) to false alarm control:

$$E_\infty[T] \geq \gamma, \quad (3.4)$$

for some pre-specified  $\gamma > 1$ .

Had we known the true value of the post-change vector  $\beta^*$ , the problem can be solved via classical CUSUM procedure in Page [62]. Mathematically, denote by  $f_{X\beta^*}$  as the probability density function (pdf) of  $N(X\beta^*, \sigma^2 I)$ , we define the following detection statistics:

$$W_t = \max \left\{ W_{t-1} + \log\left(\frac{f_{X_t\beta^*}(y_t)}{f_0(y_t)}\right), 0 \right\}, \quad (3.5)$$

for  $t \geq 1$  with  $W_0 = 0$ . The corresponding CUSUM stopping time is then defined as

$$T_{\text{CUSUM}} = \inf\{t > 0 : W_t \geq A\}, \quad (3.6)$$

for some pre-specified constant  $A > 0$ .

When the post-change regression coefficient  $\beta^*$  is unknown, an intuitive idea would be to construct an estimator  $\hat{\beta}_t$  based on the historical data and plug it into the standard CUSUM statistics in (Equation 3.5) for detection. However, there are two main challenges: (1) the subset of data to be used for the estimation is unclear. Due to the unknown change point  $\nu$ , the data used might be a mixture model from pre-change and post-change scenarios. This results in a potentially large bias of estimators and detection delay. (2) Efficient and accurate estimators remain unclear. We therefore review two signal recovery methods in high-dimensional linear regression model in the next subsection.

### 3.2.2 Review of Signal Recovery Methods in Linear Models

Consider data  $(y, X) = (y_t, X_t) \subset \mathbb{R}^m \times \mathbb{R}^p$  generated from model (Equation 3.1). A standard approach when  $p$  does not diverge with  $m$  is to estimate  $\beta^*$  via least squares

estimators (LSE), which aims to find  $\beta$  that minimizes the residual sum squares (RSS). The optimization problem can be efficiently solved via the gradient descent algorithm.

In the high-dimensional scenario when  $p$  is diverging with  $m$ , many efficient algorithms have been developed, such as Lasso proposed in [69]. Researchers recently find out that combining the over-parametrization with gradient descent can lead to a sparse solution that achieves the minimax rate, which is also known as the implicit regularization algorithm in linear models, see [73], and [74].

To be more specific, for any vector  $\beta \in \mathbb{R}^p$ , we over-parameterize  $\beta$  by two vectors  $u, v \in \mathbb{R}^p$ :

$$\beta = u \circ u - v \circ v,$$

where  $\circ$  is the Hadamard product that denotes the pointwise multiplication. Denote by  $\|\cdot\|_2$  the  $L_2$  norm, the RSS becomes

$$L(u, v) = \frac{1}{m} \|X(u \circ u - v \circ v) - y\|_2^2, \quad (3.7)$$

under the over-parameterization. We then apply the gradient descent algorithm to (Equation 3.7) to recursively update  $u$  and  $v$  by

$$\begin{aligned} \beta_\ell &= u_\ell \circ u_\ell - v_\ell \circ v_\ell, \\ u_{\ell+1} &= u_\ell - 4\eta u_\ell \circ \left[ \frac{1}{n} X^\top \{X\beta_\ell - y\} \right], \\ v_{\ell+1} &= v_\ell + 4\eta v_\ell \circ \left[ \frac{1}{n} X^\top \{X\beta_\ell - y\} \right]. \end{aligned} \quad (3.8)$$

for  $\ell = 1, \dots, L_{\max}$ , where  $L_{\max}$  is the pre-specified maximum iteration number. The estimator can then be written as  $\hat{\beta} = u_{L_{\max}} \circ u_{L_{\max}} - v_{L_{\max}} \circ v_{L_{\max}}$ . Here the gradient descent is initialized by  $u_0, v_0 \sim \text{Unif}[-\alpha, \alpha]$  for some  $\alpha > 0$ . Parameters including step size  $\eta$ , iteration number  $L_{\max}$ , and magnitude  $\alpha$  need to be properly selected to ensure the optimal rate of implicit regularization algorithm in high-dimensional linear models.

### 3.3 Implicit Regularization-Based Quickest Detection

Our proposed stopping time  $T_{\text{IR}}$  contains three key components: the estimators  $\hat{\beta}_t$ , the monitoring statistics  $W_t$ , and the candidate change point  $M(t)$ . At a high level, for each time instant  $t$ , we construct the estimators  $\hat{\beta}_t$  based on the data between the candidate change point  $M(t)$  and the current time  $t$ . Then we update the statistics  $W_t$  based on  $\hat{\beta}_t$  and choose to re-set or keep the candidate change point  $M(t)$  based on the value of  $W_t$ . For the better presentation, we define the three components separately in three subsections.

#### 3.3.1 Estimators $\hat{\beta}_t$

Let us begin with the construction of the estimator  $\hat{\beta}_t$  of the true coefficient  $\beta^*$ . At each time  $t$ , if the candidate change point  $M(t) = t - 1$ , we directly set  $\hat{\beta}_t = 0 \in \mathbb{R}^p$ , which is exactly the value of pre-change coefficient. Otherwise if  $M(t) < t - 1$ , we consider the time window between the candidate change point  $M(t)$  and current time  $t$ :  $[M(t) + 1, t - 1]$ . Denote by  $y_{M(t)+1, t-1}, X_{M(t)+1, t-1}$  the aggregation of the observed data in this time window:

$$\begin{aligned} y_{M(t)+1, t-1} &= [y_{M(t)+1}^\top, \dots, y_{t-1}^\top]^\top, \\ X_{M(t)+1, t-1} &= [X_{M(t)+1}^\top, \dots, X_{t-1}^\top]^\top. \end{aligned} \tag{3.9}$$

We define the proposed estimator  $\hat{\beta}_t$  in three steps:

1. We introduce six tuning parameters:  $\alpha_0, \eta_0, c_L > 0$  for the implementation of implicit regularization algorithm, and  $s, c, C > 0$  for the truncation of the estimator.
2. we implement the implicit regularization algorithm in (Equation 3.8) to the aggregated data  $(y_{M(t)+1, t-1}, X_{M(t)+1, t-1})$  to obtain an initial estimator  $\tilde{\beta}_t$  with the initial

value  $\alpha_t$ , step size  $\eta_t$  and the maximum iteration number  $L_t$  defined in:

$$\begin{aligned}\alpha_t &= \frac{\alpha_0}{\sqrt{t - M(t) - 1}}, \\ \eta_t &= \eta_0, \\ L_t &= \frac{c_L(t - M(t) - 1)^{1/4}}{\eta_t \sigma \sqrt{\log p/m}} \log \frac{1}{\alpha_t}.\end{aligned}$$

3. We adjust the estimator  $\tilde{\beta}_t$  to obtain the final estimator  $\hat{\beta}_t$  by truncating the  $s$  largest components of  $\tilde{\beta}_t$  with lower threshold  $c$  and upper threshold  $C$ , and setting the remaining  $p - s$  components to 0. Mathematically, if we denote by  $\hat{S}_t \subset \{1, \dots, p\}$  the index of  $s$  largest components at time  $t$ , then

$$\begin{aligned}(\hat{\beta}_t)_i &= 0, \text{ if } i \notin \hat{S}_t, \\ (\hat{\beta}_t)_i &= \min \left\{ C\sigma, \max \{ c\sigma, (\tilde{\beta}_t)_i \} \right\}, \text{ if } (\tilde{\beta}_t)_i > 0, i \in \hat{S}_t, \\ (\hat{\beta}_t)_i &= \min \left\{ -c\sigma, \max \{ -C\sigma, (\tilde{\beta}_t)_i \} \right\}, \text{ if } \tilde{\beta}_{t,i} < 0, i \in \hat{S}_t.\end{aligned}$$

The details on the selection of parameters  $\alpha_0, \eta_0, c, C, c_L$  will be postponed to Section section 3.4.

### 3.3.2 Monitoring Statistics $W_t$

With the estimator  $\hat{\beta}_t$ , we are able to plug it into the CUSUM statistics in (Equation 3.5) for detection. For better illustration, we define the monitoring statistics  $W_t$  in two steps. Firstly, we define an initial statistics  $\tilde{W}_t$  based on the classical CUSUM update for the detection of change point:

$$\tilde{W}_t = \max \left\{ W_{t-1} + \log \frac{f_{X_t \hat{\beta}_t}(y_t)}{f_0(y_t)}, 0 \right\}.$$

If  $\tilde{W}_t \geq A$  for some threshold  $A > 0$  that will be defined later, we directly set  $W_t = \tilde{W}_t$ . If  $\tilde{W}_t < A$ , we need to check the amount of data that are used for detection. For this

purpose, we introduce a new tuning parameter  $q = q(A)$ . If the length of the time window  $[M(t) + 1, t - 1]$  is too long (i.e.,  $t - M(t) - 1 \geq q$ ), we discard these data, re-set  $W_t$  back to 0 and restart the estimation. If  $t - M(t) - 1 < q$ , we still set  $W_t = \tilde{W}_t$ . To be mathematically rigorous, we define

$$W_t = \begin{cases} \tilde{W}_t, & \text{if } t - M(t) - 1 < q_A \text{ or } \tilde{W}_t \geq A, \\ 0, & \text{if } t - M(t) - 1 \geq q_A \text{ and } \tilde{W}_t < A. \end{cases} \quad (3.10)$$

### 3.3.3 Candidate change point $M(t)$

With the defined monitoring statistics  $W_t$  in (Equation 3.10), let us define the candidate change point  $M(t + 1)$  for the next time instant  $t + 1$ . If  $W_t > 0$ , which indicates the possible existence of a change point, we keep the candidate change point as it is. If  $W_t = 0$ , we reset  $M(t + 1)$  back to the current time  $t$ . This is to say,

$$M(t + 1) = \begin{cases} t + 1, & \text{if } W_t > 0, \\ M(t), & \text{if } W_t = 0. \end{cases} \quad (3.11)$$

We raise a global alarm at the stopping time  $T_{\text{IR}}$

$$T_{\text{IR}} = \inf\{t > 0 : W_t \geq A\}, \quad (3.12)$$

Note that the threshold  $A > 0$  is the same as  $A$  used for the definition of monitoring statistics, and is selected to satisfy the ARL to false alarm constraint in (Equation 3.4). Our proposed algorithm can be described as follows.

## 3.4 Theoretical Properties

In this section, we establish the theoretical guarantees for our proposed stopping time  $T_{\text{IR}}$  in (Equation 3.12). We begin by establishing a choice of the threshold  $A$  that guarantees

---

**Algorithm 1** Implicit-Regularization-Based Quickest Detection.

---

```
1: Initialize magnitude  $\alpha$ , step size  $\eta$ , tuning parameters  $s, S, c_L, q$  and threshold  $A$ .
2: Initialize  $W_0 = 0, M(1) = 0$ 
3: while  $t > 0$  do
4:   Observe data  $(y_t, X_t)$ .
5:   Obtain historical data  $y_{M(t)+1,t-1}, X_{M(t)+1,t-1}$ .
6:   Obtain the estimator  $\hat{\beta}_t$ .
7:   Update the monitoring statistics  $W_t$ .
8:   if  $W_t \geq A$  then
9:     Raise an alarm.
10:  else if  $W_t < A$  then
11:    Update the candidate change point  $M(t + 1)$ .
12:  end if
13:   $t = t + 1$ 
14: end while
```

---

that the ARL to false alarm constraint in (Equation 3.4) is met.

LEMMA 4 *For any constraint  $\gamma > 1$ , if the threshold  $A \geq \log \gamma$ , then we have that*

$$E_\infty[T_{\text{IR}}] \geq \gamma. \quad (3.13)$$

We are now ready to analyze the detection delay relationship of  $T_{\text{IR}}$ . The main challenge here as compared to the analysis in standard linear regression model without change points is how to characterize the value of candidate change point  $M(\nu + 1)$ . If  $M(\nu + 1) < \nu$ , the estimator  $\hat{\beta}_t$  for  $t > \nu$  is biased before the reset of candidate change point due to the fact that the data in the time-window  $[M(t) + 1, t - 1]$  is a mixture of both pre-change and post-change data. Under such scenario, a low percentage of pre-change data in the time-window is crucial to the accuracy of the estimation. Furthermore, the dependence between the value of  $M(\nu + 1)$  and the condition  $T_{\text{IR}} > \nu$  brings additional challenge to the theoretical analysis. To tackle this issue, we have the following lemma on  $M(\nu + 1)$ .

LEMMA 5 *For any  $n = 1, 2, \dots$ ,*

$$P_\nu[\nu - M(\nu + 1) \geq n \mid T_{\text{IR}} > \nu] \leq c_1 \exp(-c_2 n). \quad (3.14)$$

for some constants  $c_1, c_2 > 0$ .

Lemma Lemma 5 shows that  $\nu - M(\nu + 1)$  has an exponential tail bound, which means that with large probability, the candidate change point  $M(\nu + 1)$  will be close to the change point  $\nu$  and ensures the low percentage of pre-change data during the estimation for  $t > \nu$ .

Define the following Kullback-Leibler divergence of the post-change versus the pre-change distribution:

$$I(\beta^*) = \mathbb{E}_{X_t}[\mathbb{E}_{y_t}[\log \frac{f_{X_t \beta^*}(y_t)}{f_0(y_t)}]],$$

where  $y_t, X_t$  are generated from model (Equation 3.1) with  $\beta_t = \beta^*$ . We have the following theorem on the detection delay of our proposed stopping time  $T_{\text{IR}}$ .

**THEOREM 2** *If the tuning parameters for the implicit regularization algorithm  $\alpha_0, \eta_0, c_L, c, C, s, q$  satisfy:*

$$\alpha_0 \leq \min\left\{\frac{1}{5p^2}, 1, \frac{1}{2}\sqrt{\beta_{\min}^*}, \frac{c_L \sigma \sqrt{\frac{\log p}{m}}}{3(\beta_{\max}^* \sigma)^2}\right\}, \quad (3.15)$$

$$\eta_0 \leq \frac{1}{20\beta_{\max}^*}, \quad c_L \leq c_3, \quad (3.16)$$

$$|\beta_{\min}^*| \geq c\sigma, \quad |\beta_{\max}^*| \leq C\sigma, \quad s \geq s^*, \quad (3.17)$$

$$q = O(e^{c_4 A}), \quad (3.18)$$

for some constant  $c_3 > 0, 0 < c_4 < 1/2$ . the detection delay of our proposed algorithm is bounded by:

$$D[T_{\text{IR}}] \leq \frac{A}{I(\beta^*)} + \frac{C_1 m}{I(\beta^*)} \sqrt{D[T_{\text{IR}}]} + \frac{C_2 \sqrt{A/I(\beta^*)}}{I(\beta^*)} + C_3. \quad (3.19)$$

for any  $m \geq C_4 \log p$  and some constant  $C_1, C_2, C_3, C_4$  not related to  $m, \gamma, p$ .

To help better understand Theorem Theorem 2, here we add a few remarks.

1. By letting  $A = \log \gamma$  in relationship (Equation 3.19), we conclude that the detection delay of our proposed algorithm  $T_{\text{IR}}$  mainly consists of two parts: (1) the

standard delay term  $\log \gamma / I(\beta^*)$  in change point analysis, (2) the additional term  $O(m\sqrt{\mathbb{D}[T_{\text{IR}}]}/I(\beta^*))$  which results from the information loss during the estimation of the unknown coefficient.

2. When the number of observations per time step  $m \ll \log \gamma$ , our proposed stopping time  $T_{\text{IR}}$  is first-order asymptotically optimal in the sense of minimizing detection delay for each and every  $\beta^*$  in the region defined in Theorem Theorem 2. This is because it asymptotically attains the lower bound of the CUSUM procedure in (Equation 3.6) that knows the true value of  $\beta^*$ . Unfortunately if  $m \geq O(\log \gamma)$ , this conclusion no longer holds as the additional term plays a non-negligible role in the delay.
3. Our proposed algorithm included seven tuning parameters. The two most important parameters are  $\alpha_0, \eta_0$  which determines the performance of implicit regularization algorithm when estimating the post-change regression coefficient  $\beta^*$ . The other five parameters  $c, C, s, q, c_L$  are introduced for theoretical analysis and has insignificant impact in the numerical studies. In our paper, we select  $\alpha_0 = 0.001, \eta_0 = 0.1, c = 0, C = \infty, s = p, q = +\infty, c_L = 1$ .
4. We would like to comment on the computational complexity of our algorithm. At each time  $t$ , the computational complexity of  $T_{\text{IR}}$  is  $O(mpL_t)$ , and this is comparable to that of the explicit regularization methods such as LASSO. In addition, the memory requirement of our proposed method is  $mp(t - M(t) - 1)$ . We conjecture that  $t - M(t) - 1$  can be reduced to  $O(1)$  by recursively estimating  $\beta^*$ , which will be investigated elsewhere.

### 3.5 Numerical Results

In this section, we conduct Monte Carlo simulation studies to validate our theoretical results. The dimension is set to  $p = 1000$ . Assume the pre-change coefficient  $\beta_0 = 0$  and the

post-change coefficient  $\beta^* = [1, 1, 1, 1, 0, \dots, 0]^\top$ .

Two baseline methods are considered:

- $T_{\text{CUSUM}} : \hat{\beta}_t = \beta^*$  for all  $t$ , which is unrealistic in real-world application since it assumes the true post-change coefficients are given.
- $\tilde{T} : \hat{\beta}_t = [b, \dots, b]$  with  $b = \|\beta^*\|_2^2/p = 4/p$ . Here  $b$  is selected so as to maximize the Kullback information number  $E_0[\log(f_{X_t, \hat{\beta}_t}(y_t)/f_0(y_t))]$  under  $P_0$ . For the fairness of comparison, only the estimators  $\hat{\beta}_t$  at each time  $t$  are different for  $T_{\text{IR}}, \tilde{T}$ , while the construction of statistics and the detection policy remains the same.

The detailed settings are presented as follows:

- $m = 20, p = 200, 400, \sigma = 1$ .
- $\gamma = 1000, 2000, 5000, 10000, 20000$ .
- $\alpha_0 = 0.001, \eta_0 = 0.1$ .
- $c = 0, C = +\infty, c_L = 1, s = p, q_A = \infty$ .

Table 3.2: Detection Delay of  $T, T_{\text{CUSUM}}, \tilde{T}$  for  $p = 400$

| $\gamma$ | $T_{\text{IR}}$ | $T_{\text{CUSUM}}$ | $\tilde{T}$  |
|----------|-----------------|--------------------|--------------|
| 1000     | 3.165±0.038     | 1.002±0.001        | 15.484±0.185 |
| 2000     | 3.201±0.039     | 1.003±0.001        | 16.503±0.189 |
| 5000     | 3.250±0.039     | 1.004±0.001        | 17.789±0.195 |
| 10,000   | 3.287±0.039     | 1.005±0.001        | 18.746±0.198 |
| 20,000   | 3.323±0.039     | 1.007±0.001        | 19.710±0.202 |
| $\gamma$ | $T_{\text{IR}}$ | $T_{\text{CUSUM}}$ | $\tilde{T}$  |
| 1000     | 4.089±0.049     | 1.002±0.001        | 26.765±0.273 |
| 2000     | 4.147±0.049     | 1.003±0.001        | 28.460±0.277 |
| 5000     | 4.204±0.049     | 1.005±0.001        | 30.675±0.282 |
| 10,000   | 4.248±0.049     | 1.006±0.001        | 32.321±0.285 |
| 20,000   | 4.289±0.049     | 1.008±0.001        | 34.028±0.289 |

For each  $\gamma$  and  $T_{\text{IR}}, T_{\text{CUSUM}}, \tilde{T}$ , we first use the bisection method to find suitable threshold  $A$  to attain the false alarm constraint, and then simulate the detection delay under the special scenario when  $\nu = 0$ . All results are based on 10,000 Monte Carlo simulations.

From Table I,  $T_{\text{CUSUM}}$  performs the best, which is consistent with the optimality of CUSUM when the true post-change parameters are completely specified. However, this is infeasible in practice. In addition, our proposed stopping time  $T_{\text{IR}}$  has a better detection delay performance than  $\tilde{T}$  for all scenarios, which shows the effectiveness of the estimation of the unknown parameters.

### 3.6 Conclusions

In this paper, we consider the detection of a potential change-point in high-dimensional linear regression models. We develop efficient algorithms by estimating the unknown coefficient via the implicit regularization algorithm and plug-in to the CUSUM statistics for the detection. We show that our proposed algorithm asymptotically attain the optimal lower bound of detection delay while maintains the average run length to false alarm control.

There are a couple of future research directions. The first one is that we conjecture that the computational complexity of our proposed algorithm can be improved in the sense that the complexity is irrelevant to time  $t$ . One possible way is to utilize the mean of estimator obtained at each time, while the main challenge is how to analyze the detection delay relationship when the estimation might not be accurate. The second is to consider the extension from linear regression model to more complicated machine learning models such as neural network, which will be investigated in future work.

### 3.7 Technical Details

#### 3.7.1 Theoretical Properties of Implicit Regularization in Linear Model

In this subsection, we provide some theoretical results related to the implicit regularization algorithm. Consider the linear regression model with sparse vector  $\beta^*$ :

$$y = X\beta^* + \epsilon,$$

where  $X \in \mathbb{R}^{m \times p}$ ,  $y \in \mathbb{R}^m$ , and the Gaussian noise vector  $\epsilon \in \mathbb{R}^m$ .

In addition, we define the following iterative update from gradient descent in implicit regularization algorithm:

$$\begin{aligned}\beta_\ell &= u_\ell \odot u_\ell - v_\ell \odot v_\ell, \\ u_{\ell+1} &= u_\ell \odot \left( 1 - 4\eta \left( \frac{1}{m} X^\top (X\beta_\ell - y) \right) \right), \\ v_{\ell+1} &= v_\ell \odot \left( 1 + 4\eta \left( \frac{1}{m} X^\top (X\beta_\ell - y) \right) \right),\end{aligned}$$

for  $\ell = 1, \dots, L_{\max}$ , with suitable selected maximal iteration number  $L_{\max}$ , initial value  $u_0, v_0 = \alpha \mathbf{1}_{p \times 1}$ , and step size  $\eta$ .

Following the notations from [74], denote  $S$  as the support set of the sparse vector  $\beta^*$ ,  $S^+$  as the coordinates of  $\beta^*$  such that  $\beta_i^* > 0$ , and  $S^-$  as the coordinates of  $\beta^*$  such that  $\beta_i^* < 0$ . Define

$$\begin{aligned}\beta_\ell^+ &= u_\ell \odot u_\ell, \\ \beta_\ell^- &= v_\ell \odot v_\ell,\end{aligned}$$

and define

$$\begin{aligned}s_\ell &= \mathbf{1}_{S^+} \odot \beta_\ell^+ - \mathbf{1}_{S^-} \odot \beta_\ell^-, \\ e_\ell &= \mathbf{1}_{S^c} \odot \beta_\ell + \mathbf{1}_{S^-} \odot \beta_\ell^+ - \mathbf{1}_{S^+} \odot \beta_\ell^-, \\ b_\ell &= \frac{1}{m} X^\top X e_\ell - \frac{1}{m} X^\top \epsilon, \\ p_\ell &= \left( \frac{1}{m} X^\top X - I \right) (s_\ell - \beta^*).\end{aligned}$$

Then the gradient descent update in implicit regularization algorithm can be rewritten as:

$$\begin{aligned}
\beta_\ell^+ &= \beta_\ell^- = \alpha^2 \mathbf{1}_{p \times 1}, \\
\beta_\ell &= \beta_\ell^+ - \beta_\ell^-, \\
\beta_{\ell+1}^+ &= \beta_\ell^+ \odot (1 - 4\eta(s_\ell - \beta^* + p_\ell + b_\ell))^2, \\
\beta_{\ell+1}^- &= \beta_\ell^- \odot (1 + 4\eta(s_\ell - \beta^* + p_\ell + b_\ell))^2,
\end{aligned} \tag{3.20}$$

for  $\ell = 1, \dots, L_{\max}$ .

We now provide the following key lemma:

**LEMMA 6** *Consider the iterative update in (Equation 3.20). Denote by  $\beta_{\max}^*, \beta_{\min}^*$  the maximal and minimum coefficient of  $\beta^*$  respectively. Fix any  $0 < \zeta \leq \beta_{\max}^*$  and let  $\gamma = \frac{C_\gamma}{\lceil \log_2 \frac{\beta_{\max}^*}{\zeta} \rceil}$  where  $C_\gamma$  is some small enough absolute constant. Suppose that the sequences  $\{b_\ell\}_{\ell \geq 1}$  and  $\{p_\ell\}_{\ell \geq 1}$  satisfies*

$$\begin{aligned}
\|b_\ell\|_\infty &\leq C_b \zeta - \alpha, \\
\|p_\ell\|_\infty &\leq \gamma \|s_\ell - \beta^*\|_\infty,
\end{aligned} \tag{3.21}$$

for some small enough absolute constant  $C_b$ . If the step size satisfies  $\eta \leq \frac{5}{96\beta_{\max}^*}$  and the initialization satisfies  $\alpha \leq \frac{\zeta}{3(\beta_{\max}^*)^2} \wedge \frac{1}{2} \sqrt{\beta_{\min}^*} \wedge 1$ , then for  $L_{\max} = \frac{2}{\eta\zeta} \log \frac{1}{\alpha}$  we have

$$\begin{aligned}
\|s_{L_{\max}} - \beta^*\|_\infty &\leq 4\zeta, \\
\alpha^2 \prod_{i=0}^{\ell-1} (1 + 4\eta(\|b_i\|_\infty + \|p_i\|_\infty))^2 &\leq \alpha,
\end{aligned}$$

for all  $0 \leq \ell \leq L_{\max}$ .

*Proof:* Lemma Lemma 6 is a direct application of Proposition 1 and Lemma B.17 in [74] and here we omit the proof. We only need to note that the selected maximum iteration number  $L_{\max} = \frac{2}{\eta\zeta} \log \frac{1}{\alpha}$  drops in the second-to-last time interval in the induction of Lemma B.17 of [74], and that is why the upper bound here is  $4\zeta$  instead of  $\zeta$ .  $\square$

### 3.7.2 RIP conditions, Random matrix, and Concentration Inequalities

Let us start with the definition of Restricted isometry property (RIP) condition:

**Definition 1** (*Restricted isometry property*). A matrix  $X \in \mathbb{R}^{m \times p}$  is said to satisfy the  $(s, \delta)$ -restricted isometry property if, for any  $s$ -sparse vector  $u$  in  $\mathbb{R}^p$ , we have

$$(1 - \delta)\|u\|^2 \leq \frac{1}{n}\|Xu\|^2 \leq (1 + \delta)\|u\|^2.$$

The following lemma shows that random matrix satisfies RIP conditions with high probability:

**LEMMA 7** *Let  $X$  be a  $m \times p$  matrix with each component follows  $N(0, 1)$ . Then  $X$  satisfies  $(s, \delta)$ -RIP condition with at least the probability:*

$$1 - 2\left(\frac{12}{\delta}\right)^s e^{-c_0\left(\frac{\delta}{2}\right)m}.$$

Here  $c_0(x) = \frac{x^2}{4} - \frac{x^3}{6}$ .

*Proof:* See Lemma 5.1 in [75]. ■

For matrix with RIP condition, we have following lemma:

**LEMMA 8** *Suppose that  $X$  is a  $m \times p$  matrix satisfying the  $(s + 1, \delta)$ -RIP. If  $z \in \mathbb{R}^p$  is a  $s$ -sparse vector, then*

$$\left\| \left( \frac{1}{m} X^\top X - I \right) z \right\|_\infty \leq s\delta \|z\|_\infty.$$

*Proof:* See Lemma A.3 in [74]. □

We now provide some results related to the concentration inequalities under our context.

**LEMMA 9** *Let  $X$  be a  $m \times p$ -dimensional random matrix with each component  $(X)_{i,j}$  i.i.d. follows  $N(0, 1)$ , and let  $\epsilon$  be a  $m$ -dimensional noise vector of independent Gaussian*

random variables with variance  $\sigma^2$ . We then have

$$\mathbb{P}(\|\frac{1}{m}X^\top \epsilon\|_\infty \geq w\sigma) \leq 2p \exp\{-c_{e1} \min(c_{e2}w, (c_{e2}w)^2)m\}, \quad (3.22)$$

for some constants  $c_{e1}, c_{e2}$  that are not related to  $m, p, \sigma$ .

*Proof of Lemma Lemma 9* We write the  $j$ -th component of  $\frac{1}{m}X^\top \epsilon$  as

$$(\frac{1}{m}X^\top \epsilon)_j = \frac{1}{m} \sum_{i=1}^m (X)_{i,j}(\epsilon)_i.$$

Note that the random variable  $(X)_{i,j}(\epsilon)_i$  is sub-exponential, and applying Bernstein's inequality (see Corollary 2.8.3 in [76]) obtains that

$$\mathbb{P}\left\{ \left| (\frac{1}{m}X^\top \epsilon)_j \right| \geq w\sigma \right\} \leq 2 \exp\{-c_{e1} \min(c_{e2}w, (c_{e2}w)^2)m\},$$

for some constants  $c_{e1}, c_{e2}$  that are not related to  $m, p, \sigma$ . Relationship (Equation 3.22) is immediately proved by using the probability union bounds.  $\square$

### 3.7.3 Proof of Detection Delay Relationship When Change Occurs at $\nu = 0$

In this section, we establish the proof for the detection delay relationship (16) in Theorem 4.3 under a special scenario when change occurs at time  $\nu = 0$  and postpone the more complicated scenario when change occurs at general time  $\nu = k > 1$  to the next section.

When change occurs at time  $\nu = 0$ , all observed data  $(y_t, X_t)_{t \geq 1}$  are generated from the post-change model, and we define the following sequential test:

$$\mathcal{T} = \inf\{t > 0 \text{ or } t = q_A : S_t \notin (0, A)\}, \quad (3.23)$$

where the detection statistics  $S_t$  is defined as

$$S_t = \sum_{n=1}^t \log \frac{f_{X_n \hat{\beta}_n}(y_n)}{f_0(y_n)}. \quad (3.24)$$

Here  $\hat{\beta}_n = 0$  for  $n = 1$ , and for general  $n \geq 2$ ,  $\hat{\beta}_n$  is obtained via two steps: (1) applying the implicit regularization algorithm to the following data  $(y_{1:n-1}, X_{1:n-1})$  to get an initial estimator  $\hat{\beta}_n$

$$y_{1:n-1} = [y_1^\top, \dots, y_{n-1}^\top]^\top, \quad X_{1:n-1} = [X_1^\top, \dots, X_{n-1}^\top]^\top,$$

with parameters including the initial value  $\alpha_n$ , iteration number  $L_n$ , and step size  $\eta_n$  defined in

$$\begin{aligned} \alpha_n &= \frac{\alpha_0}{\sqrt{n-1}}, \\ \eta_n &= \eta_0, \\ L_n &= \frac{c_L(n-1)^{1/4}}{\eta_n \sigma \sqrt{\log p/m}} \log \frac{1}{\alpha_n}, \end{aligned}$$

(2) adjusting  $\hat{\beta}_n$  by only keeping the largest  $s$  components of  $\hat{\beta}_n$  and then truncating the non-zero components such that the absolute value of each non-zero component is in  $[c\sigma, C\sigma]$  to get the final estimator  $\hat{\beta}_n$ . In addition, the data  $(X_n, y_n)_{n \geq 1}$  in sequential test (Equation 3.23) are independent and identically distributed (i.i.d.) from the linear model:

$$y_n = X_n \beta^* + \epsilon_n.$$

Based on the deduction at the end of the manuscript, we have:

$$\begin{aligned}
\mathbb{E}_0[T_{\text{IR}} \mid T_{\text{IR}} > 0] &= \mathbb{E}_0[T_{\text{IR}}] \\
&= \mathbb{E}_0[\mathcal{T}_1 + \cdots + \mathcal{T}_w] \\
&= \sum_{i=1}^{\infty} \mathbb{E}_0[\mathcal{T}_i] \mathbb{P}_0(w \geq i) \\
&= \frac{\mathbb{E}_0[\mathcal{T}]}{\mathbb{P}_0(S_{\mathcal{T}} \geq A)}.
\end{aligned}$$

We conclude that to bound the detection delay of our proposed stopping time  $T_{\text{IR}}$  when change occurs at time  $\nu = 0$ , it suffices to bound  $\mathbb{E}_0[\mathcal{T}]$  and  $\mathbb{P}_0(S_{\mathcal{T}} \geq A)$  under  $\mathbb{P}_0$ .

LEMMA 10 Denote  $I(\beta^*)$  as the Kullback information number of the observed data  $(y_t, X_t)$  under  $\mathbb{P}_0$  :

$$I(\beta^*) = \mathbb{E}_0\left[\log \frac{f_{X_t \beta^*}(y_t)}{f_0(y_t)}\right] = \frac{m \|\beta^*\|_2^2}{2\sigma^2},$$

the average run length of  $\mathcal{T}$  is then bounded by

$$\mathbb{E}_0[\mathcal{T}] \leq V_1 + V_2 + V_3, \tag{3.25}$$

where

$$\begin{aligned}
V_1 &= \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}\right], \\
V_2 &= \mathbb{E}_0\left[\frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{2\sigma^2} \|X_n(\beta^* - \hat{\beta}_n)\|_2^2}{I(\beta^*)}\right], \\
V_3 &= \mathbb{E}_0\left[\frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{\sigma^2} (y_n - X_n \beta^*)^\top (X_n(\beta^* - \hat{\beta}_n))}{I(\beta^*)}\right].
\end{aligned} \tag{3.26}$$

*Proof of Lemma Lemma 10:* Applying Wald's equation obtains that

$$\begin{aligned}
\mathbb{E}_0[\mathcal{T}] &= \mathbb{E}_0\left[\frac{\sum_{n=1}^{\mathcal{T}} \log \frac{f_{X_n \beta^*}(y_n)}{f_0(y_n)}}{I(\beta^*)}\right] \\
&= \mathbb{E}_0\left[\frac{\sum_{n=1}^{\mathcal{T}} \log \frac{f_{X_n \hat{\beta}_n}(y_n)}{f_0(y_n)}}{I(\beta^*)} + \frac{\sum_{n=1}^{\mathcal{T}} \log \frac{f_{X_n \beta^*}(y_n)}{f_{X_n \hat{\beta}_n}(y_n)}}{I(\beta^*)}\right] \\
&= \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)} + \frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{2\sigma^2} (\|y_n - X_n \hat{\beta}_n\|_2^2 - \|y_n - X_n \beta^*\|_2^2)}{I(\beta^*)}\right] \\
&= \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)} + \frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{2\sigma^2} \|X_n(\beta^* - \hat{\beta}_n)\|_2^2}{I(\beta^*)} \right. \\
&\quad \left. + \frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{\sigma^2} (y_n - X_n \beta^*)^\top (X_n(\beta^* - \hat{\beta}_n))}{I(\beta^*)}\right] \\
&= V_1 + V_2 + V_3,
\end{aligned}$$

where  $V_1, V_2, V_3$  are defined in (Equation 3.26). □

The lemmas below focus on bounding terms  $V_1, V_2, V_3$ .

LEMMA 11

$$V_1 \leq \frac{\mathbb{A}P_0(S_{\mathcal{T}} \geq A)}{I(\beta^*)} + \frac{\mathbb{E}_0[\mathcal{T}]^{1/2} \mathbb{E}_0[\|y_n\|_2^4]^{1/2}}{2\sigma^2 I(\beta^*)} + \frac{\mathbb{A}P_0(\mathcal{T} = q_A)}{I(\beta^*)}. \quad (3.27)$$

*Proof of Lemma Lemma 11*

$$\begin{aligned}
V_1 &= \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}\right] \\
&= \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, S_{\mathcal{T}} \geq A\right] + \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, S_{\mathcal{T}} \leq 0\right] + \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right] \\
&\leq \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, S_{\mathcal{T}} \geq A\right] + \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right] \\
&= \frac{\mathbb{A}P_0(S_{\mathcal{T}} \geq A)}{I(\beta^*)} + \mathbb{E}_0\left[\frac{S_{\mathcal{T}} - A}{I(\beta^*)}, S_{\mathcal{T}} \geq A\right] + \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right] \\
&= \frac{\mathbb{A}P_0(S_{\mathcal{T}} \geq A)}{I(\beta^*)} + \mathbb{E}_0\left[\frac{\log \frac{f_{X_{\mathcal{T}} \hat{\beta}_{\mathcal{T}}}(y_{\mathcal{T}})}{f_0(y_{\mathcal{T}})}}{I(\beta)}, S_{\mathcal{T}} \geq A\right] + \mathbb{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right].
\end{aligned}$$

Regarding the term  $\mathbf{E}_0\left[\frac{f_{X_{\mathcal{T}}\hat{\beta}_{\mathcal{T}}}(y_{\mathcal{T}})}{I(\beta)}, S_{\mathcal{T}} \geq A\right]$ , we have

$$\begin{aligned}
& \mathbf{E}_0\left[\frac{\log \frac{f_{X_{\mathcal{T}}\hat{\beta}_{\mathcal{T}}}(y_{\mathcal{T}})}{f_0(y_{\mathcal{T}})}}{I(\beta)}, S_{\mathcal{T}} \geq A\right] \\
&= \mathbf{E}_0\left[\frac{\frac{1}{2\sigma^2}(\|y_{\mathcal{T}}\|_2^2 - \|y_{\mathcal{T}} - X_{\mathcal{T}}\hat{\beta}_{\mathcal{T}}\|_2^2)}{I(\beta^*)}, S_{\mathcal{T}} \geq A\right] \\
&\leq \mathbf{E}_0\left[\frac{\frac{1}{2\sigma^2}\|y_{\mathcal{T}}\|_2^2}{I(\beta^*)}, S_{\mathcal{T}} \geq A\right] \\
&\leq \mathbf{E}_0\left[\frac{\frac{1}{2\sigma^2}\|y_{\mathcal{T}}\|_2^2}{I(\beta^*)}\right] \\
&\leq \frac{1}{2\sigma^2 I(\beta^*)} \mathbf{E}_0\left[\sum_{n=1}^{\mathcal{T}} \|y_n\|_2^4\right]^{1/2} \\
&\leq \frac{1}{2\sigma^2 I(\beta^*)} \mathbf{E}_0[\mathcal{T}]^{1/2} \mathbf{E}_0[\|y_n\|_2^4]^{1/2}.
\end{aligned}$$

Regarding the term  $\mathbf{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right]$ ,

$$\mathbf{E}_0\left[\frac{S_{\mathcal{T}}}{I(\beta^*)}, \mathcal{T} = q_A\right] \leq \frac{AP_0(\mathcal{T} = q_A)}{I(\beta^*)}.$$

□

LEMMA 12

$$V_2 \leq \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=1}^{\infty} \mathbf{E}_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} \mathbf{P}_0(\mathcal{T} \geq n)^{1/2}. \quad (3.28)$$

*Proof of Lemma Lemma 12* To bound  $V_2$ , define the statistics  $V_{2,t}$ :

$$V_{2,t} = \frac{\sum_{n=1}^t \frac{1}{2\sigma^2} \|X_n(\beta^* - \hat{\beta}_n)\|_2^2}{I(\beta^*)}.$$

Note that  $\hat{\beta}_t$  is  $\mathcal{F}_{t-1}$ -adapted, and  $(X_t)_{i,j} \sim N(0, 1)$  for all  $i, j$ . We have

$$\mathbf{E}_0[V_{2,t} | \mathcal{F}_{t-1}] = V_{2,t-1} + \frac{1}{2\sigma^2 I(\beta^*)} \|\beta^* - \hat{\beta}_t\|_2^2.$$

In other words,  $V_{2,t} - \sum_{n=1}^t \frac{1}{2\sigma^2 I(\beta^*)} \|\beta^* - \hat{\beta}_n\|_2^2$  is a martingale under  $P_0$ . Applying the optional sampling theorem obtains that:

$$\mathbf{E}_0[V_{2,\mathcal{T}} - \sum_{n=1}^{\mathcal{T}} \frac{1}{2\sigma^2 I(\beta^*)} \|\beta^* - \hat{\beta}_n\|_2^2] = 0,$$

which is essentially

$$V_2 = \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=1}^{\mathcal{T}} \mathbf{E}_0[\|\beta^* - \hat{\beta}_n\|_2^2].$$

We then have

$$\begin{aligned} V_2 &= \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=1}^{\infty} \mathbf{E}_0[\|\beta^* - \hat{\beta}_n\|_2^2 1_{\mathcal{T} \geq n}] \\ &\leq \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=1}^{\infty} \mathbf{E}_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} \mathbf{P}_0(\mathcal{T} \geq n)^{1/2}. \end{aligned}$$

□

LEMMA 13

$$V_3 = \mathbf{E}_0\left[\frac{\sum_{n=1}^{\mathcal{T}} \frac{1}{\sigma^2} (y_n - X_n \beta^*)^\top (X_n (\beta^* - \hat{\beta}_n))}{I(\beta^*)}\right].$$

*Proof of Lemma Lemma 13* To characterize  $V_3$ , define the statistics:

$$V_{3,t} = \frac{1}{\sigma^2 I(\beta^*)} \sum_{n=1}^t (y_n - X_n \beta^*)^\top (X_n (\beta^* - \hat{\beta}_n)).$$

Note that  $\hat{\beta}_t$  is  $\mathcal{F}_{t-1}$ -adapted, and it is easily seen that  $V_{3,t}$  is a martingale under  $P_0$ , and thus

$$V_3 = \mathbf{E}_0[V_{3,\mathcal{T}}] = 0.$$

□

Based on Lemma Lemma 10-Lemma 13, it suffices to bound the following terms to bound  $\mathbb{E}_0[\mathcal{T}]$ :

- $\mathbb{E}_0[\|y_n\|_2^4]^{1/2}$ ,
- $AP_0(\mathcal{T} = q_A)$ ,
- $\mathbb{E}_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2}$
- $P_0(\mathcal{T} \geq n)^{1/2}$ .

The lemmas below characterize these four terms.

LEMMA 14

$$\begin{aligned} \mathbb{E}_0[\|y_n\|_2^4]^{1/2} &\leq (2m(\|\beta^*\|_2^2 + \sigma^2)^2 + (m(\|\beta^*\|_2^2 + \sigma^2))^2)^{1/2} \\ &\leq 2m(\|\beta^*\|_2^2 + \sigma^2). \end{aligned} \tag{3.29}$$

*Proof of Lemma Lemma 14:* Note that under  $P_0$ ,  $y_n = X_n\beta^* + \epsilon_n$  where  $(X_n)_{i,j} \sim N(0, 1)$  and  $\epsilon_n \sim N(0)$ . We then have

$$\mathbb{E}_0[\|y_n\|_2^4] = (\mathbb{E}_0[\|y_n\|_2^2])^2 + \text{var}(\|y_n\|_2^2),$$

and

$$\begin{aligned} \mathbb{E}_0[\|y_n\|_2^2] &= m\mathbb{E}_0[y_{n,1}^2] \\ &= m(\mathbb{E}_0[y_{n,1}])^2 + m \text{var}(y_{n,1}) \\ &= m(\|\beta^*\|_2^2 + \sigma^2), \end{aligned}$$

and

$$\begin{aligned} \text{var}(\|y_n\|_2^2) &= m \text{var}(y_{n,1}^2) \\ &= 2m(\|\beta^*\|_2^2 + \sigma^2)^2. \end{aligned}$$

Note that the last step is due to the fact that  $y_{n,1} \sim N(0, \|\beta^*\|_2^2 + \sigma^2)$ .  $\square$

LEMMA 15 *Define*

$$\zeta = \frac{2}{c_L} \sigma \sqrt{\frac{\log p}{m}} \frac{1}{(n-1)^{1/4}}.$$

Note that the maximum iteration number  $L_n$  used in our procedure can be written as  $L_n = \frac{2}{\eta \zeta} \log \frac{1}{\alpha}$  under our constructions and definitions of  $\zeta$ . Assume the following conditions hold:

$$\begin{aligned} X_{1:n-1} &\sim \text{RIP}(s^*, \frac{C_\gamma}{\sqrt{s^*}(1 + \log \frac{\beta_{max}^*}{\zeta})}), \\ \|\frac{1}{m(n-1)} X_{1:n-1}^\top \epsilon_{1:n-1}\|_\infty &\leq \frac{2C_b}{c_L} \sigma \sqrt{\frac{\log p}{m}} \frac{1}{(n-1)^{1/4}}. \end{aligned} \quad (3.30)$$

Also assume the conditions listed in Theorem 1 hold, we have

$$\begin{aligned} |\hat{\beta}_{n,i} - \beta_i^*| &\leq 4\zeta + \alpha, \text{ for } i \in S, \\ |\hat{\beta}_{n,i} - \beta_i^*| &\leq \alpha, \text{ for } i \notin S, \end{aligned}$$

for  $m \geq c_{m1} \log p$ , where  $c_{m1}$  is a constant that is not related to  $\sigma, m, p$ .

*Proof of Lemma Lemma 15:* The proof is mainly based on Lemma Lemma 6. Let us first focus on the estimator  $\tilde{\beta}_n$  before shrinkage. Note that based on our constructions,  $\alpha_n = \alpha_0/\sqrt{n-1}$  can always be bounded by our tuning parameters  $c$  and thus if  $\|\tilde{\beta}_{n,i} - \beta_i^*\|_\infty \leq \alpha$ , for  $i \notin S$ , then  $\hat{\beta}_{n,i}$  will be 0 after shrinkage and conclusions hold. And similarly for  $i \in S$ , we always have  $|\tilde{\beta}_{n,i}| \geq |\beta_i^*| - |\tilde{\beta}_{n,i} - \beta_i^*| > c$  and  $|\tilde{\beta}_{n,i}| \leq |\beta_i^*| + |\tilde{\beta}_{n,i} - \beta_i^*| < C$ . In such cases  $\tilde{\beta}_{n,i} = \hat{\beta}_{n,i}$  for  $i \in S$  after shrinkage, which means that it suffices to prove the relationship for the initial estimator  $\tilde{\beta}_n$ .

The condition on step size  $\eta_n$  also holds based on the conditions in Theorem 1. For the sequence  $\{p_\ell\}_{1 \leq \ell \leq L_n}$ , it follows directly from Lemma Lemma 8 and the RIP condition on  $X_{1:n-1}$ . For the sequence  $\{b_\ell\}_{1 \leq \ell \leq L_n}$ , it can be done by using induction methods as in the proof of Theorem 1 in [74].  $\square$

LEMMA 16

$$\mathbb{E}_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} \leq c_1 \left\{ s\sigma^2 \frac{1}{\sqrt{n-1}} + \exp(-c_2\sqrt{n-1}) \right\},$$

for  $m \geq c_{l1} \log p$ , where  $c_{l1}, c_1, c_2$  are some constants that is not related  $m, p$ .

*Proof of Lemma Lemma 16:* If the conditions in Lemma Lemma 15 holds,

$$\|\beta^* - \hat{\beta}_n\|_2^4 \leq (c_1 s\sigma^2 \frac{1}{\sqrt{n-1}})^2.$$

Otherwise,

$$\|\beta^* - \hat{\beta}_n\|_2^4 \leq (sC^2\sigma^2 + \|\beta^*\|_2^2)^2,$$

for some constant  $c_3, C$ . This is because based on our constructions, there are only  $s$  non-zero coefficients and the absolute value of each non-zero coefficient is upper bounded by the tuning parameter  $C\sigma$ . Also, we observe from Lemma Lemma 7 and Lemma Lemma 9 that we can select  $m \geq c_{l1} \log p$  for some  $c_{l1} > 0$  such that the probability of the condition not holding is bounded by  $\exp(-2c_2\sqrt{n-1})$  for some constant  $c_2 > 0$ . We then have

$$\begin{aligned} & \mathbb{E}_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} \\ & \leq \left( (c_1 s\sigma^2 \frac{1}{\sqrt{n-1}})^2 + \exp(-2c_2\sqrt{n-1})(sC^2\sigma^2 + \|\beta^*\|_2^2)^2 \right)^{1/2} \\ & \leq c_1 \sigma^2 \left\{ s \frac{1}{\sqrt{n-1}} + \exp(-c_2\sqrt{n-1}) \right\}, \end{aligned}$$

for some constant  $c_1, c_2 > 0$ . Here for notational convenience we use the same notation  $c_1, c_2$  to denote the constants. □

LEMMA 17 For  $n \geq \frac{10A}{I(\beta^*)}$ ,

$$\mathbb{P}_0(\mathcal{T} \geq n) \leq c_4 \exp(-c_5(n - \frac{10A}{I(\beta^*)})) + \frac{c_6}{n^2}.$$

for some suitable constants  $c_4, c_5, c_6$ .

*Proof of Lemma Lemma 17* Note that  $\exp(-c_2\sqrt{n-1}) \leq c_7/(n-1)^3$  for some constant  $c_6$ . It is not hard to verify that the probability of conditions in (Equation 3.30) holding for all  $i \geq \frac{n}{2}$  and  $n \geq \frac{10A}{I(\beta^*)}$  is bounded by  $c_6/n^2$  for some constant  $c_7$ . Under such scenario, we observe that

$$\begin{aligned}
& \log \frac{f_{X_n \hat{\beta}_n}(y_n)}{f_0(y_n)} \\
& \geq \log \frac{f_{X_n \beta^*}(y_n)}{f_0(y_n)} - \left| \log \frac{f_{X_n \hat{\beta}_n}(y_n)}{f_{X_n \beta^*}(y_n)} \right| \\
& = \log \frac{f_{X_n \beta^*}(y_n)}{f_0(y_n)} - c_7 \|X_n(\hat{\beta}_n - \beta^*)\|_2^2 - c_8 |\epsilon_n^\top X_n(\hat{\beta}_n - \beta^*)| \\
& \geq \log \frac{f_{X_n \beta^*}(y_n)}{f_0(y_n)} - c_7 s \sum_{i=1}^m \sum_{j \in S} (X_n)_{i,j}^2 \sigma^2 \frac{\log p}{m} \frac{1}{\sqrt{n-1}} - c_8 \sum_{i=1}^s (\epsilon_n^\top X_n)_i \sigma \sqrt{\frac{\log p}{m}} \frac{1}{(n-1)^{1/4}} := z_n
\end{aligned}$$

We then have if  $A \geq c_{13}m$  and  $m \geq c_{13} \log p$  for some  $c_{13} > 0$ ,  $z_n$  is a sub-exponential variable with positive step size. The results follows immediately from the concentration of sub exponential variable.  $\square$

LEMMA 18 *For any threshold  $A$ , and sufficiently large constant  $c_{10}$  such that  $m \geq c_{10} \log p$ , we have*

$$P_0(S_{\mathcal{T}} \geq A) > c_9 > 0.$$

for some constant  $c_9$  that is not related to  $m, p, \sigma^2$ .

*Proof of Lemma Lemma 18:* We consider the ladder variable

$$\tau_- = \inf\{t > 0 : \sum_{n=1}^t z_n \leq 0\}.$$

If  $A \geq c_{13}m$  and  $m \geq c_{13} \log p$  for some  $c_{13} > 0$ ,  $z_n$  is a sub-exponential variable with positive step size. In such scenario, we have

$$P_0(\tau_- = \infty) > 0,$$

see Chapter 8.3 in [30]. In addition, we can select large enough  $m \geq c_{13} \log p$  such that the probability of conditions in Lemma Lemma 15 NOT holding for all  $n \geq 2$  is upper bounded by  $(1 - P_0(\tau_- < \infty))/2$ . We then have

$$\begin{aligned} P_0(S_{\mathcal{T}} < A) &\leq P_0(\tau_- < \infty) + \frac{1 - P_0(\tau_- < \infty)}{2} \\ &= \frac{1 + P_0(\tau_- < \infty)}{2} < 1. \end{aligned}$$

□

Based on the above lemmas, we conclude that for  $V_1$ ,

$$\begin{aligned} V_1 &\leq \frac{AP_0(S_{\mathcal{T}} \geq A)}{I(\beta^*)} + \frac{E_0[\mathcal{T}]^{1/2} E_0[\|y_n\|_2^4]^{1/2}}{2\sigma^2 I(\beta^*)} + \frac{AP_0(\mathcal{T} = q_A)}{I(\beta^*)} \\ &\leq \frac{AP_0(S_{\mathcal{T}} \geq A)}{I(\beta^*)} + \frac{E_0[\mathcal{T}]^{1/2} (2m(\|\beta^*\|_2^2 + \sigma^2))}{2\sigma^2 I(\beta^*)} + c_8, \end{aligned}$$

where the last step is obtained by bounding  $P_0(\mathcal{T} = q_A)$  based on Lemma Lemma 17. For

$V_2$ ,

$$\begin{aligned} V_2 &\leq \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=1}^{\infty} E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} P_0(\mathcal{T} \geq n)^{1/2} \\ &\leq \left( \sum_{n=1}^{10A/I(\beta^*)} \frac{E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} P_0(\mathcal{T} \geq n)^{1/2}}{2\sigma^2 I(\beta^*)} + \sum_{n=10A/I(\beta^*)+1}^{\infty} \frac{E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} P_0(\mathcal{T} \geq n)^{1/2}}{2\sigma^2 I(\beta^*)} \right) \\ &\leq \left( \sum_{n=1}^{10A/I(\beta^*)} \frac{E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2}}{2\sigma^2 I(\beta^*)} + \sum_{n=10A/I(\beta^*)+1}^{\infty} \frac{E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} P_0(\mathcal{T} \geq n)^{1/2}}{2\sigma^2 I(\beta^*)} \right) \\ &\leq \left( \sum_{n=1}^{10A/I(\beta^*)} \frac{c_1 \sigma^2 \{s \frac{1}{\sqrt{n-1}} + \exp(-c_2 \sqrt{n-1})\}}{2\sigma^2 I(\beta^*)} + \sum_{n=10A/I(\beta^*)+1}^{\infty} \frac{E_0[\|\beta^* - \hat{\beta}_n\|_2^4]^{1/2} P_0(\mathcal{T} \geq n)^{1/2}}{2\sigma^2 I(\beta^*)} \right) \\ &\leq \frac{C_2 \sqrt{A/I(\beta^*)}}{I(\beta^*)}, \end{aligned}$$

for some constant  $C_2 > 0$ . Concluding the results above, and note that  $P_0(S_{\mathcal{T}} \geq A) >$

$c_9 > 0$  for all  $A$ , we have

$$\begin{aligned}
& \mathbb{E}_0[T_{\text{IR}} \mid T_{\text{IR}} > 0] \\
&= \frac{\mathbb{E}_0[\mathcal{T}]}{\mathbb{P}_0(S_{\mathcal{T}} \geq A)} \\
&\leq \frac{A}{I(\beta^*)} + \frac{C_1 m}{I(\beta^*)} \sqrt{\mathbb{D}[T_{\text{IR}}]} + \frac{C_2 \sqrt{A/I(\beta^*)}}{I(\beta^*)} + C_3,
\end{aligned}$$

for some suitable constants  $C_1, C_2, C_3 > 0$ .

### 3.7.4 Proof of Detection Delay Relationship When Change Occurs at $\nu = k$

In this subsection, we consider the more complicated scenario when change occurs at general time  $\nu = k$  for  $k > 0$ . Note that when change-time  $\nu = k$ , observed data  $(X_t, y_t)_{t \leq k}$  are under the pre-change scenario, and observed data  $(X_t, y_t)_{t > k}$  are under the post-change scenario. Our aim is to characterize the detection delay

$$\mathbb{E}_k[T_{\text{IR}} - k \mid T_{\text{IR}} > k],$$

for general change-time  $\nu = k$ .

The main challenge as compared to the setting when  $\nu = 0$  is that we need to characterize  $T_{\text{IR}}$  under condition  $T_{\text{IR}} > k$ , which is highly non-trivial. To see this, consider the candidate change-time  $M(k+1)$  at time  $t = k+1$ . If  $M(k+1) < k$ , then the estimator  $\hat{\beta}_t$  for  $t > k$  might contain data from the pre-change scenario, which brings additional challenge to the theoretical analysis.

To analyze the detection delay under  $\mathbb{P}_k$ , we need the following definitions:

$$\begin{aligned}
X(k) &= [X_{M(k+1)+1}^\top, \dots, X_k^\top]^\top, \\
y(k) &= [y_{M(k+1)+1}^\top, \dots, y_k^\top]^\top,
\end{aligned}$$

and  $X(k), y(k) = 0$  if  $M(k+1) = k$ . Denote by  $W_k$  the detection statistics at change-time  $\nu = k$ . We define the following sequential test:

$$\mathcal{T}_{(k)} = \inf\{t-k, t \geq k+1 : \tilde{S}_t = \sum_{n=k+1}^t \log \frac{f_{X_n \hat{\beta}_n, \sigma^2 I}(y_n)}{f_{0, \sigma^2 I}(y_n)} \notin (-W_k, A-W_k) \text{ or } t-M(k+1) = q_A\}, \quad (3.31)$$

where the estimator  $\hat{\beta}_n$  is obtained by applying the implicit regularization algorithm to the following data  $(y_{M(k+1)+1:n-1}, X_{M(k+1)+1:n-1})$ :

$$y_{M(k+1)+1:n-1} = [y(k)^\top, y_{k+1}^\top, \dots, y_{n-1}^\top]^\top, \quad X_{M(k+1)+1:n-1} = [X(k)^\top, X_{k+1}^\top, \dots, X_{n-1}^\top]^\top,$$

with parameters including the initial value  $\alpha_n$ , iteration number  $L_n$ , and step size  $\eta_n$  defined

$$\begin{aligned} \alpha_n &= \frac{\alpha_0}{\sqrt{n - M(k+1) - 1}}, \\ \eta_n &= \eta_0, \\ L_n &= \frac{c_L (n - M(k+1) - 1)^{1/4}}{\eta_n \sigma \sqrt{\log p/m}} \log \frac{1}{\alpha_n}, \end{aligned}$$

In addition, the observed data  $(X_n, y_n)_{n \geq k+1}$  in sequential test (Equation 3.31) are independent and identically distributed (i.i.d.) from the linear model:

$$y_n = X_n \beta^* + \epsilon_n.$$

The detection delay when  $\nu = k$  can then be written as

$$\begin{aligned} \mathbb{E}_k[T_{\text{IR}} - k \mid T_{\text{IR}} > k] &= \mathbb{E}_k[\mathcal{T}_{(k)} + \mathcal{T}_1 + \dots + \mathcal{T}_w \mid T_{\text{IR}} > k] \\ &= \mathbb{E}_k[\mathcal{T}_{(k)} \mid T_{\text{IR}} > k] + \mathbb{P}_k(\tilde{S}_{\mathcal{T}_{(k)}+k} < A - W_k \mid T_{\text{IR}} > k) \mathbb{E}_0[T_{\text{IR}}]. \end{aligned}$$

We have already bounded  $\mathbb{E}_0[T_{\text{IR}}]$  in the previous subsection, and it suffices to bound  $\mathbb{E}_k[\mathcal{T}_{(k)} \mid T_{\text{IR}} > k]$  and  $\mathbb{P}_k(\tilde{S}_{\mathcal{T}_{(k)}+k} < A - W_k)$  in order to characterize  $\mathbb{E}_k[T_{\text{IR}} - k \mid T_{\text{IR}} > k]$ .

LEMMA 19 *The average run length of  $\mathcal{T}_{(k)}$  can be bounded by:*

$$\mathbf{E}_k[\mathcal{T}_{(k)} \mid T_{\text{IR}} > k] \leq V_4 + V_5 + V_6, \quad (3.32)$$

where

$$\begin{aligned} V_4 &= \mathbf{E}_k \left[ \frac{\tilde{S}_{\mathcal{T}_{(k)}+k}}{I(\beta^*)} \mid T_{\text{IR}} > k \right], \\ V_5 &= \mathbf{E}_k \left[ \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \frac{1}{2\sigma^2} \|X_n(\beta^* - \hat{\beta}_n)\|_2^2}{I(\beta^*)} \mid T_{\text{IR}} > k \right], \\ V_6 &= \mathbf{E}_k \left[ \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \frac{1}{\sigma^2} (y_n - X_n\beta^*)^\top (X_n(\beta^* - \hat{\beta}_n))}{I(\beta^*)} \mid T_{\text{IR}} > k \right]. \end{aligned} \quad (3.33)$$

*Proof of Lemma 4.1:* Applying Wald's equation obtains that

$$\begin{aligned} \mathbf{E}_k[\mathcal{T}_{(k)} \mid T_{\text{IR}} > k] &= \mathbf{E}_k \left[ \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \log \frac{f_{X_n\beta^*, \sigma^2 I}(y_n)}{f_{0, \sigma^2 I}(y_n)}}{I(\beta^*)} \mid T_{\text{IR}} > k \right] \\ &= \mathbf{E}_k \left[ \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \log \frac{f_{X_n\hat{\beta}_n, \sigma^2 I}(y_n)}{f_{0, \sigma^2 I}(y_n)}}{I(\beta^*)} + \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \log \frac{f_{X_n\beta^*, \sigma^2 I}(y_n)}{f_{X_n\hat{\beta}_n, \sigma^2 I}(y_n)}}{I(\beta^*)} \mid T_{\text{IR}} > k \right] \\ &= \mathbf{E}_k \left[ \frac{\tilde{S}_{\mathcal{T}_{(k)}+k}}{I(\beta^*)} + \frac{\sum_{n=k+1}^{k+\mathcal{T}_{(k)}} \frac{1}{2\sigma^2} (\|y_n - X_n\hat{\beta}_n\|_2^2 - \|y_n - X_n\beta^*\|_2^2)}{I(\beta^*)} \mid T_{\text{IR}} > k \right] \\ &= V_4 + V_5 + V_6, \end{aligned}$$

where  $V_4, V_5, V_6$  are defined in (Equation 3.26).

Following the similar proof as in Lemma Lemma 11-Lemma 13, it is easy to obtain that

$$\begin{aligned} V_4 &\leq \frac{AP_k(\tilde{S}_{\mathcal{T}_{(k)}+k} \geq A - W_k \mid T_{\text{IR}} > k)}{I(\beta^*)} + \frac{1}{2\sigma^2 I(\beta^*)} \mathbf{E}_k[\mathcal{T}_{(k)} \mid T_{\text{IR}} > k]^{1/2} \mathbf{E}_0[\|y_n\|_2^4]^{1/2} \\ &\quad + AP_k(\mathcal{T}_{(k)} + M(K+1) = q_A \mid T_{\text{IR}} > k), \\ V_5 &\leq \frac{1}{2\sigma^2 I(\beta^*)} \sum_{n=k+1}^{\infty} \mathbf{E}_k[\|\beta^* - \hat{\beta}_n\|_2^4 \mid T_{\text{IR}} > k]^{1/2} P_k(\mathcal{T}_{(k)} \geq n - k \mid T_{\text{IR}} > k)^{1/2}, \\ V_6 &= 0. \end{aligned}$$

Based on above derivations, it suffices to bound the following terms:

- $\mathbf{E}_0[\|y_n\|_2^4]^{1/2}$

- $\mathbb{P}_k(\mathcal{T}_{(k)} + M(K + 1) = q_A \mid T_{\text{IR}} > k)$ .
- $\mathbb{E}_k[\|\beta^* - \hat{\beta}_n\|_2^4 \mid T_{\text{IR}} > k]^{1/2}$ .
- $\mathbb{P}_k(\mathcal{T}_{(k)} \geq n - k \mid T_{\text{IR}} > k)^{1/2}$ .

The lemmas below characterize terms 2-4, and term 1 has already been bounded in the previous subsection. Let us start with a lemma on the candidate change time  $M(k + 1)$  :

LEMMA 20 *Consider the candidate change-time  $M(k + 1)$ , we have*

$$\mathbb{P}_k(T_{\text{IR}} > k \mid T_{\text{IR}} > M(k + 1)) \geq c_M > 0, \quad (3.34)$$

for some constant  $c_M$ .

*Proof of Lemma Lemma 20:*

$$\begin{aligned} \mathbb{P}_k(T_{\text{IR}} \leq k \mid T_{\text{IR}} > M(k + 1)) &= \mathbb{P}_k(T_{\text{IR}} \leq k - M(k + 1)) \\ &\leq \mathbb{P}_k(T_M \leq k - M(k + 1)) \\ &\leq \sum_{n=1}^{k-M(k+1)} \frac{n}{e^A} \\ &\leq \sum_{n=1}^{q_A} \frac{n}{e^A} = 1 - c_M < 1, \end{aligned}$$

due to the fact that  $q_A = O(e^{c_4 A})$  for some  $0 < c_4 < 1/2$ . Here  $T_M$  is a stopping time defined in Section A. □

LEMMA 21 *There exists suitable constants  $c_{M,1}, c_{M,2}$  such that*

$$\mathbb{P}_k(k + 1 - M(k + 1) \geq n \mid T_{\text{IR}} > k) \leq c_{M,1} e^{-c_{M,2} n}. \quad (3.35)$$

*Proof of Lemma Lemma 21:* We have

$$\begin{aligned}
& \mathbb{P}_k(k+1 - M(k+1) \geq n \mid T_{\text{IR}} > k) \\
&= \mathbb{P}_k(k+1 - M(k+1) \geq n \mid T_{\text{IR}} > k, T_{\text{IR}} > M(k+1)) \\
&\leq \frac{\mathbb{P}_k(k+1 - M(k+1) \geq n \mid T_{\text{IR}} > M(k+1))}{\mathbb{P}_k(T_{\text{IR}} > k \mid T_{\text{IR}} > M(k+1))} \\
&\leq \frac{1}{c_M} \mathbb{P}_k(k+1 - M(k+1) \geq n) \\
&= \frac{1}{c_M} \sum_{i=1}^{k+1-n} \mathbb{P}_\infty(M(k+1) = i, W_j > 0 \text{ for all } i < j < k+1) \\
&\leq c_{M,1} e^{-c_{M,2} n},
\end{aligned}$$

for some suitable constants  $c_{M,1}, c_{M,2}$ . □

To analyze  $\hat{\beta}_n$  when change occurs at general time  $\nu = k > 1$ , We need to rewrite the gradient descent update on  $(y = y_{M(k+1)+1:n-1}, X = X_{M(k+1)+1:n-1})$ . Note that

$$\begin{aligned}
& \frac{1}{m(n - M(k+1) - 1)} X^\top (X\beta_\ell - y) \\
&= \frac{n - k - 1}{n - M(k+1) - 1} \frac{1}{m(n - k - 1)} (X_{k+1:n-1}^\top X_{k+1:n-1} \beta_\ell + X(k)^\top X(k) \beta_\ell \\
&\quad - X_{k+1:n-1}^\top X_{k+1:n-1} \beta^* - X_{k+1:n-1}^\top \epsilon_{k+1:n-1} - X(k)^\top \epsilon(k))
\end{aligned}$$

As compared to the standard gradient descent update on clean data  $(y_{k+1:n-1}, X_{k+1:n-1})$  without the contamination of  $(y(k), X(k))$ , there is an additional term,

$$\frac{1}{m(n - k - 1)} (X(k)^\top X(k) \beta_\ell - X(k)^\top \epsilon(k)).$$

The key in bounding  $\hat{\beta}_n$  in sequential test (Equation 3.31) is to show that as  $n$  increases, the

additional term converges to 0 and thus will not affect the estimation of  $\beta^*$ . Now we define

$$\begin{aligned}
\beta_\ell^+ &= u_\ell \odot u_\ell, \\
\beta_\ell^- &= v_\ell \odot v_\ell, \\
s_\ell &= 1_{S^+} \odot \beta_\ell^+ - 1_{S^-} \odot \beta_\ell^-, \\
e_\ell &= 1_{S^c} \odot \beta_\ell + 1_{S^-} \odot \beta_\ell^+ - 1_{S^+} \odot \beta_\ell^-, \\
\tilde{b}_\ell &= \frac{1}{m(n-k-1)} (X_{k+1:n-1}^\top X_{k+1:n-1} e_\ell - X_{k+1:n-1}^\top \epsilon_{k+1:n-1} + X(k)^\top X(k) \beta_\ell - X(k)^\top \epsilon(k)), \\
\tilde{p}_\ell &= \left( \frac{1}{m(n-k-1)} X_{k+1:n-1}^\top X_{k+1:n-1} - I \right) (s_\ell - \beta^*).
\end{aligned}$$

Then the gradient descent update in implicit regularization algorithm can be rewritten as:

$$\begin{aligned}
\beta_\ell^+ &= \beta_\ell^- = \alpha^2 \mathbf{1}_{p \times 1}, \\
\beta_\ell &= \beta_\ell^+ - \beta_\ell^-, \\
\beta_{\ell+1}^+ &= \beta_\ell^+ \odot (1 - 4\tilde{\eta}(s_\ell - \beta^* + \tilde{p}_\ell + \tilde{b}_\ell))^2, \\
\beta_{\ell+1}^- &= \beta_\ell^- \odot (1 + 4\tilde{\eta}(s_\ell - \beta^* + \tilde{p}_\ell + \tilde{b}_\ell))^2,
\end{aligned}$$

with

$$\tilde{\eta} = \frac{n-k-1}{n-M(k+1)-1} \eta_0 \leq \eta_0,$$

for  $\ell = 1, \dots, L_n$ .

We then have the following lemma on the characterization of  $\tilde{\beta}_n$ .

**LEMMA 22** *Consider the sequential test defined in (Equation 3.31). Define*

$$\zeta = \frac{2}{c_L} \sigma \sqrt{\frac{\log p}{m}} \frac{1}{(n-M(k+1)-1)^{1/4}}.$$

Assume the following conditions hold:

$$\begin{aligned}
X_{k+1:n-1} &\sim \text{RIP}\left(s^*, \frac{C_\gamma}{\sqrt{s^*}(1 + \log \frac{\beta_{max}^*}{\zeta})}\right), \\
\left\| \frac{1}{m(n-1-k)} X_{k+1:n-1}^\top \epsilon_{k+1:n-1} \right\|_\infty &\leq \frac{1}{3} C_b \zeta, \\
\left\| \frac{1}{m(n-1-k)} X(k)^\top X(k) \beta_\ell \right\|_\infty &\leq \frac{1}{3} C_b \zeta, \\
\left\| \frac{1}{m(n-1-k)} X(k)^\top \epsilon(k) \right\|_\infty &\leq \frac{1}{3} C_b \zeta,
\end{aligned} \tag{3.36}$$

for some constants  $c_{e2}, c_{e3}, c_{e4}$ . Assume the conditions listed in Theorem 1 hold, we have

$$\begin{aligned}
|\hat{\beta}_{n,i} - \beta_i^*| &\leq 4\zeta + \alpha, \text{ for } i \in S, \\
|\hat{\beta}_{n,i} - \beta_i^*| &\leq \alpha, \text{ for } i \notin S,
\end{aligned}$$

for  $m \geq c_{m2} \log p$ , where  $c_{m2}$  is a constant that is not related to  $\sigma, m, p$ .

*Proof of Lemma Lemma 22:* Define  $\zeta'$  :

$$\zeta' = \zeta \frac{n - M(k+1) - 1}{n - k - 1} \geq \zeta,$$

and the iteration number can be written as  $L_n = \frac{2}{\zeta' \bar{\eta}} \log \frac{1}{\alpha}$ . The remaining proof is the same as the proof of Lemma Lemma 15 by using the induction.  $\square$

The remaining problem of bounding  $\hat{\beta}_n$  is how to characterize the probability of conditions (Equation 3.36) and we have the following lemma:

**LEMMA 23** *Denote the event that conditions (Equation 3.36) do not hold as  $A_{k,n}$ , and we have*

$$\mathbb{P}_k(A_{k,n} \mid T_{\text{IR}} > k) \leq \frac{c_{14}}{(n - k - 1)^3},$$

for some constant  $c_{14}$  and sufficiently large  $n \geq n_{\ell 2}$ .

*Proof of Lemma Lemma 23:* By using the concentration inequality and the fact that the time different between the candidate change time and the change time is exponentially decaying.

□

LEMMA 24

$$\mathbb{E}_k[\|\beta^* - \hat{\beta}_n\|_2^4 | T_{\text{IR}} > k]^{1/2} \leq c_{15} \left\{ s\sigma^2 \frac{1}{\sqrt{n-k-1}} + \frac{1}{(n-k-1)^3} \right\},$$

for  $m \geq c_{11} \log p$  and  $n \geq n_{\ell 2}$ , where  $c_{11}, c_1, c_2$  are some constants that is not related  $m, p$ .

*Proof of Lemma 24:* similar proof as the proof of Lemma Lemma 16.

LEMMA 25 For  $n \geq k + 1 + \frac{10A}{I(\beta^*)}$ ,

$$\mathbb{P}_k(\mathcal{T}_{(k)} \geq n - k | T_{\text{IR}} > k) \leq c_{16} \exp\left(-c_{17}\left(n - k - \frac{10A}{I(\beta^*)}\right)\right) + \frac{c_{18}}{(n-k-1)^2},$$

for some suitable constants  $c_{16}, c_{17}, c_{18}$ .

*Proof of Lemma Lemma 25:* the proof is similar to those of Lemma 17.

Concluding the above results, we conclude that

$$\mathbb{E}_k[T_{\text{IR}} - k | T_{\text{IR}} > k] \leq \frac{A}{I(\beta^*)} + \frac{C_1 m}{I(\beta^*)} \sqrt{\mathbb{D}[T_{\text{IR}}]} + \frac{C_2 \sqrt{A/I(\beta^*)}}{I(\beta^*)} + C_3,$$

and note that the above relationship holds for all change time  $k$ , and the relationship (18) in Theorem 4.3 is then proved. ■

## CHAPTER 4

### EFFICIENT ACTIVE QUICKEST DETECTION FOR MULTI-STREAM DATA

The goal of this chapter is to investigate three research topics under the setting of active quickest detection problem in multi-stream data. In section 4.1, we consider the scenario when the post-change distributions involve unknown parameters. An efficient greedy-cyclic-sampling-based quickest detection algorithm is proposed, and we show that our proposed algorithm is first-order asymptotically optimal in the sense of minimizing the detection delay under both false alarm and sampling control constraints.

In section 4.2, we extend our results in Chapter 2 to a more complicated scenario when  $r = s = 2$ . We combine three ideas to develop efficient active quickest detection algorithms: (1) win-stay, lose-switch sampling strategy; (2) local CUSUM statistics for local monitoring; and (3) the SUM-Shrinkage technique to fuse local statistics into a global decision. Theoretical guarantees on the ARL to false alarm and detection delay are provided.

In section 4.3, we revisited the TRAS algorithm proposed in [1] which incorporates a novel idea of compensation coefficients for unobserved data streams. We investigate the impact of compensation coefficients on TRAS algorithm. Our main contributions are two-folded. On the one hand, under the general setting, we prove that if compensation coefficient is larger than  $O(q/(p - q))$ , then the TRAS algorithm is not efficient in the sense of having a larger detection delay than the optimal lower bound. On the other hand, under the special case of  $q = s = 1$ , if compensation coefficient is smaller than  $O(1/(p - 1))$ , then the TRAS algorithm is efficient.

## 4.1 Asymptotic Optimality Theory for Active Quickest Detection with Unknown Post-Change Parameters

This paper is a collaborative work with Dr. Yajun Mei, and was published in the journal *Sequential Analysis* 2023.

### 4.1.1 Problem Formulation and Methods Review

Suppose there are  $p$  statistically independent local streams in a system, and denote with  $X_t^i$  the observation from the  $i$ -th local stream at time  $t$ , where  $i = 1, \dots, p$  and  $t = 1, 2, \dots$ . Let  $f_\theta(x) = \exp(\theta x - \psi(\theta))$  be the densities of an exponential family of distributions. Initially, the system is in the in-control state and the data stream  $\{X_t^i\}$  from the  $i$ th stream produces i.i.d. samples following the density  $f_{\theta_0}(X)$  (without loss of generality we assume the pre-change distributions  $f_{\theta_0}(X)$  are the same for all  $p$  streams). At some unknown time  $\tau$ , a triggering event occurs to the system and one of its  $p$  streams, say, the  $i$ -th, changes to i.i.d. samples following a new unknown density  $f_{\theta_i}(X)$ . Specifically, if the  $i$ -th data stream is affected, then

$$X_t^i \sim \begin{cases} \exp(\theta_0 x - \psi(\theta_0)), & \text{if } t \leq \tau \\ \exp(\theta_i x - \psi(\theta_i)), & \text{if } t > \tau, \end{cases} \quad (4.1)$$

while  $X_t^j \sim f_{\theta_0}(X)$  for  $j \neq i$  and all  $t > 0$ . Here we consider in detail the one-sided change-point problem and assume that there are two positive constants  $\delta$  and  $\eta$  that characterize the minimum and maximum magnitudes of the change (ie:  $\delta \leq \theta_i \leq \zeta$  for all  $i = 1, \dots, p$ ).

Let us now discuss the sampling control constraints. To be mathematically rigorous, define a sequence of *sampling indices*  $\{R_t\}$  with  $R_t \in \{1, \dots, p\}$ , where  $R_t$  is a random variable and  $\{R_t = m\}$  means that we will sample the  $m$ -th local streams at the time instant  $t + 1$ . Under our sampling constraint, we are allowed to access only one of these  $p$  local

streams at each time  $t$ , and this can be expressed as

$$\mathbb{1}_{\{R_t=1\}} + \cdots + \mathbb{1}_{\{R_t=p\}} = 1 \text{ for all times } t = 1, 2, \dots, \quad (4.2)$$

where  $\mathbb{1}_A$  denotes the indicator function of the event  $A$ .

Thus in the multi-stream quickest detection problem under sampling control, an algorithm consists two elements: one is the sampling policies, e.g., choose  $\{R_t\}$  for all time instants  $t$  subject to (Equation 4.49), and the other is the decision policy that is defined as the stopping time  $T$  with respect to the observed data sequence  $\{X_t^{i=R_{t-1}}\}_{t \geq 1}$ . Note that the sampling decision  $R_t$  depends only on those observed data up to time  $t - 1$ , and the stopping time  $\{T = t\}$  means that we raise an alarm at time  $t$ .

Following the classical minimax formulation for quickest detection proposed by Lorden [61], we are interested in finding a procedure  $(\{R_t\}_{t=1, \dots, \infty}, T)$  that minimizes the worst-case detection delay

$$D_i(T) = \sup_{t \geq 0} \text{ess sup } E_t^i[T - t | \mathcal{F}_t^i, T > t]. \quad (4.3)$$

for any  $i = 1, 2, \dots, p$  when the  $i$ -th local stream is affected by the change, subject to the average run length to false-alarm constraint

$$E_\infty[T] \geq \gamma > 1. \quad (4.4)$$

Here  $P_t^i(\cdot), E_t^i[\cdot]$  denote the probability measure and the corresponding expectation induced by the change occurring at the  $i$ -th process at time  $\tau = t$  and  $P_\infty(\cdot), E_\infty[\cdot]$  denote the probability measure and the corresponding expectation induced by the change occurring at  $\infty$ . In addition,  $\mathcal{F}_t^i$  is the  $\sigma$ -algebra generated by observed data/information on the  $i$ -th local stream up to time  $t$ .

Let us now review some existing research that is related to our problem and proposed

algorithm. Under the simplest scenario where we had the prior knowledge on the index of  $i$  where change occurs to and the full information of post-change distribution  $f_{\theta_i}$ , it is natural to always sample the  $i$ -th stream, i.e.,  $R_t \equiv i$  for all  $t$ , and the corresponding optimal procedure will be the well-known CUSUM procedure:

$$T_{\text{oracle}}(A) = \inf\{t > 0 : W_t^i \geq A\} \quad (4.5)$$

where  $W_t^i$  is the CUSUM statistics recursively defined as

$$W_t^i = \max\{W_{t-1}^i, 0\} + \log \frac{f_{\theta_i}(X_t^i)}{f_{\theta_0}(X_t^i)} \quad \text{for } t \geq 1 \quad (4.6)$$

and  $W_0^i = 0$ , see Moustakides [63]. Here we use  $T_{\text{oracle}}$  to emphasize that this CUSUM procedure makes an oracle assumption of known affected local stream and known post-change distribution.

Note that it is highly non-trivial to develop an efficient algorithm under our setup due to two challenges: the first is that the post-change distributions are unknown. This challenge has been addressed in [77] when monitoring  $p = 1$  local stream. Their idea is to estimate the post-change parameter by the average of recent observations after the candidate change-point, and update the local statistics as in the classical CUSUM statistic in (Equation 4.6). To be more specific, when monitoring the  $i$ -th local stream, the estimator  $\hat{\theta}_{t,i}$  for the  $i$ -th stream at time instant  $t$  can be defined as

$$\hat{\theta}_{t,i} = \min\{\eta, \max\{\delta, \psi_1\left(\frac{\sum_{n=g(t)}^{t-1} X_n^i}{t-g(t)}\right)\}\}, \quad (4.7)$$

where  $\psi_1(\cdot)$  is the inverse function of  $\psi'(\cdot)$  and  $g(t) \geq 0$  is the candidate change-point that depends only on  $X_1^i, \dots, X_{t-1}^i$ . The local statistics  $\widetilde{W}_t^i$  can then be defined as in the

recursion (Equation 4.6) with  $\theta_i = \hat{\theta}_{t,i}$ , which yields to Lorden and Pollak's procedure

$$T_{\text{LP}}(A) = \inf\{t > 0 : \widetilde{W}_t^i \geq A\}. \quad (4.8)$$

The second challenge is that the index of  $i$  of affected local stream is unknown, and thus it is unclear how to choose sampling indices  $\{R_t\}$  suitably so as to detect the change quickly. A naive sampling idea is to sample each local process periodically, i.e.,  $R_t = t \bmod p + 1$  for all time instants  $t = 1, 2, \dots$ , and each local stream is visited only once during each  $p$  time instants. Combining this naive sampling policy with Lorden and Pollak's procedure in (Equation 4.8) yields the following quickest detection algorithm:

$$T_{\text{naive}}(A) = \inf\{t > 0 : \max\{\widetilde{W}_t^1, \dots, \widetilde{W}_t^p\} > A\} \quad (4.9)$$

where  $\widetilde{W}_t^i (i = 1, \dots, p)$  is only updated when  $R_{t-1} = i$ . In the sequel we will refer (Equation 4.9) as the naive algorithm with the naive periodic sampling policy.

Clearly, the naive algorithm in (Equation 4.9) seems to be inefficient, as it might spend too much time on those  $p - 1$  unaffected local streams. To the best of our knowledge, no efficient algorithms have been developed in the quickest detection literature to simultaneously address these two challenges of unknown post-change distribution and unknown index of affected local streams.

#### 4.1.2 Our Proposed Algorithm

In this section we present an efficient quickest detection algorithm under the sampling control constraint (Equation 4.2). For simplicity, we still name our proposed algorithm as  $T_{\text{MSP}}$ . Our key idea is based on a myopic sampling policy where we continue to sample a local stream until we are confident to make one of the following two decisions: either a change has occurred or no changes have occurred. For better presentation, we split this section into two parts. In the first part, we propose our algorithm including both the my-

opic sampling policy and the stopping time  $T_{\text{MSP}}$ . In the second part we investigate the theoretical properties of our proposed algorithm  $T_{\text{MSP}}$ .

### *Algorithm Development*

There are three essential components in our proposed algorithm: (i) how to construct and update local statistics for all local streams including the streams not being sampled; (ii) how to decide which local stream to be sampled based on these local statistics, and (iii) when to raise a global alarm.

At the high level, our proposed algorithm exploits the prior knowledge that there is only one local stream which changes, and adopts the myopic sampling policy that samples the local process with the maximum local statistics. By choosing local statistics as those in Lorden and Pollak's procedure (Equation 4.8), the myopic sampling policy implies that we sample each local stream until we are confident to decide whether a change has occurred or not. If there is a local change, then we stop and raise a global alarm. If there is not a local change, then we switch to sample the next local stream. We repeat these steps until we raise a global alarm.

Let us now define our proposed algorithm under the sampling control constraint (Equation 4.2). We start with the construction and update of local statistics  $\widetilde{W}_t^i$ . There are two cases, depending on whether a local stream is sampled or not. If a local stream is not sampled, then we update  $\widetilde{W}_t^i$  as  $\max\{\widetilde{W}_{t-1}^i, 0\}$ . If a local stream is sampled, then we update  $\widetilde{W}_t^i$  as in Lorden and Pollak procedure (Equation 4.8). To be more specific, let  $\hat{\theta}_{t,i}$  be the estimate of the post-change parameter for the  $i$ -th stream at time  $t$ , which will be defined in a little bit later, and the local statistics  $\widetilde{W}_t^i$  can be defined recursively as

$$\begin{aligned} \widetilde{W}_t^i &= \max\{\widetilde{W}_{t-1}^i, 0\} + \mathbb{1}_{\{i=R_{t-1}\}} \log \frac{f_{\hat{\theta}_{t,i}}(X_t^i)}{f_{\theta_0}(X_t^i)} \\ &= \begin{cases} \max\{\widetilde{W}_{t-1}^i, 0\}, & \text{if } i \neq R_{t-1} \\ \max\{\widetilde{W}_{t-1}^i, 0\} + \log \frac{f_{\hat{\theta}_{t,i}}(X_t^i)}{f_{\theta_0}(X_t^i)}, & \text{if } i = R_{t-1}, \end{cases} \end{aligned} \quad (4.10)$$

with the initial values  $\widetilde{W}_0^i = 0$  for all  $i = 1, \dots, p$ .

Our proposed stopping time  $T_{\text{MSP}}$  is then defined as

$$T_{\text{MSP}}(A) = \inf \left\{ t > 0 : \max_{1 \leq i \leq p} \widetilde{W}_t^i \geq A \right\}, \quad (4.11)$$

for some pre-specified constant  $A$ .

As for the post-change parameter estimators  $\hat{\theta}_{t,i}$ , by (Equation 4.10), we only need to pay attention to the sampled local stream and thus adopt the same idea as in Lorden and Pollak's procedure (Equation 4.8). To be more concrete, at time instant  $t$ , assume that we sample at the  $i$ -th stream, and denote by  $p(t)$  the total time instants in which we have consecutively sampled at the  $i$ -th stream. In other words, we observed data from the  $i$ -th stream during the time period of  $t - p(t) + 1$  to  $t$ . Here we propose to estimate the post-change parameter based on the observed data from the  $i$ -th stream during the time period of  $t - p(t) + 1$  to  $t - 1$ , as we save the data at the time instant  $t$  for quickest detection, not for parameter estimation. pathematically, we can define the estimator  $\hat{\theta}_{t,i}$  based on the method of moments (pOp) estimator of the distribution in (Equation 4.48):

$$\hat{\theta}_{t,i} = \min \left\{ \zeta, \max \left\{ \delta, \psi^{-1} \left( \frac{\sum_{j=t-p(t)+1}^{t-1} X_j^i}{p(t) - 1} \right) \right\} \right\} \quad (4.12)$$

with  $0/0 = -\infty$ .

It remains to define the sampling policies  $\{R_t\}$ . At the high level, the sampling policy  $\{R_t\}$  at time instant  $t$  can be defined by the local statistics  $\widetilde{W}_t^i$  at time instant  $t$ , which in turn depends on the sampling policy  $\{R_{t-1}\}$  at time instant  $t - 1$ . Here our proposed algorithm keeps sampling on the  $i$ -th stream as long as the local statistics  $\widetilde{W}_t^i > 0$  and switches to the next stream when  $\widetilde{W}_t^i \leq 0$ . pathematically, in our proposed algorithm, the sampling policy  $\{R_t\}$  at time instant  $t$  can be defined by the local statistics  $\widetilde{W}_t^i$  at time

instant  $t$ :

$$R_t = \begin{cases} R_{t-1} & \text{if } \widetilde{W}_t^{R_{t-1}} > 0 \\ R_{t-1} \bmod p + 1 & \text{if } \widetilde{W}_t^{R_{t-1}} \leq 0. \end{cases} \quad (4.13)$$

with initial values  $R_0 = 1$ .

In summary, our proposed algorithm can be summarized as follows:

**Step 1:** Sample the first stream until  $\widetilde{W}_t^1 \notin (0, A)$ . If  $\widetilde{W}_t^1 \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t^1 \leq 0$ , we switch to sampling the second stream.

**Step 2:** Sample the second stream until  $\widetilde{W}_t^2 \notin (0, A)$ . If  $\widetilde{W}_t^2 \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t^2 \leq 0$ , we switch to sampling the third stream.

$\vdots$

**Step  $p$ :** Sample the  $p$ -th stream until  $\widetilde{W}_t^p \notin (0, A)$ . If  $\widetilde{W}_t^p \geq A$ , we stop sampling and raise a global alarm; otherwise if  $\widetilde{W}_t^p \leq 0$ , we switch to sampling the first stream.

**Step  $p + 1$ :** Go back to Step 1.

### *Theoretical Properties*

In this subsection, we will investigate the theoretical properties of our proposed algorithm  $\mathbb{T}_{\text{MSP}}$  in (Equation 4.11). First, we need to make necessary standard assumptions:

$$\begin{aligned} \text{(A1): } I_\infty^i &= \int \log \frac{f_{\theta_0}(X)}{f_{\theta_i}(X)} f_{\theta_0}(X) dX > 0, \\ I_0^i &= \int \log \frac{f_{\theta_i}(X)}{f_{\theta_0}(X)} f_{\theta_i}(X) dX > 0, \\ \text{(A2): } J_\infty^i &= \int \left( \log \frac{f_{\theta_0}(X)}{f_{\theta_i}(X)} \right)^2 f_{\theta_0}(X) dX > 0, \\ J_0^i &= \int \left( \log \frac{f_{\theta_i}(X)}{f_{\theta_0}(X)} \right)^2 f_{\theta_i}(X) dX > 0, \end{aligned}$$

Now we are ready to present the theoretical properties of our proposed algorithm  $\mathbb{T}_{\text{MSP}}$  in (Equation 4.11). The main results are summarized in the following theorem and its corollary, whose high-level proofs are presented in the appendix.

**THEOREM 3** For our proposed algorithm  $\mathbb{T}_{\text{MSP}}$  in (Equation 4.11), we have

$$\mathbb{E}_\infty[\mathbb{T}_{\text{MSP}}] \geq e^A. \quad (4.14)$$

Moreover, its detection delay satisfies

$$D_i(\mathbb{T}_{\text{MSP}}) \leq \frac{A}{l_0^i} + C_0 \log A + C_1 \sqrt{A} + C_2(p-1) \quad (4.15)$$

as  $A \rightarrow \infty$  for any  $i \in 1, \dots, p$ . Here  $C_0, C_1, C_2$  are constants depending only on the distributions, not on  $A$  and  $p$ .

**COROLLARY 2** Let  $A = \log \gamma$ , then our proposed algorithm  $\mathbb{T}_{\text{MSP}}(A)$  in (Equation 4.11) satisfies both the false alarm constraint in (Equation 4.4) and the sampling control constraint in (Equation 4.2). Moreover, for each  $i = 1, \dots, p$ , its detection delay satisfies

$$0 \leq D_i(\mathbb{T}_{\text{MSP}}) - D_i^{\text{orc}} \leq C_0 \log \log \gamma + C_1 \sqrt{\log \gamma} + C_2(p-1), \quad (4.16)$$

where  $D_i^{\text{orc}}$  is the oracle detection delay achieved by the classical CUSUM procedure in (Equation 4.5) for monitoring a change in distribution of the  $i$ -th local stream under the oracle assumption that the index of affected stream and the post-change parameters are completely specified:

$$D_i^{\text{orc}} = \frac{\log \gamma}{l_0^i} + C_3 \quad (4.17)$$

and  $C_3$  is a constant that only depends on the distributions, not on  $\gamma$  and  $p$ .

It is useful to add some remarks. Note that relationship (Equation 4.16) holds for every  $p$  and  $\gamma$ . On one hand, our proposed algorithm  $\mathbb{T}_{\text{MSP}}$  has the same detection delay of the oracle or CUSUM procedure up to  $O(\sqrt{\log \gamma})$  when  $p$  is fixed as  $\gamma \rightarrow \infty$ , or when  $p = O(\sqrt{\log \gamma})$ . On the other hand when  $p$  is large but  $\gamma$  is moderately large, the additional term  $C_0 \log \log \gamma + C_1 \sqrt{\log \gamma} + C_2 p$  can be comparable to or even larger than  $D_i^{\text{orc}}$ , and

thus the performance of our proposed algorithm will be much worse than the oracle or CUSUM procedure. This is not surprising for high-dimensional setting, as the sampling control in (Equation 4.2) is too restrictive for large  $p$  and we should not be able to detect the change quickly if we only sample one out of  $p$  local streams at each time instant. In other contexts, we can evaluate the constants  $C_0, C_1$  and  $C_2$  to see the effects of the dimension  $p$  on the performance of our proposed algorithm, also see [18] for similar contexts. It remains an open problem to develop a general asymptotic optimality theory for high-dimensional streams under the sampling control.

### 4.1.3 Numerical Results

In this section, we conduct Monte Carlo simulations to demonstrate the performance properties of our proposed algorithm  $T_{\text{MSP}}$  in (Equation 4.11). In our simulation, we consider two choices on the number  $p$  of local streams:  $p = 2$  or  $p = 10$ . For each choice of  $p$  streams, we consider two different kinds of distributions, one is normal distribution and the other is exponential distribution. Due to the page constraints, here we only present the homogeneous setting (i.e.,  $f_{\theta_i} = f$  and  $f_{\phi_i} = g$  for any  $i = 1, \dots, p$ ).

- Mean shift in Normal Distribution from 0 to  $\mu \geq 0.5$ .
- Mean shift in Exponential Distribution from 1 to  $\lambda \geq 2$ .

In each case, we set the false alarm constraint  $\gamma = 50,000$ . For our proposed algorithm  $T_{\text{MSP}}(A)$  and the naive method  $T_{\text{naive}}(A)$  in (Equation 4.9), we first use the bisection method to find suitable threshold  $A$  to attain the false alarm constraint, and then simulate the worst-case detection delay under different post-change scenarios where the change occurs to the  $p$ -th stream (as our algorithm start to sample at the first stream).

Tables Table 4.1 and Table 4.2 report the detection delay of our proposed algorithm  $T_{\text{MSP}}$  and the naive method  $T_{\text{naive}}$  in (Equation 4.9). In addition, we also report the oracle detection delay of the CUSUM procedure. All numerical results are based on 50,000

Table 4.2: Comparison of Detection Delay in Exponential Distribution

| $\gamma = 50000$ |        | $M = 2$ |                  | $p = 10$ |                  |
|------------------|--------|---------|------------------|----------|------------------|
| $\mu$            | Oracle | Naive   | $T_{\text{MSP}}$ | Naive    | $T_{\text{MSP}}$ |
| 0.5              | 61.87  | 144.01  | 90.56            | 701.23   | 234.10           |
| 0.75             | 29.62  | 64.13   | 39.07            | 308.52   | 100.06           |
| 1.0              | 17.20  | 36.45   | 22.65            | 174.67   | 60.85            |
| 1.25             | 11.35  | 23.40   | 15.46            | 112.12   | 43.33            |
| 1.5              | 7.93   | 16.60   | 11.21            | 80.28    | 35.03            |

| $\gamma = 50000$ |        | $M = 2$ |                  | $p = 10$ |                  |
|------------------|--------|---------|------------------|----------|------------------|
| $\lambda$        | Oracle | Naive   | $T_{\text{MSP}}$ | Naive    | $T_{\text{MSP}}$ |
| 2.0              | 26.78  | 57.50   | 39.62            | 286.09   | 101.60           |
| 2.25             | 19.39  | 41.58   | 28.78            | 206.86   | 76.52            |
| 2.5              | 15.18  | 32.17   | 22.49            | 159.72   | 62.83            |
| 2.75             | 12.06  | 25.59   | 17.40            | 126.66   | 54.21            |
| 3.0              | 9.84   | 21.49   | 14.76            | 105.10   | 48.05            |

Monte Carlo runs. From the tables, it is clear that our proposed algorithm  $T_{\text{MSP}}$  is much better than the naive method  $T_{\text{base}}$  and can reduce the detection delay by at least 25% when  $p = 2$  and 50% when  $p = 10$ . In other words, as compared to the naive periodic sampling, our proposed myopic sampling policy can lead to a significantly improvement on the detection delay performance.

Moreover, our results also shows that as the dimension  $M$  increases from  $p = 2$  to  $p = 10$ , the detection delays of both our proposed algorithm  $T_{\text{MSP}}$  and the naive method  $T_{\text{naive}}$  in (Equation 4.9) increase significantly. We conjecture that the oracle bound of the CUSUM procedure is unattainable for high-dimensional monitoring under the sampling control, but we are unable to provide a rigorous proof.

#### 4.1.4 Proof of the Main Theorem

In this section, we provide the rigorous proof for Theorem 3. There are two elements in the proof: in the first part we study the false-alarm relationship (Equation 4.14) and in the second part we study the detection delay relationship (Equation 4.15).

### *False-Alarm Relationship*

Let us begin with the proof of false-alarm relationship in (Equation 4.14), which is the easier one. The key idea in proving (Equation 4.14) is to construct a new stopping time  $\mathbb{T}_*(A)$  such that  $\mathbb{E}_\infty[\mathbb{T}_*(A)] \leq \mathbb{E}_\infty[\mathbb{T}_{\text{MSP}}(A)]$  for all threshold  $A$  and prove that  $\mathbb{T}_*(A)$  satisfies the false-alarm constraint in (Equation 4.4). To be more specific, we define  $\mathbb{T}_*(A)$  :

$$\mathbb{T}_*(A) = \inf \left\{ t \geq 1 : S_t^* = \sum_{\ell=1}^t \prod_{m=\ell}^t \frac{f_{\hat{\theta}_{m,R_m}}(X_m^{R_m})}{f_0(X_m^{R_m})} \geq e^A \right\}. \quad (4.18)$$

It is easy to verify that  $\log S_t^* \geq \max\{\widetilde{W}_t^1, \dots, \widetilde{W}_t^p\}$  and thus  $\mathbb{E}_\infty[\mathbb{T}_*(A)] \leq \mathbb{E}_\infty[\mathbb{T}_{\text{MSP}}(A)]$  holds for any threshold  $A$ . Note that  $\{S_t^* - t\}$  is a  $\mathbb{P}_\infty$ -martingale with zero expectation, applying the optional sampling theorem obtains that

$$\mathbb{E}_\infty[\mathbb{T}_{\text{MSP}}] \geq \mathbb{E}_\infty[\mathbb{T}_*] = \mathbb{E}_\infty[S_{\mathbb{T}_*}^*] \geq e^A. \quad (4.19)$$

□

### *Detection Delay Relationship*

Consider now the detection delay relationship in (Equation 4.15). Below we write the log likelihood ratio as  $\log(f_{\theta_j}(X)/f_{\theta_0}(X)) = (\theta_j - \theta_0)X - \psi(\theta_j) + \psi(\theta_0)$ . The main idea in proving (Equation 4.15) is to present an equivalent definition of our proposed algorithm  $\mathbb{T}_{\text{MSP}}(A)$  by the sequential tests. For each  $j(j = 1, \dots, p)$ , a prototype sequential test applied to the  $j$ -th local stream is defined as

$$\mathbb{T}^j = \inf \left\{ t \geq 1 : S_t^i = \sum_{\ell=1}^t ((\hat{\theta}_\ell - \theta_0)X_\ell^j - \psi(\hat{\theta}_\ell) + \psi(\theta_0)) \notin (0, A) \right\}. \quad (4.20)$$

with the estimator  $\hat{\theta}_\ell$ :

$$\hat{\theta}_\ell = \min\{\zeta, \max\{\delta, \psi_1\left(\frac{X_1^j + \dots + X_{\ell-1}^j}{\ell - 1}\right)\}\}. \quad (4.21)$$

where  $\psi_1$  is the inverse function of  $\psi'$ . Here we use the same notation  $\hat{\theta}_\ell$  for different  $T^j$  and omit the subscript.

Define the sequence  $\{T_m\}, m = 1, 2, \dots$ , of sequential tests applied to each local stream. Then it is clear that each  $T_m$  has the same distribution as a particular prototype SPRT  $T^j$  with  $\ell$  and  $j$  related through the equation  $j = (m - 1) \bmod p + 1$ . In addition, if we define (a stopping time)  $K$  to be the first time the sequential test  $T_m$  of the local stream being tested crosses the upper boundary  $A$ , then our proposed stopping time  $T_{\text{MSP}}$  can be written as the sum

$$T_{\text{MSP}} = T_1 + T_2 + \dots + T_K = \sum_{m=1}^K T_\ell. \quad (4.22)$$

Now assume that the change occurs to the  $i$ -th local stream at change time  $\tau = k$ . We can then divide our discussions into the following two scenarios.

- Scenario I: we are not monitoring the  $i$ -th stream at time  $k$ .
- Scenario II: we switched to monitor the  $i$ -th stream before time  $k$ , and are still monitoring it at time  $k$ .

We note that analysis of detection delay under these two scenarios are completely different. Under Scenario II, the sequential test  $T^i$  on the affected stream  $i$  consists of two parts: the first part is under the pre-change distribution  $f_0(X)$  and the second part is under the post-change distribution  $f_{\theta_i}(X)$ . The estimator in the second part includes the observations in the first part, which brings additional challenges to our analysis. Below for the better presentation, we consider Scenario I and Scenario II separately.

### **Detection Delay under Scenario I:**

Let us first consider Scenario I, which is the easier one. Without loss of generality, we

assume the change occurs to the  $p$ -th local stream at change time  $\tau = 0$ , since our test needs to go through all other  $p - 1$  local streams before sampling the one that has changed and this will clearly generate the worst case detection delay. Hence, it suffices to show that  $D_p(\mathsf{T}_{\text{MSP}}) = \mathbb{E}_0^p[\mathsf{T}_{\text{MSP}}]$  satisfies relationship. Based on the cyclic nature of our proposed algorithm, by (Equation 4.22), we have

$$\begin{aligned} D_p[\mathsf{T}_{\text{MSP}}(A)] &= \mathbb{E}_0^p \left[ \sum_{\ell=1}^K \mathsf{T}_\ell \right] = \sum_{m=1}^{\infty} \mathbb{E}_0^p[\mathsf{T}_m] \mathbb{P}_0^p(K \geq m) \\ &= \Omega_p + \omega_p \Omega_p + \omega_p^2 \Omega_p + \dots \\ &= \frac{\Omega_p}{1 - \omega_p}, \end{aligned}$$

where

$$\begin{aligned} \omega_p &= \beta_p \prod_{j=1}^{p-1} (1 - \alpha_j) \\ \Omega_p &= \mathbb{E}_0^p[\mathsf{T}^1] + \mathbb{E}_0^p[\mathsf{T}^2](1 - \alpha_1) + \dots + \dots + \mathbb{E}_0^p[\mathsf{T}^p] \prod_{j=1}^{p-1} (1 - \alpha_j) \\ &= \mathbb{E}_\infty[\mathsf{T}^1] + \mathbb{E}_\infty[\mathsf{T}^2](1 - \alpha_1) + \dots + \dots + \mathbb{E}_0^p[\mathsf{T}^p] \prod_{j=1}^{p-1} (1 - \alpha_j) \end{aligned}$$

where  $\beta_p$  denotes the type-II error probability of the SPRT  $\mathsf{T}^p$  and  $\alpha_i$  denotes the type-I error probability of the SPRT  $\mathsf{T}^i$  ( $i = 1, \dots, p - 1$ ). Since  $0 \leq \alpha_i, \beta_i \leq 1$ , we have the following upper bound

$$D_p(\mathsf{T}_{\text{MSP}}(A)) \leq \frac{\mathbb{E}_0^p[\mathsf{T}^p]}{1 - \beta_p} + \frac{1}{1 - \beta_p} \sum_{j=1}^{p-1} \mathbb{E}_\infty[\mathsf{T}^j]. \quad (4.23)$$

The detection delay relationship can then be analyzed through investigating the SPRTs  $\mathsf{T}^1, \dots, \mathsf{T}^p$ . Below we break up the proofs under Scenario I into a series of lemmas.

**LEMMA 26** *For any  $j = 1, \dots, p$ , there exists constants  $C_{11}, C_{12}, C_{13}, C_{14}$  such that for*

some  $\eta > 0$  and all time  $t$ ,

$$P_0^j(|\frac{\sum_{\ell=1}^t X_\ell^j}{t} - \psi'(\theta_j)| > \eta) < C_{11}e^{-C_{12}\eta^2 t}, \quad (4.24)$$

$$E_0^j[(\frac{\sum_{\ell=1}^t X_\ell^j}{t} - \psi'(\theta_j))^4] < \frac{C_{13}}{t^2}. \quad (4.25)$$

The same relationship holds when  $\psi'(\theta_j)$  replaced by  $\psi'(\theta_0)$  and  $P_0^j, E_0^j$  replaced by  $P_\infty, E_\infty$ .

*Proof:* Standard large deviations arguments.

LEMMA 27 For the sequential test  $T^p$  in (Equation 4.20), we have:

$$E_0^p[T^p] \leq \frac{(1 - \beta_p)A}{I_0^p} + x_1 + x_2, \quad (4.26)$$

where

$$x_1 \leq C_{21}(E_0^p[\sum_{\ell=1}^{T^p} (\hat{\theta}_\ell - \theta_p)^2])^{1/2} + C_{22}(E_0^p[T^p])^{1/2} \quad (4.27)$$

$$x_2 \leq C_{23}E_0^p[\sum_{\ell=1}^{T^p} (\hat{\theta}_\ell - \theta_p)^2] + C_{24}E_0^p[\sum_{\ell=1}^{T^p} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2] \quad (4.28)$$

for some constants  $C_{21}, C_{22}, C_{23}, C_{24}$  that only depends on the parameter  $\theta_p$  but not on the threshold  $A$ .

*Proof:* Let  $\rho_m = \min(m, T^p)$ , then  $\rho_m$  is bounded and  $\rho_m \rightarrow T^p$  as  $m \rightarrow \infty$ . Applying

Wald's equation and optional stopping theorem obtains that

$$\begin{aligned}
\mathbb{E}_0^p[\rho_m] &= \frac{\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} (\theta_p - \theta_0) X_{\ell,p} - \psi(\theta_p) + \psi(\theta_0)]}{I_0^p} \\
&= \frac{\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\hat{\theta}_\ell - \theta_0) X_\ell^p - \psi(\hat{\theta}_\ell) + \psi(\theta_0))]}{I_0^p} \\
&\quad + \frac{\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell) X_\ell^p - \psi(\theta_p) + \psi(\hat{\theta}_\ell))]}{I_0^p} \\
&= \frac{\mathbb{E}_0^p[S_{\rho_m}^p]}{I_0^p} + \frac{\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell) \psi'(\theta_p) - \psi(\theta_p) + \psi(\hat{\theta}_\ell))]}{I_0^p} \\
&\leq \frac{\mathbb{P}_0^p(S_{\rho_m}^p > 0)A}{I_0^p} + \frac{\mathbb{P}_0^p(S_{\rho_m}^p > 0)\mathbb{E}_0^p[S_{\rho_m}^p - A | S_{\rho_m}^p > 0]}{I_0^p} \\
&\quad + \frac{\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell) \psi'(\theta_p) - \psi(\theta_p) + \psi(\hat{\theta}_\ell))]}{I_0^p} \\
&= \frac{\mathbb{P}_0^p(S_{\rho_m}^p > 0)A}{I_0^p} + x_{1m} + x_{2m}.
\end{aligned}$$

where  $S_{\rho_m}^p$  is the local summary statistics defined in (Equation 4.20). For  $x_{1m}$ , using optional stopping theorem implies that

$$\begin{aligned}
x_{1m} &\leq \frac{\mathbb{E}_0^p[(\hat{\theta}_{\rho_m} - \theta_0) X_{\rho_m}^p - \psi(\hat{\theta}_{\rho_m}) + \psi(\theta_0)]}{I_0^p} \\
&\leq \frac{\mathbb{E}_0^p[|\hat{\theta}_{\rho_m} - \theta_p| X_{\rho_m}^p]}{I_0^p} + \frac{\mathbb{E}_0^p[|\theta_0 - \theta_p| X_{\rho_m}^p]}{I_0^p} + \frac{\mathbb{E}_0^p[|\psi(\theta_p) - \psi(\hat{\theta}_\ell)|]}{I_0^p} \\
&\leq \frac{(\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} (\hat{\theta}_\ell - \theta_p)^2 (X_\ell^p)^2])^{1/2}}{I_0^p} + \frac{(\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} (\theta_0 - \theta_p)^2 (X_\ell^p)^2])^{1/2}}{I_0^p} \\
&\quad + \frac{\mathbb{E}_0^p[|\psi(\theta_p) - \psi(\hat{\theta}_{\rho_m})|]}{I_0^p} \\
&\leq C_{21}(\mathbb{E}_0^p[\sum_{\ell=1}^{\rho_m} (\hat{\theta}_\ell - \theta_p)^2])^{1/2} + C_{22}(\mathbb{E}_0^p[\rho_m])^{1/2}.
\end{aligned}$$

Note that  $\hat{\theta}_\ell$  is univerrally bounded by  $\delta, \zeta$  and thus  $\mathbb{E}_0^p[|\psi(\theta_p) - \psi(\hat{\theta}_\ell)|]$  is bounded. Constants  $C_{21}$  and  $C_{22}$  are suitably chosen such that the term  $\mathbb{E}_0^p[|\psi(\theta_p) - \psi(\hat{\theta}_{\rho_m})|]/I_0^p$  is omitted.

For the  $x_{2m}$  term, as  $\psi''(\theta) > 0$  for all  $\theta$ , we have:

$$\begin{aligned}
x_{2m} &= \frac{\mathbf{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell)\psi'(\theta_p) - \psi(\theta_p) + \psi(\hat{\theta}_\ell))]}{I_0^p} \\
&\leq \frac{\mathbf{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell)(\psi'(\theta_p) - \psi'(\theta_*)))]}{I_0^p} \\
&\leq \frac{\mathbf{E}_0^p[\sum_{\ell=1}^{\rho_m} ((\theta_p - \hat{\theta}_\ell)(\psi'(\theta_p) - \psi'(\theta_\ell)))]}{I_0^p} \\
&\leq C_{23}\mathbf{E}_0^p[\sum_{\ell=1}^{\rho_m} (\hat{\theta}_\ell - \theta_p)^2] + C_{24}\mathbf{E}_0^p[\sum_{\ell=1}^{\rho_m} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2]
\end{aligned}$$

for some constants  $C_{23}, C_{24}$  and some parameter  $\theta_*$  between  $\hat{\theta}_\ell$  and  $\theta_p$ . Lemma 27 is established through letting  $m \rightarrow \infty$  and the following lemmas which ensure that right hand side of Equation 4.26 is bounded.  $\square$

LEMMA 28 *There exist constants  $C_{31}$  such that*

$$\mathbf{E}_\infty[\mathbb{T}^j] < C_{31}, \quad (4.29)$$

for all  $j = 1, \dots, p$ .

*Proof:* Lorden and Pollak[77] proved that:

$$\mathbf{P}_\infty(\mathbb{T}^j \geq \ell) \leq C_{32}e^{-C_{33}\ell}, \quad (4.30)$$

for all time  $\ell$  and all  $j = 1, \dots, p$  and some constants  $C_{32}, C_{33}$ . Thus we have:

$$\mathbf{E}_\infty[\mathbb{T}^j] = \sum_{\ell=1}^{\infty} \mathbf{P}_\infty(\mathbb{T}^j \geq \ell) \leq C_{31}. \quad (4.31)$$

LEMMA 29 *For any threshold  $A$ , the type-II error probability  $\beta_p$  of the sequential test  $\mathbb{T}^p(A)$  satisfies the lower bound*

$$\beta_p \leq C_{41}, \quad (4.32)$$

where  $0 < C_{41} < 1$  is a constant that does not depend on the threshold  $A$ .

*Proof:* see Lemma 5.8 in [77].

LEMMA 30 *There exists constants  $C_{51}, C_{52}$  such that for sufficient large  $A \geq A_0$  and every  $r = 1, 2, \dots$ ,*

$$\mathbb{P}_0^p(\mathbb{T}^p > \frac{10A}{l_0^p} + r) \leq C_{51}e^{-C_{52}r}. \quad (4.33)$$

*Proof:* Let  $s = \lceil \frac{5A}{l_0^p} \rceil$ , we define the following event for some constant  $\eta$ ,

$$E = \{|\hat{\theta}_\ell - \theta_p| < \lambda \text{ for all } \ell \geq s + \frac{r}{2}\}$$

From relationship (Equation 4.24) we have  $\mathbb{P}_0^p(E^C) \leq C_{53}e^{-C_{54}r}$  for some constant  $C_{53}, C_{54}$ .

Similarly as in Lemma Lemma 29, we consider a new stopping time

$$N_*(A) = \inf\{t > 0 : \sum_{\ell=1}^t Y_\ell \geq A\}. \quad (4.34)$$

It is easy to verify that if  $\mathbb{T}^p(A) > 10A/l_0^p + r$  then  $N_*(A) > 10A/l_0^p + r$ . Moreover, it is straightforward to see that  $\mathbb{P}_0^p(N_*(A) > 10A/l_0^p + r)$  is bounded exponentially in  $r$  and we can prove that there exists constants  $C_{51}, C_{52}$  such that

$$\begin{aligned} \mathbb{P}_0^p(\mathbb{T}^p > \frac{10A}{l_0^p} + r) &= \mathbb{P}_0^p(\mathbb{T}^p > \frac{10A}{l_0^p} + r | E) \mathbb{P}_0^p(E) \\ &\quad + \mathbb{P}_0^p(\mathbb{T}^p > \frac{10A}{l_0^p} + r | E^C) \mathbb{P}_0^p(E^C) \\ &\leq \mathbb{P}_0^p(\mathbb{T}^p > \frac{10A}{l_0^p} + r | E) + \mathbb{P}_0^p(E^C) \\ &\leq \mathbb{P}_0^p(N_* > \frac{10A}{l_0^p} + r) + \mathbb{P}_0^p(E^C) \\ &\leq C_{51}e^{-C_{52}r}. \end{aligned} \quad (4.35)$$

LEMMA 31 *There exist constants  $C_{61}, C_{62}$  such that*

$$\mathbb{E}_0^p \left[ \sum_{\ell=1}^{\mathsf{T}^p} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2 \right] \leq C_{61} \log A + C_{62}. \quad (4.36)$$

*Proof:* From relationship (Equation 4.25), we obtain there exists constant  $C_{63}$  such that

$$\begin{aligned} \mathbb{E}_0^p [(\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^4]^{1/2} &= \mathbb{E}_0^p \left[ \left( \frac{X_1^p + \dots + X_{\ell-1}^p}{\ell-1} - \psi'(\theta_p) \right)^4 \right]^{1/2} \\ &= \frac{C_{63}}{\ell-1}. \end{aligned}$$

Moreover, we have,

$$\begin{aligned} \mathbb{E}_0^p \left[ \sum_{\ell=1}^{\mathsf{T}^p} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2 \right] &= \mathbb{E}_0^p \left[ \sum_{\ell=1}^{\infty} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2 \mathbb{1}_{\{\mathsf{T}^p \geq \ell\}} \right] \\ &\leq \sum_{\ell=1}^{\infty} (\mathbb{E}_0^p [(\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^4])^{1/2} (\mathbb{P}_0^p(\mathsf{T}^p \geq \ell))^{1/2} \\ &= \sum_{\ell=1}^{10A/l_0^p} (\mathbb{E}_0^p [(\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^4])^{1/2} (\mathbb{P}_0^p(\mathsf{T}^p \geq \ell))^{1/2} \\ &\quad + \sum_{\ell=10A/l_0^p+1}^{\infty} (\mathbb{E}_0^p [(\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^4])^{1/2} (\mathbb{P}_0^p(\mathsf{T}^p \geq \ell))^{1/2} \\ &\leq \sum_{\ell=2}^{10A/l_0^p} \frac{C_{63}}{\ell} + \sum_{\ell=10A/l_0^p+1}^{\infty} \frac{C_{63}}{\ell-1} (C_{51} e^{-C_{52}(\ell-10A/l_0^p)})^{1/2} \\ &\leq C_{61} \log A + C_{62}. \end{aligned}$$

LEMMA 32 *There exist constants  $C_{71}, C_{72}$  such that*

$$\mathbb{E}_0^p \left[ \sum_{\ell=1}^{\mathsf{T}^p} (\hat{\theta}_\ell - \theta_p)^2 \right] \leq C_{71} \log A + C_{72}. \quad (4.37)$$

*Proof:* As  $\hat{\theta}_\ell$  is universally bounded, we define  $\rho = \max_{\delta \leq \theta \leq \zeta} \psi''(\theta)$  and obtain that:

$$\begin{aligned} \mathbb{E}_0^p \left[ \sum_{\ell=1}^{\mathbb{T}^p} (\hat{\theta}_\ell - \theta_p)^2 \right] &\leq \frac{\mathbb{E}_0^p [\sum_{\ell=1}^{\mathbb{T}^p} (\psi'(\hat{\theta}_\ell) - \psi'(\theta_p))^2]}{\rho^2} \\ &\leq C_{71} \log A + C_{72}. \end{aligned}$$

Combining the results above, we have:

$$\begin{aligned} D_p(\mathbb{T}_{\text{MSP}}(A)) &\leq \frac{\mathbb{E}_0^p[\mathbb{T}^p]}{1 - \beta_p} + \frac{1}{1 - \beta_p} \sum_{j=1}^{p-1} \mathbb{E}_\infty[\mathbb{T}^j] \\ &\leq \frac{\frac{(1-\beta_p)A}{l_0^p} + x_1 + x_2}{1 - \beta_p} + \frac{C_{31}}{1 - C_{41}}(p - 1) \\ &\leq \frac{A}{l_0^p} + \frac{x_1 + x_2}{1 - C_{41}} + \frac{C_{31}}{1 - C_{41}}(p - 1) \\ &\leq \frac{A}{l_0^p} + C_1 \log A + C_2 \sqrt{D_p(\mathbb{T}_{\text{MSP}}(A))} + C_3, \end{aligned}$$

for some suitable constants  $C_1, C_2, C_3$ .

### Detection Delay under Scenario II

Consider now the Scenario II when the change coincidentally occurs to the stream that is under monitoring at change time. Without loss of generality, we assume the change occurs to the  $p$ -th stream at change time  $\tau = k$  and the sampling index  $R_k$  at time step  $k$  points to the  $p$ -th stream (ie:  $R_k = p$ ). As mentioned, analysis under Scenario II is more challenging and the main issue here is that we need to take the observations before the change time into consideration.

Let us begin with some notations and definitions to better characterize the SPRT on the affected stream:

- $M(k)$  : the total time instants in which we have consecutively sampled at the affected stream (the  $p$ -th stream) up to time  $k$ .
- $S(k) = \sum_{\ell=k-M(k)+1}^k X_\ell^p$  is the sum of observations on the affected stream within the window  $[k - M(k) + 1, k]$ .

- $\mathbb{T}_k^p(A)$  is the sequential test on the affected stream starting from the change time  $\tau = k$ :

$$\mathbb{T}_k^p(A) = \inf \left\{ t - k : t \geq k, S_t^p = \sum_{\ell=k+1}^t ((\tilde{\theta}_\ell - \theta_0)X_\ell^p - \psi(\tilde{\theta}_\ell) + \psi(\theta_0)) \notin (0, A) \right\} \quad (4.38)$$

where  $\tilde{\theta}_\ell$  is the estimator affected by  $M(k), S(k)$ :

$$\tilde{\theta}_\ell = \min \left\{ \zeta, \max \left\{ \delta, \frac{S(k) + X_{k+1}^p + \cdots + X_{\ell-1}^p}{M(k) + \ell - 1 - k} \right\} \right\}. \quad (4.39)$$

Here we use the notation  $\tilde{\theta}_\ell$  to distinguish it from the estimator  $\hat{\theta}_\ell$  in Scenario I which are not affected by the observations before change time.

Note that if our proposed algorithm fails to raise an alarm on the affected stream, it will switch back to sample the first stream and this return to Scenario I. The detection delay under Scenario II can then be written as:

$$D_p(\mathbb{T}_{\text{MSP}}) = \mathbb{E}_k^p[\mathbb{T}_k^p | \mathbb{T}_{\text{MSP}} \geq k] + \beta_{p,k} \mathbb{E}_0^p[\mathbb{T}_{\text{MSP}}], \quad (4.40)$$

where  $\beta_{p,k}$  is the type-II error probability of the sequential test  $\mathbb{T}_k^p(A)$ . We have bounded  $\mathbb{E}_0^p[\mathbb{T}_{\text{MSP}}]$  in Scenario I and it suffices to study  $\mathbb{T}_k^p(A)$ . We begin the analysis of Scenario II with a series of lemmas on the estimation of  $M(k)$  and  $S(k)$ .

The main challenge here is that we need to bound  $M(k)$  and  $S(k)$  under the probability measure  $\mathbb{P}_k^p(|\mathbb{T}_{\text{MSP}} \geq k)$ , while  $M(k), S(k)$  are not independent of  $\{\mathbb{T}_{\text{MSP}} \geq k\}$ . A key observation is that our detection switched to the  $p$ -th stream at time  $k - M(k) + 1$ , and things will be much easier if we could bound  $M(k)$  and  $S(k)$  under the probability measure  $\mathbb{P}_k^p(|\mathbb{T}_{\text{MSP}} \geq k - M(k) + 1)$  because  $M(k), S(k)$  are independent of  $\{\mathbb{T}_{\text{MSP}} \geq k - M(k) + 1\}$ . This clearly requires the following estimation on the conditional probability:

LEMMA 33 *There exists constants  $C_{81}$  such that:*

$$\mathbb{P}_k^p(\mathbb{T}_{\text{MSP}} \geq k | \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1) \geq C_{81} \quad (4.41)$$

for some constant  $C_{81}$  that are not related to  $k$ ,  $A$  and some sufficiently large  $A \geq A_0$ .

*Proof:* The event  $\{\mathbb{T}_{\text{MSP}} < k | \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1\}$  means that a type-I error of sequential test  $\mathbb{T}^p$  has occurred during the time period  $[k - M(k) + 1, k - 1]$ . Through standard technique of changing probability measure, we can prove the type-I error  $\alpha_p$  of the sequential test  $\mathbb{T}^p$  satisfies  $\alpha_p \leq 1/e^A \leq 1/e^{A_0}$ . Thus we have:

$$\begin{aligned} \mathbb{P}_k^p(\mathbb{T}_{\text{MSP}} \geq k | \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1) &= 1 - \mathbb{P}_k^p(\mathbb{T}_{\text{MSP}} < k | \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1) \\ &\geq 1 - \alpha_p \\ &\geq C_{81}, \end{aligned}$$

for some suitable constant  $C_{81} \in (0, 1)$ .

LEMMA 34 *There exists constant  $C_{91}$  such that for  $r = 1, 2, \dots$ :*

$$\mathbb{E}_k^p[S(k)^4 | \mathbb{T}_{\text{MSP}} \geq k] \leq C_{91}, \quad (4.42)$$

$$\mathbb{E}_k^p\left[\left(\frac{S(k)}{M(k) + r}\right)^4 | \mathbb{T}_{\text{MSP}} \geq k\right] \leq \frac{C_{91}}{r^4}, \quad (4.43)$$

*Proof:* By Holder's inequality, relationship (Equation 4.30) and Lemma Lemma 33,

$$\begin{aligned} \mathbb{E}_k^p[S(k)^4 | \mathbb{T}_{\text{MSP}} \geq k] &= \mathbb{E}_k^p\left[\sum_{u=1}^{\infty} S(k)^4 \mathbb{1}_{\{M(k)=u\}} | \mathbb{T}_{\text{MSP}} \geq k, \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1\right] \\ &\leq \frac{\sum_{u=1}^{\infty} (\mathbb{E}_k^p[(\sum_{\ell=k-u}^k X_\ell^p)^8])^{1/2} \mathbb{P}_\infty(\mathbb{T}^p \geq u)^{1/2}}{\mathbb{P}_k^p(\mathbb{T}_{\text{MSP}} \geq k | \mathbb{T}_{\text{MSP}} \geq k - M(k) + 1)} \\ &\leq \frac{\sum_{u=1}^{\infty} u^2 (C_{32} e^{-C_{33}u})^{1/2}}{C_{81}} \\ &\leq C_{91}. \end{aligned}$$

Relationship (Equation 4.43) follows directly from Relationship (Equation 4.42).

LEMMA 35 For the estimator  $\hat{\theta}_\ell$  defined in (25), there exists constant  $C_{101}$  such that:

$$\mathbb{P}_k^p(|\tilde{\theta}_r - \theta_p| > \lambda) < \frac{C_{101}}{(r-k)^4} \quad (4.44)$$

for all  $r = k+1, k+2, \dots$ .

*Proof:* See Lemma Lemma 34 and Lorden and Pollak[77]. We need to note that due to the existence of  $S(k)$  and  $M(k)$ , we no longer has the exponential bound on the deviance as in Scenario I but only a weaker bound of  $1/r^4$ . However, this weaker bound is enough to bound  $\mathbb{T}_k^p$  under  $\mathbb{P}_k^p$ .  $\square$

Similarly as in Lemma Lemma 27, we can write  $\mathbb{E}_k^p[\mathbb{T}_k^p | \mathbb{T}_{\text{MSP}} \geq k]$  as:

$$\mathbb{E}_k^p[\mathbb{T}_k^p | \mathbb{T}_{\text{MSP}} \geq k] \leq \frac{(1 - \beta_{p,k})A}{l_0^p} + x_1 + x_2, \quad (4.45)$$

where

$$\begin{aligned} x_1 &\leq C_{21}(\mathbb{E}_k^p[\sum_{\ell=k+1}^{k+\mathbb{T}_k^p} (\tilde{\theta}_\ell - \theta_p)^2 | \mathbb{T}_{\text{MSP}} \geq k])^{1/2} + C_{22}(\mathbb{E}_k^p[\mathbb{T}_k^p])^{1/2} \\ x_2 &\leq C_{23}\mathbb{E}_k^p[\sum_{\ell=k+1}^{k+\mathbb{T}_k^p} (\tilde{\theta}_\ell - \theta_p)^2 | \mathbb{T}_{\text{MSP}} \geq k] + C_{24}\mathbb{E}_k^p[\sum_{\ell=k+1}^{k+\mathbb{T}_k^p} (\psi'(\tilde{\theta}_\ell) - \psi'(\theta_p))^2 | \mathbb{T}_{\text{MSP}} \geq k] \end{aligned}$$

for some constants  $C_{21}, C_{22}, C_{23}, C_{24}$ .

LEMMA 36 There exists constant  $C_{111}$  such that for sufficient large  $A \geq A_0$  and every  $r = 1, 2, \dots$ ,

$$\mathbb{P}_k^p(\mathbb{T}_k^p > \frac{10A}{l_0^p} + r) \leq \frac{C_{111}}{r^3}. \quad (4.46)$$

*Proof:* See Lemma 30. Similarly as in Lemma 11, we no longer has the exponential bound but only a weaker bound of  $1/r^3$ . However, this weaker bound is enough to bound the sum of the deviance  $\sum_{\ell=k+1}^{k+\mathbb{T}_k^p} (\psi'(\tilde{\theta}_\ell) - \psi'(\theta_p))^2 | \mathbb{T}_{\text{MSP}} \geq k]$ .

LEMMA 37 *There exist constants  $C_{121}, C_{122}$  such that*

$$\mathbf{E}_k^p \left[ \sum_{\ell=k+1}^{k+\mathbb{T}_k^p} (\psi'(\tilde{\theta}_\ell) - \psi'(\theta_p))^2 \mid \mathbb{T}_{\text{MSP}} \geq k \right] \leq C_{121} \log A + C_{122}. \quad (4.47)$$

*Proof:* Similar proof as in Lemma 31.

Combining the results above, we obtain that

$$\begin{aligned} D_p(\mathbb{T}_{\text{MSP}}) &= \mathbf{E}_k^p[\mathbb{T}_k^p \mid \mathbb{T}_{\text{MSP}} \geq k] + \beta_{p,k} \mathbf{E}_0^p[\mathbb{T}_{\text{MSP}}] \\ &\leq \frac{(1 - \beta_{p,k})A}{l_0^p} + x_1 + x_2 + \beta_{p,k} \left( \frac{A}{l_0^p} + C_1 \log A + C_2 \sqrt{D_p(\mathbb{T}_{\text{MSP}})} \right) \\ &\leq \frac{A}{l_0^p} + C_4 \log A + C_5 \sqrt{D_p(\mathbb{T}_{\text{MSP}})} + C_6, \end{aligned}$$

for some suitable constants  $C_4, C_5, C_6$ . □

## 4.2 Active Quickest Detection When Monitoring Multi-streams with Two Affected Streams

This project was published as a conference paper in the proceedings IEEE International Symposium on Information Theory 2022.

### 4.2.1 Problem Formulation and Method Review

#### *Mathematical Formulation*

Suppose there are  $p$  statistically independent local streams in a system, and denote with  $X_t^i$  the observation from the  $i$ -th stream at time  $t$ , where  $i = 1, \dots, p$  and  $t = 1, 2, \dots$ . Initially, the system is in the in-control state and the data stream  $\{X_t^i\}$  from the  $i$ -th stream produces i.i.d. samples following the density  $f_i(X)$ . At some unknown time  $\tau$ , a triggering event occurs to the system and affects two out of  $p$  data streams in the sense that if the  $i$ -th stream is affected, the density function of its local samples  $X_t^i$  changes from  $f_i$  to  $g_i$  at time

$t \geq \tau$ . Specifically, denote the index set of the affected streams as  $R = \{i_1, i_2\}$ . Then if index  $i \in R$ ,

$$X_t^i \sim \begin{cases} f_i(X), & \text{if } t \leq \tau \\ g_i(X), & \text{if } t > \tau, \end{cases} \quad (4.48)$$

whereas  $X_t^j \sim f_j(X)$  for  $j \notin R$  and all  $t > 0$ .

Under the sampling control constraint, we are only able to observe two out of  $p$  local streams at each time instant  $t$ . To be mathematically rigorous, define the *Sample set*  $O_t = \{i_{1,t}, i_{2,t}\}$ , where the  $i$ -th stream is observed at time  $t$  if  $i \in O_t$  and the  $i$ -th stream is unobservable if  $i \notin O_t$ . The sampling constraint can then be expressed as

$$O_t \subset \{1, 2, \dots, p\} \text{ and } |O_t| = 2, \quad (4.49)$$

where  $|A|$  denotes the cardinality of set  $A$ .

Thus in the active quickest detection problem under sampling control, an algorithm consists two elements: one is the sampling policies, e.g., choose the sample set  $O_t$  for all time  $t$  subject to (Equation 4.49), and the other is the decision policy that is defined as the stopping time  $T$  with respect to the observed data sequence.

Denote by  $P_\tau^{(i_1, i_2)}$  and  $E_\tau^{(i_1, i_2)}$  the probability measure and expectation when at time  $\tau$ , the change occurs to the  $i$ -th stream for  $i = i_1, i_2$  and there are no changes at other data streams. And denote by  $P_\infty$  and  $E_\infty$  the probability measure and expectation when there the change occurs at time  $\infty$ . Following the classical minimax formulation for quickest detection proposed by [61], we are interested in finding a procedure  $(\{O_t\}_{t=1, \dots, \infty}, T)$  that minimizes the worst-case detection delay

$$D^{(i_1, i_2)}(T) = \sup_{t \geq 0} \text{ess sup } E_t^{(i_1, i_2)}[T - t | \mathcal{F}_t^i, T > t]. \quad (4.50)$$

for any combinations of affected streams  $R = \{i_1, i_2\} \subset \{1, 2, \dots, p\}$ , subject to the

average run length to false-alarm constraint

$$E_\infty[T] \geq \gamma > 1. \quad (4.51)$$

where  $\gamma$  is a pre-specified constant to control the global false alarm rate.

### *Review of Existing Methods*

Let us now review three existing methods that is related to our problem and proposed algorithm under the passive setup that we have access to the complete data set. The first method is the oracle CUSUM procedure under the oracle scenario when the index set  $R = \{i_1, i_2\}$  of affected stream is known. In such case, there is no need to sample other streams and the corresponding optimal stopping time will be

$$T_{\text{oracle}}(A) = \inf\{t > 0 : W_t \geq A\} \quad (4.52)$$

where  $W_t$  is the CUSUM statistics recursively defined as

$$W_t = \max\{W_{t-1}, 0\} + \log \frac{g_{i_1}(X_t^{i_1})}{f_{i_1}(X_t^{i_1})} + \log \frac{g_{i_2}(X_t^{i_2})}{f_{i_2}(X_t^{i_2})} \quad (4.53)$$

for  $t \geq 1$  and  $W_0 = 0$ , see [63].

The second method is the so-called MAX-scheme, which proposed to raise an alarm when the maximum of  $p$  local summary statistics triggers an alarm:

$$T_{\text{MAX}}(A) = \inf\{t > 0 : \max\{W_t^1, W_t^2, \dots, W_t^p\} \geq A\}, \quad (4.54)$$

where  $W_t^i (i = 1, \dots, p)$  is the local CUSUM statistics for the  $i$ -th stream that satisfies the

standard CUSUM-update:

$$W_t^i = \max\{W_{t-1}^i, 0\} + \log \frac{g_i(X_t^i)}{f_i(X_t^i)}, \quad (4.55)$$

and constant  $A$  is chosen to select the false-alarm constraint, see [35]. However, the MAX-scheme is shown to be only efficient when there is exactly one affected stream but inefficient when there is two or more affected streams.

The third method is the SUM-scheme proposed in [15], whose main idea is to raise an alarm based on the sum of the local CUSUM statistics from each individual data stream, and is shown to achieve the asymptotic optimality. Mathematically, [15] proposed to raise an alarm at

$$\mathsf{T}_{\text{SUM}}(A) = \inf\{t > 0 : W_t^1 + W_t^2 + \dots + W_t^p \geq A\}, \quad (4.56)$$

where  $W_t^i (i = 1, \dots, p)$  is the local CUSUM statistics for the  $i$ -th stream and constant  $A$  is chosen to select the false-alarm constraint.

Under the active learning setup of quickest detection problem, the problem becomes much more challenging as one needs to handle temporal uncertainty and spatial uncertainty simultaneously. How to actively select the two observed streams at each time and how to use the observed data to raise an alarm as quickly as possible are highly non-trivial. Also, a more challenging problem is how to accompany them with theoretical justification capable of establishing the desired form of asymptotic optimality.

### *Related Work*

#### 4.2.2 Our Proposed Algorithm

In this section, we propose two families of active quickest detection algorithm  $\mathsf{T}_0$  and  $\mathsf{T}_1$  and both of them are proved to be asymptotically optimal. One of the two algorithms  $\mathsf{T}_0$  has the renewal property and the proof is more understandable. Another algorithm  $\mathsf{T}_1$  has a better numerical performance while the proof is much more complicated. Due to the page

limit, in this paper we mainly focus on the first algorithm  $T_0$ . For the better presentation, we divide the current section into two subsections: in Section subsection 4.2.2 we present our proposed algorithm under the sampling control in (Equation 2.2) and in Section subsection 4.2.3 we establish its second-order asymptotic optimality properties.

### *Algorithm Development*

Without loss of generality, here we consider the homogeneous setting ( $f_k \equiv f, g_k \equiv g$ ). We divide our proposed algorithm into four steps.

#### *Step 1: Update of Local Summary Statistics*

At the beginning of time  $t$ , there are two cases for each local stream, depending on whether it is sampled or not. If a local stream is not sampled ( $i \notin O_t$ ), then we update  $W_t^i$  as  $\max\{W_{t-1}^i, 0\}$ . If a local stream is sampled ( $i \in O_t$ ), then we update  $W_t^i$  as in the classical CUSUM procedure in (Equation 4.55). To be more specific, at each time instant  $t = 1, 2, \dots$ , the local summary statistics  $W_t^i$  can be defined recursively as

$$\begin{aligned} W_t^i &= W_{t-1}^i + \mathbb{1}_{\{i \in O_t\}} \log \frac{g(X_t^i)}{f(X_t^i)} \\ &= \begin{cases} \max\{W_{t-1}^i, 0\}, & \text{if } i \in O_t \\ \max\{W_{t-1}^i, 0\} + \log \frac{g(X_t^i)}{f(X_t^i)}, & \text{if } i \notin O_t, \end{cases} \end{aligned} \quad (4.57)$$

with the initial values  $W_0^i = 0$  for all  $i = 1, \dots, p$ .

#### *Step 2: Decision Policy*

we propose to adopt the SUM scheme proposed in Mei[15] and raise an alarm at

$$T_0(A) = \inf \{t > 0 : W_t^1 + W_t^2 + \dots + W_t^p \geq A\}, \quad (4.58)$$

for some pre-specified constant  $A$ , which is selected so as to satisfy the false-alarm constraint in (Equation 4.51).

*Step 3: Win-Stay, Lose-Switch Sampling Policy*

If we fail to raise an alarm in Step 2, we need to update the sample set  $O_t$  based on the statistics  $W_t^1, \dots, W_t^p$  updated in Step 1. Here we adopt the win-stay, lose-switch sampling policy. On high level, for any candidate stream  $i$  in the sample set  $O_t$ , if  $W_t^i > 0$  then we assume the  $i$ -th stream “wins” and keep on observing it. Otherwise if  $W_t^i \leq 0$  then we assume the  $i$ -th stream “loses” and switch to observe another stream. Mathematically, let the sample set at time instant  $t$  be  $O_t = \{i_{1,t}, i_{2,t}\} (i_{1,t} < i_{2,t})$  and initialize  $O_{t+1} = O_t$ . For  $j = 1, 2$ , we randomly pick one index  $j_t$  from the unobservable set  $\{1, \dots, p\} \setminus O_{t+1}$  and update  $O_{t+1}$  :

$$\begin{aligned} O_{t+1} &= \{O_{t+1}/i_{j,t}, j_t\} \text{ if } W_t^{i_{j,t}} \leq 0, \\ O_{t+1} &= O_{t+1}, \text{ if } W_t^{i_{j,t}} > 0. \end{aligned} \tag{4.59}$$

*Step 4: Reset Local Summary Statistics*

If we have updated the sample set ( $O_{t+1} \neq O_t$ ) in Step 3, then we reset all local summary statistics to 0. Mathematically,

$$\begin{aligned} W_t^i &= \mathbb{1}_{\{O_{t+1}=O_t\}} W_t^i \\ &= \begin{cases} 0, & \text{if } O_{t+1} \neq O_t \\ W_t^i, & \text{if } O_{t+1} = O_t, \end{cases} \end{aligned} \tag{4.60}$$

for  $i = 1, 2, \dots, p$ .

In summary, our proposed algorithm  $T_0$  can be summarized as follows:

**Remark:** It turns out that sometimes skipping Step 4 and not resetting the local summary statistics will lead to the better performance. Here we denote the local summary statistics without the resetting steps in (Equation 4.60) as  $\widetilde{W}_t^i (i = 1, \dots, p)$  and propose

---

**Algorithm 2** Our Proposed Algorithm  $T_0$ 

---

```
1: Initialize  $O_1 = \{1, 2\}$  and  $W_0^i = 0$  for  $i = 1, \dots, p$ .
2: for each time  $t$  do
3:   Sample the streams in the sample set  $O_t$ .
4:   Update the local summary statistics  $W_t^i$  for  $i = 1, \dots, p$  as in (Equation 4.57).
5:   if  $W_t^1 + \dots + W_t^M \geq A$  then
6:     Raise an alarm at  $T_0(A) = t$ .
7:   end if
8:   Let  $O_{t+1} = O_t$ .
9:   for  $j = 1, 2$  do
10:    Randomly pick one index  $j_t$  from the unobservable set  $\{1, \dots, p\} \setminus O_{t+1}$ .
11:    if  $W_t^{i_{j,t}} \leq 0$  then
12:      Update  $O_{t+1} = \{O_{t+1}/i_{j,t}, j_t\}$ .
13:    end if
14:  end for
15:  if  $O_{t+1} \neq O_t$  then
16:    Reset all local summary statistics to 0.
17:  end if
18: end for
```

---

another family of stopping time as:

$$T_1(A) = \inf \left\{ t > 0 : \widetilde{W}_t^1 + \widetilde{W}_t^2 + \dots + \widetilde{W}_t^p \geq A \right\}, \quad (4.61)$$

### 4.2.3 Asymptotic Optimality

In this subsection, we will investigate the theoretical properties of our proposed algorithm  $T = T_0$  in (Equation 4.11) or  $T = T_1$  in (Equation 4.61). First, we need to make necessary assumptions. We assume that Kullback-Leibler information numbers are positive and finite for all  $i = 1, 2, \dots, p$ :

$$\begin{aligned} I(f, g) &= \int \log \frac{f(X)}{g(X)} f(X) dX \in (0, \infty), \\ I(g, f) &= \int \log \frac{g(X)}{f(X)} g(X) dX \in (0, \infty). \end{aligned} \quad (4.62)$$

Moreover, we assume that the second moments of log likelihood ratios are bounded away from  $\infty$ .

$$\begin{aligned} \int (\log \frac{f(X)}{g(X)})^2 f(X) dX &\in (0, \infty), \\ \int (\log \frac{g(X)}{f(X)})^2 g(X) dX &\in (0, \infty). \end{aligned} \quad (4.63)$$

Now we are ready to present the theoretical properties of our proposed algorithm. The main results are summarized in the following theorem and its corollary, whose high-level proofs are presented in the appendix.

**THEOREM 4** *For our proposed algorithm  $\mathbb{T} = \mathbb{T}_0$  in (Equation 4.11) or  $\mathbb{T} = \mathbb{T}_1$  in (Equation 4.61), we have*

$$E_\infty[\mathbb{T}] \geq e^A. \quad (4.64)$$

*Moreover, for any combinations of affected streams  $\{i_1, i_2\} \in \{1, 2, \dots, p\}$ , its detection delay satisfies*

$$D^{(i_1, i_2)}(\mathbb{T}) \leq \frac{A}{2I(g, f)} + C_0 p(p-1) + C_1 \quad (4.65)$$

*as  $A \rightarrow \infty$ . Here  $C_0, C_1$  are constants depending only on the distributions, not on  $A$ .*

**COROLLARY 3** *Let  $A = \log \gamma$ , then our proposed algorithm  $\mathbb{T} = \mathbb{T}_0$  in (Equation 4.11) or  $\mathbb{T} = \mathbb{T}_1$  in (Equation 4.61) satisfies both the false alarm constraint in (Equation 4.51) and the sampling control constraint in (Equation 2.2). Moreover, for any combinations of affected streams  $\{i_1, i_2\} \in \{1, 2, \dots, p\}$ , its detection delay satisfies*

$$0 \leq D^{(i_1, i_2)}(\mathbb{T}) - D_{\text{orc}}^{(i_1, i_2)} \leq C_0 p(p-1) + C_2 \quad (4.66)$$

*where  $D_{\text{orc}}^{(i_1, i_2)}$  is the oracle detection delay achieved by the classical CUSUM procedure for monitoring changes in distribution of the  $i_1$ -th stream and  $i_2$ -th stream subject to the false*

alarm constraint in (Equation 4.51):

$$D_{\text{orc}}^{(i_1, i_2)} = \frac{\log \gamma}{2I(g, f)} + C_3 \quad (4.67)$$

and the parameters  $C_2$  and  $C_3$  are constants depending only on the distributions, not on  $\gamma$ .

It is useful to add some remarks.

1. The second-order asymptotic optimality property is assured under the assumption that the number  $p$  of local streams is fixed. As the relationship (Corollary 3) hold for any given  $M$ , we can deduce that our proposed algorithm still enjoy first-order asymptotically optimality when  $p \rightarrow \infty$  provided that  $p^2 = o(\log \gamma)$ . It remains an open problem to develop asymptotically optimal procedures when  $p^2 \gg \log \gamma$ .
2. Our proposed algorithm including the win-stay, lose-switch sampling policy and the decision policy can be extended to the general case when there are  $r \geq 3$  affected streams. Furthermore, the second-order asymptotic optimality still holds when the dimension  $p$  is fixed but the additional term related to the dimension  $p$  in the relationship (Equation 4.66) needs to be replaced by  $\binom{p}{r}$ .
3. The reset of local summary statistics in (Equation 4.60) of Step 4 is an artificial step to ensure that the system has the renewal property and all statistics begin with 0 when we switch among the streams. However, it turns out that not resetting the local summary statistics (skipping Step 4) can still have asymptotic optimality and might lead to a better numerical performance.

#### 4.2.4 Simulation

In this section, we conduct the Monte Carlo simulation to illustrate the usefulness of our proposed algorithm  $T = T_0$  in (Equation 4.11) and  $T = T_1$  in (Equation 4.61). Assume  $f \sim N(0, 1)$  and  $g \sim N(1, 1)$ . We consider two choices of number  $p$  of available local

streams, namely  $p = 3, 5$ . For each choice of  $p$  streams, we consider the mean shift in normal distribution from  $f = N(0, 1)$  to  $g = N(\mu, 1)$  with  $\mu \geq 1$ . Due to the page constraints, here we only present the homogeneous setting ( $f_i \equiv f, g_i \equiv g$  for all  $i$ ).

In the simulation, we compute the detection delay of our proposed algorithm  $T = T_0$  in (Equation 4.11) and  $T = T_1$  in (Equation 4.61) and compare it against the MAX-scheme  $T_{\text{MAX}}$  in (Equation 4.54) and the SUM-scheme  $T_{\text{SUM}}$  in (Equation 4.56). As our detection delay is defined in the worst-case sense and our proposed algorithm always start from the first two streams, we assume the change always occurs to the last two streams. In each case, we set the false alarm constraint  $\gamma = 10,000$ .

Table 4.3: Comparison of Detection Delay of  $T = T_0$  or  $T_1$ ,  $T_{\text{SUM}}$  and  $T_{\text{MAX}}$

| $\gamma = 10000$ | $p = 3$ |       |       |       |
|------------------|---------|-------|-------|-------|
| $\mu$            | SUM     | MAX   | $T_0$ | $T_1$ |
| 1                | 9.36    | 13.09 | 15.50 | 9.85  |
| 1.25             | 6.00    | 8.52  | 10.07 | 6.43  |
| 1.5              | 4.10    | 5.94  | 7.16  | 4.55  |
| 1.75             | 2.92    | 4.29  | 5.36  | 3.34  |
| 2.0              | 2.12    | 3.83  | 4.14  | 2.53  |

| $\gamma = 10000$ | $p = 5$ |       |       |       |
|------------------|---------|-------|-------|-------|
| $\mu$            | SUM     | MAX   | $T_0$ | $T_1$ |
| 1                | 10.14   | 14.06 | 28.00 | 13.01 |
| 1.25             | 6.53    | 9.19  | 18.06 | 8.95  |
| 1.5              | 4.45    | 6.37  | 12.49 | 6.62  |
| 1.75             | 3.20    | 4.62  | 9.39  | 5.11  |
| 2.0              | 2.33    | 3.44  | 7.08  | 4.11  |

In Table 4.3 we report the detection delay of our proposed algorithm  $T$ , the MAX-scheme  $T_{\text{MAX}}$  in (Equation 4.54) and the SUM-scheme  $T_{\text{SUM}}$  in (Equation 4.56). with respect to the same average-run-length to false-alarm for  $p = 3, 5$ . From Table 4.3, it is clear that while  $T_0$  and  $T_1$  are both asymptotically optimal, the performance of  $T_1$  is always better than that of  $T_0$ . Also, even with the sampling rate control of  $2/p$  of the full-sampling scenario, the detection delay of our proposed algorithm  $T_1$  is comparable to other procedures under the full-sampling scenario. This shows that our proposed algorithm  $T$  is

quite efficient and useful.

#### 4.2.5 Technical Details

Without loss of generality, we assume the change occurs to the  $(p - 1)$ -th and the  $p$ -th streams at time  $\tau = 0$  as our proposed algorithm always started from the first and the second streams.

We first prove the false-alarm relationship (Equation 4.64), which is the easier one. The key idea in proving (Equation 4.64) is to relate our proposed stopping time  $T_0$  in (Equation 4.11) to the following sequential test  $\mathcal{T}$  :

$$\mathcal{T} = \min\{\mathcal{T}_1, \mathcal{T}_2\}, \quad (4.68)$$

where

$$\mathcal{T}_1 = \inf\{t : t > 0, S_t^1 + S_t^2 \geq A\},$$

$$\mathcal{T}_2 = \min(\mathcal{T}_3, \mathcal{T}_4)$$

$$\mathcal{T}_3 = \inf\{t : t > 0, S_t^1 \leq 0\},$$

$$\mathcal{T}_4 = \inf\{t : t > 0, S_t^2 \leq 0\},$$

where  $S_t^i = \sum_{j=1}^t \log(g(X_t^i)/f(X_t^i))$ . Our proposed stopping time  $T_0$  under  $P_\infty(\cdot)$  and  $E_\infty[\cdot]$  can then be written as:

$$\begin{aligned} E_\infty[T] &= E_\infty[\mathcal{T}] + (1 - \alpha)E_\infty[\mathcal{T}] + (1 - \alpha)^2 E_\infty[\mathcal{T}] + \dots \\ &= \frac{E_\infty[\mathcal{T}]}{\alpha} \geq \frac{1}{\alpha} \end{aligned} \quad (4.69)$$

where  $\alpha = P_\infty(\mathcal{T}_1 \leq \mathcal{T}_2)$ . We now consider another stopping time  $\mathcal{T}_5$  :

$$\mathcal{T}_5 = \inf\{t : t > 0, S_t^1 + S_t^2 \leq 0\}, \quad (4.70)$$

It is clear that  $P_\infty(\mathcal{T}_1 \leq \mathcal{T}_5)$  is the Type-I error probability of some SPRT and we have  $P_\infty(\mathcal{T}_1 \leq \mathcal{T}_5) \leq 1/e^A$  by the standard tool of changing measure. Moreover, as  $\mathcal{T}_2 \leq \mathcal{T}_5$ , we derive that  $P_\infty(\mathcal{T}_1 \leq \mathcal{T}_2) \leq P_\infty(\mathcal{T}_1 \leq \mathcal{T}_5) \leq 1/e^A$ . Thus,

$$E_\infty[T_0] \geq \frac{1}{\alpha} \geq e^A. \quad (4.71)$$

We then focus on the proof of detection delay relationship (Equation 4.65). Due to the page limit, we consider the simplest scenario when the number of streams  $p = 3$  and give the high-level idea for the scenario of general  $p$  local streams.

**Case 1: the number of streams  $p = 3$**

When  $p = 3$ , we take advantage of the fact that there is always one affected stream under monitoring after the change occurs. The main idea of the proof for Case 1 is to divide the system into two states:

- State 1: all two affected streams are under monitoring. The system in State 1 can be characterized as  $\mathcal{T}$  in (Equation 4.68) when the first and the second streams get affected.
- State 2: one affected stream and one unaffected stream are under monitoring. The system in State 2 can be characterized as  $\mathcal{T}$  in (Equation 4.68) when the second and the third streams get affected.

It is clear that the system will always be in one of these two states and below we analyze the detection delay of our proposed algorithm by characterizing the system's performance on these two states.

We now define a new stopping time  $T'_0$  that has only one difference on the decision policy from our proposed algorithm  $T_0$  in (Equation 4.11):  $T'_0$  will only raise an alarm on State 1 but  $T_0$  can raise an alarm on both states. That is, when the system is in State 2, we replace the sequential test  $\mathcal{T}$  with the sequential test  $\mathcal{T}_2$ . It is then clear that  $T'_0 \geq T_0$  and it suffices to prove that  $T'_0$  satisfies the relationship (Equation 4.65).

For the proposed stopping time  $T'_0$ ,

- When the system is in State 1, either it will raise an alarm or it enters State 2.
- When the system is in State 2, either it will enter State 1 or it stays in State 2.

The system can be characterized in a sequence of states: State 2, State 2,  $\dots$ , State 1, State 2, State 2,  $\dots$ , State 1,  $\dots$ , Raise an alarm. And the detection delay of  $T'_0$  when the second and third streams get affected can be written as

$$\begin{aligned}
D^{(2,3)}(T'_0) &= E_0^{(2,3)}[T_2] + \beta E_0^{(2,3)}[T_2] + \beta^2 E_0^{(2,3)}[T_2] + \dots \\
&+ E_0^{(1,2)}[T] + \gamma E_0^{(2,3)}[T_2] + \gamma \beta E_0^{(2,3)}[T_2] + \dots \\
&+ \gamma(E_0^{(1,2)}[T] + \gamma E_0^{(2,3)}[T_2] + \gamma \beta E_0^{(2,3)}[T_2] + \dots) \quad (4.72) \\
&+ \dots \\
&= \frac{E_0^{(1,2)}[T]}{1 - \gamma} + \frac{E_0^{(2,3)}[T_2]}{(1 - \beta)(1 - \gamma)}
\end{aligned}$$

where  $\beta = P_0^{(2,3)}(\mathcal{T}_3 > \mathcal{T}_4)$  and  $\gamma = P_0^{(1,2)}(\mathcal{T}_1 > \mathcal{T}_2)$ . Furthermore we can show that

$$\begin{aligned}
\frac{E_0^{(1,2)}[T]}{1 - \gamma} &= \frac{A}{2I(g, f)} + C_4 \\
\frac{E_0^{(1,2)}[T_2]}{(1 - \beta)(1 - \gamma)} &= C_5,
\end{aligned}$$

for some constants  $C_4, C_5$  that only depend on the distribution but not  $A$ . The detection delay relationship (Equation 4.65) under the scenario when  $p = 3$  is proved.

### Case 2: general $p \geq 4$

For general number  $p \geq 4$  of available streams, we need to add one more state to the system:

- State 3: two unaffected streams are under monitoring.

It is clear that the system will always be in one of these three states and we can do a similar analysis of  $T_0$  by writing  $T_0$  as the sum of stopping time on each state. The  $O(p(p - 1))$

term results from the fact that the system takes  $O(p)$  time to switch from State 3 to State 2 and takes another  $O(p)$  time to switch from State 2 to State 1.

### 4.3 Impact of Compensation Coefficients on Active Quickest Detection

This is a collaborative project with Dr. Jianjun Shi, and was submitted to the Journal Sequential Analysis for possible publication in February 2024.

#### 4.3.1 Background

##### *Problem Formulation*

Suppose there are  $p$  independent local streams in a system, and denote with  $X_{t,i}$  the observation from the  $i$ -th stream at time  $t$ , where  $i = 1, \dots, p$  and  $t = 1, 2, \dots$ . Initially, the system is in the in-control state and the data stream  $\{X_{t,i}\}$  from the  $i$ -th stream produces i.i.d. samples following the density  $f(X)$ . At some unknown time  $\nu$ , an undesired event occurs to the system and affects  $s$  out of  $p$  data streams in the sense of changing the density function of affected local streams from  $f$  to  $g$  at time  $t \geq \nu$ . Mathematically, denote the index set of the affected streams as  $S = \{i_1^*, i_2^*, \dots, i_s^*\}$ . Then if index  $i \in S$ ,

$$X_{t,i} \sim \begin{cases} f(X), & \text{if } t \leq \nu \\ g(X), & \text{if } t > \nu, \end{cases} \quad (4.73)$$

whereas  $X_{t,j} \sim f(X)$  for  $j \notin S$  and all  $t > 0$ .

Under the sampling control constraint, we are only able to observe  $q$  out of  $p$  local streams at each time instant  $t$ . Here we assume that the sampling constraint is greater than the number of affected data streams, i.e.,  $q \geq s$ . To be rigorous, define the *Sampling set*  $S_t = \{i_{1,t}, i_{2,t}, \dots, i_{q,t}\}$  which points to the set of streams that will be sampled during time

instant  $t$ , and the sampling constraint can be expressed as

$$S_t \subset \{1, 2, \dots, p\} \text{ and } |S_t| = q, \quad (4.74)$$

where  $|A|$  denotes the cardinality of set  $A$ . In addition, we denote by  $X_{S_t} = \{X_{t,i_1,t}, \dots, X_{t,i_q,t}\}$  the set of observations at time  $t$ .

In the active quickest detection problem under sampling control, an algorithm consists two elements: one is the sampling policies, e.g., choose the sample set  $S_t$  for all time  $t$  subject to (Equation 4.49), and the other is the decision policy that is defined as the stopping time  $T$  with respect to the observed data sequence.

Denote by  $P_t$  and  $E_t$  the probability measure and expectation when change occurs at time  $t$ , and denote by  $P_\infty$  and  $E_\infty$  the probability measure and expectation when there are no changes, or equivalently, when the change occurs at time  $\infty$ . Denote by  $\mathcal{F}_t$  the filtration generated by the observations and sampling sequences, i.e.,

$$\mathcal{F}_t = \sigma(X_{S_1}, \dots, X_{S_t}, S_1, \dots, S_t).$$

To evaluate the performance of the detection procedure  $(\{S_t\}_{t=1,\dots,\infty}, T)$ , we first assume the procedure is designed to satisfy the false alarm constraint

$$E_\infty[T] \geq \gamma > 1, \quad (4.75)$$

for some pre-specified value  $\gamma$ . Following [61], we are interested in finding a procedure  $(\{S_t\}_{t=1,\dots,\infty}, T)$  that minimizes the worst case detection delay:

$$D(T) = \sup_{\nu \geq 0} \text{ess sup } E_\nu[\max(T - \nu, 0) \mid \mathcal{F}_\nu], \quad (4.76)$$

for any set  $S$  of affected streams, subject to the sampling control constraint in (Equa-

tion 4.49) and the average run length to false-alarm constraint in (Equation 4.51). In practice, in most cases the detection delay occurs at time  $\nu = 0$ , and below we will follow the statistical process control (SPC) literature to focus on the detection delay properties when change occurs at time  $\nu = 0$ .

### *Review of the TRAS algorithm*

The TRAS algorithm is an efficient and widely used algorithm proposed in [1], which is motivated from the SUM-scheme in [15] for effective monitoring of the high-dimensional process in the context of limited resources. A key idea of TRAS is to introduce a compensation coefficient to those local data streams that are not taking observations. That allows us to increase the chance of exploring unobserved local streams when using greedy sampling to compare with the local CUSUM statistics in [62] from those observed local streams. This novel idea of compensation coefficient has been extended to other more complicated scenarios, see [78, 79, 80].

The TRAS algorithm first constructs a local statistic for each observed stream (i.e.,  $i \in S_t$ ). For those unobserved streams (i.e.,  $i \notin S_t$ ), the local statistics are updated by adding a compensation parameter  $\Delta$ . Mathematically, the TRAS algorithm recursively updates the local statistics by

$$W_{t,i} = \begin{cases} \max\{W_{t-1,i} + \log \frac{g(X_{t,i})}{f(X_{t,i})}, 0\}, & \text{if } i \in S_t, \\ W_{t-1,i} + \Delta, & \text{if } i \notin S_t, \end{cases} \quad (4.77)$$

where  $\Delta > 0$  is the compensation coefficient for unobserved data streams. Next, the TRAS algorithm combines the top- $r$  local statistics to decide whether to raise an alarm. In other words, the stopping time of the TRAS algorithm is defined as

$$T_{\text{TRAS}} = \inf\{t > 0 : W_{t,(1)} + W_{t,(2)} + \cdots + W_{t,(r)} \geq A\}, \quad (4.78)$$

for some pre-specified parameter  $r$ . Here  $\{W_{t,(i)}\}_{i=1,\dots,p}$  are the order statistics of  $\{W_{t,i}\}_{i=1,\dots,p}$  in descending order. Regarding the sampling policy at time  $t$ , the TRAS algorithm adaptively deploys the sensors to the streams with  $q$  largest local statistics in (12) at the time instant  $t$ .

As pointed out by [1], the advantage of introducing the compensation coefficient  $\Delta$  is that no data streams will be left unobserved for a long period and thus can help the sensors localize the affected data streams quickly. It is of great importance to select a suitable  $\Delta$ . If  $\Delta$  is too large, the sampling policy will be dominated by  $\Delta$ . As a result, a sensor will be redistributed onto the stream that currently does not have a sensor deployed after each time no matter whether this stream is affected by the change or not. On the other hand, if  $\Delta$  is too small, then the system may take a longer time to reallocate sensors from the in-control streams onto the affected streams. However, in these existing research, the choices of compensation coefficients are currently based on Monte Carlo numerical studies, and it is unclear how to choose them appropriately from the theoretical point of view.

#### 4.3.2 Our Main Results

*When the compensation coefficient  $\Delta \geq O(q/(p - q))$*

Before presenting the main results, we need to make some assumptions. We assume that Kullback-Leibler information numbers are positive and finite:

$$\begin{aligned} \text{(A1): } I(f, g) &= \int \log \frac{f(x)}{g(x)} f(x) dx > 0, \\ I(g, f) &= \int \log \frac{g(x)}{f(x)} g(x) dx > 0, \end{aligned} \tag{4.79}$$

for all  $i = 1, \dots, p$ . Apart from Assumption (Equation 4.79), we need to make an additional assumption that the absolute value of log likelihood ratio  $\log \frac{g(x)}{f(x)}$  is uniformly

bounded by some constant. Mathematically,

$$(A2): \quad -\eta \leq \log \frac{g(x)}{f(x)} \leq \eta \text{ for all } x, \quad (4.80)$$

for some constant  $\eta > 0$ . Some examples of distributions include Bernoulli distribution or truncated Gaussian distribution.

To evaluate the performance of TRAS algorithm under different  $\Delta$ , we define the optimal lower bound of detection delay:

$$D_{\text{CUSUM}} = \frac{\log \gamma}{sI(g, f)} + C_0, \quad (4.81)$$

for some constant  $C_0$ , which is achieved by the standard CUSUM procedure when the true indexes of all affected data streams are pre-specified. The proof of exact optimality of CUSUM can be found in [63]. Here we define a procedure to be efficient if the corresponding detection delay when change occurs at time  $\nu = 0$  asymptotically attains the optimal lower bound of CUSUM.

The following theorem characterizes the relationship between the performance of  $\mathbb{T}_{\text{TRAS}}$  and  $\Delta$  for large  $\Delta$ .

**THEOREM 2** *When the compensation coefficient  $\Delta$  satisfies:*

$$\Delta > \frac{I(f, g)q}{p - q}, \quad (4.82)$$

*then the ARL to false alarm and the detection delay of  $\mathbb{T}_{\text{TRAS}}$  satisfy:*

$$\mathbf{E}_\infty[\mathbb{T}_{\text{TRAS}}] \leq C_1 A, \quad (4.83)$$

$$\mathbf{E}_0[\mathbb{T}_{\text{TRAS}}] \geq C_2 A, \quad (4.84)$$

*for all possible data streams  $S$  and any parameter  $r$  when the dimension  $p$  is fixed. Here*

$C_1, C_2$  are two constants that are not related to the threshold  $A$ .

The rigorous proof of Theorem Theorem 4 will be presented later. By Theorem Theorem 4, we conclude that  $\mathbb{T}_{\text{TRAS}}$  is NOT asymptotically optimal for large  $\Delta$ .

**COROLLARY 1** *Under the condition of Theorem Theorem 4, we have*

$$D(\mathbb{T}_{\text{TRAS}}) \geq \frac{C_2}{C_1} \gamma.$$

*In other words, for large  $\Delta$  in (Equation 4.82),  $\mathbb{T}_{\text{TRAS}}$  is not efficient in the sense that the detection delay is of order  $\gamma$ , which is much larger than the optimal lower bound of order  $\log \gamma$  in (Equation 4.81) as  $\gamma$  goes to infinity.*

*Proof:* To see this, let  $E_\infty[\mathbb{T}_{\text{TRAS}}] = \gamma$ , then applying relationship (Equation 4.83) obtains that  $A \geq \gamma/C_1$ . Using relationship (Equation 4.84) we have

$$D(\mathbb{T}_{\text{TRAS}}) \geq E_0[\mathbb{T}_{\text{TRAS}}] \geq \frac{C_2}{C_1} \gamma,$$

which is clearly not efficient as compared to the optimal lower bound  $O(\log \gamma)$ . □

*When the compensation coefficient  $\Delta \leq O(1/(p-1))$*

In this subsection, we focus on a special case when there is only one affected stream, say, the  $j$ -th stream, and we are only allowed to observe one stream at each time  $t$  (i.e.,  $q = s = r = 1$ ). The stopping time  $\mathbb{T}_{\text{TRAS}}$  can then be rewritten as:

$$\mathbb{T}_{\text{TRAS}} = \inf\{t > 0 : \max_{i=1, \dots, p} W_{t,i} \geq A\}. \quad (4.85)$$

When the compensation coefficient  $\Delta$  equals 0, the asymptotic optimality of  $\mathbb{T}_{\text{TRAS}}$  in (Equation 4.85) has already been established in [22]. Thus, it is natural to conjecture that when  $\Delta$  is upper bounded by some upper threshold, the TRAS algorithm might still achieve good detection delay performance.

The following theorem presents the theoretical properties of  $T_{\text{TRAS}}$  when the compensation coefficient  $\Delta \leq O(1/(p-1))$ :

**THEOREM 3** *When the compensation coefficient  $\Delta$  satisfies,*

$$\Delta < \min \left( \frac{\lambda}{p-1}, \frac{I(g, f)}{2} \right), \quad (4.86)$$

*for some constant  $\lambda > 0$ , we have:*

$$E_{\infty}[T_{\text{TRAS}}] \geq \frac{C_4 e^A}{p}. \quad (4.87)$$

*Moreover, the detection delay of the TRAS algorithm when change occurs at time  $\nu = 0$  satisfies*

$$E_0[T_{\text{TRAS}}] \leq \frac{A}{I(g, f)} + C_5, \quad (4.88)$$

*for any affected data streams  $j$ . Here  $C_4, C_5$  are some constants that do not depend on the threshold  $A$ .*

It is useful to add some remarks.

1. By letting  $E_{\infty}[T_{\text{TRAS}}] \geq \gamma$ , we have

$$E_0[T_{\text{TRAS}}] \leq \frac{\log \gamma}{I(g, f)} + O(\log p).$$

In other words, the TRAS algorithm achieves the same order of detection delay when change occurs at  $\nu = 0$  as compared to the optimal delay achieved by CUSUM procedure, under the condition that the dimension  $p$  is fixed.

2. The constant  $\lambda$  in Theorem Theorem 3 is a loose lower bound to ensure the efficient detection of the change-point. For the general sampling control  $s$  and the number  $s$

of affected data streams, we can select the compensation coefficient by

$$\Delta \leq \min\left(\frac{qI(f, g)}{2(p - q)}, \frac{I(g, f)}{2}\right). \quad (4.89)$$

3. The assumption on the uniformly bounded log likelihood ratio is to ensure that the overshoot term in the sequential test is always uniformly bounded. We conjecture that this assumption is not necessary, but have not been able to do without it in proofs.

### 4.3.3 Numerical Studies

In this section, we conduct several numerical studies to demonstrate the performance properties of  $T_{\text{TRAS}}$  under different setting. Below we consider two types of numerical examples: in subsection subsection 4.3.3 we do Monte-Carlo simulations to show the ARL to false alarm performance of  $T_{\text{TRAS}}$  for different  $\Delta$  and in subsection subsection 4.3.3 we compare the detection delay performance of  $T_{\text{TRAS}}$  for different  $\Delta$ .

#### *The study on the ARL to false alarm*

We first consider the ARL to false alarm performance of  $T_{\text{TRAS}}$  with respect to different compensation coefficients  $\Delta$ . Below are the detailed settings:

- $p = 11, q = s = r = 1$ .
- $f \sim N(0, 1), g \sim N(1, 1)$ .
- The threshold  $A$  ranges from 3 to 7.

We note that we select the dimension  $p = 11$  as the upper bound of  $\Delta$  in Theorem Theorem 4 under such case has a nice form. The upper bound of  $\Delta$  in Theorem 1 equals  $\frac{qI(f, g)}{p - q} = 0.05$ , and we consider two values of  $\Delta$  : 0.03, 0.07. As a comparision, we also consider the greedy-cyclic sampling policy proposed in [22], which corresponds to the TRAS algorithm  $\Delta = 0$ . All results are based on 50,000 Monte Carlo simulations.

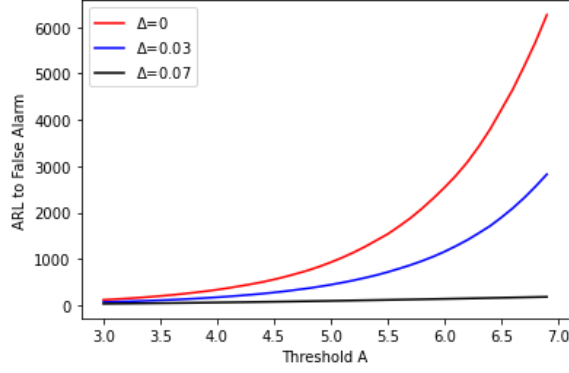


Figure 4.1: The ARL to false alarm for different  $\Delta$

In Figure Figure 4.1 we plot the ARL to false-alarm using different  $\Delta$  with respect to the same threshold  $A$ . We observe that when  $\Delta$  is small ( $\Delta = 0, 0.03$ ), the ARL to false-alarm of  $T_{\text{TRAS}}$  performs similar to the exponential function. While for large  $\Delta$  ( $\Delta = 0.07$ ), the ARL to false-alarm is much smaller than that of small  $\Delta$  and performs like a linear function. This indicates that to keep the same ARL to false alarm rate, the scenario with large  $\Delta$  such as  $\Delta = 0.07$  needs to adopt a much larger threshold  $A$  than that with small  $\Delta$  such as  $\Delta = 0, 0.03$ , and thus the corresponding detection delay performance would not be satisfying. This accords with our theoretical results in relationships (Equation 4.82) and (Equation 4.87).

#### *The study on the detection delay*

In this section, we compare the detection delay performance of  $T_{\text{TRAS}}$  with different  $\Delta$ .

Below are the detailed settings:

- $p = 11, q = s = r = 1$ .
- $f \sim N(0, 1), g \sim N(1, 1)$ .
- The change occurs to the first, the  $(\lfloor p/2 \rfloor + 1)$ -th stream, and the  $p$ -th stream.
- The ARL to false alarm  $\gamma$  ranges from 2000 to 10,000.

we first use the bisection method to find suitable threshold  $A$  to attain the false alarm constraint, and then simulate the detection delay with different compensation coefficient  $\Delta$ . All the results are based on 50,000 Monte Carlo simulations.

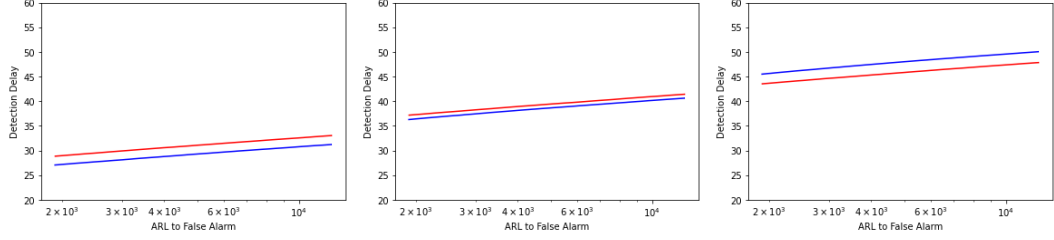


Figure 4.2: Detection delay relationship for  $\Delta = 0$  (blue) and  $\Delta = 0.03$  (red)

Table 4.4: Detection delay for  $\Delta = 0, 0.03, 0.07$  when the change occurs to Stream 1

| $\gamma$ | Delay ( $\Delta = 0$ ) | Delay ( $\Delta = 0.03$ ) | Delay ( $\Delta = 0.07$ ) |
|----------|------------------------|---------------------------|---------------------------|
| 1000     | $25.44 \pm 0.12$       | $27.18 \pm 0.11$          | $58.81 \pm 0.12$          |
| 2000     | $27.17 \pm 0.12$       | $28.85 \pm 0.11$          | $94.43 \pm 0.13$          |
| 5000     | $29.28 \pm 0.13$       | $31.70 \pm 0.11$          | $206.34 \pm 0.16$         |
| 10,000   | $30.77 \pm 0.13$       | $32.64 \pm 0.12$          | $384.15 \pm 0.20$         |

The results of the detection delay performance are summarized in Figure Figure 4.2 and Table Table 4.4. We have the following observations:

- From Figure Figure 4.2 we observe that for different affected data streams, the gap between the blue line and the red line remains as a constant as the ARL to false alarm increases from  $\gamma = 10^3$  to  $10^4$  for different affected data stream. This shows that the TRAS algorithm  $\mathbb{T}_{\text{TRAS}}$  with small  $\Delta = 0.03$  is second-order asymptotically optimal, due to the fact that the asymptotic optimality of  $\Delta = 0$  has already been established in [22].
- We can also observe in Table Table 4.4 that the detection delay of  $\mathbb{T}_{\text{TRAS}}$  with  $\Delta = 0.07$  is much larger than that of  $\Delta = 0, 0.03$ . Specifically, the delay when  $\Delta = 0.07$  grows almost linearly with the ARL to false-alarm constraint  $\gamma$ , and this accords with our results in relationship (Equation 4.83). Table Figure 4.2 shows that the TRAS algorithm with large  $\Delta$  is not asymptotically optimal and is inefficient.

#### 4.3.4 Conclusions

In this part, we study the impact of compensation coefficients of TRAS algorithm. On the one hand, if compensation coefficient  $\Delta$  is larger than  $I(f, g)q/(p - q)$ , then the TRAS algorithm is not efficient in the sense of having a larger detection delay than the optimal lower bound. On the other hand, under a special scenario of  $q = s = 1$ , if compensation coefficient  $\Delta$  is bounded by  $O(1/(p - 1))$ , the TRAS algorithm asymptotically attains the optimal lower bound of detection delay when change occurs at time  $\nu = 0$ . These results provide theoretical insights on how to numerically tune the compensation coefficient  $\Delta$  for the TRAS algorithm and its extensions.

There are a couple of future research directions on the impact of compensation coefficient on the TRAS algorithm. The first one is that we conjecture that the bound  $O(q/(p-q))$  is sharp to ensure the asymptotic optimality. We proved it for the specific scenario of  $q = s = 1$  but it remains an open problem for the general scenario of  $q, s > 1$ . The main challenge is that the analysis of switches among different data streams is highly non-trivial under the change-point setting. The second is what happens when dimension  $p$  goes to infinity simultaneously at a suitable rate with the false alarm constraint  $\gamma$ . When  $p \ll \log \gamma$  our results can still go through. However, when  $p = O(\gamma)$  or  $p \gg \gamma$ , our arguments do not work since we do not know what is the sharp lower bound of detection delay. Thus, it is unclear how to investigate the impact of compensation coefficient in the high-dimensional cases.

### 4.3.5 Proof of the Main Theorem

*Proof of ARL to false-alarm relationship (Equation 4.83)*

The key idea in the proof is to consider the sum of all  $p$  local statistics. Let us start with the definition of random variable  $Y_{t,i}$  for each time  $t$  and the  $i$ -th stream:

$$Y_{t,i} = \mathbb{1}_{\{i \notin S_t\}} \Delta + \mathbb{1}_{\{i \in S_t\}} \log \frac{g(X_{t,i})}{f(X_{t,i})},$$

The local statistics  $W_{t,i}$  can then be written as:

$$W_{t,i} = \max_{\ell=1, \dots, t} (Y_{\ell,i} + \dots + Y_{t,i}). \quad (4.90)$$

It is clear that

$$W_{t,i} = \max_{\ell=1, \dots, t} (Y_{\ell,i} + \dots + Y_{t,i}) \geq M_{t,i} = Y_{1,i} + \dots + Y_{t,i}. \quad (4.91)$$

Define a new stopping time  $\mathbb{T}_{*,1}$ :

$$\mathbb{T}_{*,1} = \inf\{t > 0 : W_{t,1} + \dots + W_{t,p} \geq pA\}. \quad (4.92)$$

Then we have:

$$\mathbb{T}_{\text{TRAS}} \leq \mathbb{T}_{*,1}. \quad (4.93)$$

This is because

$$\begin{aligned} W_{t,(1)} + \dots + W_{t,(r)} &\geq \max_{i=1, \dots, p} W_{t,i} \\ &\geq \frac{W_{t,1} + \dots + W_{t,p}}{p}. \end{aligned}$$

Define another stopping time  $\mathbb{T}_{*,2}$  :

$$\mathbb{T}_{*,2} = \inf\{t > 0 : M_{t,1} + \cdots + M_{t,p} \geq pA\}, \quad (4.94)$$

and it is clear that

$$\mathbb{E}_\infty[\mathbb{T}_{*,1}] \leq \mathbb{E}_\infty[\mathbb{T}_{*,2}], \quad (4.95)$$

as  $M_{t,i} \leq W_{t,i}$  for all  $i$  and all time instant  $t$ .

We define a random variable  $Z_t$  for each time instant  $t$ , which is the sum of log-likelihood from each observed streams and compensation coefficients from the unobservable streams:

$$Z_t = \sum_{i \in S_t} \log \frac{g(X_{t,i})}{f(X_{t,i})} + \Delta(p - q),$$

and the stopping time  $\mathbb{T}_{*,2}$  can be rewritten as

$$\mathbb{T}_{*,2} = \inf\{t > 0 : \tilde{Z}_t = Z_1 + \cdots + Z_t \geq pA\}. \quad (4.96)$$

Note that  $Z_t$  is i.i.d., and applying Wald's equation obtains that:

$$\mathbb{E}_\infty[\mathbb{T}_{*,2}] = \frac{\mathbb{E}_\infty[\tilde{Z}_{\mathbb{T}_{*,2}}]}{\mathbb{E}_\infty[Z_\ell]} \leq \frac{pA + q\eta + \Delta(p - q)}{\mathbb{E}_\infty[Z_\ell]} \leq \frac{pA}{\Delta(p - q) - qI(f, g)} + O(1),$$

where the last inequality is based on the assumptions (1)  $\Delta > \frac{qI(f, g)}{p - q}$ , (2) the log-likelihood is uniformly bounded and thus the overshoot term is bounded by  $q\eta + \Delta(p - q)$ . Relationship (Equation 4.83) is then immediately proved based on the relationship  $\mathbb{T}_{\text{TRAS}} \leq \mathbb{T}_{*,1} \leq \mathbb{T}_{*,2}$ .

□

*Proof of detection delay relationship :*

We define random variables  $B_{t,i}$  for all  $t, i$  and the sum  $B_t$  of all  $B_{t,i}$  :

$$\begin{aligned} B_{t,i} &= \max(\Delta, X_{t,i}), \\ B_t &= B_{t,1} + \cdots + B_{t,p}. \end{aligned}$$

Consider a new stopping time  $\mathbb{T}_{*,3}$  :

$$\mathbb{T}_{*,3} = \inf\{t > 0 : \tilde{B}_t = B_1 + \cdots + B_t \geq A\}. \quad (4.97)$$

Then we have:

$$\mathbb{E}_0[\mathbb{T}_{*,3}] \geq \frac{\mathbb{E}_0[\tilde{B}_{\mathbb{T}_{*,3}}]}{\mathbb{E}_0[B_t]} \geq \frac{A}{\mathbb{E}_0[B_t]}. \quad (4.98)$$

It is not difficult to see that  $W_{t,i} \leq \sum_{\ell=1}^t B_{\ell,i}$  and thus

$$W_{t,(1)} + \cdots + W_{t,(r)} \leq B_1 + \cdots + B_t. \quad (4.99)$$

We then have  $\mathbb{T}_{\text{TRAS}} \geq \mathbb{T}_{*,3}$  and

$$\mathbb{E}_0[\mathbb{T}_{\text{TRAS}}] \geq \mathbb{E}_0[\mathbb{T}_{*,3}] \geq \frac{A}{\mathbb{E}_0[B_t]}. \quad (4.100)$$

Relationship (Equation 4.83) is then immediately proved with  $C_2 = 1/\mathbb{E}_0[B_t]$ .

□

*Proof of Theorem 3:*

Let us start with one concentration inequality on uniformly bounded random variables.

**COROLLARY 2** *Suppose  $X_1, X_2, \dots, X_n$  are uniformly bounded i.i.d. random variables, i.e.,  $a \leq X_i \leq b$  for all  $i = 1, \dots, n$  and some constant  $a, b$ . Let  $S_n = X_1 + \cdots + X_n$  and*

we have:

$$\begin{aligned} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq \beta) &\leq \exp\left(-\frac{\beta^2}{2n(b-a)^2}\right), \\ \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -\beta) &\leq \exp\left(-\frac{\beta^2}{2n(b-a)^2}\right). \end{aligned} \tag{4.101}$$

*Proof:* See [76]. □

Recall that the stopping time  $\mathbb{T}_{\text{TRAS}}$  when  $q = s = r = 1$  is defined as:

$$\mathbb{T}_{\text{TRAS}} = \inf\{t > 0 : \max_{i=1, \dots, p} W_{t,i} \geq A\}. \tag{4.102}$$

To bound  $\mathbb{E}_\infty[\mathbb{T}_{\text{TRAS}}]$ , let us start with a series of stopping time

$$\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k, \dots$$

where

$$\mathcal{T}_k = \inf\{t > 0 : S_{\mathcal{T}_1 + \dots + \mathcal{T}_{k-1} + 1} \neq S_{\mathcal{T}_1 + \dots + \mathcal{T}_{k-1} + t + 1}\}.$$

In other words, each stopping time is a stochastic process on a new stream before the corresponding statistics is exceeded by some other streams. Note that our sampling set  $S_t$  is  $\mathcal{F}_{t-1}$ -adapted, and thus  $\mathcal{T}_k$  is a stopping time.

We then define a random variable  $K$ , which denotes the first time when the maximum of  $p$  local statistics during  $\mathcal{T}_1 + \dots + \mathcal{T}_{k-1} + 1$  to  $\mathcal{T}_1 + \dots + \mathcal{T}_k$  exceeds  $A$ :

$$K = \inf\{k > 0 : \max_{t=\mathcal{T}_1 + \dots + \mathcal{T}_{k-1} + 1, \dots, \mathcal{T}_1 + \dots + \mathcal{T}_k} \max_{i=1, \dots, p} W_{t,i} \geq A\}. \tag{4.103}$$

It is easy to see that:

$$\mathbb{E}_\infty[\mathbb{T}_{\text{TRAS}}] \geq \mathbb{E}_\infty[K]. \tag{4.104}$$

Similarly, for each individual stream  $i$ , we consider a series of stopping time:

$$\mathcal{T}_1^i, \dots, \mathcal{T}_k^i, \dots \quad (4.105)$$

where

$$\begin{aligned} \mathcal{T}_k^i &= \inf\{t > 0 : i \in S_{\mathcal{T}_1^i + \dots + \mathcal{T}_{k-1}^i + t + 1}\}, \text{ if } k \text{ is odd,} \\ \mathcal{T}_k^i &= \inf\{t > 0 : i \notin S_{\mathcal{T}_1^i + \dots + \mathcal{T}_{k-1}^i + t + 1}\} \text{ if } k \text{ is even,} \end{aligned}$$

That is, when  $k$  is odd,  $\mathcal{T}_k^i$  is the waiting time before the sensor switched to the Stream  $i$ , and when  $k$  is even,  $\mathcal{T}_k^i$  is a stochastic process on Stream  $i$  before the sensor switches from Stream  $i$  to some other streams, i.e., when the  $i$ -th local statistics is exceeded by some other local statistics.

Based on  $\mathcal{T}_k^i$ , we then define a random variable  $K_i$  for each  $i = 1, \dots, p$ :

$$K_i = \inf\{k > 0 : \max_{t=\mathcal{T}_1^i + \dots + \mathcal{T}_{k-1}^i + 1, \dots, \mathcal{T}_1^i + \dots + \mathcal{T}_k^i} W_{t,i} \geq A\}.$$

We observe that:

$$\{K \leq k\} \subset \cup_{i=1}^p \{K_i \leq k\},$$

and thus

$$\begin{aligned} \mathbb{P}_\infty(K < k) &\leq \mathbb{P}_\infty(\cup_{i=1}^p \{K_i < k\}) \\ &\leq \sum_{i=1}^p \mathbb{P}_\infty(K_i < k). \end{aligned}$$

If we are able to show that

$$\mathbb{P}_\infty(K_i < k) \leq (k-1)C_7 e^{-A}, \quad (4.106)$$

for all  $i$  and some constant  $C_7$  that does not depend on the threshold  $A$ , we then have

$$\begin{aligned} \mathbb{P}_\infty(K \geq k) &= 1 - \mathbb{P}_\infty(K < k) \\ &\geq 1 - p(k-1)C_7e^{-A}, \end{aligned}$$

and thus

$$\begin{aligned} \mathbb{E}_\infty[K] &= \sum_{k=1}^{\infty} \mathbb{P}_\infty(K \geq k) \\ &\geq \sum_{k=1}^{\lceil e^A/2pC_7 \rceil} (1 - p(k-1)C_7e^{-A}) \\ &\geq \sum_{k=1}^{\lceil e^A/2pC_7 \rceil} \frac{1}{2} \\ &\geq \frac{C_8e^A}{p}, \end{aligned}$$

for some suitable constant  $C_8 > 0$ . Relationship (Equation 4.87) is then immediately proved.

We now focus on the probability  $\mathbb{P}_\infty(K_i < k)$ . Note that

$$\begin{aligned} \mathbb{P}_\infty(K_i < k) &= \mathbb{P}_\infty\left(\bigcup_{w=1}^{k-1} \max_{t=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1, \dots, \mathcal{T}_1^i+\dots+\mathcal{T}_w^i} W_{t,i} \geq A\right) \\ &\leq \sum_{w=1}^{k-1} \mathbb{P}_\infty\left(\max_{t=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1, \dots, \mathcal{T}_1^i+\dots+\mathcal{T}_w^i} W_{t,i} \geq A\right). \end{aligned}$$

It now suffices to show that

$$\mathbb{P}_\infty\left(\max_{t=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1, \dots, \mathcal{T}_1^i+\dots+\mathcal{T}_w^i} W_{t,i} \geq A\right) \leq C_7e^{-A}, \quad (4.107)$$

for some constant  $C_7 > 0$  and any  $w = 1, 2, \dots$ .

When  $w$  is odd,  $W_{t,i}$  is strictly increasing (by the compensation coefficient  $\Delta$  at each

time) for  $t = \mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i + 1, \dots, \mathcal{T}_1^i + \dots + \mathcal{T}_w^i$ , and thus

$$\begin{aligned} \mathbb{P}_\infty\left(\max_{t=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1,\dots,\mathcal{T}_1^i+\dots+\mathcal{T}_w^i} W_{t,i} \geq A\right) &= \mathbb{P}_\infty(W_{\mathcal{T}_1^i+\dots+\mathcal{T}_w^i,i} \geq A) \\ &= \mathbb{P}_\infty(e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_w^i,i}} \geq e^A) \\ &\leq e^{-A} \mathbb{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_w^i,i}}]. \end{aligned}$$

When  $w$  is even, we need to consider the stochastic process occurred on the  $i$ -th stream: we keep observing from the  $i$ -th stream until  $W_{t,i}$  is exceeded by some other local statistics. For this purpose, we define the following stopping time without the impact of the remaining  $p - 1$  streams:

$$\mathbb{T}_w^{i,*} = \inf\{t > 0 : S_{t,i} = \sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+t} \log \frac{g(X_{\ell,i})}{f(X_{\ell,i})} \geq A - W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i,i}\}, \quad (4.108)$$

and it is clear that

$$\begin{aligned} \mathbb{P}_\infty\left(\max_{t=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1,\dots,\mathcal{T}_1^i+\dots+\mathcal{T}_w^i} W_{t,i} \geq A\right) &\leq \mathbb{P}_\infty(\mathbb{T}_w^{i,*} < \infty) \\ &\leq e^{-A} \mathbb{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i,i}}], \end{aligned}$$

where the last step applies the standard technique of changing measure in sequential analysis, see Page 10 in [30].

Based on the above analysis, We observe that the key is to bound  $\mathbb{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_w^i,i}}]$  for all  $w$  is odd and all  $i$ . In other words, the key is to ensure that every time the sensor switches to Stream  $i$ , the expectation of the corresponding local statistics can be upper bounded by some constants. The proof below is divided into several lemmas.

**LEMMA 38** *Recall that for any even  $w$ ,  $\mathcal{T}_w^i$  is a stochastic process on Stream  $i$  before the local statistics  $W_{t,i}$  is exceeded by the accumulating compensation coefficients from other statistics. Denote by  $W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i,(1)}$ ,  $W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i,(2)}$  the largest and the second-largest*

local statistics at time  $\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i$ , and denote by  $D_w$  the corresponding difference of these two values:

$$D_w = W_{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i, (1)} - W_{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i, (2)}.$$

We then have

$$0 \leq D_w \leq \Delta + \eta. \quad (4.109)$$

Furthermore, the stopping time  $\mathcal{T}_w^i$  can be rewritten as:

$$\mathcal{T}_w^i = \inf\{t > 0 : S_{t,i} = \sum_{\ell=\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + 1}^{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + t} \log \frac{g(X_{\ell,i})}{f(X_{\ell,i})} \leq \Delta t - D_w\}. \quad (4.110)$$

Under the condition  $\Delta < \frac{I(g,f)}{2}$ , we have:

$$\mathbb{E}_\infty[e^{S_{\mathcal{T}_w^i,i}}] \leq C_8 < 1, \quad (4.111)$$

$$\mathbb{E}_\infty[e^{S_{\mathcal{T}_w^i,i}} \mathbb{1}_{\{S_{\mathcal{T}_w^i,i} > 0\}}] \leq C_8 < 1, \quad (4.112)$$

$$S_{\mathcal{T}_w^i,i} \geq -C_9, \quad (4.113)$$

for some constants  $0 < C_8 < 1, C_9 > 0$ .

*Proof:* Applying the standard technique of changing measure obtains that

$$\begin{aligned} \mathbb{E}_\infty[e^{S_{\mathcal{T}_w^i,i}}] &= \sum_{t=1}^{\infty} \int_{\mathcal{T}_w^i=t} \prod_{\ell=\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + 1}^{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + t} \frac{g(X_{\ell,i})}{f(X_{\ell,i})} \prod_{\ell=\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + 1}^{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + t} f(X_{\ell,i}) dX \\ &= \sum_{t=1}^{\infty} \int_{\mathcal{T}_w^i=t} \prod_{\ell=\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + 1}^{\mathcal{T}_1^i + \cdots + \mathcal{T}_{w-1}^i + t} g(X_{\ell,i}) dX \\ &= \mathbb{P}_0(\mathcal{T}_w^i < \infty) \leq C_8 < 1. \end{aligned}$$

for some constants  $0 < C_8 < 1$ . Note that  $E_0[\log \frac{g(X_{\ell,i})}{f(X_{\ell,i})} - \Delta] > 0$ , and the last step is a classical conclusion in the theory of random walk with positive step size, see Lemma 2.5.1 in [35]. Relationship (Equation 4.112) is a direct application of relationship (Equation 4.111):

$$E_\infty[e^{S_{\mathcal{T}_w^i,i}} \mathbb{1}_{\{S_{\mathcal{T}_w^i,i} > 0\}}] \leq E_\infty[e^{S_{\mathcal{T}_w^i,i}}] \leq C_8.$$

For relationship (Equation 4.113), we note that the overshoot term of the sequential test is bounded by  $\Delta + \eta$  due to the fact that the log-likelihood is uniformly bounded by  $\eta$ , and we have

$$S_{\mathcal{T}_w^i,i} \geq \Delta \mathcal{T}_w^i - D_w - \text{overshoot} \geq -D_w - \text{overshoot} \geq -C_9$$

for some constant  $C_9 > 0$ . □

After characterizing  $S_{\mathcal{T}_w^i,i}$ , we are now able to characterize  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i,i}$ . Note that

$$W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i,i} = \max\{W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i} + S_{\mathcal{T}_w^i,i}, 0\}, \quad (4.114)$$

and we have the following lemma:

LEMMA 39 *For any even  $w$ ,*

$$E_\infty[e^{W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i,i}}] \leq C_8 E_\infty[e^{W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i}}] + (1 + C_8)e^{C_9}. \quad (4.115)$$

*Proof:* We divide the proof into two parts based on the value of  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i}$

1. If  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i} \geq C_9$ , then due to relationships (Equation 4.113) and (Equation 4.114), we obtain that  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i,i} = W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i} + S_{\mathcal{T}_w^i,i}$ .
2. If  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i} < C_9$ , we then have  $W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i,i} \leq W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i,i} + S_{\mathcal{T}_w^i,i} \mathbb{1}_{\{S_{\mathcal{T}_w^i,i} \geq 0\}}$ .

Combining the results obtains that

$$\begin{aligned}
& \mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i < C_9\}} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}] \\
& \leq \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i < C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mathbf{E}_\infty[e^{S_{\mathcal{T}_w^i}^i} \mathbb{1}_{\{S_{\mathcal{T}_w^i}^i \geq 0\}} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}] \\
& = \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i < C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mathbf{E}_\infty[\mathbb{1}_{\{S_{\mathcal{T}_w^i}^i < 0\}} + e^{S_{\mathcal{T}_w^i}^i} \mathbb{1}_{\{S_{\mathcal{T}_w^i}^i \geq 0\}} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}] \\
& \leq (1 + C_8) \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i < C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \\
& \leq (1 + C_8) e^{C_9},
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i \geq C_9\}} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}] \\
& = \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i \geq C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mathbf{E}_\infty[e^{S_{\mathcal{T}_w^i}^i} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}] \\
& \leq C_8 \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i \geq C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i}.
\end{aligned}$$

We then have

$$\begin{aligned}
\mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i}] & = \mathbf{E}_\infty[\mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} \mid \mathcal{F}_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}]] \\
& = \mathbf{E}_\infty[C_8 \mathbb{1}_{\{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i \geq C_9\}} e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i} - (1 + C_8) e^{C_9}] \\
& \leq C_8 \mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i}^i}] + (1 + C_8) e^{C_9}.
\end{aligned}$$

□

LEMMA 40 For any odd  $w$ , we have for  $t \geq \frac{6(\eta+\Delta)(p-1)}{I(f,g)}$

$$\mathbf{P}_\infty(\mathcal{T}_w^i > t) \leq \exp\left(-\frac{tI(f,g)^2}{32\eta^2}\right), \tag{4.116}$$

And by selecting  $\Delta < \frac{I(f,g)^2}{64\eta^2}$ , we have

$$\mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i + \dots + \mathcal{T}_w^i, i}^i}] \leq (e^{C_{10}\Delta(p-1)} + e^{-C_{11}p})\mathbf{E}_\infty[e^{W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i, i}^i}], \quad (4.117)$$

for some constants  $C_{10}, C_{11} > 0$ .

*Proof:* Consider the observed i.i.d. data sequence starting from  $\mathcal{T}_1^i + \dots + \mathcal{T}_j^i : \{t > \mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i : X_{S_t}\}$  and define the sum of log likelihood of observations:

$$S_{i,w,t}^* = \sum_{\ell=\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i + 1}^{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i + t} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})}. \quad (4.118)$$

We consider a new stopping time:

$$\mathcal{T}_w^{i,*} = \inf\{t > 0 : S_{i,w,t}^* < -3(\eta + \Delta)(p-1)\}. \quad (4.119)$$

We state that

$$\mathcal{T}_w^{i,*} \geq \mathcal{T}_w^i.$$

Let's say on the contrary  $\mathcal{T}_w^{i,*} < \mathcal{T}_w^i$ . In other words, when  $t = \mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i + \mathcal{T}_w^{i,*}$ , there still exists a stream  $j$ , such that  $W_{t,j} > W_{t,i}$ . We have the following observations:

1.

$$W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i, i} - W_{\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i, (1)} \geq -(\Delta + \eta). \quad (4.120)$$

In other words, the maximum difference between the  $i$ -th local statistics and the other statistics after switching from stream  $i$  to other stream is upper bounded by  $\eta + \Delta$ . Again, this is due to the fact that the log-likelihood ratio is uniformly bounded and thus the overshoot term in the sequential test is upper bounded.

2. If  $\mathcal{T}_w^{i,*} < \mathcal{T}_w^i$ , then for any Stream  $n \neq i$ ,

$$\sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+\mathcal{T}_w^{i,*}} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})} \mathbb{1}_{\{n \in S_\ell\}} \geq -2.5(\eta + \Delta).$$

Let's say for contradictory the above relationship does not hold for some  $n$ . Then we can find an earlier time point  $t < \mathcal{T}_w^{i,*}$  such that

$$\sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+t} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})} \mathbb{1}_{\{n \in S_\ell\}} \leq -1.5(\eta + \Delta).$$

This is because the log-likelihood ratio is strictly bounded. Then this contradicts due to the fact that after this time point  $\mathcal{T}_1^i + \dots + \mathcal{T}_{w-1}^i + t$ , the  $i$ -th local statistics will be greater than the  $n$ -th local statistics (recall relationship (Equation 4.120) and the sensor will not switch to stream  $n$  anymore before visiting stream  $i$ .

We immediately obtains that the above observation 2 contradicts the construction of  $\mathcal{T}_w^{i,*}$  and thus we obtain that  $\mathcal{T}_w^{i,*} \geq \mathcal{T}_w^i$ . For  $t \geq \frac{6(\eta+\Delta)(p-1)}{I(f,g)}$ , applying the concentration inequalities in Corollary 2:

$$\begin{aligned} \mathbb{P}_\infty(\mathcal{T}_w^i > t) &\leq \mathbb{P}_\infty(\mathcal{T}_w^{i,*} > t) \\ &\leq \mathbb{P}_\infty\left(\sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+t} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})} \geq -3(\eta + \Delta)(p - 1)\right) \\ &\leq \mathbb{P}_\infty\left(\sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+t} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})} + tI(f, g) \geq -3(\eta + \Delta)(p - 1) + tI(f, g)\right) \\ &\leq \mathbb{P}_\infty\left(\sum_{\ell=\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+1}^{\mathcal{T}_1^i+\dots+\mathcal{T}_{w-1}^i+t} \log \frac{g(X_{S_\ell})}{f(X_{S_\ell})} + tI(f, g) \geq \frac{tI(f, g)}{2}\right) \\ &\leq \exp\left(-\frac{tI(f, g)^2}{32\eta^2}\right). \end{aligned}$$

By selecting  $\Delta < \frac{I(f,g)^2}{64\eta^2}$ , we have

$$\begin{aligned}\mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_w^i}^i}] &= \mathbf{E}_\infty[\mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_w^i}^i} \mid \mathcal{F}_{\tau_1^i+\dots+\tau_{w-1}^i}]] \\ &= \mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_{w-1}^i}^i} \mathbf{E}_\infty[e^{\Delta\tau_w^i} \mid \mathcal{F}_{\tau_1^i+\dots+\tau_{w-1}^i}]].\end{aligned}$$

For the term  $\mathbf{E}_\infty[e^{\Delta\tau_w^i} \mid \mathcal{F}_{\tau_1^i+\dots+\tau_{w-1}^i}]$ ,

$$\begin{aligned}\mathbf{E}_\infty[e^{\Delta\tau_w^i} \mid \mathcal{F}_{\tau_1^i+\dots+\tau_{w-1}^i}] &\leq e^{\Delta\frac{6(\eta+\Delta)(p-1)}{I(f,g)}} + \sum_{t=\frac{6(\eta+\Delta)(p-1)}{I(f,g)}+1}^{\infty} e^{\Delta t} \exp\left(-\frac{tI(f,g)^2}{32\eta^2}\right) \\ &\leq e^{\Delta\frac{6(\eta+\Delta)(p-1)}{I(f,g)}} + \sum_{t=\frac{6(\eta+\Delta)(p-1)}{I(f,g)}+1}^{\infty} \exp\left(-\frac{tI(f,g)^2}{64\eta^2}\right) \\ &\leq e^{C_{10}\Delta(p-1)} + e^{-C_{11}p},\end{aligned}$$

for some constants  $C_{10}, C_{11} > 0$ . We then have

$$\begin{aligned}\mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_w^i}^i}] &\leq \mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_{w-1}^i}^i} \mathbf{E}_\infty[e^{\Delta\tau_w^i} \mid \mathcal{F}_{\tau_1^i+\dots+\tau_{w-1}^i}]] \\ &\leq (e^{C_{10}\Delta(p-1)} + e^{-C_{11}p}) \mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_{w-1}^i}^i}].\end{aligned}$$

□

Combining Lemma Lemma 39 and Lemma Lemma 40, we conclude that for any odd  $w$ ,

$$\mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_w^i}^i}] \leq C_8(e^{C_{10}\Delta(p-1)} + e^{-C_{11}p}) \mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_{w-2}^i}^i}] + (1+C_8)e^{C_9}(e^{C_{10}\Delta(p-1)} + e^{-C_{11}p}).$$

By selecting  $\Delta < \frac{\lambda}{p-1}$  for some constant  $\lambda > 0$  such that  $C_8(e^{C_{10}\Delta(p-1)} + e^{-C_{11}p}) < 1$ , we obtain that there exists a constant  $0 < C_{12} < 1, C_{13}$  such that:

$$\mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_w^i}^i}] \leq C_{12} \mathbf{E}_\infty[e^{W_{\tau_1^i+\dots+\tau_{w-2}^i}^i}] + C_{13}, \quad (4.121)$$

Based on relationship (Equation 4.121), we have that

$$E_{\infty}[e^{W_{\tau_1^i + \dots + \tau_w^i, i}}] \leq C_{14},$$

for some constant  $C_{14} > 0$  and any odd  $w$  due to the fact that  $C_{12} < 1$  and thus the sequence of expectations converge. We then complete the proof of relationship (Equation 4.107) and the proof of ARL to false alarm relationship (Equation 4.87).

□

*Proof of detection delay relationship (Equation 4.88):*

Without loss of generality, we assume that the change occurs to the  $j$ -th stream. Recall that under the scenario  $q = s = r = 1$ , the stopping time  $T_{\text{TRAS}}$  is defined as:

$$T_{\text{TRAS}} = \inf\{t > 0 : \max_{i=1, \dots, p} W_{t,i} \geq A\}, \quad (4.122)$$

Following the idea in the proof of ARL to false alarm relationship (Equation 4.87), we consider the individual stopping time  $T_0^j$  for stream  $j$  that will only raise an alarm when the sensor points to stream  $j$ :

$$T_0^j = \inf\{t > 0 : j \in S_t \text{ and } W_{t,j} \geq A\}. \quad (4.123)$$

Then it is clear that:

$$E_0[T_{\text{TRAS}}] \leq E_0[T_0^j]. \quad (4.124)$$

When change occurs at time  $\nu = 0$ , we still consider a series of stopping time:

$$z_1^j, T_1^j, z_2^j, T_2^j, \dots, z_k^j, T_k^j, \dots, \quad (4.125)$$

where  $z_k^j, \mathcal{T}_k^j$  are mathematically defined as:

$$\begin{aligned}
z_k^j &= \inf\{t > 0 : j \in S_{z_1^j + \mathcal{T}_1^j + \dots + z_{k-1}^j + \mathcal{T}_{k-1}^j + t + 1}\}, \\
\mathcal{T}_k^j &= \min(\mathcal{T}_{k,1}^j, \mathcal{T}_{k,2}^j), \\
S_t &= \sum_{\ell=z_1^j + \mathcal{T}_1^j + \dots + z_k^j + 1}^{z_1^j + \mathcal{T}_1^j + \dots + z_k^j + t} \log \frac{g(X_{\ell,j})}{f(X_{\ell,j})} \\
\mathcal{T}_{k,1}^j &= \inf\{t > 0 : S_t \geq A - W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, j}\} \\
\mathcal{T}_{k,2}^j &= \inf\{t > 0 : S_t - \Delta t \leq W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, (2)} - W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, j}\},
\end{aligned}$$

That is,  $z_k^j$  is the waiting time before the sensor switches to the affected Stream  $j$ , and  $\mathcal{T}_k^j$  is a stochastic process on the affected stream  $j$  (either the local statistics  $W_{t,j}$  exceeds  $A$  and raising a global alarm ( $\mathcal{T}_{k,1}^j$ ), or the  $j$ -th local statistics is exceeded by another local statistics ( $\mathcal{T}_{k,2}^j$ )). Here  $W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, (2)}$  is second largest local statistics after switching to stream  $j$ .

Define  $K_*$  as the first time when  $\mathcal{T}_k^j = \mathcal{T}_{k,1}^j$ , and the ARL of  $\mathbb{T}_0^j$  can written as:

$$\begin{aligned}
\mathbb{E}_0[\mathbb{T}_0^j] &= \mathbb{E}_0[z_1^j + \mathcal{T}_1^j + z_2^j + \mathcal{T}_2^j + \dots + z_{K_*}^j + \mathcal{T}_{K_*}^j] \\
&= \sum_{k=1}^{\infty} \mathbb{E}_0[(z_k^j + \mathcal{T}_k^j) \mathbb{1}_{\{K_* \geq k\}}] \\
&\leq \sum_{k=1}^{\infty} (\mathbb{E}_0[(z_k^j)^2] \mathbb{P}_0(K_* \geq k))^{1/2} + \sum_{k=1}^{\infty} \mathbb{E}_0[\mathcal{T}_k^j \mathbb{1}_{\{K_* \geq k\}}].
\end{aligned}$$

We have the following lemma on  $\mathcal{T}_k^j$  :

LEMMA 41 *The stopping time  $\mathcal{T}_k^j$  satisfies:*

$$\mathbb{E}_0[\mathcal{T}_k^j \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}] = (A + \eta + \Delta) \frac{\mathbb{E}_0[\mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} + C_{15}, \quad (4.126)$$

where  $C_{15}$  is a constant that does not depend on  $A$  and  $\mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_{k-1}^j}$ .

*Proof:*

$$\begin{aligned}
\mathbb{E}_0[\mathcal{T}_k^j \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}] &= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} + S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,2}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&= B + C
\end{aligned}$$

where

$$\begin{aligned}
B &= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)}, \\
C &= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,2}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)}.
\end{aligned} \tag{4.127}$$

Note that the log-likelihood ratio is uniformly bounded, and so is the overshoot term in sequential tests. We then have

$$\begin{aligned}
B &= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&\leq (A + \eta + \Delta) \frac{\mathbb{E}_0[\mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)}.
\end{aligned}$$

Note that  $0 \geq W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, (2)} - W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, j} \geq -(\eta + \Delta)$  due to the fact that the log-likelihood ratio is uniformly bounded. Regarding the term  $C$ , we have

$$\begin{aligned}
C &= \frac{\mathbb{E}_0[S_{\mathcal{T}_k^j} \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,2}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&\leq \frac{\mathbb{E}_0[\Delta \mathcal{T}_k^j \mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,2}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&\leq \Delta \frac{\mathbb{E}_0[\mathcal{T}_{k,2}^j \mathbb{1}_{\{\mathcal{T}_{k,2}^j < \infty\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \\
&= \Delta \frac{\sum_{t=1}^{\infty} t \mathbb{P}_0[\mathcal{T}_{k,2}^j = t \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)}
\end{aligned}$$

Note that for any  $t \geq 1$ ,

$$\begin{aligned}
\mathbb{P}_0(\mathcal{T}_{k,2}^j = t \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}) &\leq \mathbb{P}_0(S_t - \Delta t \leq W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, (2)} - W_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j, j} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}) \\
&\leq \mathbb{P}_0(S_t - \Delta t \leq 0) \\
&= \exp(-C_{16}t),
\end{aligned}$$

where the last step results from the concentration inequality and  $C_{16}$  is some constant that does not depend on  $\mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}$ . Thus we have:

$$C \leq \Delta \frac{\sum_{t=1}^{\infty} t \mathbb{P}_0[\mathcal{T}_{k,2}^j = t \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} \leq C_{17}.$$

for some constant  $C_{17}$  that does not depend on  $\mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}$ . Combining the results above we conclude that:

$$\mathbb{E}_0[\mathcal{T}_k^j \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}] = (A + \eta + \Delta) \frac{\mathbb{E}_0[\mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]}{I(g, f)} + C_{15},$$

where  $C_{15}$  is a constant that does not depend on  $A$  and  $\mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_{k-1}^j}$ . □

Based on Lemma Lemma 41, we obtain that

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}_0[\mathcal{T}_k^j \mathbb{1}_{\{K_* \geq k\}}] &= \sum_{k=1}^{\infty} \mathbb{E}_0[\mathbb{E}_0[\mathcal{T}_k^j \mathbb{1}_{\{K_* \geq k\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]] \\
&\leq \frac{A + \eta + \Delta}{I(g, f)} \sum_{k=1}^{\infty} \mathbb{E}_0[\mathbb{E}_0[\mathbb{1}_{\{\mathcal{T}_k^j = \mathcal{T}_{k,1}^j\}} \mathbb{1}_{\{K_* \geq k\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]] \\
&\quad + C_{15} \sum_{k=1}^{\infty} \mathbb{E}_0[\mathbb{E}_0[\mathbb{1}_{\{K_* \geq k\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]] \\
&= \frac{A + \eta + \Delta}{I(g, f)} \sum_{k=1}^{\infty} \mathbb{E}_0[\mathbb{E}_0[\mathbb{1}_{\{K_* = k\}} \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}]] + C_{15} \mathbb{E}_0[K_*] \\
&= \frac{A + \eta + \Delta}{I(g, f)} + C_{15} \mathbb{E}_0[K_*]
\end{aligned}$$

LEMMA 42

$$\mathbb{E}_0[K_*] \leq C_{18}, \quad (4.128)$$

for some constants  $C_{18}$  that does not depend on the threshold  $A$ .

*Proof:*

$$\begin{aligned} \mathbb{P}_0(\mathcal{T}_{k,1}^j = \mathcal{T}_k^j \mid \mathcal{F}_{z_1^j + \mathcal{T}_1^j + \dots + z_k^j}) &\geq \mathbb{P}_0(S_t > \Delta t \text{ for all } t) \\ &\geq C_{19} \end{aligned}$$

for some constant  $C_{19} > 0$  that only depends on the distributions if  $\Delta < I(g, f)$ , which is a classical conclusion in random walk theory with positive step size. Thus we have:

$$\mathbb{P}_0(K_* \geq k) \leq (1 - C_{19})^{k-1}, \quad (4.129)$$

and

$$\mathbb{E}_0[K_*] \leq \frac{1}{C_{19}}. \quad (4.130)$$

Relationship (Equation 4.128) then holds with  $C_{18} = 1/C_{19}$ .  $\square$

LEMMA 43

$$\mathbb{E}_0[(z_k^j)^2] \leq C_{20}, \quad (4.131)$$

for some constant  $C_{20}$  that only depends on the threshold  $A$ .

*Proof:* Similar proof as in Lemma Lemma 40.  $\square$

Combining the results above,

$$\begin{aligned} \mathbb{E}_0[\mathbb{T}_0^j] &\leq \sum_{k=1}^{\infty} (\mathbb{E}_0[(z_k^j)^2] \mathbb{P}_0(K_* \geq k))^{1/2} + \sum_{k=1}^{\infty} \mathbb{E}_0[\mathcal{T}_k^j \mathbb{1}_{\{K_* \geq k\}}] \\ &\leq \frac{A}{I(g, f)} + C_5, \end{aligned}$$

for some constant  $C_5$  that does not depend on the threshold  $A$ . This completes the proof of detection delay relationship in (Equation 4.88).  $\square$

## CHAPTER 5

### CONCLUSIONS AND FUTURE RESEARCH

#### 5.1 Conclusions

This dissertation focuses on making comprehensive progress on methodology, theory, and application of active quickest detection problem in multi-stream data under the sampling or resource constraints. Our specific research aims are to design new algorithms with theoretical guarantees and develop an asymptotic optimality theory to characterize sharp information bound.

Chapter 2 investigated the active quickest detection problem under a specific scenario when there is only one affected local stream and we are only able to sample from one stream at a time. Under the assumption that the post-change distributions are pre-specified, Chapter 2 shows that a simple myopic sampling scheme which always samples the streams that are most likely to contain a local change is second-order asymptotically optimum when the number  $p$  of local streams is fixed. To the best of our knowledge, a second-order asymptotically optimum result is proved for the first time for the active quickest detection problem in multi-stream data.

Chapter 3 tackles the problem of detecting changes in high-dimensional linear regression models - one of the most fundamental predictive models. By treating linear regression models as two-layers neural networks, we develop efficient implicit regularization-based quickest detection for linear regression models, and provide theoretical guarantees on the average run length to false alarm and detection delay. Our approach can be easily extended to the context of monitoring other advanced predictive machine learning or artificial intelligent algorithms for future research directions.

Chapter 4 discusses three research topics under the setting of active quickest detection

problem. In Chapter 4.1, we consider the scenario when the post-change distributions involve unknown parameters. An efficient greedy-cyclic-sampling-based quickest detection algorithm is proposed, and we show that our proposed algorithm is first-order asymptotically optimal in the sense of minimizing the detection delay under both false alarm and sampling control constraints. In Chapter 4.2, we extend our results of  $r = s = 1$  in Chapter 1 to a more complicated scenario when  $r = s = 2$ . We combine three ideas to develop efficient active quickest detection algorithms: (1) win-stay, lose-switch sampling strategy; (2) local CUSUM statistics for local monitoring; and (3) the SUM-Shrinkage technique to fuse local statistics into a global decision. Theoretical guarantees on the ARL to false alarm and detection delay are provided. In Chapter 4.3, we revisited the TRAS algorithm proposed in [1] which incorporates a novel idea of compensation coefficients for unobserved data streams. However, it is unclear how to choose the compensation coefficients suitably from theoretical point of view so as to balance the trade-off between the detection delay and false alarm. In this context, we investigate the impact of compensation coefficients on TRAS algorithm. Our main contributions are two-folded. On the one hand, under the general setting, we prove that if compensation coefficient is larger than  $O(q/(p - q))$ , then the TRAS algorithm is not efficient in the sense of having a larger detection delay than the optimal lower bound. On the other hand, under the special case of  $q = s = 1$ , if compensation coefficient is smaller than  $O(1/(p - 1))$ , then the TRAS algorithm is efficient.

## 5.2 Future Research

The active sequential change-point analysis of multi-stream data is an important yet challenging research problem. In this dissertation, we have made efforts to tackle some of the challenges. However, there are still many interesting research opportunities. Some of which are summarized in what follows.

1. The first direction is to extend the myopic sampling policy proposed in Chapter 2 to the scenario of general  $q, s > 1$ . One conjecture is that by always sampling  $q$

streams with  $q$  largest local statistics, we are able to establish second-order asymptotic optimality results under the special case when  $q = s$ . The main challenge during the proof is how to characterize and monitor the switch among different streams. For example, from the view point of information theory, one necessary condition of achieving optimality is to capture information from all affected streams, and thus the time point when all affected streams are monitored by one sensor is crucial for the analysis of detection delay. However, the characterization of such stopping time is highly non-trivial and currently we are unable to provide the rigorous proof.

2. The second direction is to extend the methodology proposed in Chapter 3 and Chapter 4.1 to the more complicated models with unknown parameters. One may consider the scenario when the unknown parameter  $\theta$  is multi-dimensional. For example, for multinomial distribution with  $p$  features, by incorporating a simultaneous estimation of all  $p$  features with the standard CUSUM statistics, the asymptotic optimality can be obtained. In addition, one can also consider the detection of potential change in some machine learning models like neural network. While classical methods such as LASSO highly relies on the structure of linear regression, we believe that implicit regularization methods have more potential to play an important role in the future research of active quickest detection problems.
3. The third direction is to consider a new scenario when the number of sensors is greater than the number of affected streams, i.e.,  $q \geq s$ . Under such setting, one might be interested in how to monitor streams so as to capture the local stream that involves those most severe changes. One possible method is to randomly switch to some those unobserved streams after some pre-determined time-interval and switch to them if the observed data reveals a more severe change.

## REFERENCES

- [1] K. Liu, Y. Mei, and J. Shi, “An adaptive sampling strategy for online high-dimensional process monitoring,” *Technometrics*, vol. 57, no. 3, pp. 305–319, 2015.
- [2] H. B. Lim, V. T. Lam, M. C. Foo, and Y. Zeng, “An adaptive distributed resource allocation scheme for sensor networks,” in *Mobile Ad-hoc and Sensor Networks: Second International Conference, MSN 2006, Hong Kong, China, December 13-15, 2006. Proceedings 2*, Springer, 2006, pp. 770–781.
- [3] M. Zoghi and M. H. Kahaei, “Adaptive sensor selection in wireless sensor networks for target tracking,” *IET Signal Processing*, vol. 4, no. 5, pp. 530–536, 2010.
- [4] E. Kagan and I. Ben-Gal, *Probabilistic Search for Tracking Targets: Theory and Modern Application*. Wiley Online Library, 2013.
- [5] K. Cohen and Q. Zhao, “Active hypothesis testing for anomaly detection,” *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1432–1450, 2015.
- [6] B. Huang, K. Cohen, and Q. Zhao, “Active anomaly detection in heterogeneous processes,” *IEEE Transactions on information theory*, vol. 65, no. 4, pp. 2284–2301, 2018.
- [7] M. Naghshvar and T. Javidi, “Active sequential hypothesis testing,” 2013.
- [8] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, “Controlled sensing for multihypothesis testing,” *IEEE Transactions on automatic control*, vol. 58, no. 10, pp. 2451–2464, 2013.
- [9] Y. Wang, K. Paynabar, and Y. Mei, “Thresholded multivariate principal component analysis for multi-channel profile monitoring,” *arXiv preprint arXiv:1603.05265*, 2016.
- [10] R. Zhang, Y. Mei, and J. Shi, “Wavelet-based profile monitoring using order-thresholding recursive cusum schemes,” *New Frontiers of Biostatistics and Bioinformatics*, pp. 141–159, 2018.
- [11] J. Patterson and A. Gibson, *Deep learning: A practitioner’s approach*. ” O’Reilly Media, Inc.”, 2017.
- [12] A. O. Hero and D. Cochran, “Sensor management: Past, present, and future,” *IEEE Sensors Journal*, vol. 11, no. 12, pp. 3064–3075, 2011.

- [13] Q. Huang and J. Shi, "Stream of variation modeling and analysis of serial-parallel multistage manufacturing systems," *J. Manuf. Sci. Eng.*, vol. 126, no. 3, pp. 611–618, 2004.
- [14] J. R. Frost and L. D. Stone, "Review of search theory: Advances and applications to search and rescue decision support," 2001.
- [15] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.
- [16] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," in *2013 Information Theory and Applications Workshop (ITA)*, IEEE, 2013, pp. 1–20.
- [17] H. P. Chan, "Optimal sequential detection in multi-stream data," *The Annals of Statistics*, pp. 2736–63, 2017.
- [18] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, 2015.
- [19] R. Zhang and Y. Mei, "Asymptotic statistical properties of communication-efficient quickest detection schemes in sensor networks," *Sequential Analysis*, vol. 37, no. 3, pp. 375–396, 2018.
- [20] R. Zhang, J. Wang, and Y. Mei, "Search for evergreens in science: A functional data analysis," *Journal of Informetrics*, vol. 11, no. 3, pp. 629–644, 2017.
- [21] A. Shiryaev, "On the detection of disorder in a manufacturing process. i.," *Theory of Probability & Its Applications*, vol. 8, no. 3, pp. 247–265, 1963.
- [22] Q. Xu, Y. Mei, and G. V. Moustakides, "Optimum multi-stream sequential change-point detection with sampling control," *IEEE Transactions on Information Theory*, vol. 67, no. 11, pp. 7627–7636, 2021.
- [23] Q. Xu and Y. Mei, "Asymptotic optimality theory for active quickest detection with unknown postchange parameters," *Sequential Analysis*, pp. 1–32, 2023.
- [24] V. V. Veeravalli, G. Fellouris, and G. V. Moustakides, "Quickest change detection with controlled sensing," *IEEE Journal on Selected Areas in Information Theory*, 2024.
- [25] J. Li and J. ( Jin, "Optimal sensor allocation by integrating causal models and set-covering algorithms," *IIE Transactions*, vol. 42, no. 8, pp. 564–576, 2010.

- [26] Z. Li and P. Qiu, “Statistical process control using a dynamic sampling scheme,” *Technometrics*, vol. 56, no. 3, pp. 325–335, 2014.
- [27] K. Liu, X. Zhang, and J. Shi, “Adaptive sensor allocation strategy for process monitoring and diagnosis in a bayesian network,” *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 452–462, 2014.
- [28] S. Arora, N. Cohen, W. Hu, and Y. Luo, “Implicit regularization in deep matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [29] S. L. Smith, B. Dherin, D. G. Barrett, and S. De, “On the origin of implicit regularization in stochastic gradient descent,” *arXiv preprint arXiv:2101.12176*, 2021.
- [30] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 2013.
- [31] M. Woodroffe, *Nonlinear renewal theory in sequential analysis*. SIAM, 1982.
- [32] B. Yakir, *Extremes in random fields: a theory and its applications*. John Wiley & Sons, 2013.
- [33] H. V. Poor and O. Hadjiladis, *Quickest Detection*. Cambridge University Press, 2008.
- [34] V. V. Veeravalli and T. Banerjee, “Quickest change detection,” in *Academic press library in signal processing*, vol. 3, Elsevier, 2014, pp. 209–255.
- [35] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC press, 2014.
- [36] T. Banerjee and V. V. Veeravalli, “Data-efficient quickest change detection in minimax settings,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6917–6931, 2013.
- [37] H. Chernoff, *Sequential design of experiments*. Springer, 1992.
- [38] S. A. Bessler, “Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments: Part i—theory,” *Appl. Math. Statist. Lab., Stanford Univ., Stanford, CA, USA, Tech. Rep*, vol. 55, 1960.
- [39] S. A. Bessler, “Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments—part ii: Applications,” *Technical Report 56*, 1960.

- [40] A. E. Albert, “The sequential design of experiments for infinitely many states of nature,” *The Annals of Mathematical Statistics*, vol. 32, no. 3, pp. 774–799, 1961.
- [41] J. Kiefer and J. Sacks, “Asymptotically optimum sequential inference and design,” *The Annals of Mathematical Statistics*, pp. 705–750, 1963.
- [42] S. Lalley and G. Lorden, “A control problem arising in the sequential design of experiments,” *The Annals of Probability*, pp. 136–172, 1986.
- [43] R. Keener, “Second order efficiency in the sequential design of experiments,” *The Annals of Statistics*, pp. 510–532, 1984.
- [44] S. Nitinawarat and V. V. Veeravalli, “Controlled sensing for sequential multihypothesis testing with controlled markovian observations and non-uniform control cost,” *Sequential Analysis*, vol. 34, no. 1, pp. 1–24, 2015.
- [45] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, “Sequential controlled sensing for composite multihypothesis testing,” *Sequential Analysis*, vol. 40, no. 2, pp. 259–289, 2021.
- [46] K. Cohen and Q. Zhao, “Asymptotically optimal anomaly detection via sequential testing,” *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2929–2941, 2015.
- [47] A. Gurevich, K. Cohen, and Q. Zhao, “Sequential anomaly detection under a nonlinear system cost,” *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3689–3703, 2019.
- [48] B. Hemo, T. Gafni, K. Cohen, and Q. Zhao, “Searching for anomalies over composite hypotheses,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 1181–1196, 2020.
- [49] N. K. Vaidhiyan and R. Sundaresan, “Learning to detect an oddball target,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 831–852, 2017.
- [50] A. Tsopelakos, G. Fellouris, and V. V. Veeravalli, “Sequential anomaly detection with observation control,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019, pp. 2389–2393.
- [51] H. Robbins, “Some aspects of the sequential design of experiments,” 1952.
- [52] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

- [53] T. L. Lai, “Adaptive treatment allocation and the multi-armed bandit problem,” *The annals of statistics*, pp. 1091–1114, 1987.
- [54] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Conference on learning theory, JMLR Workshop and Conference Proceedings*, 2012, pp. 39–1.
- [55] S. Bubeck, N. Cesa-Bianchi, *et al.*, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.
- [56] G. Burtini, J. Loeppky, and R. Lawrence, “A survey of online experiment design with the stochastic multi-armed bandit,” *arXiv preprint arXiv:1510.00757*, 2015.
- [57] N. Cesa-Bianchi and G. Lugosi, “Combinatorial bandits,” *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.
- [58] L. Tang, R. Rosales, A. Singh, and D. Agarwal, “Automatic ad format selection via contextual bandits,” in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 1587–1594.
- [59] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, “Quickest search over multiple sequences,” *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5375–5386, 2011.
- [60] A. N. Shiryaev, “Quickest detection problems: Fifty years later,” *Sequential Analysis*, vol. 29, no. 4, pp. 345–385, 2010.
- [61] G. Lorden, “Procedures for reacting to a change in distribution,” *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.
- [62] E. S. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
- [63] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.
- [64] G. Lorden, “On excess over the boundary,” *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 520–527, 1970.
- [65] A. Wald, “Sequential tests of statistical hypotheses,” in *Breakthroughs in statistics: Foundations and basic theory*, Springer, 1992, pp. 256–298.
- [66] Y. Mei, “Sequential change-point detection when unknown parameters are present in the pre-change distribution,” 2006.

- [67] A. Chaudhuri, G. Fellouris, and A. Tajer, “Round robin active sequential change detection for dependent multi-channel data,” *arXiv preprint arXiv:2403.16297*, 2024.
- [68] L. Lambert, *Zillow’s 6 billion home flipping business was a disaster. now, a cooling housing market could foil its comeback plan*, <https://fortune.com/2022/06/02/zillow-6-billion-home-flipping-business-housing-market-fortune-500/>, Last accessed on 2022-06-02, 2022.
- [69] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.
- [70] C. Zou and P. Qiu, “Multivariate statistical process control using lasso,” *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1586–1596, 2009.
- [71] G. Ciuperca, “Real time change-point detection in a model by adaptive lasso and cusum,” *Journal de la Société Française de Statistique*, vol. 156, no. 4, pp. 113–132, 2015.
- [72] M. Pollak, “Optimal detection of a change in distribution,” *The Annals of Statistics*, pp. 206–227, 1985.
- [73] P. Zhao, Y. Yang, and Q.-C. He, “High-dimensional linear regression via implicit regularization,” *Biometrika*, vol. 109, no. 4, pp. 1033–1046, 2022.
- [74] T. Vaskevicius, V. Kanade, and P. Rebeschini, “Implicit regularization for optimal sparse recovery,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [75] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive approximation*, vol. 28, pp. 253–263, 2008.
- [76] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge university press, 2018, vol. 47.
- [77] G. Lorden and M. Pollak, “Sequential change-point detection procedures that are nearly optimal and computationally simple,” *Sequential Analysis*, vol. 27, no. 4, pp. 476–512, 2008.
- [78] X. Xian, A. Wang, and K. Liu, “A nonparametric adaptive sampling strategy for online monitoring of big data streams,” *Technometrics*, vol. 60, no. 1, pp. 14–25, 2018.

- [79] H. Ye and K. Liu, “A generic online nonparametric monitoring and sampling strategy for high-dimensional heterogeneous processes,” *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 1503–1516, 2022.
- [80] H. Ye, Z. Zheng, J.-R. C. Cheng, B. Hable, and K. Liu, “Online monitoring of high-dimensional asynchronous and heterogeneous data streams for shifts in location and scale,” *International Journal of Production Research*, vol. 62, no. 3, pp. 720–736, 2024.

## VITA

Qunzhi Xu was born in Bengbu, Anhui Province, China in August 1997. He received the Bachelor's degree in Statistics from University of Science and Technology of China in 2019. Afterwards, he joined the Ph.D. program in The H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology, under supervision of Prof. Yajun Mei and Prof. Jianjun Shi. With a fulfilling period and many stimulating experiences, he has completed his Ph.D. studies and now he is ready for new adventures. These will take him to New York University, where he will be a postdoc in the School of Global Public Health.